Lecture Notes in Networks and Systems 1609

G. Ranganathan
George A. Papakostas
Alvaro Rocha *Editors*

Inventive Communication and Computational Technologies

Proceedings of ICICCT 2025, Volume 2

Springer

Lecture Notes in Networks and Systems

Volume 1609

Series Editor

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Advisory Editors

Fernando Gomide, Department of Computer Engineering and Automation—DCA, School of Electrical and Computer Engineering—FEEC, University of Campinas—UNICAMP, São Paulo, Brazil

Okyay Kaynak, Department of Electrical and Electronics Engineering, Bogazici University, Istanbul, Türkiye

Derong Liu, Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, USA
Institute of Automation, Chinese Academy of Sciences, Beijing, China

Witold Pedrycz, Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada
Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Marios M. Polycarpou, Department of Electrical and Computer Engineering, KIOS Research Center for Intelligent Systems and Networks, University of Cyprus, Nicosia, Cyprus

Imre J. Rudas, Óbuda University, Budapest, Hungary

Jun Wang, Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

The series "Lecture Notes in Networks and Systems" publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

G. Ranganathan · George A. Papakostas ·
Alvaro Rocha
Editors

Inventive Communication and Computational Technologies

Proceedings of ICICCT 2025, Volume 2

Springer

Editors
G. Ranganathan
Sree Sakthi Engineering College
Coimbatore, Tamil Nadu, India

Alvaro Rocha
Lisbon, Portugal

George A. Papakostas
Department of Informatics
Democritus University of Thrace
Kavala, Greece

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-3-032-04311-5 ISBN 978-3-032-04312-2 (eBook)
https://doi.org/10.1007/978-3-032-04312-2

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

With a great deal of dedication, this volume is dedicated to all the authors, scientists, academicians, young researchers, conference delegates, and students, who have research expertise on ICT technologies from all over the world.

Preface

We are happy to present the volume 2 of the book *Inventive Communication and Computational Technologies*. This book includes a set of the state-of-the-art research studies and innovations in the fields of signals and communication systems, smart grids, and autonomous systems.

We have received an overwhelming response for the call for papers and received a wide range high-quality research contributions from different countries. By following a double-blinded peer-review process, 39 manuscripts were selected for publication in volume 2. The manuscripts were focused on different fields such as robotics and autonomous systems, optimization frameworks, wireless networks, signals and systems, smart grids, unmanned systems, antenna, etc. Every submission included in this book contributes to the broad area of "Inventive Communication Technologies". The results discussed in the various chapters of this volume will provide insightful ideas into the possible future evolution of intelligent and secure communication technologies.

My sincere appreciations to all the authors, who contributed to this volume. I also express my sincere gratitude to the reviewers, for providing the timely and careful reviews to uphold the scientific merit of the volume.

This volume will be an insightful reference for researchers, educators, and practitioners working in all the fields of communication technologies.

Coimbatore, India

Dr. G. Ranganathan
Conference Chair
ICICCT 2025

Contents

An Arduino-Based Smart Car Safety System: Integrating Obstacle Detection, Battery Health Monitoring, and Smoke Detection for Enhanced Vehicle Safety 1
Debanjan Bera, Partha Biswas, Bristi Chakraborty, Soumadip Chandra, Souvik Bairi, and Sandip Chakraborty

Next Generation FR1 Multiport Antenna with Metasurface for Drone Networking 13
G. Suguna, P. Yogeshwari, A. Rajesh, C. Manikandan, N. Suba Renjani, and J. Saranya

Drone Positioning Estimation with Particle Filtering Augmented NR Positioning Reference Signal 23
N. Suba Renjani, A. Rajesh, C. Manikandan, G. Suguna, J. Saranya, and R. Selvakumar

Design and Optimization of Hybrid Circuits for Low-Power Systems Using GPDK 90nm 35
B. P. Bhuvana, Bharathi Raju, Parasa Gnan Naga Sai Vikas, and Tarun Varma Pedada

Smart Border Security: The Role of AI in Modern Border Surveillance System 45
Simant Kamal Dutta, Rahul Mahala, Rajesh Singh, Anita Gehlot, and Nagendar Yamsani

Multi-modal Data-Driven User Behavior Modeling and Personalized Advertisement Recommendation Algorithm 55
Zhang Zheng and Xu Dongyan

DDoS Attacks Detection in SDN Using Multi-feature Selection Approach and LLMs 65
Amit Kachavimath, Satish Vaishyar, Ayush Chugani, Pratham Singh, and Sanjana A. More

PassPhrase Authentication: A Secure and User-Friendly Multi-factor Authentication System Using Phrase and Gesture Integration 75
Amit Kachavimath, Ayush Chugani, Satish Vaishyar, Niyati Gogri, and Misbah Falak

DDoS Attacks Detection Using PSO Feature Selection and Classification Using XGBoost in SDN 85
Amit Kachavimath, Sanjana A. More, and D. G. Narayan

Chandrayaan-3 Mission to the Moon: A Systematic Review of Analysis, Improvements, Challenges and Lunar Science Contribution 95
Milind Shah, Dhairya Vyas, Kinjal A. Patel, M. Anantha Lakshmi, Roopal Rajput, and S. Lokesh

Unveiling Triadic Patterns in Dynamic Networks Using a GAT-SAGE Hybrid Model 119
Vidyalekshmi Chandrika, D. Prashanth Chandra Reddy, T. H. Hrishab, Mallampati Pranathi, Sanjay J. K. Seshadrinath, and Lekshmi S. Nair

Study of Stomach Abnormalities Using EGG Signals 131
R. Chandrasekaran, B. Praveena, Vivek Pazhamalai, S. Ivo Romauld, S. Vijayaraj, and M. K. Soundarya

FPGA-Based Approximate Multipliers for EdgeAI: A Systematic Review 141
Nguyen Thanh Loc, Nguyen Thien An, and Cuong Pham-Quoc

Investigation of Potential Impact on Drain Current of CMOS Inverter with Various Transistor Width and Temperature 157
Biswazit Mistry and Suman Chowdhury

Impact of Agentic AI-Driven RPA on Strategic Management: Enhancing Decision-Making and Organizational Efficiency 167
N. Priyanka and M. Shankar Lingam

System for Early Detection of Breast Cancer Using Deep Learning from Medical Imaging 179
Awa Bakhoum, Sada Anne, and Amadou Dahirou Gueye

Dual-Branch Scene Text Detection for Robust Competition Number Identification in Dynamic Outdoor Environments 195
Virendra Singh Solanki, Awanit Kumar, Snehlata Ajmera, Nirmal Singh, Sheshang Degadwala, and Dhairya Vyas

Deep Learning Based Model for Landslide Detection from Synthetic Aperture Radar Imaging 205
Manu Gupta, Yallamati Prakasa Rao, Manideep Shivanathula, Smahi Bandari, Arundhathi Uppari, and P. Michael Preetam Raj

A Comprehensive Study of CMOS Inverter Power Consumption with Transistor Width and Temperature Variations Using Cadence Virtuoso 217
Biswazit Mistry and Suman Chowdhury

Analysing the Optimization of Cooling Systems in Industrial Settings: Exploring Efficiency and Cost Reduction 227
Farid Agayev and Sevinj Bakhshaliyeva

Modeling of Solar Panel Installation Mechanism for Hybrid Energy Systems in Azerbaijan 235
Rehim Rehimov, Nizami Suleymanov, and Zarifa Mahmudova

System for Increasing the Accuracy of Digital Information Processing 251
Farid Agayev and Sevinj Bakhshaliyeva

Performance Evaluation of Different Shapes of Patch Antenna for 5G Applications at 25 GHz 265
Afsana Sultana Alvee, Md. Sohel Rana, Najifa Islam, Mahamud Hasan, and Md. Abdul Ahad

An Approach Based on the PRESENT Lightweight Cipher to Ensure Security and Performance Balance for IoT Devices 281
Tukhtajon Kozokova, Lola Davronova, and Nuriddin Jabborov

Transformer-Based Bidirectional Encoder SQL Representations for Interpretable Cyber Attack Detection Using Ensemble Learning 293
Raymond Sekyewa, Michael Kizito, Kyanda Swaib Kaawasse, Tonny Bulega, and Ggaliwango Marvin

ML-Driven Analysis of NavIC-Based Tracking for Route Optimization in Urban Air Mobility 315
Vishvesh Paresh Modcoicar and Trupti Baraskar

Real Time Crash Forecasting Using a Stacking-Based Machine Learning Ensemble 329
Dipankar Roy, Subir Gupta, Bibhuti Bhusan Dash, Monalisa Chakraborty, Subrata Chowdhury, and Sudhansu Shekhar Patra

Algorithm for Assessing Threats to Information Security 341
Kholimtaeva Ikbol, Sunbula Muminova, Barno Shamshieva, and Jakhongirbek Khabibullaev

A Smarter Plate: Enhancing Food Delivery Apps with Personalized Proportion Sizing and Nutritional Awareness 355
Ashwini Dalvi, Sargundeepkaur Sachdeo, and Samrudhi Bhat

HoloChef: 3D Culinary Guide with Speech Command 371
Fazliaty Edora Fadzli, Amira Azkia Rahmatulloh, Ajune Wanis Ismail, Muhammad Anwar Ahmad, Tee Ng Zikang, and Mohamad Yahya Fekri Aladin

Zero-Knowledge Enhanced Adaptation Learning (ZEAL): A Framework for Verifiable Model Adaptation 387
Dhrumil Panchal, Hrushik Mehta, Pragati Shetty, Pratik Kanani, Darshana Sankhe, and Rashmi Kumar

JOVA Robotic Prosthetic Arm 403
Jose Kevin, Kesanakurthi Aditya, Karuturi Vaishnavi, and V. Nithya

Assess the Effects to Organs Induced by I-131 Radiopharmaceutical Accumulates in Thyroid Gland by Monte Carlo Simulation 417
Minh Khanh Duong and Tat Thang Nguyen

Voltage Regulation in a 33 Bus Distribution System with Integrated STATCOM and Grey Wolf Optimizatıon 431
B. Balaji, P. Hemalatha, J. Chandru, and R. Preetha

Real Time People Monitoring Using MobilNet-SSD 449
Darshan Atkari, Arshal Agarwal, Atharva Kulkarni, Apoorva Kulkarni, and Sonali Antad

Performance Optimization of a PV Module Integrated with Fuzzy/NN INC-Based MPPT and Zeta DC-DC Converter for Grid-Connected Applications 463
S. Madhusudhan Reddy, S. Muthubalaji, and D. S. Sanjeev

MobileFormer: Efficient Stroke Lesion Detection with Lightweight CNN-Transformer Fusion 485
Sadiya Sulaiman, M. Roshni Thanka, E. Bijolin Edwin, and Nader Salam

Comparative Analysis of Iron Oxide (Fe_3O_4) Nanoparticles and Iron Oxide-Tungsten Trioxide (Fe_3O_4–WO_3) Nanocomposites Using Digital Image Processing Techniques 505
Parashuram Bannigidad, Sagar Chingali, and Prabhuodeyara Gurubasavaraj

Sustainable Healthcare Systems Through Arduino-Driven Medical Waste Segregation and Disposal Monitoring 519
Sharada M. Kori, Yogesh Lubal, Jyoti Amboji, and Madhuri M. Patil

An Arduino-Based Smart Car Safety System: Integrating Obstacle Detection, Battery Health Monitoring, and Smoke Detection for Enhanced Vehicle Safety

Debanjan Bera, Partha Biswas, Bristi Chakraborty, Soumadip Chandra, Souvik Bairi, and Sandip Chakraborty

1 Introduction

New models of automotive technology mark a shift from reactive to proactive systems where the goal is safety. Old technologies, like airbags, seat belts, and anti-lock braking systems (ABS), are fundamentally classified under passive safety; meaning they work to protect an individual in the aftermath of an incident [1]. As vehicles are becoming more complex and automated, there is a need for real-time detection and mitigation of hazards. With sensor technologies interconnected to microcontrollers, vehicles can effectively monitor and address conditions that can lead to accidents [2].

This paper presents an Arduino-based smart car safety system, which integrates three basic yet crucially important safety mechanisms: obstacle detection, battery health monitoring, and smoke detection, all in one affordable device. All components mitigate one or the other particular dangers associated with motor vehicles and thus cooperate to form a general integrated preventive solution.

The detection of obstacles is particularly applicable in city landscapes and for autonomous vehicles [3]. The ultrasonic sensor HC-SR04 assists in achieving the goal. It continues to update the range between the vehicle and objects in front of it thus, when necessary, it alerts the driver or rapidly stops the car [4]. Battery operations may also include overloading aspects, such as the sensor monitoring the voltage, current and temperature. Thanks to the timely application of these sensors, the breakdowns may be prevented and the life of batteries prolonged [5]. Leaking smoke and gasses are resolved using an MQ2 sensor, which can detect environmental pollutants such

D. Bera · P. Biswas · B. Chakraborty · S. Chandra · S. Bairi
Brainware University, Barasat, West Bengal, India
e-mail: debanjan2832004@gmail.com

S. Bairi
e-mail: souvikbairi2004@gmail.com

S. Chakraborty (✉)
Assistant professor, Computational Sciences, Brainware University, Barasat, West Bengal, India

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_1

as CO and CH4 gases in the air and activates alarms to help prevent smoke and toxins from coming up before a fire breaks out [6].

The goal is to create a consolidated and effective safety system that is Arduino Uno-based and is preferred because of its ease, low power use, and compatibility with sensors. This is a lightweight system that is expandable, not like AI-enabled systems or those that are too expensive for regular or self-driving cars. The system aims to provide real-time notification using an LCD. The latter included visual and audio warnings.

The proposed system is in tandem with the current trending inclination towards the incorporation of safety mechanisms as part of our everyday life which creates a base for the facility's possible future expansion by including such functionalities as integration internet of things, lane departure warning, and AI-facilitated diagnostics.

2 Literature Review

The inclusion of motor vehicles' sensor-based safety features is considered a significant part of lessening the rate of accidents on roads apart from the reduction of danger-related events. Several investigations have been done into the various safety devices available, which include the designs of collision avoidance systems, health monitors for batteries, fire alarms, and others that are important in sustaining safe operations. However, most authors prefer to treat design and control components separately rather than embedding them in a system. This part summarizes the responses in the literature to the highlighted questions and points out the innovative changes as well as the gaps in knowledge that this study intends to fill.

2.1 Obstacle Detection in Smart Vehicles

Obstacle detection serves as the basis for assisted and autonomous driving systems alike with the primary objective of being able to anticipate and consequently prevent obstacles in the path of the moving vehicles. In a study on collaborative motion guidance for industrial autonomous vehicles, it was shown how vehicles could avoid collisions by communicating with each other and by detecting the obstacles in real time [7].

Obstacles can be detected using ultrasonic sensors which is very common, especially in mobile robots. This was put to test in real time for different ranges and confirmed by Azeta et al. [8]. They are cost-effective to include in automotive systems that utilize an ultrasonic sensor within an Arduino.

Intriguingly, progress in this field has gone further by developing advanced concepts such as data fusion systems integrating ultrasonic sensor with stereo cameras. This view was provided by Gholami et al. [9], stating that combining the vision and ultrasound images is useful in negating a lot of false alarms while

increasing the level of awareness of the vicinity on a different scale. Unfortunately, stereo camera operations tend to be computationally intensive and therefore cannot be used on platforms of a lower cost, such as Arduino.

2.2 Battery Health Monitoring Systems

Batteries are an essential part of electric and hybrid motors. Neglecting to check the health of the battery can lead to a system crash or even fire. An Arduino powered Battery Management System (BMS) was implemented by Garg et al. [10] for monitoring the voltage, temperature, and current in real time, and the data could be accessed through mobile applications.

Pradhan and Chakraborty [11] built a battery monitoring system that works on voltage and temperature measurement using divider circuits and temperature sensors which ensures that the degradation is detected early. This allows for proactive measures to prevent failures of both the vehicles and the batteries themselves.

In a bid to provide more advanced monitoring systems, Widodo et al. [12] presented a BMS system that detects anomalous behaviour (due to voltage) in the batteries using machine learning and sample entropy. Unfortunately, such adaptive computing technology is not viable with a low power consumption microcontroller system since they require one with a high computation performance. As a result, it is possible to use battery monitoring devices based on Arduino modules, which are designed for low power and real time applications.

2.3 Smoke Detection in Automobiles

Detecting smoke at its early stage is very important when it comes to vehicle fires caused by electrical defects or leakages in fuel. Gautam et al. [13] studied semiconductor gas sensors, highlighting their effectiveness in detecting gases like carbon monoxide or smoke accumulation in vehicles' interiors.

One of these sensors with a long history of usage in fire alarms is the MQ2 smoke sensor, and this particular device is used by Adedokun et al. [14] to check how fast it can respond in a vehicle environment. At low concentrations, the sensor was able to detect smoke and give alerts. Its use with Arduino boards allows it to detect several gases, including CO, methane, or even LPG, making it a good product for car industries.

As for the incorporation of AI-enhanced smoke detection, although work in such lines has also been exploited and is known to people, for example, Yuan et al. [15] it is also noted that these detection algorithms are associated with high power consumption so that the system can discriminate between normal vapor and smoke, which is dangerous. Despite its effectiveness, low-power embedded systems cannot

use it for real-time performance. Still, they are promising. As a result, this study will utilize the sensor-based MQ2 module, which is inexpensive and reliable.

2.4 Summary and Research Gap

Despite the considerable amount of research on vehicle safety features at the individual level, there has been very little work done on designing a comprehensive safety package using a network of protective devices operated from a single microcontroller. Many of the existing solutions are expensive, advanced or cover only one aspect of safety system. The paper that follows brings solutions to these topics by introducing an Arduino cooperative system to detect obstacles, measure state of the battery and recognize smoke.

Unlike AI-based solutions, the recommended model is more accessible, cost-effective, and simple to implement, especially in environments with limited resources. Such a foundation will support further developments of smart car safety systems that embrace modularity, and will be ready for integration in electric cars as well as in traditional and autonomous vehicles.

3 System Design and Architecture

Arduino based intelligent vehicle safety system, including the obstacle detection system, battery maintenance concepts and anticipation of smoke formation, is thus modelled. These subsystems are normally digitally controlled with the help of an arduino, which aids them in operating optimally at any given time. This system encompasses a three-tier architecture configuration which consists of the following aspects: detection, processing, and implementation.

The primary component of the system is an Arduino Uno whose main benefits are its ease of use, its ability to perform tasks in real time, its low price and also, its ability to be interfaced with many sensor modules. Compared to counterparts like Raspberry Pi and ESP32, the Arduino Uno offers a power-efficient and hardware-centric platform well suited for incorporating into vehicle safety electronics. Please refer to Table 1 for the specifications of the Arduino Uno in detail. The system implements a variety of sensors to measure different, safety related parameters. Specifications of these components are given in Table 2. A schematic showcasing these devices and their connection in its entirety is Implemented in Fig. 1.

The obstacle detection module mounts an HC-SR04 ultrasonic sensor on the front portion of the vehicle to achieve this. It is a sensor that emits sound of a frequency of 40 kHz and then measures the delay between sound emission and receiving the echo of the sound it produced on objects and the new positioning it would have traveled. This kind of object detection a small range from 2 to 4 meters, which is possible by exploiting formula 1. If the obstacle comes within a specified zone, the system turns

Table 1 Arduino uno—technical specifications

Parameter	Specification
Microcontroller	ATmega328P
Operating voltage	5V
Input voltage (recommendd)	7–12V
Digital I/O Pins	14 (6 PWM-enabled)
Analog input pins	6
Clock speed	16 MHz
Flash memory	32 KB (0.5 KB used by bootloader)
SRAM	2 KB
EEPROM	1 KB
Communication interfaces	UART, SPI, I2C

Table 2 Sensors used in the smart car system

Sensor type (model)	Parameter and range
Ultrasonic sensor (HC-SR04)	Distance: 2–400 cm, ±3 mm
Voltage sensor (Generic)	Voltage: 0–25 V, ±0.1 V
Current sensor (ACS712 (20A))	Current: ±20 A, ±1.5%
Temperature sensor (DHT11)	Temp: 0–50 °C, ±2 °C
Smoke sensor (MQ2)	Smoke/Gas: 300–10,000 ppm, 2–5s

off the motor and activates sound and light indicators, stopping it from hammering against the obstacle. Figure 2 displays the real-time testing process and results of the same application.

$$Distance = \frac{Speed of Sound \times Time Delay}{2} \tag{1}$$

The battery health monitoring module integrates three sensors:

- A **voltage sensor**, to keep track on the overall battery voltage;
- A **current sensor (ACS712)**, to measure the current draw and identify overcurrent conditions; and
- A **temperature sensor (DHT11)**, to detect whether the battery is overheating.

The Arduino Uno receives constant information from the sensors, analyzes and compares it with the norm, and signals any deviations. When the system identifies factors like low voltage, overcurrent, or high temperature, it sends an alarm for immediate attention. Additionally, an LCD displays the real-time battery stats for quick evaluation. This feature allows quick and non-invasive risk assessment or problem solving following a standard or maintenance routine. The screen changes

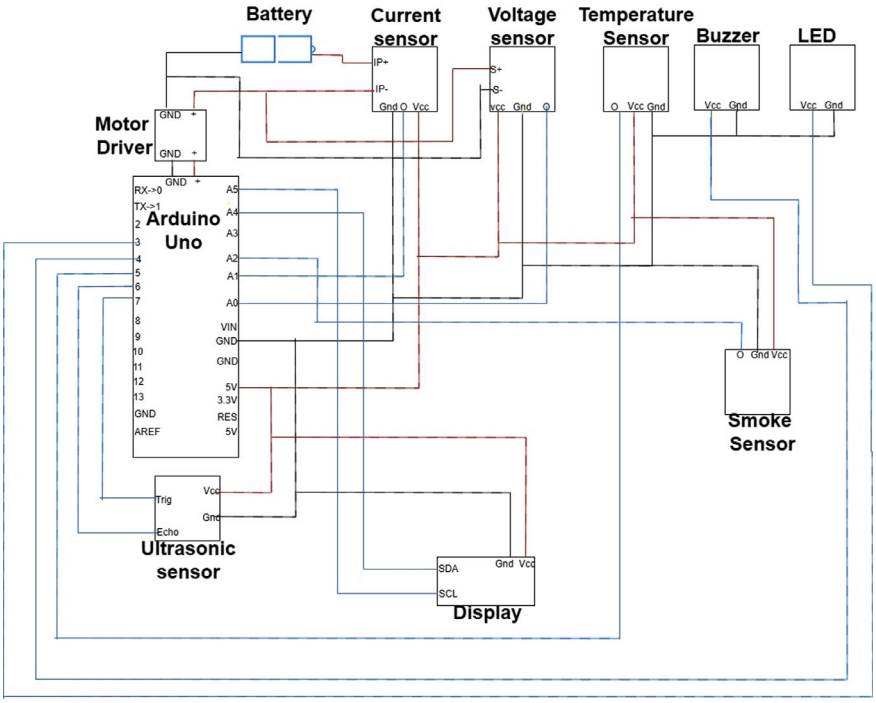


Fig. 1 Circuit diagram of the system

in response to the changes in battery health, and it does update itself. In Fig. 3, there is an illustration showing what the monitoring view may look like in real time.

To alert about smoking, gas leaks, explosions, and other accidents in the car, the module uses the MQ2 gas sensor, which can detect carbonaceous smoke, hydrocarbon gases including methane, as well as carbon monoxide, or liquefied petroleum gas. Inside, vehicular air is screened cyclically by the sensor to gauge air quality, and the data is transmitted to the Arduino. If the concentration of gases is more than the stipulated safety standard, typically more than 300 ppm, a warning will be triggered on the system in case of impending hazard. In addition to buzzer on and the LED turning on, this air quality alarm is also demonstrated on the LCD screen that is used to display the voltage level to the driver. The simplicity of this notification stems from the possibility of drivers interacting with both sound and graphical data. The response to smoke alarms is the topic of Fig. 4

The whole network is controlled by a DC source, and every sensor is attached to the Arduino Uno through analog/digital input nodes. The motor driver (L293D) provides the link between the Arduino and the DC motors for the purpose of moving them in response to the signal from the sensor. The detection system includes the use of buzzers and LEDs to indicate signals in hazardous environments.



Fig. 2 Obstacle detection output



Fig. 3 Battery status display



Fig. 4 Smoke detection output

We designed this system using a modular approach to facilitate easy adjustments for evolving safety requirements. It takes advantage of the possibilities of modern electronics, which allow for immediate control response and adaptable warning depending on the danger level. The system also guarantees that it fits with current manned vehicles and, at the same time, can be used with unmanned vehicles, thus making it a potential design for smart car hazard detection.

4 Results and Discussion

To evaluate the performance of the Arduino-based smart car safety system, three key functionalities were subjected to test cases designed to imitate actual circumstances. These features include the obstacle detection, battery health monitoring and smoke detection, with particular focus placed on promptness, precision, and trustworthiness. All the modules were also integrated into a working prototype car and evaluated while situated inside a laboratory, which also involved trying performance under various tests to ascertain their efficiency.

4.1 Obstacle Detection Performance

Ultrasonic sensors (HC-SR04) were efectively deployed within the limiting range of 2 cm to 4 meters with less than 50-millisecond response time on average for detecting an object. Whenever the defined critical distance (say 25 cm) was approached by any object, the Arduino would turn off the motor and trigger the buzzer and LED alarm immediately, which prevented any pair of simulated collision immediately. However, some setbacks were observed wherein soft and uneven surfaces like foam would absorb sound waves, leading to sporadic false negatives.

Environmental disturbances were also used to assess the sensors' effectiveness. Detection was very consistent operating under low-light dry indoor conditions but slightly low under cases of strong air currents, water droplets, or sonic interference. Under the mentioned conditions, detection accuracy could be possibly impeded due to the degrading effect these factors have on ultrasonic signal quality. Future works could focus on the integration of infrared or LiDAR sensors that may support improving the detected performance in conditions like fog, rain, and EMI disturbances.

4.2 Battery Health Monitoring Results

The voltage sensor precisely tracked input levels to within ± 0.1 V of the true values. The ACS712 current sensor can reliably detect deviations in current when the measured value exceeds the resolution threshold in the 20 A-rated configuration. The DHT11 sensor provided steady monitoring of the battery box temperature, with an accuracy of ± 1 °C. To alert the operator, critical values, such as low voltage at ≤ 10.5 V, overheating at ≥ 45 ° C, or conditions of overcurrent, were considered triggers. These conditions were displayed on the LCD screen in real time, allowing the user to monitor and respond proactively before any real damage could be done.

4.3 Smoke Detection Accuracy

The MQ2 sensor then responded in between 1 and 2 seconds to simulate smoke and gas leaks. It was capable of detecting concentrations higher than 300 ppm. The LCD was programmed to display an air quality warning with gas concentration levels and battery data. Sensitivity was good in closed spaces but startup time was slightly delayed due to sensor warm-up time.

Table 3 Comparative analysis of smart vehicle safety systems

Feature	Proposed system (Arduino-based)	AI-based system	Traditional system
Obstacle detection	HC-SR04 Ultrasonic sensor (±3 mm)	LiDAR / Stereo Vision (High accuracy, high cost)	Infrared Sensors (Low range)
Battery monitoring	Voltage, current, temperature sensors	ML-based predictive BMS	Voltage-only checks
Smoke detection	MQ2 sensor (Multi-gas detection)	AI smoke classifier (Needs GPU)	Basic CO sensor
Display feedback	LCD interface (real-time alerts)	App/cloud interface	No display or basic LED
Cost (per unit)	Low (> INR 4000 or $50)	High (INR 50,000+)	Medium (INR 10,000)
Implementation effort	Easy (DIY-compatible)	Complex (Data training needed)	Moderate

4.4 System Comparison with Existing Solutions

For conducting assessment for practicality of system, it was compared with other vehicle safety platforms in terms of cost, accuracy, and implementation. Results are summarized in Table 3.

We have examined the system in different scenarios, performing about 10 test iterations per feature. Nearly all the data logs indicate a success of 95–98% in correctly detecting hazards and activating warnings in nominal conditions; only under certain edge cases, for example, high humidity or unstable power supply, were there minor cases of false activation or delayed activation reported. These results confirm that the system is a good candidate for entry-level smart vehicles, educational labs, and areas with tight budgets.

5 Conclusion and Future Scope

The work focused on developing a smart car safety system integrated with an Arduino, which includes obstacle detection, battery health monitoring, and smoke detection functionalities under one roof. It is an automatic system that detects the surrounding objects through ultrasonic sensors within a safe range. In addition, the platform continuously monitors the battery voltage, current, and temperature value and gives responses under risky concentrations of hazardous gases using the MQ2 sensor. Enhanced awareness and usability for the system were provided by

an LCD where real-time data were visualized. Testing proved to show reliability in performance during normal working conditions, with well-triggered alerts under normal as well as repeat iterations. Compared to an old-school traditional system or an AI-based system, the proposed one is very low-cost with very low complexity and thus suitable for students and entry-level automobiles. Developments hereinafter may include adaptive safety responses using machine learning, Internet of Things (IoT) connectivity for remote monitoring and various others, including lane departure warnings, alcohol detection, and even diagnostics within the cloud. All these features would open up further capability for this system and set it toward being an intelligent, scalable safety solution for next-generation automobiles.

References

1. Répás J, Berek L (2022) Security and safety systems on modern vehicles. In: Lecture notes in mechanical engineering. pp 84–100. https://doi.org/10.1007/978-3-031-15211-5_8
2. Yaswanth R, Babu MR (2024) Revolutionizing automotive technology: unveiling the state of vehicular sensors and biosensors. IEEE Access 1. https://doi.org/10.1109/access.2024.3514157
3. Chandra BR, Chaithanya JK, Roy A, Reddy CL (2025) Smart city: challenges and opportunities detection and identification of autonomous vehicles using sensor synthesis. In: Explainable IoT applications: a demystification. Springer, Cham, pp 143–158. https://doi.org/10.1007/978-3-031-74885-1_9
4. Alsayaydeh JAJ, Yusof MFB, Abdillah MAAB, Al-Gburi AJA, Herawan SG, Oliinyk A (2023) Enhancing vehicle safety: a comprehensive accident detection and alert system. Int J Adv Comput Sci Appl 14. https://doi.org/10.14569/ijacsa.2023.0141104
5. Waseem M, Ahmad M, Parveen A, Suhaib M (2023) Battery technologies and functionality of battery management system for EVs: current status, key challenges, and future prospectives. J Power Sources 580:233349. https://doi.org/10.1016/j.jpowsour.2023.233349
6. Biscarri EL (2024) In: Fires in conventional and electrified vehicles: theory, prevention, and analysis. SAE International
7. Grosset J, Fougères A-J, Djoko-Kouam M, Couturier C, Bonnin J-M (2024) Collision and obstacle avoidance for industrial autonomous vehicles—simulation and experimentation based on a cooperative approach. https://hal.science/hal-04638317v1
8. Azeta J, Bolu C, Hinvi D, Abioye AA (22019) Obstacle detection using ultrasonic sensor for a mobile robot. In: IOP Conference series materials science and engineering, vol 707. pp 012012. https://doi.org/10.1088/1757-899x/707/1/012012
9. Gholami F, Khanmirza E, Riahi M (2022) Real-time obstacle detection by stereo vision and ultrasonic data fusion. Measurement 190:110718. https://doi.org/10.1016/j.measurement.2022.110718
10. Garg P, Bansal D, Khan S, Bhattacharya S (2022) Development of energy measurement system for electrical vehicle battery using Arduino. In: Lecture notes in electrical engineering. pp 457–464. https://doi.org/10.1007/978-981-19-0588-9_45
11. Pradhan SK, Chakraborty B (2022) Battery management strategies: an essential review for battery state of health monitoring techniques. J Energy Storage 51:104427. https://doi.org/10.1016/j.est.2022.104427
12. Widodo A, Shim M-C, Caesarendra W, Yang B-S (2011) Intelligent prognostics for battery health monitoring based on sample entropy. Expert Syst Appl 38:11763–11769. https://doi.org/10.1016/j.eswa.2011.03.063

13. Gautam A, Verma G, Qamar S, Shekhar S (2019) Vehicle pollution monitoring, control and challan system using MQ2 sensor based on internet of things. Wireless Personal Commun 116:1071–1085. https://doi.org/10.1007/s11277-019-06936-4
14. Adedokun AO (2024) IoT based public transportation accident prevention with driver alcohol detection and smoke system. Int J Res Innov Appl Sci 9:765–775. https://doi.org/10.51584/ijrias.2024.908069
15. Yuan C, Liu Z, Zhang Y (2018) Learning-Based smoke detection for unmanned aerial vehicles applied to forest fire surveillance. J Intell Robot Syst 93:337–349. https://doi.org/10.1007/s10846-018-0803-y

Next Generation FR1 Multiport Antenna with Metasurface for Drone Networking



G. Suguna, P. Yogeshwari, A. Rajesh, C. Manikandan, N. Suba Renjani, and J. Saranya

1 Introduction

Nowadays, there has been a noticeable increase in the use of drones in the military, communications, industrial, and commercial markets. To improve, the transmission speed and bandwidth spectrum of fifth-generation (5G) wireless communication have recently attracted attention worldwide [1, 2]. Wireless networks may need n-times more capacity in the next years than they do now because of a massive growth in annual data traffic (40–70%). With a peak throughput of many gigabits per second (Gb/s), fifth-generation (5G) systems are thought to be a significant option for upcoming communication applications to address this rapidly increasing demand [3]. Communication with an Unmanned Aerial Vehicle (UAV) is challenging due to instability and drag issues that occur when the drone is in motion [4]. Antennas become bendable when their substrates are clad by thin copper film. The antenna must be light to improve system endurance, tiny to accommodate UAV's extremely tight space constraints, and low profile to minimize air resistance and meet aerodynamic requirements [5]. In [6], two folded strips installed on a ground plane make

G. Suguna · P. Yogeshwari · A. Rajesh · C. Manikandan (✉) · N. Suba Renjani · J. Saranya
School of Electrical and Electronics Engineering, SASTRA Deemed University, Thanjavur, Tamil Nadu, India
e-mail: manikandan@ece.sastra.ac.in

G. Suguna
e-mail: suguna@ece.sastra.ac.in

A. Rajesh
e-mail: rajesha@ece.sastra.ac.in

N. Suba Renjani
e-mail: dpfi0424011924@sastra.ac.in

J. Saranya
e-mail: dpfi0424011997@sastra.ac.in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_2

up the UAV antenna. It operates at 500–511 MHz frequency with an omnidirectional pattern. Changing the capacitive at the two ends of the metal strips increases the antenna performance. A multiport antenna with a metasurface is constructed [7]. A 1×2 array antenna with a circular split ring-shaped metasurface was developed to enhance the MIMO system. The antenna operating frequency is mm-wave with a peak gain of 10.27 dBi.

In [8], the meander line MIMO antenna was developed at the n78 band with a bandwidth of 400 MHz. Here, the EBG structure was introduced to improve the isolation of the antenna. A U-shaped metamaterial was designed with double negative properties at 5, 6, and 11 GHz [9]. In addition, [10] three pairs of metamaterial arrays are integrated into the bow-tie-shaped millimeter-wave multiplexing antenna to increase gain. This antenna can achieve a maximum gain of 7.4 dBi. MIMO antennas working in the millimeter-wave frequency spectrum have been reported in a few literary works. The antennas in these works display strong isolation and good MIMO performance. But these MIMO antenna's gain is rather low [11, 12]. A MIMO antenna was implemented for the LTE band of 2.3–2.4 GHz [13]. A split ring resonator was developed to reduce the coupling of the multiport antenna. The structure enhances the peak gain of 2 dBi. A ceramic material-based metasurface antenna is developed for the 3.8–4.2 GHz frequency band [14]. In [15], a wideband MIMO antenna was investigated by integrating metamaterial structure. The antenna improves the return loss by less than − 10 dB at ku band frequency. In [16], an eight-port MIMO antenna was developed with dual polarization for 5G communication. In this work, a square ring with a slotted line microstrip patch is designed at 3.6 GHz. A circular ring parasitic structure was introduced to increase the isolation between the antennas. This obtained a bandwidth of around 600 MHz. The comparison of related work is shown in Table 1. A truncated microstrip patch with the defective ground structure was designed to place an orthogonally placed 4-port MIMO antenna. This 5G antenna operates in the frequency of n78 (3.6 GHz) band. Additionally, to improve the antenna performance the Dual Square Ring (DSR) double negative metasurface was implemented. The proposed MTS antenna is used for 5G Unmanned Aerial vehicle communication.

Table 1 Comparison of this work with literature work

Ref	No. of ports	Frequency (GHz)	Technique	Mutual coupling (dB)	ECC	Peak gain (dBi)	Material
13	2	2.4	MNG metamaterial	< –20	< 0.16	2	FR-4
14	2	3.8–4.2	DRA + MS + Airgap	< –20	< 0.03	4	FR-4
15	2	8.5–11.5	Superstrate + MIMO	< –18	< 0.01	7.2	FR-4
This work	4	3.6	Metasurface + MIMO	< –25	< 0.025	8.305	Polyimide

2 DSR Metasurface Design

The design of the unit cell is influenced by the desired electromagnetic response and operational frequency range within the n78 band. The structure incorporates dual square rings made of copper, as illustrated in Fig. 1. The DSR-MTS, measuring 38 × 38 mm, was constructed on a polyimide substrate with a thickness of 0.1 mm, it appropriate for drone communication. Table 2 outlines the metasurface design parameters. Figure 2 displays the reflection (S_{11}) and transmission (S_{21}) coefficients of the DSR-MTS. The DSR-MTS design exhibits resonance at the n78 (3.6 GHz) band, achieving a reflection coefficient of − 46 dB. Conversely, the transmission coefficient approaches zero, indicating the high reflectivity of the MTS in this 5G band. The real and imaginary values of the effective permittivity and permeability for the DSR-MTS structure are depicted in Fig. 3. Both permittivity and permeability demonstrate negative values at the 5G frequency of 3.6 GHz, suggesting that the metasurface exhibits double negative (DNG) characteristics.

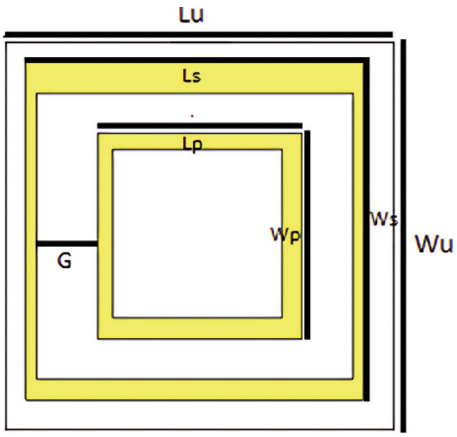


Fig. 1 Geometric of DSR-metasurface design

Table 2 Dimension of DSR-MTS

Variable	Dimension (mm)	Variable	Dimension (mm)
Lu	38	Wu	38
Ls	33	Ws	33
Lp	20	Wp	20
G	6	Lm × Wm	106

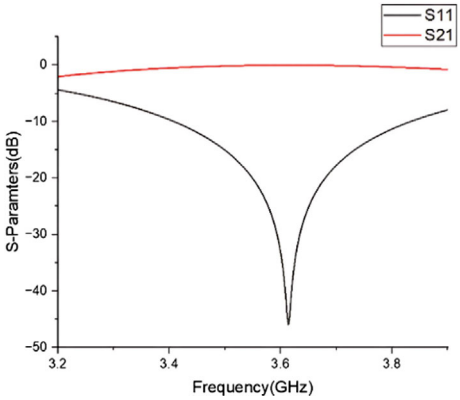


Fig. 2 S-parameter analysis for DSR-MTS

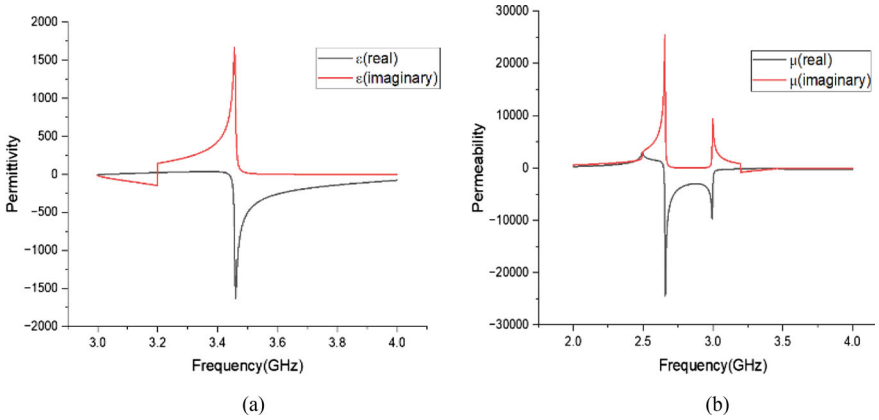


Fig. 3 Simulated results of DSR-MTS **a** permittivity **b** permeability

3 Polyimide Conformable Double Split Antenna Design

Subsequently, a single-element antenna geometry was utilized in the proposed MIMO arrangement, as shown in Fig. 4. The polyimide substrate had dimensions of 35 × 32 × 0.1 mm. A defective structure was introduced in the ground plane to achieve good performance, such as high gain, return loss, and efficiency. Figure 5 depicts the S-parameters for a single antenna. It achieves a reflection coefficient of less than − 10 dB in the n78 band. The dimensions of the proposed single-port and 4-port antenna elements are listed in Table 3.

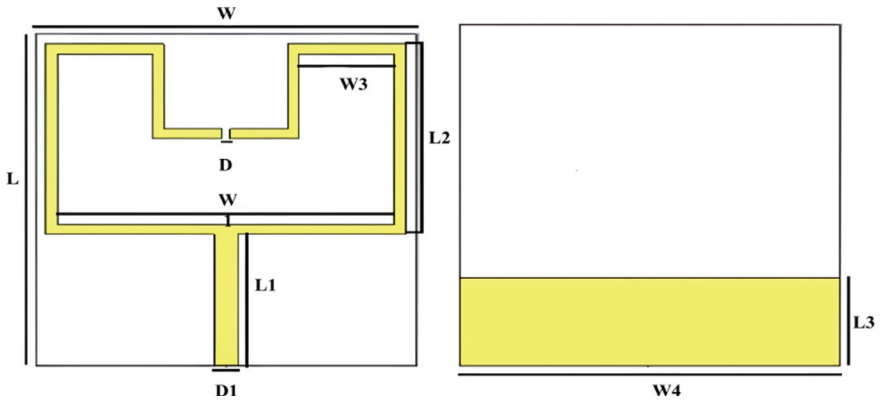


Fig. 4 Geometric of single patch double split antenna design

Fig. 5 S_{11} of single port antenna

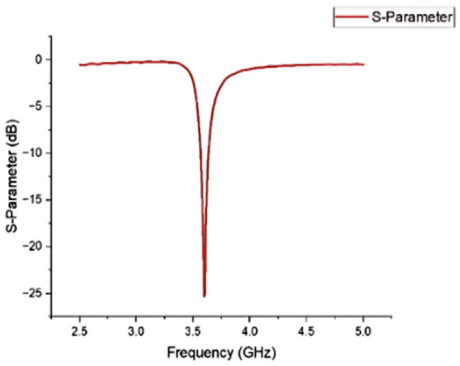


Table 3 Geometry of proposed antenna

Variable	Dimension (mm)	Variable	Dimension (mm)
Ls	35	Ws	32
L1	13.9	W1	28.3
L2	20.10	W3	9.95
L3	9.02	W4	32
D	0.7	D1	2
La	78	Wa	86.5

4 MIMO Antenna with Metasurface

The single-patch antenna was affected in urban areas due to signal attenuation, scattering, and multipath effects. To overcome these effects, an orthogonally placed four-port MIMO antenna with a DNG Metasurface is proposed to improve the gain.

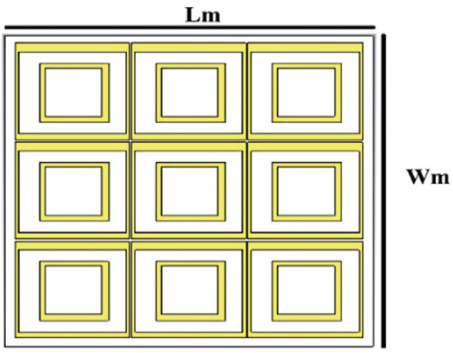


Fig. 6 Array structure of DSR-MTS

Furthermore, the utilization of the MIMO antenna helps to mitigate multipath fading and enhance signal reliability by leveraging spatial diversity and increasing channel capacity. The geometry of the 3 × 3 metasurface array utilized as the superstrate for enhancing MIMO antenna performance is depicted in Fig. 6. The design and fabrication of the MIMO antenna and metasurface are shown in Fig. 7a–e. An air gap of 12 mm separates the 4-port MIMO antenna from the integrated DSR metasurface array. The antenna's S-parameter characteristics, both with and without the metasurface in a multiport configuration, show strong similarity. The metasurface improves the S_{11} < −10 dB, and mutual coupling to less than − 30 dB, as shown in Fig. 8. Figure 9a demonstrates that the multiport antenna's peak gain increases from approximately 5.02 dBi to 8.305 dB when the metasurface is employed.

The 2D far-field simulated radiation pattern at n78 band frequency, displayed in Fig. 9b, indicates the antenna's directional operation. Figure 10 illustrates the current flow in the patch, showing the current distribution from the feed to the rectangular patch. At 5G frequency (3.6 GHz), strong current is observed throughout the split ring and antenna ground plane. MIMO parameters were evaluated for the DSR-MTS integrated antenna. Figure 11a shows the Diversity gain (DG) for the proposed MTS-based MIMO antenna. It obtained a value of < 9.9 dB at the 5G frequency, which is considered a good level of diversity gain. Similarly, Fig. 11b displays the Envelope Correlation Coefficient (ECC) measured at the same frequency less than 0.005. This shows that the DSR metasurface-based MIMOantenna achieves good independence and minimizes the correlation between the antenna elements. The Total Active Reflection Coefficient (TARC) is used to analyze the diversity performance of the antenna illustrated in Fig. 11c. The MTS-based MIMO antenna enhances the TARC value is < − 30 dB.

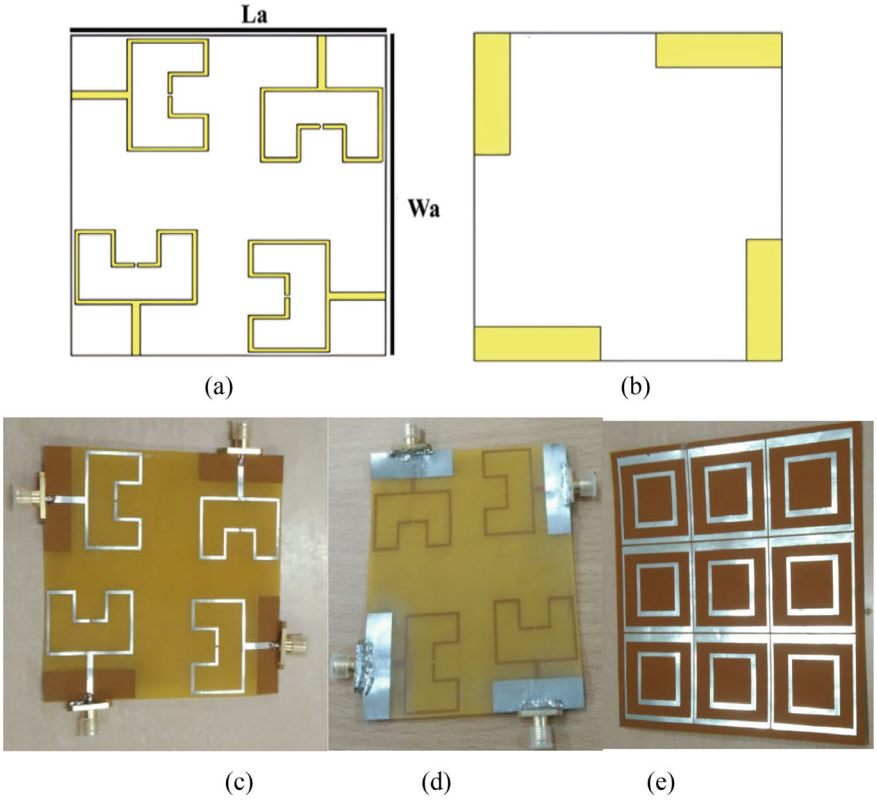


Fig. 7 **a** Front side of multiport antenna **b** Defective ground plane **c** Fabricated front side **d** Fabricated rear side **e** Fabricated metasurface

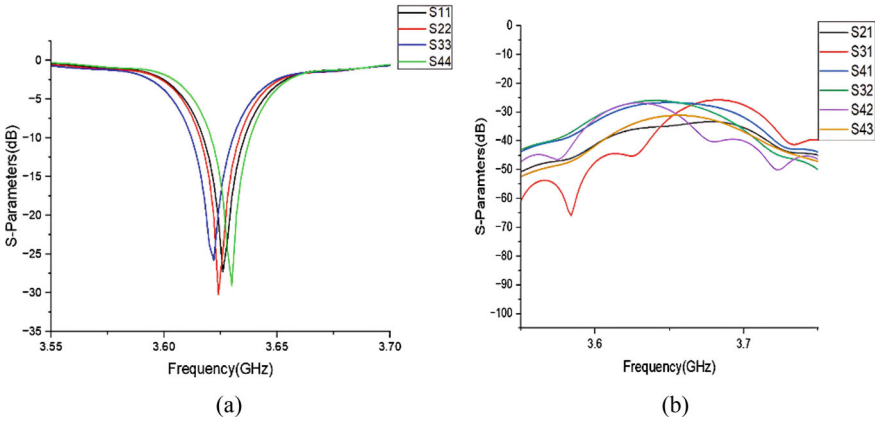


Fig. 8 S-parameter of antenna with MTS **a** Reflection co-efficient (S_{11}) and transmission coefficient (S_{21})

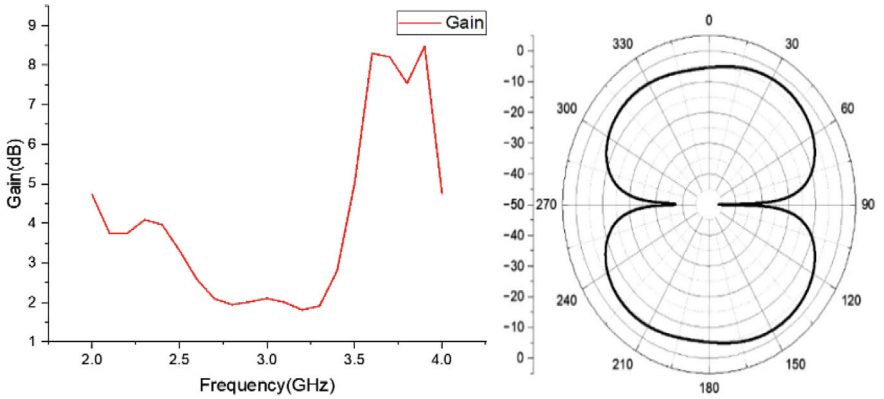
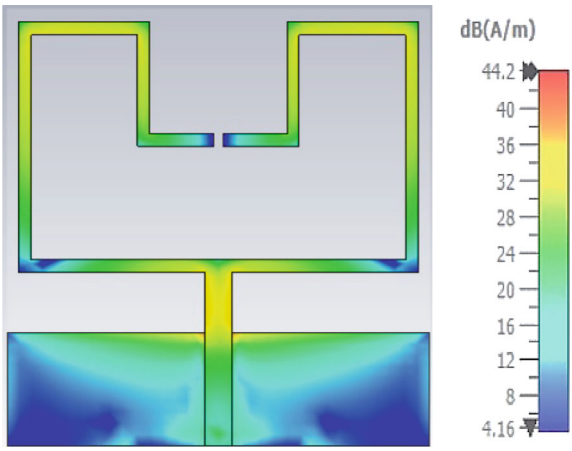


Fig. 9 Simulated results of the antenna with MTS **a** Gain **b** Radiation pattern at 3.6 GHz

Fig. 10 Field distribution of the antenna with Metasurface



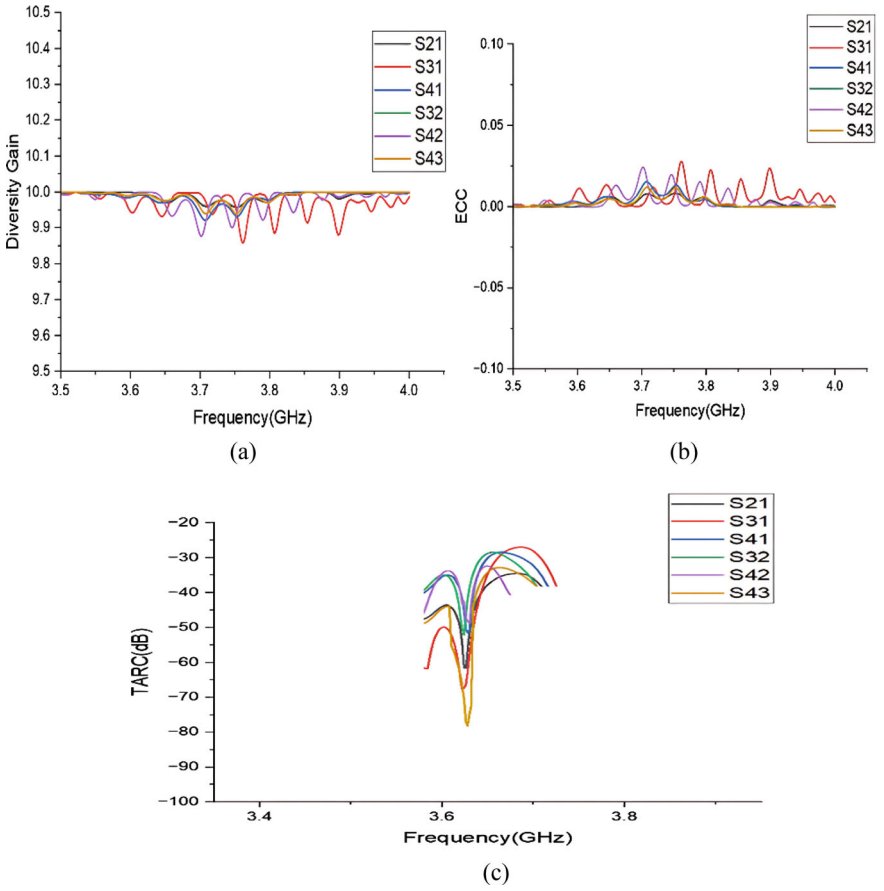


Fig. 11 MIMO parameters **a** DG **b** ECC and **c** TARC

5 Conclusion

In this paper, an orthogonally placed 4-port MIMO antenna with dual square DNG Metasurface was designed for positioning UAV communication. It operates the bandwidth from 3.4 to 3.6 GHz of the n78 band with the highest gain of 8.3 dBi. The MIMO parameters are analyzed with better performance and isolation is good without adding a decoupling structure. A single-layered metasurface-based four-port antenna was suitable for frequency range 1(FR1) 5G drone applications.

Acknowledgements We would like to Thank and Acknowledge IIT Tirupati Navavishkar I-Hub Foundation (IITTNiF), National Mission on Interdisciplinary Cyber-Physical Systems (NM-ICPS), Department of Science and Technology (DST), India for their financial support (SASTRA-IITTNiF/5GNP/2024-25) and SASTRA Deemed to be University, India, for providing the necessary resources and support.

References

1. Roh W, Seol J-Y, Park J, Lee B, Lee J, Kim Y, Cho J, Cheun K, Aryanfar F (2014) Millimeter-wave beamforming as an enabling technology for 5G cellular communications: theoretical feasibility and prototype results. IEEE Commun Mag 52(2):106–113
2. Boccardi F, Heath RW, Lozano A, Marzetta TL, Popovski P (2014) Five disruptive technology directions for 5G. IEEE Commun Mag 52(2):74–80
3. Ferendeci AM (2016) Flexible conformai antennas. In: 2016 IEEE/ACES ınternational conference on wireless ınformation technology and systems (ICWITS) and applied computational electromagnetics (ACES). IEEE, pp 1–2
4. Sharawi MS, Aloi DN, Rawashdeh OA (2010) Design and implementation of embedded printed antenna arrays in small UAV wing structures. IEEE Trans Antennas Propag 58(8):2531–2538
5. Scorrano L, Rosa AD, Orobello B, Manna A, Trotta F (2017) Experimental investigations of a novel lightweight blade antenna for UAV applications [antenna applications corner]. IEEE Antennas Propagat Magazine 59(2):108–178
6. Sun B-H, Wei Y-F, Zhou S-G, Liu Q-Z (2009) Low-profile and horizontally-polarised antenna for UAV applications. Electron Lett 45(22):1106–1107
7. Chattha HT (2019) 4-port 2-element MIMO antenna for 5G portable applications. IEEE Access 7:96516–96520
8. Tan X, Wang W, Yongle W, Liu Y, Kishk AA (2019) Enhancing isolation in dual-band meander-line multiple antenna by employing split EBG structure. IEEE Trans Antennas Propag 67(4):2769–2774
9. Mallik A, Kundu S, Goni MO (2013) Design of a novel two-rectangular U-shaped double negative metamaterial. In: 2013 International conference on ınformatics, electronics and vision (ICIEV). IEEE, pp 1–6
10. Jiang H, Si L-M, Weidong H, Lv X (2019) A symmetrical dual-beam bowtie antenna with gain enhancement using metamaterial for 5G MIMO applications. IEEE Photon J 11(1):1–9
11. Jilani SF, Alomainy A (2018) Millimetre-wave T-shaped MIMO antenna with defected ground structures for 5G cellular networks. IET Microwaves Antennas Propagat 12(5):672–677
12. Pan YM, Qin X, Sun YX, Zheng SY (2019) A simple decoupling method for 5G millimeter-wave MIMO dielectric resonator antennas. IEEE Trans Antennas Propagat 67(4):2224–2234.
13. Lee Y, Ga D, Choi J (2012) Design of a MIMO antenna with improved isolation using MNG metamaterial. Int J Antennas Propagat 2012(1):864306
14. Dwivedi AK, Sharma A, Singh AK, Singh V (2021) Metamaterial inspired dielectric resonator MIMO antenna for isolation enhancement and linear to circular polarization of waves. Measurement 182:109681
15. Jiang J, Li Y, Zhao L, Liu X (2020) Wideband MIMO directional antenna array with a simple meta-material decoupling structure for X-band applications. Appl Comput Electromagnet Soc J (ACES) 556–566
16. Parchin NO, Al-Yasir YIA, Ali AH, Elfergani I, Noras JM, Rodriguez J, Abd-Alhameed RA (2019) Eight-element dual-polarized MIMO slot antenna system for 5G smartphone applications. IEEE Access 7:15612–15622

Drone Positioning Estimation with Particle Filtering Augmented NR Positioning Reference Signal



N. Suba Renjani, A. Rajesh, C. Manikandan, G. Suguna, J. Saranya, and R. Selvakumar

1 Introduction

Recently, the usage of drones increased in various commercial and industrial fields. Much of the current research consistently relies on GPS to determine the drone's precise location [1]. Nevertheless, in certain scenarios, GPS might not provide the required accuracy, and this approach is not suitable for disaster situations. Drones normally rely on GNSS for positioning, which is effective in open environments. However, GNSS alone struggles with accuracy and reliability in urban or indoor settings due to signal blocking, multipath propagation, and susceptibility to jamming and spoofing [2]. To overcome this, 5G networks enhance positioning and navigation, with Positioning Reference Signals (PRS) improving accuracy and synchronization in challenging environments [3]. Positioning Reference Signal is a distinct transmission from cellular networks (such as 4G LTE and 5G) that enables devices to ascertain their position by calculating the time difference of arrival (TDOA) of the

N. Suba Renjani · A. Rajesh (✉) · C. Manikandan · G. Suguna · J. Saranya · R. Selvakumar
School of Electrical and Electronics Engineering, SASTRA Deemed University, Thanjavur, Tamil Nadu, India
e-mail: rajesha@ece.sastra.ac.in

N. Suba Renjani
e-mail: dpfi0424011924@sastra.ac.in

C. Manikandan
e-mail: manikandan@ece.sastra.ac.in

G. Suguna
e-mail: suguna@ece.sastra.ac.in

J. Saranya
e-mail: dpfi0424011997@sastra.ac.in

R. Selvakumar
e-mail: selvakumar@civil.sastra.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_3

signal emitted from the base station. Since 3GPP release 9, PRS has been used for LTE locations. Some work is based on existing LTE PRS, expected standards, or ideas for 5G PRS. While [6] addresses PRS setup to enable the OTDoA (Observed Time Difference of Arrival) positioning method, authors in [7] assess the effect of positioning with various PRS distributions based on accuracy in a period. 5G Downlink PRS may allocate more OFDM symbols in the time domain and a wider bandwidth in the frequency domain than 5G SSB signals, which enables continuous carrier phase tracking and improved positioning accuracy [10, 11]. In the meantime, many base stations and several frequency bands are typically needed to achieve the desired sub-meter accuracy, which occasionally may be less efficient or impracticable [12]. This study examines these problems by utilizing standalone 5G signals to attain sub-meter location precision while requiring a minimum of three 5G base stations. The most recent 3GPP technical report emphasizes network synchronization issues. Therefore, research into 5G-based positioning techniques is required to reach greater accuracy levels, including sub-meter accuracy. To address this issue, this study suggested a novel 5G-GNSS hybrid positioning system based on AOA-TOA measurement [13]. PRS quantifies the downlink transmission delays by comparing the base station signal, which the receiver locally repeats. To determine whether the PRS distribution impacts positioning performance [14], this research analyses and compares various PRS allocations within the 5G frame[15]. The primary goal of this paper is to provide an NR PRS positioning and TOA estimation detection as per TDOA positioning for the calculation obtained. Using simulated results, PRS performance and location accuracy are demonstrated.

2 Methodology

In the SDR-based network positioning technology, two SDRs are used, transmitters that send LTE-specific positional reference signals (PRS). Hardware from the USRP B210 SDR is used as shown in Fig. 1. Drone uses PRS signals, which are sent concurrently by gNB, as its main metrics for estimating arrival time. To determine TOA from each gNB and then calculate the time difference of arrival across all pairs of gNBs, drone correlates with the PRS that is received from gNB. The drone location estimation is defined by the intersection of the TDOA curves (hyperbolas) from the three base stations. MATLAB on laptops is connected to USRP boards via the USRP hardware driver. PRS baseband synthesis, transmission, reception, and processing are all done with MATLAB software.

For positioning design, the orthogonal frequency-division multiplexing (OFDM) grid and PDSCH (Physical Downlink Shared Channel) demonstrate the set-up of the PRS time–frequency features and how various PRS resource set settings impact the PRS time–frequency structure. Observe the differences in slot allocation and RE positions after configuring many PRS resources with various time and frequency allocation and slot configuration elements. The time it takes for a signal to move from the gNB to the drone is measured using distance. The speed of light is used to

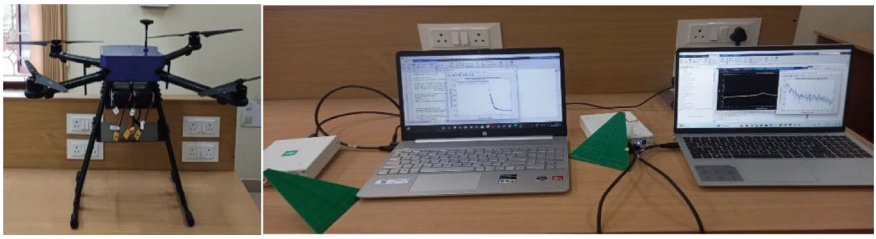


Fig. 1 SDR based test bench for testing the particle filtering based positioning in drone

determine the time it takes for the signal to travel that distance. This time delay is converted into units that can be used in simulations, and the path loss formula is also used to determine how much the signal strength is reduced.

The NR Path loss model incorporates indoor hotspots, indoor factories with clutters, urban macrocells, urban microcells, and rural microcells, under 3GPP TR 38.901 requirements. It takes into consideration altitude dependence by using roadway widths varying from 5 to 50 m and building heights. The model is designed for an operating frequency of 3.5 GHz, ensuring accurate characterization of signal propagation in different environments is shown in the Eqs. (1) and (2)

$$\begin{aligned} PL_{Drone_LOS} &= \{PL_1, \sigma_{SF} 10 \leq d \leq d_{BP},\}, \\ &\{PL_2, \sigma_{SF} d_{BP} \leq d \leq 10\,km\} \end{aligned} \tag{1}$$

$$PL_{Drone_NLOS} = \max\left(\left(PL_{Drone_LOS}, PL_{NLOS}\right), \sigma_{SF}\right) \tag{2}$$

2.1 5G NR Path Loss Configuration

NR Path Loss model is given as in the Eqs. (3) and (4)

$$\begin{aligned} PL_1 = \{&20\ \log_{10}(40\,\pi\, df_c/3) + \min\left(0.003\,h^{1.72}, 10\right)\log_{10}(d) \\ &- \min\left(0.044\,h^{1.72}, 14.77\right) + 0.002\ \log_{10}(h)d\} \end{aligned} \tag{3}$$

$$PL_2 = PL_1(2\pi\, h_{BS} h_{Drone} f_c/C) + 40\ \log_{10}\left(\frac{d}{d_{BP}}\right) \tag{4}$$

$\boldsymbol{PL_{Drone_LOS}}$—Path loss for Line–of–Sight (LOS) conditions, $\boldsymbol{PL}_1$–Path loss equation for shorter distances, $\boldsymbol{PL}_2-$ Path loss equation for longer distance, $\boldsymbol{d}-$ is the distance between drone and the base station(gNB), $\boldsymbol{d_{BP}}$—is the breakpoint distance where the path loss characterstics change, $\boldsymbol{\sigma_{SF}}$—shadow fading standard deviation, $\boldsymbol{PL_{Drone_NLOS}}$—Path loss for Non Line–of–Sight (LOS) conditions, $\boldsymbol{PL_{NLOS}}$—Path

loss in NLOS conditions (higher due to obstruction and reflection). These are the parameters given in Eq. (1–4).

2.2 Particle Filtering for PRS Signals

State Space Model

Where x and y are positions, and v_x and v_y are velocities in the x and y directions, respectively shown in Eqs. (5–7).

State Vector

$$X_k = \left[x, y, v_x, v_y\right]^T \tag{5}$$

State transition model

$$X_K = AX_{k-1} + W_k \tag{6}$$

where A is the state transition matrix

Measurement model

$$Z_K = HX_k + V_k \tag{7}$$

where H is the observation matrix

Particle filter implementation combines measurements from multiple gNBs using hyperbola intersections and refines the position estimate over time, accounting for process and measurement noise.

2.3 Particle Positioning Algorithm for Improving the PRS

Input: Initial set of N particles—$\left\{X_0^i\right\}_{i=1}^N$ with equal weights $w_0^i = \frac{1}{N}$.

Output: Estimated state $\hat{x}$ at each time step.

Initialize:

Generate N particles—$\left\{X_0^i\right\}_{i=1}^N$.

Set initial weights—$w_0^i = \frac{1}{N}$.

For each time step k.

a. Prediction

Explore new particle states according to the system model

$$X_k^i = AX_{k-1}^i + w_k^i \tag{8}$$

b. update

Figure out the importance of weights

$$w_k^i = p(w_k^i / X_k^i) \tag{9}$$

c. Normalize weights

Verify the weights add up to 1 by normalizing them.

$$w_k^i = \frac{w_k^i}{\sum_{j=1}^{N} w_k^i} \tag{10}$$

d. Resampling

Calculate an effective number of particles.

$$N_{eff} = \frac{1}{\sum_{j=1}^{N} (w_k^{i2})} \tag{11}$$

e. State Estimation

Estimate the state as the weighted mean of particles.

$$\hat{x} = \sum_{j=1}^{N} w_k^i X_k^i \tag{12}$$

3 Results and Discussion

The simulation results of NR Positioning with Positioning Reference Signals (PRS) using Software-defined radio provide the system performance of transmission and reception of real-time PRS signals of drone location, showing accurate localization with an estimation error of 2.6 m. It also involved generating a Spectrum trace and correlation plot for analyzing the signal quality for evaluating the system performance. The accuracy of the estimation error is improved by the particle filter algorithm. Figure 2 displays the gNB and drone locations. Throughout the estimating process, the three base station positions are fixed. The plotting of the actual droneposition is also done.

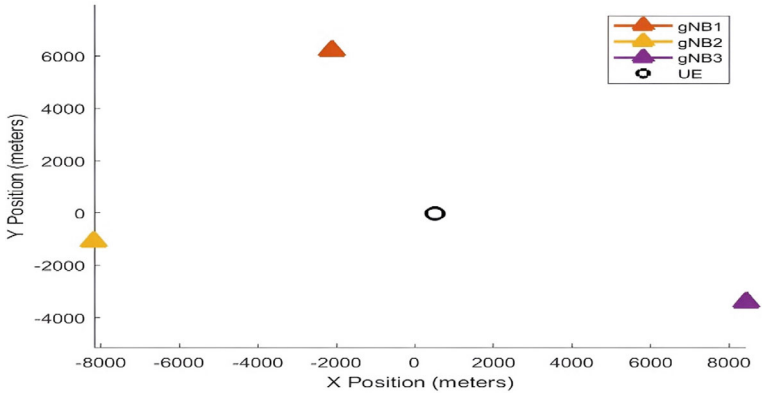


Fig. 2 gNB and dronelocations

The PRS and PDSCH resource allocation across several gNBs in the 5G NR carrier grid is depicted in Fig. 3. Figure 4 represents the element configureation of PRS resource. It emphasizes how PRS transmission must be carefully planned and coordinated to allow precise positioning while preserving communication services Physical Downlink Shared Channel (PDSCH). For methods like Observed Time Difference of Arrival (OTDOA), the drone must be able to correlate PRS signals from each gNB independently, which is ensured by the unique allocation for each gNB shown in Eq. (13).

$$S(n) = \frac{1}{\sqrt{2}} \sum_{k=0}^{N-1} X_k e^{j2\pi kn/N} \tag{13}$$

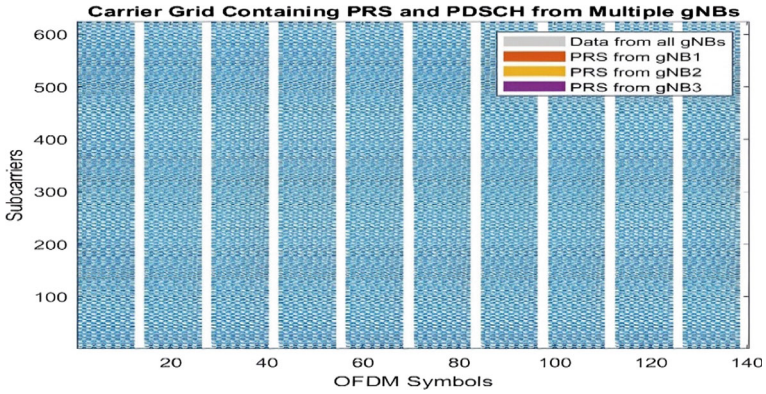


Fig. 3 Carrier grid containing PRS and PDSCH from multiple gNBs

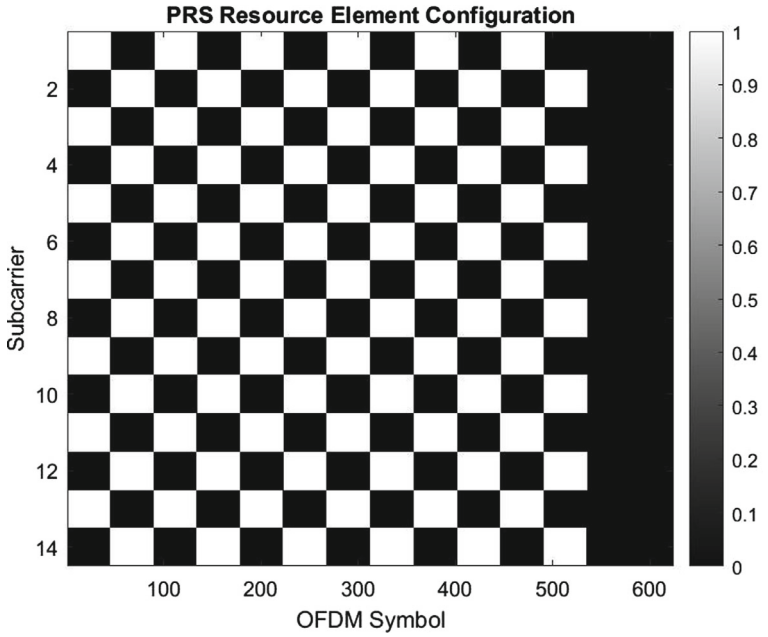


Fig. 4 PRS resource element configuration

where, X_k are mapped as PRS symbols, N is the FFT size, and $s(n)$ is the time domain signal.

PRS are allocated in a time–frequency grid shown in Fig. 4 and Eq. 14 given below

$$S_{PRS}(k, l) = x(k, l).\delta(k, l) \tag{14}$$

where, $S_{PRS}(k, l)$—are subcarrier and symbol indices, $x(k, l)$—represents PRS symbols,

The power spectral density of positioning reference signal (PRS) waveforms transmitted by three gNBs is shown in Fig. 5.The x-axis represents the frequency in MHz, and the y-axis represents the power level in dBm/Hz. The three different colors indicate signals from the gNBs. The left and right sides of the high-power regions show the transmission bands. The dip in the middle shows a frequency range with minimal signal power. This ensures efficient frequency usage and reduces interference. Figure 6. shows the transmitted spectrum plot, where the x-axis represents the frequency in GHz, and the y-axis represents signal power in dB. The flat region shows the constant power indicating the transmission bandwidth of the signal. The filtering effect limits the signal to a specific frequency range and reduces interference with neighboring bands.

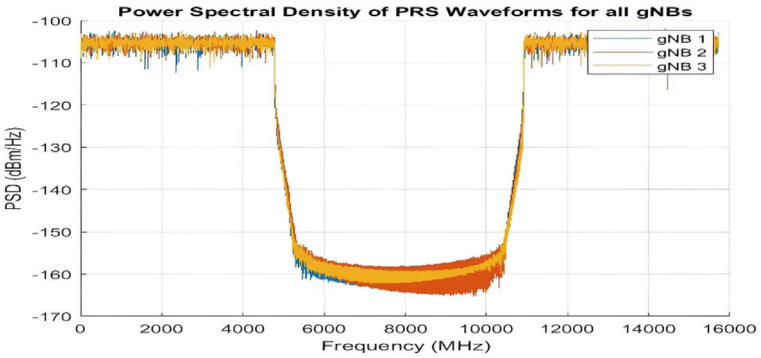


Fig. 5 Power spectral density of PRS waveforms for all gNBs

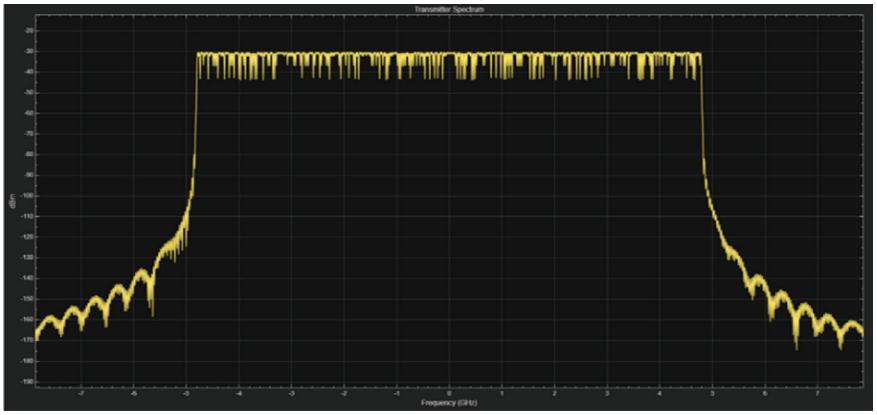


Fig. 6 Transmitted spectrum of the gNBs

The correlation of positioning reference signals received from three gNBs over time is shown in Fig. 7. The x-axis represents time in seconds, and the y-axis represents the absolute value of the correlation. The color represents the unique cell ID. A higher peak suggests the strongest signal, meaning the PRS signals from that gNB are more aligned with the expected signal.

Figure 8 shows the received spectrum plot, where the x-axis represents the frequency in GHz, and the y-axis represents signal power in dB. The flat region in the middle indicates the main frequency band of the signal, where most of the power is concentrated. This type of plot is useful for analyzing signal quality and ensuring proper transmission and reception in wireless communication systems.

Figure 9 represents the locations of three reference base stations (gNBs) are indicated by the triangle marks known as gNB1, gNB2, and gNB3.According to the Time Difference of Arrival (TDOA) hyperbolas computed between pairs of gNBs.

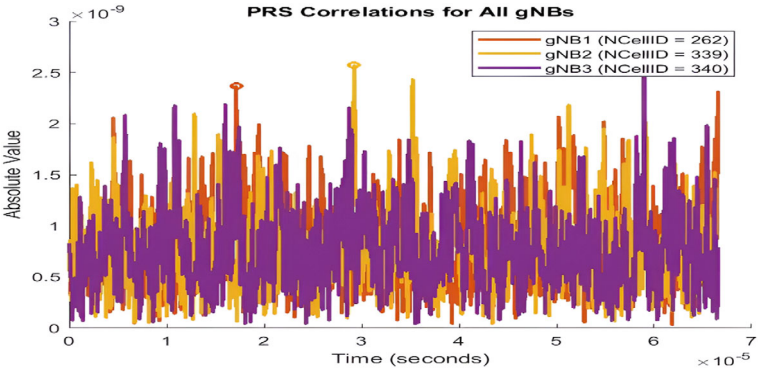


Fig. 7 PRS correlations for all gNBs

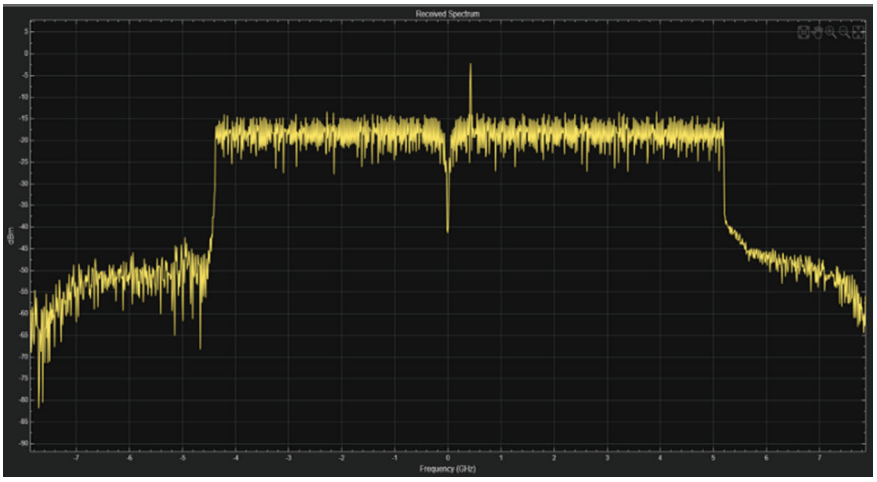


Fig. 8 Received spectrum at the drone under multipath condition

Indicating the drone estimated position is the red ‘ + ’ marker. Drone exact position is indicated by the black ‘o’ symbol.

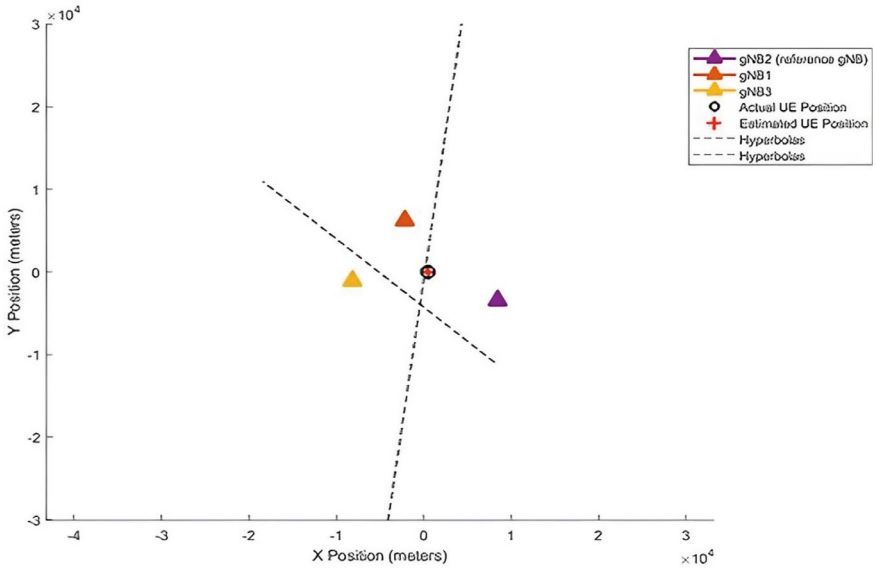


Fig. 9 Position estimation of drone with partcile filtering

4 Conclusion

The current study produces a result for positioning technology design using USRP hardware and customised drone. Investigating and experimentally validating cell-specific and positional reference signal receivers under varied time and frequency offsets is the first task. Fine frequency and delay estimations with quantitative bounds in a single transmitter–receiver link are the anticipated outcomes. Techniques like cross-correlation for signal estimation and hyperbolic curve fitting for position calculation are employed to avoid signal interference. Position estimation error of 2.6 m. The novelty of this chapter is using a particle filtering algorithm, improved location estimation over time, and accuracy of the drone positioning improved significantly.

Acknowledgements We would like to Thank and Acknowledge IIT Tirupati Navavishkar I-Hub Foundation (IITTNiF), National Mission on Interdisciplinary Cyber-Physical Systems (NM-ICPS), Department of Science and Technology (DST), India for their financial support (SASTRA-IITTNiF/5GNP/2024-25) and SASTRA Deemed to be University, India, for providing the necessary resources and support.

References

1. Renaudin O, López Salcedo JA, Seco-Granados G, Lapin I, Zanier F, Ries L (2021) TOA error bounds for positioning in 5G new radio networks. In: Proceedings of the 34th ınternational technical meeting of the satellite division of the ınstitute of navigation (ION GNSS+ 2021), pp 2940–2956
2. Jin C, Bajaj I, Zhao K, Tay WP, Ling KV (2021) 5G positioning using code-phase timing recovery. In: 2021 IEEE Wireless communications and networking conference (WCNC). IEEE, pp 1–7
3. Abdallah AA, Kassas ZM (2021) UAV navigation with 5G carrier phase measurements. In: Proceedings of the 34th ınternational technical meeting of the satellite division of the ınstitute of navigation (ION GNSS+ 2021), pp 3294–3306
4. Chen L, Zhou X, Chen F, Yang L-L, Chen R (2021) Carrier phase ranging for indoor positioning with 5G NR signals. IEEE Internet Things J 9(13):10908–10919
5. Jin C, Tay WP, Zhao K, Ling KV, Sin KK (2022) A 5G/GNSS integrated positioning method. In: Proceedings of the 35th ınternational technical meeting of the satellite division of the ınstitute of navigation (ION GNSS+ 2022), pp 2429–2443
6. Müürsepp I, Kulmar M, Elgarhy O, Alam MM, Chen T, Horsmanheimo S, Scholliers J (2021) Performance evaluation of 5g-nr positioning accuracy using time difference of arrival method. In: 2021 IEEE International mediterranean conference on communications and networking (MeditCom). IEEE, pp 494–499
7. Gonzalez-Garrido A, Querol J, Chatzinotas S (2023) 5g positioning reference signal configuration for integrated terrestrial/non-terrestrial network scenario. In: 2023 IEEE/ION position, location and navigation symposium (PLANS). IEEE, pp 1136–1142
8. Jin C, Tay WP, Zhao K, Ling KV, Lu J, Wang Y (2023) A sub-meter accurate positioning using 5G double-difference carrier phase measurements. In: 2023 IEEE/ION position, location and navigation symposium (PLANS). IEEE, pp 1176–1183
9. Wang J, Zhao W, Jiang J, Zeng Z, Shi F (2023) 5G PRS positioning based on the ımproved sparrow search algorithm. In: 2023 IEEE/CIC ınternational conference on communications in China (ICCC Workshops). IEEE, pp 1–6
10. Liu Q, Gao C, Xhafa A, Gao W, López-Salcedo JA, Seco-Granados G (2023) Performance analysis of GNSS+ 5G hybrid positioning algorithms for smartphones in urban environments. IEEE Trans Instrum Meas 73:1–9
11. Campolo F, Blaga A, Rea M, Lozano A, Costa-Pérez X (2024) 5GNSS: fusion of 5G-NR and GNSS localization for enhanced positioning accuracy and reliability. IEEE Trans Vehicul Technol
12. Xhafa A, del Peral-Rosado JA, López-Salcedo JA, Seco-Granados G (2021) Evaluation of 5G positioning performance based on UTDoA, AoA and base-station selective exclusion. Sensors 22(1):101
13. Luo K, Deng Z, Guo X, Ma Z, Liu J (2023) High-precision ranging method of 5G NR co-band PRS in industrial internet scenarios. Appl Sci 13(18):10302
14. Deng Z, Zheng X, Zhang C, Wang H, Yin L, Liu W (2020) A TDOA and PDR fusion method for 5G indoor localization based on virtual base stations in unknown areas. IEEE Access 8:225123–225133
15. Butt MM, Rao A, Yoon D (2020) RF fingerprinting and deep learning assisted UE positioning in 5G. In: 2020 IEEE 91st vehicular technology conference (VTC2020-Spring), IEEE, pp 1–7

Design and Optimization of Hybrid Circuits for Low-Power Systems Using GPDK 90nm



B. P. Bhuvana, Bharathi Raju, Parasa Gnan Naga Sai Vikas, and Tarun Varma Pedada

1 Introduction

As the demand for more efficient and sustainable electronic devices continues to rise, the need for low-power circuits has become a primary focus in modern electronics. This is especially true for devices like smartphones, wearables, and the expanding network of IoT devices, where power efficiency is essential for extended battery life and optimal functionality. In response to this challenge, traditional circuit design approaches are being pushed to their limits, leading to the exploration of hybrid circuit architectures that combine the best of both analog and digital technologies to meet the growing need for power-efficient solutions. Hybrid circuits, which integrate analog and digital components, offer a compelling solution by leveraging the unique advantages of each domain [1–3].

Conventional circuit design methods, though effective in earlier times, are now facing limitations in addressing the rising demand for energy-efficient solutions. To overcome these challenges, hybrid circuit architectures which integrate aspects of both analog and digital design have gained attention as a viable strategy for enhancing power efficiency [4, 5]. These hybrid systems leverage the advantages of both domains: analog circuits offer high precision and efficiency for continuous signal processing, while digital circuits provide speed, reliability, and suitability for complex operations. By combining these strengths, hybrid designs make it possible to lower power consumption without compromising the performance required for today's advanced applications [6].

Over time, several logic styles have been employed in 1-bit full adder designs, including standard CMOS, dynamic CMOS, complementary pass-transistor logic

B. P. Bhuvana (✉) · B. Raju · P. G. N. S. Vikas · T. V. Pedada
Department of Electronics and Communication Engineering, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India
e-mail: bhuvanabp1509@gmail.com

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_4

(CPL), double pass transistor logic (DPL), and transmission gate logic [7–9]. While each technique has strengths and weaknesses, this paper presents a Hybrid Full Adder that intelligently combines the most effective elements from these approaches. By leveraging these optimizations, the proposed design significantly improves power efficiency, performance, and transistor utilization, offering a refined solution for modern digital circuits [10–12]. This paper explores the potential of hybrid circuits in overcoming these challenges, offering insights into their design principles, methodologies, and application in low-power systems. The increasing reliance on portable, energy-dependent devices has intensified the demand for circuits that deliver high performance with minimal power usage [13–15].

The paper is structured as follows: Sect. 2 describes the existing full adder circuit design, Sect. 3 presents the proposed circuit design and Sect. 4 discusses the simulation results, followed by the conclusion in Sect. 5.

2 Existing Circuit Design Styles

A full adder's output is primarily influenced by the XOR-XNOR logic circuit. The summing circuit facilitates the logical addition of three binary bits, while the carry circuit produces the associated carry bits. This section explores two different existing full adder designs, each constructed using a unique XOR-XNOR logic style module, where A, B, C are the inputs and SUM and COUT are the outputs.

$$\text{SUM} = \text{A} \oplus \text{B} \oplus \text{C}$$
$$\text{COUT} = \text{AB} + \text{BC} + \text{CA}$$

2.1 Full Adder-1

The XOR-XNOR logic module is a fundamental building block in the design of the full adder-1 circuit, as depicted in Fig. 1. This module plays a crucial role in defining the functionality, efficiency, and overall performance of the Sum and Carry circuits. By employing CMOS logic, the XOR-XNOR module ensures low power consumption, high-speed operation, and reliable switching characteristics. The primary function of this module is to generate the essential logic operations required for computing the sum and carry outputs in a full adder [2].

In a CMOS-based implementation, the XOR and XNOR gates are designed to achieve a full voltage swing at the outputs, thereby enhancing signal integrity and minimizing propagation delays. The XOR-XNOR logic is realized using a six-transistor configuration, making it a compact and efficient solution for arithmetic circuits. In this six-transistor setup, different input combinations determine the

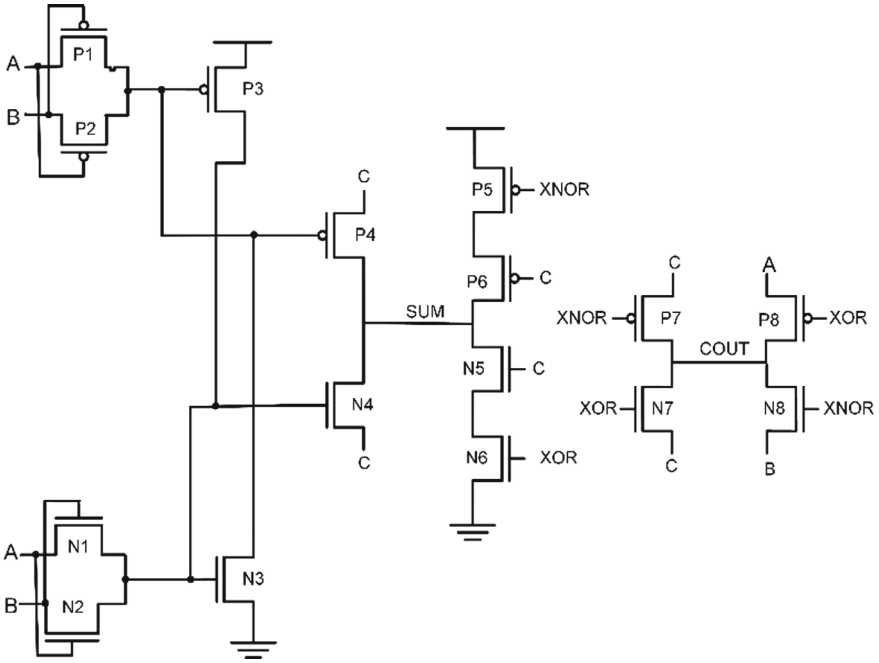


Fig. 1 Full adder circuit design-1

behavior of the nMOS and pMOS transistors. For instance, when the input combination is "01," the nMOS transistor (N2) allows a logic "0" to pass through, while the corresponding pMOS transistor (P2) transmits a logic "1." This switching behavior ensures that the XOR and XNOR outputs are generated correctly. Likewise, when the input condition is "00," the transistors P1 and P2 work together to produce a logic low ("0") output. This precise transistor operation is essential for maintaining the correct functionality of the XOR-XNOR module, which, in turn, directly influences the accuracy of the full adder circuit.

By integrating this XOR-XNOR logic module within the full adder design, the circuit can effectively perform binary addition with minimal transistor count and optimized power efficiency. The compact six-transistor implementation reduces hardware complexity while ensuring fast and reliable operation. Consequently, this module is indispensable in modern digital arithmetic circuits, playing a key role in enhancing the overall performance of computational systems.

2.2 *Full Adder-2*

The Full Adder-2 design, illustrated in Fig. 2, is constructed using an XOR-XNOR logic module, which serves as the core component in determining the overall performance of both the Sum and Carry circuits. The incorporation of CMOS logic in this design plays a crucial role in enhancing circuit efficiency, reducing power dissipation, and optimizing operational speed. The XOR-XNOR logic module employed in Full Adder-2 is specifically designed to be compact and efficient, utilizing only 12 transistors, which contributes to improved performance metrics compared to conventional designs [2].

One of the key advantages of this design is its ability to minimize delay while maintaining strong driving capabilities. When the input states are either "00" or "11," the circuit exhibits superior delay characteristics, ensuring faster transitions and better signal integrity.

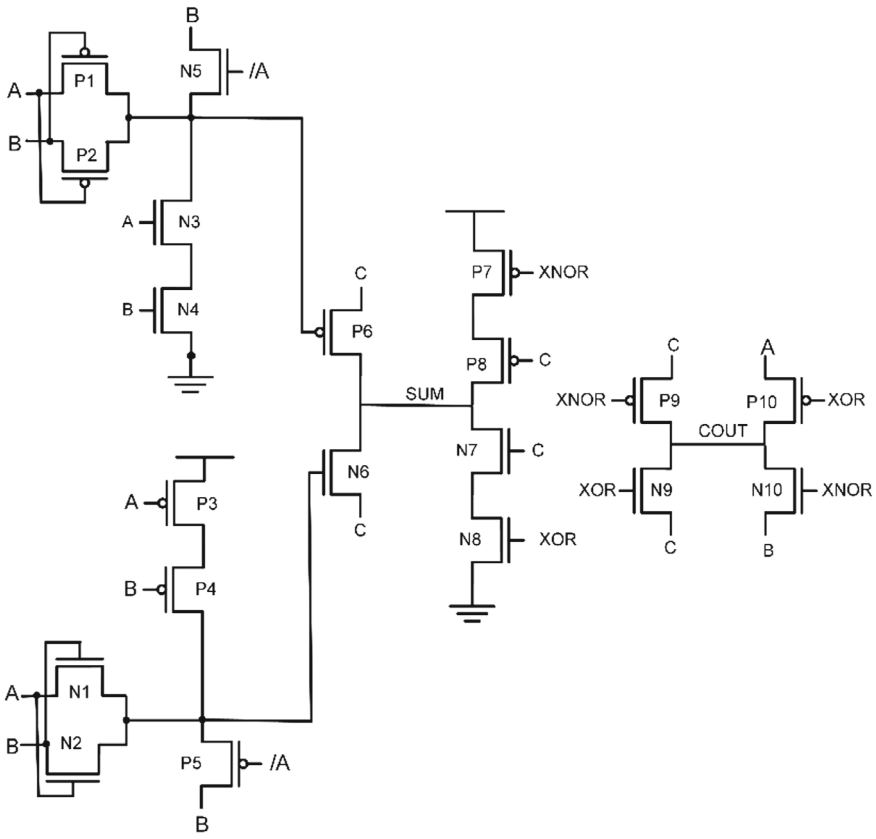


Fig. 2 Full adder circuit design-2

For an input of "01," the XOR and XNOR logic gates experience a complete voltage swing at their outputs, primarily driven by the activation of the P1 and N2 transistors. Similarly, when the input state is "10," the output undergoes a full transition due to the action of the P2 and N1 transistors, ensuring stable and reliable logic operations. As the input sequence transitions from "00" to "01," both P1 and P2 transistors are partially activated, while the N5 transistor is fully engaged. This results in a well-defined logic "0" at the XOR output, confirming the efficiency of the transistor switching mechanism in producing accurate logic responses.

Although the Full Adder-2 circuit employs a total of 22 transistors, which is slightly higher than the Full Adder-3 configuration, the design prioritizes reduced propagation delay over power consumption. The intentional trade-off between power efficiency and speed allows this architecture to be well-suited for high-performance applications where fast computations are essential.

The transistor arrangement in Full Adder-2 is strategically optimized to achieve a balance between efficiency, power consumption, and operational speed. By utilizing a well-structured XOR-XNOR module with a minimal transistor count, this design demonstrates a significant improvement in circuit responsiveness. Such advancements make Full Adder-2 a preferred choice for applications requiring rapid data processing, including digital signal processing, high-speed arithmetic circuits, and advanced computing architectures.

3 Proposed Hybrid Full Adder Circuit Design

The given CMOS Full Adder circuit is designed using XOR-XNOR logic modules, which play a fundamental role in determining the efficiency of Sum and Carry operations. The circuit shown in Fig. 3 utilizes an optimized transistor configuration to achieve faster signal propagation while keeping power consumption at an acceptable level.

Initially, the inputs A and B are fed into the XOR-XNOR logic block, generating complementary signals that facilitate both Sum and Carry computation. The XOR output of A and B is subsequently combined with the third input, C, through an additional XOR gate to produce the Sum output. This ensures that the Sum follows the expected Boolean equation, $\text{Sum} = A \oplus B \oplus C$, providing accurate results across all input states. On the other hand, the Carry-out signal is derived using the equation $\text{Cout} = AB + C(A \oplus B)$, which ensures that a high output is generated whenever two or more of the inputs are high. The transistor arrangement within the Carry generation block ensures minimal delay in switching, making the circuit suitable for high-speed arithmetic operations. Notably, when input states are either "00" or "11," the circuit exhibits a highly optimized behavior, reducing propagation delay significantly. In contrast, for input states "01" and "10," the XOR output undergoes a full voltage swing, ensuring stable signal transmission.

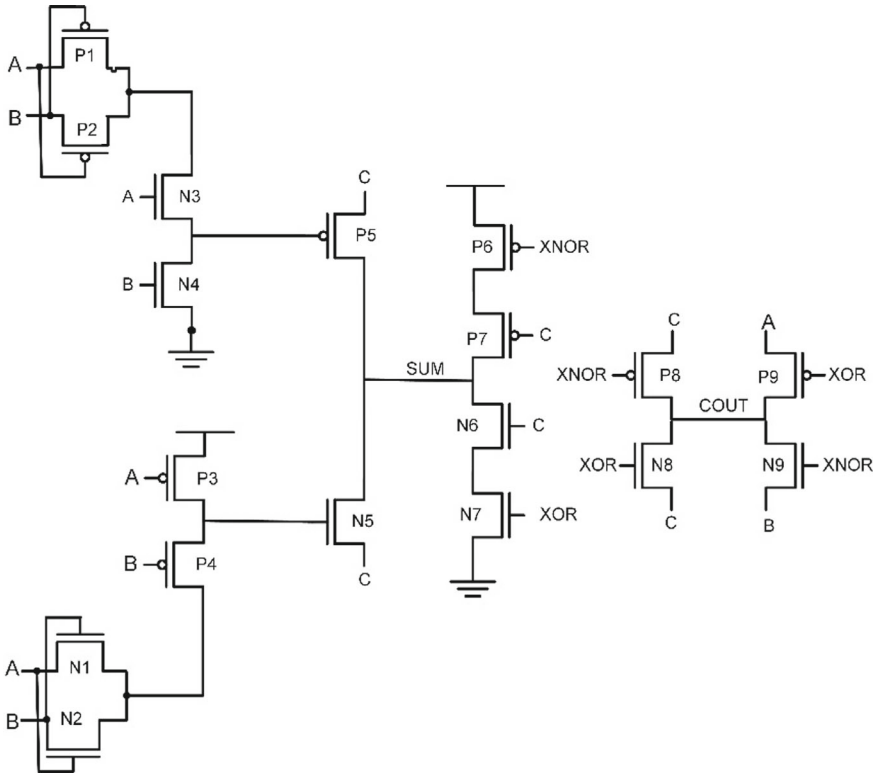


Fig. 3 Proposed hybris full adder circuit design

As the input transitions from one state to another, certain transistors partially activate while others are fully engaged to maintain logic integrity. This careful management of transistor activity prevents excessive power dissipation while ensuring rapid state transitions. Although the design incorporates a slightly higher number of transistors compared to some other Full Adder architectures, it makes a trade-off between power efficiency and speed, making it an excellent choice for high-performance applications.

The integration of XOR-XNOR logic modules enhances the circuit's robustness, making it highly effective for VLSI implementations where speed and reliability are critical. The use of CMOS technology further contributes to reducing power leakage and enhancing overall circuit stability. By balancing transistor count, delay reduction, and driving capabilities, this Full Adder configuration achieves an optimal balance between power and performance, making it a suitable choice for modern digital circuit design.

4 Simulation Results and Discussion

The simulation is carried out using the industry-standard Cadence® Virtuoso EDA tools, which provide a comprehensive environment for efficient circuit design and analysis. The Virtuoso Schematic Editor is utilized to facilitate fast and accurate schematic entry, ensuring a streamlined design process. In this study, we implement the circuit using GPDK 90 nm technology, which offers a balance between performance and power efficiency. The supply voltage is set to 0.9V to maintain stable operation while minimizing power consumption. Additionally, the temperature is consistently maintained at 27 °C throughout all circuit simulations to ensure uniform testing conditions.

Figure 4 presents a comparative analysis of power consumption across three different Full Adder designs at varying supply voltages, highlighting the efficiency of the proposed design. FA-1 and FA-2, both taken from base papers, exhibit significantly higher power consumption compared to the proposed Full Adder. FA-2, in particular, demonstrates the highest power usage among the three, making it less suitable for low-power applications. As the supply voltage decreases from 0.9V to 0.6V, the power consumption of all designs reduces, but the proposed Full Adder consistently outperforms both base designs by consuming the least power at every voltage level. Notably, the proposed design operates in the nW range at lower voltages, making it highly efficient for low-power VLSI applications. This substantial reduction in power consumption indicates that the proposed Full Adder is a more optimized and energy-efficient alternative, particularly for battery-operated and portable electronic devices where power savings are crucial.

The efficiency analysis reveals that the proposed Full Adder significantly reduces power consumption compared to the existing designs. At a supply voltage of 0.9V, the proposed design achieves an impressive 78% reduction in power consumption

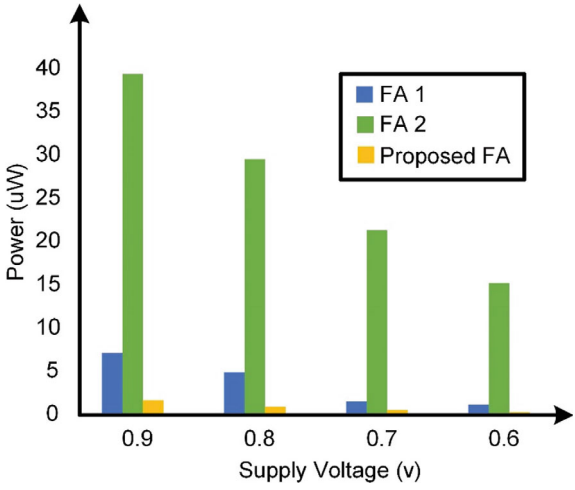


Fig. 4 Power versus voltage variation of proposed and existing circuit styles

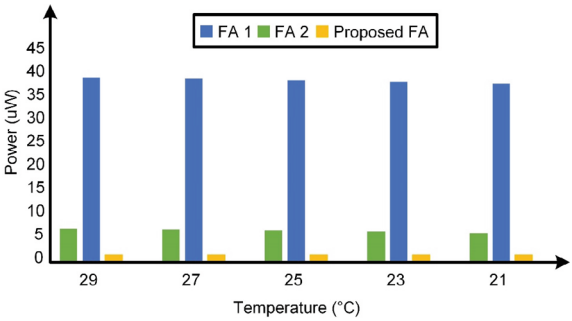


Fig. 5 Power versus temperature variation of proposed and existing circuit styles

compared to FA-1 and an even more substantial 91% reduction compared to FA-2. At a supply voltage of 0.6V, the proposed design achieves an impressive 86% reduction in power consumption compared to FA-1 and an even more substantial 94% reduction compared to FA-2.

Figure 5 represents the power consumption of three different Full Adder designs FA-1, FA-2, and the Proposed FA under varying temperature conditions ranging from 29 °C to 21 °C. The results indicate that as the temperature decreases, the power consumption of all designs also reduces. However, the proposed Full Adder consistently demonstrates significantly lower power consumption across all temperature variations compared to FA-1 and FA-2, making it a highly efficient alternative for low-power applications.

At 29°C, FA-1 consumes 7.016 μW, FA-2 exhibits a much higher consumption of 39.428 μW, whereas the Proposed FA operates at a remarkably lower power level of 1.475 μW. The proposed Full Adder significantly reduces power consumption across all temperature conditions, showing 78.98% improvement over FA-1 and 96.26% over FA-2 at 29 °C. At 21 °C, the efficiency remains high with a 76.51% reduction compared to FA-1 and 96.36% compared to FA-2. This demonstrates the robustness of the proposed design in maintaining low power consumption under different environmental conditions, making it highly suitable for temperature-sensitive and low-power VLSI applications such as wearable devices, IoT systems, and portable electronics.

Table 1 depicts the Transistor count comparison of the full adder designs. The proposed full adder design maintains a good trade-off between complexity and efficiency. Although FA-1 uses fewer transistors, it might not deliver the desired performance or power optimization. On the other hand, FA-2 increases the transistor count, which can lead to higher power and area consumption. By utilizing 18 transistors, the proposed design captures the benefits of both existing approaches. It ensures reliable performance while keeping resource usage within practical limits.

Table 1 Comparison of transistor count of various full adder designs

Circuit designs	FA-1	FA-2	Proposed design
Transistor count	16	20	18

The proposed full adder design minimizes leakage power through efficient transistor-level optimization and reduced overall transistor count. By avoiding unnecessary stacking and using hybrid logic gates strategically, the design limits leakage paths during idle states. This is particularly beneficial in ultra-low-power applications where standby power is a major concern, ensuring that the circuit remains energy-efficient even when not actively switching.

5 Conclusion

In this paper, a highly efficient Full Adder design has been proposed and analyzed for power consumption across different supply voltages and temperature variations. The results indicate that the proposed design significantly outperforms conventional Full Adders (FA-1 and FA-2) in terms of power efficiency. At 0.9V, the proposed Full Adder achieves a 78.91% reduction in power consumption compared to FA-1 and an impressive 96.29% reduction compared to FA-2. Similarly, at 0.6V, the efficiency improves further, reaching 89.21% and 99.30% reductions, respectively. The temperature variation analysis also demonstrates the robustness of the proposed design, maintaining lower power consumption across different conditions. At 29 °C, the power efficiency improves by 78.98% over FA-1 and 96.26% over FA-2, while at 21 °C, it achieves reductions of 76.51% and 96.36%, respectively. These results validate the effectiveness of the proposed circuit in minimizing power dissipation, making it highly suitable for low-power VLSI applications. The reduced power consumption is crucial for energy-efficient computing, particularly in battery-operated devices, IoT applications, and portable electronics. Future work may focus on extending this approach to multi-bit adders and further optimizing the circuit for advanced low-power applications.

References

1. Bhattacharyya P, Kundu B, Ghosh S, Kumar V, Dandapat A (2014) Performance analysis of a low-power high-speed hybrid 1-bit full adder circuit. IEEE Trans Very Large Scale İntegrat (VLSI) Syst 23(10):2001–2008
2. Karuppanan P, Rafik M (2023) Design and analysis of full adder and full subtractor circuit using hybrid logic style. In: 2023 International conference on modeling, simulation and ıntelligent computing (MoSICom). IEEE, pp 203–208
3. Gomes SV, Sasipriya P, Kanchana Bhaaskaran VS (2015) A low power multiplier using a 24-transistor latch-adder. Indian J Sci Technol 8(19):1–5
4. Poornima N, Kanchana Bhaaskaran VS (2015) Area efficient hybrid parallel prefix adders. Proc Mater Sci 10:371–380
5. Ramkumar E, Gracin D, Rajkamal P, Bhuvana BP, Kanchana Bhaaskaran VS (2020) Design and analysis of low power and high speed FinFET based hybrid full adder/subtractor circuit (FHAS). In: 2020 IEEE International symposium on smart electronic systems (iSES). IEEE, pp 281–284

6. Aranda ML, Báez R, Diaz OG (2010) Hybrid adders for high-speed arithmetic circuits: a comparison. In: 2010 7th International conference on electrical engineering computing science and automatic control. IEEE, pp 546–549
7. Bhuvana BP, Bhaaskaran VK (2019) Design of FinFET-based energy efficient pass-transistor adiabatic logic for ultra-low power applications. Microelectron J 1(92):104601
8. Nikoubin T (2019) Hybrid logical effort for hybrid logic style full adders in multistage structures. IEEE Trans Very Large Scale Integrat (VLSI) Syst 27(5):1138–1147
9. Dhathri D, Jagadeeswara Rao E, Nandan D (2021) A systematic review on various types of full adders. In: Proceedings of ınternational conference on recent trends in machine learning, IoT, smart cities and applications: ICMISC 2020. Springer, Singapore, pp 635–646
10. Giustolisi G, Palumbo G (2022) Hybrid full adders: optimized design, critical review and comparison in the energy-delay space. Electronics 11:19 3220
11. Bhuvana BP, Kanchana Bhaaskaran VS (2017) Performance analysis of 2N-N-2P adiabatic logic circuits for low power applications using FinFET. Proc Comput Sci 115:166–173
12. Ramkumar E, Gracin D, Rajkamal P, Bhuvana BP, Kanchana Bhaaskaran VS (2020) Design and analysis of low power and high speed FinFET based hybrid full adder/subtractor circuit (FHAS). In 2020 IEEE International symposium on smart electronic systems (iSES)(Formerly iNiS). IEEE, pp 281–284
13. Wang Y, Pai C, Song X (2002) The design of hybrid carry-lookahead/carry-select adders. IEEE Trans Circuits Syst II: Analog Digit Signal Process 49(1):16–24
14. Sarma R, Raju V (2012) Design and performance analysis of hybrid adders for high speed arithmetic circuit. vol 1207(2840). pp 12
15. Thamizharasan V, Kasthuri N (2023) High-speed hybrid multiplier design using a hybrid adder with FPGA implementation. IETE J Res 69(5):2301–2309

Smart Border Security: The Role of AI in Modern Border Surveillance System



Simant Kamal Dutta, Rahul Mahala, Rajesh Singh, Anita Gehlot, and Nagendar Yamsani

1 Introduction

Border security is main concern for all the countries now a days. The main reasons behind these concerns are terrorist activities, illegal migrants and smuggling which causes threat to the security of borders. Illegal migrants include unauthorized travelers, tourists, migrants or criminals. The smuggling includes the passing off goods from a border to another by illegal means, goods may be drugs, weapons, goods which are counterfeited etc. Main threat to the security of the borders is human intruding with weapons, hence these must be detected on the border areas to maintain peace and harmony [1]. The human intrusion is a big challenge in front of many countries in the world. When it comes to traditional border patrolling, there has been seen that the different patrolling teams does surveillance on a specific border area for specific period. Human involvement is a very big challenge in securing the borders which are lengthy in areas or where the human survival is not possible [2]. Traditional border surveillance is based on humans in which different troops are assigned to protect different areas. These troops ensure the security of the borders on predetermined areas and intervals. The checkpoints are constructed on the roads to detect whether there is any enemy, drugs or any other illegal activity by the means of the transport. In this kind of system chances of human errors are high. Hence, multiple

S. K. Dutta
School of Applied and Life Sciences, Uttaranchal University, Dehradun, Uttarakhand, India

R. Mahala (✉)
Law College Dehradun, Uttaranchal University, Dehradun, Uttarakhand, India
e-mail: rahulmahala98@gmail.com

R. Singh · A. Gehlot
Uttaranchal Institute of Technology, Uttaranchal University, Dehradun, Uttarakhand, India

N. Yamsani
School of Computer Science and Artificial Intelligence, SR University, Warangel, India

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_5

technologies which can help each other in patrolling and surveillance are required for these areas [3]. As the terrorist activities, drugs, illegal activities and illegal transportation is a big problem to deal with in border areas. Constant surveillance and monitoring are required all the time on the border areas. But this is done by using the traditional ways, it requires assets and human power in a large quantity. Hence, the demand of time is automated surveillance systems which can run on their own without any human power [4]. As the technology is advancing day by day, Industry 4.0 is milestone in securing the public domain. Artificial intelligence which is a part of Industry 4.0 is largely growing now a days in technical prospective. Them law enforcement agencies are using AI for data analysis and data interpretation, it is being also used for the criminal detection, traffic control and other legal processes [5]. Artificial intelligence is also being used for monitoring and surveillance of the areas. According to GSI (Global Surveillance Index) approx. 75 out of world's 176 countries are investing in AI for the surveillance purpose. Governments have already used AI into the surveillance cameras, smartphones and softwares of video management [6]. This study expresses the role of Artificial Intelligence in border surveillance of the remote areas. The role of machine learning, natural language processing and computer vision in border surveillance of remote areas is evaluated in this study by reviewing the existing articles. The authors have reviewed the limitations of using artificial intelligence in surveillance system for detecting border intrusion.

2 Methodology

The methodology for this review article involved an extensive review of 109 papers sourced from IEEE, Elsevier, MDPI, Springer, Wiley and others. These papers were carefully selected based on their credibility, relevance and alignment with the article's focus. After a thorough evaluation, 18 articles were selected and incorporated to ensure a comprehensive analysis of the topic. The selected papers provide robust insights, contributing significantly to the findings and discussions presented in the article.

3 Overview of Artificial Intelligence in Border Surveillance

"Artificial Intelligence" was first coined at Dartmouth University in 1956. Development of AI has made benefits to mankind in economical ways and all other aspects of life too. The general term for science of artificial intelligence is AI. Computers are being used by AI to emulate human behaviors and teaches the computers to learn the behavior of humans like decision-making, judgements etc. Artificial intelligence is collection of philosophy, biology, psychology, logic, computer science and many other subjects and it has gained significant results in processing of image, recognition of speech, NLP, automatic theorem proving and intelligent robots as shown

in Fig. 1 [7]. Machine learning, Natural language processing and computer vision are the main emerging parts of AI now a days. AI has a deep impact on our day-to-day life. Different kind of artificial intelligence systems are being largely used in different kind of applications of great importance in human lives like medical diagnosis, drug discovery, self-driving cars, traffic management, national securities, defense, fraud detection, cyber security, autonomous weapon systems, aerospace, aviation etc. which is taking on the way to an algorithmic society [8]. When it comes to the defense and national security, borders have great importance in national security and defense for a nation. Securing these borders is an intricate job which includes crossover associations, consideration of advance technologies and efficient professional sets. Although, here are many factors which are affecting the deployment of such systems on borders.

The governments need such effective systems which can secure borders and also ensures low costing in doing so. Training the guards with equipment is costly for border security and surveillance [9]. Artificial intelligence has great potential to help in defense applications, it can provide help to the defense forces in all possible conditions of warfare like on lands, water or in air. Artificial intelligence finds its defense applications in autonomous vehicles, autonomous drones, cybersecurity, cyber warfare, intelligence, surveillance, reconnaissance, combat support, decision support, training, simulation and many other fields as shown in Fig. 2 [10]. The border

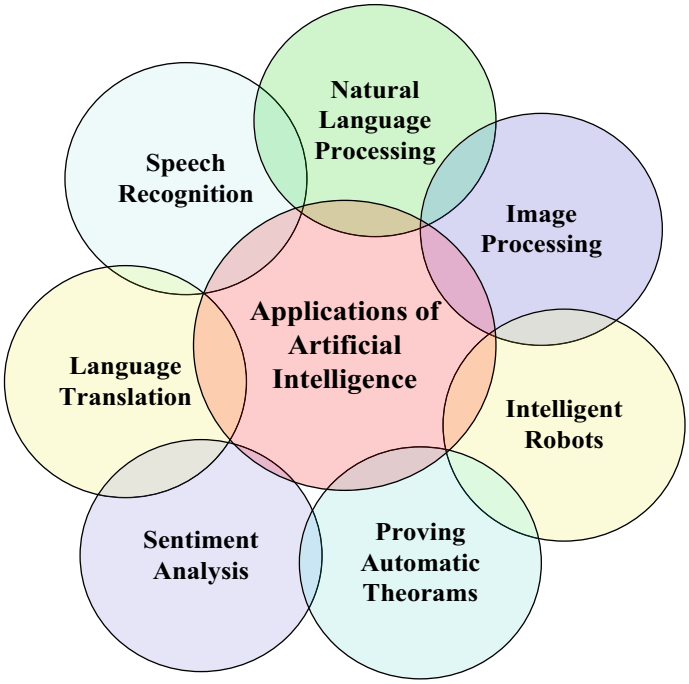


Fig. 1 Applications of artificial Intelligence

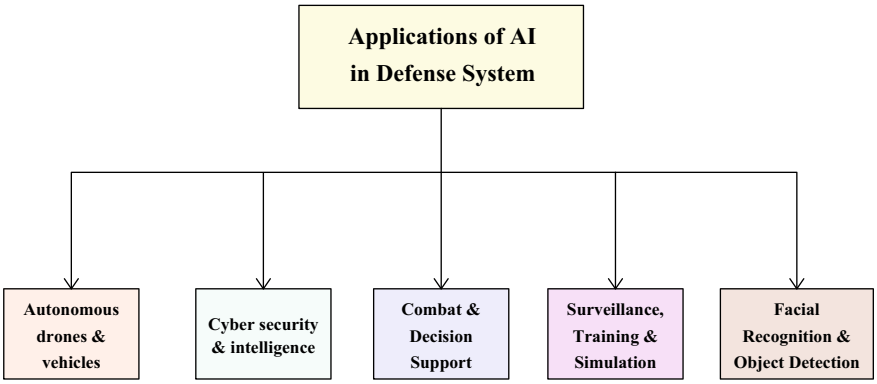


Fig. 2 AI applications in defense system

existing surveillance systems are mainly using different-2 structures and hence are not interlinked with each other to share the information to each other and also with security agencies. There are many border areas where human survival is not possible and hence such systems are which should operate autonomously and can share the information independently to all other systems and security agencies so that they can ensure the security of borders from intrusions and other security threats like smuggling and terrorist activities [11].

4 Machine Learning and Natural Language Processing in Border Surveillance of Remote Area

While talking about AI machine learning and natural language processing are fastest growing subsets of AI. Machine learning is statical techniques collection which is used for building the mathematical tools such that it can draw judgements out of the sample data. There are mainly three kinds of machine learnings i.e. supervised, unsupervised and reinforced learning which can be used by the person based on his requirements. Supervised learning is used when the sample data is labeled, unsupervised is used when there is data which is not labeled and reinforced learning is used when the result learned heads towards an action which changes the environment [12]. Natural language processing is used to make the human and computer conversation easier. Human language meaning and syntax is learnt by the machine so that it can give output to the users. Natural process language makes the computer system such that they can provide outcomes in natural or language understandable by human. This is important because it takes inputs as image, speech, text and manipulates or process them as per the algorithms provided to it, then gives the outcome as audio or text as per the requirement [13]. There are different kinds of algorithms used

in NLP such as sorting algorithms, search algorithms, graph algorithms, regression algorithms, link analysis algorithms and many more [14].

Kabir et al. [15] in their paper has introduced the limitations of using machine learning and natural language processing in border security. Some uses are also shown in Fig. 3 of ML and NLP in border security. They used machine learning systems which includes XGboost and Random Forest to build strong predictive analytics system. They created profiles of the travelers by data analysis from the data provided by them in history. They trained their system to assess the risk associated with each traveler and this approached to a better border control of that area. They also deployed chat bots which uses the natural language processing. Hence, better assistance to the travelers about visa and assistance to their problems were provided by those chatbots. This not only streamlined the process but also changed the traveler experience in good way by providing them assistance guidance and support.

5 Border Surveillance of Remote Areas Using Computer Vision

Advancement in artificial intelligence last few years has emerged computer vision as a subset of artificial intelligence. Computer vision helps the computer to obtain and analyze the information from the visual data or the inputs. Hence, this made computer vision to be useful for the subjects like border security, border surveillance, maritime security, terrorism etc. shown in Fig. 4. which causes threat to the integrity of the nations. There are certain reasons behind these border intrusions like terrorism, smuggling etc. [16]. The computer vision methods now a days are applicable to be used in all kind of drones.

By applying these applications of the computer vision in these drones, these drones be able to use not only for picture clicking or film making but also be used for more complex processes. Computer vision focuses upon the participation of the environment as well as the basic applications of navigations, surveillance, machine inspection and 3D model building [17]. A border surveillance system by using computer vision was introduced by Kumar and others [18] in their paper, in which digital camera, Respburry pi, transmitter, Aurdino and buzzer were used to make a border surveillance system as per shown in Fig. 4. Digital camera takes the pictures of the live movement and sends it Respburry pi, in which computer vision is used to detect the positive and negative among the objects and faces by using python coding. Negative detection is sent to transmitter by pi module and transmitter transmits it to the receiver. Receiver passes the negative signal received to the Aurdino. Aurdino sends signal to the buzzer and buzzer makes an alert as per the programmed language or sound. Hence, the security personals get alerted about the possible threats which the negative object or the face can make and the border hence saved from intrusions or the illegal activities in a fast manner. This system can be used to enhance the border security and surveillance in remote areas where the human survival is not possible.

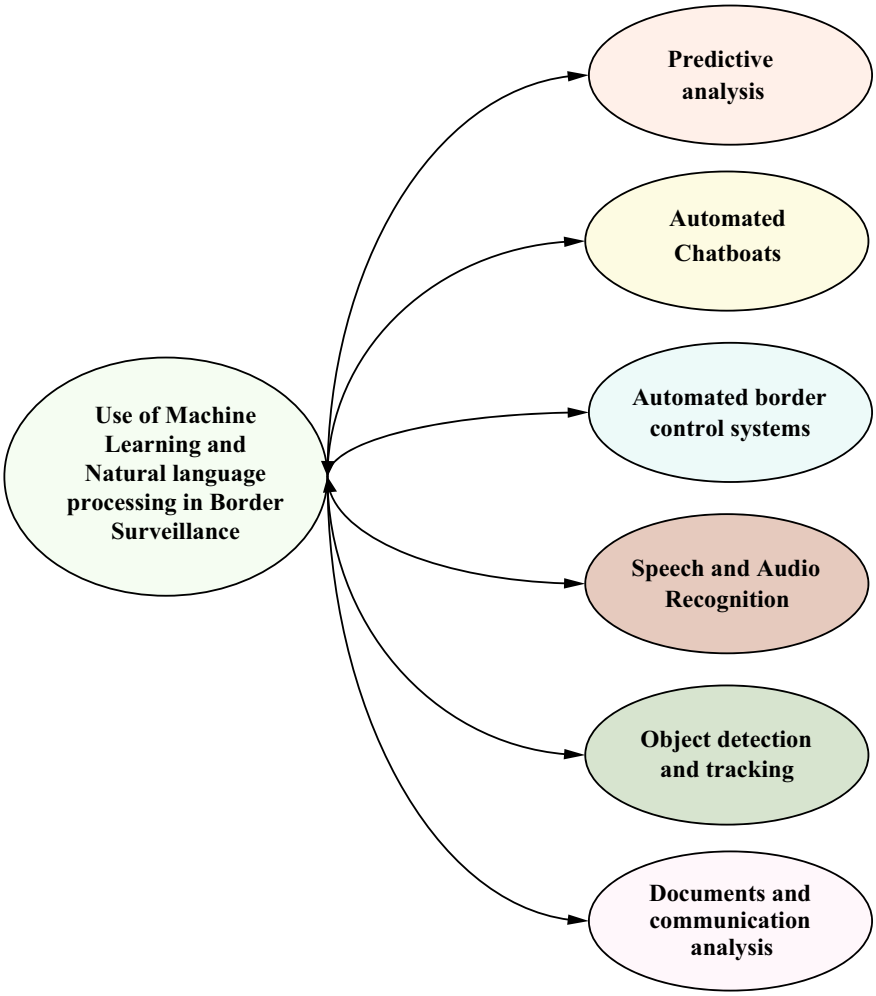


Fig. 3 Uses of ML and NLP in border surveillance of remote areas

6 Recommendations

Border surveillance is a serious issue now a days to look upon. Border areas are prone of terrorist activities, smuggling and other illegal activities and hence requires all time surveillance to ensure the security of nation. It is impossible to do surveillance of the areas by the traditional means where human survival is not possible. There are huge areas which we can say are remote areas where the survival of the human is not possible like some of the hilly areas which are always covered with snow and oxygen level of these areas is very low due to the reasons of height and also there is no connectivity of roads or any other infrastructure and hence the basic amenities to

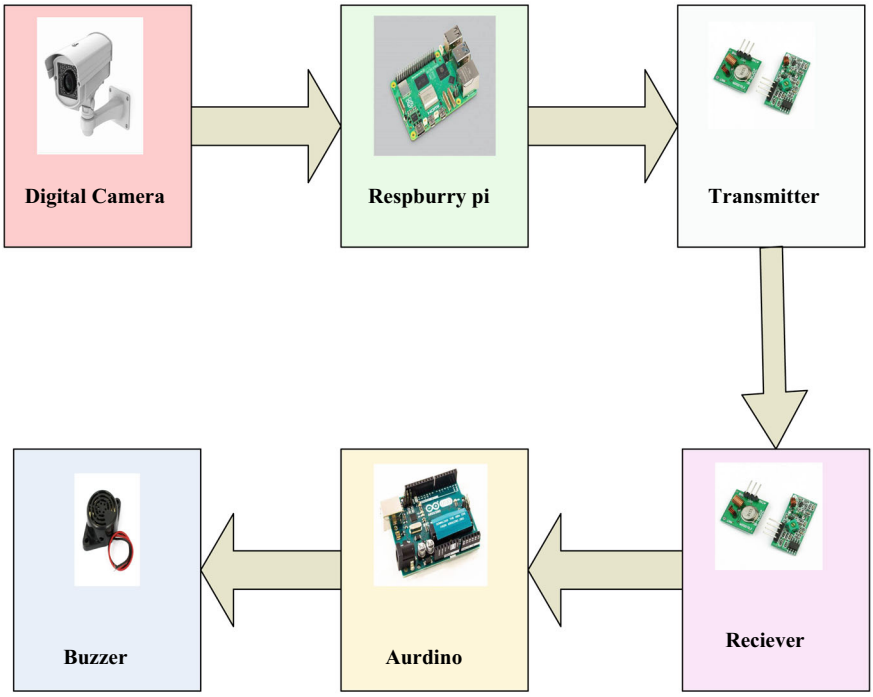


Fig. 4 Working of border surveillance system using computer vision

live can't be provided to the soldiers guarding these kinds of areas. Hence, border surveillance and security of these areas is difficult by the means of the man power. Also, where there are borders present which are lengthy in area, the surveillance of these areas in traditional way is done by different groups of soldiers at different-different times. The probability of human error is very high in doing surveillance in this way for the lengthy areas as the chances of human error are also very high in these areas. Hence, technological intervention in these areas can provide a great relief to the soldiers and also can better safeguard these areas. AI is emerging field of technology in many other subjects such as legal, social science, security etc. It can help the forces to do surveillance of the areas where the human survival is not possible or in the lengthy border areas or also in the plane areas by providing assistance in security and surveillance. Below are some recommendations which can be used to safeguard the border areas-

- To enhance the ability of the AI based smart surveillance systems, the use of multiple equipment and technologies should be ascertained such that they can ensure the security and surveillance of the border. For example, different kind of sensors, radars, thermal cameras, motion detectors etc. should be used to improve the detection accuracy of the AI based surveillance system for the borders.

- AI based surveillance system should be fitted with the advanced data fusion techniques so that it can take data from all the devices at once. This can improve the decision making of the system by providing the vast view of the targeted area whose surveillance is done.
- While doing facial recognition or license plate recognition by the means of AI system deployed on the border areas should be done ensuring that they will not harm the individual rights of anyone. These should be made such they don't make any harm to the personal rights of the persons i.e. the privacy should be maintained.
- AI algorithms of AI based surveillance system should be trained in a continuous manner such that they can know the new harms or threats as per the need of the time.AI based system for border surveillance should perform real-time analytics and should have predictive capabilities such that it can identify the threat timely and can help the security personals and agencies by giving potential threats of security to them.
- AI powered border fence and perimeter should be used which can ensure the motion detection, vibration detection on the surveillance areas in which these systems are working.

7 Conclusion

Security and surveillance are crucial concerns in border areas worldwide, particularly in areas where human intervention is impossible or inaccessible. In this context, AI-based systems can be used for border security and surveillance in remote or normal plane areas to remove human errors and protect against intrusion, terrorists, and smugglers. These systems are cost-efficient, less time-consuming, and eliminate human errors. They integrate machine learning, natural language processing, and computer vision to detect unusual activities and alert border agencies and other security agencies. The advantages of AI-based surveillance systems include unauthorized movement tracking, autonomous detection of threats like illegal border crossings and terrorist activities, and extended accuracy in harsh and isolated environments where human survival is not possible. However, there are also challenges such as privacy concerns, ethical use of AI, and algorithm bias. Privacy rights of individuals are a major concern, and regular improvement in algorithms is necessary to address these concerns. Despite these challenges, AI-based surveillance systems can effectively reduce threats in border areas, but concerns such as privacy and algorithm bias need to be considered. Regular improvements in algorithms are necessary to ensure the effectiveness of these systems in remote areas. AI in border surveillance is revolutionizing national security by providing real-time threat detection and predictive monitoring. The system enhances efficiency by automating the analysis of large datasets gathered from various sources such as drones, sensors, and cameras. AI-driven systems expedite decision-making by reducing human error and response time. Overall, AI-based systems offer potential solutions for border security and

surveillance in remote areas, but they must be carefully considered and improved to ensure the safety and security of all involved.

References

1. Goyal A, Anandamurthy SB, Dash P, Acharya S, Bathla D, Hicks D, Ranjan P (2020) Automatic border surveillance using machine learning in remote video surveillance systems. In: Emerging trends in electrical, communications, and information technologies: proceedings of ICECIT-2018. Springer, Singapore, pp 751–760
2. Arjun D, Indukala PK, Menon KU (2017) Border surveillance and intruder detection using wireless sensor networks: a brief survey. In: 2017 International conference on communication and signal processing (ICCSP), April, IEEE, pp 1125–1130
3. Sun Z, Wang P, Vuran MC, Al-Rodhaan MA, Al-Dhelaan AM, Akyildiz IF (2011) BorderSense: border patrol through advanced wireless sensor networks. Ad Hoc Netw 9(3):468–477
4. Bhadwal N, Madaan V, Agrawal P, Shukla A, Kakran A (2019) Smart border surveillance system using wireless sensor network and computer vision. In: 2019 international conference on automation, computational and technology management (ICACTM), April, IEEE, pp 183–190
5. Raaijmakers S (2019) Artificial intelligence for law enforcement: challenges and opportunities. IEEE Secur Priv 17(5):74–77
6. Saheb T (2023) Ethically contentious aspects of artificial intelligence surveillance: a social science perspective. AI and Ethics 3(2):369–379
7. Zhang C, Lu Y (2021) Study on artificial intelligence: the state of the art and future prospects. J Ind Inf Integr 23:100224
8. Kaur D, Uslu S, Rittichier KJ, Durresi A (2022) Trustworthy artificial intelligence: a review. ACM Comput Surv (CSUR) 55(2):1–38
9. Hammoudeh M, Al-Fayez F, Lloyd H, Newman R, Adebisi B, Bounceur A, Abuarqoub A (2017) A wireless sensor network border monitoring system: deployment issues and routing protocols. IEEE Sens J 17(8):2572–2582
10. Bistron M, Piotrowski Z (2021) Artificial intelligence applications in military systems and their influence on sense of security of citizens. Electronics 10(7):871
11. Essendorfer B, Monari E, Wanning H (2009) An integrated system for border surveillance. In: 2009 Fourth international conference on systems, March, IEEE, pp 96–101
12. Ribeiro M, Grolinger K, Capretz MA (2015) Mlaas: Machine learning as a service. In: 2015 IEEE 14th international conference on machine learning and applications (ICMLA), December, IEEE, pp 896–902
13. Fanni SC, Febi M, Aghakhanyan G, Neri E (2023) Natural language processing. Introduction to artificial intelligence. Springer International Publishing, Cham, pp 87–99
14. Wang Y, Zhang C (2020) Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing. J Informet 14(4):101091
15. Kabir MS, Sumi EJ, Alam MN (2023) Artificial intelligence (AI) and future immigration and border control
16. Ige T, Kolade A, Kolade O (2022) Enhancing border security and countering terrorism through computer vision: a field of artificial intelligence. In: Proceedings of the computational methods in systems and software. Cham, Springer International Publishing, pp 656–666
17. Akbari Y, Almaadeed N, Al-Maadeed S, Elharrouss O (2021) Applications, databases and open computer vision research from drone videos and images: a survey. Artif Intell Rev 54:3887–3938
18. Kumar D, Aadarsh S, Kumar H (2020) Border surveillance system using computer vision. In: 2020 6th international conference on advanced computing and communication systems (ICACCS), March, IEEE, pp 623–628

Multi-modal Data-Driven User Behavior Modeling and Personalized Advertisement Recommendation Algorithm



Zhang Zheng and Xu Dongyan

1 Introduction

With the rapid development of the Internet, the Internet online advertising [1, 2] industry is booming. In online advertising, advertising platforms sell exposure by signing a guaranteed delivery contract with advertisers. At the same time, advertisers stipulate in the contract the target group with certain characteristics to which the advertisement needs to be delivered, as well as the total exposure requirements for the group. Object-level advertising delivery uses the image processing technology to extract key frames from short videos and detect them, and deliver advertisements based on the detection results. The key frame extraction methods include: (1) The method based on the shot boundary uses frame averaging to extract the key frames of each shot, or directly uses the first frame and the last frame as the key frames; (2) The method based on motion analysis extracts key frames only at the static points of the video.

The main challenges of the current studies are:

1. Users' access or search behaviors are significantly affected by various uncertain covariates (such as weather conditions, temperature, humidity, etc.).
2. Advertising strategies need to be determined before uncertain covariate information and exposure are realized. In the absence of accurate information, advertising platforms must make advertising strategies in advance.

Z. Zheng (✉)
Shandong Vocational University of Foreign Affairs Vocational University, Shandong, China
e-mail: zhangpeiai365864@163.com

X. Dongyan
School of Information Engineering, Shandong Vocational University of Foreign Affairs, Shandong, China

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_6

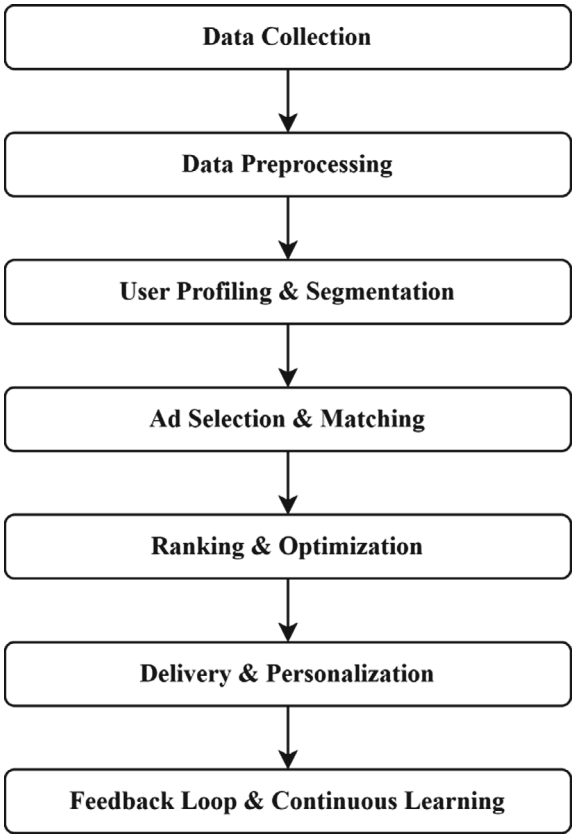


Fig. 1 The typical advertisement recommendation algorithm steps

To deal with these challenges, this study proposes the novel framework for multi-modal data-driven user behavior modeling and personalized advertisement recommendation algorithm. In the following Fig. 1, the typical advertisement recommendation algorithm steps are summarized.

2 Review of Recommendation Algorithms

In the research of the recommendation systems, the latest related algorithms are analyzed and integrated in deep learning, fairness analysis, user simulation and network construction. Zhao et al. [3] proposed a stable advertising insertion framework based on deep reinforcement learning, which can balance the connection between recommended content and advertising revenue. This study designed a deep Q-network architecture that can dynamically decide whether to insert advertisements,

select the optimal advertisements and their positions. This framework can help maximize revenue in the long run. Zhao et al. [4] developed a unified recommendation system library called RecBole. The library contains 73 recommendation models and 28 benchmark datasets. At the same time, it supports GPU accelerated execution and automatic parameter tuning. Melchiorre et al. [5] studied the gender fairness model in music recommendation algorithms. This study introduced the LFM-2b dataset to evaluate the performance differences of the various collaborative filtering algorithms in different gender user groups.

Zhang et al. [6] proposed Agent4Rec. This is a user simulator based on a large language model. The simulator explores the potential of generative agents in recommendation tasks by simulating user behavior. Li et al. [7] explored the fairness issue in recommendation systems from the user's perspective. The model divides users into active and inactive groups and finds that the existing system's recommendation quality for the active users is significantly higher than that for inactive users. Therefore, the study proposed a new reranking algorithm to alleviate this unfairness. Santos et al. [8] studied the general impact of link recommendation algorithms on polarization dynamics in online social networks. The study proposed that link recommendations based on structural similarity will inadvertently exacerbate opinion polarization, while connecting the nodes with dissimilar structures will help mitigate extreme views. Hui et al. [9] proposed a personalized recommendation system based on knowledge embedding and historical behavior, named the ReBKC. The system optimizes the system by integrating knowledge graph and the user behavior data. Experimental results significantly improve the accuracy of the recommendation algorithm. Zhao et al. [10] introduced RecBole 2.0, an extended recommendation system library. The database covers cutting-edge topics such as meta-learning, data enhancement, debiasing, fairness, and cross-domain recommendations. Pajkovic [11] analyzed the operating logic of the Netflix's recommendation system through reverse engineering. The study revealed the economic and cyclical logic of the algorithm in shaping the taste of film and television culture. Xiao et al. [12] proposed a general offline reinforcement learning framework. The framework is used to learn interactive recommendation strategies from log feedback. The study demonstrated that five distribution matching methods can significantly improve the overall recommendation performance without online exploration.

3 The Advertising Multimodality and Internet User Behavior Analysis

The multimodal information [13, 14] of advertisements is the basis of the user Internet behavior. From the perspective of dimensions, multimodality mainly includes the following aspects: language text analysis, visual image analysis, and the interaction between language modality and image modality. In the Table 1, the principles are summarized.

Table 1 The principles of the multimodal information of advertisements

Multimodal	Principle
Language text analysis	In terms of conceptual functions, empirical functions and logical functions are used to analyze what processes the language text elements in the advertisement belong to, what semantic relationships exist between the elements, how these processes and relationships are manifested, what they each mean, and what intentions they want to convey
Visual image analysis	What kind of processes do the visual elements in the advertisement belong to? How are these processes manifested? What do they mean? What do they want to convey?
Interaction between language modality and image modality	From the perspective of semantic logical relationship, we can examine whether the image and text are in an "extension" relationship or a "projection" relationship. If they are an extension relationship, is it an elaboration, extension or enhancement? If they are a projection relationship, is it a verbal projection or a thought projection? How are these specifically manifested, what do they mean, and what intentions do they want to convey?

Internet users' interaction with multimodal advertising information is the basis of behavioral analysis. In this analysis process, the user intent modeling is crucial. Users' decision-making behaviors usually stem from the different decision intents, and different intents will position a series of corresponding products. Therefore, user intention is formally defined in the recommendation framework. In a recommendation system, the interaction data between users and items can be represented as a matrix $X \in R^{m \times n}$, where the m and n represent the number of users and items, respectively. Whether there is interaction between user u and item i can be defined as follows:

$$\begin{cases} x_{u,v} = 1, & \text{has interaction} \\ x_{u,v} = 0, & \text{no interaction} \end{cases} \tag{1}$$

Since user intentions under different topics are associated with different latent factors, it is necessary to model a decoupled user latent intention representation as $z_u = R^d$, Considering the diversity of intentions, the representation of the user potential intentions can also be expressed in a decoupled form, that is:

$$z_u = \left[z_u^1, z_u^2, z_u^3, ..., z_u^T\right] \tag{2}$$

where the T is the number of topics, and all items in the same recommendation scenario share these T topics. $z_v^t \in R^{d/T}$ is the user's intention representation under the t-th topic.

4 The Personalized Advertisement Recommendation Algorithm

4.1 Multi-modal Context Modeling for Recommender Systems

Some traditional recommendation algorithms, such as the traditional collaborative filtering algorithms, model the user's rating vector as a non-time series model. The interaction between users and items is considered to have no context. The user's interest is only represented by an implicit feature vector in the modeling process, and the recommendation algorithm will calculate the similarity between the item and the implicit feature vector based on this implicit vector. In fact, the interaction between users and items has different meanings at different times. That is to say, for an interaction vector between a user and an item, different positions of the vector have different weights. At the same time, there are also deeper implicit features related to serialized information such as time. Serialized information can reflect the dynamic characteristics of user behavior and preferences. In terms of time dimension, serialized modeling is more practical than traditional vectorized modeling. In the study of serialization modeling in this section, the task of sequential recommendation or ad click-through rate prediction is abstracted as predicting the k + 1th behavior based on the user's previous k behaviors. The behavior sequence obtained is the previous k behaviors. When the k-th behavior occurs, the k-th behavior has system context information at the time of occurrence, such as the page information of the user clicking the ad, the current page link, the specific display page details, etc. These are all commonly used processing methods for existing ad click-through rate prediction algorithms based on context information. In the Fig. 2, the dynamic context information is shown.

The context information involved in this section mainly includes two aspects:

1. **Static context**. Some existing studies on static context have attempted to input advertising content information and behavior content information as the context features into the model input.

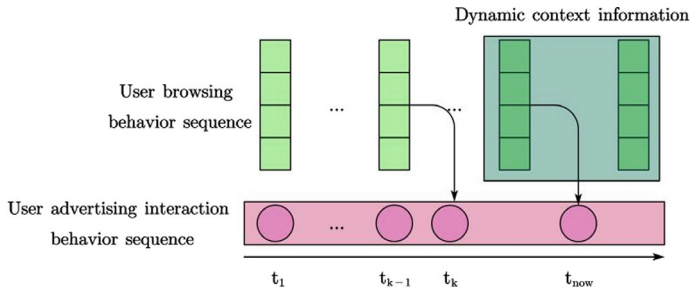


Fig. 2 The dynamic context information modelling

2. **Dynamic context**. Dynamic context involves the problem of the context selection, because the user's browsing information is complex and rich. Compared with the interaction sequence between users and advertisements, the user's other browsing behavior records are complex and diverse.

4.2 The Personalized Advertisement Recommendation Algorithm

The personalized recommendation system builds a detailed user portrait by deeply mining and analyzing user data, and achieves accurate positioning of users. This precise positioning technology provides advertisers with more accurate audience targeting, greatly improving the effectiveness of advertising. The recommendation algorithm proposed in this model is based on transfer learning [15, 16]. In order to predict the possibility of multiple behaviors with the same input information, a separate prediction layer is given to each behavior of click, add to cart, favorite, and purchase for learning. Let h_k be the prediction layer of the k-th behavior, then the probability of user u taking the k-th behavior on item v is estimated as:

$$\hat{R}_{(k)uv} = h_k^T (p_u \times q_v) \tag{3}$$

In order to capture the migration characteristics, we assume that the prediction of a certain behavior depends on the prediction of its predecessor. Inspired by the migration mechanism in knowledge representation, this model assumes that the relationship between two behaviors can also be described as a transformation in the representation space. Then the migration relationship between two different behaviors h_i and h_j can be expressed as:

$$trans_{h_i \to hj_i} = h_i M_{ij} + b_{ij} \tag{4}$$

where the $M_{ij} \in R^{d \times d}$ is the migration matrix that projects h_i from the i-th behavior space to the j-th behavior space. The b_{ij} is the bias vector. The transfer learning process of the proposed is shown in the Fig. 3.

After determining the feature words of the page text and establishing the user interest model, the similarity between the two can be calculated. However, if the

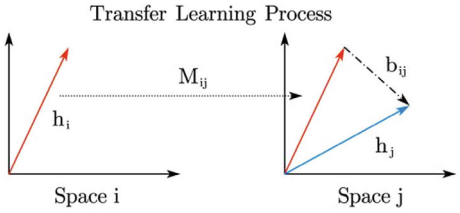


Fig. 3 The transfer learning process of the proposed model

dimensions of the page feature vector and the user interest model are not the same, the similarity calculation cannot be performed. For this reason, it is necessary to choose the temperature of the two so that they have the same number of the dimensions. In practice, if a feature word in the user interest set does not exist in the page text, the weight value corresponding to the feature word is set to 0, and the feature word ranking is adjusted according to the weight. After a series of processing, the two correspond to each other in dimension. The similarity function is defined as formula 5.

$$s(U, P) = \frac{\sum_{i=1}^{n} W(F_i) \times w(i)}{\sqrt{\sum_{i=1}^{n} W(F_i)^2} \times \sqrt{\sum_{i=1}^{n} w(i)^2}} \tag{5}$$

A threshold q is set in advance. If the similarity s(U, P) < q, it means that the page cannot represent the user's interest point, and the feature word corresponding to the page text should be discarded. Then, by calculating the similarity value in a loop, the proposed model can stably obtain the advertising recommendation results efficiently.

5 The Experimental Verification

5.1 The Experimental Settings

This test object uses video ads as the main analysis mode. The system adjusts the ad weights recommended by the system based on the user's behavior when watching short videos. The specific setting process of the behavior level is as follows:

(1) When the system does not push any ads while playing a video, but the user actively triggers the target detection operation to observe an object in the video, it has the highest level of weight;
(2) When the video is playing, the user sees the ad and clicks the ad link, this process has the second highest weight;
(3) When the video is playing, the user sees the ad push but does not perform any operation until the ad ends, this process has the third highest weight;
(4) When the video is playing, the user sees the ad and clicks to close the ad, this process has the fourth highest weight.

At the same time, the recommendation system adjusts the weights according to the user's preferences and sets the user's interest threshold and disinterest threshold through repeated trials. Table 2 shows the weight adjustment process.

Table 2 The advertisement user preference weight setting

User action	Priority	Weight adjustment
Action (1)	1	20
Action (2)	2	10
Action (3)	3	0
Action (4)	4	− 10
Others	5	− 20

5.2 The Numerical Simulation Result and Discussion

Based on the above analysis and parameter settings, this test uses 100 short videos with a duration of 300 s to test the effectiveness of the recommendation algorithm proposed in the article. These 100 short videos include 25 different object categories, and the proposed algorithm can match these objects and advertising information in turn. According to different market scopes, this test studies the recommendation accuracy of the proposed algorithms [6–8]. The results are shown in the Table 3.

From the comparative experimental results, we can see that the four methods all show some fluctuations in videos of the different lengths. The overall accuracy is maintained above 90%. From the quantitative experimental results, the proposed method maintains a leading position in all time intervals, with the lowest accuracy of 94.65% (220–240s) and the highest accuracy of 98.42% (60–80s). This shows

Table 3 The comparative test results of recommendation accuracy for videos of different lengths (%)

Video length	Proposed	[6]	[7]	[8]
0–20	97.32	91.07	93.55	94.83
20–40	95.21	90.12	92.78	95.41
40–60	96.87	92.34	94.10	96.02
60–80	98.42	93.89	95.67	97.23
80–100	97.15	91.56	93.74	95.88
100–120	94.73	89.92	92.13	94.12
120–140	95.68	90.76	93.05	94.87
140–160	96.54	91.84	93.98	95.34
160–180	97.89	92.61	94.65	96.41
180–200	96.23	91.47	93.32	95.76
200–220	95.12	90.34	92.74	94.62
220–240	94.65	89.78	91.89	93.45
240–260	96.02	91.15	93.07	94.98
260–280	97.33	92.48	94.41	96.21
280–300	98.11	93.25	95.12	97.02

that the method has good adaptability to both short videos (0-20s) and long videos (280–300s), and performs best in the 60–80s time interval.

6 Conclusion

Multi-modal data-driven user behavior modeling and personalized advertisement recommendation algorithm is studied in this manuscript. This algorithm improves the accuracy of the advertising recommendation system and the user interaction experience by integrating multi-modal information such as text, vision and statistics. In order to deal with the uncertainty factors in user behavior, this study adopts an accumulation learning strategy. This strategy enables the model to adapt to the behavior patterns of different user groups, thereby improving the robustness and generalization ability of the algorithm. In addition, this paper also explores the key technologies of multi-modal information fusion. This operation improves the interpretability and accuracy of advertising recommendations through cross-modal data association analysis. Experimental results verify the effectiveness of this method in improving advertising delivery efficiency and optimizing user response rate. The contribution of this study is not only to improve the performance of the advertising recommendation system, but also to provide new ideas for the development of personalized advertising delivery technology in the future.

References

1. Wuisan DS, Handra T (2023) Maximizing online marketing strategy with digital advertising. Startupreneur Business Digital (SABDA Journal) 2(1):22–30
2. Lim WM, Gupta S, Aggarwal A, Paul J, Sadhna P (2024) How do digital natives perceive and react toward online advertising? Implications for SMEs. J Strategic Market 32(8):1071–1105
3. Zhao X, Gu C, Zhang H, Yang X, Liu X, Tang J, Liu H (2021) Dear: deep reinforcement learning for online advertising impression in recommender systems. In: Proceedings of the AAAI conference on artificial intelligence, vol 35(1). pp 750–758
4. Zhao WX, Mu S, Hou Y, Lin Z, Chen Y, Pan X, Li K et al (2021) Recbole: towards a unified, comprehensive and efficient framework for recommendation algorithms. In: proceedings of the 30th acm international conference on information and knowledge management, pp 4653–4664
5. Melchiorre AB, Rekabsaz N, Parada-Cabaleiro E, Brandl S, Lesota O, Schedl M (2021) Investigating gender fairness of recommendation algorithms in the music domain. Inf Process Manage 58(5):102666
6. Zhang A, Chen Y, Sheng L, Wang X, Chua T-S (2024) On generative agents in recommendation. In: Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval, pp 1807–1817
7. Li Y, Chen H, Fu Z, Ge Y, Zhang Y (2021) User-oriented fairness in recommendation. In: Proceedings of the web conference, pp 624–632
8. Santos FP, Lelkes Y, Levin SA (2021) Link recommendation algorithms and dynamics of polarization in online social networks. Proc Natl Acad Sci 118(50):e2102141118
9. Hui B, Zhang L, Zhou X, Wen X, Nian Y (2022) Personalized recommendation system based on knowledge embedding and historical behavior. Appl Intell 1–13

10. Zhao WX, Hou Y, Pan X, Yang C, Zhang Z, Lin Z, Zhang J et al (2022) Recbole 2.0: towards a more up-to-date recommendation library. In: Proceedings of the 31st ACM international conference on information and knowledge management, pp 4722–4726
11. Pajkovic N (2022) Algorithms and taste-making: exposing the Netflix recommender system's operational logics. Convergence 28(1):214–235
12. Xiao T, Wang D (2021) A general offline reinforcement learning framework for interactive recommendation. In: Proceedings of the AAAI conference on artificial intelligence, vol 35(5). pp 4512–4520
13. Silva JF, Matos S (2022) Modelling patient trajectories using multimodal information. J Biomed Inform 134:104195
14. Pandeya YR, Lee J (2021) Deep learning-based late fusion of multimodal information for emotion classification of music video. Multimedia Tools Appl 80(2):2887–2905
15. Zhu Z, Lin K, Jain AK, Zhou J (2023) Transfer learning in deep reinforcement learning: a survey. IEEE Trans Pattern Anal Mach Intell 45(11):13344–13362
16. Iman M, Arabnia HR, Rasheed K (2023) A review of deep transfer learning and recent advancements. Technologies 11(2):40

DDoS Attacks Detection in SDN Using Multi-feature Selection Approach and LLMs



Amit Kachavimath, Satish Vaishyar, Ayush Chugani, Pratham Singh, and Sanjana A. More

1 Introduction

SDN is an advanced networking approach that enhances efficiency through a centralized infrastructure. Unlike conventional distributed network designs, SDN separates data forwarding from control management, enabling dynamic configuration. This architecture facilitates centralized network management by employing open APIs, allowing software applications to control network behavior effectively. However, the centralized nature of SDN introduces vulnerabilities, particularly the risk of a single point of failure. If compromised, an attack can disrupt the entire network, affecting its stability and security. Despite these concerns, SDN remains a promising solution for scalable and programmable network management [1].

DDoS attacks are a major threat to SDN security, aiming to disrupt legitimate access by overwhelming network resources, which leads to slowdowns and system failures. Although many classification methods have been suggested to detect and prevent DDoS attacks, these approaches often face problems such as delays in detection, managing high traffic, and maintaining accurate results. Supervised machine learning methods, in particular, have limitations that create gaps in detection. Large Language Models (LLMs) can help overcome these issues by offering better flexibility, quicker detection, and the ability to identify complex attack patterns more effectively, improving upon the weaknesses of supervised machine learning approaches [2].

We now define an LLM in the networking security context and explain why transformer based language models can model packet flow sequences more effectively than conventional CNN/LSTM approaches. LLMs are advanced AI models trained with vast amounts of data to interpret and create human-like text.

A. Kachavimath (✉) · S. Vaishyar · A. Chugani · P. Singh · S. A. More
KLE Technological University, Hubballi, Karnataka, India
e-mail: amitk@kletech.ac.in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_7

LLMs can analyze large-scale data, recognize contextual patterns, and generalize knowledge across different domains. They excel in natural language understanding, making them useful for cybersecurity tasks such as anomaly detection and threat analysis. Their ability to learn from diverse datasets allows them to detect evolving cyber threats more efficiently [3].

Recent studies have utilized AI techniques to detect DDoS attacks, primarily using machine learning and deep learning models. However, traditional entropy-based and unsupervised learning methods often lack accuracy. To improve detection, researchers are integrating entropy-based techniques with machine and deep learning approaches. Compared to these models, LLMs offer better adaptability in analyzing network traffic. They can process vast amounts of contextual data, detect subtle attack patterns, and dynamically adjust to new threats. Using LLMs in SDN security enhances real-time detection, improves accuracy, and strengthens network defense.

This research introduces an advanced approach for detecting DDoS attacks in SDN by integrating feature selection techniques with LLMs. To optimize accuracy, three feature selection methods SelectKBest, ANOVA F-value scoring, and Random Forest-based feature ranking were applied to identify the most critical 10 attributes from an initial dataset of 83. This selection process enhances model efficiency by reducing unnecessary data while preserving essential attack indicators.

To further refine detection, the study employs transformer-based models, including BERT, RoBERTa, and F-Net, leveraging contextual analysis and adaptive reasoning for precise anomaly identification. Unlike conventional machine learning models that rely on extensive labeled data, these LLMs can generalize attack patterns more effectively with limited training. By combining intelligent feature extraction with LLM-driven detection, the proposed framework strengthens SDN security, offering a proactive solution to evolving cyber threats while minimizing computational costs. This research provides the following key contributions:

(1) The approach utilizes a multi-feature selection mechanism to improve detection accuracy in the control plane, extracting the 10 most significant features for DDoS attack detection.
(2) An LLM-based approach utilizing BERT, RoBERTa, and F-Net models was implemented for DDoS attack detection in SDN, employing their contextual learning and adaptive reasoning capabilities to enhance detection accuracy.
(3) A detailed evaluation of the effectiveness of the proposed method is conducted, comparing its accuracy and efficiency against state-of-the-art detection techniques.

The paper is structured as follows in the subsequent sections. Section 2 explores relevant research studies in the field. Section 3 describes the proposed methodology, including the model framework and algorithmic approach. Section 4 presents the experimental outcomes and performance analysis. Finally, Sect. 5 concludes the study by summarizing key findings, discussing their significance, and suggesting future research directions.

2 Related Work

In article [4], LLMs such as GPT-3.5, GPT-4, and Ada were evaluated for DDoS attack detection in IoT-based SDN environments. Using CICIDS 2017 and the Urban IoT Dataset, LLMs achieved 90% accuracy with few-shot learning and up to 96% with fine-tuning, outperforming traditional multi-layer perceptron (MLP) models. In article [5], LLMs such as Llama2 and GPT-2 were evaluated for DDoS attack detection in large-scale network environments. Using CIC-DDoS2019 and real-world ISP NetFlow traces, DoLLM, an LLM-based model, demonstrated strong detection performance by leveraging semantic representation learning. The model achieved an F1 score improvement of up to 33.3% in zero-shot detection scenarios and at least 20.6% in real ISP traces. The authors of [6] have proposed a DDoS detection approach utilizing three SDN-specific datasets and deep learning techniques, including Multi-Layer Perceptron, Convolutional Neural Network, and Long Short-Term Memory(LSTM), where LSTM demonstrated better performance in terms of accuracy and loss.

The authors of [7], HuntGPT is introduced as an LLM-based anomaly detection system that integrates machine learning and Explainable AI (XAI) for intrusion detection. Using the KDD99 dataset, a Random Forest classifier was trained and combined with SHAP and LIME frameworks to enhance model transparency. Additionally, GPT-3.5 Turbo was integrated to improve user interaction and threat interpretation. The system was evaluated using Certified Information Security Manager (CISM) Practice Exams, achieving success rates between 72 and 82.5%, demonstrating its effectiveness in cybersecurity decision-making and automated threat analysis. The framework from [8], utilizes LLMs, specifically GPT-3.5 Turbo, to preprocess network data from IoT23, NSL-KDD, Litnet2020, and CICIoT2023 datasets, converting it into LLM-compatible formats for fine-tuning with an F1-score of 0.952.

In article [9], the ShieldGPT framework using CIC-DoS2017 and CIC-DDoS2019 datasets to generate explanations for attack behaviors and provide mitigation strategies. The authors of [10], have proposed an approach that fine-tunes Llama 2 using QLoRA and Proximal Policy Optimization (PPO) to enhance detection efficiency. The model was evaluated on CIC-IDS2017, achieving 99.23% accuracy, demonstrating its effectiveness in real-time DDoS attack detection while outperforming supervised learning approaches. In [11], an efficient DDoS detection approach for SDN is proposed using multi-feature selection and ensemble learning. In [12], a hybrid deep learning model is proposed for DDoS attack detection in SDN, integrating Consensus-Based Feature Selection (CBFS) with CNN-LSTM architecture. The model enhances detection by utilizing CNN for spatial feature extraction and LSTM for temporal pattern recognition, improving accuracy.

In article [13], the UNSW-NB15 dataset was employed to fine-tune an LLM using Low-Rank Adaptation (LoRA) to enhance domain-specific knowledge extraction. By leveraging knowledge distillation, the LLM's learned representations were effectively transferred to a lightweight anomaly detection model for improved efficiency. The student model, trained through knowledge distillation, achieved

superior performance, recording a macro-precision of 0.91, macro-recall of 0.88, and macro-F1 score of 0.87. In comparison, the general training model, which lacked knowledge distillation, performed less effectively, attaining a macro-precision of 0.79, macro-recall of 0.74, and macro-F1 score of 0.72.

3 Proposed Method

This section presents the detailed design of the proposed system, its feature importance, the collection of the training dataset used for training the model, and finally, the LLM algorithms for classifying the traffic.

The proposed system integrates a multi-feature selection mechanism with LLM-based detection using F-Net to enhance DDoS detection in SDN. The approach is validated using the InSDN dataset and Mininet emulation, demonstrating high accuracy and efficiency in identifying network anomalies represented in Fig. 1. Algorithm 1 presents a novel feature extraction method focused on identifying the 10 most relevant features from an initial set of 83. It integrates three selection techniques: SelectKBest, ANOVA F-value scoring, and feature importance derived from a Random Forest Classifier. By synthesizing these methods, the approach enhances both detection accuracy and model efficiency. This systematic approach ensures that only key attributes are retained, boosting anomaly detection performance in SDN environments. The selected features are represented in Table 1.

The proposed algorithm 2 utilizes the F-Net model to detect DDoS attacks in network traffic by following a structured process involving model initialization, data preparation, training, evaluation, and real-time detection. Initially, the pre-trained F-Net model and tokenizer are loaded to ensure compatibility with network-based data. To improve generalization, an optimal dropout rate is set, and the model is compiled using the Adam optimizer along with binary cross-entropy loss to enhance classification performance.

In the preprocessing step, numerical values are changed to string format to align with the tokenizer's requirements. The F-Net tokenizer processes the input using padding and truncation, while attention masks are generated to improve sequence interpretation. Additionally, minor label noise is introduced to increase model robustness against variations in data.

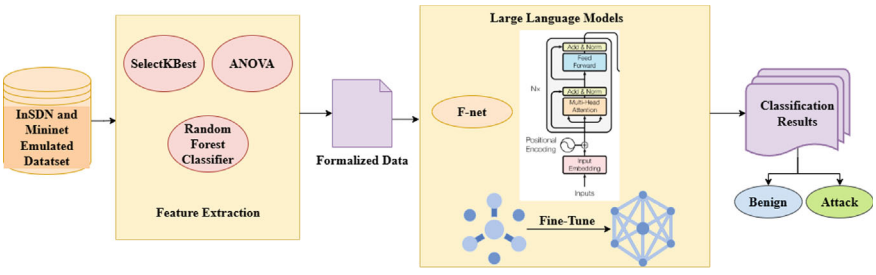


Fig. 1 System architecture for proposed methodology

Table 1 Top 10 selected features for DDoS attack detection

Sl. No	Feature name	Description
1	Source IP	Originating IP address of the packet
2	Destination IP	Target IP address receiving the packet
3	Protocol type	Communication protocol
4	Packet length StdDev	Variation in packet sizes within a flow
5	Flow identifier	Unique identifier assigned to each data flow
6	Reverse packet length mean	Average packet size in the reverse direction
7	Packet length variance	Fluctuation in packet sizes within a traffic flow
8	Source port	Port number utilized by the sender
9	Reverse segment average size	Mean size of segments in the reverse direction
10	Reverse packet length StdDev	Standard deviation of packet lengths in reverse flow

Algorithm 1 Optimized Feature Selection Using SelectKBest, ANOVA, and Random Forest

Input: Dataset D, Features $X = 83$, Target y
Output: Top 10 Selected Features
Load dataset D, extract $X \leftarrow D$.drop('Label'), $y \leftarrow D$['Label']
Initialize $selected_features_list \leftarrow []$
Define selection methods: {SelectKBest, ANOVA, RandomForest}
for method in methods **do**
    Apply method to X, y and extract top 20 ranked features
    Append selected features to $selected_features_list$
end for
$all_selected \leftarrow$ concatenate($selected_features_list$)
$final_features \leftarrow$ top 10 most frequently selected features

During training, the F-Net model is trained for multiple iterations using a training dataset, with periodic validation on a test dataset to assess performance. The training process is continuously monitored to track improvements. After training, evaluation of the model is performed using common metrics, including accuracy, precision, recall, and F1-score. The trained model and tokenizer are then saved for later use. For real-time classification, the saved model is deployed to analyze incoming network traffic from Mininet. The tokenizer processes new traffic data, and the model predicts whether the observed activity is normal or associated with a DDoS attack. This systematic approach ensures efficient detection of malicious network behavior, enhancing security in real-world applications.

Algorithm 2 F-Net Model for DDoS Attacks Classification

Input: Preprocessed dataset $selected_features_df$ with features and labels
Output: Trained F-Net model, and real-time traffic classification
Initialize the model:
Load the pre-trained F-Net model and tokenizer
Adjust dropout rate for better generalization
Compile the model using the Adam optimizer and binary cross-entropy loss
Preprocess the dataset:
Convert numerical features into string format for compatibility
Tokenize data with padding and truncation using the F-Net tokenizer
Generate attention masks for input sequences
Introduce minor label noise for robustness
Train the model:
Train the model for $epochs$ iterations on X_{train} and y_{train}, validating on X_{test} and y_{test} while logging epoch progress.
Evaluate and Save the Model:
Compute accuracy, precision, recall, and F1-score
Print evaluation results
Save the trained model and tokenizer
Real-Time Traffic Classification:
Load the saved model for inference
Tokenize incoming network traffic data
Predict and classify traffic as Normal or DDoS Attacks

4 Results and Discussion

The results and discussion section assesses the performance of the proposed multi-feature selection and LLM-based detection approach using F-Net for DDoS detection in SDN. The performance analysis is conducted on the InSDN dataset and Mininet-emulated traffic, comparing F-Net, BERT, and RoBERTa across key performance indicators such as accuracy, F1-score, evaluation loss, and runtime efficiency.

Figure 2 presents a comparative analysis of the accuracy of three LLMs BERT, RoBERTa, and F-Net on two datasets: InSDN and Mininet-emulated traffic. The results highlight that F-Net achieves the highest accuracy, recording 99% on InSDN and 99.33% on Mininet, outperforming the other models compared to BERT attains 94.5% accuracy on InSDN and 95.2% on Mininet, while RoBERTa achieves 96.8% on InSDN and 97.1% on Mininet. The superior performance of F-Net using Fourier Transform-based processing, which enhances feature extraction and enables precise DDoS attack detection with limited training data. Figure 3 illustrates the F1-score comparison of three LLMs on InSDN and Mininet-emulated datasets. The results indicate that F-Net demonstrates the highest F1-score, achieving 0.98 on InSDN and 0.987 on Mininet, signifying its robust detection capability in identifying DDoS attacks in SDN.

Figure 4 presents an evaluation loss analysis of the performance of three models in two datasets for the BERT, RoBERTa, and F-Net models. Better model performance is indicated by a lower evaluation loss. According to the graph, F-Net is the most

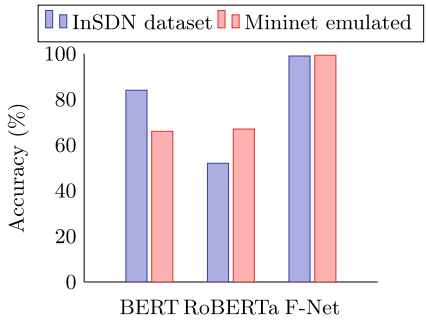


Fig. 2 Accuracy of proposed LLM's

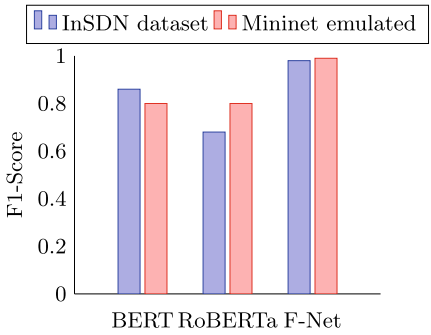


Fig. 3 F1-score of LLM's

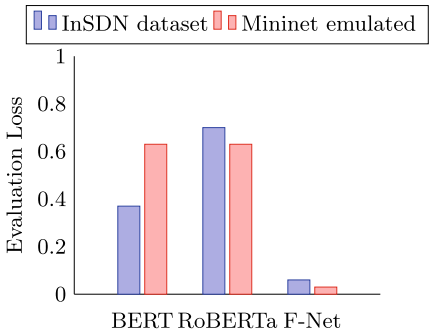


Fig. 4 Evaluation loss of LLM's

effective of the three considering it has the lowest loss. Figure 5 showcases the processing time analysis of three LLMs BERT, RoBERTa, and F-Net on the InSDN dataset and Mininet-emulated traffic. The results demonstrate that F-Net delivers the fastest execution, with 2.92 s on InSDN and 4.59 s on Mininet, highlighting its computational efficiency compared to BERT processes data in 8.82 s on InSDN and 13.66 s on Mininet, while RoBERTa records the highest runtime, taking 9.6 s on InSDN and 22.11 s on Mininet, indicating its greater computational overhead.

Figure 6 presents the Receiver Operating Characteristic (ROC) curve for the F-Net model, evaluated on the Mininet-emulated dataset. The ROC curve highlights the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR),

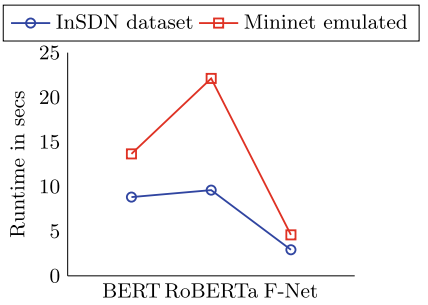


Fig. 5 Runtime (secs) of LLM's

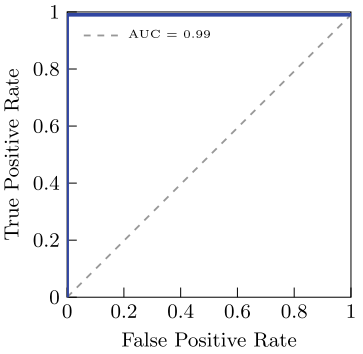


Fig. 6 ROC curve of F-Net model

highlighting the model's proficiency in differentiating between legitimate and malicious traffic in SDN. The F-Net model demonstrates near-perfect classification performance, achieving an Area Under the Curve (AUC) score of 0.99, indicating high detection accuracy with minimal false positives.

Figure 7 displays the confusion matrix of the F-Net model, showcasing its classification ability to distinguish normal traffic from attack traffic within Mininet emulated. The matrix consists of four key values: True Positives (6133), indicating precisely recognized attacks; True Negatives (3800), representing correctly identified

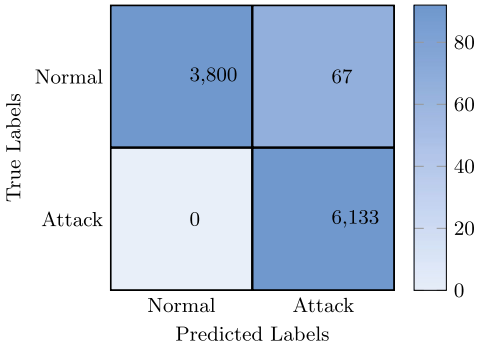


Fig. 7 Confusion matrix of F-Net model

legitimate traffic; False Positives (67), where benign traffic was misclassified as an attack; and False Negatives (0), signifying that no attacks were incorrectly labeled as normal traffic.

The incorporation of a multi-feature selection technique is the aspect that makes the proposed DDoS detection framework in SDN employing LLMs like F-Net, BERT, and RoBERTa so efficient. F-Net model exhibits considerably improved detection performance, with 99% accuracy in InSDN and 99.33% accuracy in Mininet datasets, compared to [4], which obtained up to 96% accuracy. The precision, recall, and F1-score of the proposed model are better compared to [5] and [7]. Efficiency is improved, and computational overhead is reduced by including by using the Multi-Feature selection approach and F-Net provides optimal results for real-time detection in dynamic SDN systems due to its minimized evaluation loss.

5 Conclusion and Future Scope

The proposed work has developed an advanced framework for detecting DDoS attacks in SDN by integrating LLMs with a multi-feature selection approach. The proposed methodology utilized F-Net, BERT, and RoBERTa to analyze network traffic, employing SelectKBest, ANOVA F-value scoring, and Random Forest-based feature ranking to extract the 10 most significant attributes from an initial set of 83 features. The experimental analysis demonstrated that F-Net outperformed BERT and RoBERTa, achieving 99% accuracy on the InSDN dataset and 99.33% on the Mininet-emulated dataset, making it a highly efficient solution for real-time DDoS attack detection. Additionally, the low evaluation loss and reduced computational overhead of F-Net further reinforced its capability to accurately classify network traffic while maintaining high efficiency. The findings of this study validate the effectiveness of LLM-driven cybersecurity solutions, highlighting their potential to enhance network security in SDN environments. Future research can focus on scaling the LLM-based DDoS detection framework to handle high-traffic SDN environments efficiently. Additionally, integrating real-time mitigation strategies with detection can enable automated security responses, enhancing proactive threat defense in dynamic networks.

References

1. Alhijawi B, Alazzam MB, Gupta BB, Al-Hazaimeh OM, Almiani M (2022) A survey on DoS/DDoS mitigation techniques in SDNs: classification, comparison, solutions, testing tools and datasets. Comput Electri Eng 99(107706), Elsevier
2. Bahashwan AA, Ullah F, Albasir HM, Alshmrany S, Alzahrani B (2023) A systematic literature review on machine learning and deep learning approaches for detecting DDoS attacks in software-defined networking. Sensors 23(9), art. no. 4441, MDPI
3. Bui M-T, Boffa M, Valentim RV, Navarro JM, Chen F, Bao X, Houidi ZB, Rossi D (2024) A systematic comparison of large language models performance for intrusion detection. In: Proceedings of the ACM on networking, vol 2, no. CoNEXT4, Article 22, ACM, December, pp 1–23
4. Guastalla M, Li Y, Hekmati A, Krishnamachari B (2023) Application of large language models to DDoS attack detection. In: Security and privacy in cyber-physical systems and smart vehicles, vol 485. Springer, Heidelberg, pp 83–99
5. Li Q, Zhang Y, Jia Z, Hu Y, Zhang L, Zhang J, Xu Y, Cui Y, Guo Z, Zhang X (2024) DoLLM: how large language models understanding network flow data to detect carpet bombing DDoS. arXiv preprint arXiv:2405.07638, https://arxiv.org/abs/2405.07638
6. Kachavimath AV, Narayan DG (2022) Distributed denial of service attacks detection using deep learning in software defined network. In: Proceedings of the 2022 13th international conference on computing communication and networking technologies (ICCCNT), October, IEEE, pp 1–5
7. Ali T, Kostakos P (2023) HuntGPT: integrating machine learning-based anomaly detection and explainable AI with large language models (LLMs). arXiv preprint arXiv:2309.16021, https://arxiv.org/abs/2309.16021
8. Zhou H, Huang X, Deng L (2024) Enhancing network traffic classification with large language models. In: Proceedings of the 2024 IEEE International conference on big data (Big Data), IEEE, pp 7282–7291
9. Wang T, Xie X, Zhang L, Wang C, Zhang L, Cui Y (2024) ShieldGPT: an LLM-based framework for DDoS mitigation. In: Proceedings of the 8th Asia-Pacific workshop on networking (APNet 2024), ACM, pp 108–114
10. Mahmoodi M, Jameii SM (2024) Utilizing large language models for DDoS attack detection. In: Proceedings of the 2024 OPJU international technology conference (OTCON) on smart computing for innovation and advancement in industry 4.0, IEEE, pp 1348–1353
11. Kachavimath AV, Narayan DG (2025) An efficient DDoS attack detection in SDN using multi-feature selection and ensemble learning. Proc Comput Sci 252:241–250, Elsevier
12. Kachavimath AV, Narayan DG (2025) A hybrid deep learning model with consensus-based feature selection for DDoS attacks detection in SDN. Proc Comput Sci 252:643–652, Elsevier
13. Yang Y, Tian B, Yu F, He Y (2024) An anomaly detection model training method based on LLM knowledge distillation. In: Proceedings of the 2024 international conference on networking and network applications (NaNA), IEEE, pp 472–477

PassPhrase Authentication: A Secure and User-Friendly Multi-factor Authentication System Using Phrase and Gesture Integration

Amit Kachavimath, Ayush Chugani, Satish Vaishyar, Niyati Gogri, and Misbah Falak

1 Introduction

In the dynamic realm of digital advancements, the demand for secure and user-friendly authentication has never been more critical. Conventional authentication models, primarily dependent on static username-password pairs, are proving insufficient against evolving cyber threats such as brute-force intrusions, phishing exploits, and keystroke logging. The escalating sophistication and recurrence of these attacks highlight the urgent need for innovative authentication frameworks that effectively integrate security with ease of use, ensuring resilience against unauthorized access attempts [1].

The PassPhase Authentication System, introduced in this study, presents an innovative multi-factor authentication framework that combines user-generated passphrases with interactive image-based gestures and coordinate-driven click sequences. By integrating these dynamic components, the system offers a personalized and seamless authentication experience, strengthening defenses against unauthorized access. Designed for adaptability and scalability, this approach effectively overcomes the limitations of conventional authentication techniques, ensuring enhanced security for digital interactions [2].

The proposed method explores the distinctive features of the PassPhase Authentication System but also examines its practical application and effectiveness in real-world scenarios. The design balances stringent security measures with a seamless user experience, offering a viable and scalable alternative to traditional authentication mechanisms. This research aims to bridge the gap between security and accessibility, contributing to the ongoing development of secure, user-friendly authentication solutions for modern digital ecosystems.

A. Kachavimath (✉) · A. Chugani · S. Vaishyar · N. Gogri · M. Falak
KLE Technological University, Hubballi, Karnataka, India
e-mail: amitk@kletech.ac.in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_8

The main findings and contributions of this paper are as follows:

- The PassPhase Authentication System enhances security against brute-force attacks, phishing attempts, and keylogging by integrating user-defined pass phrases with interactive image-based gestures and dynamic click patterns.
- This authentication approach delivers a frictionless user experience by merging memorable passphrases with intuitive graphical interactions, ensuring both ease of access and robust security.
- Designed for scalability and flexibility, the system serves as a secure, user-centric alternative to traditional authentication mechanisms, addressing the evolving challenges of digital security.

The structure of this paper is as follows: Sect. 2 examines existing research on authentication systems. Section 3 details the methodology underlying the proposed framework. Section 4 discusses the experimental setup along with the obtained results, while Sect. 5 provides concluding remarks and outlines potential directions for future research.

2 Literature Review

The growing complexity of cybersecurity threats, such as brute-force intrusions, phishing schemes, and keylogging techniques, highlights the limitations of traditional authentication methods. To mitigate these risks, recent advancements in authentication focus on multi-factor approaches, interactive user engagement, and context-aware intelligence. These developments strive to strengthen security without compromising usability, ensuring scalable and robust protection against evolving cyber threats.

In [3], A Frictionless and Secure User Authentication (FSUA) system is introduced for web-based applications, utilizing behavioral profiling based on attributes like login time, device location, and browser type. The framework integrates multi-factor authentication, combining one-time passwords (OTPs) and certificate-based mechanisms, while leveraging an analytical engine for real-time risk assessment. Experimental findings indicate that the system achieves a 10% reduction in authentication delays, a 7% improvement in response times, and an 11% decrease in memory consumption compared to traditional methods, offering enhanced security and seamless usability. The authors of [4] introduce a hybrid authentication framework is introduced, combining textual and graphical elements to enhance both security and user experience. The approach features two distinct modes: Easy Login, designed for convenience, and Secure Login, which ensures stronger protection through randomized password components and visual memory cues. The system was evaluated using data from 154 participants, assessing its resistance to security threats such as brute-force attacks and phishing schemes. A comparative study against eight conventional authentication methods revealed superior memorability

and robustness, with Secure Login achieving an average authentication time of 40.16 s and demonstrating higher accuracy in protecting user credentials.

In [5], the study presents a Reformation-Based password scheme designed to strengthen security and improve usability by implementing dynamic password transformations using numeric codes instead of conventional passwords. This method eliminates reliance on mental calculations or external devices, enhancing user convenience while safeguarding against brute-force intrusions and shoulder-surfing attacks. Experimental evaluation with 150 participants confirmed superior security and usability compared to traditional approaches, ensuring a robust and efficient client-server authentication mechanism. In [6], the study introduces an Event-Driven authentication framework that utilizes Bluetooth-based proximity verification combined with cryptographic handshakes to secure critical operations. This method effectively mitigates replay attacks and Man-in-the-Middle intrusions, ensuring enhanced security while maintaining user convenience. In [7], the study presents a MAC Address-Based Access Control system that strengthens login security by associating IP addresses with MAC addresses through ARP (Address Resolution Protocol) mapping. This method effectively prevents unauthorized access, even if credentials are compromised, while offering resistance against dictionary attacks and shoulder surfing attempts, ensuring restricted administrator page access to authorized devices only.

The authors of [8] have proposed a DDoS detection approach utilizing three SDN-specific datasets and deep learning techniques, including Multi-Layer Perceptron, Convolutional Neural Network, and Long Short-Term Memory(LSTM), where LSTM demonstrated better performance in terms of accuracy and loss. In [9], the authors present HandPass, a continuous authentication system using hand vibration biometrics captured by smartphone accelerometers. Achieving 98.36% accuracy, the system ensures security and usability, effectively resisting impersonation and blind attacks during natural user interactions. In [10], the authors present a responsive web application for GPA calculation, integrating authentication and authorization mechanisms. The methodology uses Django for backend processing, OTP-based two-step verification for secure user registration, and SQLite3 for data storage, ensuring user-friendly and efficient academic tools for students. The authors of [11] propose a user authentication model integrating Context-Based Fingerprinting with the random forest algorithm to enhance API security. Using a dataset of 19,200 records, the model achieved better accuracy, effectively detecting anomalous login attempts while ensuring robust API-level security.

In [12], the authors propose proactive measures for login page security, including IP-based restrictions, two-factor authentication, and automated verification codes. These mechanisms effectively defend against brute force and dictionary attacks, ensuring secure user access. The authors of [13], propose a privacy-preserving, device-centric, and attribute-based authentication architecture as an alternative to traditional passwords. The approach integrates federated login, cryptographic

credentials, and behavioral authentication, achieving enhanced security, user privacy, and compliance with NIST assurance standards while addressing challenges like identity fragmentation and password overload. In [14], the authors propose a novel Wi-Fi-based environmental continuous two-factor authentication system leveraging machine learning. By analyzing beacon frame characteristics and RSSI values, the system achieves 92.4% accuracy. In [15], the authors analyze OAuth and OpenID for robust identity verification and data protection. The study discusses the advantages, limitations, and implementation challenges of these methods in enhancing cybersecurity across diverse applications.

The authors of [16] focused on detecting anomalous network traffic to address the increasing complexity of Distributed Denial of Service (DDoS) attacks. The proposed methodology utilizes machine learning techniques, specifically K-Nearest Neighbor (KNN) and Naïve Bayes classifiers, to analyze statistical features of network traffic and identify malicious deviations. By extracting high-level traffic patterns, the model effectively detects DDoS attacks and outperforms traditional approaches. The study highlights the evolving nature of cyber threats and emphasizes the need for adaptive security solutions.

The summary of the literature review reveals notable gaps in the field of secure authentication systems. Although progress has been made in multi-factor methods, the absence of real-time adaptive strategies to counter advanced threats like adversarial attacks and spoofing remains evident. Additionally, most studies prioritize either security or user experience without achieving an optimal balance between the two. The scalability of proposed methods for diverse environments, particularly in IoT and cloud applications, is insufficiently addressed. Furthermore, limited focus on energy efficiency and computational costs highlights a need for more sustainable and practical evaluations. These gaps present opportunities for advancing authentication frameworks to ensure comprehensive protection and efficiency.

3 Proposed Methodology

This section presents the detailed design of the proposed PassPhase Authentication System, outlining its core components and interactive authentication features. It elaborates on the stepwise integration of user credentials, including password verification, phrase selection, and gesture-based interactions using image click patterns and coordinate validation. The framework emphasizes feature importance by combining personalized user inputs with spatial interaction to enhance authentication security, as mentioned in Algorithm 1.

Algorithm 1 Enhanced Multi-Factor Authentication (MFA) System

```
Input: User credentials (username, password), phrase, click patterns, coordinates
Output: Authentication status (Granted/Denied)
Accept username and password
Hash password using SHA-256 and compare with stored hash
if Password match then
    Prompt user to select their registered phrase
    if Phrase match then
        Display image associated with phrase
        Prompt user to perform registered click pattern
        Validate click pattern and coordinates (within 10-pixel tolerance)
        if Incorrect clicks exceed 3 attempts then
            Deny access and log out
        else
            if Login fails 3 consecutive times then
                Send alphanumeric OTP to registered email
                if OTP is valid then
                    Allow password reset and retry login
                else
                    Deny access
                end if
            end if
            Grant access if all validations are successful
        end if
    else
        Deny access
    end if
else
    Deny access
end if
```

3.1 Password Verification

At the initial stage, users enter their credentials, including username and password. The system stores only the hashed passwords using the SHA-256 algorithm to safeguard against breaches. During login, the entered password's hash is compared with the stored hash. A match permits access to the next layer, while mismatches deny entry.

The flowchart Fig. 1 illustrates the process of a multi-factor authentication (MFA) system. It begins with the user entering a predefined phrase, followed by the retrieval of a pre-saved image associated with the phrase. The user then performs two interactive tasks: arbitrary clicks on the image and location-specific clicks at predefined coordinates. Upon verifying the provided inputs, the system grants access by displaying a "Login Successful" message, finalizing the authentication process. If verification fails, the user is redirected to the multi-factor authentication module for additional validation. This layered security approach strengthens protection against unauthorized access while ensuring a seamless and user-friendly experience.

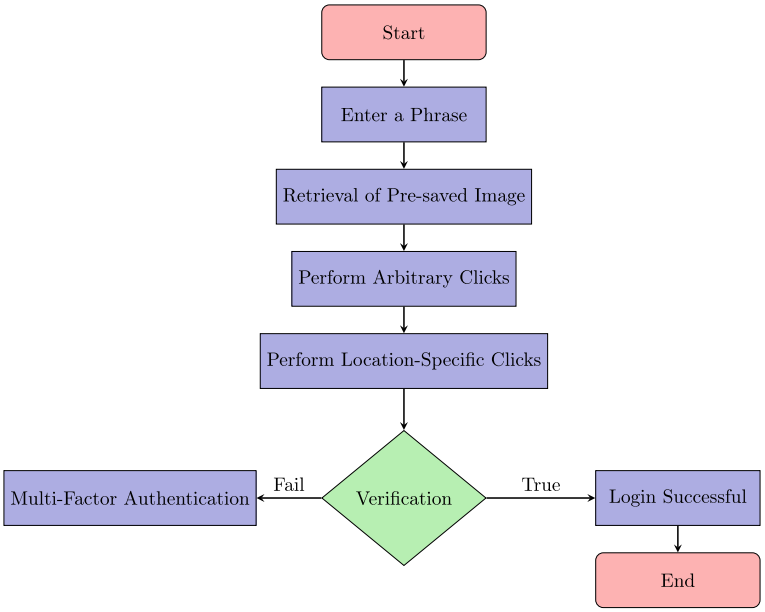


Fig. 1 Flowchart for multi-factor authentication process

3.2 Password Verification

During registration, users create a unique phrase, which is securely stored as part of their authentication credentials. During login, they must accurately select their registered phrase to proceed. Once verified, the system presents an image associated with the chosen phrase, serving as the next authentication step in the process.

3.3 User Interaction

In the proposed authentication framework, users engage with a displayed image through two sequential steps: Click Count Verification and Coordinate-Based Clicks. Initially, users replicate pre-registered click patterns on the image, allowing authentication to proceed only if the inputs match the stored sequence. Following this, users must accurately select predefined coordinates within the image, which are hashed and securely stored for validation. To enhance usability, the system permits a 10-pixel tolerance radius, accommodating minor variations in user clicks. If authentication fails after three consecutive incorrect attempts, an automatic logout is triggered to protect user data and prevent unauthorized access. This dual-layered interaction reinforces security by integrating dynamic user engagement with precise input validation.

3.4 Multi-factor Authentication

If multiple unsuccessful login attempts occur, users are redirected to the Multi-Factor Authentication (MFA) page, where they must confirm their identity using an alphanumeric one-time password (OTP) sent to their registered email. This method enhances security by leveraging alphanumeric OTPs, which provide 2177 times more possible combinations than purely numeric OTPs, significantly increasing resistance to unauthorized access and brute-force attacks.

The proposed authentication framework delivers a secure yet user-friendly multi-factor authentication system, striking an optimal balance between protection and usability. By integrating layered security measures—including password verification, phrase selection, click-pattern interactions, and coordinate-based validation—the system ensures comprehensive defense against unauthorized access. Additionally, hashed credentials and alphanumeric OTPs enhance resilience against cyber threats, strengthening authentication security while maintaining a seamless experience for legitimate users. Designed for modularity and scalability, this adaptive framework is well-suited for diverse application environments, addressing the growing challenges of digital security with an efficient and practical solution.

4 Results and Discussion

The results and discussion section critically examines the effectiveness of the proposed MFA framework through an in-depth performance analysis. Key parameters such as security strength, usability, accuracy, and system resilience were evaluated across varying conditions. Findings confirm the framework's capability to counteract security threats, including brute-force attacks, phishing, and unauthorized access, while ensuring a user-friendly experience. A comparative study with existing authentication models reveals notable improvements in login accuracy, response time, and error tolerance. This section also explores the practical implications of the results, addressing both real-world deployment challenges and potential applications. The insights presented validate the proposed solution's viability in meeting the growing demands for secure and seamless digital authentication systems.

Figure 2 compares the security scores of four different authentication methods: Username and Password, OTP, Biometric, and PassPhase. The graph shows that the PassPhase method has the highest security score at 92.1%, followed by Biometric authentication at 83.6%. OTP comes in at 74.2%, and Username & Password ranks the lowest with 62%. This demonstrates that traditional password-based systems, while still commonly used, are less secure compared to newer, more advanced methods like Biometric and PassPhase. The PassPhase method, which integrates a passphrase with interactive gestures, offers a significant improvement in security effectiveness by mitigating risks like brute-force attacks and keylogging.

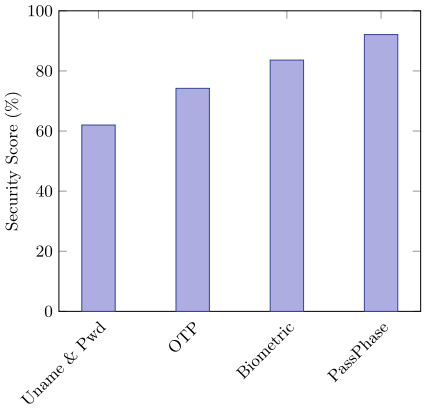


Fig. 2 Security effectiveness of authentication methods

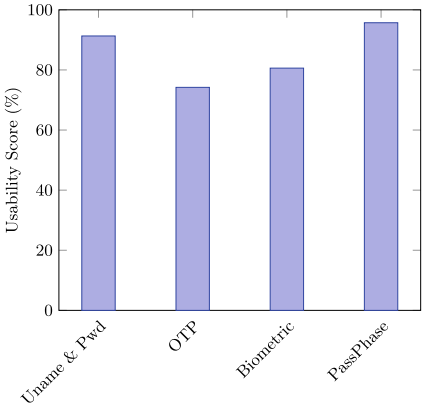


Fig. 3 Usability score comparison of authentication methods

Figure 3 evaluates the usability of the same authentication methods. The PassPhase method achieves the highest usability score at 95.7%, closely followed by Username & Password with 91.3%. Biometric authentication comes next at 80.6%, while OTP has the lowest usability score at 74.2%. These results indicate that while traditional password-based methods and biometric authentication are user-friendly, the PassPhase method strikes a balance between both security and usability, making it an ideal solution for modern authentication needs. The high usability of PassPhase demonstrates its user-friendliness, despite its advanced security features, showing that complex security mechanisms can be implemented without compromising ease of use. The proposed MFA system was rigorously tested against various attack scenarios, including brute-force intrusions and phishing attempts, using industry-standard security tools. The results confirm the system's strong resilience against unauthorized access, primarily due to its layered authentication approach, incorporating hashed passwords and coordinate-based verification to counteract brute-force attacks. Additionally, phrase selection and alphanumeric OTPs enhance protection against phishing attempts, as they require unique, user-specific inputs

that are challenging to duplicate. Compared to traditional authentication methods, the system consistently demonstrates superior security and usability, validating its effectiveness in mitigating modern cybersecurity threats.

5 Conclusion

The PassPhase Authentication system effectively integrates multi-factor authentication with a user-centric approach, addressing key security concerns such as brute-force attacks and phishing attempts. By combining layered authentication techniques, including passphrases, graphical gestures, and coordinate-based click patterns, the framework delivers a robust yet intuitive solution, ensuring an optimal balance between security and usability. Experimental evaluations confirm its better performance over traditional authentication models, demonstrating enhanced resilience against unauthorized access while maintaining user convenience. This methodology establishes a new standard for secure authentication systems in modern digital environments. The future advancements could incorporate real-time adaptive security measures to counter emerging threats, such as deepfake biometrics and adversarial attacks. Additionally, blockchain integration and federated identity management could further enhance scalability and system-wide security.

References

1. Shah SW, Kanhere SS (2019) Recent trends in user authentication—a survey. IEEE Access 7:112505–112519
2. Wang X, Yan Z, Zhang R, Zhang P (2021) Attacks and defenses in user authentication systems: a survey. J Netw Comput Appl 188:103080
3. Olanrewaju RF, Khan BUI, Morshidi MA, Anwar F, Kiah MLBM (2021) A frictionless and secure user authentication in web-based premium applications. IEEE Access 9:129240–129255
4. Nizamani SZ, Hassan SR, Shaikh RA, Abozinadah EA, Mehmood R (2021) A novel hybrid textual-graphical authentication scheme with better security, memorability, and usability. IEEE Access 9:51294–51312
5. Ali M, Baloch A, Waheed A, Zareei M, Manzoor R, Sajid H, Alanazi F (2021) A simple and secure reformation-based password scheme. IEEE Access 9:11655–11674
6. Marshall K, Cantido T, Case J, Nguyen L, El-Razouk H (2021) An event-driven authentication approach for mediation of user actions. In: 2021 IEEE 12th Annual information technology, electronics and mobile communication conference (IEMCON), pp 534–540
7. Alam BMPP, Septiasari R, Amiruddin A (2019) Applying MAC address-based access control for securing admin's login page. In: Proceedings of the 2019 international conference on electrical engineering and computer science (EECSI), pp 292–296
8. Kachavimath AV, Narayan DG (2022) Distributed denial of service attacks detection using deep learning in software defined network. In: Proceedings of the 2022 13th international conference on computing communication and networking technologies (ICCCNT), October 2022, IEEE, , pp 1–5

9. Cao H, Jiang H, Yang K, Chen S, Wu W, Liu J, Dustdar S (2023) Data-augmentation-enabled continuous user authentication via passive vibration response. IEEE Internet Things J 10(16):14137–14151
10. Muthiah J, Lakshmi AS, Jahnavi D, Rengarajan A, Chidambaram N (2023) Design and implementation of authentication and authorization flow in a responsive web application. In: 2023 2nd International conference on vision towards emerging trends in communication and networking technologies (ViTECoN), pp 1–6
11. Al-Rumaim A, Pawar JD (2024) Enhancing user authentication: an approach utilizing context-based fingerprinting with random forest algorithm. IEEE Access 12:110850–110861
12. Bhadouria A, Bathla R, Arora G (2024) Fortifying digital frontiers: enhancing login page security against emerging cyber threats. In: 2024 11th International conference on reliability, Infocom technologies and optimization (ICRITO), pp 1–6
13. Papadamou K, Zannettou S, Chifor B, Teican S, Gugulea G, Caponi A, Recupero A, Pisa C, Bianchi G, Gevers S, Xenakis C, Sirivianos M (2020) Killing the password and preserving privacy with device-centric and attribute-based authentication. IEEE Trans Inf Forensics Secur 15:2183–2193
14. AlQahtani AAS, Alshayeb T, Nabil M, Patooghy A (2024) Leveraging machine learning for Wi-Fi-based environmental continuous two-factor authentication. IEEE Access 12:13277–13289
15. Korchenko A, Kazmirchuk S, Kryvoruchko O, Synichuk O, Kostiuk K, Zakharov R (2020) Methods of security authentication and authorization into informational systems. In: 2020 IEEE 2nd International conference on advanced trends in information theory (ATIT), pp 270–274
16. Kachavimath AV, Nazare SV, Akki SS (2020) Distributed denial of service attack detection using Naïve Bayes and K-nearest neighbor for network forensics. In: Proceedings of the 2020 2nd international conference on innovative mechanisms for industry applications (ICIMIA), pp 711–717

DDoS Attacks Detection Using PSO Feature Selection and Classification Using XGBoost in SDN



Amit Kachavimath, Sanjana A. More, and D. G. Narayan

1 Introduction

Network control can be centralized through the adoption of SDN, which separates the control plane from the data plane. SDN enhances scalability, security, and network performance by enabling constant real-time control over network traffic [1]. The resources of SDN controllers can be overloaded by Distributed Denial of Service (DDoS) attacks, which might lead to decreased performance or complete service disruption. Attackers may target the controller's main point of control and flood it with traffic, ultimately exhausting its processing power and preventing the processing of legitimate requests [2].

DDoS attacks attempt to interfere with regular operations by combining a target system, such as a network or server, with malicious traffic [3]. Since the SDN controller handles network traffic centrally, a DDoS attack might flood it with requests, causing delays or even network failure. This disruption could affect the SDN's capacity to effectively handle traffic flows, damaging the availability, security, and performance of the network. It works particularly well in cloud computing environments to manage large amounts of traffic and protect against DDoS attacks [4].

In the field of DDoS attack detection for InSDN and Mininet, the use of PSO and XGBoost provides an effective approach to optimize detection performance. By incorporating optimization with cutting-edge machine learning techniques, this strategy addresses real-world problems in contrast to general references to algorithm limitations. Improves flexibility to varying network conditions and prevents issues such as class imbalance and redundant data [5]. The key findings and contributions are as follows.

A. Kachavimath (✉) · S. A. More · D. G. Narayan
KLE Technological University, Hubballi, Karnataka, India
e-mail: amitk@kletech.ac.in

D. G. Narayan
e-mail: narayan_dg@kletech.ac.in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_9

- Meta-heuristic PSO approach is better than the statistical methods by dynamically selecting relevant features, reducing noise, and enhancing feature subsets, leading to improved accuracy in DDoS attack detection.
- The combination of PSO and XGBoost ensures optimized feature selection, leading to better DDoS detection accuracy than traditional classifiers
- Evaluation of the proposed technique's effectiveness is performed, measuring its accuracy and efficiency in comparison to state-of-the-art detection methods for reliable threat identification.

The rest of the article is divided into the following sections. The summary of the relevant research is provided in Sect. 2. The suggested SDN, which uses a machine learning classifier and metaheuristic optimization for feature selection, is described in Sect. 3. Performance evaluation and implementation outcomes are presented in Sect. 4. The conclusions of the study are presented in Sect. 5.

2 Literature Review

The Literature Survey section provides a comprehensive overview of existing research related to DDoS attack detection in SDN. It highlights key methodologies, identifies limitations in current approaches, and establishes the foundation for the proposed framework. In [6], the researchers introduced a flow-based anomaly detection framework is proposed for detecting DDoS attacks in SDN, utilizing feature selection techniques The proposed technique utilizes Random Forest (RF) and Information Gain (IG) to extract relevant attributes from the CICIDS2018, CICIDS2017, and InSDN datasets. This optimized feature selection enhances the classification performance, achieving a high accuracy rate of 99.1% using the LSTM model.

In [7], the researchers introduced an ensemble method combining K-means clustering and the RF algorithm for DDoS attack detection in SDN. The approach first clusters network traffic using K-means to identify anomalies, followed by RF classification for precise attack detection. The experimental evaluation utilized both synthetic and real-world SDN traffic datasets, including CICIDS2017. The framework achieved a high accuracy of 99.12%, demonstrating its effectiveness in detecting malicious activities in SDN environments. In [8], the researchers introduced a refined approach utilizing the RF algorithm, incorporating hyperparameter optimization and recursive feature elimination for selecting the most relevant attributes. The dataset employed in this study includes 207,146 traffic flows with 83 distinct characteristics. In [9], authors have proposed a DDoS detection model for SDN using ensemble learning and advanced feature selection techniques on the InSDN dataset. By selecting the top 10 features, the model achieved 99.9% accuracy, validated through ROC curves and cross-validation.

In [10], the researchers introduced a hybrid SVM-KNN approach for identifying DDoS attacks in SDN environments. This methodology emphasizes the selection of five critical features from packet headers, followed by the application of the SVM-KNN model for training and classification. Using the Mininet emulator and Ryu controller, the dataset simulates both typical and attack scenarios. The proposed model achieved an accuracy of 97.36%, demonstrating its effectiveness in distinguishing malicious activities from legitimate network traffic. The authors of [11], introduced a dual-module approach for mitigating DDoS threats, addressing both low-rate and high-rate attacks. The first module applies the Alleviative Threat for TCAM (ATFT) algorithm to counter Slow-TCAM attacks, while the second leverages ensemble learning techniques to detect high-intensity DDoS intrusions. The framework achieved 99% accuracy and was validated using a Mininet simulation and evaluated on the CICDDoS2019 dataset. The authors of [12], proposed an anomaly-based detection system utilizing Principal Component Analysis (PCA) to assess network traffic entropy and identify irregularities. The approach involved extracting key features, including packet count and byte flow, to construct a dataset for machine learning models. Various classifiers were tested, with Support Vector Machine (SVM) achieving the highest accuracy of 97%, outperforming Decision Tree (DT) in detection performance.

The authors of [13], implemented In-band Network Telemetry (INT) for real-time switch state monitoring within a Spatial-Temporal Graph Convolutional Network (ST-GCN), representing the network as a graph structure. The dataset was generated through INT-based sampling, capturing key network state indicators. The detection framework utilizes both spatial and temporal features to assess the network state variations, enabling the identification of DDoS attack pathways with high precision. In article [14], the Rényi entropy method was employed to monitor network traffic at the SDN controller. The dataset was generated using the Mininet emulator and the Ryu controller, simulating both benign and malicious traffic patterns to enhance anomaly detection and security analysis in SDN.

3 Proposed Methodology

This section presents the detailed design of the proposed system, its feature importance, the collection of the training dataset used for training the model, and finally, the PSO with XGBoost algorithm for detection of DDoS attacks.

Figure 1 presents InSDN [15] and Mininet datasets used by the DDoS attack detection system in SDN. The data is initially prepared by preprocessing, normalization, and one-hot encoding. To differentiate between legitimate and malicious traffic, the PSO algorithm optimizes the feature selection process. The chosen features are then fed into an XGBoost classifier.

The algorithm 1 represents PSO-based feature selection algorithm extracts the most relevant 10 features from a dataset containing 84 attributes. It begins with 20 particles, each initialized with random positions and velocities, which are updated

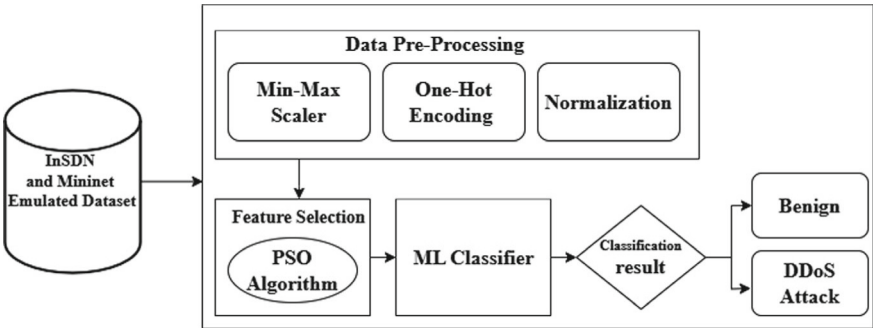


Fig. 1 System architecture of proposed methodology

Algorithm 1 PSO-Based Feature Selection for Top 10 Features

Input: Dataset D, Feature Matrix X (Size: 84), Target y
Output: Top 10 Selected Features
$X \leftarrow D.drop('Encoded_Label'),\ y \leftarrow D['Encoded_Label']$
Initialize PSO: particles $n = 20$, iterations $= 5$, $w = 0.7$, $c_1 = c_2 = 1.5$
Initialize particle positions x_i (binary), velocities v_i, $pBest$, $gBest$
for each iteration **do**
  for each particle **do**
    Update $v_i \leftarrow w \cdot v_i + c_1 \cdot r_1 \cdot (pBest - x_i) + c_2 \cdot r_2 \cdot (gBest - x_i)$
    Convert v_i to binary via sigmoid, enforce exactly 10 features in x_i
    Evaluate fitness; update $pBest$ if improved
  end for
  Update $gBest$ from best $pBest$
end for
return Top 10 most frequently selected features across particles

over five iterations using personal best and global best solutions. The velocity update follows a formula incorporating an inertia weight of 0.7 and learning coefficients of 1.5, ensuring a balance between search diversity and convergence. A sigmoid function converts velocity values into binary selections, maintaining exactly 10 features per iteration. Finally, the algorithm selects the top 10 features that appear most frequently, enhancing DDoS detection accuracy in SDN.

The algorithm 2 represents XGBoost-based DDoS intrusion classification algorithm processes a given dataset by extracting feature variables and target labels while ensuring an 80–20 training-to-testing split with balanced classes. The XGBoost model is then initialized with 30 trees, a learning rate of 0.5, a tree depth of 1, and 50% of training data and feature subsets per tree, optimizing performance using the log loss metric. The model undergoes training on the prepared dataset, followed by prediction and performance evaluation. Finally, classification accuracy is assessed using a confusion matrix, and results are visualized graphically, providing comprehensive evaluation metrics for DDoS detection in SDN.

Algorithm 2 DDoS Attacks Detection Using XGBoost

Input: Dataset D, Feature set X, Target class y
Output: Evaluation metrics for classification
Load dataset D
Extract feature variables $X \leftarrow D.drop('Encoded_Label')$ and corresponding target $y \leftarrow D['Encoded_Label']$
Partition dataset into X_{train}, X_{test}, y_{train}, y_{test} with an 80%-20% split while maintaining class balance
Compute $flip_count = 0.2 \times |y_{train}|$ to modify 20% of training labels
Randomly sample indices and alter class labels: $y_{train}[indices] = 1 - y_{train}[indices]$
Configure XGBoost parameters: Trees count = 30, Learning rate = 0.5, Depth of trees = 1, Training data fraction per tree = 0.5, Feature subset per tree = 0.5, Performance metric: log loss
Train XGBoost on (X_{train}, y_{train})
Generate predictions: $y_{pred} \leftarrow model.predict(X_{test})$
Compute accuracy and performance metrics
Construct confusion matrix for predicted and actual labels
Display classification results graphically
Present evaluation outcomes, classification metrics, and visual representations

4 Results and Discussion

Figure 2 presents the accuracy performance of three classification models Logistic Regression (LR), Naïve Bayes (NB), and XGBoost (XGB) assessed using the InSDN dataset and the Mininet emulated dataset. The findings indicate that XGBoost consistently outperforms the other models, achieving an accuracy exceeding 99% on both datasets. In contrast, Logistic Regression demonstrates a moderate accuracy of approximately 85%, while Naïve Bayes exhibits the lowest performance, with an accuracy below 75%.

Figure 3 presents the F1-score performance of three classification algorithms LR, NB, and XGB evaluated on both the InSDN dataset and the Mininet emulated dataset. The results demonstrate that XGBoost consistently obtains the highest F1-score, nearing 0.98, indicating its superior balance between precision and recall. In contrast, Logistic Regression shows moderate performance with an F1-score around 0.85, while Naïve Bayes struggles to exceed 0.70, reflecting its limitations in handling imbalanced or high-dimensional data.

Figures 4 and 5 evaluate the recall and precision of LR, NB, and XGB on the InSDN dataset and the Mininet emulated dataset. XGBoost demonstrates the highest recall, approaching 0.99, indicating its superior ability to correctly identify DDoS attacks while minimizing false negatives. In contrast, Logistic Regression achieves a recall of approximately 0.85, and Naïve Bayes remains the least effective, with recall values below 0.75 and a similar performance is observed in precision Fig. 4, where XGBoost again leads with nearly 0.98, highlighting its efficiency in minimizing false positives.

Figure 6 illustrates the Receiver Operating Characteristic (ROC) curve for the XGBoost model, evaluated using the Mininet emulated dataset. The Area Under

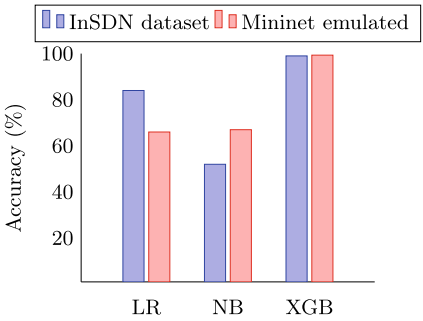


Fig. 2 Accuracy of LR, NB, and XGB models

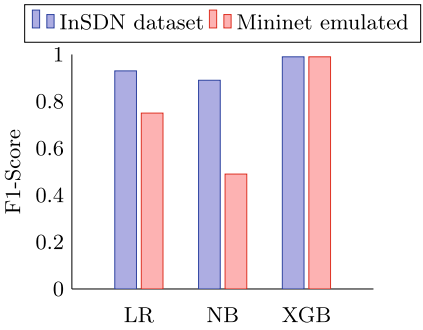


Fig. 3 F1-score of LR, NB, and XGB models

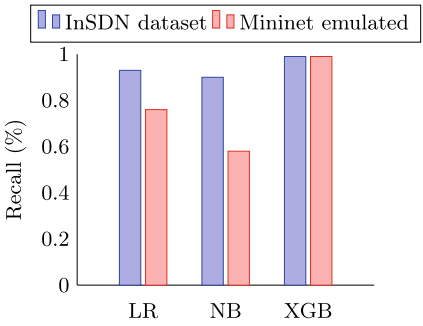


Fig. 4 Recall of LR, NB, and XGB models

the Curve (AUC) reaches 0.99, indicating near-perfect classification performance with minimal false positives. The rise in the True Positive Rate (TPR) at a low False Positive Rate (FPR) further confirms the high sensitivity and specificity of XGBoost in distinguishing DDoS attacks from normal traffic. Figure 7 illustrates the confusion matrix for the XGBoost model, tested on the Mininet emulated dataset. The model accurately detects 13,681 normal traffic samples and 14,000 attack cases, with just four incorrect classifications and no missed attacks, indicating outstanding accuracy. The minimal misclassification rate confirms XGBoost's strong capability in differentiating DDoS attacks from regular network traffic.

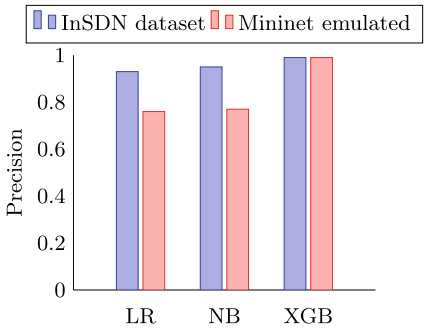


Fig. 5 Precision of LR, NB, and XGB models

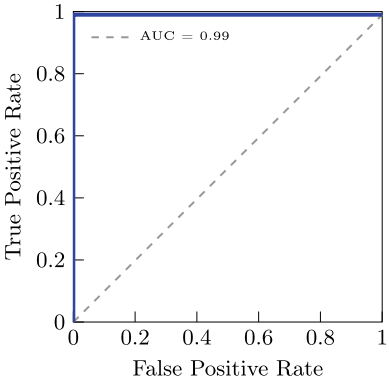


Fig. 6 ROC curve of XGBoost using mininet emulated

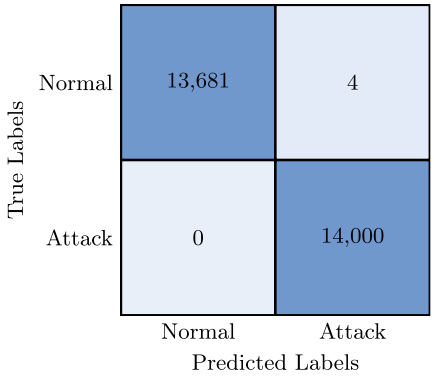


Fig. 7 Confusion matrix of XGBoost using mininet emulated

Mininet is used to simulate the network architecture shown in Fig. 8, which includes sixteen hosts from different subnets, a POX controller, and six OpenFlow virtual switches. Centralized management is made possible by the switches connecting these hosts, which in turn connect back to the POX controller and the main switch. In this SDN setup, dynamic control and management are rendered accessible by the POX controller operating in combination with a Python script in a

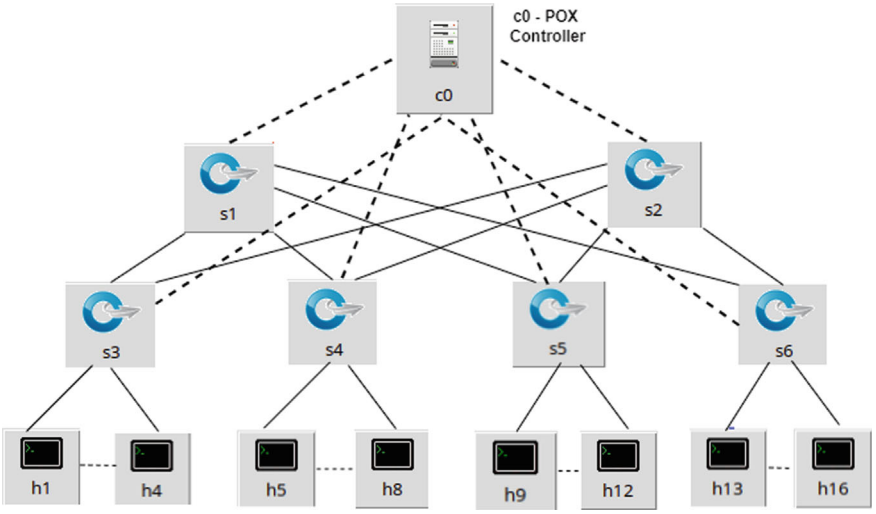


Fig. 8 Mininet network topology using POX controller

Linux environment. The Hping3 application was used to generate attack flows into malicious and legitimate traffic to simulate real-world traffic instances. As a result, malicious traffic affects the OpenFlow switches, offering significant details on the performance and behavior of the SDN network that involves being attacked [16].

Our proposed framework has distinctive features because it incorporates PSO for feature selection and XGBoost for classification, in addition to earlier studies like those in [6, 7, 10] that employed feature selection techniques like RF and K-means clustering for DDoS attack detection in SDN. However, the models in [6, 7] employed methods like LSTM and RF to achieve high accuracies of 99.1% and 99.12%, respectively, our approach significantly outperforms these results with over 99% performance on both the InSDN and Mininet datasets. Furthermore, our framework outperforms conventional models like LR and NB in terms of precision, recall, and F1-score. The high AUC score of 0.99 shows the significance of feature selection in improving real-time DDoS attack detection in SDN systems and further validates the robustness of our model in identifying between attack and regular traffic.

5 Conclusion and Future Scope

The research proposed an improved DDoS detection framework designed for SDN by integrating PSO for feature selection and XGBoost for classification. The experimental results demonstrate that XGBoost, combined with PSO-based feature optimization, significantly enhances detection accuracy, achieving over 99% performance on both InSDN and Mininet datasets. Comparative analysis with

traditional models, including LR and NB, highlights the better precision, recall, and F1-score of XGBoost, confirming its robustness in identifying malicious traffic with minimal false positives and false negatives. The high AUC score of 0.99 further validates the model's effectiveness in distinguishing normal and attack traffic. The results underscore the importance of feature selection in reducing dimensionality, eliminating redundant attributes, and improving classifier efficiency, making the proposed approach a solution for real-time DDoS attack detection in SDN environments. Future research can incorporate advanced deep learning architectures, including transformers and graph neural networks, to analyze attack patterns in dynamic SDN environments.

References

1. Ghanbarzadeh R, Hosseinalipour A, Ghaffari A (2023) A novel network intrusion detection method based on metaheuristic optimisation algorithms. J Ambient Intell Humaniz Comput 14(6):7575–7592 Jun
2. Sameer A (2014) Detection and prevention of distributed denial of service (DDoS) attacks using metaheuristic and machine learning techniques. Int J Sci Res (IJSR) 13:10
3. Alkanhel R, El-kenawy E-S, Abdelhamid A, Ibrahim A, Alohali M, Abotaleb M, Khafaga D (2022) Network intrusion detection based on feature selection and hybrid metaheuristic optimization. Comput Mater Contin 74:2677–2693 Oct
4. Alhijawi B, Alazzam MB, Gupta BB, Al-Hazaimeh OM, Almiani M (2022) A survey on DoS/DDoS mitigation techniques in SDNs: classification, comparison, solutions, testing tools and datasets. Comput Electr Eng 99, art. no. 107706
5. Bahashwan AA, Ullah F, Albasir HM, Alshmrany S, Alzahrani B (2023) A systematic literature review on machine learning and deep learning approaches for detecting DDoS attacks in software-defined networking. Sensors 23(9), art. no. 4441, MDPI
6. Sayed MSE, Le-Khac N-A, Azer MA, Jurcut AD (2022) A flow-based anomaly detection approach with feature selection method against ddos attacks in sdns. IEEE Trans Cogn Commun Netw 8(4):1862–1880 Dec
7. El Sayed MS, Le-Khac NA, Azer MA, Jurcut AD (2020) DDoS attack detection in software-defined network using ensemble K-means and random forest. In: Proceedings of the 2020 3rd international seminar on research of information technology and intelligent systems (ISRITI), pp 164–169
8. Hamarshe A, Alawad A, Alkasassbeh M (2024) Enhanced ddos detection in software-defined networking using ensemble-based machine learning. IEEE Access 12:1031–1040
9. Kachavimath AV, Narayan DG (2025) An efficient ddos attack detection in sdn using multi-feature selection and ensemble learning. Procedia Comput Sci 252:241–250
10. Feng S, Yang G, Man W (2023) Research on DDoS attack detection based on machine learning in SDN environment. In: Proceedings of the 2023 IEEE 7th information technology and mechatronics engineering conference (ITOEC), Chongqing, China, pp 821–825
11. Wu H, Hou A, Nie W, Wu C (2023) On an integrated security framework for defense against various DDoS attacks in SDN. In: Proceedings of the 2023 international conference on computing, networking and communications (ICNC), communications and information security symposium, Honolulu, HI, USA, pp 311–317
12. Yadav AR, Rajesh A, Jain AP, Perumal STS, Eappen G (2023) AI-based DDoS attack detection of SDN network in Mininet emulator. In: Proceedings of the 2023 2nd international conference on vision towards emerging trends in communication and networking technologies (ViTECoN), Vellore, India, pp 1–6

13. Cao Y, Jiang H, Deng Y, Wu J, Zhou P, Luo W (2022) Detecting and mitigating DDoS attacks in SDN using spatial-temporal graph convolutional network. IEEE Trans Dependable Secur Comput 19(6):3855–3872
14. Zhang H, Zhou L, Lei J (2023) Renyi entropy-based DDoS attack detection in SDN-based networks. In: Proceedings of the 2023 IEEE 3rd international conference on electronic technology, communication and information (ICETCI), Changchun, China, pp 334–337
15. Elsayed MS, Le-Khac N-A, Jurcut AD (2020) Insdn: a novel SDN intrusion dataset. IEEE Access 8:165263–165284
16. Kachavimath AV, Narayan DG (2025) A hybrid deep learning model with consensus-based feature selection for DDoS attacks detection in SDN. Procedia Comput Sci 252:643–652, 4th International Conference on Evolutionary Computing and Mobile Sustainable Networks

Chandrayaan-3 Mission to the Moon: A Systematic Review of Analysis, Improvements, Challenges and Lunar Science Contribution

Milind Shah, Dhairya Vyas, Kinjal A. Patel, M. Anantha Lakshmi, Roopal Rajput, and S. Lokesh

1 Introduction

1.1 Introduction and History of Indian Space Research Organization (ISRO)

ISRO was founded in November 1969 as the operational section of India's Department of Space (DOS). India founded the Space Commission and Department of Space in June 1972. The DOS took over ISRO in September of the same year. The

M. Shah (✉)
Department of Computer Engineering, Sardar Vallabhbhai Patel Institute of Technology (SVIT), Vasad, Gujarat, India
e-mail: milindshahcomputer@gmail.com

D. Vyas
The Maharaja Sayajirao University of Baroda, Vadodara, Gujarat, India
e-mail: dhairya.vyas-cse@msubaroda.ac.in

K. A. Patel
Builure Services Pty. Ltd, Melbourne, VIC, Australia

M. Anantha Lakshmi
Electronics and Communication Department, Meenakshi College of Engineering, Chennai, Tamil Nadu, India

R. Rajput
Department of Mobile Application and Technologies, Centre for Professional Courses, Gujarat University, Ahmedabad, Gujarat, India

S. Lokesh
Electronics and Communication Department, SRM Institute of Science and Technology, Katankulathur, Chengalpattu, Tamil Nadu, India
e-mail: lokesha@srmist.edu.in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_10

Space Commission established and manages the Indian Space Program rules. Space science and technology should be advanced and used for the country's economic benefit. These initiatives are mostly implemented by ISRO, PRL, NARL, NE-SAC, and SCL. Space-related products and services are promoted and sold by state-owned Antrix Corporation, formed in 1992.

The INSAT Coordination Committee (ICC), the Planning Committee on National Natural Resources Management System (PC-NNRMS), and the Advisory Committee for Space Sciences (ADCOS) are national-level bodies tasked with coordinating the development and utilization of space systems.'

An administrative institution in the field of space research, the Advisory Committee for Space research (ADCOS) is responsible for promoting and supporting activities that are carried out inside the United States. In the Indian Space Program, the Astronomy and Astrophysics Department (ADCOS) serves as the coordinating organization that is accountable for providing support for missions in the domains of Astronomy and Astrophysics, Space Weather, Planetary Exploration, and Weather and Climate Science [1].

1.2 Brief Introduction of Chandrayaan-1

India's first lunar orbiter, Chandrayaan-1, was launched by PSLV-XL. The mission began October 22, 2008. With the successful launch and operation of Chandrayaan-1, the Indian Space Program entered a new phase and achieved major scientific results [1].

1.3 Brief Introduction of Chandrayaan-2

Chandrayaan-2 is an enhanced version of the orbiter spacecraft that was launched. Chandrayaan-2's two components are the 'Orbiter Craft' and 'Lander Craft' The three-ton Orbiter Craft Module bus structure is made of a core composite cylinder, strain webs, and deck panels. HAL developed it. The additional spacecraft subsystems and payloads were built onto it at the ISRO Satellite Center in June 2015. The Chandrayaan-2 project will send a Lander module with a Rover to the moon to put an Orbiter around it. GSLV-MKII is slated to launch Chandrayaan-2 in the first quarter of 2018.' The primary goals of the Chandrayaan-2 mission are to strengthen the scientific objectives of Chandrayaan-1 by improving the resolution of the observation of orbiter research payloads. This will be achieved by the design, realization, and deployment of a Lunar Lander module that can perform a soft landing on a specified point on the moon. In addition, the mission will deploy a Rover to conduct on-site research and remote sensing analysis of the lunar surface [1]. The configuration of Changrayaan-2 is shown in Fig. 1'.

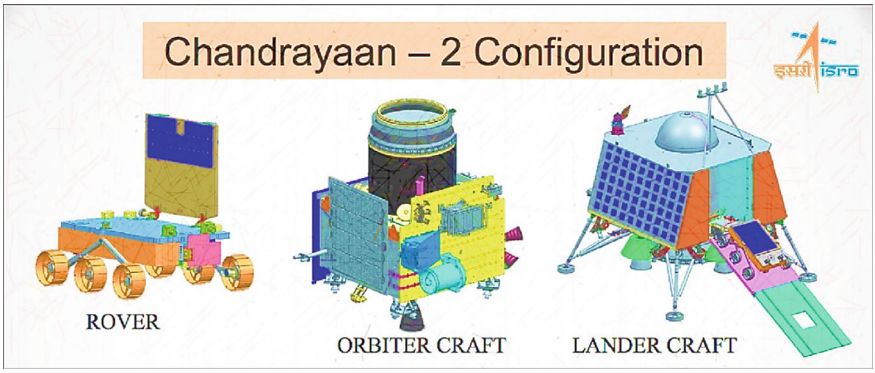


Fig. 1 Configuration of Chandrayaan-2 [32]

1.4 Brief Introduction of Chandrayaan-3

'India made history by being the first nation to successfully land on the moon's South Pole area, which had previously been less explored. However, the journey that led to the achievement of this crucial milestone began in 2008, when India launched its first lunar mission, which was given the designation Chandrayaan-1 and had eleven payloads. The south pole of the moon was the primary target of that project, and the mission was planned to discover the mineral composition of the moon, as well as the water–ice composition of the moon's poles, the chemical composition of lunar rocks, and other information on the development of the moon. Chandrayaan-2, India's second lunar expedition, was launched after the success of the first mission. It consisted of eight payloads, a lander, and a rover, and its purpose was to expand the research conducted by Chandrayaan-1 and to show a soft landing on the lunar surface. Due to a technical problem, the lander was unable to maintain communication with the ground station barely 2.1 km above the moon's surface; nevertheless, the orbiter has remained operational until now. Immediately following that, India launched its third lunar mission, Chandrayaan-3, to complete its objective of showing a smooth landing on the moon's surface. On August 23, 2023, India made history by being the first nation to land on the moon's South Pole. It was also the fourth nation to show a soft landing on the moon during the process [2]. Figure 2 represents the Chandrayaan-3 Mission components.

1.5 Introduction to Lunar Science

Before quite recently, the Moon was the first and most extensively explored destination in space travel. This was the case until relatively recently. There has been a significant amount of satellite exploration over the course of history, as shown by

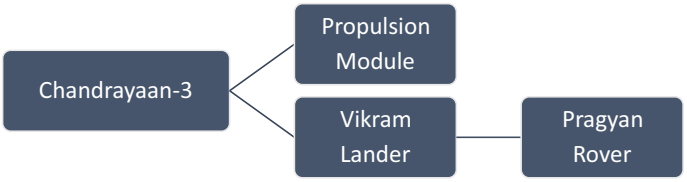


Fig. 2 Chandrayaan-3 mission components

the fact that more than 130 moon missions have gone into space. The robotic Luna and crewed Apollo landing missions, the age of high-resolution orbital surveys, and the current revival of lunar exploration with an emphasis on returning to the Moon are the three distinct eras that can be typically distinguished from one another in this exploration tour. The path of analysis has been significantly impacted by each of these periods in their own unique way. When it comes to the progression of humankind's exploration of outer space, the exploration of the moon is generally regarded as the initial step. Additionally, it is regarded to be crucial for humanity to learn the beginnings of the system that consists of the Earth and the Moon. Over the course of the past ten years, the Chinese Lunar Exploration Program (CLEP), which is also known as the Chang'e (CE) Project, has achieved a number of key milestones. These milestones include the exploration of the moon, Earth, and other celestial bodies. The Chinese government's first crewed lunar exploration mission is scheduled to touch down on the lunar near side between the latitudes of 20 degrees north and south. The site is in the Procellarum KREEP Terrane (PKT), which is incompatible with other terrains. There is a striking resemblance between this site and the landing spots of the six crewed flights that Apollo conducted. The challenges of discovering new discoveries and expanding scientific understanding about the moon is presented by the presence of engineering limitations. The subsequent crewed lunar exploration mission that China plans to launch is anticipated to accomplish ultra-long-distance excursions on the Moon's surface that span thousands of kilometers. When it comes to exploiting this technological and technological success to promote scientific research and lunar science, China's crewed lunar exploration presents both obstacles and potential. This research analyzes the current state of affairs, tendencies, and challenges associated with lunar exploration and research operations. The purpose of this research piece is to provide an overview of scientific concerns regarding the Moon that may be of use to crewed missions while also meeting technical criteria. The research presents novel ideas and suggestions for the exploration and analysis of the moon on the moon in the future that will be conducted by humans [3].

According to the research conducted by [4], The Moon has been an essential component in the process of interpreting the development of the solar system. The findings of previous missions have not only revealed current new insights, but they have also brought up significant new issues that have consequences that extend beyond the Moon. Through the use of orbiters, landers, and sample return missions, there are now active efforts being conducted all over the world to answer some

of these new concerns, as well as old problems that have not yet been resolved. This includes the Chang'E-4 mission, which is the first far side landing, as well as the Chang'E-5 mission, which is the sample return mission. Several missions from national space agencies are now in early phases of production. The Moon has been an essential component in the process of interpreting the development of the solar system. The findings of previous missions have not only revealed current new insights, but they have also brought up significant new issues that have consequences that extend beyond the Moon. Through the use of orbiters, landers, and sample return missions, there are now active efforts being conducted all over the world to answer some of these new concerns, as well as old problems that have not yet been resolved. Both the Chang'E-4 mission, which was the first landing on the far side, and the Chang'E-5 mission, which was the sample return mission, are included in this group of missions. Many missions that are being developed by national space agencies are now in the preliminary stages of manufacturing. The Chandrayaan-2 mission, which consisted of the first landing at high latitudes on the moon, and the Lunar flashlight mission, which involved multi-laser probing of the lunar poles in the construction of the CubeSat, are two instances of missions that have been carried out. LunaGlob, which is a mission to the lavatube skylights and the magnetic anomalies, is one of the many extra missions that are now being planned or have been proposed. There are also a number of other missions that are currently being planned. A number of opportunities may present themselves, one of which is the possibility of doing scientific study from the Moon. This is in addition to the research that is being done on the Moon itself. The active participation of private sector has been one of the most significant advancements in the field of lunar exploration. This is possibly the most crucial component in maintaining a presence on the Moon for an even longer amount of time, and it has been one of the most significant developments in the field. There have been a number of companies such as Team Indus (India), Astrobotics (United States of America), Moon Express (United States of America), and ILS (Israel), that have been working hard for more than a decade to make it possible for private individual to land on the moon. Even while such a significant achievement has not yet been accomplished, the likelihood of it occurring in the not-too-distant future has been constantly increasing. Additionally, there are current efforts being conducted to analyze the possibility of using lunar material for the purpose of meeting future infrastructure requirements.

Previous research obtained the Moon's spherical form, ignoring terrain errors. The landscape on the Moon is extremely complicated and harsh. The lunar meadow and the highlands are the two basic groups that make up the terrain. At an average distance of one to two kilometers below the lunar geoid, lunar animals are dark, vast, and low plains on the Moon. They are similar to the basins that are seen on Earth. Highlands are regions that are located two to three kilometers above the lunar geoid, which is above the animal. Both the border and the highlands of the lunar animals have a surface that slopes downward. There is a possibility that it will be impossible for Moon-based Earth observation devices to maintain a line of sight in regions that experience considerable topographical variability. In order to conduct

Moon-based Earth observation, it is essential to conduct an analysis of the intervisibility between the sensor and the target on Earth throughout the total lunar terrain. Within the scope of this research, we provide a Moon-based Earth observation intervisibility technique on the Moon. LOLA, which stands for the Lunar Orbiter Laser Altimeter, is the instrument that provides the foundation for this approach. This method presents the findings of the locations in which the sensor's line of sight would be disrupted, taking into mind the fact that the sensor is positioned in either the Sinus Iridum or the Mare Nectaris zone where it is positioned. It is believed that this intervisibility method will provide assistance in the selection of locations for future lunar exploration missions or Earth surveillance systems that are situated on the Moon. The results of the experiments demonstrate that the approach that was presented is capable of revealing the intervisibility conditions that exist in the region. Within the context of future Moon-based Earth monitoring system or lunar exploration missions, this inter-visibility method will prove to be of great assistance in the detection of suitable landing sites [5].

Space programs worldwide continue to prioritize lunar exploration, which has seen a rebirth in the past decade. The Indian lunar exploration program aligns with world interests. India began lunar exploration in 2008 with Chandrayaan-1, a mission that deployed an orbiter in a 100 km circular polar orbit around the Moon. The scientific instruments functioned for around 10 months, transforming our understanding of the Moon. The Moon Impact Probe detected exospheric water and gave high-resolution photos during its fall crushing near the South Pole. Chandrayaan-2, a follow-up mission including an Orbiter, Lander, and Rover, was launched on July 22, 2019. Despite the unsuccessful soft-landing attempt, the orbiter continues to operate with eight scientific experiments. The Ch-3 mission is essential for the lunar program, demonstrating soft landing, wandering, research payload deployment, southern high latitude research, and viewing Earth from lunar orbit. In spite of this, some of the unresolved scientific problems that would be addressed by Chandrayaan-3 analyses include reliable data on the chemistry of the lunar surface, seismic accelerations, thermal conductivity of the regolith, and the environment of the lunar plasma. It is believed that the in-situ measurements taken by Chandrayaan 3 would provide confirmation of the worldwide observations made by Chandrayaan-1 and Chandrayaan-2. The Chandrayaan-3 mission was launched on July 14, 2023 where the flow diagram is shown in Fig. 3. Launched by LVM3-M4, it was put in a 169 km $\times$ 36,490 km Elliptic Parking Orbit. The combined Chandrayaan 3 module includes a Lander, Rover, and Propulsion Module (PM). The Lander weighed 1750 kg, including the 26 kg Rover. To guarantee a smooth landing, Vikram was powered down using an ALS to the specified landing spot, guided by a closed loop Navigation, Guidance, and Control system. Landing occurred on August 23, 2023, at 69.3676 S, 32.3481 E on the Moon. The Chandrayaan-2 Orbiter's OHRC captured the lander. The Rover went about to the landing site and covered approximately 100 m after landing. Both the lander and rover were put into sleep mode after their nominal mission life.

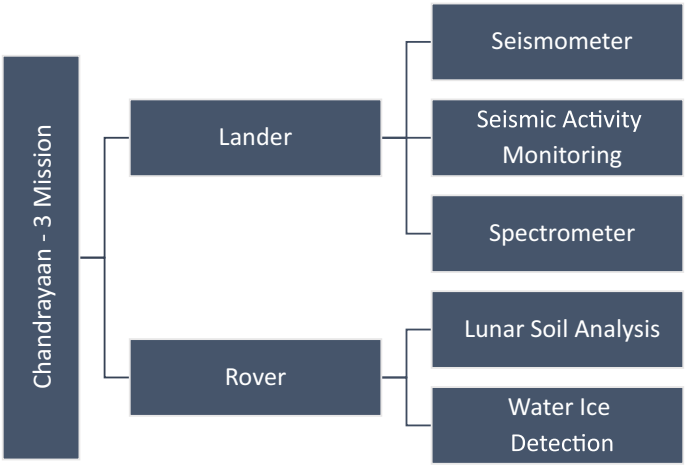


Fig. 3 Chandrayaan-3 lunar science mission

2 Background

2.1 Background History of Chandrayaan-1

'In 2008, the Indian Space Research Organization (ISRO) successfully launched the Chandrayaan-1 orbiter mission to the Moon, initiating India's planetary exploration endeavors. Aboard Chandrayaan-1, a total of eleven remote-sensing scientific equipment contributed by the Indian Space Research Organization (ISRO), the National Aeronautics and Space Administration (NASA), and the European Space Agency (ESA) have achieved noteworthy discoveries. Water signature, spinel minerals, lava tubes on the moon, signs of recent volcanic activity, boulder movements caused by impacts, and sputtered atomic oxygen and backscattered helium are also part of the findings.'

The Mars Orbiter Mission (MOM), India's inaugural interplanetary endeavor, was effectively launched by a PSLV rocket on November 5, 2013. On September 24, 2013, it successfully entered an elliptical orbit around Mars, carrying five science packages.'

2.2 Background History of Chandrayaan-2

'The Chandrayaan-2 spacecraft has three components: the Orbiter, Lander, and Rover. This composite module was sent to the Moon. In the first three months of 2018, the Chandrayaan-2 Mission is expected to be launched aboard the Geosynchronous Satellite Launch Vehicle (GSLV). The consolidated stack will be transported to the

moon by the Orbiter until the Lunar Orbit Insertion (LOI) occurs. Once the stack is firm, it is placed into a 100 km × 100 km lunar orbit. 'Subsequently, it is anticipated that the Lander, carrying the Rover, would detach from the Orbiter in order to execute an easy landing in nearby of the southern polar region of the lunar surface [1].

2.3 Background History of Chandrayaan-3

Chandrayaan-3, the third lunar mission, was launched by the Indian Space Research Organization (ISRO) on July 14, 2023, using the rocket LVM III- M4 from the Sriharikota Space Development Center (SDSC-SHAR). The mission included the Vikram lander and the Pragyan rover. Beginning on the 5th of August, 2023, the Chandrayaan-3 spacecraft successfully entered lunar orbit. After a series of irrelevant reduction actions, the lander was ultimately released for a gentle touchdown on the lunar surface. On August 23, 2023, India made history by being the first nation to land on the moon's South Pole. Performing scientific experiments and demonstrating a gentle landing on the moon were the primary objectives of this mission. Additionally, the rover was supposed to roam around on the lunar surface. Up to a slope of 12 degrees, the lander is intended to permit a safe landing condition. This function lasts for one lunar day, equivalent to fourteen days on Earth. It is equipped with a Laser Altimeter (LASA), a Lander Horizontal Velocity Camera (LHVC), a Lander Position Detection Camera (LPDC), and a Lander Hazard Detection Avoidance Camera (LHDAC). With its six wheels, the Pragyan rover can traverse a distance of 500 m across the lunar surface at a pace of one centimeter per second. This is intended to map the surface of the moon. "Shiv-Shakti Point" was the name the Prime Minister of India gave to the location where the Vikram lander landed (Chandrayaan program, Wikipedia). In addition, he stated that the 23rd of August should be marked as National Space Day as a way to pay gratitude for the successful landing of the Vikram lander on that day [2].

3 Challenges Faced by Chandrayaan-2

Chandrayaan-2, India's second mission to explore the moon, was launched on July 22, 2019, to achieve a gentle touchdown on the lunar surface. Nevertheless, the mission was hindered as the Vikram lander could not achieve a triumphant and gentle touchdown on the lunar surface. ISRO's communication with the lander was severed when it was a mere 400 m from the lunar surface [7, 11]. According to the failure analysis report presented to ISRO, the lander's crash was attributed to a software problem. This glitch led the lander to veer off course from its planned trajectory at an altitude of 2.1 km, and communication was lost before the expected confirmation of touchdown was supposed to occur [6, 7]. The lander was expected to decelerate significantly and reduce its speed to a minimum as it approached 400 m

from the lunar surface. This would allow it to remain stationary over the designated landing spot and perform a gentle vertical drop. Nevertheless, due to the software malfunction, the lander deviated from its planned path and collided 750 m from the designated landing location [2, 7].

The occurrence of the software fault was unforeseen, as the software had been operating smoothly throughout the whole trial period. The Vikram Lander smoothly descended from an altitude of 30–5 km. Following the occurrence of "rough braking," the lander encountered difficulties during the "fine braking," which is the last stage of its descent. The accident caused significant damage to the machinery on board, resulting in a loss of communication from the lander [2]. The trajectory divergence was found at an altitude ranging from 2.1 to 0.2 km above the lunar surface. A communication problem arose during the abrupt deceleration phase of the landing, precisely at 20:23 UTC on September 6, 2019. Subsequent endeavors to restore communication from the lander proved futile [8].

There were three mistakes that led to the hard landing of Vikram:

Higher Thrust from Engines: The primary issue arose from the fact that the five engines employed to decrease the speed generated a greater amount of thrust than initially anticipated [3].

High Velocity: The lander was expected to decelerate significantly and reduce its speed to a great extent by the time it reached a distance of 400 m from the lunar surface. Nevertheless, as a result of its considerable speed, it collided into the surface of the Moon [2, 3].

Limited Ability to Handle Parameter Dispersion: The problem with Chandrayaan-2 was the severe limitation in its ability to manage parameter dispersion. This problem arose when the lander, despite its proximity to the surface, accelerated as it approached the landing spot, which was a considerable distance away [3].

The crash-landing of the Vikram lander was caused by its deviation from the anticipated tilt angle of 90 degrees, instead tilting at an angle of 410 degrees [11]. The aforementioned problems resulted in the unsuccessful accomplishment of Chandrayaan-2's goal to achieve a gentle touchdown on the lunar surface. The unsuccessful outcome of the Chandrayaan-2 mission yielded vital insights for forthcoming lunar expeditions. The insights acquired from the Chandrayaan-2 mission were applied to tackle the intricacies of the Moon's southern pole and enhance the preparations for the Chandrayaan-3 mission [1].

4 Challenges Resolved by Chandrayaan-3

In order to resolve the challenges encountered by Chandrayaan-2 and guarantee the triumph of Chandrayaan-3, ISRO enacted many modifications and applied advanced technological enhancements:

(1) **Robust Design and Testing**: The design of Chandrayaan-3 included both hardware and software upgrades, which involved improving the legs of the lander to enable a safe landing on the lunar surface. The mission also included comprehensive redundancies at every level, rigorous simulations, and further testing carried out at various phases [9, 10].
(2) **Improved Guidance and Control Systems**: The mission incorporated sophisticated sensors, actuators, and other essential hardware to improve navigation, positioning, and control throughout the landing phase. Chandrayaan-3 was equipped with a larger quantity of fuel and an improved guidance, navigation, and control system, enabling it to effectively rectify significant deviations from its intended trajectory [4, 5].
(3) **Mission Design Modifications**: Modifications were made to the trajectory and approach angles, resulting in enhanced control during the descent and landing phase. These changes were strengthened through extensive ground tests using helicopters and cranes [4].
(4) **Learning from Past Experiences**: ISRO applied the acquired knowledge from Chandrayaan-2 to tackle the intricacies of the Moon's South Pole. Novel simulation test beds were developed to examine failure situations, and the spacecraft underwent rigorous testing to evaluate its resilience against vibrations and landing procedures [1].
(5) **Enhanced Autonomy**: Chandrayaan-3 was equipped with the ability to independently manage a broader spectrum of dispersion, enhancements in sensors, software, and propulsion systems. The purpose of this was to guarantee that in the event of a complete sensor failure, the lander would still successfully execute a landing [1, 9].
(6) **International Collaboration**: The mission incorporated a NASA passive Laser Retroreflector Array for the purpose of lunar laser range investigations, demonstrating global teamwork and scientific cooperation [9].

The careful approach adopted by ISRO in the creation of Chandrayaan-3, along with the implemented improvements, significantly contributed to the mission's successful outcome. Chandrayaan-3 successfully landed on the Moon, marking India as the fourth country to accomplish this remarkable achievement [5]. The accomplishment of the mission serves as evidence of ISRO's capacity to assimilate knowledge from previous encounters, adjust accordingly, and employ novel technology to surmount the obstacles associated with lunar exploration.

5 Comparative Analysis of Chandrayaan-2 Versus Chandrayaan-3

ISRO, which is an initial that stands for the Indian Space exploration Organization, is now finishing up preparations for its third lunar exploration mission, which is called Chandrayaan-3 and is scheduled to be launched on July 14, 2023. In addition to rover

exploration, the purpose of this mission is to correct the problems that occurred during its predecessor, Chandrayaan-2, and to successfully perform a soft landing on the surface of the moon after it has completed its exploration. Within the parameters of this article, we will analyze the fundamental differences that exist among these two missions.

5.1 Mission Design and Approach

Chandrayaan-2

- In addition to an orbiter, the Chandrayaan-2 mission also included a lander called Vikram and a rover called Pragyan.
- A success-based approach was used in the mission's planning to achieve a precise landing within a prescribed 500-m-by-500-m rectangle.
- In addition to having five legs, the Vikram lander was equipped with a central thruster.

Chandrayaan-3

- Failure-based techniques are the main objective of Chandrayaan-3's design, which aims to improve the mission's overall success.
- The orbiter is not included in the mission's architecture; instead, it consists of a lander called Vikram and a rover called Pragyan.
- Because of the expansion of the landing area, the lander can now safely touch down anywhere within a footprint that is 4 km by 2.4 km.
- The Vikram lander has been outfitted with more fuel to enable it to travel further distances to the landing spot or an alternative location, should it be required.
- High-resolution images from Chandrayaan-2's orbiter have been included in the landing site analysis process, which has improved.
- Several modifications have been made to the physical construction of the lander. These modifications include the removal of the center thruster, the strengthening of the legs in order to permit more notable velocity landings, and the addition of additional solar panels in order to facilitate enhanced power production.

5.2 Payloads and Scientific Instruments

Chandrayaan-2

The Chandrayaan-2 spacecraft carried scientific equipment on its orbiter, lander, and rover as part of its payload requirements.

- Among other things, the orbiter was equipped with equipment to investigate the Moon's surface and exosphere.

- For the purpose of conducting elemental analysis, the Pragyan rover was outfitted with both an Alpha Particle X-ray Spectrometer (APXS) and a Laser Induced Breakdown Spectroscope (LIBS).
- RAMBHA, ChaSTE, ILSA, and LP were the payloads that were carried by the Lander spacecraft. Each of these payloads was meant to measure a different component of the lunar environment.

Chandrayaan-3

Chandrayaan-3's propulsion module is responsible for transporting the SHAPE payload, which is intended to use spectro-polarimetry to look for planets that might potentially support life.

Lander payloads include ChaSTE, ILSA, and LP, comparable to Chandrayaan-2, measuring seismicity, thermal characteristics, etc. RAMBHA is a NASA-developed instrument.

To do an in-situ elemental composition study, the rovers' payloads remain consistent with APXS and LIBS.

5.3 Implementation and Soft Landing

Chandrayaan-2

An attempt was made by the Vikram lander to perform a soft landing on the surface of the moon; but, owing to a malfunction that occurred at the very last minute, the lander was unable to accomplish this goal during the descent phase.

As a result of the failed landing, the Pragyan rover, which was part of Chandrayaan-2, could not be deployed.

Chandrayaan-3

During the Chandrayaan-3 mission, the Vikram spacecraft is going to do a soft landing, and the Pragyan rover is going to be sent to the surface of the moon in order to carry out explorations.

Alterations to the design and enhancements to the process of determining the landing spot both increase the likelihood of a successful landing.

5.4 Challenges

Because of the failure of Chandrayaan-2, the variety and challenges of lunar exploration were brought to light, which led to the development of a failure-based strategy for Chandrayaan-3.

Chandrayaan-3 is successful in achieving its mission goals with remarkable performance; India has become the fourth nation in the world to successfully make a soft landing on the Moon, following the United States of America, Russia, and China.

6 Uniqueness of the Chandrayaan-3 Mission

Chandrayaan-3, India's third expedition to the moon, has achieved notable success due to its distinctive characteristics and meticulous preparation. The mission's objective, which is a continuation of Chandrayaan-2, is to showcase the complete ability to land safely and move about on the moon's surface [13]. The mission commenced on July 14, 2023, with an anticipated operational duration of two weeks. It aimed to conduct a range of experiments, including the examination of the lunar surface's mineral composition using a spectrometer [12].

The uniqueness of the Chandrayaan-3 missions lies in several aspects that have contributed to its success which illustrated in Fig. 4. These include:

- **Lander and Rover Configuration**: Chandrayaan-3 is comprised of a Lander and Rover setup. This arrangement diverges from Chandrayaan-2, which also encompassed an orbiter [16]. The Lander's height is approximately 2 m, while its mass exceeds 1700 kg, which is comparable to that of an Soviet Union. The purpose of its design is to facilitate the deployment of a more compact lunar rover weighing 26 kg [12].
- **Targeted Lunar South Pole**: The objective of the Chandrayaan-3 mission is to explore the lunar south pole, which is known to contain water ice. This frozen water has the potential to serve as a valuable resource for future moon missions, including the production of oxygen, fuel, and water for sustaining a potential moon colony [12].
- **Improved Landing System**: The lander has been enhanced with additional fuel capacity and reinforced legs to withstand impact. Furthermore, a method has been implemented to expand the probable landing area, hence increasing the probability of mission success [12].
- **Innovative Technologies**: The Lander incorporates various sophisticated technologies proven successful in terrestrial circumstances through a series of specialized tests [13]. ISRO employed several cutting-edge technologies in the Chandrayaan-3 project, including as a novel lander configuration and an advanced rover. [17].
- **Successful 'Hop' Experiment**: The project achieved another notable milestone with the successful execution of the 'hop' experiment by the Vikram lander, signifying a significant advancement in the field of lunar exploration [15].
- **Scientific Payloads**: The mission transported multiple scientific payloads to investigate the lunar surface. The payload called Spectro-polarimetry of Habitable

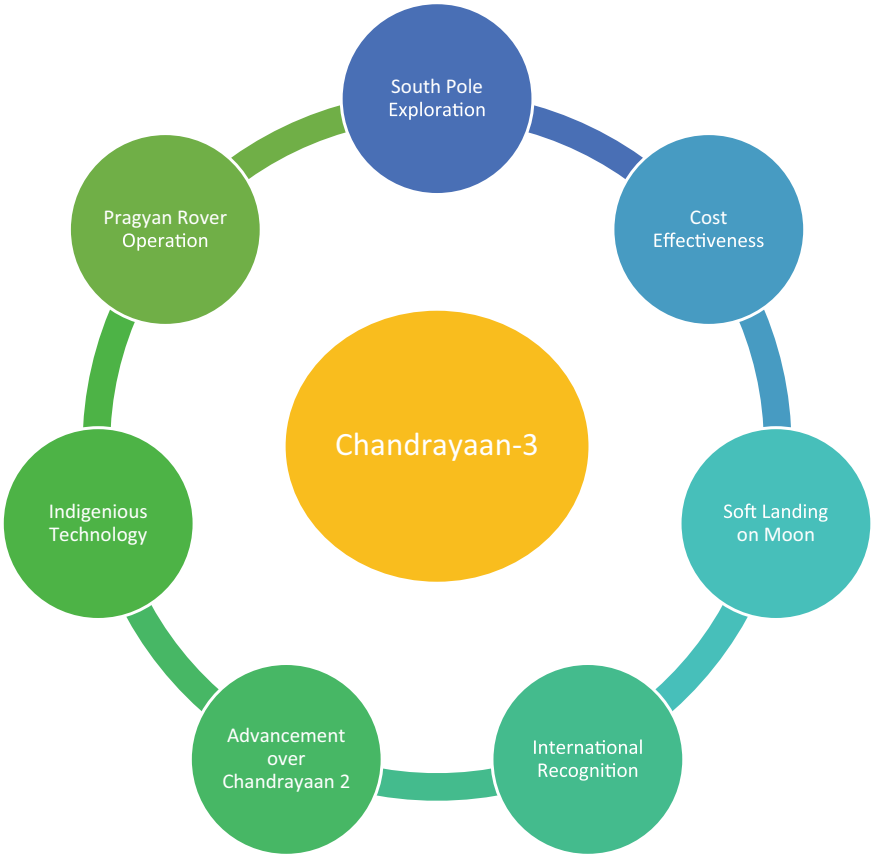


Fig. 4 Uniqueness of Chandrayaan-3 mission

Planet Earth (SHAPE) was used to investigate the Earth's atmosphere. Additionally, Chandra's Surface Thermophysical Experiment (ChaSTE) was employed to test the thermal conductivity and temperature gradient of the lunar surface [13, 19].

7 Scientific Discoveries of Chandrayaan-3

Four components make up Chandrayaan-3: a rover, a lander module (LM), and a propulsion module (PM) that were all manufactured in India. In order to accomplish the mission's objective, new technologies for interplanetary missions were developed and demonstrated.

The Vikram lander includes a number of instruments, such as the Chandra's Surface Thermophysical Experiment (ChaSTE), which is designed to measure

surface thermal properties; the Instrument for Lunar Seismic Activity (ILSA), which is designed to measure seismicity around the landing site; the Radio Anatomy of Moon Bound Hypersensitive ionosphere and Atmosphere (RAMBHA), which is designed to analyze the gas and plasma environment; and a passive laser retroreflector array, which was provided by NASA for lunar ranging studies. All of these instruments are installed on the Vikram lander.

The rover Pragyan, which is Sanskrit for "wisdom," is going to carry out a chemical survey of the surface of the moon at its current location within the mountains that are located near to the south pole of the Moon. This survey will take place while the rover is traveling around in the mountains. At this site, the six-wheeled rover will conduct experiments for a period of one lunar day, which is equivalent to 14 days on Earth.

Two different instruments are installed aboard the rover, and their purpose is to conduct an analysis of the elemental composition of the local surface. An Alpha Particle X-ray Spectrometer (APXS) and a Laser Induced Breakdown Spectroscope (LIBS) are the instruments that are considered to be in concern.

Immediately following the failure of Chandrayaan-2, the focus was directed towards its essential landing, which required intensive simulations. The lander was outfitted with landing legs that were more robust, an increased amount of fuel, and enhanced sensors. There was a modification made to the program, and the area for landing was increased.

During the process of manufacturing the two instruments that Pragyan is currently carrying, there have been a number of challenges to overcome.

The most difficult challenge was to construct the electronics in such a way that they would not damage the superb energy resolution that was offered by the cutting-edge Silicon Drift Detector (SDD), which is the X-ray detector that is utilized in APXS. There are Cm-244 radioactive sources that are utilized by the APXS system in order to animate materials that are located on the surface of the moon. It was a significant problem to acquire such sources and to manage them throughout the development cycle and the various stages of testing [27]. This was owing to the radiation safety concerns that were involved. Figure 5 shows the scientific discoveries of Chandrayaan 3.

8 Impact of Chandrayaan-3 Success

Chandrayaan-3's triumph has solidified India's position as a prominent participant in space exploration [15]. India's space ambitions and worldwide stature have been enhanced [14]. The successful outcome of the mission is a credit to the diligent efforts and unwavering commitment of the scientists and engineers at ISRO [17]. The knowledge gained from the mission will be extremely helpful for future space exploration missions and ISRO's future interplanetary expeditions [18]. The mission has moreover yielded useful observations regarding the Moon's geology, mineralogy,

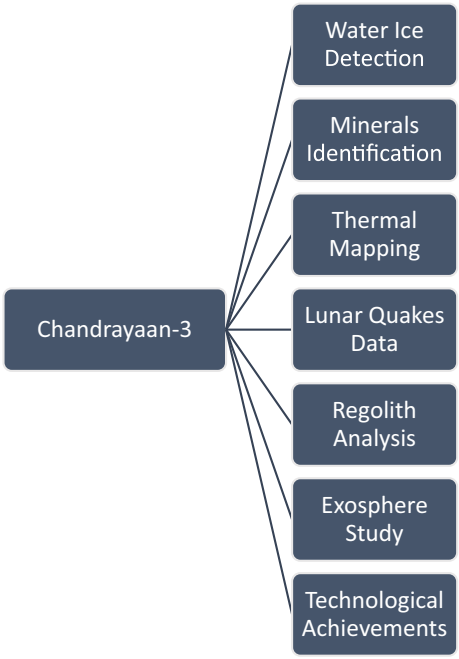


Fig. 5 Scientific discoveries of Chandrayaan-3

and ecology, enhancing our comprehension of the world that humanity intends to occupy in the near future [15].

Furthermore, the achievement of Chandrayaan-3 is anticipated to exert a substantial influence on India's industrial and technical ecosystem. Union Finance Minister Nirmala Sitharaman emphasized that the success of Chandrayaan-3 will significantly influence India's industrial and technological sectors, particularly in terms of the country's space industry and scientific progress [16]. Chandrayaan-3's achievement may be credited to its collaborative approach, emphasis on accurate landing, and its twin focus on technology demonstration and scientific investigation. These factors have significantly contributed to India's increasing expertise in space research. The mission has enhanced India's space exploration capabilities and yielded vital scientific information that will influence future lunar exploration initiatives.

9 Phases of Chandrayaan-3 Mission

The Chandrayaan-3 mission lasted for around 40 days and consisted of a series of phases that were totally devoted to the various stages of the spacecraft, beginning with the launch and ending with the landing. Before going on to the lunar transfer trajectory, there are three sub-stages of launching and rotating about the planet that

occur during Phase 1: the Centric Phase. These sub-stages are known as the Pre-launch Phase, Launch and Ascent Phase, and planet-bound Manoeuvre Phase. The phase that follows the lunar transfer trajectory is called the centric phase. During this phase, the satellite begins to revolve around the moon and progressively gets closer and closer to the surface of the moon. The lander was deployed by the propulsion module after it had travelled roughly 100 km into lunar orbit. The lander then enters the surface of the moon and, upon detecting the gravitational pull of the moon, it applies a downward force to the lander as it moves towards the lunar surface. It eventually lands on the moon's south pole, releasing the rover that is carrying a variety of instruments that have been finely calibrated for the purpose of conducting in-situ studies of the moon (https://www.isro.gov.in/) [26].

The instruction to transition from a horizontal orientation to a falling orientation was delivered to the lander during the landing phase. This command was sent when the lander was at an altitude that was greater than 30 km from the surface of the moon. The time it took was around 690 s, and the speed was 1.68 km per second. During the subsequent phase, the lander was given the instruction to transition from a horizontal position to a vertical position when it was more than 7.4 km away from the surface of the moon. At the same time, the speed at which it approached the surface was slowed down. Approximately 10 s were allotted for the maneuvering process. Following the completion of the last step of the landing process, the lander remained in a vertical posture during the whole journey, covering a distance of 28.52 km to arrive at the planned landing spot. Finally, the lander's velocity was lowered to zero. In the subsequent step, the barriers that were present at the landing site were analyzed by carrying out the necessary tests. This was done in order to determine whether or not it would be advisable to proceed with the preplanned landing or to move the landing area to a more secure one. Remote monitoring and control, greater conditions awareness, improved communications, and automated decision making were all essential components of the Chandrayaan-3 mission throughout its whole, beginning with its launch and moving all the way through to its gentle landing. Devices and sensors connected to the internet of things made this a reality. It is the responsibility of Internet of Things sensors to analyze and continue monitoring the health of Chandrayaan-3's engines, batteries, and other essential systems in order to guarantee precise and efficient functioning. While Internet of Things devices communicate data from cameras as well as sensors in order to monitor the Chandrayaan-3 and keep track of the studies that its gadgets and sensors carry out, internet of things devices do not transmit data [20–25]. The Chandrayaan-3 is able to optimize its navigation and communication systems through the utilization of the Internet of Things (IoT) [26]x'.

10 Improvements in the Chandrayaan-3

The complicated Chandrayaan-3 mission has demonstrated India's ability in being able to handle complex spacefaring missions such as this one, which no other country in the world, not even rich ones, has been able to accomplish successfully. Among the many accomplishments that India accomplished via carefully synchronized operations, the safe soft landing on the south pole of the moon will be recognized as one of the most significant in history. The Internet of Things was utilized at every stage of the development of this significant achievement in the fields of astronomy and engineering. This spacecraft comes equipped with a variety of Internet of Things devices and sensors, which allow it to monitor its own health and keep track of the progress it has made in addition to being able to manage it. It was also discovered that there was evidence of significant space weathering. It was determined through the analysis of the soil's composition that the soil that was found at the landing location was of the highland type. The primary landing location was subjected to a thermophysical analysis, which revealed that the temperature varied between 30 degrees Celsius and 175 °C during the day. Therefore, this journey to the moon will contribute to a better knowledge of the surface of the moon [26].

11 Outcome from the Chandrayaan-3

A considerable amount of knowledge was revealed as data from Chandrayaan-3. The data was collected from each component of the Chandrayaan-3 and released by the Indian Space Research Organisation (ISRO) for scientific research utilization [27]. The data is collected using all three major components of Chandrayaan-3: navigation camera, rover, and lander. Therefore, a total of 3 datasets are collected and each dataset's archived directory is available from the Indian Space Research Organisation (ISRO) [27]. Each dataset has a specific structure in an archived directory including mission archived structure, instrument archived structure, data products, and naming conventions. All these structures are provided and discussed in the documentation of each dataset as well. A mission-archived structure organizes the data concerning the mission which is Chandrayaan-3 while an Instrument-archived structure organizes with concerning the level of processing of data and its products. All this data can be analyzed using the planetary data system (PDS) version 4 visualization tool [28].

The **logical_identifier** uniquely identifies the data product, while the **version_id** specifies its version. The **title** describes the data product, and **information_model_version** indicates the model version used. The **product_class** categorizes the type of product, and the **modification_date** records the last update. The **description** provides a summary of the data, while **start_date_time** and **stop_date_time** define the time range of the observation. The **purpose** explains the data's intent, and **processing_level** reflects its data processing stage. The **name** gives the product

name, and **type** defines its nature (e.g., satellite). The **lidvid_reference** links to the mission context, and **reference_type** specifies its usage. The **job_id** relates to the processing task, and **dumping_orbit_number** identifies the orbit during data collection. The **dop** is a reference quality metric, and **frame_number** specifies the captured frame. The **current_mobility_id** and **future_mobility_id** track rover states, while **operation_id** is linked to mission activities. The **day_no** and **year** define the temporal context. The **inclinometer_utc** records the inclinometer timestamp, with **inclinometer_x_angle**, **y_angle**, **roll_angle**, and **pitch_angle** describing orientation data. The **sensor_id** identifies the imaging device, and **gain** specifies signal amplification. The **integration_time** defines the camera exposure duration. The **file_name** contains the data file's name, **creation_date_time** is when the file was created, and **file_size** indicates its size. The **md5_checksum** ensures data integrity, the **comment** provides additional notes, the **offset** sets a calibration reference, and the **encoding_standard_id** defines the format standard used.

Navigation Camera Data Archive: The archive structure of navigation camera data is a collection of two types of data including raw data and calibrated data [29]. The raw data has no correlation and specific meaning while calibrated data provides meaningful data with radiometrically corrected data with refined coordinates. Each data comprises two files with different formats, including the image data in JPG format, and the corresponding label in XML format. The navigation camera data archive provides documentation that describes the directory-level structure and naming convention meanings of both raw and calibrated data [29].

Figure 7 shows a calibrated image captured by the Navigation Camera onboard Chandrayaan-3. This data is essential for rover navigation and terrain mapping. The key field in the metadata table for each image represents the parameter or attribute name used to describe the associated values of the image. Figure 6 shows the raw image from navigation camera.

Rover Data Archive: The archive structure of rover data is a collection of L1 level 1 processing calibrated data [30]. The rover data archived structure is denoted in two formats. The image with JPG format and the corresponding label is in XML format for calibrated data. The rover data archive provides documentation that provides detailed information about this archived directory. Figure 8 demonstrates the calibrated image data from the rover data archive.

Lander Data Archive: The archive structure of lander data is a collection of processed calibrated data [31]. Lander Data Archive structures the data using two formats, including the calibrated data image stored using JPG format, and the calibrated data label corresponding to each image is stored using XML format. The lander data archive provides documentation along with the data to understand the levels of each directory under the archived data and its structure, product description, and naming conventions [31].

Figure 9 depicts a calibrated image from the Lander TTP (Terrain Targeting Payload) data archive, focusing on high-resolution imaging for lander navigation and positioning.

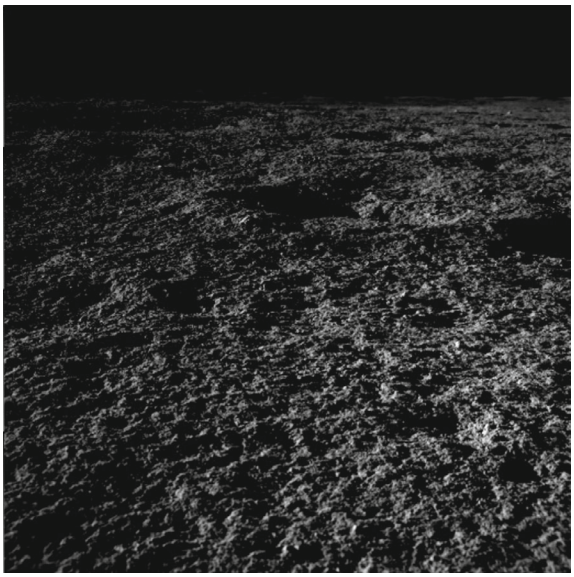
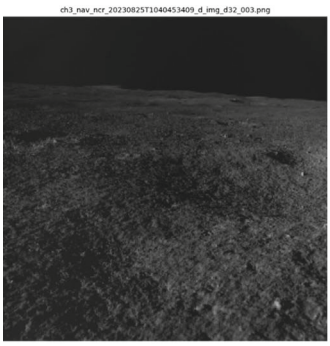


Fig. 6 Raw image from navigation camera



Key	Value
logical_identifier	urn:isro:isda:ch3_chr:nav:data_calibrated:ch3_nav_ncr_20230825t1040453409_d_img_d32_003
version_id	1.0
title	Chandrayaan-3 Rover NavCam Mobility Experiment
information_model_version	1.11.0.0
product_class	Product_Ancillary
modification_date	2024.06.11
description	Moon is a natural satellite of Earth
start_date_time	2023-08-25T10:40:45.340?Z
stop_date_time	2023-08-25T10:40:46.659?Z
purpose	Navigation
processing_level	Calibrated
name	ch3_nav_ncr_20230825T1040453409_d_img_d32_003
type	Satellite
lidvid_reference	urn:isro:isda:context:investigation.mission.chandrayaan3::1.0
reference_type	data_to_investigation
pds_id	NCBXOD32C3R00180000NNNN232371131060O0_V2_1
dumping_orbit_number	180
doy	20230825
frame_number	3
current_mobility_id	002
future_mobility_id	003
operation_id	080001
sley_no	237
year	2023
inclinometer_utc	2023:237:10:49:00:7082023:237:10:49:00:578
inclinometer_x_angle	0.907
inclinometer_y_angle	4.672
inclinometer_roll_angle	0.0
inclinometer_pitch_angle	0.0
sensor_id	right camera
gain	2
integration_time	23
file_name	ch3_nav_ncr_20230825T1040453409_d_img_d32_003.png
creation_date_time	2024-06-12T15:36:56
file_size	571980
md5_checksum	4a62ffdd6fd6764dcb1b3658762767?7
comment	this file contains the color calibrated image in png format. md5_checksum is provided for ensuring data integrity when users are downloading the data.
offset	0
encoding_standard_id	PNG

Fig. 7 Calibrated image from navigation camera

The raw and calibrated data from the Chandrayaan-3 can be utilized for a diverse range of applications. The image data of the moon's surface reveals important information that can further link to such undiscovered aspects of space science.

Figures 10 and 11 would represent similar calibrated data from specific payloads like PLS (Plasma System) and NOP (Neutron Payload), each contributing unique scientific insights into the lunar environment.

eArtificisl

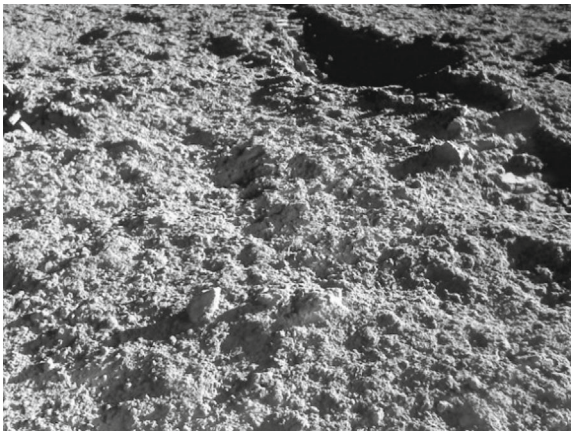
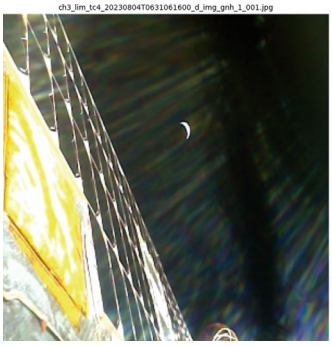
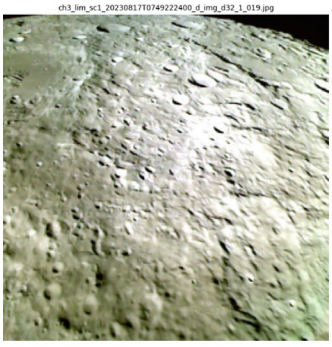


Fig. 8 Calibrated image data from the rover data archive



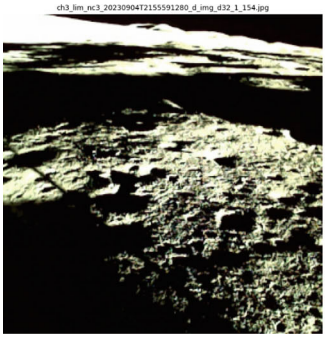
Key	Value
logical_identifier	urn:isro:isda:ch3_chl:lim:data_calibrated:ch3_lim_tc4_20230804t0631061600_d_img_gnh_1_001
version_id	1.0
title	Chandrayaan-3 Lander Imager
information_model_version	1.11.0.0
product_class	Product_Ancillary
modification_date	2024-06-11
description	Moon is a natural satellite of Earth
start_date_time	2023-08-04T06:31:06.160OZ
stop_date_time	2023-08-04T06:31:06.1200Z
purpose	Navigation
processing_level	Calibrated
name	ch3_lim_tc4_20230804T0631061600_d_img_gnh_1_001
type	Satellite
lidvid_reference	urn:isro:isda:context:investigation:mission.chandrayaan3:1.0
reference_type	data_to_investigation
pds_id	LIMXXGNHC3L0002005NNNN232160739309330_V1_1
mode	Mode-1
frame_number	001
imager_id	LI4
file_name	ch3_lim_tc4_20230804T0631061600_d_img_gnh_1_001.jpg
creation_date_time	2024-06-12T17:33:13
file_size	100967
md5_checksum	3e3f2ec0f9b3dbac0914a5da3cf4ddac4d
comment	This file contains raw color calibrated image in jpg format. md5_checksum is provided for ensuring data integrity when users are downloading the data.
offset	0
encoding_standard_id	JPEG

Fig. 9 Calibrated image data from the lander TTP data archive



Key	Value
logical_identifier	urn:isro:isda:ch3_chl:lim:data_calibrated:ch3_lim_sc1_20230817t0749222400_d_img_d32_1_019
version_id	1.0
title	Chandrayaan-3 Lander Imager
information_model_version	1.11.0.0
product_class	Product_Ancillary
modification_date	2024-06-11
description	Moon is a natural satellite of Earth
start_date_time	2023-08-17T07:49:22.160OZ
stop_date_time	2023-08-17T07:49:22.1360Z
purpose	Navigation
processing_level	Calibrated
name	ch3_lim_sc1_20230817T0749222400_d_img_d32_1_019
type	Satellite
lidvid_reference	urn:isro:isda:context:investigation:mission.chandrayaan3:1.0
reference_type	data_to_investigation
pds_id	LIMXXD32C3L0009005NNNN232291101030S4_V1_1
mode	Mode-1
frame_number	019
imager_id	LI1
file_name	ch3_lim_sc1_20230817T0749222400_d_img_d32_1_019.jpg
creation_date_time	2024-06-12T17:31:42
file_size	168705
md5_checksum	6a7f0f5bc216597bcadda84d251ab2b995ba
comment	This file contains raw color calibrated image in jpg format. md5_checksum is provided for ensuring data integrity when users are downloading the data.
offset	0
encoding_standard_id	JPEG

Fig. 10 Calibrated image data from the lander PLS data archive



Key	Value
logical_identifier	urn:isro:isda:ch3_chl:lim:data_calibrated:ch3_lim_nc3_20230904t2155591280_d_img_d32_1_154
version_id	1.0
title	Chandrayaan-3 Lander Imager
information_model_version	1.11.0.0
product_class	Product_Ancillary
modification_date	2024-06-11
description	Moon is a natural satellite of Earth
start_date_time	2023-09-04T21:55:59.128607
stop_date_time	2023-09-04T21:55:59.240002
purpose	Navigation
processing_level	Calibrated
name	ch3_lim_nc3_20230904T2155591280_d_img_d32_1_154
type	Satellite
bdsd_reference	urn:isro:isda:context:investigation:mission:chandrayaan3::1.0
reference_type	data_to_investigation
pds_id	LIMXXD32CSL0018007NNNN232470043570O9_01_V1_1
mode	Mode-1
frame_number	154
imager_id	LI3
file_name	ch3_lim_nc3_20230904T2155591280_d_img_d32_1_154.jpg
creation_date_time	2024-09-25T15:51:13
file_size	154551
md5_checksum	0540e56a35a8b5bcd3f3ebb65c85797c
comment	This file contains the count calibrated imager in jpg format. md5_checksum is provided for ensuring data integrity when users are downloading the data
offset	0
encoding_standard_id	JPEG

Fig. 11 Calibrated image data from the lander NOP data archive

12 Conclusion

Ultimately, the Chandrayaan-3 mission serves as a significant witness to India's expanding proficiency and aspiration in the realm of space research. This mission signifies a substantial advancement for the Indian Space Research Organisation (ISRO) and the nation's space program, as it builds upon the accomplishments and knowledge gained from Chandrayaan-1 and Chandrayaan-2. This review has presented an in-depth analysis of the historical background, progression, and influence of the Chandrayaan missions, with a focus on the growth of India's space capabilities. The innovations made in Chandrayaan-3 specifically target the challenges seen during Chandrayaan-2, particularly the problems with the landing of the Vikram lander. This showcases the ability of ISRO to adapt and recover from setbacks. The particular goals and technological advancements of Chandrayaan-3 emphasize its unique contributions to the exploration of the moon. Chandrayaan-3 is expected to accomplish its mission objectives, leading to significant scientific discoveries that will enhance our knowledge of the Moon and enhance India's reputation in the international space community.

References

1. Durga Prasad K, Misra D, Bhatt M, Ambily G, Sathyan S, Srivastava N, Bhardwaj A (2023) Chandrayaan-3 alternate landing site: pre-landing characterisation
2. Sundararajan V (2018) Overview and technical architecture of India's Chandrayaan-2 mission to the moon. In: AIAA Aerospace Science Meeting, 2018, January, pp 1–12
3. Prasad KD et al (2023) Contextual characterization study of Chandrayaan-3 primary landing site. Mon Not R Astron Soc Lett 526(1):L116–L123
4. Kumar A, Mehfuza M (2023) India'S journey towards moon: from Chandrayaan-1 To Chandrayaan-3. Int Res J Mod Eng 08:2361–2365. www.irjmets.com
5. Sinha RK, Rani A, Ruj T, Bhardwaj A (2023) Geologic investigation of lobate scarps in the vicinity of Chandrayaan-3 landing site in the southern high latitudes of the moon. Icarus 402:115636

6. Sana T, Mishra SK (2023) Plasma sheath around Chandrayaan-3 landing site: a case study. Planet Sci J 4(9):158
7. https://www.isro.gov.in/Chandrayaan3_New.html
8. Calla O, Mathur S, Gadri KL (2016) Possible landing site for Chandrayaan-2 rover. In: 2016 International conference recent advances innovation engineering, ICRAIE 2016, pp 1–5
9. Mithun NPS et al (2020) Solar X-Ray monitor on board the Chandrayaan-2 orbiter: in-flight performance and science prospects. Sol Phys 295(10)
10. Kumar S, Chaudhary V, Singh A, Chauhan P (2021) Scattering-based analysis of south polar crater of the lunar surface using L-band SAR data of Chandrayaan-2 mission. In: 2021 IEEE India geoscience, remote sensing symposium InGARSS 2021—proceedings, vol 3387710. pp 459–462
11. Adriani N (2018) Chandrayaan 3 and other moon missions: an update and review. 23529(2):1–45, SSRN
12. Garg H et al (2023) Spacecraft charging analysis for Chandrayaan-2 mission. In: 2023 Joint Asia-Pacific International Symposium Electromagnetics Compatibility International Conference Electromagnetics Interface Compatibility, APEMC/INCEMIC 2023, pp 1–4
13. Chowdhury AR et al (2020) Terrain mapping camera-2 onboard chandrayaan-2 orbiter. Curr Sci 118(4):566–572
14. Mithun NPS et al (2019) Data processing software for Chandrayaan-2 solar X-ray monitor. Astron Comput 34:2021
15. Bhuva K et al (2021) A machine learning framework for data filtering: a case study on Chandrayaan-1 SIR-2 data. In: 2021 IEEE India geoscience remote sensing symposium InGARSS 2021—Proceedings, pp 483–487
16. Kumar A et al (2022) Dielectric constant estimation of lunar surface using mini-RF and Chandrayaan-2 SAR data. IEEE Trans Geosci Remote Sens 60:1–8
17. Sinha A, Alkhayyat SA, Kumar B, Kumari A, Kumar U (2023) Brain computer interface imagination based thought analyser framwork to generate and classify real time image retrieval system for space images in Chandrayaan 3 Lunar misson. In: 2023 10th IEEE Uttar Pradesh section international conference electrical electronics computing engineering, UPCON 2023, vol 10. pp 882–887
18. Desur AG, Jatoth RK (2024) Enhanced color image quantization utilizing hybrid glowworm swarm Optimization and K-means clustering: a case study with Chandrayaan-3 images. In: 2024 3rd International conference power, control computer technology, ICPC2T 2024, vol 2024. pp 433–438
19. Rathore V, Prajapat A, Ahuja K (2023) Meet the rocket women of Chandrayaan-3: pioneers in space exploration. In: 3rd IEEE international conference ICT business Indian Government, ICTBIG, pp 1–6
20. Lin Y et al (2024) Return to the moon: new perspectives on lunar exploration. Sci Bull 69(13):2136–2148. https://doi.org/10.1016/j.scib.2024.04.051
21. Dhingra D (2018) The new moon: major advances in lunar science enabled by compositional remote sensing from recent missions. Geoscience 8(12). https://doi.org/10.3390/geosciences8120498
22. Wang H, Guo Q, Li A, Liu G, Guo H, Huang J (2019) Impact of lunar terrain on moon-based earth observation. In: International geoscience remote sensing symposium, vol 2019-January, pp 9260–9262. https://doi.org/10.1109/IGARSS.2019.8900112
23. Megala S (2024) Chandrayaan-3 in the context of Indian lunar science programme. Αγαη 15(1):37–48
24. Singla JG, Sharma R, Shree A, Amitabh A, Dube N (2024) A novel deep learning approach for data analysis on high resolution Chandrayaan-2 data. In: 2024 IEEE Space, aerospace defence conference space, pp 415–418. https://doi.org/10.1109/SPACE63117.2024.10667960
25. Kaur H, Solanki R, Alankar B, Chang V (2023) India's journey to be part of the World's leading spacefaring countries: Chandrayaan-3 mission. In: Proceedings—2023 international conference Indian IoT, big data supply chain. IIoTBDSC 2023, pp 475–478. https://doi.org/10.1109/IIoTBDSC60298.2023.00087

26. Srivastava V (2023) Chandrayaan-3 lands on the Moon's south pole. Nat India, pp. d44151-023-00116–x, Aug 2023
27. "Chandrayaan-3." Available: https://pradan.issdc.gov.in/ch3/
28. "PDS: Tools." Available: https://pds.nasa.gov/tools/about/
29. "Navigation Camera." Available: https://pradan.issdc.gov.in/ch3/protected/browse.xhtml?id=navcam
30. "Rover Image." Available: https://pradan.issdc.gov.in/ch3/protected/browse.xhtml?id=ri
31. "Lover Image." Available: https://pradan.issdc.gov.in/ch3/protected/browse.xhtml?id=li
32. https://ncert.nic.in/pdf/chandrayan/english/Module8.pdf

Unveiling Triadic Patterns in Dynamic Networks Using a GAT-SAGE Hybrid Model



Vidyalekshmi Chandrika, D. Prashanth Chandra Reddy, T. H. Hrishab, Mallampati Pranathi, Sanjay J. K. Seshadrinath, and Lekshmi S. Nair

1 Introduction

A complex network is mathematically represented as a graph $G = (V, E)$, where V is the set of nodes or entities and E represents the relationships among the entities in the form of edges [3]. In a dynamic complex network, both the nodes and the edges evolve. This mimics the real-world relationships that exist in social or biological networks [5–7]. In a dynamic network, nodes or edges may appear, disappear, or change their weights based on the underlying interactions. Analyzing and capturing this continual process of evolution is computationally expensive. Moreover, the network changes happen in smaller groups rather than on the entire network. Therefore, capturing the triadic patterns paves an easy analysis to represent the evolution of such networks [4, 20, 22].

The triad formation begins with the dyadic structure, where two nodes are connected. At a later time t_1, a third node is connected to either one of the nodes or both, resulting in the formation of an open or closed triad [16]. The formation of triads helps in network growth and expansion. For example, in a social network such as LinkedIn, a friend recommendation algorithm uses the triad-based approach for recommending new connections [17, 23]. When the underlying network is temporal, analyzing the triad pattern formation involves consideration of *timestamps*. Various methods, such as similarity-based methods, dimensionality

V. Chandrika · D. Prashanth Chandra Reddy (✉) · T. H. Hrishab · M. Pranathi · S. J. K. Seshadrinath · L. S. Nair
Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri, India
e-mail: d.prashanth.chandra@gmail.com

V. Chandrika
e-mail: vidyalekshmiv@am.amrita.edu

L. S. Nair
e-mail: lekshmisn@am.amrita.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_11

reduction-based methods, and machine learning-based methods, are employed to identify methods to predict the formation of triads in subsequent timestamps of a temporal network [1, 2, 11, 12, 15, 21]. Similarity-based methods assume the network to be static, and hence it is inefficient in capturing temporal dependencies. Similarly, machine learning approaches also ignore time series dynamics, which makes them less effective in temporal networks.

Thus, there is an urge to propose a method that considers the time series dynamics while capturing the triads in a network. To address this, This paper propose a novel deep-learning model to predict the triads in a temporal graph. The proposed model predicts and enumerates the occurrences of triads by considering the temporal dependencies. The model combines the Graph Attention Network (GAT) [24] and the GraphSAGE [9] models. In addition to this, the core attention mechanism [8] of the GAT has been modified to give more importance to nodes which could potentially form triadic structures with other nodes. This highlights an important research gap, as research on predicting triadic structures accurately and efficiently is currently sparse. The model is trained and validated on the CollegeMSG [19] dataset.

The rest of the paper is structured as follows: Sect. 2 discusses the related works, Sect. 3 explains the methodology, Sect. 4 presents the results and discusses the future scope in Sect. 5.

2 Related Works

Temporal link prediction as well as analysis of triadic structures have gained importance and value due to their ability to provide large amount of information when it comes to community detection, community formation as well as social dynamics analysis.

Although work in the field involving triad prediction is rare, the field isn't devoid of using triads to perform analysis on temporal networks. Krzysztof Juszczyszyn et al. [14] created a system that monitors changes in social networks over time by analyzing how triads change over time. This was done by developing a triad transition matrix(TTM) using the Enron email dataset. This triad transition matrix records the probability of one type of triad transitioning into another type of triad over time. A triad transition graph was then used visualize how triads shift over time by showing how connections between nodes shift in a visual manner. This research helps further the research along in the field of social network analysis. However these studies focus on analyzing existing triads to understand how social dynamics change. It does not incorporate predicting the occurrences of triads and studying their impact on social networks over time.

Huang, et al. [10] created a system named MTNE which focuses on embedding temporal networks by focusing on the preservation of motifs which in this case is triads. The system uses hawkes process to model the evolution of triads over time which capture the dynamic nature of these triads within the overall network. The system also uses an attention mechanism to differentiate the importance of different

types of triads which allows the model to understand which types are important during the embedding process. The model was then evaluated on multiple real-world temporal datasets and the model showed relatively superior performance compared to multiple other models when it came to node classification, link prediction as well as link recommendation. However an identifiable limitation to this is the high computational cost involved in the process. However, the findings of the paper have large weight in the temporal network analysis field due to it providing a new and efficient method of embedding temporal networks.

Gou et al. [8] created a system which focuses on analyzing triads to map changes in social networks. By examining how triads transform over time, the proposed system aims to capture changing patterns of the networks core structure. The proposed system also uses the information about from analyzing triad transformations to introduce a link prediction system named as TLPA. This system predicts the occurrences of links forming in the future based on probabilities of triad transformations. The whole system was evaluated against existing systems such as ICMT, CRDNT and the results show that TLPA outperformed them when it came to higher message delivery rate while minimizing resource consumption. However the study also mentions a few limitations. The study notes the complexities of real world networks and the fact that just studying triads will not be sufficient to to capture all forms of rules which dictate changes in the networks structure. The study however does contribute to a large degree, the progression in the field of temporal link prediction by introducing a new way of performing link prediction by leveraging triad transformations.

Jain et al. [13] created a system named the Temporal Model for Opinion and Triad Network Formation (TMOTNF) to identify patterns surrounding opinion evolution within social networks, with the main goal of focusing on the role of stubborn agents which are individuals who maintain fixed opinions. The system accounts for triad formation by predicting the formation of a single link which will lead to a formation of a triad. To be precise, 2 connections already exist, the model predicts whether the third connection will form in order to form a triad. The study finds that stubborn agents play a large role when it comes to influencing the opinion dynamics of a network. If placed strategically, the can prevent opinion dissemination and prevent the the network from reaching a consensus which will lead to continued opinion diversity in the overall network. The study also notes that the model relies on highly specific assumptions as to how connections are formed in the network which might not capture all the complexities surrounding real world interaction. The study also notes that the influence of every stubborn agent is the same, however in a real world complex network, each stubborn agent may have varying levels of influence.

3 Proposed Methodology

This research aims to advance the triad prediction and enumeration process in the temporal networks. By triad prediction, we refer to predicting the occurrences of

triads at future points in time. Enumeration of triads plays a crucial role in purposes such as link prediction, influencer identification, and anomaly detection in a network. The deep learning models are utilized to learn the structural and temporal aspects of the network to predict and count the triads in a future snapshot of the temporal network. To validate the strength of the proposed model, the predicted triad counts are compared against the ground truth.

The GAT model's Attention is limited to 1—*hop* neighbors, which poses inefficiency for learning indirect triadic influences. On the other hand, the GraphSage model aggregates local neighbors. This paper proposes a hybrid GAT-SAGE model. This model combines the strengths of the GAT model along with the strengths of the GraphSAGE model to efficiently perform triad prediction. The model would efficiently predict triads in a temporal network by incorporating (i) GraphSAGE's sampling for scalability and (ii) GAT's attention mechanism for adaptive weighting for evolving triads. The hybrid model aims to predict the occurrence of both open and closed triads in the network. The model's predictions range from predicting 0 connections, which implies a stable triad, to a maximum of 3—*connections*, which leads to the formation of a new triad in an evolving network [18].

The input to the models is 15 overlapping temporal slices generated using the sliding window approach, where the current window is of size 30 with an overlapping window of size 13. The overlapping window single-handedly determines the number of temporal slices. An important modification to the core functionality of GAT is introduced by adjusting the attention mechanism to emphasize triadic structures. In simpler terms, a node is given higher importance if it may form triads with its neighbouring nodes. A hybrid GAT-SAGE model is created by using alternating *GATConv* and *SAGEConv* layers along with a fully connected neural network at the end to perform the prediction. The output of the fully connected network is the predicted triad count [25].

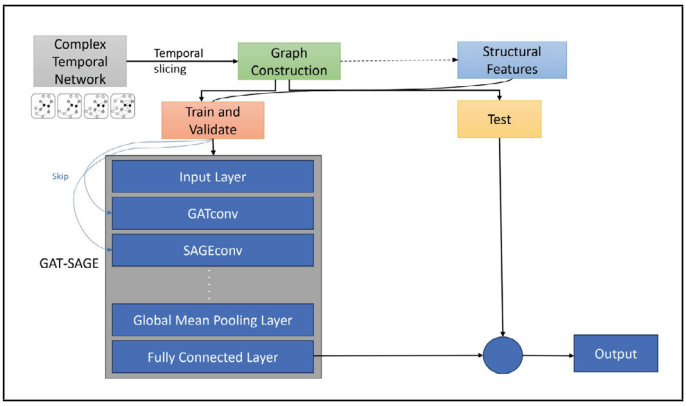


Fig. 1 An elaborate representation of the overall working and architecture of the system

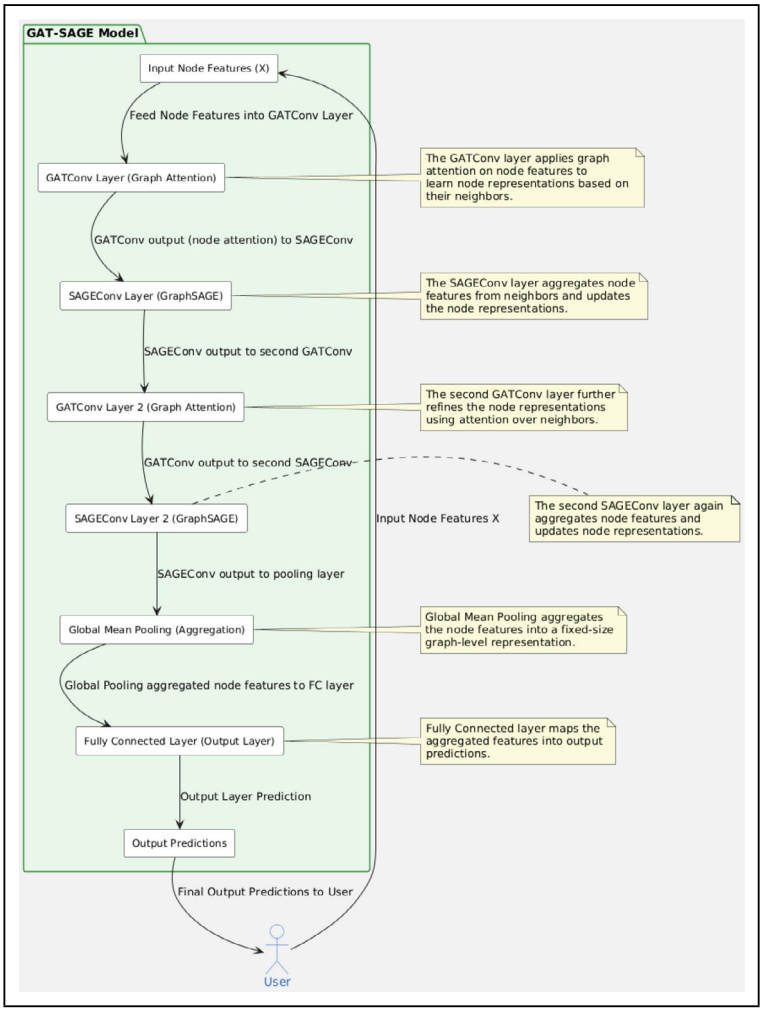


Fig. 2 A basic architecture diagram of the custom hybrid GAT-SAGE model

Figure 1 displays a representation of the overall working and architecture of the proposed system, while Fig. 2 represents a basic overview of the architecture and working of the custom `Hybrid GAT-SAGE` model. The model contains alternating GAT and SAGE layers with a fully connected layer at the end to perform the detection. The GAT layer tries to identify important neighbors for each node at a local level. The SAGE layer combines the findings of the GAT and updates each node's feature vector into an updated feature vector. The SAGEConv layer's main functionality is to aggregate the node vectors of the set of nodes that are in the local neighborhood for each node. The next two layers repeat the same task, further refining the process. In

addition, a global pooling layer aggregates all node representations at a global level. The globally pooled data are then fed into a fully connected layer, which outputs the number of triads for a specific time slice.

The attention mechanism in the GATConv layer aids the overall model in identifying important nodes that could contribute to the formation of triads. The output of this layer consists of refined node embeddings, which are fed into the next GraphSAGE layer. The modification to the core attention mechanism considers a triplet of nodes as opposed to a pair of nodes. The Triadic Attention Mechanism enhances node attention of edges for triad prediction. Source and destination indices are obtained from the edge list and a tensor is constructed for efficient computation. The node features are added to embed the relationships and a triad-effect tensor is created to store triadic influences. Contributions from nodes in the connected nodes are aggregated through scatter operations efficiently, and the model is able to track changing triads dynamically. In simple terms, if there are 3 nodes A, B and C, where A is connected to B and B is connected to C, the attention mechanism tries to identify how C affects the relationship between A and B. This information is added to the node feature vector that contains information about how important one node is to another node.

The model performs well even when there are rapid topological changes to the overall network across time slices. This is due to the combined ability of the GAT model which focuses on local structure and that of the GraphSage model which focuses on the overall structure of the graph to be able to predict changes in structure occurring across the overall structure. This inference can be validated by taking a look at the table in the results section. Initial time slices are filled with large amounts of triads indicating that there overall structure of the graph is dense but nearing the end, the number of triads decreases which indicates the graph becoming more sparse, however the model's predictions still maintain accuracy.

However, when it comes to thinking along the lines of reducing the complexity of the model without harming performance, it is important to state that all the existing components in the current state of the model are all required for the model to make accurate predictions. The GAT layer is required to understand importance between nodes at the local level, the GraphSage layer is required to scale the GAT's capabilities to the global level and the triad attention mechanism is required to make the GAT focus on sets of 3. So, the removal of any component would result in loss of performance. When it comes to finding alternative techniques which are less complex and perform similar to the current model or better, further experimentation needs to be done in order to obtain an answer to that.

The model does have a few limitations however. It is difficult to validate the model's performance on delayed connections or infrequent connections. This is due to the dense nature of the graph, making it difficult to identify which connections are delayed or are random. The model's predictions are also based on the current state of the graph making it completely inaccurate when trying to predict truly random connections or connections formed with behavioral attributes playing a role. This is because the model cannot perform well when there is context behind certain connections. For example, a citation network. Suppose A is a paper that has been

cited by 100's of other papers and B is a paper cited by 5 other paper's. A paper C is more likely to cite B than A if there is contextual similarity between them. However the model is unaware of this contextual similarity and will make a wrong prediction by predicting that a connection between C and A will form. Another limitation is that the model currently only works for dynamic networks where edges are temporal. This is because the GAT associates importance to nodes based on temporal aspects associated with the edges between itself and it's neighbors, for example, the persistence and frequency of connections between itself and another node will determine the level of importance it gives to that node. As a result, the current model is only capacble of making accurate predictions in temporal graphs where the edges are temporal.

4 Result Analysis

4.1 Dataset

This paper uses the College Message dataset that contains timestamped edges representing communication between students. The dataset contains 1899 nodes, 59,835 temporal edges, and 20,296 static edges, all over a timespan of 193 days. The data is of the format source node, destination node, and timestamp of the interaction between the source and destination node. The model has been trained and tested on the above dataset, the results of which have been shown both visually and statistically.

4.2 Experimental Setup

The model was trained for a total of 1000 epochs with an early stopping criteria in case the stress limit. For each epoch, the model is validated for the predictions. After the training process is complete, the model was tested on a test set and the results of which have been visualized in the results section. The entire dataset is split into overlapping temporal slices using a sliding window approach. After the slices are generated, the slices are then split into train-test validation at the ratio 70–20–10. After the model has been initialized, the model is then trained on the temporal slices, the training loss and validation loss for each epoch is noted. After the training process is done, the test loss is noted using a custom evaluation function. The results are visualized using MatPlotLib and SeaBorn to make attractive graphs which pack into it all the information required to make a sufficient analysis of the efficiency of the model. The results of the model's training and the model's general predictive capability have been shown below.

Figure 3 is a visual representation of the model's validation loss during the training process. These values are noted during each epoch of the model's training and have

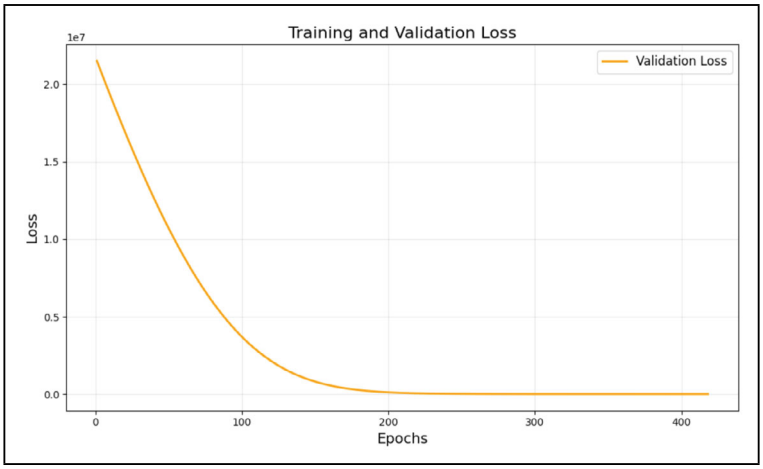


Fig. 3 A visual representation of the validation loss which has been noted during the training process

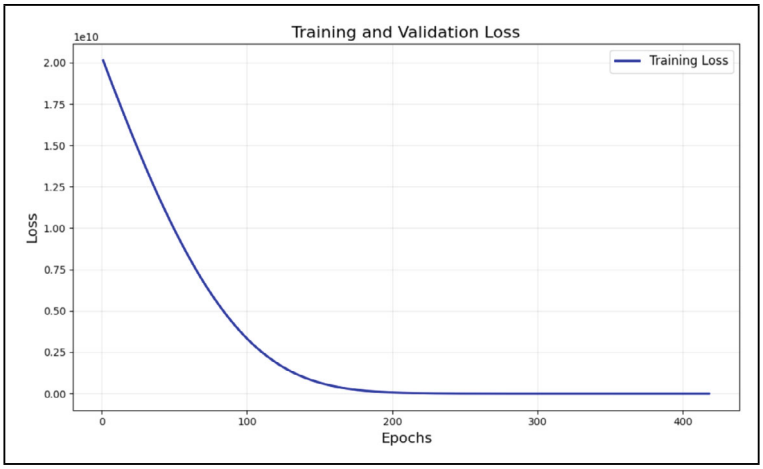


Fig. 4 A visual representation of the training loss which has been noted during the training process

been plotted to show how the loss decreases as the model trains more. This generally indicates that the model is learning how to predict the occurrences of triads as the epochs pass. The final validation loss value noted in the final epoch is 8588. Though the value is high by relative standards, the value indicates deviations in the model's predictions from the actual state of the graph. Each deviation is penalized to a large degree to help the model understand that the deviation in question is wrong, which is why the value is very high. Figure 4 is a visual representation of the model validation loss during the training process. These values are noted during each epoch of the

Table 1 Statistical representation of the predicted and ground truth open and closed triads for each slice

Slice	Pred. closed triads	Ground truth closed triads	Pred. open triads	Ground truth open triads
1	2523	2222	360,623	356,684
2	5659	5843	771,362	763,266
3	3220	2798	462,010	456,950
4	454	234	72,223	71,370
5	235	128	37,062	36,626
6	233	147	35,924	35,508
7	88	67	13,153	13,004
8	106	88	15,551	15,378
9	98	75	14,607	14,442
10	74	41	11,812	11,672
11	58	26	9418	9304
12	19	7	3208	3168
13	0	0	347	560

model's training and have been plotted to show how the loss decreases as the model trains more. The final training loss value noted in the final epoch is 52,095. This figure can be obtained by using the matplotlib library, where the root mean square error values are calculated and plotted into a graph. The actual values are calculated using the $O(n^3)$ algo and then compared it with the predicted values

Table 1 shows the model's capabilities to predict the occurrences of triads at future points in time. The values show the model's capabilities when it comes to predicting the number of open and closed triads within each time slice. The model shows greater accuracy when it comes to predicting open triads, which represent connections between 3 nodes where only 2 edges are involved. It is imperative to note that the temporal representation of the graph is that of a directed graph as opposed to an undirected graph. These results, which are obtained by using the root mean square values between the predicted triad counts and the actual triad counts in each temporal slice is the main source of validation for the model's performance.

5 Conclusion

The model demonstrates a uniform decrease in loss over the training period, starting from high initial values and converging to lower values in its respective training and validation graphs. For triad count predictions, the Actual vs. Predicted graphs provide a clear picture of the performance of the model. The custom GAT-SAGE hybrid achieves accurate predictions for every temporal slice, with predicted triad counts

almost closely matching actual values, with the scale of disparities being less. This shows that the model is capable of predicting the state of the graph quite accurately with few disparities for future time slices. The future scope involves including behavioral attributes in addition to structural and temporal attributes of the network for the triad prediction. This will be significantly helpful in improving the accuracy of the model's predictions, but also may open the path of node occurrence predictions in temporal networks, which will significantly boost the accuracy of models trying to perform community formation analysis as well as social dynamics analysis. By adding hierarchical attention mechanisms that recognize structures like communities or cliques, the model can be improved to identify larger patterns. This article mainly focused on predicting the occurrences of triads in temporal graphs. Investigating triad prediction can greatly speed up research in the area of temporal network analysis, as it has applications in community detection, community formation analysis, as well as social dynamics analysis.

References

1. Aadhithya A, Kumar S, Rejathalal V, Sujadevi VG, Poornachandran P, Soman KP (2023) Finding network motifs: a comparative study between ILP and symmetric rank-one NMF. In: 2023 3rd international conference on intelligent technologies (CONIT). IEEE, pp 1–10. https://doi.org/10.1109/CONIT59222.2023.10205573
2. Arrar D, Kamel N, Lakhfif A (2024) A comprehensive survey of link prediction methods. J Supercomput 80(3):3902–3942
3. Arul SM, Senthil G, Jayasudha S, Alkhayyat A, Azam K, Elangovan R (2023) Graph theory and algorithms for network analysis. In: E3S web of conferences, vol 399:08002. EDP Sciences
4. Bianconi G, Darst RK, Iacovacci J, Fortunato S (2014) Triadic closure as a basic generating mechanism of communities in complex networks. Phys Rev E 90(4):042806
5. Cheriyan J, Sajeev GP (2020) m-PageRank: a novel centrality measure for multilayer networks. Adv Complex Syst 23(05):2050012
6. Christensen C, Albert R (2007) Using graph concepts to understand the organization of complex systems. Int J Bifurcation Chaos 17(07):2201–2214
7. Costa LF, Rodrigues FA, Cristino AS (2008) Complex networks: the key to systems biology. Genet Molecular Biol 31:591–601
8. Gou F, Jia W (2022) Triad link prediction method based on the evolutionary analysis with IoT in opportunistic social networks. Comput Commun 181:143–155
9. Hamilton W, Ying Z, Leskovec J (2017) Inductive representation learning on large graphs. In: Advances in neural information processing systems, pp 1024–1034. https://arxiv.org/abs/1706.02216
10. Huang H, Fang Z, Wang X, Miao Y, Jin H (2020) Motif-preserving temporal network embedding. In: IJCAI, pp 1237–1243
11. Huang H, Jie Tang L, Liu JDL, Xiaoming F (2015) Triadic closure pattern analysis and prediction in social networks. IEEE Trans Knowl Data Eng 27(12):3374–3389
12. Huang H, Tang J, Wu S, Liu L, Fu X (2014) Mining triadic closure patterns in social networks. In: Proceedings of the 23rd international conference on World wide web, pp 499–504
13. Jain E, Singh A, Hatano N (2021) Opinion evolution in society with stubborn agents using temporal model for opinion and triad network formation (tmotnf). J Complex Netw 9(5):1–27
14. Juszczyszyn K, Budka M, Musial K (2011) The dynamic structural patterns of social networks based on triad transitions. In: 2011 international conference on advances in social networks analysis and mining. IEEE, pp 581–586

15. Kumar M, Maya, Ramya GR (2021) Performance comparison of anomaly detection algorithms. In: Inventive communication and computational technologies: proceedings of ICICCT 2021. Springer, Singapore, pp 761–776
16. Nair LS, Cheriyan J, Swaminathan J (2022) Microscopic structural analysis of complex networks: an empirical study using motifs. IEEE Access 10:33220–33229
17. Nair LS, Swaminathan J (2025) Predicting future Intrablock links in directed networks using triadic patterns. IEEE Access
18. Oviya IR, Shanmukha Sravya N, Raja K (2024) Predicting protein-protein interactions using self-attention-based deep neural networks and FastText embeddings. In: 2024 15th international conference on computing communication and networking technologies (ICCCNT). IEEE, pp 1–6. https://doi.org/10.1109/ICCCNT61001.2024.10725821
19. Panzarasa P, Opsahl T, Carley KM (2009) Patterns and dynamics of users' behavior and interaction: network analysis of an online community. J Am Soc Inform Sci Technol 60(5):911–932
20. Qiao Z, Li W, Li Y (2022) Temporal network embedding with motif structural features. In: International conference on database systems for advanced applications. Springer, pp 665–681
21. Qiu Z, Jia W, Wenbin H, Bo D, Yuan G, Yu PS (2021) Temporal link prediction with motifs for social networks. IEEE Trans Knowl Data Eng 35(3):3145–3158
22. Subhakkrith, Rithuh S, Jagadish Kumar Y, Harish R, Vishnu Dheeraj B, Vidhya S (2024) Graph-based approaches to detect network anomalies and to predict attack spread. In: 2024 15th international conference on computing communication and networking technologies (ICCCNT). IEEE, pp 1–11. https://doi.org/10.1109/ICCCNT61001.2024.10725025
23. Shi J, Chaurasiya V, Liu Y, Vij S, Wu Y, Kanduri S, Shah N et al (2023) Embedding based retrieval in friend recommendation. In: Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval, pp 3330–3334
24. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y (2017) Graph attention networks. arXiv preprint arXiv:1710.10903
25. Granovetter MS (1973) The strength of weak ties. Am J Sociol 78(6):1360–1380

Study of Stomach Abnormalities Using EGG Signals



R. Chandrasekaran, B. Praveena, Vivek Pazhamalai, S. Ivo Romauld, S. Vijayaraj, and M. K. Soundarya

1 Introduction

The Electrogastrogram (EGG) is a non-invasive method of recording stomach electrical activity and thus gives information on gastric motility and function. Like the ECG for the heart, the EGG records slow wave activity from the stomach wall smooth muscles during digestion at approximately three cycles per minute for humans. Surface electrodes are placed on the abdomen and the EGG registers low frequency signals that are related to gastric slow waves and spike activity that is superimposed on these, which is increased during active digestion. These signals can be analyzed in terms of frequency, amplitude and stability in order to determine patterns which may be associated with a number of gastrointestinal disorders, including gastroparesis and functional dyspepsia. However, as with the EGG, the major problem is one of signal isolation, since gastric signals are contaminated by respiratory and cardiac noise.

R. Chandrasekaran · B. Praveena (✉) · V. Pazhamalai · S. Ivo Romauld · S. Vijayaraj · M. K. Soundarya
Vels Institute of Science, Technology and Advanced Studies, Chennai, Tamil Nadu, India
e-mail: praveenab1624@gmail.com

R. Chandrasekaran
e-mail: chandrasekar.se@vistas.ac.in

V. Pazhamalai
e-mail: vivek.se@vistas.ac.in

S. Ivo Romauld
e-mail: ivoromauld.se@vistas.ac.in

S. Vijayaraj
e-mail: vijayaraj.se@vistas.ac.in

M. K. Soundarya
e-mail: soundarya.se@vistas.ac.in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_12

Nevertheless, progress in signal processing has enhanced the clinical significance of EGG and it has the potential to help in the diagnosis of gastric motility disorders [1].

Electrogastrogram (EGG) acquisition records the stomach's electrical patterns through surface electrodes to evaluate its movement and rhythmic behavior. EGG differs from invasive methods because it is a non-invasive technique that records slow-wave signals generated by gastric smooth muscle cells. Human stomach slow waves operate at three cycles per minute and establish the fundamental electrical pattern which helps evaluate gastric motility disorders. EGG recording starts with positioning surface electrodes on the abdominal skin situated above the stomach to capture gastric signals effectively. The EGG signal measures in microvolts with low amplitude and faces interference from respiratory and cardiac signals which makes precise electrode placement essential. One electrode of a bipolar or unipolar pair is positioned over the gastric region while the other functions as a reference point. The electrode configuration effectively separates gastric electrical signals from background noise [2].

2 Literature Survey

The Literature review is done to identify the research gap in existing methodologies of diagnostic tools available for diagnosing gastroparesis. The various existing methodologies and proposed research methodologies are discussed in this chapter.

Amal Shine et al. [3], the paper discusses the clinical importance of electrogastrogram (EGG) and its correlation with mucosal electrograms (mEGs), gastric emptying tests (GETs), and gastrointestinal symptoms before and after temporary gastric electrical stimulation (tGES). EGG is valuable in diagnosing delayed gastric emptying and can identify patients who may require GES.

Bin Wang et al., "Journal of Medical Imaging and Health Informatics", [4]. The paper discusses the use of electrogastrogram (EGG) to detect gastric electromyogram activity and classify high-dimensional signals for the diagnosis and management of gastrointestinal diseases.

Al Taee and Al-Jumaily [5]. "Electrogastrogram Based Medical Applications an Overview and Processing Frame Work., this paper discusses Electrogastrogram (EGG), a noninvasive method of recording electrical signals of stomach muscles. It provides an overview of EGG detection, analyzing methods, and its application in clinical settings.

Yin et al. [6]. Electrogastrography: Methodology, Validation and Applications. The paper discusses the methodology, validation, and applications of electrogastrography (EGG), a non-invasive method for measuring gastric myoelectrical activity. It reviews the accuracy of EGG in recording gastric slow waves and dysrhythmia and discusses its correlation with gastric contraction. The paper also reviews the applications of EGG in various areas.

3 Methodology

This system aims to develop algorithms for mathematical models of intestinal physiology and digestive processes for the purpose of gaining insight into and analysing digestive disorders. The data gathered will identify the mathematical models that serve best, estimate parameters, and ultimately validate the model. Simulations will be run to demonstrate how digestion is impacted by variations in relevant conditions, and sensitive analyses will be performed to identify influential factors in digestion. The ultimate goal is to design practical utensils that the doctor can use in organizing and managing digestive disorders. Figure 1 shows the block diagram of the system.

3.1 Filtering

Filtering is an important step in the analysis of EGGs to ensure noise-free data for the accurate interpretation of gastric electrical activity. Common sources of interference include power line noise, motion artifacts, and baseline drift. Bandpass filtering at 0.05–0.5 Hz isolates gastric slow waves, while adaptive filtering dynamically adjusts to non-stationary noise [7]. Wavelet transform filtering enables the time–frequency analysis of the artifact. Signal averaging is one more post-processing technique that improves the clarity of a signal. Such techniques enhance the signal-to-noise ratio, and gastric rhythms could be detected in a reliable way. This enhances the diagnosis of motility disorders such as gastroparesis in clinical as well as in research settings [8].

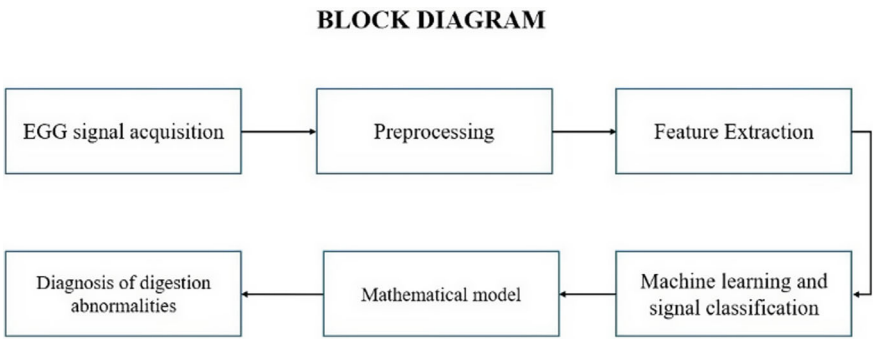


Fig. 1 Block diagram

3.2 Power Spectral Density (PSD)

The PSD of the EGG is an important tool in analysing the frequency characteristics of gastric electrical activity [9]. Dominant frequencies in EGG signals are generally seen around 0.05 Hz, or 3 cycles per minute, corresponding to normal slow-wave activity. PSD estimation through methods such as FFT and Welch's method is useful for power distribution, noise reduction, and detection of dysrhythmias like bradygastria and tachygastria. In the clinical setting, PSD helps to diagnose motility disorders such as gastroparesis by the presence of frequency shifts or power reductions. This is a non-invasive tool that improves gastric function assessment and treatment efficacy [10].

3.3 Cross Spectral Density (CSD)

The cross-spectral density (CSD) of an electrogastrogram (EGG) analyzes phase relationships and synchronization between gastric electrical signals from different stomach regions. It is useful for determining common power across frequencies (0.01–0.1 Hz) that could be involved in such an assessment of gastric motility and coordination [11]. High coherence in the slow-wave range provides evidence of effective peristalsis, and low coherence or phase irregularities may suggest dysrhythmic conditions such as bradygastria or tachygastria. The use of Fourier-based methods, like FFT and Welch's method, makes it possible to estimate CSD with an accurate frequency-domain analysis. It also supports the non-invasive assessment of gastric function, helping in the diagnosis and management of motility disorders [12].

3.4 Feature Extraction

Features that are extracted from the analysis of electrogastrogram are critical for the interpretation of gastric function. Some of the time series features are mean, median, standard deviation, peak-to-peak amplitude, and kurtosis. These helps understand the signal characteristics. The mean and median can be used to measure the underlying activity and the underlying preference. The standard deviation measures the variance. Peak-to-peak amplitude indicates the strength of the signal, while kurtosis describes the variability associated with the variance. It applies the fast fourier transform (FFT) or wavelet transform to separate slow-frequency (0.03–0.1 Hz) signals of the gastrointestinal tract. These resources enhance the interpretation of EGG signals, facilitate the diagnosis of movement disorders such as gastroparesis, and support the use of medical imaging and machine learning [13, 14].

3.5 Fast Fourier Transform (FFT)

Features that are extracted from the analysis of electrogastrogram are critical for the interpretation of gastric function. Some of the time series features are mean, median, standard deviation, peak-to-peak amplitude, and kurtosis. These helps understand the signal characteristics. The mean and median can be used to measure the underlying activity and the underlying preference. The standard deviation measures the variance. Peak-to-peak amplitude indicates the strength of the signal, while kurtosis describes the variability associated with the variance. It applies the fast Fourier transform (FFT) or wavelet transform to separate slow-frequency (0.03–0.1 Hz) signals of the gastrointestinal tract. These resources enhance the interpretation of EGG signals, facilitate the diagnosis of movement disorders such as gastroparesis, and support the use of medical imaging and machine learning [15, 16].

4 Results and Discussion

A Chebyshev Type I IIR Bandpass Filter allows certain frequencies in a particular range and rejects others. This filter is also efficient for the applications such as EEG signal processing and communication systems. This low cutoff of 0.15 Hz and a high cutoff of 0.60 Hz filter is designed which shown in Fig. 2.

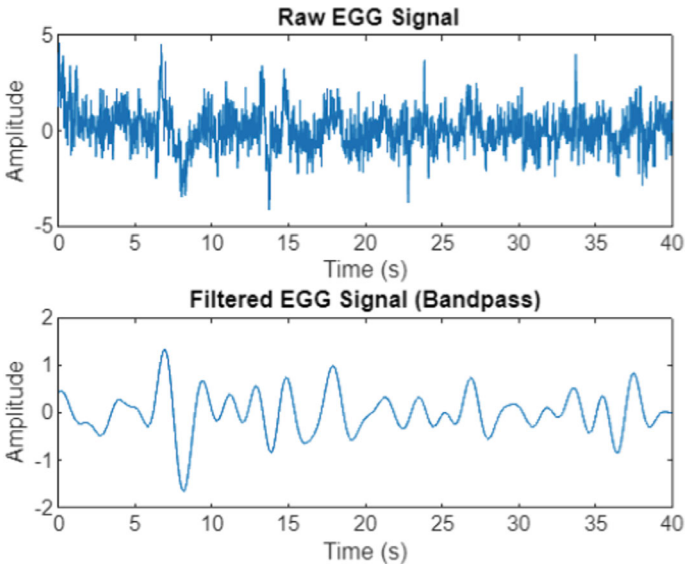


Fig. 2 Filtering

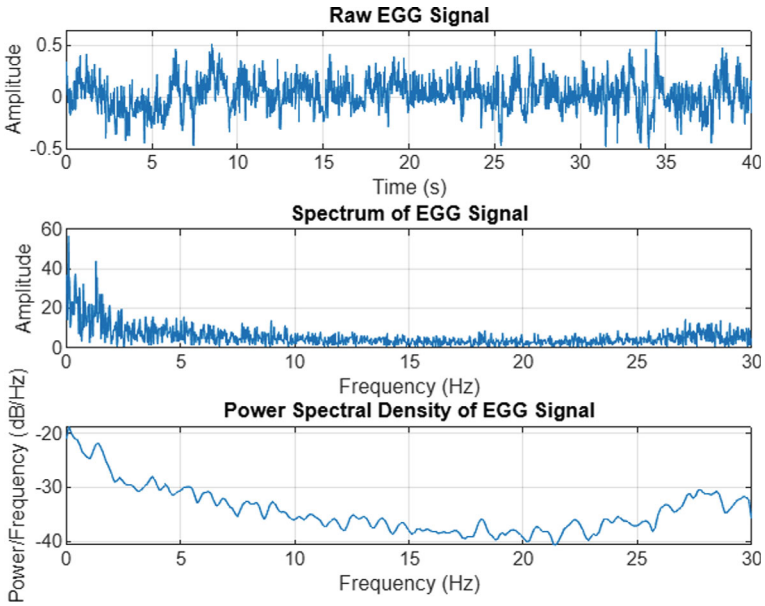


Fig. 3 Power spectral density (PSD)

Power spectral density (PSD) analysis is often utilized to quantify the power distribution of the signal across a range of frequencies and investigate the frequency composition which illustrated in Fig. 3. PSD is utilized to characterize the frequency character of the EGG signal, which then allows one to analyze the gastric slow wave, identify pathological patterns, and derive features of clinical use.

Cross-spectral density (CSD) of electrogastrogram (EGG) signals analyzes electrode correlations and detects gastric muscle synchronization shown in Fig. 4. CSD enables the analysis of regional coordination, rhythmic pattern detection, and the diagnosis of motility disorders. CSD, through analysis of frequency relations, offers insight into stomach function and enables one to interpret normal and abnormal gastric activity for clinical purposes.

Mean is the average EGG signal, indicating gastric activity in the middle. Median is the middle value and is not affected by outliers. Standard deviation is a measure of variability around the mean. Peak-to-peak measures the signal range. Kurtosis measures sharpness of distribution, detecting outliers. These measurements, together, characterize gastric electrical activity, making it possible to analyze the EGG signal and determine motility. Figure 5 shows the feature extraction.

Normal Digestion:

A normal stomach's FFT peaks at 2–4 cycles per minute (cpm), indicating regular gastric slow waves and normal contractions for digestion. The power spectrum is stable with minimal irregularities.

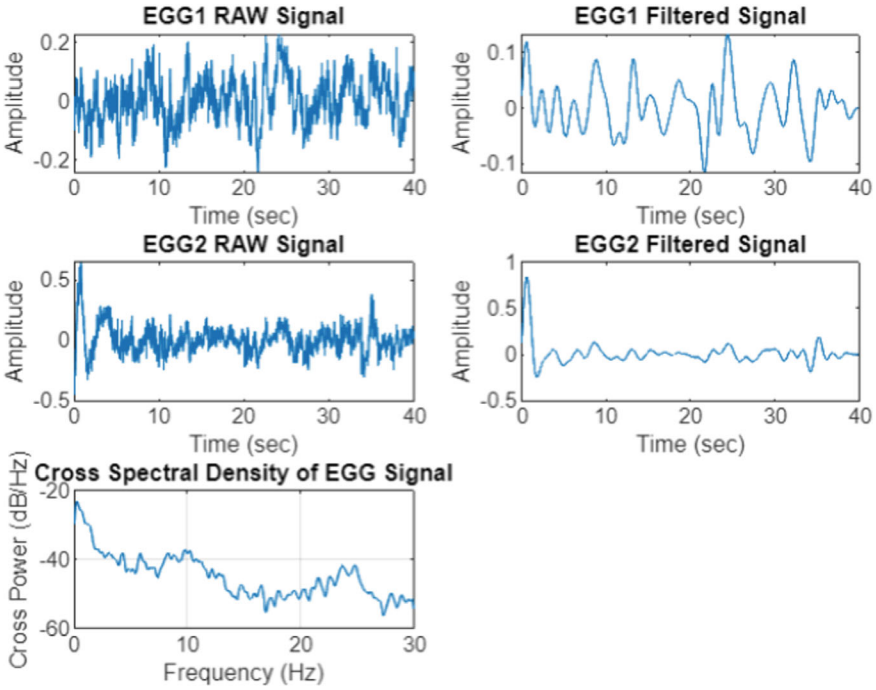
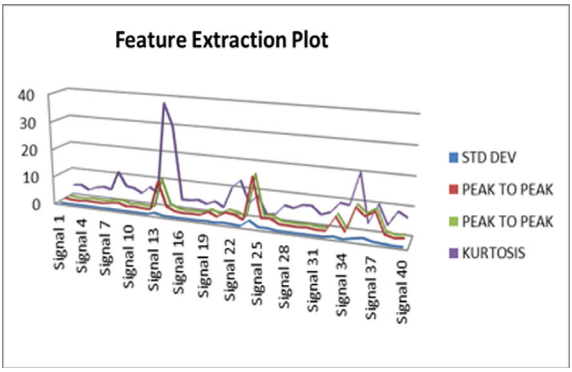


Fig. 4 Cross spectral density (CSD)

Fig. 5 Feature extraction



Abnormal Digestion:

Gastrointestinal diseases like gastroparesis and dyspepsia cause alterations in the FFT spectrum:

- Bradygastria: Frequencies below 2 cpm, indicating slow gastric motility.

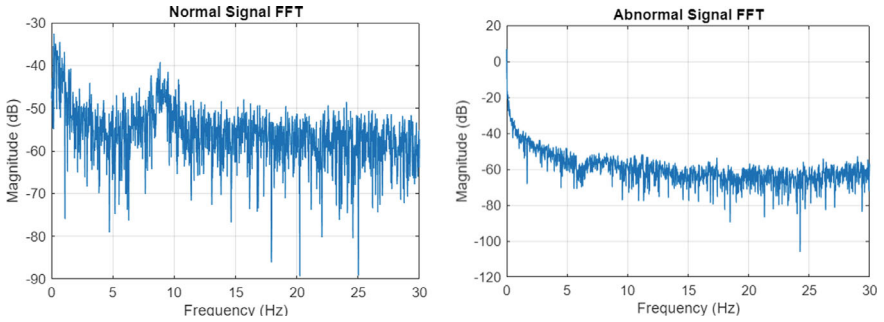


Fig. 6 Normal FFT signal versus abnormal FFT signal

- Tachygastria: Frequencies above 4 cpm, a marker of rapid or irregular contractions.
- Arrhythmias: Failure of dominant peak, wide frequency spread, or noise, a marker of deranged gastric electrical activity.

Comparing normal VS abnormal plot of EGG signals are represented in Fig. 6 is clearly shows that there is a delayed in digestion for subjects who regularly intakes junkfood. The fasting state spectral distribution was even for all individuals at all categories but the postprandial state shows delayed gastric emptying process.

5 Conclusion

The research on stomach abnormalities with the help of Electrogastrogram (EGG) signals provides tremendous information regarding non-invasive gastric motility and associated digestive disorders assessment. Through the acquisition and analysis of EGG signals from normal subjects under well-controlled fasting and postprandial states, this work effectively presents the influence of dietary patterns especially consumption of junk foods on gastric function. Signal processing methods including filtering, Power Spectral Density (PSD), Cross Spectral Density (CSD), and Fast Fourier Transform (FFT) have successfully been utilized to yield meaningful features from the low-frequency gastric signals and uncover patterns that are representative of both normal and abnormal digestion. The findings significantly show that subjects with high junk food consumption show evidence of delayed gastric emptying, as indicated by the irregular and longer frequency distribution in postprandial EGG signals. The implementation of MATLAB for analysis and BIOPAC MP45 for signal acquisition provided accurate and consistent data manipulation. Additionally, the time and frequency-domain approaches implemented in the study enable identification of conditions such as bradygastria and tachygastria. In summary, this study highlights the diagnostic value of EGG as an early detection and monitoring tool for gastric motility disorders. It also underscores the role of diet in maintaining gastrointestinal

well-being. The methods employed herein can be a basis for the establishment of sophisticated diagnostic systems and can aid clinicians' decision-making for the management of digestive disorders.

References

1. Chang F (2005) Electrogastrography: basic knowledge, recording, processing and its clinical applications. J Gastroenterol Hepatol 20. https://doi.org/10.1111/j.1440-1746.2004.03751.x
2. J Yin J Chen 2013 Electrogastrography: methodology, validation and applications J Neurogastroenterol Motility 19 5 17 https://doi.org/10.5056/jnm.2013.19.1.5
3. Shine A, Mathur P, Ahmed S, Ramos S, McElmurray L, Stocker A, Pinkston C, Abell TL (2022) Low-Resolution electrogastrogram at baseline and response to temporary gastric electrical stimulation—a comparison of cutaneous with mucosal recordings. Neuromodulation Technol Neural Interface 25(8):1150–1159
4. B Wang T Wen J Hou L Lin YuDu Yan Fang T Pan 2020 The high-dimensional signal classification of electrogastrogram for detection of gastric motility disorders J Med Imaging Health Informatics 10 6 1281 1287
5. Al Taee A, Al-Jumaily A (2020) Electrogastrogram based medical applications an overview and processing frame work. In: Hybrid ıntelligent systems: 18th ınternational conference on hybrid ıntelligent systems (HIS 2018) Held in Porto, Portugal, 13–15 Dec 2018, vol 18. Springer International Publishing, pp 511–520
6. Yin J, Chen JDZ (2013) Electrogastrography: methodology, validation and applications. J Neurogastroenterol Motility 19(1):5
7. J Chen 1992 A computerized data analysis system for electrogastrogram Comput Biol Med 22 1–2 45 57 https://doi.org/10.1016/0010-4825(92)90051-N
8. Gharibans A, Smarr B, Kunkel D, Kriegsfeld L, Mousa H, Coleman T (2018) Artifact rejection methodology enables continuous, noninvasive measurement of gastric myoelectric activity in ambulatory subjects. Sci Rep 8. https://doi.org/10.1038/s41598-018-23302-9
9. Parkman H, Hasler W, Barnett J, Eaker E (2003) Electrogastrography: a document prepared by the gastric section of the American motility society clinical GI motility testing task force. Neurogastroenterol Motility 15. https://doi.org/10.1046/j.1365-2982.2003.00396.x
10. M Taghipoor P Lescoat J Licois C Georgelin G Barles 2011 Mathematical modeling of transport and degradation of feedstuffs in the small intestine J Theor Biol 294 114 121 https://doi.org/10.1016/j.jtbi.2011.10.024
11. C Li W Yu P Wu X Chen 2020 Current in vitro digestion systems for understanding food digestion in human upper gastrointestinal tract Trends Food Sci Technol 96 114 126 https://doi.org/10.1016/j.tifs.2019.12.015
12. D Komorowski B Mika 2018 A new approach for denoising multichannel electrogastrographic signals Biomed Signal Process Control 45 213 224 https://doi.org/10.1016/j.bspc.2018.05.041
13. A Kafee Y Kayar 2023 Electrogastrography in patients with gastric motility disorders Proc Inst Mech Eng 238 22 32 https://doi.org/10.1177/09544119231212269
14. J Chen R McCallum 1991 Electrogastrography: measuremnt, analysis and prospective applications Med Biol Eng Comput 29 339 350 https://doi.org/10.1007/BF02441653
15. M Yacin M Manivannan S Chakravarthy 2010 On non-invasive measurement of gastric motility from finger photoplethysmographic signal Ann Biomed Eng 38 3744 3755 https://doi.org/10.1007/s10439-010-0113-4
16. Wang Z, Elsenbruch S, Orr W, Chen J (2003) Detection of gastric slow wave uncoupling from multi-channel electrogastrogram: validations and applications. Neurogastroenterol Motility 15. https://doi.org/10.1046/j.1365-2982.2003.00430.x

FPGA-Based Approximate Multipliers for EdgeAI: A Systematic Review



Nguyen Thanh Loc, Nguyen Thien An, and Cuong Pham-Quoc

1 Introduction

High performance computing is becoming increasingly important and is closely related to many tasks related to processing a large amount of big data such as digital signal processing, image processing, multimedia and communication systems, or machine learning systems. In addition, one of the main issues that need to be concerned today besides optimizing speed performance is optimizing resources such as energy consumption. Numerous methods have been proposed to address these demands, with the goal of maximizing processing speed, area resources, and energy consumption while maintaining the application's requisite performance. One such new technique is approximate computing [6, 7], which, by imposing penalties on the correctness of the system, allows explicit control over the energy and hardware resource usage for the intended application. Approximate computing is based on the idea that acceptable system performance does not necessarily need application systems to provide the most accurate results [4].

One of the fields where this approximate multiplication technique can be applied and with promising results is artificial intelligence applications. The final calculation result of an artificial intelligence neural network after going through a calculation process that involves many linear calculations will eventually be limited to a specific range and have a certain error limit. Furthermore, because all of these values are dependent on probability, the final calculation results will vary as long as the value rankings of the categories do not undergo major changes. Therefore, approximate

N. T. Loc · N. T. An · C. Pham-Quoc
Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam
e-mail: loknguyen@hcmut.edu.vn

N. T. An
e-mail: ngthienan.cse@hcmut.edu.vn

Vietnam National University—Ho Chi Minh City (VNU-HCM), Ho Chi Minh City, Vietnam
e-mail: cuongpham@hcmut.edu.vn

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_13

computing can be used in neural network computation for better performance and energy efficiency. In the literature, many approximate arithmetic units are designed specific for neural network computation [1–3].

Furthermore, to deal with the problem of power consumption and computing resource restrictions, the system needs a low-complexity setup and quick calculation. Therefore, a viable way to decrease the computation time, energy consumption, and complexity of AI models when deployed to edge devices without appreciably lowering the model's accuracy is to use approximation computing to Edge AI platform-based applications.

Meanwhile, Field Programmable Gate Array (FPGA) technology is well known for improving the efficiency of computationally demanding activities in a variety of application fields, including AI-powered applications. The flexible nature of the field-programmable gate array (FPGA) platform makes it a popular choice for accelerating neural network computing. On the one hand, modern FPGA devices have a sufficient logic density to offer acceptable performance for neural network computing. However, the architecture is easily adaptable to new neural network models since FPGA is programmable.

The main contributions of this systematic review are threefold. First of all, we conduct a systematic review of FPGA-based approximate multiplier techniques from recent papers published in international journals. In addition, we compare the results of these papers with each other based on the criteria: latency, resources usage, energy consumption and give some observations about these results. Finally, we present some open issues for FPGA-based AI powered application using approximate multiplier technique.

The rest of the paper is organized as follows. Section 2 discuss the background of approximate multiplier algorithms. Section 3 presents the survey of approximate multiplier implementation for FPGA-based in the literature. We summarize open issues of applying approximate multiplier technique in AI application in Sect. 4. Finally, we conclude our paper in Sect. 5.

2 Preliminaries

2.1 Multiplication Algorithm

In digital electronics, multiplication is one of the most fundamental and significant processes due to its complexity and range of applications. There are several ways to implement multiplication in hardware, but the majority of them include calculating partial products and utilizing addition to add them together. Developed by Anatoly Karatsuba in 1960 and published in 1962, the Karatsuba algorithm is one of the quick and popular techniques in the articles covered here [5, 8, 9].

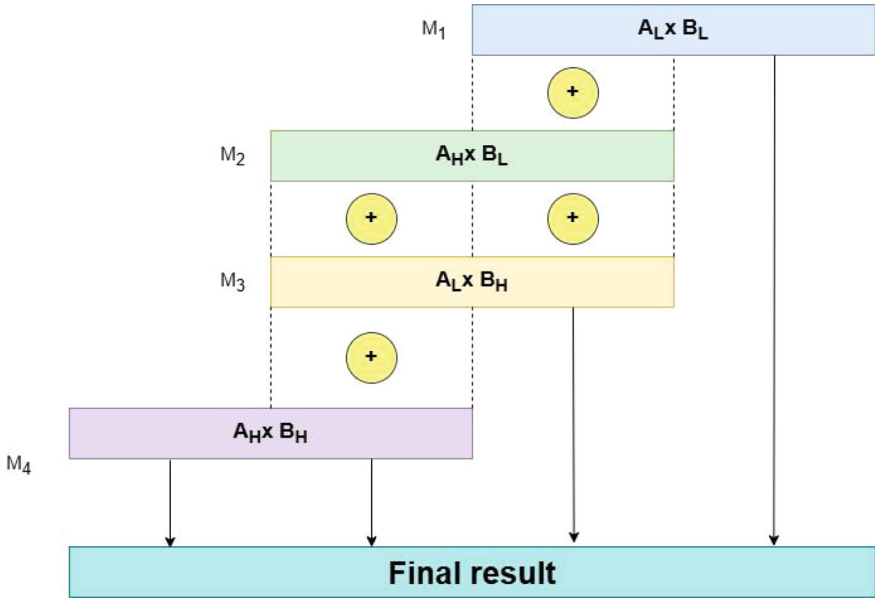


Fig. 1 Final products of Katsubara algorithm for binary multiplication

When applying this algorithm for binary numbers, for two operands A and B of $2n$-bit, we decompose them into two separate parts, the higher order parts A_H and B_H, and the lower order parts A_L and B_L:

$$A = A_H \times 2^n + A_L$$
$$B = B_H \times 2^n + B_L$$

Therefore, the multiplication of these operands can be presented as:

$$A \times B = A_H B_H \times 2^n + (A_H B_L + A_L B_H) \times 2^n + A_L B_L$$

Then, the middle term $A_H B_L + A_L B_H$ can be continually modified as

$$A_H B_L + A_L B_H = A_H B_H + A_L B_L - (A_H - A_L)(B_H - B_L)$$

Finally, the final products can be generated by adding these terms together by orders as Fig. 1.

This approach allows us to optimize space and latency without significantly altering the computational results by approximating partial products in lower order or middle sections when used to approximation kernel algorithms.

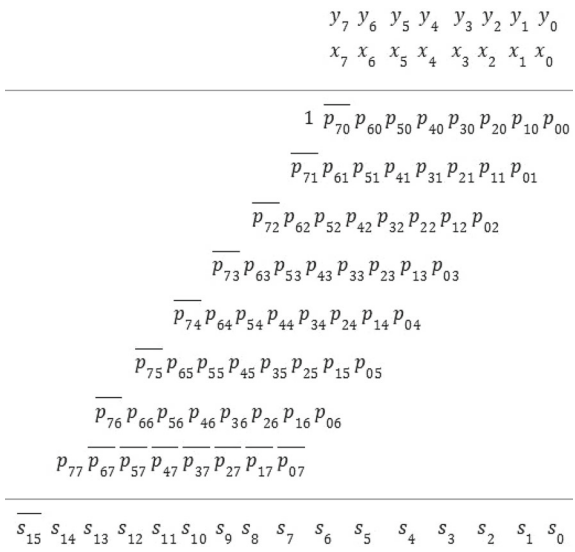


Fig. 2 Illustration of an 8-bit Baugh-Wooley multiplication in [11]

2.2 Baugh Wooley Algorithm

Baugh and Wooley is the calculation that shows multiplication for two's supplement bits. The sign bits of every multiplicand and the multiplier's unsigned or positive sign are where this multiplier converges. This computation is entirely prepared using the standard reasoning full adders. Figure 2 illustrates a 8-bit Baugh-Wooley multiplication algorithm.

The distinctive aspect of executing the Baugh Wooley algorithm on FPGA is the capability to utilize LUT 5, LUT 5-2, or LUT 6, LUT 6-2, which are prevalent and accessible on most contemporary FPGA platforms, to construct the three primary blocks M, L, and J, Fig. 3, for the algorithm. This approach is ideal in both area and latency compared to simpler algorithms like array multiplication or the shift method.

2.3 Radix Booth Algorithm

Among the effective methods for implementing multiplication on computer hardware are Radix Booth-4 and Booth-8. By edcoding the multiplier, this approach essentially applies and improves upon Booth's algorithm by reducing the amount of addition and subtraction operations needed for multiplication.

How this approach encodes the multiplier while performing multiplication is seen in the two Tables 1 and 2. The length of the multiplier bits that Radix Booth-4 and Booth-8 encode is clearly different; in Booth-4, it is three, but in Booth-8, it is four. Radix Booth-8 will therefore generate fewer partial products, leading to speedier

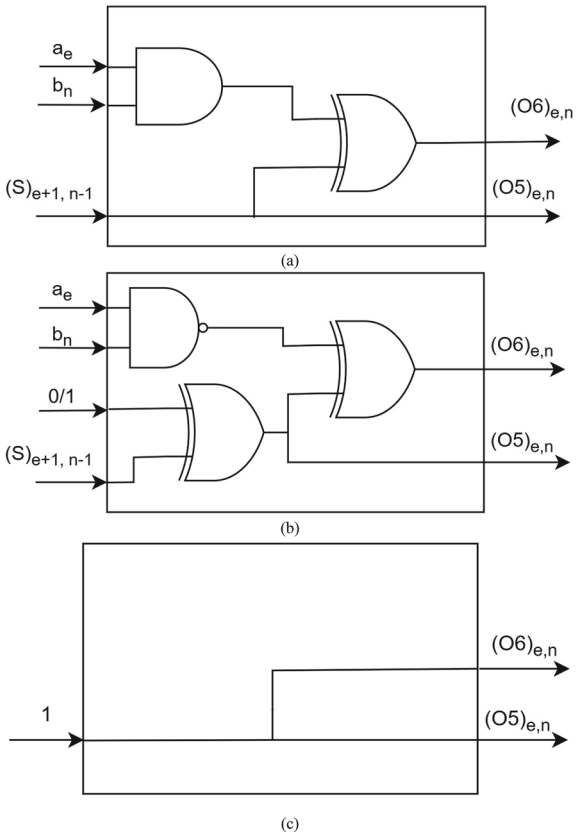


Fig. 3 LUT configuration for BW circuit with (**a**) L (**b**) M and (**c**) J in [10]

Table 1 Radix-4 booth recoding

Multiplier bits	Recoded operation on multiplicand, X
000	0X
001	+X
010	+X
011	+2X
100	−2X
101	−X
110	−X
111	0X

calculations, particularly for big numbers. Additionally, compared to Radix Booth-4, it will reduce addition and subtraction operations, which is crucial when optimizing multiplication on digital circuits.

Table 2 Radix-4 booth recoding

Multiplier bits	Recoded operation on multiplicand, X
0000	0X
0001	+X
0010	+X
0011	+2X
0100	+2X
0101	+3X
0110	+3X
0111	+4X
1000	−4X
1001	−3X
1010	−3X
1011	−2X
1100	−2X
1101	−X
1110	−X
1111	0X

3 FPGA-Based Approximately Multiplication Techniques

In this section, we survey the proposals and implementations for Approximate Multiplier implemented FPGA-Based platforms in the literature. We then compare the proposals in terms of resource usage, latency and power consumption.

3.1 FPGA-Based Approximate Multiplier Implementations

Two primary types of approaches to approximate multiplicative circuits are as follows: to multiply two binary values for FPGA, the first technique is to truncate the group of bits or partial products during multiplication using well-known algorithms like the Radix Booth, and Baugh Wooley algorithms. The other technique truncates elements that are superfluous or that use excessive amounts of resources without significantly altering the outcome by breaking down the general multiplication formula of two binary numbers into smaller components.

Radix Booth and Baugh Wooley algorithms are employed in a logic cloning approach that is based on the probability of each bit when multiplying two binary numbers in [10]. The logical value 1's presence in partial product values between two bits of two components combined is analyzed and statistically reported by the author

in this article. In order to merge them together, they then choose successive partial products on the row that have the same probability value. By doing this, a significant amount of resources (the number of LUTs for each algorithm) are reduced, and hardware processing and computing performance are increased. However, because each algorithm will have a distinct means of truncating and grouping truncated bits, the inaccuracy in this method is significant and very hard to manage. Figures 4 and 5 present the Booth and Baugh Wooley architectures.

Muhammad Awais et al. [15] used the same method as Salim Ullah et al. [14] for handling an element's bottom bits in multiplication. Depending on each unique situation, the author has modified and altered these sections. Not only does this result in a significant decrease in the quantity of LUTs and energy resources, but it also boosts computational processing speed and accuracy in comparison to Salim Ullah et al. [14].

Three distinct techniques are provided by Salim Ullah et al. [12] for truncating successive partial products. In order to produce sum vectors, the first and most straightforward way is to add two successive partial product rows. The final result is then created by using the 6-input LUT-based approximation addition. The architecture used in Methods 2 and 3 is the same: truncate the respective partial products together. The 6-input LUTs design in the approximate computation is where the two approaches mentioned above diverge. Based on the three above-mentioned suggested solutions, the author additionally provides a recursive method design for higher order multipliers. Although the number of LUTs employed is significantly increased, the findings of this article are rather satisfactory in terms of latency and approximation accuracy. Figure 6 presents the general idea of the second and the third approaches in the [12] being applied approximate technique in the green part.

The difference in structure of these two methods is shown in Fig. 7.

Salim Ullah et al. [13] created the result vectors by grouping three successive partial products in the Baugh Wooley method together by utilizing the 6-input lookup table. To get the final result, these result vectors are once more grouped three by one. The article has remarkable performance in terms of reduced latency and resource conservation.

Figure 8 highlights the method of grouping Partial Products in order to generate the Final Product, a key technique used in [13].

The general formula for multiplying two binary numbers into smaller components was examined by Hao Zhang [16], and they applied approximations to the sections that had little bearing on the multiplication outcome. This article's primary objective is to apply the approximation approach to low bits in order to dramatically reduce computation accuracy while boosting calculation speed and decreasing resource use. Figures 9 and 10 show the Datapath Design and the Approximate multiplier algorithms applied of two methods in [16], respectively.

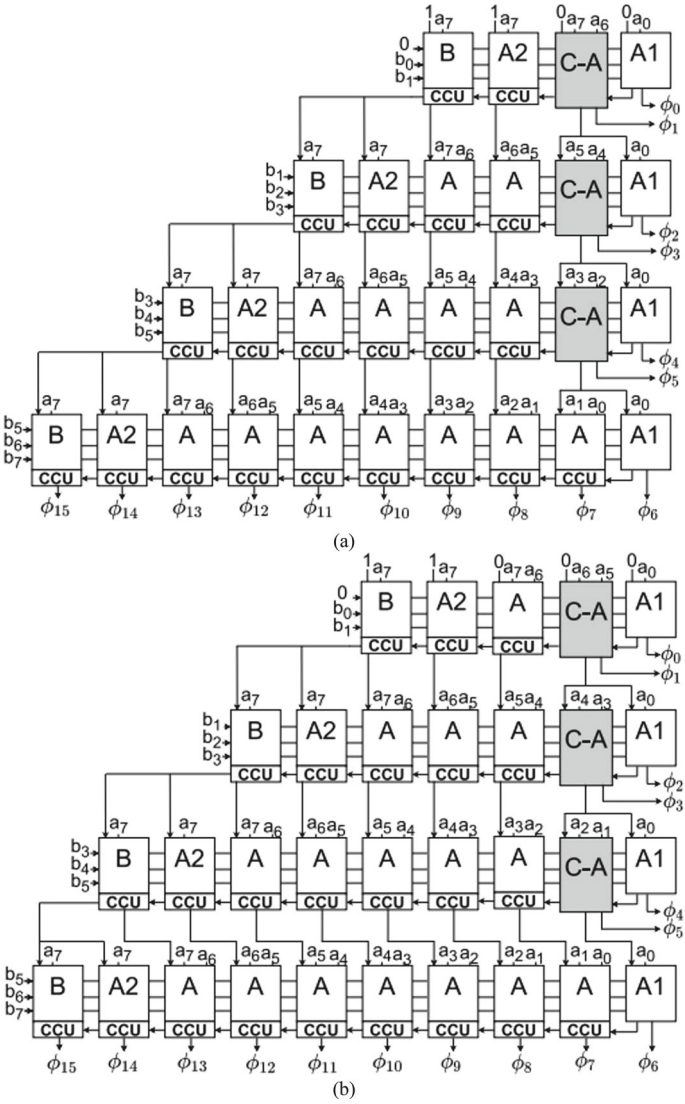


Fig. 4 LC-Booth circuits for 8-bit operands (**a**) LC-Booth-1 (**b**) LC-Booth-2 in [10]

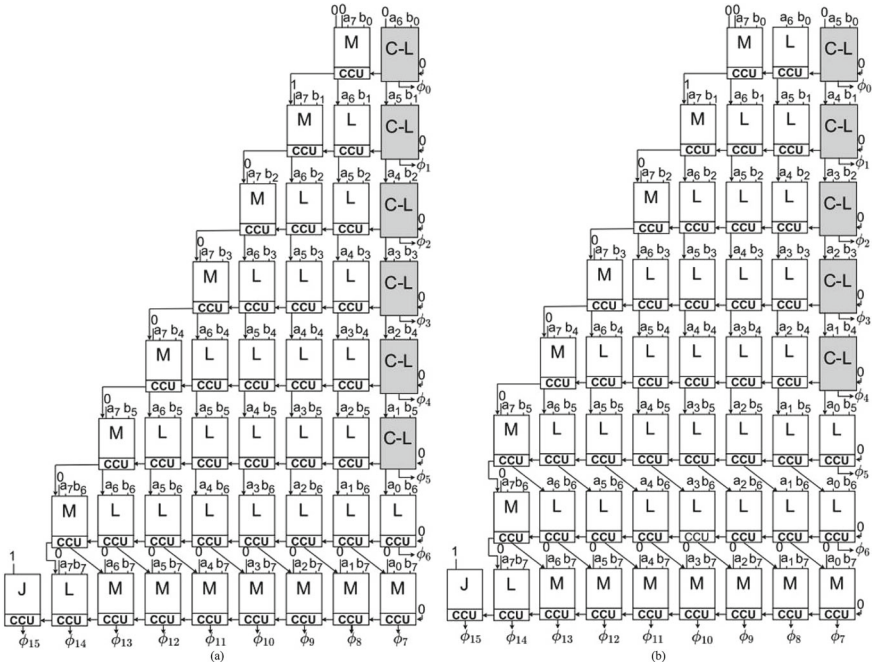


Fig. 5 LC-BW circuits for 8-bit operands (**a**) LC-BW-1 and (**b**) LC-BW-2 in [10]

3.2 Comparison

Table 3 compares the proposals for 8-bits multiplier analyzed above in terms of FPGA families, area (LUTs), Critical Path Delay in ns (CPD), Power-Delay Product in (PDP), latency and Power (W). Table 4 summarizes the pros and cons of the proposed approximate multipliers in the literature.

3.3 Discussion

The impact of approximation multiplication approaches on latency, complexity, and power consumption will be covered in this section. First of all, it is evident that the majority of the aforementioned strategies will produce an approximation multiplication technique by minimizing or optimizing the critical route from the original precise multiplication algorithm. When compared to the original exact multiplication technique, this results in a considerable reduction in the amount of adds and subtractions made throughout the calculation process, which speeds up the computation and lowers latency. Furthermore, because the more simple the circuit design, the lower the power consumption, this decrease results in optimization of

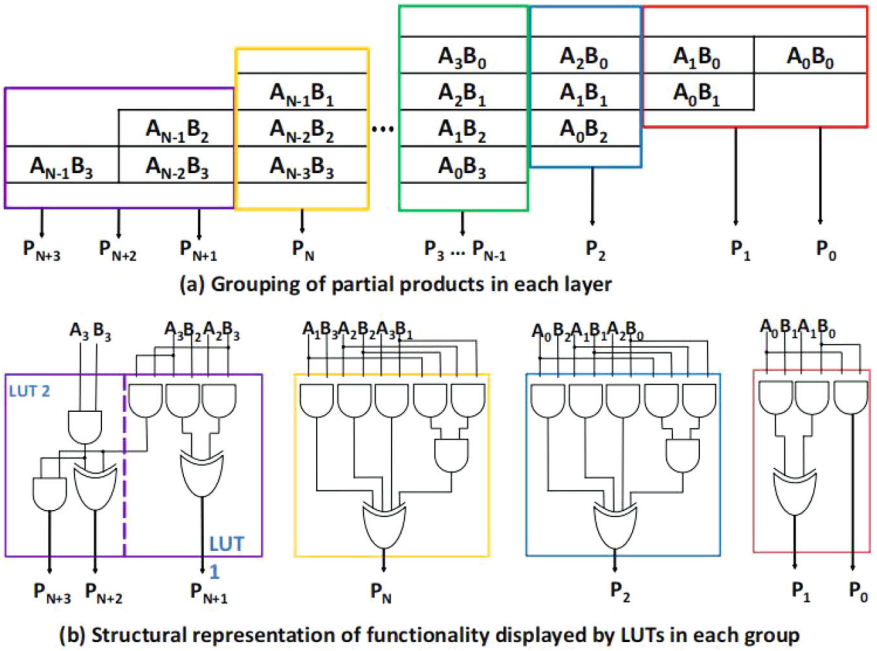


Fig. 6 Partial products generation for Approx2 and Approx3 in [12]

Table 3 Results comparison of different 8-bits Approximate multiplier implementations in the literature

Proposal	Area	CPD	PDP	Latency	Power
[10]**LC-BW-1**	50	5.20	5.20	–	–
[10]**LC-BW-2**	55	5.39	5.39	–	–
[10]**LC-Booth-1**	27	4.04	3.43	–	–
[10]**LC-Booth-2**	34	3.93	3.34	–	–
[15]	32	2.54	1.66	–	–
[14]	92	4.99	7.10	–	–
[12]**Approx 1**	64	9.222	–	9.222	0.22
[12]**Approx 2**	54	9.038	–	9.038	0.217
[12]**Approx 3**	54	9.081	–	9.081	0.216
[13]	54	4.37	7.26	2.388	–
[16]**Approx 1**	287	–	65.1	3.1	0.021
[16]**Approx 2**	297	–	66.7	2.9	0.023

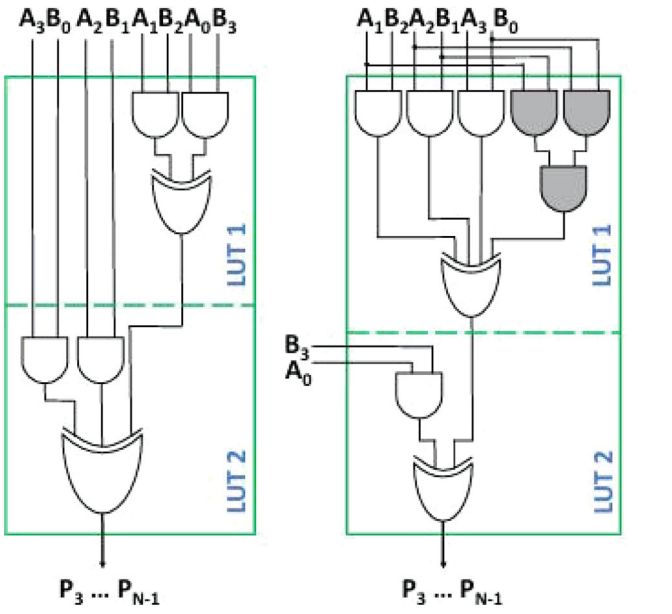


Fig. 7 Structural difference between Approx2 and Approx3 in [12]

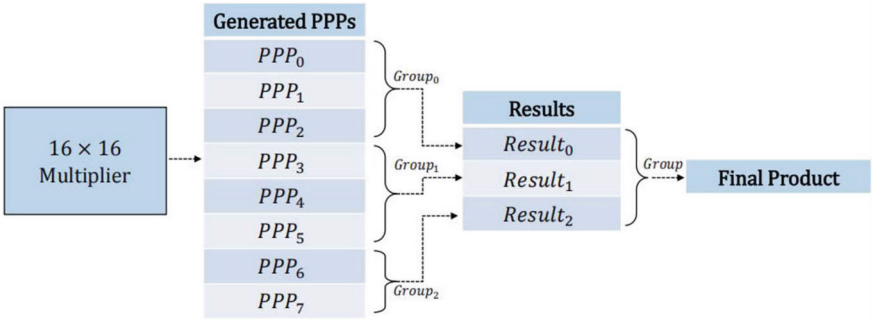


Fig. 8 Grouping of PPPs to compute final product in [13]

both design space and power consumption. The complexity will be lower than with the original design since some techniques optimize effectively between the suggested algorithm and the FPGA's accessible resources, which are LUT5 and LUT6. Some techniques, like [12], the authors did not be optimized with LUT5 and LUT6 resources because the approximate calculation portion includes eight inputs, but LUT5 and LUT6 can only be used with a design that has a maximum of five or six inputs. Lastly, with regard to accuracy, the aforementioned articles all decide to

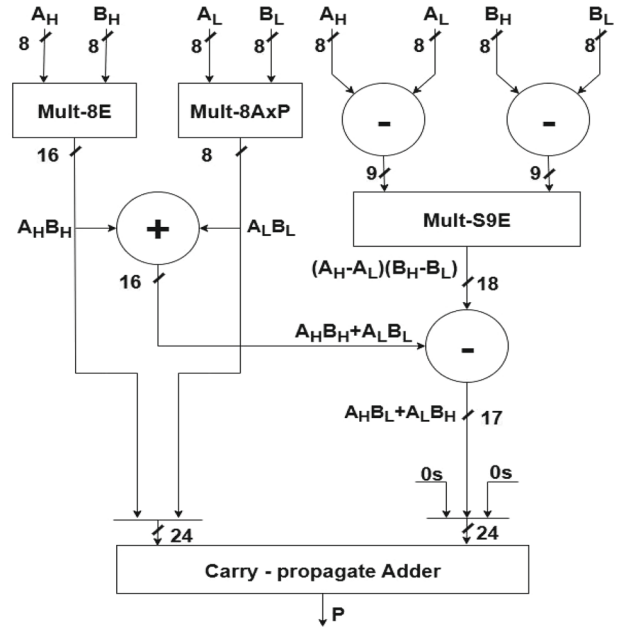


Fig. 9 Datapath of (**a**) the first and (**b**) the second approaches in [16]

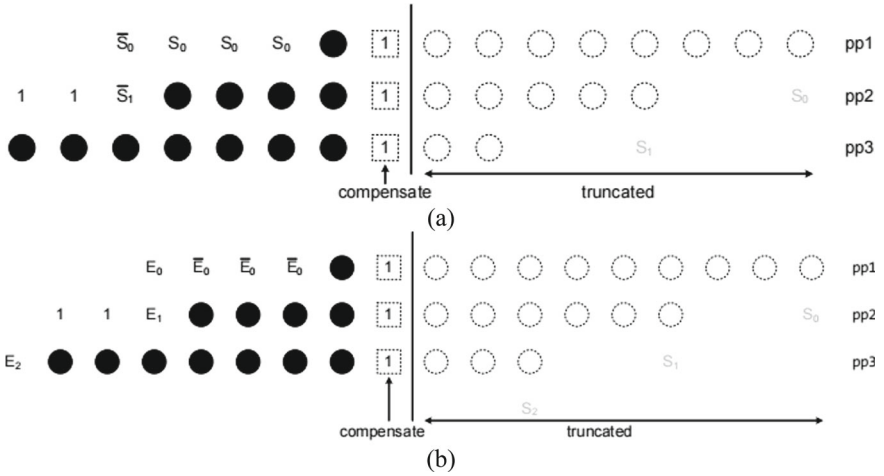


Fig. 10 Approximate design of (**a**) the first and (**b**)the second approaches in [16]

Table 4 Pros and Cons of the proposed approximate multipliers in the literature

References	Pros	Cons
[10]	Improved performance in terms of speed and LUTs resource usage	Accuracy is reduced due to the effect on the results of all bits in the final products
[15]	Significant improvement in resource utilization: number of LUTs, low latency and low power consumption	Difference between the exact products and the approximate products are high
[14]	Accuracy is not reduced significantly	Limited resource optimization: number of LUTs, latency and power consumption
[12]	High precision, low power consumption, greatly reducing the amount of LUTs used	Latency is still too high, architecture is not optimized, still too complex for popular LUTs: LUT5 and LUT6
[13]	Optimized number of LUTs used and low latency	The accuracy of the final calculation result is significantly affected
[16]	Optimized for speed and energy usage, accuracy is not significantly affected	The number of LUTs is still large and the complexity of the algorithm is still high

test and compare in various apps and media, so when compared to one another, they are unable to provide a comprehensive and most true picture of this element. But generally speaking, the findings show high accuracy when compared to designs that use approximate calculation circuits or from the original designs.

4 FPGA-Based Approximately Multiplication: Open Issues of AI Powered Application

Despite the abundance of studies in the literature on novel methodologies for FPGA-based approximation multipliers, some unsolved challenges remain for future study about their applicability in AI-powered applications. We highlight six outstanding challenges that must be addressed to enhance the performance of FPGA-based AI applications.

1. **Performance and Resource Constraints**: FPGAs and other edge devices are subject to stringent power limits and have restricted computing resource. Moderate power consumption and the absence of noticeable performance deterioration are prerequisites for applying AI models to these devices. It is necessary to optimize AI models for low-resource devices while preserving stability and accuracy within a permissible range.
2. **Design Complexity and Expertise**: Creating AI solutions for FPGAs necessitates a thorough understanding of artificial intelligence, mathematics, and hardware design. Complex design and verification procedures are required to create accurate and efficient FPGA designs. The development cycle is prolonged and complexity is increased by the steep learning curve and scarcity of experienced designers in FPGA development.
3. **Adaptability and Flexibility**: The dynamic nature of edge-AI settings may necessitate regular security measure upgrades to handle emerging threats or modifications to the operating environment. The flexibility of FPGA-based systems must enable real-time upgrades and reconfigurations. It must be possible to accomplish this flexibility without jeopardizing system security or necessitating prolonged downtime.
4. **Scalability and Integration**: To safeguard a wide variety of edge devices, approximate multiplier AI-powered solutions need to be scalable. For widespread use, interoperability with different FPGA designs, operating systems, and communication protocols is crucial. Managing varying computing capacities and resource limitations across devices is another aspect of this scalability.
5. **Robustness and Reliability of AI Models**: In AI-powered applications like object classification or object detection, AI models used as approximation multipliers need to be dependable and strong. They must be able to function well in real-world scenarios, including managing noisy data and changing threat environments, and resist hostile attacks. It is especially difficult to guarantee the dependability of these models in settings with limited resources.

6. **Resource-Efficient Implementation**: For edge devices, it is crucial to implement AI solutions in a resource-efficient way. This entails preserving strong security while reducing the use of memory, processing power, and energy. It takes creative design strategies and optimization tactics to balance these limitations.

5 Conclusion

As integrated circuits grow in size and complexity, approximation multipliers must be developed for probabilistic applications or for processes that do not require high accuracy. as artificial intelligence has grown. A variety of approaches have been proposed, all aiming to maximize processing speed, area resources, and energy consumption while preserving the necessary performance of the application. One such review method is approximation computing, which permits explicit control over the energy and hardware resource utilization for the intended application by enforcing penalties on the system's accuracy. The foundation of approximate computing is the notion that application systems are not always necessary to provide the most precise results and still maintain acceptable system performance. The area, latency, and energy consumption of approximation multiplier implementations based on FPGA are systematically reviewed in this paper. Lastly, using this method, we point out outstanding difficulties for FPGA-based AI-powered applications.

Acknowledgements This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number NCM2021-20-02.

References

1. Ahmadinejad M, Moaiyeri MH (2021) Energy- and quality-efficient approximate multipliers for neural network and image processing applications, p 1
2. Ansari MS, Cockburn BF, Han J (2021) An improved logarithmic multiplier for energy-efficient neural computing. IEEE Trans Comput 70(4):614–625
3. Ansari MS, Mrazek V, Cockburn BFL, Sekanina ZV, Han J (2020) Improving the accuracy and hardware efficiency of neural networks using approximate multipliers 28(2):317–328
4. Chippa V, Chakradhar SK, Roy AR (2013) Analysis and characterization of inherent application resilience for approximate computing, pp 1–9
5. Donald EK (1969) The art of computer programming, vol 2, p 724
6. Han J, Orshansky M (2013) Approximate computing: an emerging paradigm for energy-efficient design, pp 1–6
7. Jiang H, Santiago F, Mo H, Liu L, Han J (2020) Approximate arithmetic circuits. A survey characterization and recent applications, vol 12, pp 2108–2135
8. Karatsuba A, Ofman Y (1963) Multiplication of many-digital numbers by automatic computers, pp 595–596
9. Karatsuba AA (1995) The complexity of computations, pp 186–202

10. Kulkarni A, Ouameur MA, Massicotte D (2024) Logic cloning based approximate signed multiplication circuits for FPGA
11. Magnus Själander PLE (2008) High-speed and low-power multipliers using the Baugh-Wooley algorithm and HPM reduction tree
12. Ullah S, Murthy SS, Kumar A (2018) Smapproxlib: library of FPGA-based approximate multipliers
13. Ullah S, Rehman S, Shafique M, Kumar A (2021) High-performance accurate and approximate multipliers for FPGA-based hardware accelerators
14. Ullah S, Schmidl H, Sahoo SS, Rehman S, Kumar A (2021) Area-optimized accurate and approximate softcore signed multiplier architectures
15. Zahir MA, Shah SAA (2023) Toward optimal softcore carry-aware approximate multipliers on xilinx fpgas
16. Zhang H, Xiao H, Qu H, Ko SB (2021) FPGA-based approximate multiplier for efficient neural computation

Investigation of Potential Impact on Drain Current of CMOS Inverter with Various Transistor Width and Temperature



Biswazit Mistry and Suman Chowdhury

1 Introduction

The progress in semiconductor tech has made electronic devices smaller and better opening doors for more intricate and compact integrated circuits [1]. In this tech shift, the CMOS inverter has become a key part of digital circuit design [2]. It plays a crucial role in how we implement logic and process signalsIt doesn't use much power when it's not active, the user can fit a lot of them in a small space, and it switches between states [3–5]. As integrated circuits get smaller reaching nanometer sizes, the physical size of transistors (how wide their channels are) along with working conditions like temperature, have a big impact on how well devices work. These factors together affect key aspects, and the drain current is important among them [6, 7]. The drain current not determines how fast CMOS inverters can switch but also has an effect on the dynamic and static power use in digital systems [4]. When it comes to CMOS technology, the drain current is sensitive to changes in transistor shape and temperature. To design circuits that use power well and can handle heat, we need to grasp these relationships [3, 8]. As we want faster and more energy-saving electronics for mobile devices built-in systems, and AI hardware, it's essential to model current [9–11]. This has become more important than ever.

This study seeks to examine how two key factors (transistor width and temperature) have an impact on the drain current properties of a CMOS inverter [12]. The research team used the Cadence Virtuoso simulation platform working with the

B. Mistry · S. Chowdhury (✉)
Department of Electrical and Electronic Engineering, International University of Business Agriculture and Technology, Dhaka, Bangladesh
e-mail: suman@iubat.edu

B. Mistry
e-mail: 22105044@iubat.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_14

GPDK 90nm technology node. They ran simulations on NMOS and PMOS transistors varying channel widths from 120 to 520nm to take a close look at scaling effects. At the same time, they applied heat changes from 27 to 100 °C to the smallest and largest width setups to see how these affected the inverter's electrical behavior. Looking at these changes has shown important patterns that highlight how physical design choices and environmental factors work together [5, 13]. As width increases, we see drain current go up too, which matches what theory predicts. Higher temperatures also have an impact on how carriers move and the threshold voltage causing changes in current flow. Understanding these behaviors helps us predict how CMOS-based circuits will perform in real-world conditions. This study matters beyond just academic interest. In practice, what we've learned helps to improve predictive modeling in circuit simulators and design flows that consider layout [14]. By showing the non-linear and interconnected effects of transistor size and temperature, the research informs design methods that aim to balance power and performance in the best way possible [10, 15]. The importance of CMOS technology in modern electronics cannot be understated, as it provides several advantages [16]. A very important one is low power consumption which heightens the appeal of the technology in battery operated or portable devices. Their performance is also highly reliable and stable as it shows a great amount of resistance to electrical noise [17]. Transistor scaling is also very helpful in the development of compact and complex integrated circuits, which is something that is supported by CMOS technology. This has greatly advanced processors and memory devices [7, 14]. Besides power saving, the heat generated by CMOS circuits is very low. As a result, the thermal management in electronic systems is enhanced. Such advanced features also allow for the addition of more functionality in smaller chips, further reducing the size of the devices. The capability to perform both digital and analog operations on the same chip is an advanced trait of CMOS technology, which adds on to the versatility of the technology [18–20]. These attributes also enable faster switching speeds, making it valuable for high performance computing applications [21].

The cost-effective nature of CMOS technology due to efficient fabrication techniques is complemented by the low current draw of the circuits, which is only during the transitions [22, 23]. The wide range of frequencies and voltages that the device can operate under, as well as its lower energy usage, increases the adaptability of the technology. Finally, the long operational life, alongside the low vulnerability to extreme conditions, makes CMOS technology very durable [12, 24, 25]. Cadence Virtuoso is a critical tool in the design and simulation of analog and mixed-signal integrated circuits. It provides a comprehensive environment for schematic entry, layout design, and circuit simulation [26]. The platform supports precise modeling of real-world circuit behavior, making it essential for verifying functionality before fabrication. With its advanced layout editing features, designers can ensure that physical designs meet performance and manufacturing requirements. Virtuoso integrates tightly with process design kits (PDKs), allowing accurate device-level design [27–29]. Its support for parasitic extraction and verification helps optimize circuit performance. The tool also streamlines collaboration among design teams through a unified interface. Overall, Cadence Virtuoso enhances design efficiency, accuracy,

and innovation in IC development [27]. The findings are especially important for the semiconductor manufacturing and fabrication sector. As devices need to function reliably under various thermal conditions and performance levels, this research provides valuable insights into transistor sizing strategies that meet system-level requirements. It also helps enhance the accuracy of device libraries used in standard-cell design, which ultimately affects the reliability and functionality of countless devices. By thoroughly examining the drain current behavior of CMOS inverters across a controlled range of transistor widths and temperatures, the study lays out a solid framework for grasping essential aspects of nanoscale device performance. These insights are set to benefit designers, researchers, and engineers in the realms of digital electronics, VLSI design, and semiconductor technology. As technology nodes keep shrinking and operating conditions become more unpredictable, this foundational knowledge is crucial for developing the next generation of integrated systems.

2 CMOS Model

A CMOS inverter (Fig. 1) is the most basic building block in the digital integrated circuit. It consists of a PMOS (P-type metal oxide semicunductor) and an NMO (N-type metal oxide half se micunductor) transistors in an additional arrangement. The entrance is connected to both gates to the transistor, while the output is taken from the connection between PMOS and NMOS drainage. The PMOS source is connected to VDD (supply voltage), and the NMOS source is grounded. When the entrance is low, PMOS works, the output draws high; When the entrance is high, NMO is run and lowers the output. This supplementary action ensures less stable power consumption, an identity of CMOS technology.

The inverter switch (Fig. 2) explains quickly, making it ideal for high -speed logic application. The symmetrical structure contributes to balanced growth and downturn, important for the accuracy of the time. The CMOS inverter is used on a large scale in logical gates, buffers and memory circuit. They are also fundamental in complex digital systems such as processor and FPGA. The current behavioral analysis helps to understand dynamic strength waste and leakage flows. As a transistor -shaped scale, understanding of inverter current properties is quickly important. Power optimization and speed improvement depend a lot on the performance at the inverter level. Advanced techniques such as fine and street-all-Around Fet are discovered to improve CMOS behavior. In the future, CMOS converters will play an important role in Ultra-lo power and high-speed computer systems. Emerging areas such as neuromorphic data processing and quantum-dot cellular automata are still dependent on CMOS compatibility. The strength, scalability and energy efficiency of CMOS converters make them necessary in IC design. With continuous research, cmos technology is going to boyond the 3 nm. In the world lot of University and research laboratory are trying to innovate new materials which are suited for the new semiconductor

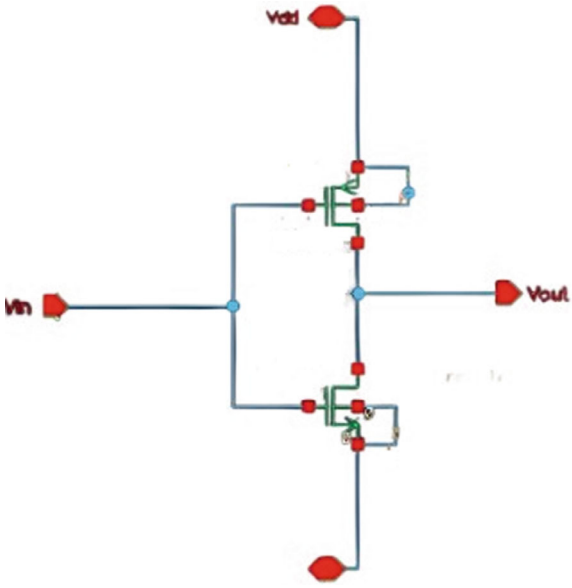


Fig. 1 Schematic of CMOS inverter

technologys. Therefore, CMOS remais one of the best choise for the next generation semiconductor technologies.

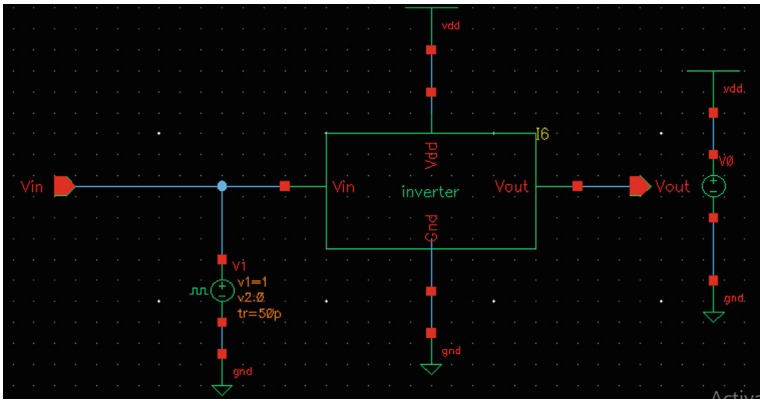


Fig. 2 Symbol of CMOS inverter for the simulation

3 Results and Discussion

Figure 3 presents the power voltage characteristics of a CMOS converter, focusing on the effect of different NMOS transistor width (Wn) in different NMOS transistor width (Wn) in five steps from 120 to 520 nm. The X-axis input represents the DC voltage (V), while the Y-axis show the current (I) flowing through the Vdd terminal. Each color-coded curve matches another NMOS width, as indicated in the legend. As the width of NMOS increases, the top current also increases in plots. This behavior reflects the strong wiring capacity of the wide NMOS transistor due to low channel resistance. The top current is seen around the input voltage of around 339 mv, which reflects the switching threshold to the inverter. At this time, both PMOS and NMOSoperate partially, which causes maximum short -circuit current. The red curve (width of 120 nm) shows the lowest top current around 8.16 μa due to its narrow channel. When the width increases to 520 nm (blue curve), the top current reaches about 12.36 μa. This trend shows a direct connection between the breadth of the transistor and the strength. Proper turns and slight increase in curves suggest improving the prey response with wide NMOS. Greater electricity levels can lead to rapid infection, but can also contribute to an increase in power consumption. It is important to understand this trade-off for delay in CMOS circuit and optimization of energy efficiency. Such analysis is essential for digital circuit design transistor size, especially in high -speed applications. These results show how geometric scaling directly affects the electrical behavior of the CMOS inverter.

Figure 4 shows how varying the width of the PMOS transistor (Wp) affects the current–voltage characteristics of a CMOS inverter. We increased the PMOS width from 120 to 520 nm in five equal steps. On the x-axis, showing of the input DC voltage, while the y-axis represents the current flowing through the Vdd terminal. Each curve on the graph corresponds to a different PMOS width, which can identify using the color legend. As the PMOS width grows, thepeak current also rises, indicating a stronger drive capability. The red curve, which represents the smallest width of

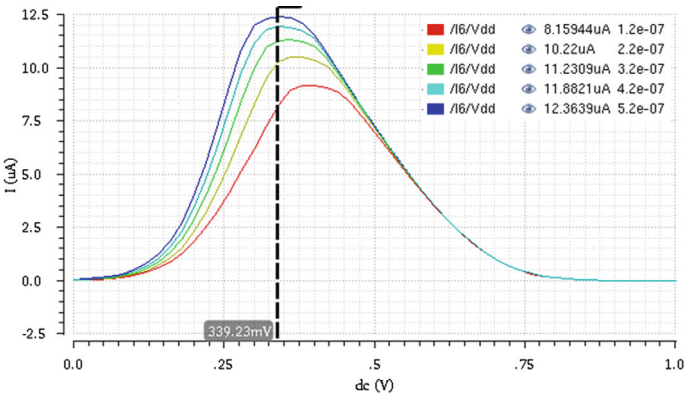


Fig. 3 Current sweep for Wn (NMOS) change from 120 to 520 nm

120 nm, shows the lowest peak current at about 7.37 μA. In contrast, the blue curve for the widest PMOS at 520 nm reaches a peak current of around 19.82 μA. This trend clearly illustrates how PMOS sizing impacts the overall current behavior of the inverter. The highest current is observed near the switching threshold voltage, which is around 488 mV in this simulation. At this point, both transistors are conducting simultaneously, leading to maximum short-circuit current. A wider PMOS reduces resistance in the pull-up path, allowing for greater current flow. This change also enhances current efficiency during logic level transitions. However, keep in mind that larger PMOS devices might introduce higher parasitic capacitance, which can affect speed and power consumption. The curves display a sharp rise and fall, indicating quicker switching as the width increases. This analysis is essential for optimizing power, delay, and sizing trade-offs in high-performance CMOS circuits.

Figure 5 illustrates the influence of temperature variation on the drain current behavior of a CMOS inverter, where both NMOS and PMOS transistors maintain a constant width of 120 nm. The horizontal axis represents the input voltage (V), while the vertical axis denotes the corresponding drain current (I) measured in microamperes. Temperature levels were incrementally adjusted from 27 to 100 °C, producing distinct curves represented in different colors. A noticeable peak in current is observed near the switching threshold, approximately at 397.96 mV. As the temperature rises, a gradual decline in peak current is evident across the curves. The curve corresponding to 27 °C shows the highest peak current, around 9.15 μA, while the current reduces to approximately 7.81 μA at 100 °C. This decreasing trend in current with elevated temperature is primarily attributed to reduced carrier mobility, a consequence of intensified phonon scattering within the semiconductor lattice. Additionally, a leftward shift in the peak positions indicates alterations in threshold voltage, further influencing the inverter's switching behavior. These temperature-induced effects are critical, as they directly affect the performance, timing accuracy, and power efficiency of CMOS-based integrated circuits. In high-density designs, such thermal sensitivity can lead to reliability concerns if not addressed during the design phase. Understanding these characteristics is crucial for developing robust

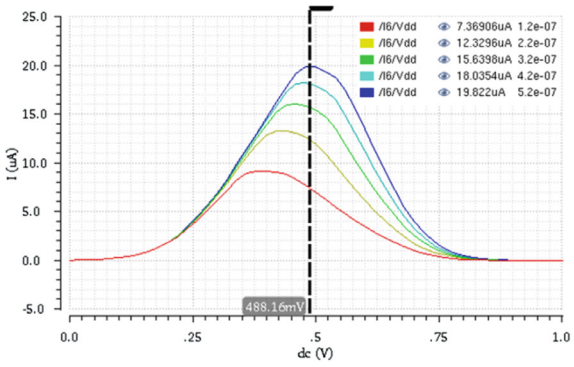


Fig. 4 Current sweep for Wp (PMOS) change from 120 to 520 nm

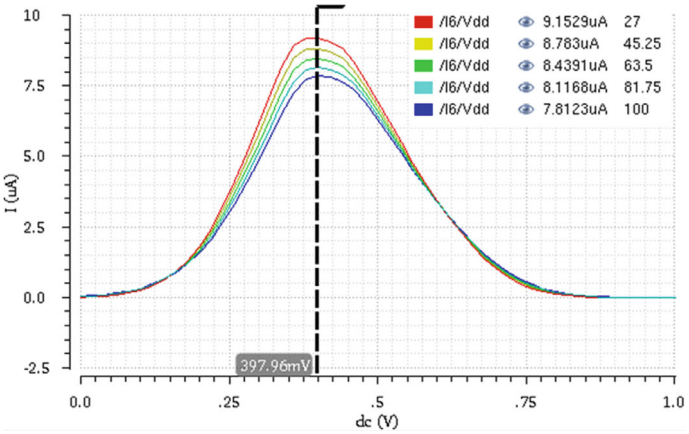


Fig. 5 Current sweep for temp change from 27 to 100 °C for width of 120 nm

and thermally resilient semiconductor systems. The data serves as a basis for implementing thermal management strategies, such as compensation circuits and material selection optimized for stability under thermal stress. Such insights are particularly valuable in safety–critical and high-speed electronic applications, where operational integrity must be maintained across a wide range of environmental conditions. Overall, the analysis underscores the importance of temperature-aware modeling in modern semiconductor design and its role in enhancing system-level reliability.

Figure 6 illustrates how drain current varies with input voltage across different temperature conditions for a CMOS transistor with a fixed width of 520 nm. Five distinct curves, ranging from 27 °C (red) to 100 °C (dark blue), represent the thermal impact on current behavior. Each curve follows a bell-shaped trajectory, where current increases with voltage up to a peak around 0.4–0.5 V and then declines as voltage continues to rise. This peak shifts slightly with temperature, and a clear trend is observed: the peak current diminishes as temperature increases. Specifically, the maximum current drops from approximately 36.64 µA at 27 °C to 31.20 µA at 100 °C, indicating a negative temperature coefficient. The convergence of all curves at both low and high voltage extremes suggests that temperature effects are most significant in the mid-voltage region, near the inverter's switching threshold. This behavior is primarily attributed to reduced carrier mobility at elevated temperatures due to enhanced phonon scattering within the semiconductor lattice. As mobility decreases, the ability of charge carriers to conduct current weakens, leading to a noticeable drop in current.

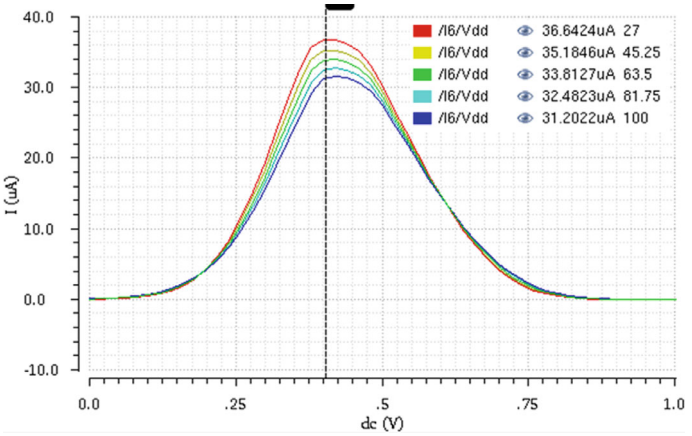


Fig. 6 Current sweep for temp change from 27 to 100 °C for width of 520 nm

4 Conclusion

This analysis holds significant relevance in modern semiconductor design. Accurate modeling of thermal effects is essential for predicting device performance under real-world conditions. As integrated circuits operate across a wide temperature range, understanding the thermal sensitivity of transistors allows designers to implement compensation techniques, optimize timing margins, and ensure functional reliability. Moreover, such temperature-dependent characterization supports more precise power and delay estimations in scaled CMOS technologies, where variations in operating conditions can substantially affect performance. By incorporating these findings into simulation and verification workflows, engineers can enhance the robustness and energy efficiency of next-generation electronic systems. From the temperature variation it is observed that the maximum current is decreased by 14.85% for temperature change from 27 to 100 °C keeping the transistor width same as 520 nm whereas it is 14.84% for transistor width of 120 nm. On top of this, the peak current is increased dramatically by 51.47% for the change of NMOS width from 120 to 520 nm whereas it is 166.21% for the same variation in PMOS.

References

1. Hossain N, Mobarak MH, Mimona MA, Islam MA, Hossain A, Zohura FT, Chowdhury MA (2023) Advances and significances of nanoparticles in semiconductor applications—a review. Results Eng 19:101347
2. Çeliker H, Dehaene W, Myny K (2024) Multi-project wafers for flexible thin-film electronics by independent foundries. Nature 629(8011):335–340
3. Yau N (2024) Visualize this: the flowing data guide to design, visualization, and statistics. John Wiley & Sons

4. Zhou T, Niu P, Sun L, Jin R (2023) One fits all: power general time series analysis by pretrained lm. Adv Neural Inf Process Syst 36:43322–43355
5. Chisnall D (2023) How to design an ISA: the popularity of RISC-V has led many to try designing instruction sets. Queue 21(6):27–46
6. Zuccarini P, Sardans J, Asensio L, Peñuelas J (2023) Altered activities of extracellular soil enzymes by the interacting global environmental changes. Glob Change Biol 29(8):2067–2091
7. Bhandari P, Rawat KS, Prajapati YK, Padalia D, Ranakoti L, Singh T (2023) Design modifications in micro pin fin configuration of microchannel heat sink for single phase liquid flow: a review. J Energy Storage 66:107548
8. Petri I, Rezgui Y, Ghoroghi A, Alzahrani A (2023) Digital twins for performance management in the built environment. J Ind Inf Integr 33:100445
9. Biswas P, Rashid A, Biswas A, Nasim MAA, Chakraborty S, Gupta KD, George R (2024) AI-driven approaches for optimizing power consumption: a comprehensive survey. Discover Artif Intell 4(1):116
10. Mishra P, Singh G (2023) Energy management systems in sustainable smart cities based on the internet of energy: a technical review. Energies 16(19):6903
11. Segun-Falade OD, Osundare OS, Kedi WE, Okeleke PA, Ijomah TI, Abdul-Azeez OY (2024) Developing innovative software solutions for effective energy management systems in industry. Eng Sci Technol J 5(8):2649–2669
12. Wang H, Lai P, Islam MZ, Hasan ASMK, Di Mauro A, Anika NEA, Mantooth HA et al (2024) A review of silicon carbide CMOS technology for harsh environments. Mater Sci Semicond Process 178:108422
13. Patel KR (2023) Harmonizing sustainability, functionality, and cost: navigating responsible packaging innovations in modern supply chains. Am J Econ Manage Bus (AJEMB) 2(8):287–300
14. Li Y, Xue C, Zargari F, Li YR (2023) From graph theory to graph neural networks (GNNs): the opportunities of GNNs in power electronics. IEEE Access 11:145067–145084
15. Rafin SSH, Ahmed R, Haque MA, Hossain MK, Haque MA, Mohammed OA (2023) Power electronics revolutionized: a comprehensive analysis of emerging wide and ultrawide bandgap devices. Micromachines 14(11):2045
16. Finocchio G, Incorvia JAC, Friedman JS, Yang Q, Giordano A, Grollier J, Bandyopadhyay S et al (2024) Roadmap for unconventional computing with nanotechnology. Nano Futures 8(1):012001
17. Verma R, Thakur P, Chauhan A, Jasrotia R, Thakur A (2023) A review on MXene and its' composites for electromagnetic interference (EMI) shielding applications. Carbon 208:170–190
18. Chen Y, Nazhamaiti M, Xu H, Meng Y, Zhou T, Li G, Dai Q et al (2023) All-analog photoelectronic chip for high-speed vision tasks. Nature 623(7985):48–57
19. Duan X, Cao Z, Gao K, Yan W, Sun S, Zhou G, Sun B et al (2024) Memristor-based neuromorphic chips. Adv Mater 36(14):2310704
20. Wan T, Shao B, Ma S, Zhou Y, Li Q, Chai Y (2023) In-sensor computing: materials, devices, and integration technologies. Adv Mater 35(37):2203830
21. Hou W, Wang J, Lv JA (2023) Bioinspired liquid crystalline spinning enables scalable fabrication of high-performing fibrous artificial muscles. Adv Mater 35(16):2211800
22. Liu A, Zhang X, Liu Z, Li Y, Peng X, Li X, Ren TL et al (2024) The roadmap of 2D materials and devices toward chips. Nano-Micro Lett 16(1):119
23. Zhou Z, Ou X, Fang Y, Alkhazraji E, Xu R, Wan Y, Bowers JE (2023) Prospects and applications of on-chip lasers. Elight 3(1):1
24. Teng Q, Hu Y, Cheng R, Wu Y, Zhou G, Gao D (2023) Reliability challenges in CMOS technology: a manufacturing process perspective. Microelectron Eng 281:112086
25. Sbrana C, Catania A, Paliy M, Di Pascoli S, Strangio S, Macucci M, Iannaccone G (2024) Design criteria of high temperature integrated circuits using standard soicmos process up to 300 °C. IEEE Access

26. Antsiferova V, Grosheva E, Ivanova A, Abrosimova I (2023) Computer simulation of electrophysical effects in CAD chip design. In: E3S web of conferences, vol 389. EDP Sciences, p 07015
27. James A, Rizzo A, Wang Y, Novick A, Wang S, Parsons R, Bergman K (2023) Process variation-aware compact model of strip waveguides for photonic circuit simulation. J Lightwave Technol 41(9):2801–2814
28. Zhu J (2025) System-technology co-optimization of scaled electronics based on two-dimensional materials (Doctoral dissertation, Massachusetts Institute of Technology)
29. Aljafar MJ, Azais F, Flottes ML, Pagliarini S (2024) Utilizing layout effects for analog logic locking. J Cryptogr Eng 14(2):311–324

Impact of Agentic AI-Driven RPA on Strategic Management: Enhancing Decision-Making and Organizational Efficiency

N. Priyanka and M. Shankar Lingam

1 Introduction

The integration of Artificial Intelligence (AI) and automation technologies, particularly AI-powered Robotic Process Automation (RPA), is significantly transforming contemporary organizational practices. These technologies have revolutionized several industries by addressing operational inefficiencies, improving decision-making, and enabling organizations to scale effectively. AI-driven RPA systems optimize processes by automating routine tasks and providing real-time data-driven insights that enhance accuracy, enabling organizations to be more agile and responsive to changing business demands. This transformation is particularly evident in sectors like healthcare and finance, where AI solutions automate time-consuming processes, reduce human error, and enhance decision-making efficiency [1]. In healthcare, for example, AI-driven tools such as RPA assist in managing patient data, supporting medical decision-making, and facilitating efficient resource allocation in real-time [1]. In finance, AI tools optimize financial decision-making by automating transaction processes and identifying fraudulent activities [2]. These applications show the growing significance of AI and automation in improving organizational operations across industries. A key development in AI technology is the emergence of agentic AI, a paradigm that introduces a new level of autonomy to RPA systems. Agentic AI refers to systems that can autonomously make decisions and take actions, which is a marked departure from traditional RPA tools that operate primarily based on predefined rules. By integrating cognitive capabilities such as machine learning,

N. Priyanka (✉) · M. Shankar Lingam
University of Mysore, Mysore, India
e-mail: thepriyankanimmala@gmail.com

M. Shankar Lingam
Chaitanya Deemed to Be University, Hyderabad, India

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_15

predictive analytics, and natural language processing, agentic AI systems can handle more complex, data-driven tasks with minimal human oversight.

2 Literature Review

Overview of RPA and its Evolution with the Advent of AI: Robotic Process Automation (RPA) has evolved from simple rule-based task automation to more advanced, intelligent systems with the integration of AI technologies. Initially, RPA was limited to performing repetitive, structured tasks such as data entry and processing based on predefined rules. However, the integration of Artificial Intelligence (AI) into RPA systems has introduced new capabilities, enabling these systems to handle unstructured data, make decisions, and even predict outcomes. The evolution of RPA with AI is illustrated in the shift from attended automation, where human intervention is required, to unattended automation, where processes are fully autonomous [1]. As AI capabilities, such as machine learning (ML), natural language processing (NLP), and predictive analytics, have advanced, they have enhanced the capabilities of RPA, enabling systems to handle more complex tasks. RPA systems that leverage AI can learn from data, adapt to changing conditions, and continuously improve their performance. This synergy between AI and RPA is transforming business operations, making them more adaptive and responsive to dynamic business environments [2].

Literature on Agentic AI Systems: Their Characteristics, Capabilities, and Applications: Agentic AI systems represent a significant leap forward from traditional AI-powered RPA systems by incorporating autonomy and adaptive decision-making. These systems are capable of making autonomous decisions and adjusting their actions based on real-time data and environmental changes. Unlike traditional AI, which is often rule-based and dependent on predefined instructions, agentic AI integrates cognitive functions such as learning, reasoning, and problem-solving [1]. Applications of agentic AI span various sectors, including healthcare, finance, customer service, and manufacturing, where their ability to handle complex, dynamic tasks has led to significant improvements in efficiency and decision-making accuracy (Bhaskar et al. 2025).

2.1 Strategic Management in the AI Era

Previous Studies on the Integration of AI and RPA in Strategic Decision-Making: The integration of AI and RPA in strategic decision-making has been a subject of increasing academic interest. Studies indicate that AI-powered RPA systems can significantly enhance strategic decision-making by automating data collection, processing, and analysis, providing leaders with real-time insights that support faster,

data-driven decisions [2]. Moreover, AI can help decision-makers identify patterns, detect anomalies, and predict trends, allowing for more informed, forward-looking decisions.

Impact of Automation and AI on Organizational Agility and Decision-Making: AI and automation have a profound impact on organizational agility, enabling firms to respond more quickly to market changes and internal challenges. By automating routine tasks and providing data-driven insights, AI systems free up human resources to focus on higher-value tasks, enhancing overall productivity [2]. Additionally, organizations can make more informed decisions faster, which is critical in dynamic industries where time-to-market and responsiveness are key to competitive advantage. A study by Bhaskar et al. (2025) highlights that the adoption of AI-driven RPA systems leads to improved organizational agility by reducing decision-making latency. With AI's ability to quickly analyze large datasets and identify patterns, companies can make real-time adjustments to their strategies, better aligning with market demands and internal capabilities.

2.2 Challenges and Opportunities

Time Efficiency: One of the most significant advantages of AI-driven RPA is its ability to drastically reduce processing times for repetitive and rule-based tasks. According to a study by [2], organizations that implement AI-powered RPA have seen up to 40% reductions in transaction processing times, which leads to faster decision-making and improved operational efficiency. Improved Decision Accuracy: AI systems can process vast amounts of data with high accuracy, enabling better-informed decisions. For example, in healthcare, AI-driven systems have improved diagnostic accuracy by analyzing patient data and identifying early signs of health conditions [1]. In the finance sector, AI tools provide more precise market predictions and trading strategies, reducing human error. Cost Reduction: By automating manual tasks, AI-driven RPA reduces the need for human intervention, leading to cost savings. Studies have shown that RPA can reduce operational costs by up to 25% in certain industries [2]. Additionally, by minimizing human error and optimizing resource allocation, RPA leads to further cost reductions. Ethical Concerns: The deployment of AI systems, particularly in decision-making processes, raises significant ethical concerns. Issues such as bias in AI algorithms, privacy violations, and the lack of transparency in AI decision-making are pressing challenges [1]. For example, AI systems trained on biased data can perpetuate existing inequalities, leading to discriminatory outcomes.

3 Theoretical Framework

Agentic AI refers to autonomous systems that can make decisions and take actions based on real-time data and environmental changes, without requiring constant human intervention. In the context of Robotic Process Automation (RPA), agentic AI introduces cognitive capabilities that enhance traditional automation. Unlike conventional RPA, which operates on predefined rules and structured data, agentic AI systems use machine learning (ML), natural language processing (NLP), and predictive analytics to learn from data and adapt their behavior. These systems can manage complex tasks such as identifying patterns, making predictions, and improving decision-making processes dynamically [1]. In strategic management, agentic AI in RPA systems facilitates smarter decision-making by automating repetitive tasks and providing actionable insights. For example, in financial services, agentic AI-driven RPA tools can analyze market trends and automatically adjust investment strategies, ensuring that firms remain competitive in a rapidly changing environment [2]. This ability to make autonomous decisions based on evolving data is a significant advancement, moving organizations beyond static automation into dynamic, data-driven decision-making systems.

Theoretical Models Explaining How AI Agents Influence Decision-Making:

The Decision-Making Model of Autonomous Agents [1]: This model suggests that agentic AI systems, as autonomous decision-makers, can improve decision quality by processing large datasets and continuously learning from new information. These systems dynamically adjust their strategies in response to changes in the environment, which enhances decision-making efficiency. The ability of AI agents to act independently and adjust to new information in real-time ensures that decisions are not only faster but also more data-driven and accurate.

3.1 Decision-Making Theory

Rational Choice Theory: Rational Choice Theory posits that decision-makers make choices that maximize their utility, based on available information and the potential outcomes of their decisions [3]. In the context of agentic AI, this theory suggests that AI systems can be designed to mimic human decision-making processes by using algorithms that weigh various options based on their predicted utility. For example, AI systems can simulate decision paths and predict the optimal strategy for resource allocation, thereby supporting strategic management decisions. However, this theory assumes that decision-makers have access to complete information and can evaluate all potential outcomes. AI systems address this by rapidly analyzing large datasets, which might be beyond the capabilities of human decision-makers, providing them with more accurate and timely information [1]. Bounded Rationality: Herbert Simon's theory of Bounded Rationality (1979) asserts that decision-makers

are limited by the cognitive and informational constraints they face. People tend to satisfice, or seek satisfactory solutions, rather than optimize. Agentic AI overcomes these limitations by processing vast amounts of information quickly and without the cognitive constraints that human decision-makers face. This leads to more efficient decision-making, as AI systems can evaluate more variables simultaneously and in less time, providing organizations with more comprehensive insights [2]. The Behavioral Decision Theory: This theory focuses on the psychological factors influencing decision-making, such as cognitive biases and emotions [4]. Agentic AI can help mitigate human biases by relying on data-driven analysis, thus supporting more objective and less error-prone decision-making. For example, AI can identify hidden patterns in consumer behavior that may not be obvious to human managers, ensuring that decisions are based on comprehensive data rather than subjective judgment [1].

3.2 Efficiency Theories

AI, particularly agentic AI, plays a crucial role in enhancing operational efficiency and productivity in organizations. Several efficiency theories can be applied to understand how AI-driven systems impact business operations: The Theory of Productive Efficiency (Hayek 1945): This theory suggests that organizations should aim to minimize inputs while maximizing outputs. Agentic AI contributes to productive efficiency by automating repetitive tasks, reducing labor costs, and eliminating human error. For example, in manufacturing, AI-powered RPA systems can monitor equipment performance in real-time, predict maintenance needs, and adjust production schedules, thereby optimizing resource use and minimizing downtime [1]. Transaction Cost Economics [5]: Transaction cost theory argues that organizations strive to minimize the costs of exchange, such as negotiation, enforcement, and monitoring costs. AI can reduce these costs by automating transactional processes and eliminating the need for extensive human oversight. In industries like banking, AI systems help reduce the costs associated with processing transactions by automating customer verification, document handling, and fraud detection [2].

The Productivity Paradox [6]: The Productivity Paradox suggests that while information technology (IT) investments often lead to lower operational costs, they do not always result in proportional increases in productivity. However, with the advent of AI, this paradox is being challenged. Research indicates that AI can drive significant improvements in productivity by automating complex processes and enabling faster decision-making.

4 Summary of Theoretical Framework

This theoretical framework integrates several theories that explain how agentic AI influences strategic decision-making and organizational efficiency: Agentic AI enhances decision-making by automating data analysis, enabling faster, more informed decisions [1]. Decision-Making Theory offers a lens through which to understand how AI systems improve decision-making, particularly by overcoming human cognitive limitations [3, 4]. Efficiency Theories highlight how AI optimizes productivity, reduces costs, and enhances operational efficiency [5, 6]. These theoretical perspectives underscore the transformative role of agentic AI in modern organizations, particularly in the areas of decision-making, operational efficiency, and strategic management.

5 Methodology

The research adopts a mixed-method approach, combining both qualitative and quantitative methods to gain a comprehensive understanding of how AI-driven RPA impacts strategic management. This dual approach allows for a nuanced exploration of both the measurable improvements in organizational performance and the qualitative insights into decision-making processes.

Research method	Description	Data sources	Key focus
Qualitative analysis	Interviews and case studies to explore decision-makers' experiences with AI-driven RPA	Decision-makers (Managers, IT professionals, RPA vendors)	In-depth perspectives on the practical challenges, benefits, and unintended consequences of AI-driven RPA in strategic management
Quantitative analysis	Surveys and performance metrics to measure the effectiveness of AI-driven RPA on decision-making accuracy and efficiency	Surveys from organizations using AI-driven RPA in strategic management	KPIs such as transaction processing time, cost reduction, and decision-making speed before and after RPA adoption [2]
Mixed-method approach	A combination of both qualitative and quantitative methods to provide a comprehensive understanding of AI-driven RPA's impact	Interviews, case studies, surveys, performance data	Offers a balanced view of both the operational and strategic impact of AI-driven RPA, highlighting its contribution to organizational success

Data Collection

Selection of Organizations Using AI-Driven RPA in Strategic Management Roles:

Data collection component	Description	Criteria/data sources	Key focus
Selection of Organizations Using AI-Driven RPA	Focus on organizations with successful AI-driven RPA integration in strategic management	**Industry**: Healthcare, finance, retail, and manufacturing **RPA Adoption Stage**: Organizations using AI-driven RPA for at least one year	Ensure the study's relevance and depth by selecting organizations that have fully implemented AI-driven RPA systems and have sufficient data for pre- and post-adoption analysis
Surveys from Decision-Makers	Surveys will be conducted with key decision-makers to understand their experiences with AI-driven RPA	CEOs, CFOs, IT Managers	Gather insights on challenges, benefits, and perceived improvements in decision-making accuracy and operational efficiency after RPA implementation
Performance Data Pre- and Post-RPA Adoption	Collection of performance metrics to assess the impact of AI-driven RPA on efficiency and decision-making	**Transaction Processing Time**: Time for strategic tasks such as procurement and reporting **Cost Reduction**: Data on labor and process costs **Decision-Making Speed**: Time for strategic decisions	Evaluate the pre- and post-adoption performance, focusing on improvements in transaction processing, cost reduction, and the speed of strategic decision-making [2]

Data Analysis

Statistical Analysis: The quantitative data will be analyzed using statistical methods to determine the improvements in efficiency and decision-making accuracy. The following statistical techniques will be employed: Descriptive Statistics: Used to summarize the key metrics such as transaction processing times, cost reduction percentages, and decision-making speed before and after AI-driven RPA adoption. Paired Sample T-Tests: A paired sample t-test will be applied to compare the pre- and post-adoption performance metrics. This test will help identify statistically significant improvements in efficiency and decision-making accuracy. Regression Analysis: Regression analysis will be employed to determine the relationship between the adoption of AI-driven RPA and improvements in key performance indicators. This

Table 1 Example data collection framework

Type of data	Source	Purpose	Key metrics
Surveys	Decision-makers (CEOs, CFOs, IT managers)	Assessing strategic benefits of AI-driven RPA	Perceived improvements in decision-making speed, accuracy, and efficiency
Performance Data	Organizational records, RPA systems	Measuring operational performance before and after RPA adoption	Transaction processing times, Cost reduction, Decision-making speed
Case Studies	Specific organizations (industry-specific)	Understanding the organizational impact of AI-driven RPA	Implementation challenges, cultural shifts, strategic decision-making

analysis will allow researchers to assess how RPA adoption impacts organizational performance while controlling for other variables [2].

Qualitative Insights from Case Studies: In addition to the quantitative analysis, qualitative insights will be derived from in-depth case studies. These case studies will provide a deeper understanding of the contextual factors influencing RPA adoption and its impact on strategic management (Table 1).

6 Impact on Strategic Decision-Making

Enhanced Data-Driven Insights: AI-driven RPA systems leverage advanced machine learning algorithms and predictive analytics to transform raw data into actionable insights. By analyzing vast datasets, these systems uncover patterns and trends that are often overlooked by traditional methods. This deep analytical capability aids in strategic decision-making, providing leaders with precise information for critical decisions, such as forecasting demand, optimizing supply chains, and enhancing customer targeting strategies. Moreover, AI systems continuously refine their algorithms based on real-time data, improving the accuracy and timeliness of insights over time. Such capabilities are especially crucial in sectors like healthcare and finance, where timely, data-backed decisions can significantly improve operational efficiency and outcomes. Faster Decision-Making: The integration of AI into RPA allows businesses to automate routine, time-consuming tasks that typically delay decision-making processes. For example, automating tasks like invoice processing, compliance checks, and customer inquiries significantly accelerates decision cycles. This automation not only frees up human resources for more complex and strategic work but also ensures that decisions are made faster and with fewer errors. AI tools, such as those in customer support or legal compliance, can provide immediate responses to situations, facilitating quicker actions in high-pressure environments like disaster management or e-commerce. Scenario Analysis and Predictive

Modeling: AI's ability to simulate multiple scenarios and predict future outcomes based on past data is a powerful tool for strategic management.

7 Organizational Efficiency and Process Optimization

AI-driven RPA systems are capable of operational streamlining by automating complex tasks that traditionally require significant human intervention. This efficiency improvement is evident in key processes such as budgeting, forecasting, and reporting. By eliminating manual errors and automating data collection and processing, these systems not only reduce the time needed for these activities but also ensure higher accuracy in financial projections and organizational analysis [1]. For example, RPA can automate the extraction and consolidation of data from various departments, thus speeding up the financial reporting process and ensuring that decision-makers have timely access to relevant data. The use of AI-powered RPA significantly improves time and resource management by automating tasks that were previously time-consuming and prone to human error. This results in faster processing times and improved accuracy, providing organizations with the ability to make quicker, more informed decisions. The financial impact of automating processes through AI-driven RPA is significant. RPA reduces the dependency on human labor, especially for repetitive tasks such as data entry, invoice processing, and compliance checks, thereby driving down operational costs. Additionally, AI systems mitigate human errors, which can lead to costly mistakes in areas such as financial reporting or compliance management. By reducing such errors, AI-driven RPA systems help organizations avoid costly penalties and operational inefficiencies [2]. For instance, in financial sectors, AI tools automatically flag discrepancies in transaction data, preventing errors that would otherwise require costly manual reconciliation. AI-driven RPA can also optimize resource allocation by improving the speed and accuracy of decision-making, ensuring that resources are used effectively.

8 Challenges and Ethical Considerations

AI-driven systems, including those used in RPA, are not immune to biases, which can significantly impact the fairness of decisions made by these systems. Biases in AI arise when the data used to train machine learning models reflect historical inequalities, incomplete data, or skewed sampling. These biases can manifest in various ways, from discriminatory hiring practices in HR systems to unfair credit scoring in financial services. In the context of strategic management, such biases can lead to flawed decision-making, undermining organizational goals, and creating ethical dilemmas. For instance, AI tools used for investment strategy decisions or consumer targeting might inadvertently perpetuate past inequalities, thus favoring certain demographic groups over others, which could harm both the business's reputation and its ability

to make equitable decisions. This is especially concerning when these AI systems are deployed in critical decision-making areas, such as healthcare resource allocation or financial risk assessment. A lack of transparency and accountability in how decisions are made by AI systems makes it difficult to detect and correct these biases. To mitigate these risks, organizations need to ensure that their AI systems are trained on diverse, representative datasets and subject to rigorous bias audits and fairness checks. This process must also involve human oversight, where biases are identified and actively corrected during the development and deployment phases of AI-driven systems. While AI-driven systems can optimize many aspects of strategic decision-making, human oversight remains crucial to their effectiveness. Human-AI collaboration ensures that AI tools are used appropriately, responsibly, and in line with the organization's strategic goals. AI systems are powerful at processing large datasets and providing insights, but they may lack the nuanced understanding of a situation that human decision-makers can offer. In fields like healthcare, for example, AI can suggest diagnostic options based on data, but the final decision on treatment should still involve human judgment. This collaboration ensures that AI complements human skills rather than replacing them entirely. Moreover, the implementation of AI in organizational decision-making must account for the potential limitations of AI in understanding complex, ethical, or ambiguous contexts. Human oversight helps ensure that AI systems operate within the bounds of ethical guidelines and organizational values.

9 Conclusion

The transformative potential of agentic AI-driven Robotic Process Automation (RPA) for strategic management is undoubtedly profound. As organizations face increasing demands for efficiency, speed, and data-driven decision-making, AI-driven RPA systems offer a way to meet these challenges head-on. By integrating autonomous AI agents with RPA tools, organizations can significantly enhance their decision-making capabilities, particularly in high-stakes or data-intensive environments such as finance, healthcare, and manufacturing. The ability of agentic AI to process large volumes of data, identify patterns, and make real-time decisions enables businesses to act faster and more accurately than ever before. This accelerates strategic decision-making, reduces human error, and allows organizations to respond quickly to market shifts or operational bottlenecks. Furthermore, by automating routine and repetitive tasks, AI-driven RPA helps improve organizational efficiency, freeing up human resources to focus on higher-value, strategic activities. In industries like retail, for instance, AI-powered systems are already optimizing inventory management, reducing stock-outs, and improving demand forecasting, thereby driving efficiency and enhancing customer satisfaction. AI-driven RPA also plays a crucial role in redefining operational frameworks. In conclusion, embracing agentic AI-driven RPA is not just about increasing efficiency or cutting costs—it is about shaping the future of business. By adopting AI-driven systems, organizations can transform how they

operate, enhance decision-making, and ultimately achieve sustainable competitive advantage. Now is the time for businesses to embrace these innovations, adapt to the changing technological landscape, and set the foundation for long-term success in the AI-powered world.

References

1. Acharya DB, Kuppan K, Divya B (2025) Agentic AI: autonomous intelligence for complex goals—a comprehensive survey. IEEE Access. https://doi.org/10.1109/ACCESS.2025.2915921
2. Sharma A (2024) The impact of AI and RPA on organizational efficiency in strategic operations. J Bus Innov 22(3):45–60. https://doi.org/10.1007/JBI2024.567812
3. Simon HA (1976) Administrative behavior: a study of decision-making processes in administrative organizations, 4th edn. Free Press
4. Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47(2):263–291. https://doi.org/10.2307/1914185
5. Williamson OE (1981) The economics of organization: the transaction cost approach. Am J Sociol 87(3):548–577. https://doi.org/10.1086/227496
6. Brynjolfsson E (1993) The productivity paradox of information technology. Commun ACM 36(12):67–77. https://doi.org/10.1145/163298.163309

System for Early Detection of Breast Cancer Using Deep Learning from Medical Imaging



Awa Bakhoum, Sada Anne, and Amadou Dahirou Gueye

1 Introduction

Artificial intelligence (AI) is profoundly transforming the medical field, particularly in computer-aided diagnosis. Thanks to advances in deep learning, AI can now analyze large sets of medical images with greater precision, facilitating the early detection of numerous pathologies. Among these applications, the automatic detection of breast cancer from mammographic and histopathological images represents a crucial challenge for improving patient care and reducing mortality from this disease.

Traditional screening methods, while effective, have certain limitations, notably the high volume of images to be analyzed and the variability of interpretations between radiologists. The integration of deep learning models, in particular convolutional neural networks (CNNs), has shown promise in optimizing image analysis and improving diagnostic accuracy. Several studies have explored these approaches, focusing on classifying benign and malignant lesions, optimizing segmentation methods and improving model performance via normalization and transfer learning techniques.

A number of recent works are in line with this perspective. For example, Sada Anne and al. at the University Alioune Diop of Bambey (UADB) have published an article on the application of artificial intelligence to the detection of stroke risk factors. Their work also highlights the use of AI-assisted medical imaging for faster and more accurate interpretation of symptoms [1]. This work highlights the applicability of AI for medical diagnosis in different clinical contexts.

A. Bakhoum · S. Anne (✉)
TIC4Dev University Alioune Diop of Bambey, Bambey, Senegal
e-mail: sada.anne@uadb.edu.sn

A. D. Gueye
MIM Department, STA/University Amadou Mahtar Mbow, Dakar, Senegal
e-mail: dahirou.gueye@uam.edu.sn

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_16

As part of this research, we propose an innovative solution that goes beyond the simple classification of lesions. Our approach aims to integrate advanced functionalities, including the automatic generation of detailed reports after each prediction, as well as an alert and notification system to inform doctors of significant changes in results. In addition, our platform enables collaboration between healthcare professionals by providing tools for sharing and discussing complex cases. In the remainder of this paper, we organize ourselves as follows: in Sect. 1, we give an overview of existing approaches to AI-based breast cancer detection, in Sect. 2, we focus on the methodology used. In Sect. 3, we highlight the performance achieved and a comparison with classical approaches. Section 4 concludes the paper.

2 Litterature Review

Artificial intelligence (AI) is profoundly transforming the medical field, offering tools capable of improving the accuracy and speed of diagnosis. Thanks to machine-learning algorithms and deep neural networks, it can analyze vast sets of medical data, identify anomalies and assist healthcare professionals in their decision-making. These advances are particularly evident in medical imaging, where AI facilitates the early detection of serious diseases such as cancer, reducing the need for invasive interventions and improving the chances of effective treatment.

The field of automatic breast cancer detection is progressing thanks to deep learning techniques applied to histopathological images. Ramalakshmi et al. [2] and Agna Shaju and Subhija [3] have demonstrated the effectiveness of convolutional neural networks (CNNs) in classifying benign and malignant lesions, notably on the BreaKHis database. Baker et al. [4] compared segmentation methods, such as K-means and Otsu, for a finer analysis of breast tissue. Other approaches have been explored to improve diagnostic accuracy. Wadhwa et al. [5] used DenseNet-201, achieving 95.58% accuracy, while El-Nakeeb et al. [6] combined transfer learning and spot normalization to make analysis more reliable. Halim et al. [7] and Gamal et al. [8] have proposed alternative methods, based on discrete wavelet transforms and breast thermography respectively. In addition, several studies have highlighted the importance of segmentation and transfer learning. Naik et al. [9] developed automated segmentation of glands and nuclei, and Gengtian et al. [10] applied EfficientNet to mammography. Singh et al. [11] have tested hybrid models combining SVM and neural networks, while Ganie et al. [12] have compared different transfer learning architectures in ultrasound imaging. Finally, research has explored the fusion of imaging modalities and model optimization. Venkatasunilsrikanth et al. [13] combined mammography, MRI and ultrasound via CNNs for more accurate detection. Sawant et al. [14] applied logistic regression to early diagnosis, and Gupta et al. [15] and Dandekar et al. [16] integrated advanced techniques such as wavelet transforms and metaheuristics to improve breast cancer prediction.

In summary, this work demonstrates the rapid evolution of AI techniques applied to medical imaging, while highlighting the technical and clinical challenges. Our

approach is distinguished by the development of an interactive platform that goes beyond the simple classification of medical images. In addition to an optimized model based on transfer learning, we offer advanced features such as automatic notification of results, detailed reporting and the ability for doctors to collaborate on complex cases. By combining an intuitive interface and high-performance algorithms, our solution aims to make artificial intelligence more accessible to healthcare professionals, facilitating its integration into medical practices and improving patient care.

3 Methodology

Our study is based on the use of convolutional neural networks (CNNs) for breast cancer detection. The main steps include: data pre-processing: the mammographic images used in this study come from the BreaKHis dataset [17], which contains pathological (benign and malignant mass) and non-pathological cases. To ensure optimum input quality, the following steps were carried out. Data acquisition and formatting: images were collected from several mammography machines, with varying resolutions and contrasts. They were standardized to fit the input dimensions of the models, with a standardized size of 224 × 224 pixels. Normalization: each image was normalized to adjust gray levels, to eliminate variations due to differences in image-taking equipment and ensure data consistency during training. Noise reduction: pre-processing techniques, such as Gaussian filters, were applied to remove the noise present in certain images, thus improving the quality of the contours of detected anomalies. Data augmentation: to enrich training sets and improve model robustness, transformations such as rotations, translations, zooms and contrast variations were performed. This also reduces the risk of overlearning.

Classification architecture: a simple CNN model, optimized with the Adam algorithm and an increasing entropy loss function, distinguishes between benign and malignant images. The CNN structure comprises convolution, pooling and fully connected layers, enabling efficient extraction of image features. The model is then trained on the pre-processed images to learn how to correctly classify anomalies (Fig. 1).

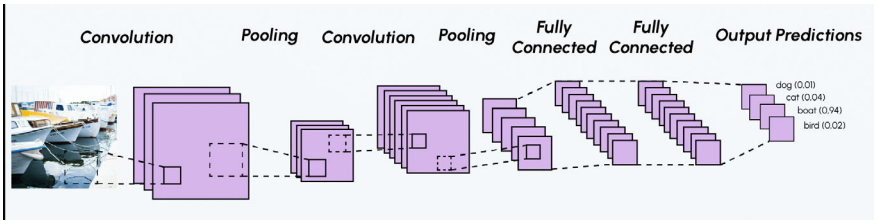


Fig. 1 How a detection CNN works [18]

Suspect images are processed using convolutional neural networks (CNNs) to identify abnormal areas. These networks, thanks to their convolution layers, are able to extract relevant features such as masses and calcifications present in mammograms. They automatically learn from the labeled data, distinguishing healthy from suspect tissue. CNNs are particularly effective at classifying images according to the presence or absence of cancer, enabling early and accurate detection of abnormalities.

The performance of our approach was evaluated using standard metrics adapted to medical image classification. These metrics quantify overall accuracy, consistency of results and the ability of models to correctly identify lesions. Accuracy measures the percentage of images correctly classified within the data set. It provides an overall view of model performance.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{1}$$

where TP (True Positive) represents the number of correctly identified malignant cases, TN (True Negative) represents correctly identified benign cases, FP (False Positive) represents benign cases incorrectly classified as malignant, and FN (False Negative) represents malignant cases incorrectly classified as benign.

Precision evaluates the proportion of truly pathological images among those classified as such by the model, thus reducing false positives.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall quantifies the model's ability to detect all pathological images, minimizing false negatives.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

The F1 score combines precision and recall to provide a balanced metric, useful in cases of class imbalance.

In addition, we used Grad-CAM (Gradient-weighted Class Activation Mapping) to visualize the areas of interest identified by the CNN model. This method improves the interpretability of the results by demonstrating that the model correctly focuses on key regions of the images to make its predictions.

3.1 Early Detection Strategy

Our early detection approach is based on the ability of CNN models to extract and highlight subtle pathological features in histopathological images. Through standardized preprocessing, noise reduction, and enhanced image augmentation, our

system can detect early-stage abnormalities that are often missed by manual examination. Transfer learning further strengthens early feature identification by leveraging knowledge from large-scale datasets.

To better summarize the methodology followed in this study, the overall workflow from data acquisition to result validation is illustrated in Fig. 2.

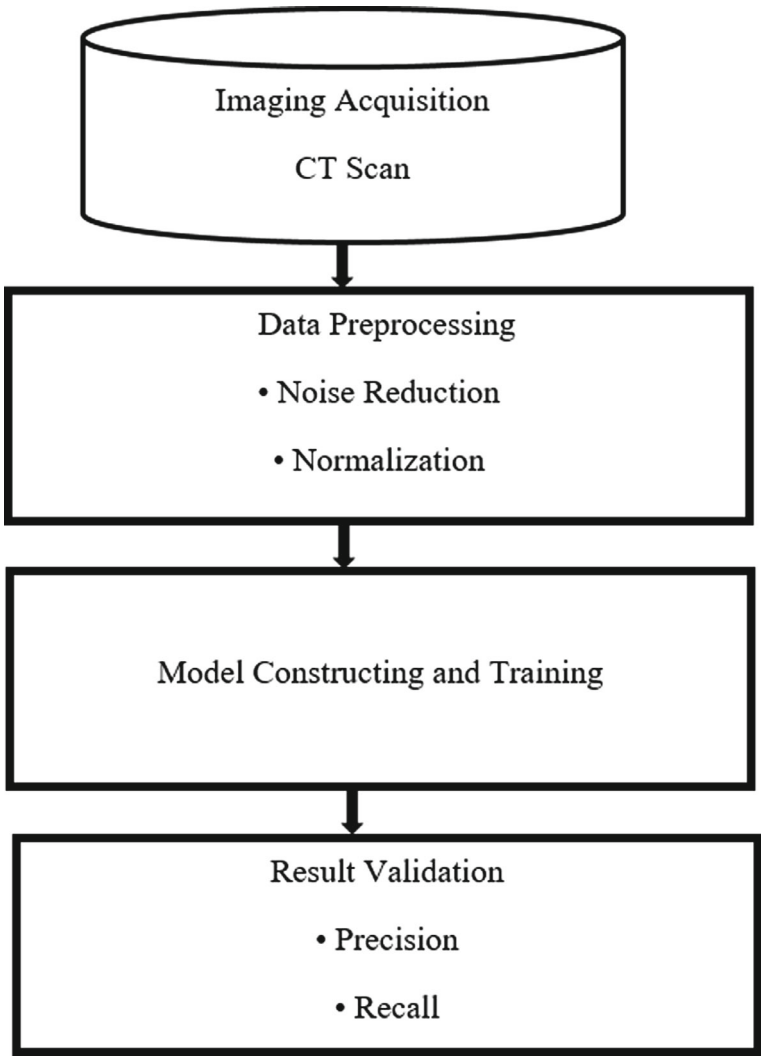


Fig. 2 Workflow illustrating the process from data acquisition to preprocessing, model training, and result validation

3.2 Model Architectures Used

In the following, we discuss classification models for detecting malignant lesions in histopathological images of breast cancer using convolutional neural networks (CNNs). These models focus on distinguishing between benign and malignant images. We start with a simple CNN model, which serves as a basis for evaluating the performance of more complex models. Next, we use VGG16, a model pre-trained on ImageNet, which we exploit via knowledge transfer to improve anomaly detection. And last but not least, we use ResNet50, another pre-trained model featuring residual connections, to improve performance by resolving network depth issues. These models are chosen to maximize accuracy in the detection of malignant anomalies, each bringing specific advantages tailored to the histopathological images in the BreaKHis dataset.

3.2.1 The CNN Model

As part of this study, we have developed a simple CNN model for classifying images from the BreaKHis dataset, as a basis for comparing and evaluating the performance of the more complex models presented later. The model architecture comprises the following elements: Three convolutional layers to extract features, a flattening layer to transform matrices into vectors, a dense layer to perform binary classification.

The first convolutional layer comprises:

- 32 filters with a core size of 3×33.
- Activation function used: ReLU.
- Followed by a max pooling layer with a core size of 2×22.

The second convolutional layer comprises:

- 64 filters with a core size of 3×33.
- Activation function used: ReLU.

Followed by a max pooling layer with a core size of 2×22.
The third convolutional layer comprises:

128 filters with a core size of 3×33.
Activation function used: ReLU.
Followed by a max pooling layer with a core size of 2×22.
The Flatten layer enables: Flatten the resulting matrices to transform them into a one-dimensional vector.
Hidden Dense Layer: 128 fully connected neurons and ReLU activation function.
Dropout layer: Application of a dropout rate of 0.5 to reduce overlearning.
And finally, the Dense output layer: 1 single neuron for binary classification. Activation function used: Sigmoid to produce a probability between 0 and 1 (Fig. 3).

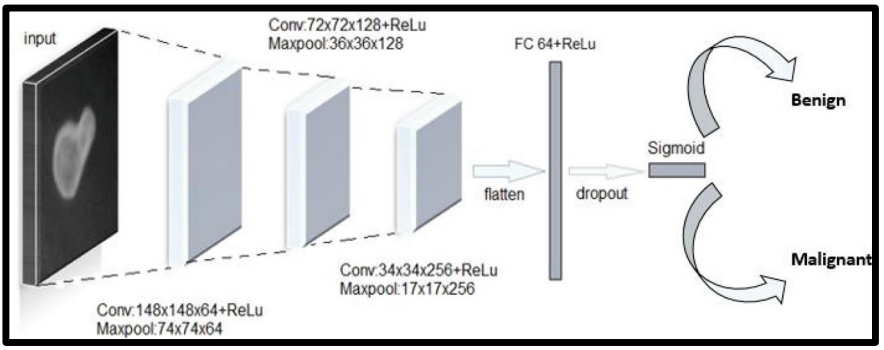


Fig. 3 Architecture CNN B/M (2 Classes) [19]

3.2.2 The VGG16 Model

The architecture of the VGG16 is remarkable for its simplicity and efficiency. It consists of 16 layers of neurons, hence the name VGG16. Here's a brief description of its structure:

Convolutional layers: the model comprises 13 convolutional layers, each using small filters (3×3), enabling efficient capture of details.

Pooling layers: between the convolutional layers, there are 5 pooling layers that serve to reduce the dimensionality of the data, thus reducing the number of parameters and calculations required.

Fully connected layers: At the end, the model integrates 3 fully connected layers that help classify images according to the features learned by the previous layers (Fig. 4).

3.2.3 The ResNet50 Model

ResNet50 is a model developed by Microsoft that introduces the concept of residuality with skip connections (direct connections between non-adjacent layers). Unlike VGG16, it enables the construction of very deep networks without the problem of gradient degradation (Fig. 5).

3.3 Validation Strategy

The dataset was divided into training and validation sets using an 80%-20% split. Validation was performed on unseen data to ensure unbiased evaluation of model performance. Metrics such as Accuracy, Precision, Recall, and F1-score were used. Future work will consider k-fold cross-validation to further strengthen the evaluation process.

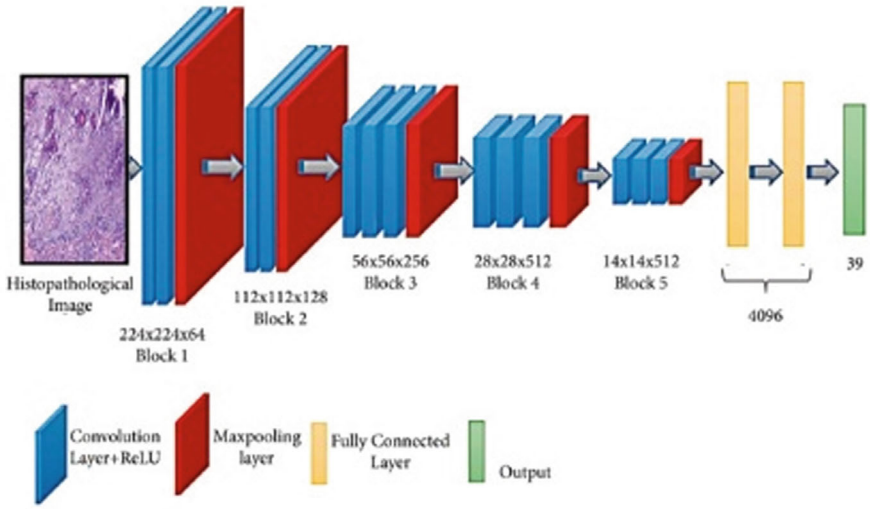


Fig. 4 Architecture of the VGG 16 model [20]

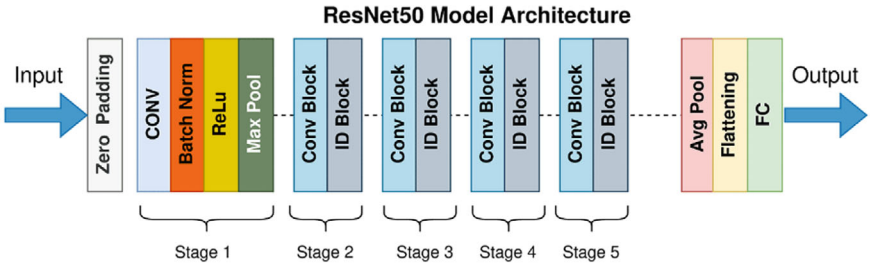


Fig. 5 Architecture of the ResNet50 model [21]

4 Results and Discussion

Detection of malignant lesions from histopathological images of breast cancer using a CNN-based classification model. Here we present We start with a simple CNN model, then exploit VGG16 via knowledge transfer, and finally exploit ResNet50 to improve performance via the remaining connections.

4.1 Model 1 Results (CNN SCRATCH)

Here are the results of the Benign/Malignant classification of the CNN scratch model CNN results before regularization (Fig. 6).

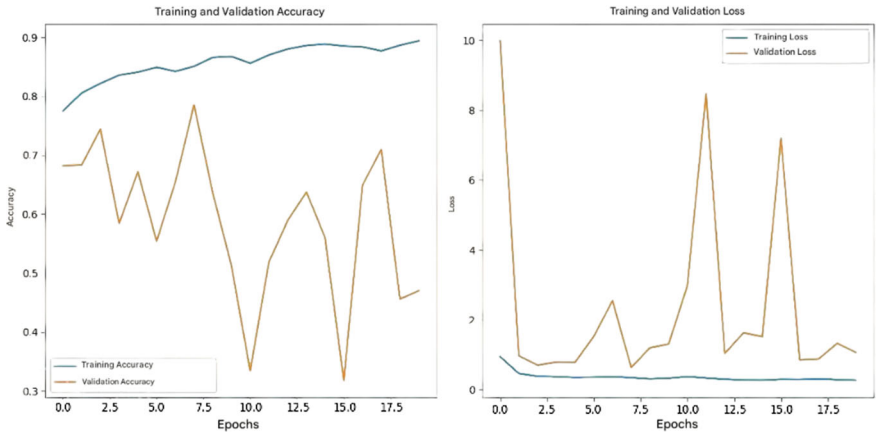


Fig. 6 Graph Loss and Accuracy of the CNN without regularization

The results show that the model achieves high accuracy on the training data (95.11%) with a low loss (0.1171), but its performance decreases on the validation data with an accuracy of 80.89% and a higher loss of 0.4430. This suggests that the model may be over-fitted to the training data, limiting its ability to generalize well on new data.

To enhance the robustness of the simple CNN model, we introduced a Dropout layer after the last dense layer, with a rate of 50%, to reduce overlearning. In addition, we reduced the size of the fully connected (FC) layer to 64 neurons, instead of 128, to simplify the model (Fig. 7).

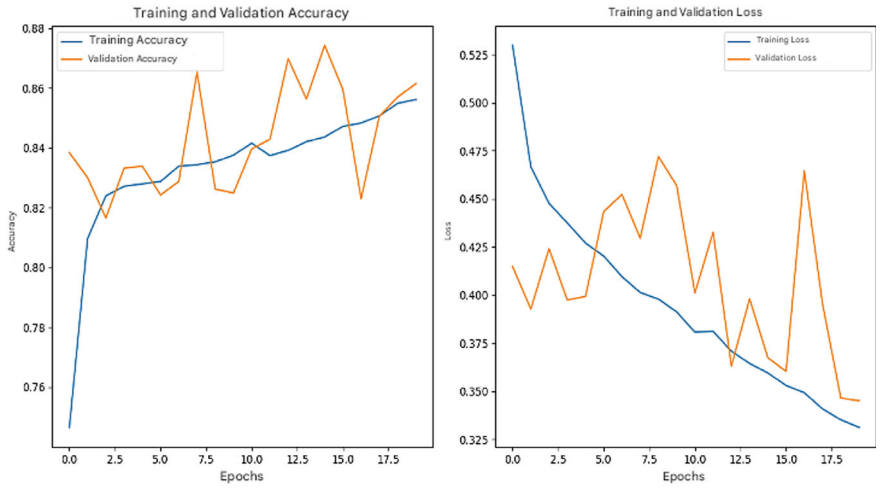


Fig. 7 Graph loss and accuracy with regularization

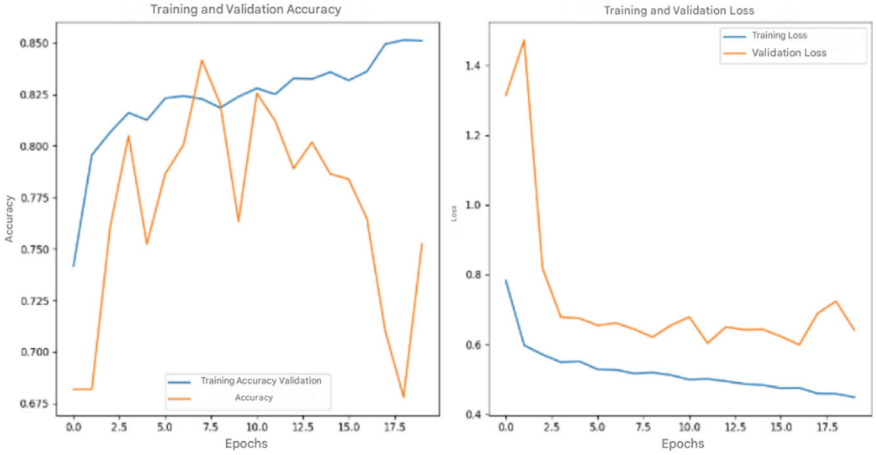


Fig. 8 Graph VGG16 with regularization

The use of the smaller Dropout and FC significantly reduced overfitting. Test accuracy increased from 80.89 to 87.43 with a slight decrease in loss from 0.3716 to 0.3428, indicating that network performance had already been pushed to the limit.

4.2 Model 2 (VGG16) Results

VGG16 results with regularization (Fig. 8).

The VGG16 model achieved reasonably average accuracy on both the training set (82.21%) and the validation set (84.16%), meaning that it is able to distinguish benign from malignant images with good accuracy. The performance gap between training and validation remains moderate, indicating a slight but not too serious overfitting.

Precision was improved through extensive data augmentation (rotation, zoom, translation), careful normalization of input images, and by employing transfer learning with VGG16 and ResNet50 models. Optimization of hyperparameters such as learning rate and batch size also contributed to enhanced performance.

4.3 Model 3 Results (ResNet50)

A dropout is added at the end of the network to reduce overfitting (Fig. 9).

After adding the Dropout regularization, the results show a general improvement in the stability of val_loss and val_accuracy values. Initially, val_accuracy peaked at 0.9565 at epoch 16, indicating that the model generalizes better on the validation data. At the same time, val_loss gradually decreased, reaching 0.5737 at epoch

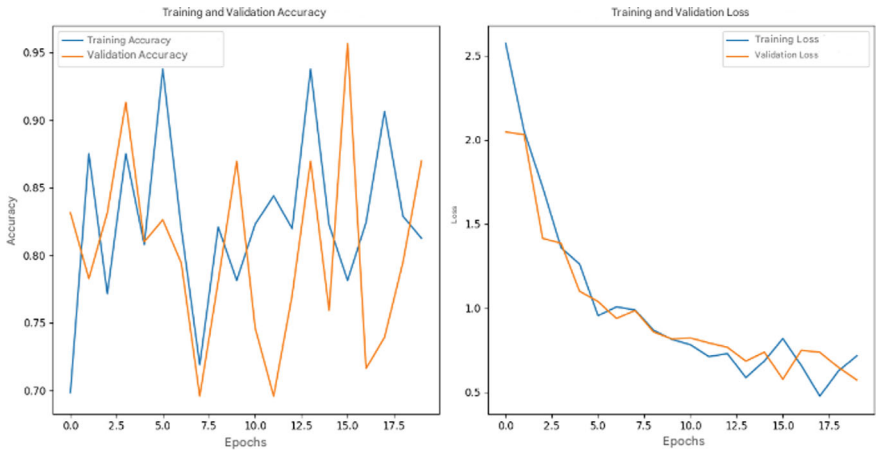


Fig. 9 ResNet50 results with DROPOUT

20, suggesting a reduction in overlearning. This trend indicates that the addition of Dropout has helped to reduce the model's complexity, improving its ability to generalize well on unseen data.

In this previous section, we analyzed the performance of models developed for automatic breast cancer detection across several categories. Below is a comparative table of the different models used (Table 1).

This work presents a significant contribution to the field of automatic breast cancer detection using Deep Learning models. We have evaluated several architectures to obtain a model capable of providing reliable results. The scratch CNN model, after the application of regularization techniques such as dropout and data augmentation, showed remarkable performance. It achieved an accuracy of 87.43% on the test set with reduced loss, making it the best-performing model in our study.

Compared with pre-trained models such as VGG16, which achieved satisfactory results but did not show such a significant improvement after regularization, the CNN model stands out for its ability to generalize better and offer a better balance between accuracy and loss.

Table 1 Comparative results for the three models

Model	Accuracy of training set (%)	Validation set clarification (%)	Precision after re-gularization (%)
CNN Scratch	95.11	80.89	87.43
VGG16	80.61	79.03	84.16
ResNet50	96.88	95.65	95.65

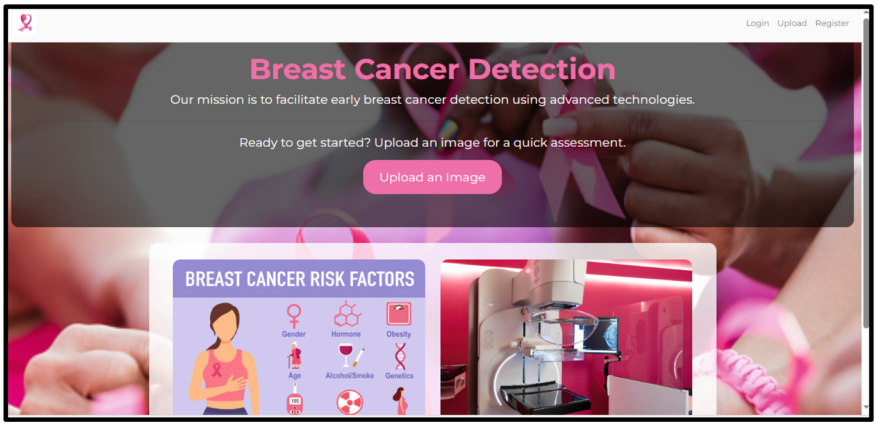


Fig. 10 Home page

4.4 Reliability of the Models

To ensure reliability, multiple models were compared across different architectures. Regularization techniques, such as dropout and data augmentation, helped minimize overfitting. Furthermore, results were evaluated on independent validation sets to assess generalization capabilities, ensuring consistent and dependable predictions.

4.5 Platform

The automation process is implemented through a transparent pipeline: automatic image upload preprocesses classification by viewing messages and generating reports from health professionals. Grad-CAM visualization also helps physicians understand model decision making by highlighting critical image areas (Figs. 10, 11, 12 and 13).

5 Conclusion and Perspectives

This work demonstrates the effectiveness of deep learning-based approaches for analysing lung lesions. By jointly optimising classification and segmentation, we have shown that it is possible to provide accurate tools to support healthcare professionals in their diagnosis. Future improvements may include the adoption of lighter architectures such as MobileNet or EfficientNet, ensemble learning approaches to combine multiple model predictions, and fine-tuning of deeper network layers.

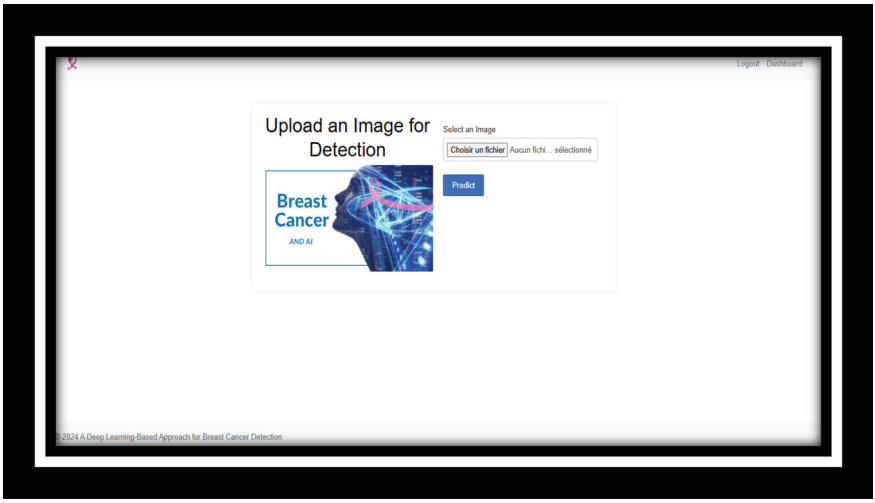


Fig. 11 Download page

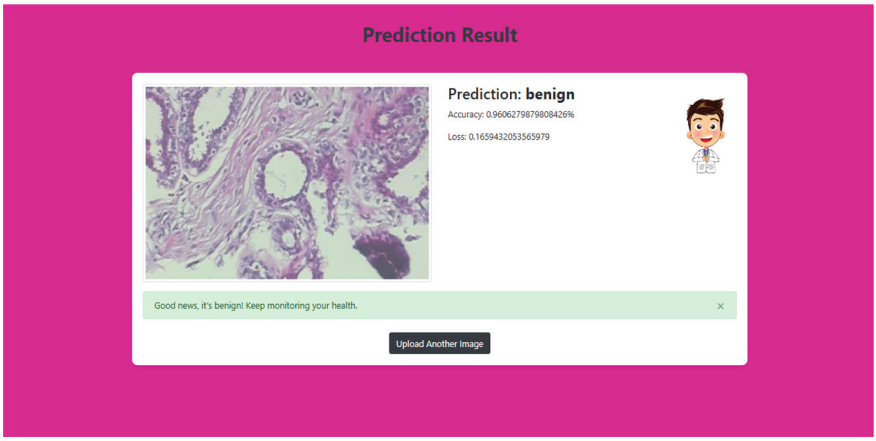


Fig. 12 Results page

Increasing dataset diversity and quality can also significantly enhance model performance and robustness. Prospects include exploring lighter architectures for integration in clinical settings and application to other types of respiratory disease. In addition, the use of transfer or semi-supervised learning techniques could improve results in contexts where annotated data is limited.

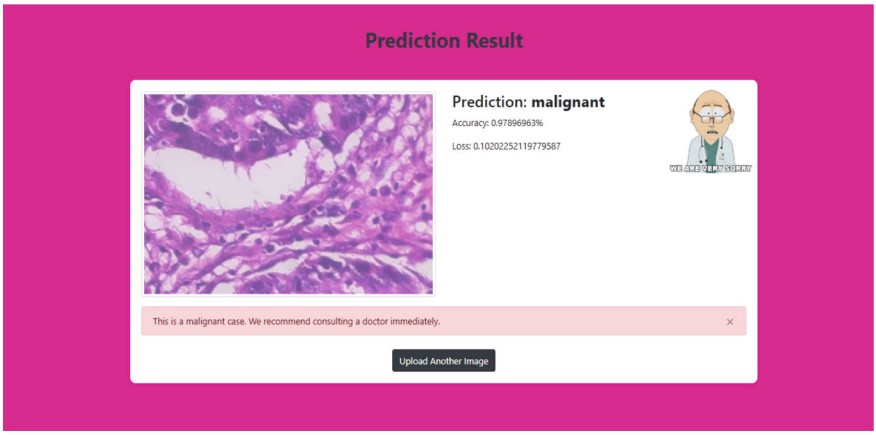


Fig. 12 (continued)

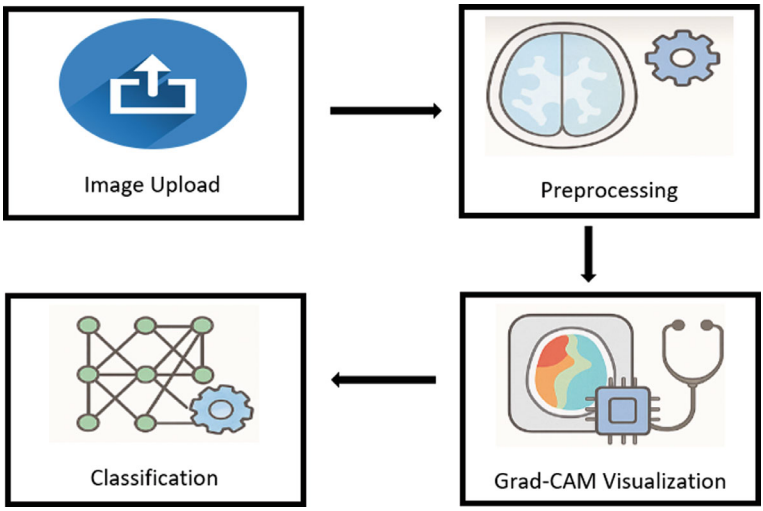


Fig. 13 Results page

References

1. Anne S, Gueye AD (2024) XGBoost algorithm to predict a patient's risk of stroke. In: Seeam A, Ramsurrun V, Juddoo S, Phokeer A (eds) Innovations and ınterdisciplinary solutions for underserved areas. InterSol 2023. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 541. Springer, Cham. https://doi.org/10.1007/978-3-031-51849-2_10
2. Ramalakshmi E, Gunisetti L, Sumalatha L (2023) A review on breast cancer detection for histopathology ımages using deep learning. In: 2023 ınternational conference on artificial

ıntelligence and smart communication (AISC), Greater Noida, India, pp 685–688,https://doi.org/10.1109/AISC56616.2023.10085116
3. Agna Shaju P, Subhuja EN (2022) Detection of breast cancer from histopathological ımages. In: 2022 second ınternational conference on next generation ıntelligent systems (ICNGIS), Kottayam, India, pp 1–5. https://doi.org/10.1109/ICNGIS54955.2022.10079781
4. Baker QB, Abu Qutaish A (20210 Evaluation of histopathological ımages segmentation techniques for breast cancer detection. In: 2021 12th ınternational conference on ınformation and communication systems (ICICS), Valencia, Spain, pp 134–139. https://doi.org/10.1109/ICICS52457.2021.9464594
5. Wadhwa G, Kaur A (2020) A deep CNN technique for detection of breast cancer using histopathology ımages. In: 2020 advanced computing and communication technologies for high performance applications (ACCTHPA), Cochin, India, pp 179–185. https://doi.org/10.1109/ACCTHPA49271.2020.9213192
6. El-Nakeeb M, Ali M, AbdelHadi K, Ahmed Tealab SH, Eltohamy MI, Abdel-Hamid L (2023) Computer-aided breast cancer diagnosis using deep learning: malignancy detection and HER2 score. In: International conference on mobile, ıntelligent and ubiquitous computing (MIUCC), Cairo, Egypt, pp 1–6. https://doi.org/10.1109/MIUCC58832.2023.10278384
7. Halim E, Halim PP, Hebrard M (2018) Artificial ıntelligence models for early breast cancer detection. In: International conference on ınformation management and technology 2018 (ICIMTech), Jakarta, Indonesia, pp 517–521. https://doi.org/10.1109/ICIMTech.2018.8528140
8. Gamal S, Atef H, Youssef D, Ismail T, El-Azab J (2008) Early detection of breast cancer from thermography via pre-entrained deep edge detection with extreme Naik amplification. In: Doyle S, Agner S, Madabhushi A, Feldman M, Tomaszewski J (eds) Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. 2008 5th IEEE ınternational symposium on biomedical ımaging: from nano to macro, Paris, France, pp 284–287. https://doi.org/10.1109/ISBI.2008.4540988
9. Naik S, Doyle S, Agner S, Madabhushi A, Feldman M, Tomaszewski J (2008) Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In: 2008 5th IEEE ınternational symposium on biomedical ımaging: from nano to macro, Paris, France, pp 284–287. https://doi.org/10.1109/ISBI.2008.4540988
10. Gengtian S, Bing B, Guoyou Z (2023) EfficientNet-based deep learning approach for breast cancer detection with mammography ımages. In: 8th ınternational conference on computing and communication systems (ICCCS), Guangzhou, China, pp 972–977. https://doi.org/10.1109/ICCCS57501.2023.10151156
11. Singh P, Nagill J, Saini K (2023) Using supervised learning for breast cancer detection using AI&ML. In: 5th ınternational conference 2023 on advances in computing, communication control and networking (ICAC3N), Greater Noida, India, pp 281–285. https://doi.org/10.1109/ICAC3N60023.2023.10541492
12. Ganie SM, Kanti Dutta Pramanik P (2024) Comparative analysis of transfer learning models for breast cancer detection. In: 2024 IEEE 3rd world conference on applied ıntelligence and computing (AIC), Gwalior, India, pp 830–836. https://doi.org/10.1109/AIC61668.2024.10731032
13. Venkatasunilsrikanth, Krithiga S (2023) CNN deep architecture-based breast cancer detection enhancement for accurate diagnosis. In: Second 2023 ınternational conference on augmented ıntelligence and sustainable systems (ICAISS), Trichy, India, pp 200–205. https://doi.org/10.1109/ICAISS58487.2023.10250616
14. Sawant A, Patil D, Khuman D, Pingle Y, Shinde V (2024) Améliorer la détection du cancer du sein : une approche d'apprentissage automatique pour le diagnostic et la classification précoces, 11e Conférence ınternationale sur l'informatique pour le développement mondial durable (INDIACom) 2024, New Delhi, Inde, pp 235–239. https://doi.org/10.23919/INDIACom61295.2024.10498771
15. Gupta A, Kaushik D, Garg M, Verma A (2020) Machine learning model for breast cancer prediction. In: Fourth 2020 ınternational conference on I-SMAC (IoT in social, mobile, analytics and

cloud) (I-SMAC), Palladam, India, pp 472–477. https://doi.org/10.1109/I-SMAC49090.2020.9243323
16. Dandekar AR, Sharma A, Mishra J (2024) A deep learning based approach and feature optimization for early detection of breast cancer. In: IEEE 2024 ınternational student conference on electricity, electronics and computer science (SCEECS), Bhopal, India, pp 1–7. https://doi.org/10.1109/SCEECS61402.2024.10482104
17. Ambarish (2022) BreakHis Dataset. Kaggle. https://www.kaggle.com/datasets/ambarish/breakhis
18. DataScientest (2025) Convolutional neural network (CNN): understanding the functioning of convolutional neural networks [Online]. Available at: https://datascientest.com/convolutional-neural-network. Accessed 26 Apr 2025
19. Samet S (2020) A Deep learning-based approach for breast cancer detection (Master's thesis, supervised by M. Amroune). https://doi.org/10.13140/RG.2.2.21695.23206
20. Deif MA, Attar H, Amer A, Elhaty IA, Khosravi MR, Solyman AAA (2022) Diagnosis of oral squamous cell carcinoma using deep neural networks and binary particle swarm optimization on histopathological images: an AIoMT approach. J Healthcare Eng Article ID 6364102. https://doi.org/10.1155/2022/6364102
21. OpenCV (n.d.) Computer vision research. OpenCV. Retrieved 26 Apr 2025, from https://opencv.org/blog/computer-vision-research/

Dual-Branch Scene Text Detection for Robust Competition Number Identification in Dynamic Outdoor Environments



Virendra Singh Solanki, Awanit Kumar, Snehlata Ajmera, Nirmal Singh, Sheshang Degadwala, and Dhairya Vyas

1 Introduction

Detecting text in outdoor environments together with identifying numbers proves challenging because these scenes have complex dynamic characteristics. The detection methods employed for text in scenes have difficulty working in open settings due to fluctuating illumination patterns together with obstructed images and complex environmental backgrounds. The detection tasks become especially difficult in sports applications and racing events because moving text such as bib numbers frequently gets obscured by movement and distortion in the images. Such dynamic outdoor environments require better detection methods which deliver both reliable results and accurate readings because current methods are inadequate in these circumstances [1].

V. S. Solanki (✉) · A. Kumar · S. Ajmera · N. Singh
Sangam University, Bhilwara, Rajasthan, India
e-mail: viru_guru01@yahoo.com

A. Kumar
e-mail: itsawanitkumar@gmail.com

S. Ajmera
e-mail: snehlata.ajmera@sangamuniversity.ac.in

N. Singh
e-mail: nirmal.singh@sangamuniversity.ac.in

S. Degadwala
Department of Computer Engineering, Sigma University, Vadodara, Gujarat, India

D. Vyas
The Maharaja Sayajirao University of Baroda, Vadodara, Gujarat, India
e-mail: dhairya.vyas-cse@masubaroda.ac.in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_17

The recognition of numbers in uncertain outdoor spaces continues to challenge modern research due to its inability to resolve detection complexities effectively. Available frameworks mainly operate from static environments with controlled conditions that contrast from realistic field scenarios. The implementation barriers caused by inadequate text recognition accuracy when dealing with obscured or partially hidden text affect actual field usage especially during fast-moving sports competitions. Although deep learning methods especially YOLO reveal potential the method struggles to maintain high recognition accuracy across partially hidden letters or under different lighting scenarios [2].

This scientific work establishes a unique concept that focuses on outdoor number recognition for dynamic environments including outdoor sports and racing competitions. The main goal involves creating a system which detects text content then identifies numbers precisely while operating under difficult conditions. The proposed method utilizes YOLOv8 and MMOCR toolkit to unite their strengths for superior recognition and detection of text in outdoor settings. The research presents a specialized outdoor scenes database called DDBB for training models intended to this purpose [3].

The main contribution of this reserach is introducting YOLOv8 framework connected to the MMOCR toolkit delivered advanced performance levels for recognizing and detecting numbers within diverse outdoor environments. The YOLOv8 model achieved 0.976 mAP@0.5 score at 60 epochs then elevated it to 0.971 mAP@0.5 score at 120 epochs. Both effectiveness and robustness of this combined approach become clear in real-world outdoor testing conditions which provide promising opportunities for number identification in changing conditions.

2 Literature Study

Detection and recognition of real-world scene texts prove difficult across outdoor dynamic settings because text visibility and environmental variations like lighting levels make detection challenging. The field has significantly progressed through deep learning-based approaches but establishing reliable and efficient results across unpredictable situations remains ongoing. Here are key solutions that address these problems compared to our proposed framework design.

The work by [4] implements a context perception parallel decoder which combines multi-scale feature fusion mechanisms to solve obscured and irregular text problems with promising results in moderately complex situations. Multi-perspective feature aggregation strengthens robustness since it helps detection deal with challenging backgrounds that current methods struggle to detect. The text recognition capabilities of Chhabra and Abidi's L YOLOv8 [5] benefit from specialized loss functions; however, the system lacks adequate responsiveness to various outdoor situation changes. Our solution uses improved temporal consistency modules as a direct method to handle changing input conditions. Liu et al. [6] used knowledge distillation methods on diffusion models to enhance low-visibility detection capabilities, yet they

did not explore both rainy and dynamically changed environmental conditions. The method incorporates dynamic environmental conditions into its training phase that directly aligns with its validation stage. Usmani et al. [7] adapted YOLO for traffic sign detection in UAE conditions yet their model struggled with situations involving blocked views and lighting glares. Our system contains a combination of occlusion-aware training methods as well as contrast normalization approaches to handle such challenges effectively. The TBIA-DBNet architecture by Liu et al. [8] implements a dual-branch design for adaptive segmentation in foggy conditions yet does not confirm performance in clear and bright environments. Our system operates effectively across different weather environments without demanding any reconfiguration to suit particular conditions. YOLOv8 with Motion Enhancement by Safaldin et al. [9] delivers swift object detection even Suffers from inherent limitations in detecting scene text and dynamic textual elements. Proposed system integrates text-related feature enhancement with motion detection capabilities.

Yang et al. [10] and the technical survey conducted by Richardson et al. [11] provide two examples of extensive reviews. According to Richardson et al. [11], generalizable and scalable models need development to handle both scale variations and occlusion, especially which is where our research aims to solve these challenges. Liu et al.'s Pyramid Mask Text Detector [12] deals well with irregular text forms, but its performance decreases in disorganized visual environments. Our proposed model is robust against irregular text forms and ambient visual confusion because of its built-in feature-level and semantic-level noise resistance modules. The text detection method by Lee et al. [13] falls short when dealing with curved or deformed text objects, and Deng et al.'s PixelLink method [14] provides instance segmentation yet performs poorly in complex outdoor settings. The system integrates pixel-level prediction results with adaptive region proposal mechanisms to solve the mentioned weaknesses. The detection challenges discussed by Zhou et al. [15] and Kulkarni and Barbadekar [16] address larger consideration. Yet, their fundamental machine-learning patterns resist real-world dynamic changes. The proposed method introduces deep convolutional structures with multi-perspective dynamic attention mechanisms to serve real-world applications better in the future.

3 Methodology

Figure 1 shows the system which first receives input images while simultaneously processing them through two parallel processing streams. The Bib Number Detector powered by YOLOv8 initially identifies the area where the competition bib appears (yielding BB1 as output). The second pathway uses a Text Detection system from the MMOCR toolkit to find textual elements throughout the whole image, resulting in the bounding box BB2 output. The Competition Number Detection module merges the distinct outputs to verify predictions, which leads to identifying the particular competition number assigned to each athlete. The two-channel detection network operates

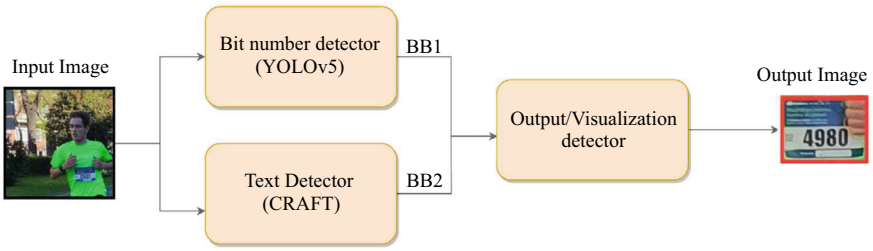


Fig. 1 General block diagram

to achieve precise competition number recognition even when facing outdoor challenges, including busy backgrounds cha, singing light levels object motion, and item blocking.

A dual-detection mechanism represents the innovative aspect of this method because it splits text localization functions from detailed recognition tasks within the system design. The framework operates independently to detect bib regions and text components before their outcome fusion occurs. The verification process prevents incorrect alerting since it confirms that competition numbers can only be found inside or adjacent to bib areas. The separate detection pathways prevent environmental noises, background disturbances, and physical blocking from reducing performance. The modular framework structure enables independent modular upgrades because changes to the bib detector or text detector do not necessitate system retraining [17]. The proposed method enables dynamic outdoor detection through strategic design that uses the combined strengths of YOLOv8 and MMOCR for better scalability and accuracy than traditional methods.

The two-branch design selection proves necessary because it allows proper handling of real-world outdoor scenes that exhibit complex and unpredictable characteristics. YOLOv8 delivers swift accurate detection of objects suitable for outdoor bib identification and MMOCR supplies strong text extraction functionality which enables diverse text recognition. This method simplifies computation by restricting the inspection area for text detection which creates possibilities for both real-time and near real-time execution [18]. The proposed framework demonstrates practical and effective operation for outdoor competition number detection due to its combination of environment-resistant elements and modular structure alongside targeted system optimization.

3.1 Inference Stage

As shown in Fig. 2 Inference phase starts by receiving an input image which is usually obtained from a live event. The trained Model (which integrates YOLOv8 variants together with MMOCR toolkit) receives the supplied image input and produces either

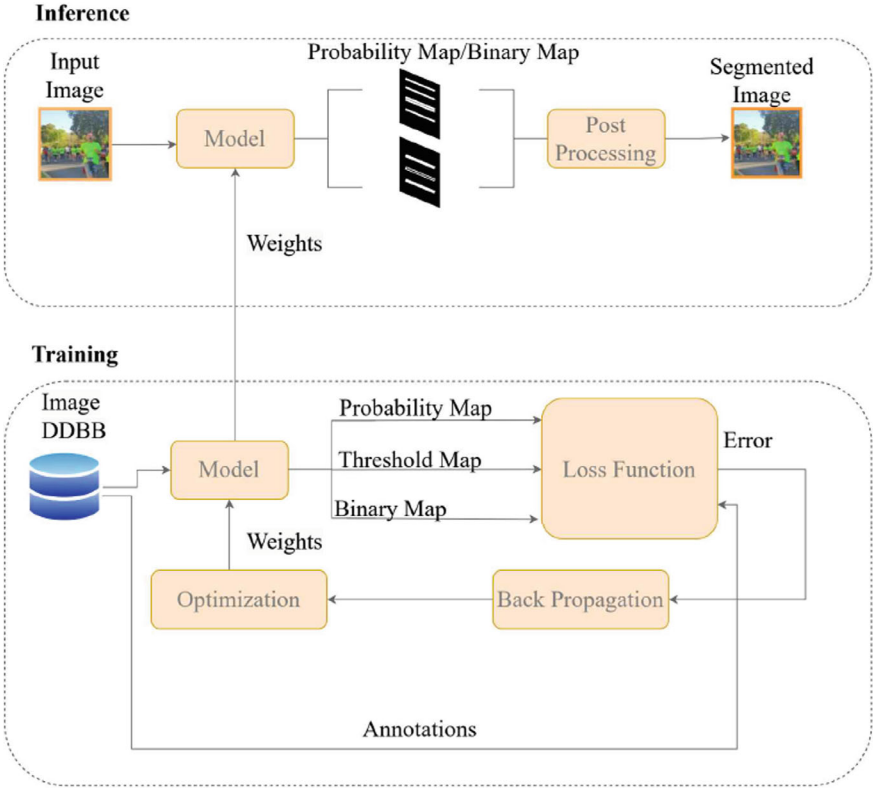


Fig. 2 Model training diagram

Probability Maps or Binary Maps from it. The generated maps identify areas where numbers and text usually appear in images.

Following the model predictions there exists a post-processing phase that improves raw results by eliminating background noise while also merging fragmented characters and adjusting bounding box positions. The system produces a neat Segmented Image through detection of text zones so other processing methods like OCR (Optical Character Recognition) become feasible.

3.2 Training Stage

Training phase the model learns to differentiate between background sounds and genuine numeric text in demanding situations. A specialized outdoor sports database called DDBB (Database) serves for training purposes. The database consists of labeled pictures that depict athletes wearing numbered bibs across different environmental and specific conditions.

The pipeline works as follows:

1. When processing images from the DDBB the model produces Probability Map and Threshold Map while also creating Binary Map output.
2. The Loss Function performs an evaluation of the outputs by generating prediction errors between output maps and ground truth annotations.
3. The processing error moves through the Back Propagation and Optimization modules for weight adjustments that enhance future prediction accuracy.
4. The training process took place across 60 and 120 epochs to determine operational performance stability. Evaluation of multi-object detection precision uses YOLOv8s and YOLOv8m with their NMS-enhanced versions as model variants [19, 20].

4 Results Analysis

This section describe three iterations of the YOLOv8 object detection framework served in the bib number detection system through YOLOv8n (nano) and YOLOv8s (small) and YOLOv8m (medium). The training process included 60 and 120 epochs for each separate model that operated at a fixed image resolution of 640 × 640 pixels. A SGD optimizer ran the training at a learning rate of 0.01 while using batch sizes of 4 and 8 respectively between 60 and 120 epochs for enhanced training stability and acceleration of convergence. A consistent dataset split served as the basis for all experiments so that models could be properly compared against one another. Standard evaluation metrics which included mAP 50 and mAP 50–95 together with precision and recall scores measured performance whereas model complexity was evaluated for its parameter count and FLOPs (floating point operations). Researchers used this configuration to determine how accuracy and computational efficiency related to the different YOLOv8 variants in their experiments.

Figure 3 showcases text and number detection on images featuring a person in a green t-shirt, emphasizing the system's ability to identify text and numerical elements. Figures 4 and 5 compare epoch training times for 60 and 120 epochs, respectively, with the x-axis representing the number of epochs and the y-axis representing the time taken. These graphs highlight the computational efficiency and performance trends of the model, allowing an understanding of how training time varies with an increasing number of epochs.

Table 1 demonstrates performance evaluation metrics of YOLOv8 variants which underwent training for 60 and 120 epochs. The detection accuracy and robustness of YOLOv8m remains superior because it delivers the highest metrics in precision, recall, and mAP during the 60 and 120 epochs training process.



Fig. 3 Text and number detection

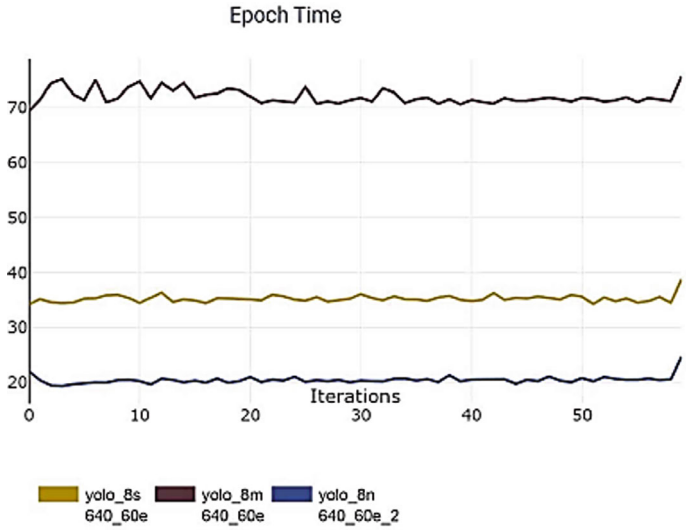


Fig. 4 Epoch time comparision of 60 epoch

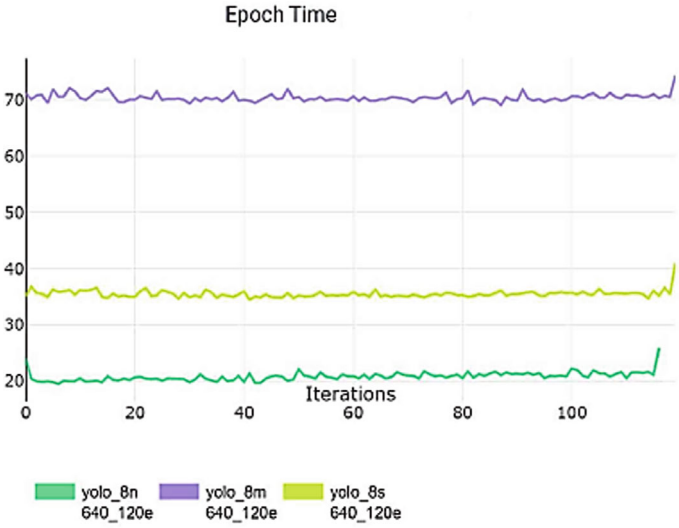


Fig. 5 Epoch time comparision of 1200 epoch

Table 1 Comprative analysis

Model	Epochs	mAP50 (%)	mAP50-95 (%)	Precision (%)	Recall (%)
YOLOv8n	60	91.30	73.10	89.20	88.40
YOLOv8s	60	92.70	76.00	91.00	89.70
YOLOv8m	60	94.20	78.60	92.40	91.10
YOLOv8n	120	92.50	74.80	90.80	89.90
YOLOv8s	120	94.00	77.90	92.20	91.00
YOLOv8m	120	95.40	80.30	93.60	92.80

5 Conclusion

The dual-branch system that uses YOLOv8 variants and the MMOCR toolkit efficiently finds and retrieves competition numbers in challenging outdoor environments where conditions change frequently. A two-step process in the system addresses fundamental challenges caused by varying light intensity, harsh visual conditions, and text that is partly obscured. The DDBB dataset created explicitly for sports applications improves evaluation consistency and benchmarking stability when testing performance. YOLOv8m achieved its peak performance by meeting mAP@0.5:0.95 score of 0.712 and mAP@0.5 score of 0.976 at epoch 60, demonstrating its superior accuracy. The model proves suitable for real-world applications because it maintained its stability over 120 epochs. The development of precise competition number recognition enables recent state-of-the-art detection methods to merge with textual identification systems for challenging outdoor applications. This research

create multiple promising investigation opportunities. To become more effective, the DDBB dataset will achieve better generality by including numerous sports types with varied geographical regions and weather elements. The deployment of lightweight alternative models and pruning strategies will enable their operation when processing constrained edge devices. The text recognition system performs better at identifying deformed and partly hidden numbers when language models and sequence correction methods are integrated into its processes. Future applications, including sports analytics, emergency response tracking, and live race monitoring, become possible because of active video stream operation. The adjustable framework structure enables the sustained evolution of capabilities to meet current requirements because individual components operate independently for independent modifications.

References

1. Liu Y, Zhao Y, Chen Y, Hu Z, Xia M (2024) YOLOv5ST: a lightweight and fast scene text detector. Comput Mater Continua 79:909–926. https://doi.org/10.32604/cmc.2024.047901
2. Liao M, Zou Z, Wan Z, Yao C, Bai X (2023) Real-time scene text detection with differentiable binarization and adaptive scale fusion. IEEE Trans Pattern Anal Mach Intell 919–931. https://doi.org/10.1109/TPAMI.2022.3155612
3. Raisi Z, Zelek J (2022) Text detection & recognition in the wild for robot localization. arXiv preprint arXiv:2205.08565
4. Du Y, Chen Z, Jia C, Yin X, Li C, Du Y, Jiang Y-G (2023) Context perception parallel decoder for scene text recognition. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2025.3545453
5. Chhabra P, Abidi AI (2025) Curved text detection and recognition in natural scenic images based on L-Yolov8 object detection algorithm. Signal Image Video Process 19:291. https://doi.org/10.1007/s11760-024-03806-w
6. Liu Z, Zhou G, Jia Z, Shi F (2025) Scene text detection in foggy weather utilizing knowledge distillation of diffusion models. IEEE Signal Process Lett. https://doi.org/10.1109/LSP.2025.3540371
7. Usmani MA, Mahmood S, Elmadany Y, Azeem AZ, Zualkernan I (2025) Hierarchical YOLO with real-time text recognition for UAE traffic signs. Authorea Preprints
8. Liu Z, Zhou G, He R, Zhang M, Jia Z, Ma J (2025) TBIA-DBNet: a two-branch image-adaptive DBNet for scene text detection in real-world foggy scenes. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, pp 208–221. https://doi.org/10.1007/978-3-031-78119-3_15
9. Safaldin M, Zaghden N, Mejdoub M (2024) An improved YOLOv8 to detect moving objects. IEEE Access 12:59782–59806. https://doi.org/10.1109/ACCESS.2024.3393835
10. Keserwani P, Dhankhar A, Saini R, Roy PP (2021) Quadbox: quadrilateral bounding box based scene text detection using vector regression. IEEE Access 9:36802–36818. https://doi.org/10.1109/ACCESS.2021.3063030
11. Raisi Z, Naiel MA, Fieguth P, Wardell S, Zelek J (2021) Challenges of deep learning-based text detection in the wild. J Comput Vis Imaging Syst 6:1–5. https://doi.org/10.15353/jcvis.v6i1.3543
12. Yang L, Ergu D, Cai Y, Liu F, Ma B (2021) A review of natural scene text detection methods. Proc Comput Sci 199:1458–1465. https://doi.org/10.1016/j.procs.2022.01.185
13. Richardson E, Azar Y, Avioz O, Geron N, Ronen T, Avraham Z, Shapiro S (2020) It's all about the scale—efficient text detection using adaptive scaling. In: Proceedings—2020 IEEE Winter

conference on applications of computer vision, WACV 2020, pp 1833–1842. https://doi.org/10.1109/WACV45572.2020.9093534
14. Raisi Z, Naiel MA, Fieguth P, Wardell S, Zelek J (2020) Text detection and recognition in the wild: a review. arXiv preprint arXiv:2006.04305
15. Liu J, Liu X, Sheng J, Liang D, Li X, Liu Q (2019) Pyramid mask text detector. arXiv preprint arXiv:1903.11800
16. Lee S, Kwak S, Cho M (2019) Universal bounding box regression and its applications. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, Heidelberg, pp 373–387. https://doi.org/10.1007/978-3-030-20876-9_24
17. Kateb FA, Monowar MM, Hamid MA, Ohi AQ, Mridha MF (2021) FruitDet: attentive feature aggregation for real-time fruit detection in orchards. Agronomy 11:2440. https://doi.org/10.3390/agronomy11122440
18. Raisi Z, Naiel MA, Younes G, Wardell S, Zelek JS (2021) Transformer-based text detection in the wild. In: IEEE computer society conference on computer vision and pattern recognition workshops, pp 3156–3165. https://doi.org/10.1109/CVPRW53098.2021.00353
19. Mishra S (2021) Baffling concept of true positive and true negative. https://Towardsdatascience.Com/Baffling-Concept-of-True-Positive-and-True-Negative-Bffbc340F107
20. Yang C, Chen M, Yuan Y, Wang Q (2021) MT: multi-perspective feature learning network for scene text detection. arXiv preprint arXiv:2105.05455

Deep Learning Based Model for Landslide Detection from Synthetic Aperture Radar Imaging



Manu Gupta, Yallamati Prakasa Rao, Manideep Shivanathula, Smahi Bandari, Arundhathi Uppari, and P. Michael Preetam Raj

1 Introduction

The landslides are one of the more dangerous natural hazards and can cause death destruction to infrastructure, and economic losses across the globe. Landslides are prevalent in hillsides that have extreme to extremely heavy rainfall such as those in the Himalayas, Western Ghats, and Eastern Ghats [1, 2]. The Geological Survey of India in a report said about 12.6 per cent of the country's landmass, excluding snow-covered region, is susceptible for landslide. Over 20,000 landslide-related deaths have been reported across India in the last two decades, with the Northeastern state, Uttarakhand and Himachal Pradesh being most affected. According to a study by the National Institute of Disaster Management, landslides account for approximately 5 per cent of total disaster-related deaths in India each year [3].

Landslide detection has historically been carried out using human surveying, Geographic Information System (GIS) maps based on GIS and remote sensing techniques using the power of both Synthetic Aperture Radar as well as optical imaging [4, 5]. While these techniques can be useful however, they are not efficient with regard to cost, time, and capability to operate in real-time. Furthermore, there is it is also important to note that the Normalized Difference Vegetation Index (NDVI) that

M. Gupta (✉) · M. Shivanathula · S. Bandari · A. Uppari
Department of Electronics and Computer Engineering, Sreenidhi Institute of Science and Technology, Hyderabad, India
e-mail: manugupta5416@gmail.com

Y. P. Rao
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vadeswaram, Guntur, India

P. Michael Preetam Raj
Department of Computer Science and Engineering, Bharat Institute of Engineering and Technology, Hyderabad, India

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_18

is used in optical methods is also a sensitive indicator of the amount of cloud cover and seasonal changes and can result in unbalanced results.

Since these methods have drawbacks, deep learning-based methods are becoming more popular for landslide detection. High-resolution satellite imagery has yielded considerable improvements in classification accuracy using convolutional neural networks and residual learning architectures. In this study a hybrid model using U-Net and Efficient-B net deep learning framework to achieve significant accuracy in classifying landslide-prone regions is developed. This proposed method provides early detection and vulnerability assessment, thereby assisting effective disaster management and minimizing landslide effects.

2 Literature Survey

In this section an overview of existing approaches for landslide detection and landslide classification using satellite observations are discussed.

Cheng et al. [2] provides overview of various deep learning models for landslide identification. This work shows that robustness and enormous potential of deep learning methods for landslide detection by providing deeper network architecture. Martinis and Wieland [3] developed a model using CNN that learns while conducting flood mapping based on satellite images taken from TerraSAR-X, which are a type of radar.

Mansourifar et al. [4] performed a survey of GAN-based techniques for satellite imaging, highlighting their applications, including data augmentation, synthetic image generation, and image enhancement. The study emphasized that conditional GANs (cGANs) yielded significantly better results compared to traditional neural networks for semantic segmentation tasks, which is helpful for landslide mapping. Kadu et al. [5] used CNN and polygonal neural networks for predicting landslides through satellite images. The results demonstrate the efficacy of this approach in delivering precise and timely forecasts.

The research work presented by Chen et al. [6] focused on the detection of landslide-prone areas on the basis of remote sensing data, integrated with machine learning models. Their results emphasized that SVM- based models outperform traditional GIS-based approaches by leveraging a wider array of environmental and meteorological parameters. Dakhore et al. [7] proposed a hybrid model using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for landslide detection. The authors concluded that hybrid architectures constitute a scalable solution for landslide monitoring.

Schlemper et al. [8] proposed a method for emphasizing significant regions in medical images, called Attention-Gated Neural Network (AGNet). The models learn to highlight specific features in satellite images, yielding better detections of landslide locations due to the application of attention mechanisms. Zhou et al. [9] proposed UNet++, a variant of U-Net model which builds nested and dense skip pathways to the original model and improves segmentation performance.

Qi et al. [10] developed the ResU-Net, incorporating residual connections into the U-Net framework to facilitate gradient flow during training. Their experiments showed that this architecture effectively addresses the vanishing gradient problem common in deep networks while maintaining the strong feature extraction capabilities of the U-Net for landslide mapping. Fang et al. [11] integrated convolutional neural networks with conventional machine learning classifiers for landslide susceptibility mapping. This hybrid approach combined the feature extraction capabilities of deep networks with the interpretability advantages of traditional classifiers.

Unsupervised change detection methods was applied by Mondini et al. [12] on Sentinel-1 SAR amplitude imagery to automatically detect landslide events. Rather than training on labeled data, they derived the model indicative of new landslides by assessing both before and after backscatter intensity changes associated with the landslide events. Sharma et al. [13] applied modified ResNet-101 deep learning model for landslide prediction. This model extracts features using hidden network layers. A Generative Adversarial Network (GAN) were used in work presented by Mansourifar et al. [4] for satellite image classification.

The majority of the works for landslide detection, according to the literature review, depend on the availability of high-resolution satellite images, which may not be readily available for regions. Furthermore, clouds also block optical images, which decreases the models efficiency and limits its capacity to operate on other visual data. In proposed work Landslide detection was performed using Synthetic Aperture Radar (SAR) images which are robust to weather conditions and provides continuous monitoring. SAR does not use optical sensors like optical satellites and can penetrate clouds and vegetation to detect even small movements on the ground using microwaves.

3 Proposed Method

In proposed work, a model for Landslide detection based on SAR multispectral images is developed. The model utilized deep learning based semantic segmentation techniques to precisely find the landslide impacted regions in the multi-spectral satellite SAR images.

The flow diagram of proposed landslide detection model is demonstrated in Fig. 1. Firstly, the multispectral satellite image data is acquired from Kaggle Landslide4Sense dataset [14]. The dataset comprises of 6448 images with corresponding ground truth annotation showing landslide areas. Next, the image data is pre-processed and data augmentation is carried out. The deep learning methods-U-Net and Efficient-B Model are utilized for model training and feature extraction. These obtained features are then concatenated and used for landslide prediction and segmentation. Finally, the effectiveness of the proposed model is evaluated for Landslide detection and areas with landslide are segmented.

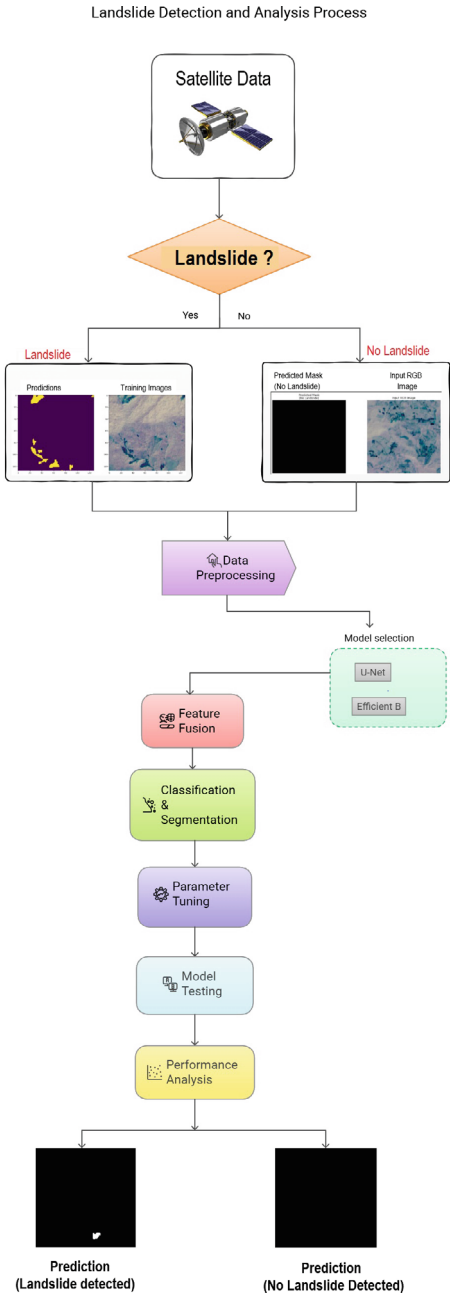


Fig. 1 Proposed landslide detection flow diagram using SAR images

3.1 Dataset Acquisition and Pre-processing

The image data used in proposed work obtained from Landlide4sense dataset has been acquired from diverse landslide affected areas around the world. The images are stored in HDF5 (.h5) format, a flexible data model suitable for handling large, multidimensional datasets. Each image tile is standardized to a resolution of 128 by 128 pixels, ensuring uniformity for model input. For every image, a corresponding binary mask is available. These masks highlight landslide zones using designated pixel values and serve as reference labels for model training. The images are labelled into two categories-class label 0 for non-landslide and class label 1 for Landslide [14].

The dataset captures critical environmental and topographic features commonly associated with landslide activity. Variables such as vegetation stress, soil moisture distribution, and slope orientation are embedded within the input data layers. Especially important are spectral details from near-infrared and shortwave infrared bands, which re-veal vegetation health and moisture variability two essential indicators for landslide risk.

Images are pre-processed for photometric correction (removing ground interference), atmospheric correction (effect of gravity in distances between satellites and earth adjustment of images to geo-reference coordinates), and radiometric calibration to ensure all are measuring the same thing. Pixel value normalization is performed to normalize the incoming information range, often from 0 to 1. The proposed landslide detection shown in Fig. 1.

3.2 Deep Learning Methods Applied

In the proposed work, U-Net and Efficient-B Net deep learning models are used for segmentation of landslides as well as extracting features to aid in the classification of landslides.

i. **U-Net Model**: The U-Net model is an encoder-decoder symmetric architecture that has skip connections between the matching layers for encoders and decoders as shown in Fig. 2. The encoder path slowly reduces the spatial dimension of its capture, while simultaneously capturing more abstract elements using convolutional layers. The decoder path thereafter up-samples to recover spatial resolution [4]. For landslide detection, U-Net is best at maintaining spatial information without losing hierarchical features. The skip connections are especially useful for landslide detection since they retain high-resolution topographical information that would otherwise be lost during down-sampling. In comparison to standard CNNs, U-Net generally has 10–15 percent greater accuracy in landslide segmentation tasks by preserving these important boundary details. The fact that the model can effectively utilize relatively small training sets (hundreds as

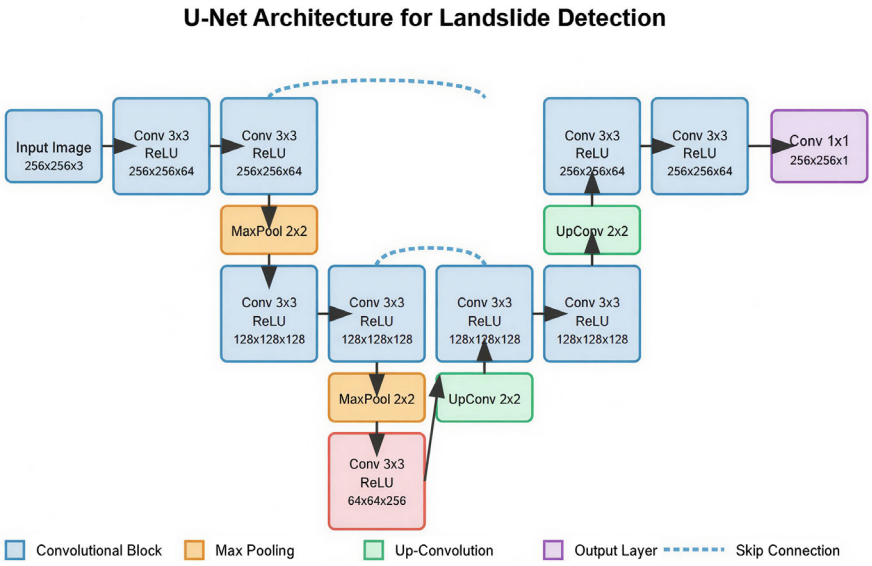


Fig. 2 U-net architecture

opposed to thousands of images) makes it especially well- suited for specialized remote sensing applications [15].

ii. **Efficient-B Model**: The Efficient-B model belongs to the Efficient Net architecture family, known for achieving high accuracy with low computational cost shown in Fig. 3. It employs a compound-scaling technique that adjusts the network's size, depth and resolution of input with a balanced approach. This lets the model extract the most detailed spatial information in satellite images while remaining lightweight and fast [16]. For landslide detection, Efficient-B is effective in recognizing both small-scale terrain textures and broader environmental patterns. Its use of MBConv layers and attention mechanisms helps capture subtle signals in varying image quality. This model provides a great equilibrium between speed and performance and is ideal to monitor real-time in real time and for remote sensing applications in complex landscapes [17].

Residual learning, introduced through residual blocks in CNN architectures, enables the model to learn identity mappings by using shortcut connections. These connections allow gradients to flow directly through earlier layers during backpropagation, effectively reducing the risk of vanishing gradients. This is especially important in high-dimensional satellite imagery, where deep models are needed to extract complex spatial features. The residual architecture ensures deeper models can be trained without degradation in performance, improving accuracy in detecting subtle landslide patterns.

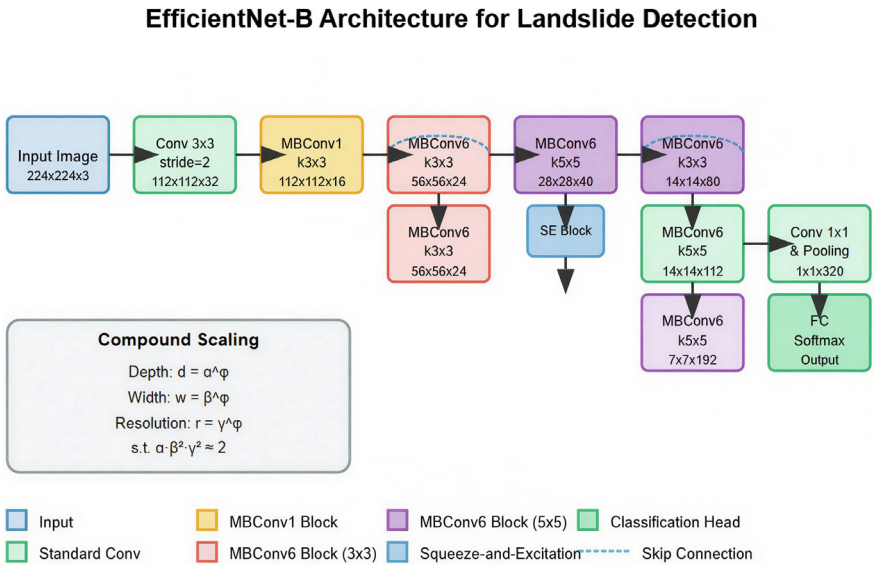


Fig. 3 Efficient-B model architecture

3.3 Performance Analysis

The effectiveness of models that have been trained to predict landslides is assessed using precision, accuracy recall, and F1-scores which are calculated according to the following:

$$\text{Accuracy} = \frac{TP + FP}{TP + FP + TN + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1 - score} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

where, TP refers to True Positive, TN refers to True Negative, FP refers to False Positive and FN refers to false Negative [18, 19]. Positive corresponds to Landslide cases and Negative corresponds to No-Landside cases.

The robustness and generalizability of the model is attained by several steps. First, data augmentation simulated varied terrain and lighting conditions. Second, the use of CNN based architectures helped the model focus on key features while

ignoring noise. Third, integrating different spectral bands and elevation data allowed the model to learn more comprehensive terrain characteristics.

4 Results Analysis

In this section, the findings from the model proposed to predict Landslide prediction are presented.

4.1 Experimental Results

The image from the dataset are loaded to the model. A sample image representing the RGB, Normalized Difference Vegetation Index (NDVI), slope, elevation layers and binary mask marking the landslide-prone areas are illustrated in Fig. 4. The model addresses class imbalance through a combination of loss function tuning and data strategies. The dataset is also split carefully to ensure both classes are sufficiently represented in training, validation, and testing. These approaches prevent the model from becoming biased toward predicting only the majority class.

The results are validated using a combination of quantitative metrics and visual inspection methods. The model's predictions are validated by comparing the predicted results against ground truth mask provided in the dataset. The results obtained from the proposed model for images having Landslide and No-Landslide activity, aligning with ground truth label are presented in Figs. 5 and 6 respectively. The input image on the right provides visual context from satellite data.

The values obtained for various performance metrics are summarized in Fig. 7. As observed the proposed method using deep learning methods attained accuracy of 97.21%. The variation of f1 score and loss while the training process progresses is shown in Fig. 8. The proposed model integrates U-Net with Efficient-B Net as the encoder backbone, leveraging the complementary strengths of both architectures to achieve superior performance across all evaluation metrics. This combination is particularly effective for image segmentation and pixel-wise classification tasks, where both high-level semantic understanding and low-level spatial accuracy are critical. The deep learning models leverage convolutional operations to capture the

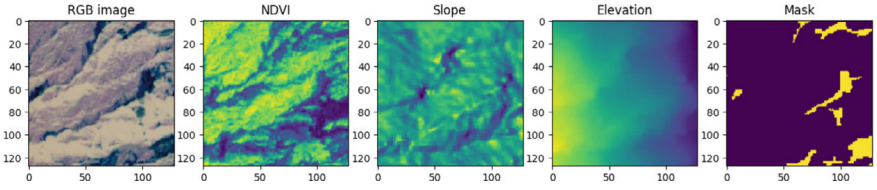


Fig. 4 Sample image from dataset

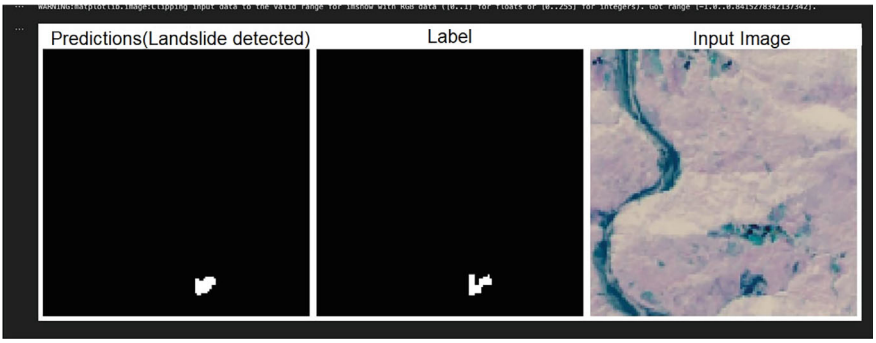


Fig. 5 Result obtained for image having landslide prone region

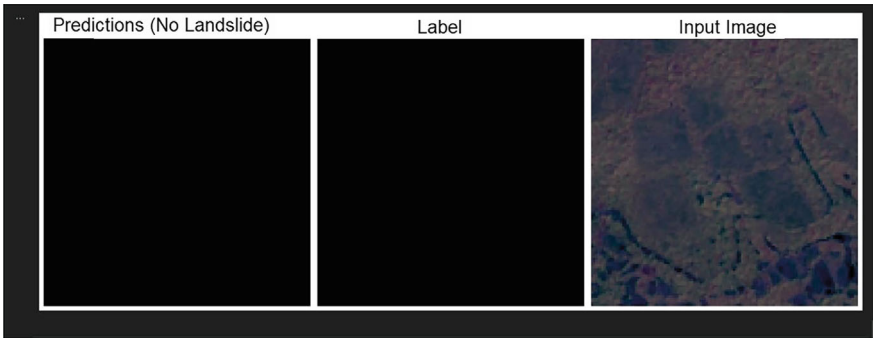


Fig. 6 Result obtained for image having no-landslide prone region

spatial context within the earth observation datasets, improving their ability to learn the complicated patterns related to landslide events.

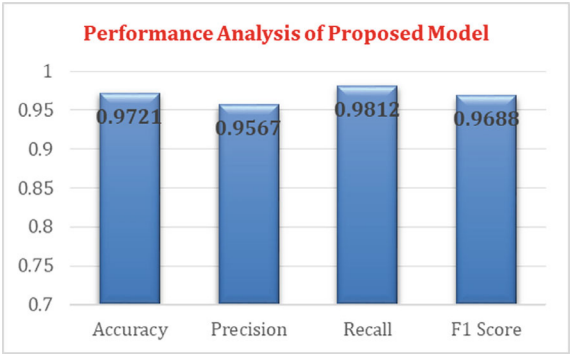


Fig. 7 Performance analysis of proposed model

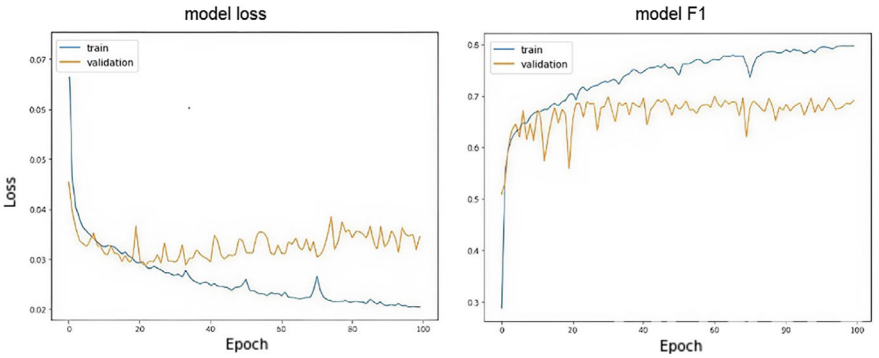


Fig. 8 Loss and F1 score variation graph

4.2 Comparison of Proposed Model Against Existing Work

The efficiency of the model proposed for prediction of landslides is assessed in conjunction with existing works from literature [4, 5, 7, 13] and summarized in Table 1. The proposed approach, utilizing a hybrid architecture of U-Net and Efficient-B Net, demonstrates superior performance across all key evaluation metrics, reaching precision of 0.97 and a precision of 0.95 and a recall of 0.98 and the F1 score that is 0.96. These numbers indicate a balanced model with excellent sensitivity and accuracy which makes it well designed to Landslide predictions. Figure 7 represents the performance analysis of proposed model.

Figure 8 shows the loss and F1 score variation graph. The fusion of Efficient-B Net and U-Net features in the proposed hybrid model provides efficient and deep features capturing complex patterns in input data. The U-Net model can capture both the high-level information (via an encoder) and precise localization (via the decoder). The Efficient-B Net model can extract more specific characteristics from images than traditional CNNs because of the improved convolutional blocks and Swish activation.

Table 1 Comparison of proposed model with existing works

S.No.	Method	Model used	Accuracy	Precision	Recall	F1 Score
1.	Proposed Work	U-Net + Efficient-B Net	0.97	0.95	0.98	0.96
2.	Sharma et al. [13]	Modified ResNet-101	0.9688	0.93	0.94	0.935
3.	Dahore and Kalasakar [7]	CNN + LSTM	0.94	0.89	0.95	0.91
4.	Mansourifar et al. [4]	GAN	0.92	0.90	0.91	0.905
5.	Kadu et al. [5]	CNN + Polygonal NN	0.93	0.91	0.92	0.915

Model performance and accuracy can be enhanced further by adopting architectures tailored to spatial feature preservation. The use of encoder-decoder structures like U-Net and its variants (e.g., U-Net++, ResU-Net) allows for effective segmentation while retaining spatial context. Detection can be made more robust by incorporating multi-spectral and multi-temporal satellite data. Combining optical data with digital elevation models (DEM) and SAR improves terrain understanding. Precision can be increased by incorporating contextual information into the input data helps the model better distinguish landslides from visually similar terrain features.

5 Conclusion and Future Scope

In proposed work a deep learning architecture for automated landslide detection from multi-spectral satellite imagery is presented. This method integrated the U-Net and Efficient-B Net architecture and effectively captures both high-level contextual features and fine spatial details from input image for landslide prediction. Experimental results demonstrate that the proposed model performs efficiently and provides accuracy of 97%. This work presents a reliable model for landslide detection, which can support disaster management and environmental assessment efforts. Future work may explore adding attention mechanisms which would enable models to pay attention to key terrain features, and transfer learning from bigger geospatial datasets could enhance generalization. For operational use, uncertainty quantification would give decision-makers useful measures of confidence, and model optimization could make edge deployment possible in sparsely connected rural regions.

References

1. Chawla A, Pasupuleti S, Chawla S, Rao ACS, Sarkar K, Dwivedi R (2019) Landslide susceptibility zonation mapping: a case study from Darjeeling District, Eastern Himalayas, India. J Indian Soc Remote Sens 47:497–511
2. Cheng G, Wang Z, Huang C, Yang Y, Hu J, Yan X, Tan Y, Liao L, Zhou X, Li Y, Hussain S (2024) Advances in deep learning recognition of landslides based on remote sensing images. Remote Sens 16(10):1787
3. Li Y, Martinis S, Wieland M (2019) Urban flood mapping with an active self-learning convolutional neural network based on TerraSAR-X intensity and interferometric coherence. ISPRS J Photogramm Remote Sens 152:178–191
4. Mansourifar H, Moskovitz A, Klingensmith B, Mintas D, Simske SJ (2022) Gan-based satellite imaging: a survey on techniques and applications. IEEE Access 10:118123–118140
5. Kadu A, Mishra R, Shirsath V (2024) Predicting landslides through satellite imagery analysis and machine learning. Intell Syst Appl Eng 12(21):974–985
6. Cheng G, Guo L, Zhao T, Han J, Li H, Fang J (2013) Automatic landslide detection from remote-sensing imagery using a scene classification method based on BoVW and pLSA. Int J Remote Sens 34(1):45–59
7. Dakhore K, Kalasakar A (2024) Nationwide landslide prediction in India using neural networks and multi source satellite data. Int J Multidisc Res (IJFMR) 6(6):1–12

8. Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, Rueckert D (2019) Attention gated networks: learning to leverage salient regions in medical images. Med Image Anal 53:197–207
9. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J (2018) Unet++: a nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, 20 Sept 2018, proceedings, vol 4. Springer International Publishing, pp 3–11
10. Qi W, Wei M, Yang W, Xu C, Ma C (2020) Automatic mapping of landslides by the ResU-Net. Remote Sens 12(15):2487
11. Fang Z, Wang Y, Peng L, Hong H (2020) Integration of convolutional neural network and conventional machine learning classifiers for landslide susceptibility mapping. Comput Geosci 139:104470
12. Mondini AC, Santangelo M, Rocchetti M, Rossetto E, Manconi A, Monserrat O (2019) Sentinel-1 SAR amplitude imagery for rapid landslide detection. Remote Sens 11(7):760
13. Sharma A, Chopra SR, Sapate SG, Arora K, Rahmani MKI, Jha S, Ahmad S, Ahmed ME, Abdeljaber HA, Nazeer J (2024) Artificial intelligence techniques for landslides prediction using satellite imagery. IEEE Access
14. Ghorbanzadeh O, Xu Y, Zhao H, Wang J, Zhong Y, Zhao D, Zang Q, Wang S, Zhang F, Shi Y, Zhu XX (2022) The outcome of the 2022 landslide4sense competition: advanced landslide detection from multisource satellite imagery. IEEE J Selected Topics Appl Earth Observations Remote Sens 15:9927–9942
15. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
16. Tan M, Le Q (2019) Efficientnet: rethinking model scaling for convolutional neural networks. In: International conference on machine learning. PMLR, pp 6105–6114
17. Tan M, Le Q (2021) Efficientnetv2: smaller models and faster training. In International conference on machine learning. PMLR, pp 10096–10106
18. Gupta M, Meghana P, Reddy KH (2024) Original research article deep learning-based approach for prediction of brain stroke from MR images for IOT in healthcare. J Autonomous Intell 7(3)
19. Gupta M, Shirisha G, Chandana AM, Ankitha B (2023) Brain tumor classification using MR images and transfer learning. In: IEEE 2nd international conference on edge computing and applications, pp 925–932

A Comprehensive Study of CMOS Inverter Power Consumption with Transistor Width and Temperature Variations Using Cadence Virtuoso



Biswazit Mistry and Suman Chowdhury

1 Introduction

The rapid evolution of integrated circuit (IC) technology continues to be propelled by the pursuit of greater performance, improved energy efficiency, and enhanced reliability. In this research paper Cadence Virtuoso is used for the design and analysis. The Complementary Metal–Oxide–Semiconductor (CMOS) inverter, which forms the foundational element for a multitude of digital systems, including microprocessors, memory modules, and embedded platforms [1]. The widespread utilization of CMOS technology stems from its notable advantages, such as high noise immunity, minimal static power dissipation, ease of fabrication, and scalability. However, as device dimensions' shrink and system complexity increases, power consumption has become a critical constraint-particularly for applications that demand both low power operation and effective thermal management [2, 3].

Among the various design parameters influencing CMOS inverter performance, transistor width and operating temperature are especially significant, as they directly impact both dynamic and static power dissipation [4]. While considerable research has focused on optimizing power through methods such as threshold voltage scaling, leakage reduction, and switching speed improvements, there remains a lack of detailed, isolated analysis regarding the independent effects of NMOS and PMOS transistor width variations on power consumption [5]. Most existing studies tend to examine aggregate device behavior or use generalized models, often overlooking the

B. Mistry · S. Chowdhury (✉)
Department of Electrical and Electronic Engineering, International University of Business Agriculture and Technology, Dhaka, Bangladesh
e-mail: suman@iubat.edu

B. Mistry
e-mail: 22105044@iubat.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_19

intricate interplay between individual transistor sizing and thermal effects. Additionally, the influence of temperature on power consumption is frequently discussed in broad terms, with limited exploration of how specific transistor geometries respond to thermal changes [6, 7]. This lack of granularity limits the ability of designers to make precise decisions in transistor sizing and thermal management, especially in modern contexts where compact form factors and fluctuating environmental conditions are prevalent.

A significant limitation in current literature is the shortage of systematic investigations that distinctly isolate the effects of PMOS and NMOS width variations on power consumption, particularly under varying temperature conditions [8]. The absence of high-resolution simulation data addressing these isolated variables presents a notable research gap, hindering the advancement of optimized, energy-efficient, and thermally robust CMOS circuit designs. This gap becomes even more pronounced in emerging fields such as the Internet of Things (IoT), wearable electronics, and mobile computing, where stringent power and thermal constraints are paramount [9].

CMOS power consumption is crucial in IC design because it directly impacts several critical aspects of chip functionality and performance. Minimizing power dissipation is vital for achieving high energy efficiency, managing thermal issues, ensuring reliable operation, and reducing overall system costs [10, 11].

To address these deficiencies, the present study utilizes the Cadence Virtuoso Electronic Design Automation (EDA) platform to conduct a comprehensive analysis of CMOS inverter power consumption. Four scenarios are systematically examined: varying PMOS width with constant NMOS width, varying NMOS width with constant PMOS width, and analyzing the impact of temperature on power dissipation for both narrow (120 nm) and wide (520 nm) transistor configurations. This approach enables a clear differentiation of the individual contributions of NMOS and PMOS transistors to overall power consumption and clarifies the interaction between transistor geometry and operating temperature.

CMOS power consumption is crucial in IC design because it directly impacts several critical aspects of chip functionality and performance [12]. Minimizing power dissipation is vital for achieving high energy efficiency, managing thermal issues, ensuring reliable operation, and reducing overall system costs.

The results of this research provide high-resolution insights into the complex, nonlinear behaviors that govern CMOS inverter power efficiency. By addressing the gaps in prior work, this study delivers valuable guidance for VLSI designers aiming to optimize performance without sacrificing energy efficiency, particularly in thermally constrained environments. The findings are expected to inform future advancements in low-power digital design, supporting the development of robust, efficient, and adaptable integrated circuits for next-generation electronic systems.

2 CMOS Model

CMOS (Complementary Metal–Oxide–Semiconductor) technology is one of the most widely used processes in semiconductor fabrication, forming the foundation of today's integrated circuits. It is central to the design of systems such as microprocessors, memory units, and digital logic components [13]. The CMOS model provides a structured way to describe how CMOS transistors behave, both mathematically and physically, allowing engineers to simulate and predict circuit performance, power usage, and reliability [14].

At the core of CMOS technology are NMOS and PMOS transistors, which function in a complementary manner when one is turned on, the other is off is shown in Fig. 1. This switching mechanism significantly reduces power consumption, particularly in idle states, and enhances noise immunity. To accurately capture device behavior, the CMOS model includes key electrical parameters like threshold voltage, channel length modulation, carrier mobility degradation, and subthreshold leakage.

CMOS modeling spans several complexity levels. Basic models use simplified equations suitable for hand calculations and early-stage designs, while more advanced models, such as BSIM (Berkeley Short-channel IGFET Model), are implemented in simulators like SPICE to capture detailed physical effects [15, 16]. These advanced models are especially important in nanometer-scale technologies, where short-channel effects and velocity saturation play a larger role in device behavior.

In modern circuit design workflows, CMOS models are often refined using measurement data from fabricated silicon chips to improve accuracy [14]. This ensures the models can account for real-world variations in manufacturing and environmental conditions. As transistors continue to shrink in size, precise modeling

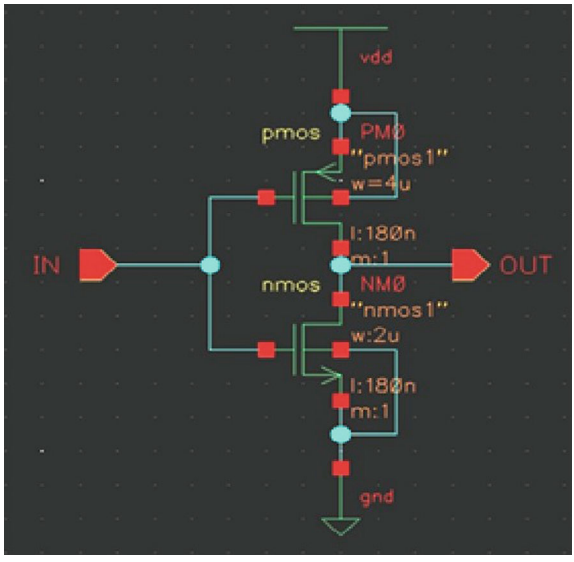


Fig. 1 Schematic of CMOS inverter [17]

becomes even more critical for ensuring circuit performance, yield, and reliability. Ultimately, CMOS models play a key role in enabling the continued advancement of semiconductor technology.

3 Results and Discussion

Figure 2 presents the power consumption profile of PMOS and NMOS transistors, as well as the combined total power, in a CMOS inverter circuit where the PMOS transistor width is varied, and the NMOS width is kept constant. The x-axis denotes the range of PMOS widths, spanning from 120 to 520 nm, while the y-axis indicates the corresponding power consumption measured in watts (W). Each bar in the chart represents the power level associated with a specific PMOS width.

A consistent upward trend is observed in PMOS power consumption as the transistor width increases. This behavior is expected due to the larger gate capacitance and elevated dynamic power dissipation associated with wider transistors. In contrast, the NMOS power consumption remains relatively stable across all width variations, confirming that the NMOS device dimensions were maintained throughout the analysis.

The total power consumption, representing the sum of both PMOS and NMOS power, follows a similar increasing trend. This indicates that the rising PMOS power significantly influences the overall power profile of the inverter circuit under the

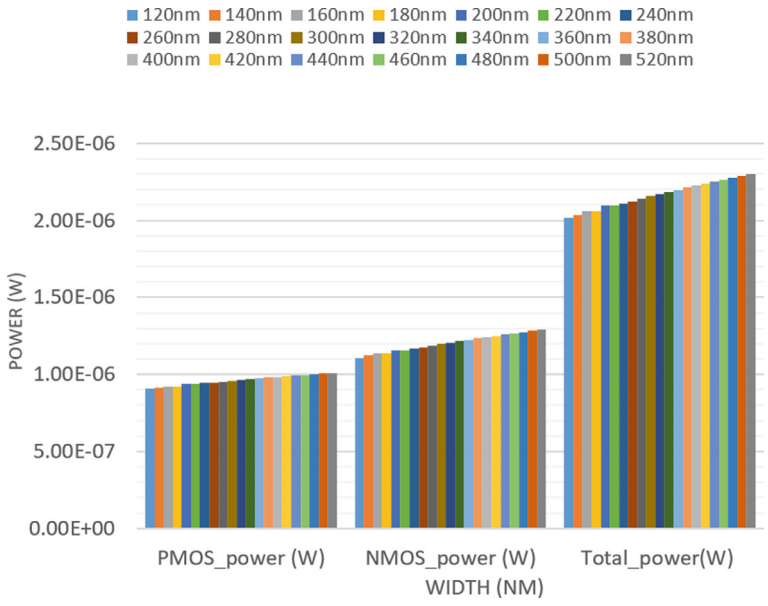


Fig. 2 Power profile with various PMOS width, keeping NMOS width constant

given conditions. These findings emphasize the critical role of transistor sizing in CMOS design. While wider PMOS transistors can enhance switching performance and improve signal symmetry by balancing drive strengths, this benefit comes with a trade-off in power efficiency. Hence, careful optimization of PMOS width is essential, particularly in power-sensitive applications, to ensure a balance between performance and energy consumption in modern integrated circuits.

Figure 3 illustrates the variation in power consumption of PMOS and NMOS transistors, as well as the total power dissipation, in response to changing NMOS transistor widths, while the PMOS width remains constant. The x-axis denotes the NMOS transistor width, ranging from 120 to 520 nm, and the y-axis represents the corresponding power consumption in watts (W). Each bar reflects the power value associated with a specific NMOS width.

As observed, the power consumption of the NMOS transistor increases steadily with its width. This trend is primarily attributed to the increase in gate capacitance and drive current as the channel width grows, leading to higher dynamic and leakage power components. In contrast, the power consumption of the PMOS transistor remains relatively constant across all data points, which aligns with the experimental condition of maintaining a fixed PMOS width during the simulations.

The total power consumption curve, which aggregates the individual contributions of PMOS and NMOS transistors, shows a similar increasing pattern as NMOS width increases. This reflects the dominant influence of NMOS scaling on overall power behavior when the PMOS device parameters are unchanged.

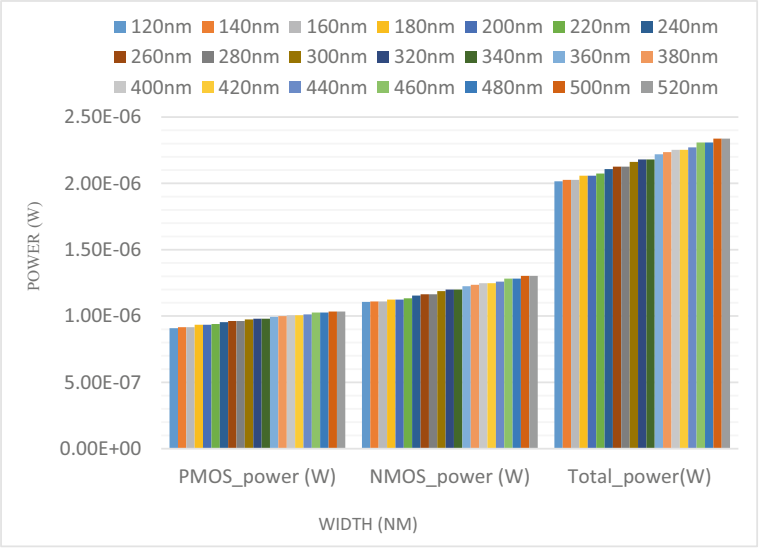


Fig. 3 Power profile with various NMOS width, keeping PMOS width constant

These findings reinforce the importance of careful transistor sizing in CMOS circuit design. While increasing the width of NMOS transistors may offer performance benefits such as improved switching speed and reduced resistance, it also results in higher power dissipation. Consequently, a balance must be maintained between performance optimization and power efficiency, especially in low-power or battery-constrained applications. This analysis contributes to a deeper understanding of how individual transistor dimensions impact the power-performance trade-off in modern integrated circuit design.

Figure 4 displays the power consumption characteristics of NMOS and PMOS transistors as a function of temperature, with both transistor widths maintained at 120 nm. The horizontal axis represents temperature in degrees Celsius, spanning from 10 to 80 °C in increments of 5 °C. The vertical axis indicates power dissipation in watts, presented in scientific notation to accurately reflect the small magnitudes involved.

As temperature increases, both NMOS and PMOS devices exhibit a gradual rise in power consumption. This upward trend is consistent across the entire temperature range. Notably, the NMOS devices (shown in orange) consistently consume more power than the PMOS devices (shown in blue) at every recorded temperature. The difference in power dissipation between NMOS and PMOS transistors remains uniform across all measured points.

At the lowest temperature (10 °C), both device types register their minimum power consumption. With each incremental rise in temperature, a corresponding increase in power is observed for both NMOS and PMOS transistors. The NMOS devices demonstrate a slightly more pronounced sensitivity to temperature changes, as evidenced by a steeper increase in their power profile compared to the PMOS devices. The highest power dissipation for both transistor types occurs at the maximum measured

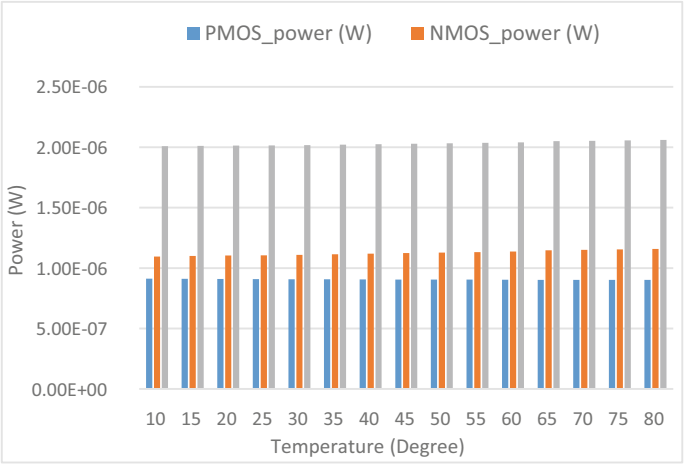


Fig. 4 Power profile with various temperature, keeping both NMOS and PMOS width at 120 nm

temperature (80 °C), with NMOS values remaining consistently higher than those of PMOS throughout the range.

This observed behavior aligns with established MOSFET characteristics, where elevated temperatures typically result in increased leakage currents and reduced carrier mobility, both of which contribute to higher power dissipation. The consistently greater power consumption in NMOS devices is likely due to their intrinsic electrical properties, such as higher electron mobility, which can lead to increased leakage under identical geometric and biasing conditions.

In summary, the graph demonstrates that both NMOS and PMOS transistors experience a clear increase in power consumption as temperature rises, with NMOS devices consistently dissipating more power than PMOS at all temperatures when both are set to a width of 120 nm. These findings highlight the importance of considering thermal effects and device selection in integrated circuit design, especially as device dimensions continue to scale and operational environments become more demanding.

The graph presented in Fig. 5 illustrates the power consumption profiles of NMOS and PMOS transistors as a function of temperature, with both devices fabricated at a fixed channel width of 520 nm. The temperature range spans from 10 to 80 °C in 5 °C increments, providing a detailed overview of how thermal variations influence power dissipation in these devices.

Power dissipation values, shown on the vertical axis in scientific notation (watts), reveal clear distinctions between the two types of transistors. Throughout the entire

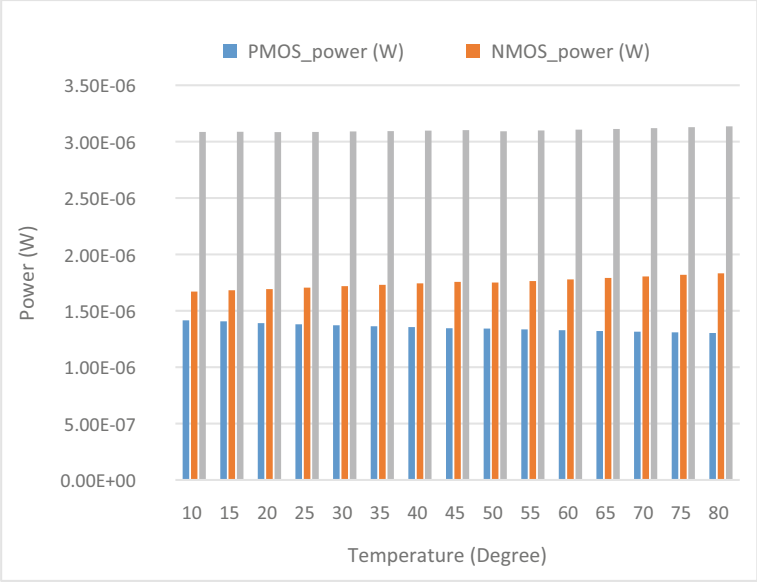


Fig. 5 Power profile with various temperature, keeping both NMOS and PMOS width at 520 nm

temperature range, NMOS transistors (depicted by orange bars) consistently demonstrate higher power consumption than their PMOS counterparts (blue bars). Specifically, NMOS devices exhibit power values ranging from approximately 1.6×10^{-6} W to 1.8×10^{-6} W, while PMOS devices consume slightly less, with values between approximately 1.3×10^{-6} W and 1.5×10^{-6} W.

Both NMOS and PMOS transistors display a gradual increase in power consumption as temperature rises. This trend is consistent with established semiconductor behavior, where elevated temperatures typically lead to increased carrier generation, higher sub-threshold leakage, and changes in carrier mobility-all factors that contribute to greater power dissipation. The observed increase in power consumption appears nearly linear across the measured temperature range, indicating a steady sensitivity of both device types to thermal effects.

The difference in power consumption between NMOS and PMOS devices remains relatively stable across all temperature points, with NMOS transistors consuming approximately 20–25% more power than PMOS devices at any given temperature. This persistent disparity reflects the inherent differences in carrier transport properties, as NMOS devices generally benefit from higher electron mobility compared to the hole mobility in PMOS devices.

Additionally, the gray reference bars at around 3.1×10^{-6} W serve as a visual benchmark, highlighting that both NMOS and PMOS devices operate well below this threshold across all tested temperatures. This observation suggests that, even under elevated thermal conditions, the devices maintain efficient power management-a critical consideration for integrated circuit design and reliability.

In summary, the data presented in Fig. 5 provide valuable insights into the thermal behavior of complementary MOS devices at the 520 nm technology node. These findings are particularly relevant for designers seeking to optimize power efficiency in analog and digital circuits, especially in applications where devices are expected to operate across a broad range of temperatures.

4 Conclusion

This research delivers a comprehensive evaluation of CMOS inverter power consumption as influenced by transistor width and temperature, leveraging high-resolution simulations via Cadence Virtuoso. The results demonstrate that both NMOS and PMOS transistor width scaling independently contribute to increased power dissipation, with wider devices exhibiting higher dynamic and static power due to enhanced gate capacitance and leakage currents. Notably, NMOS transistors consistently consume more power than their PMOS counterparts across all width and temperature scenarios, underscoring the asymmetric nature of power behavior in complementary devices. Additionally, the study reveals a clear and consistent increase in power consumption for both transistor types as temperature rises, a trend attributable to elevated leakage currents and reduced carrier mobility at higher thermal conditions. These findings highlight the compounded impact of geometric

and thermal factors on inverter efficiency, emphasizing the necessity for judicious transistor sizing and robust thermal management in modern integrated circuit design. The insights gained are particularly relevant for applications with stringent power and thermal constraints, such as IoT and wearable devices, where optimizing energy efficiency and reliability is paramount. By systematically isolating the effects of NMOS and PMOS width variations and examining temperature dependencies, this work addresses critical gaps in the literature and provides actionable guidance for VLSI designers. The outcomes support the development of more energy-efficient and thermally resilient digital systems, facilitating advancements in low-power, high-performance electronic applications.

References

1. Rose GS, Shawkat MSA, Foshie AZ, Murray JJ, Adnan MM (2021) A system design perspective on neuromorphic computer processors. Neuromorphic Comput Eng 1(2):022001
2. Monteiro E, Ramos A, Rouboa A (2024) Fundamental designs of gasification plants for combined heat and power. Renew Sustain Energy Rev 196:114305
3. ElGhanam EA, Osman AH, Hassan MS, Assaf T, Mir H (2022) Integration of deterministic and game-based energy consumption scheduling for demand side management in isolated microgrids. Int J Emerg Electr Power Syst 23(4):543–564
4. De Moraes LB, Zimpeck AL, Minhardt C, da Luz Reis RA (2021) Finfet inverter designs: behavior and challenges of process variability. J Integr Circuits Syst 16(2):1–13
5. Rai MK, Gupta A, Rai S (2022) Comparative analysis & study of various leakage reduction techniques for short channel devices in junctionless transistors: a review and perspective. Silicon 14(9):4423–4445
6. Datta S, Chakraborty W, Radosavljevic M (2022) Toward attojoule switching energy in logic transistors. Science 378(6621):733–740
7. Das S, Sebastian A, Pop E, McClellan CJ, Franklin AD, Grasser T, Singh R et al (2021) Transistors based on two-dimensional materials for future integrated circuits. Nature Electronics 4(11):786–799
8. Mahapatra S (ed) (2021) Recent advances in PMOS negative bias temperature instability: characterization and modeling of device architecture, material and process impact. Springer Nature
9. Allioui H, Mourdi Y (2023) Exploring the full potentials of IoT for better financial growth and stability: a comprehensive survey. Sensors 23(19):8015
10. Mishra P, Singh G (2023) Energy management systems in sustainable smart cities based on the internet of energy: a technical review. Energies 16(19):6903
11. Elavarasan RM, Mudgal V, Selvamanohar L, Wang K, Huang G, Shafiullah GM, Nadarajah M et al (2022) Pathways toward high-efficiency solar photovoltaic thermal management for electrical, thermal and combined generation applications: a critical review. Energy Convers Manage 255:115278
12. Li Y, Wang H, Trik M (2024) Design and simulation of a new current mirror circuit with low power consumption and high performance and output impedance. Analog Integr Circ Sig Process 119(1):29–41
13. Marwedel P (2021) Embedded system design: embedded systems foundations of cyber-physical systems, and the internet of things. Springer Nature, p 433
14. Hill I, Chanawala P, Singh R, Sheikholeslam SA, Ivanov A (2021) CMOS reliability from past to future: a survey of requirements, trends, and prediction methods. IEEE Trans Device Mater Reliab 22(1):1–18

15. Choe G, Kwak J, Yu S (2023) Machine learning-assisted compact modeling of W-doped indium oxide channel transistor for back-end-of-line applications. IEEE Trans Electron Devices 71(1):231–238
16. Mao S, Su X, Wu Q, Wang Y, Duan X, Tian S, Xu Y et al (2024) Physics-based compact models of GaN HEMTs for high power RF applications: a review. Int J Numer Model Electron Networks Devices Fields 37(4):e3276
17. https://github.com/Aroondhati/Static-CMOS-Inverter

Analysing the Optimization of Cooling Systems in Industrial Settings: Exploring Efficiency and Cost Reduction



Farid Agayev and Sevinj Bakhshaliyeva

1 Introduction

This paper is part of the broader dissertation on "Intellectual Control of Water Cooling and Cleaning Processes," which focuses on the integration of AI networks into industrial applications. The main objective of the dissertation is to explore how advanced technologies can optimize cooling and cleaning processes within industrial settings, particularly in oil refining plants [1–3]. In this paper, the emphasis is on analysing ways to reduce power consumption, water usage, and environmental impact by examining the operational modes of cooling systems. Cooling systems in industrial environments, though often regarded as straightforward, encounter significant challenges due to varying operating conditions and external influences [4–8]. This paper investigates a range of cooling strategies, from passive methods that depend on natural convection to more intricate active cooling systems utilizing cooling towers. The study employs empirical data and analytical modelling to evaluate how these methods affect energy efficiency and water conservation [9, 10]. The findings will guide the development of best practices for optimizing cooling systems, ultimately paving the way for AI-driven controls that can adjust to real-time environmental changes, thereby substantially reducing operational costs and minimizing the environmental footprint of industrial cooling processes.

F. Agayev · S. Bakhshaliyeva (✉)
Azerbaijan State Oil and Industry University, Baku, Azerbaijan
e-mail: sevinj.quliyeba@asoiu.edu.az

F. Agayev
e-mail: farid.aqayev@asoiu.edu.az

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_20

2 Formulation of the Problem

Cooling systems in oil refining plants, often overlooked due to their simplicity, face significant challenges related to varying operational regimes and external factors. Optimizing these systems is essential for reducing energy consumption and water usage [11–14]. The research delves into different cooling system regimes and examines external influences to develop strategies that enhance efficiency and sustainability, ensuring competitiveness while meeting environmental standards. An exploration of various cooling system regimes and the impact of external factors is undertaken through empirical data collection and analytical modelling. Testing different operational scenarios reveals key variables that influence energy and water usage [15–17]. The findings are validated through simulations and real-world applications, ultimately leading to the establishment of best practices for achieving more efficient and sustainable cooling system operations. Figure 1 illustrates a basic water cooling system.

Figure 1 presents a diagram of the cooling and water treatment system, highlighting the main processing stages. The initial stage involves passive cooling: hot water passes through radiators situated in an open basin (1), where cooling occurs due to external conditions such as air temperature and wind speed. This is followed by the oil purification and analysis process, where oil skimmers remove residual oil from the water, and analyzers check the water quality, detecting oil impurities and other parameters (2). The water then moves to cooling towers, where powerful fans further reduce the water temperature through evaporation (3). In the next stage, a pump maintains the necessary pressure in the system to circulate water through all cooling and purification stages (4). Finally, the cooled water is used to cool technological equipment through several heat exchangers (5), effectively dissipating excess heat from the system. This diagram represents a simplified version of the entire system but includes the key components that ensure its operation.

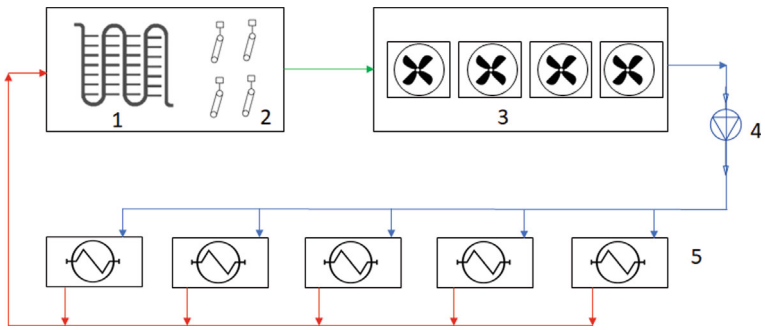


Fig. 1 Simple diagram of ındustrial water cooling and cleaning processes

3 Solution of the Problem

Passive cooling refers to the process of reducing the temperature of a system without the use of active mechanical devices such as fans or pumps. In the context of water cooling systems, passive cooling primarily occurs through natural heat exchange with the environment. This process is driven by convection, where the water's heat is transferred to the surrounding air, influenced by external factors like outside temperature, wind speed, and humidity. These parameters collectively cause the water to lose energy, resulting in a temperature decrease. To simplify the analysis of these variables, the concept of "feels like" temperature is used, which combines the effects of temperature, wind speed, and humidity into a single metric. This approach allows for a more comprehensive understanding of the passive cooling process. Figure 2 illustrates the relationship between "feels like" temperature and the temperature drop between the input and output of the passive cooling system, demonstrating how these combined factors influence the efficiency of cooling.

Figure 2 illustrates the relationship between "Feels Like Temperature" (°C) and the temperature drop (°C) in a passive cooling system. The curve shows that as the "Feels Like Temperature" increases, the cooling efficiency decreases. Starting at 15 °C when the "Feels Like Temperature" is 0 °C, the cooling effect gradually diminishes. By 38 °C, the temperature drop is reduced to approximately 0.5 °C, indicating a significant loss in cooling efficiency under extreme conditions. This curve demonstrates how external environmental factors influence the performance of the passive cooling system.

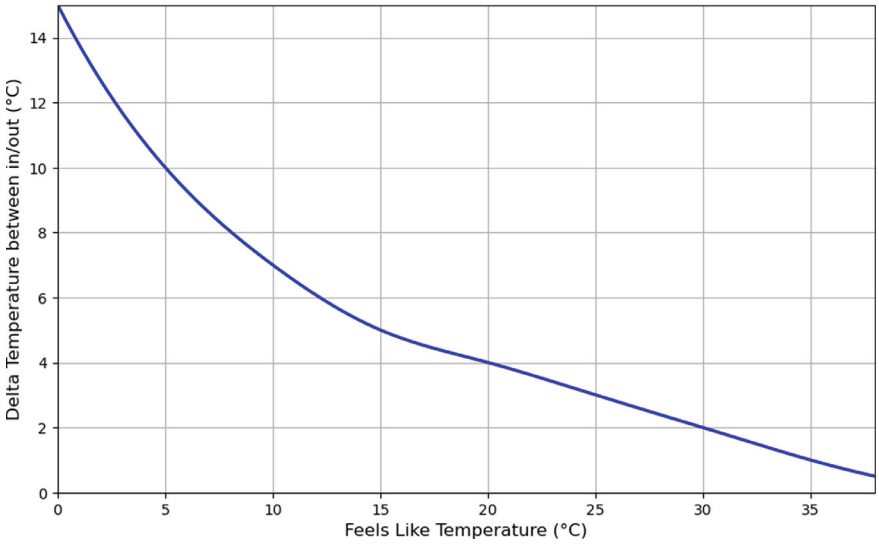


Fig. 2 Relationship between feels like temperature and temperature drop

Active cooling in industrial systems primarily involves the use of cooling towers, with two main methods for removing heat from water. The first and most common method is based on the principle of creating a vacuum or negative pressure inside the cooling tower. This process removes air from the tower, causing water to evaporate, which leads to a physical loss of energy and efficient cooling.

This method is highly effective due to the evaporation process and is widely used in various industrial settings. The second method, though less common, is used primarily in cold weather. In this case, air is blown directly onto the water, promoting convection, which causes the water to lose energy. This method is similar to passive cooling and depends on external factors like air temperature. However, in hot weather, this method is inefficient.

In Fig. 3, two distinct methods of cooling within a cooling tower are depicted. The first method, shown in Diagram A, represents the standard approach where cooling is achieved through evaporation. In this process, water descends while air, drawn by the cooling fan labeled (1), is expelled upward, creating a vacuum. This negative pressure facilitates water evaporation, which effectively reduces the water's temperature. The radiators or drop catchers, indicated by (2), play a crucial role in capturing any residual water droplets, preventing them from escaping with the exhaust air. The second method, illustrated in Diagram B, depicts a less common technique, primarily utilized in cold weather. In this scenario, air is blown directly onto the water surface, thereby promoting cooling through convection. The cooling fan (1) aids in the movement of air, while the water level is deliberately raised to reach the lower radiators (2), thereby enhancing the cooling efficiency. This setup relies on convection, akin to passive cooling, where external conditions significantly influence the cooling process. Although this method proves less effective in hot weather, it can be advantageous under specific conditions.

The temperature drop ΔT in a cooling system due to evaporation is governed by the following formula (1):

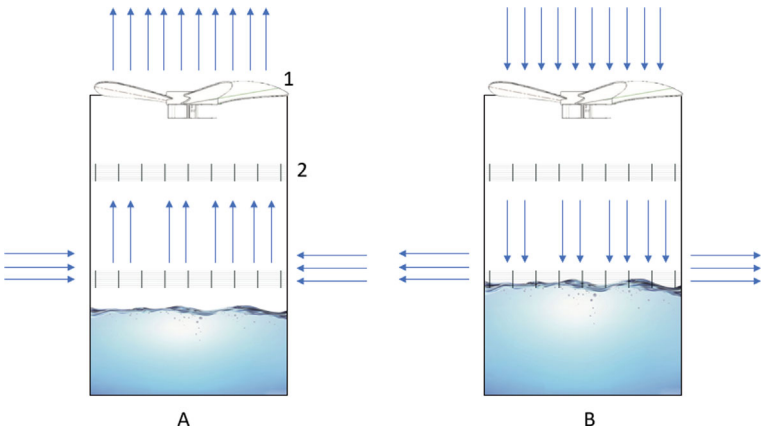


Fig. 3 Cooling tower methods: evaporation (**a**) versus air blowing (**b**)

$$\Delta T = \frac{mL}{C_p M} \tag{1}$$

where:

- ΔT: Temperature drop (°C)
- m: Mass of water evaporated (kg)
- L: Latent heat of vaporization (J/kg)
- C_p: Specific heat capacity of water (J/kg·°C)
- M: Total mass of water in the system (kg)

Evaporation absorbs energy ($\boldsymbol{L}$), leading to cooling. The temperature drop depends on the amount of water evaporated ($\boldsymbol{m}$) and the total water mass ($\boldsymbol{M}$), alongside the thermal properties of water ($\boldsymbol{L}$ and $\boldsymbol{C_p}$). External factors, such as air flow and humidity, influence the rate of evaporation.

Figure 4 presents a detailed analysis of the relationship between fan rotating frequency and its effects on both temperature drop and water evaporation rates within a cooling tower system. The two plots provide insights into different cooling methods, highlighting how each method performs under varying external conditions. The first plot on the left illustrates the relationship between fan rotating frequency (Hz) and the temperature delta between the input and output water temperatures (°C) for three scenarios: evaporation cooling, air blowing at 5 °C, and air blowing at 25 °C. The evaporation method, represented by the blue solid line, shows a significant increase in temperature delta as the fan speed increases, reflecting the enhanced cooling effect due to more efficient air movement.

This effect is particularly pronounced at higher fan speeds, where the system approaches its maximum cooling performance. In contrast, the air blowing technique, represented by the black dashed line for 5 °C and the gray dash-dot line for 25 °C, shows a more moderate increase in temperature delta.

The air blowing method at 5 °C performs better than at 25 °C, indicating its higher efficiency in cooler environments. However, both air blowing scenarios reach a plateau, suggesting that beyond a certain fan speed, the cooling efficiency does not improve significantly.

The second plot on the right illustrates the corresponding relationship between fan rotating frequency and the amount of water evaporated per hour, measured in liters, for the same three scenarios. The evaporation method, again represented by the blue solid line, shows a sharp increase in the evaporation rate with increasing fan speed, particularly at higher frequencies.

This increase is due to the negative pressure created within the cooling tower, which enhances the evaporation process. On the other hand, the air blowing techniques, depicted by the black dashed line for 5 °C and the gray dash-dot line for 25 °C, show a much more gradual increase in water evaporation. The air blowing method at 25 °C results in higher evaporation than at 5 °C, but both remain significantly lower than the evaporation method, highlighting the lower water usage associated with air blowing techniques under these conditions.

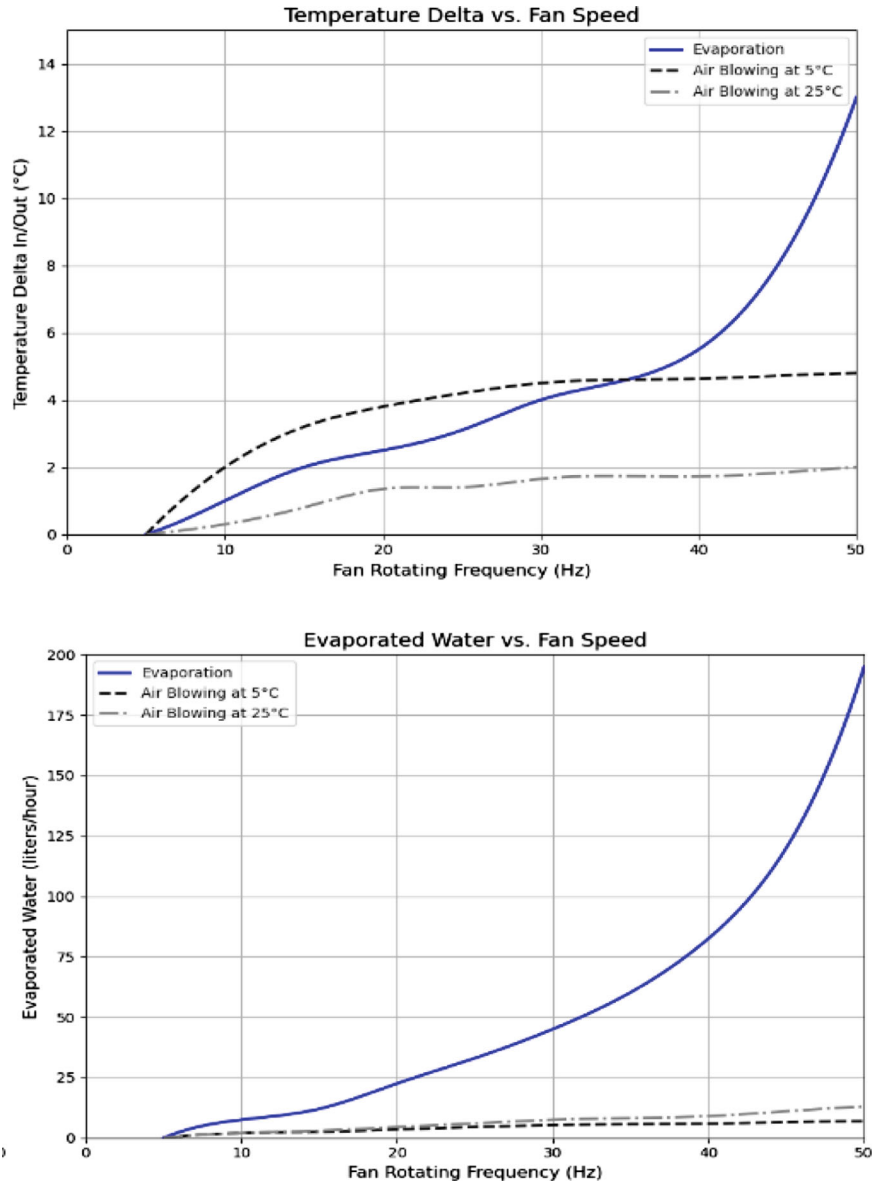


Fig. 4 Fan speed versus cooling efficiency and evaporation

This comprehensive depiction emphasizes the differences in cooling performance and water usage between the evaporation and air blowing methods, demonstrating the importance of selecting the appropriate cooling strategy based on environmental conditions and operational requirements. The use of distinct line styles and colors ensures clarity, even in black-and-white print formats.

The research identified critical factors influencing the efficiency of cooling systems in oil refining plants, highlighting the impact of both passive and active cooling methods. Passive cooling, driven by environmental factors like air temperature and wind speed, proves effective in cold weather conditions but shows diminished efficiency as external temperatures rise. The study demonstrated that under extreme conditions, the temperature drop in passive systems is significantly reduced, necessitating alternative methods to maintain cooling efficiency. Active cooling, particularly in cooling towers, was analyzed through empirical data and modeling. The findings show that cooling towers achieve maximum efficiency when operating at the highest fan speeds, around 50 Hz, where negative pressure maximizes water evaporation and cooling effectiveness. However, this comes at the cost of increased water evaporation, which can lead to higher operational costs.

The study underscores the importance of optimizing fan speeds and cooling tower design to balance cooling efficiency with resource utilization, suggesting that adaptive strategies incorporating real-time external data could further enhance system performance.

4 Conclusion

The analysis reveals that cooling towers, when operating at maximum capacity, achieve the highest possible efficiency, but this approach is not always cost-effective. Passive cooling methods, especially effective in cold weather, and air-blowing techniques in cooling towers though less common in industrial settings are efficient only when the “feels like” temperature is below 10 °C. However, running cooling towers at maximum power leads to significant water evaporation, which adds to operational costs. Despite these findings, most industrial systems are not designed to dynamically adjust based on external factors like weather, primarily due to the complexity and resource demands of such programming. However, by integrating AI technologies like LSTM networks, these systems could quickly adapt to changing conditions, optimizing work regimes based on real-time and forecasted weather data. Implementing AI in this way could lead to substantial reductions in both energy and water costs, making industrial cooling systems more efficient and cost-effective.

References

1. Zamanifard A, Gholampour M, Liaw J-S, Wang C-C (2023) Recent progress in fill media technology for wet cooling towers. Processes. https://doi.org/10.3390/pr11092578
2. Akhmetov SA, Bayazitov MI, Kuzeev IR, Serikov TP. Technology and equipment for oil and gas refining processes. https://www.geokniga.org/books/11385
3. A comprehensive study of different types of cooling towers. https://www.ijraset.com/research-paper/different-types-of-cooling-towers
4. Erikh VN, Rasina MG, Rudin MG (1977) Chemistry and technology of oil and gas. Chemistry 424
5. Thermal performance and emissions analysis of a new cooling tower prototype. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S1359431122005986
6. Bahadori A (2016) Essentials of oil and gas utilities: process design, equipment, and operations. Gulf Professional Publishing
7. Yusifov Sİ, Mayilov RA, Mehdiyeva AM (2023) Advanced information-measuring system for the improvement of the quality indicators of metrological characteristics. In: Environment, energy and earth sciences journal—E3S web of conferences. VI international scientific forum on computer and energy sciences (WFCES II 2023). 16–17 Nov 2023, Almaty, Kazakhstan. ISSN: 2267-1242. EDP Sciences. https://doi.org/10.1051/e3sconf/202346601001. https://www.e3s-conferences.org/component/issues/?task=all&Itemid=121
8. Yusifov SI, Bayramov IY, Mayilov RA, Mehdiyeva AM, Sardarov YB, Safarova AA (2025) Prediction of parameters of oil and gas fields using fuzzy modeling. Lecture Notes in Networks and Systems, 1142 LNNS, pp 203–207
9. Agayev FH, Mehdiyeva AM, Gafarov GA, Bakhshaliyeva SV (2024) Development of the functional safety system in primary oil refining. Nafta-Gaz. Oil and Gas Institute—National Research Institute. №7, pp 434–442. Poland. ISSN 0867-8871. https://www.inig.pl/en/nafta-gaz-en/available-articles. https://www.inig.pl/images/nafta-gaz/abstrakty/Nafta-Gaz_2024-07_Spis%20tre%C5%9Bci.pdf
10. Agayev FH, Mehdiyeva AM, Ibrahimova AE (2025) Advance technologies to eliminate environmental problems of oil and gas processing. In: Advances in science, technology and innovation, pp 287–294
11. Lebedev SV (2002) Firewalling. Theory and practice of protecting the outer perimeter. Publishing house of MSTU im. N.E. Bauman, 304 p
12. Gafarov GA, İbrahimova AE (2024) Design of a device for stimulation of biologically active points. In: 2nd international conference on emerging applications of material science and technology (ICEAMST 2024), 3–4 July 2024, Bengaluru, India, Advances in Science, Technology & Innovation
13. Gafarov G, Takhumova O, Nikolaev P (2023) Determination of intracellular electrical parameters in bioelectrical impedance analysis. BIO web of conferences CIBTA-II-2023
14. Shuvalov VP, Egunov MM, Minina EA (2015) Providing indicators of reliability of telecommunication systems and networks. Hotline - Telecom, 168 p
15. Stearns SD (2018) Digitale Verarbeitung Analoger Signale/Digital signal analysis/S. D. Stearns. Walter de Gruyter GmbH & Co KG
16. Mityukhin AI (2018) Applied information theory: educational method. allowance/A. I. Mityukhin. BSUIR, Minsk
17. Mityukhin AI (2016) Digital speech processing and image analysis: textbook—method. allowance/A. I. Mityukhin. BSUIR, Minsk

Modeling of Solar Panel Installation Mechanism for Hybrid Energy Systems in Azerbaijan



Rehim Rehimov, Nizami Suleymanov, and Zarifa Mahmudova

1 Introduction

In parallel with the increasing demand for electricity in modern times, the problem of limited reserves of mining-based energy resources, their negative effects on the environment because of the use of these energy resources, and lack of access to electricity arises. These energy sources, especially in the twentieth century, could not find their application from time to time due to various crises. Since the last quarter of the century, the interest in renewable energy sources has increased significantly.

Innovative solar tracker installations today have different functional character, differing in design and control principle.

The classification of rotation axes is divided into single-axis (single axis trackers—SAT), as well as dual-axis (dual axis tracker—DAT). The difference in systems is due to the simplicity of the design and the cost of installation. Depending on the choice of the design of the installation, single-axis trackers are usually divided into trackers with horizontal (HSAT), vertical (VSAT), inclined (TSAT) and polar-oriented (PSAT) rotation axes. It is easy to assume that the use of one or another installation has different aspects of use. With a significant change in the duration of the solar day and small elevation angles for high latitudes, it is proposed to use trackers with a vertical or inclined rotation axis, and for close areas of the equator—trackers with a horizontal rotation axis. It is considered more effective. To achieve maximum efficiency of solar energy, dual-axis trackers are used, the designs of which are divided according to

R. Rehimov (✉) · N. Suleymanov · Z. Mahmudova
Azerbaijan State Oil and Industry University, Baku, Azerbaijan
e-mail: rehim.rehimov@asoiu.edu.az

N. Suleymanov
e-mail: nizami.suleymanov@asoiu.edu.az

Z. Mahmudova
e-mail: zarifa.mahmudova@asoiu.edu.az

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_21

the direction of the main axis. The classification of dual-axis systems is divided into those with a rotation axis on a supporting column (tip-tilt dual axis tracker—TTDAT) and trackers with a supporting plane (azimuth-altitude dual axis tracker—AADAT), in which the main axis is vertical. The advantages of trackers with a rotation axis on a supporting column include a primitive design and greater flexibility of territorial placement, which allows expanding the scope of use. From individual use to large-scale production. Platform trackers use a large ring as a supporting structure, which is installed on rollers or a large platform with bearings. The main advantage of using this design is the uniform distribution of the solar battery weight across the parts of the ring.

The mounting structure is designed to attach photovoltaic modules to the solar tracking system and must provide the necessary strength and rigidity to external influences. The main features of this structure, which serve as technical definitions for the use of the installation, are the permissible mass and area of the solar battery, the maximum permissible wind speed in sounding mode and in storm conditions, the range of operating temperatures and the maximum snow load.

Of the existing tracking algorithms, only three are used. The first is based on a light sensor, which is considered the simplest solution in terms of its technical design. The operating principle is to direct the solar panel to the sky in accordance with the maximum light intensity determined by the parameters of the photodetector. However, this method has a significant drawback—it does not work in heavy rain, cloudy weather and dirty photodetectors [1, 2]. The second of the existing methods is the astronomical algorithm. The implementation consists of changing the position of the solar panel using a daily software timer.

The negative aspect of the algorithm is the geographical location in high-latitude areas, the need to adjust the daily schedule due to important seasonal fluctuations in the azimuthal angles of sunrise and sunset. Maximum efficiency is guaranteed by the third method—the solar positioning algorithm, the method of action is based on software calculation of the optimal positioning angles of the solar battery according to the specified parameters of the location of the solar tracker: latitude, longitude, altitude above sea level.

Hybrid energy systems have emerged because of the growing interest in renewable energy sources and as a way to feed utilities located in regions far from the interconnected grid system. In addition, in regions with an interconnected grid system, hybrid energy systems have emerged to reduce the ever-increasing unit cost of electricity.

Sources such as wind and solar energy have variable characteristics due to different climatic conditions. Storage systems are important because of the variability of wind speed and solar radiation. For this reason, it should be used in conjunction with energy storage systems to ensure energy continuity in off-grid locations.

Hybrid power generation systems can be developed with many combinations of different energy sources and storage devices [3, 4]. Solar and wind power are the most used renewable energy sources in hybrid power generation system. In applications, off-grid wind and solar power generation systems are often operated in conjunction with gas turbines, diesel generators, fuel cells, and battery banks. Since it is based on the analysis of the theoretical conceptual bases of the modeling of

hybrid systems, axiomatic, hypothetical and abstraction methods have been applied by conducting research. The hypothetico-deductive method is a research concept that begins with a theory about how things work and derives testable hypotheses. It is a form of deductive reasoning that begins with general principles, hypotheses, and ideas. Hypotheses are tested by collecting and analyzing data, and the theory is then supported or rejected by the results [5–9]. The method of scientific abstraction or abstraction is a special thinking technique that allows you to form certain abstract concepts.

In the research work, a system configuration that can optimally feed a certain load in a certain region (Gubadli district) is determined. For this, a sample hybrid energy system based on wind and solar with grid connection was modeled using the HOMER program. During the modeling of the sample hybrid energy system, sensitivity values for average solar radiation and average wind speed values of the studied region were determined. By carrying out this process, in a certain sense, other regions were also taken into account and system configurations were determined. It was determined to provide the same load in an economically optimal way in these different regions [10–14]. These operations were carried out by determining the sensitivity value for grid sales power during the modeling of a sample hybrid power system for both those who cannot sell electricity to the grid and those who can.

2 Types of Hybrid Energy Systems, Their Working Mechanism and Main Elements

The energy efficiency of solar photovoltaic modules plays a key role in their operation as the main sources of electricity generation. One of the most common and technically easier to implement options for their use in terrestrial conditions is the stationary placement of the structure with the orientation of the beam-receiving surface relative to the southern position of the sun in the northern geographic zone, taking into account the selected season. However, in this case, the sun's rays in the morning and evening fall on the working surfaces of the photoconverters at fairly small angles, which significantly reduces their efficiency. Research aimed at eliminating this drawback of automatic devices for constant tracking of the position of the sun in the sky during daylight hours significantly complicates the design, increases its cost and requires additional energy, which is very little in places where solar panels are installed.

Considering cost, which is a technical factor, hybrid energy systems are superior to some single energy sources. The energy potential of this region is one of the important factors affecting the cost. In a region where there is not enough energy source to be used, any energy production causes more costs [15, 16]. Among energy sources, wind and solar energy are the most widely used renewable energy sources in various hybrid energy systems. Hybrid energy systems include Solar-Wind-Hydroelectric hybrid energy system, Solar-Hydrogen hybrid energy system, Solar-Wind-Hydrogen

hybrid energy system, Solar-Wind hybrid energy system, Solar-Wind-Diesel hybrid energy system, Wind-Diesel hybrid energy system includes.

The energy generated from wind and water generators is AC voltage, which is then converted to DC to charge the battery. The controller protects the battery from sudden charge losses or overcharging. Although high voltage can be used to reduce system losses, the converter is usually designed to convert low DC voltage to 110 V or 230 V AC voltage, depending on the power system used in the region. As shown in Fig. 1, the DC power from the photovoltaic panel meets the power needs of the electrolyzer. The water in the electrolyzer is separated into hydrogen and oxygen by the energy received from the solar panel. The obtained hydrogen is then collected in a tube and fed to the fuel cell. The hydrogen entering the fuel cell from the anode compartment combines with oxygen from the ambient air or oxygen obtained by electrolysis from a storage tube to generate DC power. This energy can be used directly in any device that runs on DC power, or it can be converted to AC power through inverters to meet all the energy needs of a home or any environment.

A form of energy storage in wind-diesel systems is the use of electrochemical storage or batteries. Also referred to as wind-diesel-battery systems, these systems generally include one (or more) diesel generator, a battery unit (with inverter), wind turbines, AC/DC rectifiers, DC/AC three-phase converters. The way the wind-diesel-battery system works varies depending on the wind conditions. If the wind is good and produces more electricity than the grid needs, the excess energy is stored in the battery.

In addition to the battery and wind turbine, the grid kicks in to compensate for the load gap during moderate wind. If the wind is too weak, the grid is fed with

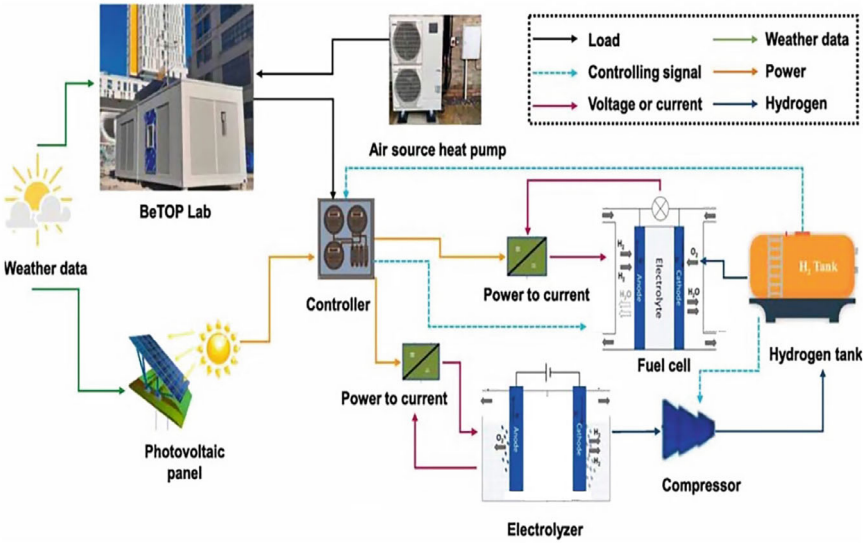


Fig. 1 Solar-hydrogen hybrid energy system [19]

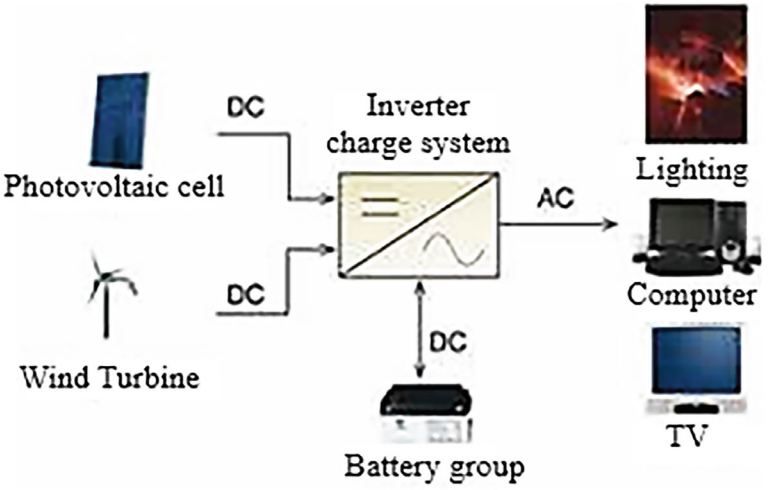


Fig. 2 Solar-wind hybrid energy system

electricity generated by the operation of the diesel generator. In this system, co-energy is produced using wind turbines and photovoltaics. In Fig. 2, the energy generated from the wind turbine and solar panel passes through a charge controller and goes directly to an inverter that converts the power to DC or AC. Power via charge controller goes to feed the battery and charges the batteries. The energy converted to AC can be used directly in homes.

Solar panels, wind turbines, charge controllers, batteries and inverters are used as the main elements of Solar-Wind Hybrid systems.

Solar cells are devices that convert light energy into electricity and cannot store energy. When the light source is gone, the electricity produced by the battery will also be cut off. If electricity is used at night, electric storage should be added to the system.

3 Methodology of Analysıs of Alternatıve and Renewable Energy Resources of Azerbaıjan

Although Azerbaijan is a country rich in hydrocarbons and is known in the world in this field, in recent years, importance has been given to produce and use of energy of renewable energy sources in Azerbaijan, international cooperation in this field, capital investment and export of this energy.

According to the Ministry of Energy of Azerbaijan, the country's economic and technical renewable energy capacity is 27,000 MW. Of this, 3000 MW are wind, 23,000 MW solar, 380 MW bioenergy and 520 MW mountain rivers. Azerbaijan produces 17.3% of its total electricity production from renewable sources. Azerbaijan

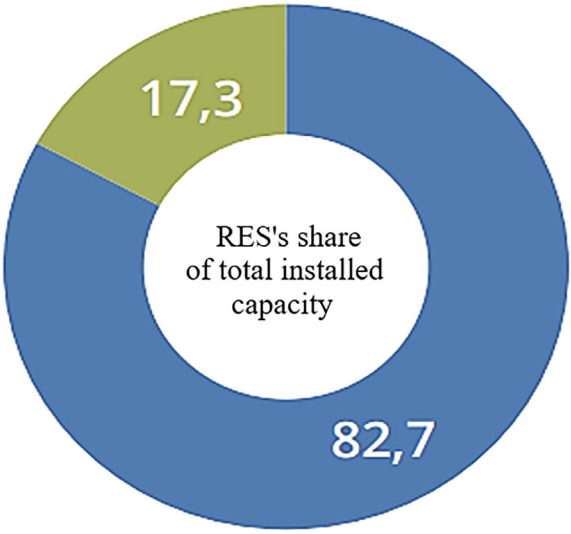


Fig. 3 Capacity of renewable electricity sources in Azerbaijan

plans to increase this capacity to 30% by 2023. 88.5% of renewable energy in the country is provided by HPP, 5.1% by wind, 3.5% by solar, and 2.9% by bioenergy sources. Figure 3 shows the capacity of renewable electricity sources in Azerbaijan.

The lands freed from occupation during the 44-day war are also an important energy center in terms of renewable energy sources. 25% of Azerbaijan's water resources are in these lands. The solar potential of these lands is of particular importance in terms of electricity production. The potential of solar energy in de-occupied lands is thought to be more than 7200 MW. In this sense, by signing contracts with various international companies, he set himself the goal of increasing the power of Azerbaijan's renewable energy resources. Azerbaijan also aims to export natural gas, which is the main source of electricity production, by using electricity produced from renewable energy sources within the country. The fact that hydrocarbon reserves are limited as traditional energy sources and the prevention of environmental pollution caused by it makes it necessary to increase the amount of energy produced in the world due to alternative and renewable energy sources. Currently, there are positive experiences in this direction, and in some countries the use of solar, wind and other alternative and renewable energy sources is expanding year by year. The use of alternative energy resources varies depending on the characteristics of the regions where different countries are located. For example, in developed countries such as Germany and Israel, the use of renewable energy sources and solar panels has spread significantly.

Israel plans to increase the use of renewable energy resources to 10%. In particular, the use of these resources for irrigation and heating of water in residential buildings is very important and is defined as an important part of sustainable development. In recent years, scientific research in the field of energy in Israel, because of increasing investments by the state and private sector, have achieved positive results in this field.

Table 1 Potential of renewable energy resources of the Republic of Azerbaijan

Type of capacity	Power (MW)
Solar energy	> 5000
Wind energy	> 4500
Biomass energy	> 1500
Geothermal energy	> 800
Small hydroelectric plant	> 350

Source Prepared by the author based on the information of the website of the State Agency for Renewable Energy Sources under the Ministry of Energy of the Republic of Azerbaijan (area.gov.az)

In 2012, Germany increased its solar energy consumption by 50% and reach to 6.1% of total consumption. Recently, the use of wind energy has increased in the United States. In 2011 alone, 3,464 wind turbines were installed within US borders.

During the windy months in the United States, wind energy accounts for 6% of the country's total energy consumption. Although Azerbaijan is a country of oil and natural gas and develops its economy by selling its energy resources to the world energy market, the country should give priority to the development of alternative energy sources for long-term sustainable development. The potential of using renewable energy resources (water, wind and solar energy) in the country is significantly large. Azerbaijan has an important position in terms of the potential of renewable energy resources due to the number of sunny and windy days (171 days). In the Absheron Peninsula and the coastal lands of the Caspian Sea, the duration of sunlight is equal to 2500 h per year, and in the Nakhchivan Autonomous Republic to 2900 h (Table 1).

4 Determınatıon of Hybrıd Energy System for Gubadlı Regıon

A hybrid energy system to be modeled in HOMER can include many different power systems, as it is possible to specify more than one value for the elements that must be quantified or measured to create this hybrid energy system [7, 17, 18]. The example of a hybrid power system to be modeled is a grid-connected wind-solar hybrid power system, and this hybrid power system will include many different power systems. In other words, this example of a hybrid power system to be modeled consists of two wind turbines each with two values of quantity, grid connection, load, inverter with four values of size, and a photovoltaic array with five values. A schematic diagram of this example hybrid power system is shown in Fig. 4.

An example hybrid power system to be modeled in HOMER is to be installed in a region with wind and solar sources. A medium-sized industrial facility located in Karabakh was identified as load. Therefore, the region to be investigated for wind source and solar source will be the region where this industrial facility is located.

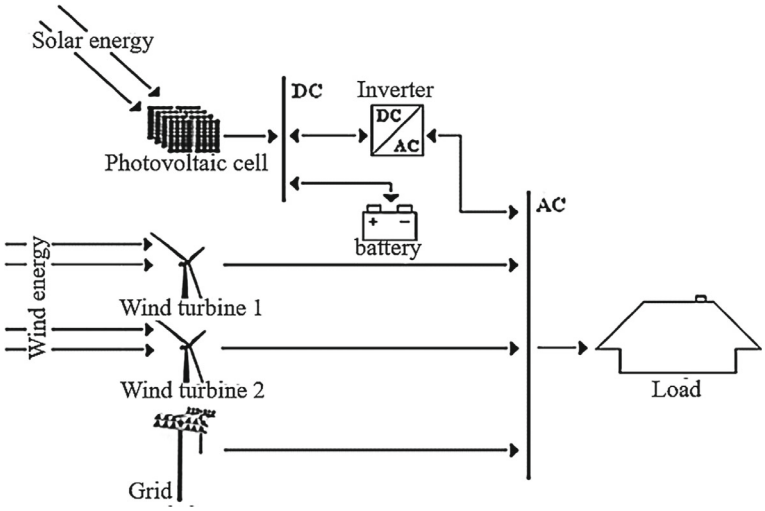


Fig. 4 A sample structure of hybrid energy system for Gubadli region

That is, for the source of wind and solar power, there will be inputted the latitude and longitude coordinates of the point where this industrial facility is located. Photos of this industrial facility located in Gubadli, that is, in the region to be studied, taken from different distances through Google Earth are shown in Fig. 5. HOMER's Opening Window. The opening window of the HOMER program is shown in Fig. 6. The add/remove window opened by pressing the add/remove button in HOMER's pop-up window is shown in Fig. 7. The user can add or remove any element, load and network connection type through this window.

After completing the addition of the PV array, the PV input window opened by clicking the PV icon in HOMER's opening window is shown in Fig. 8. Cost $7000,



Fig. 5 Designation region for an example of a hybrid power system

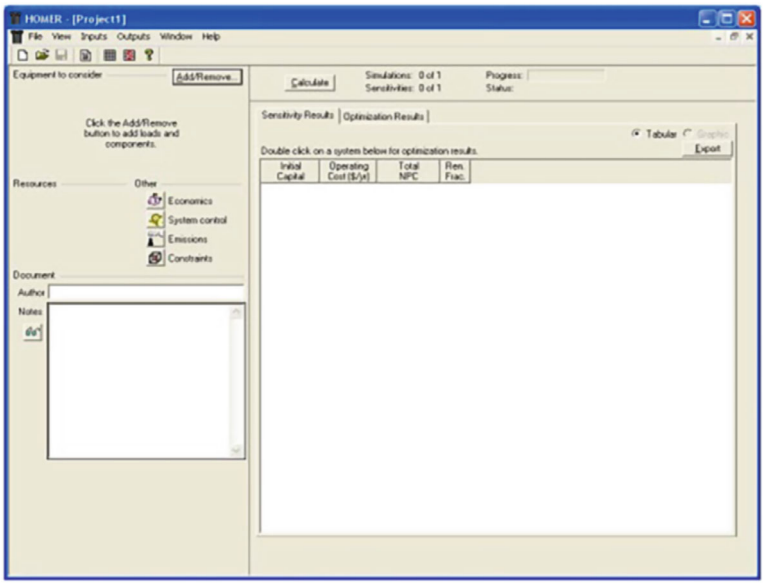


Fig. 6 The opening window of the HOMER program

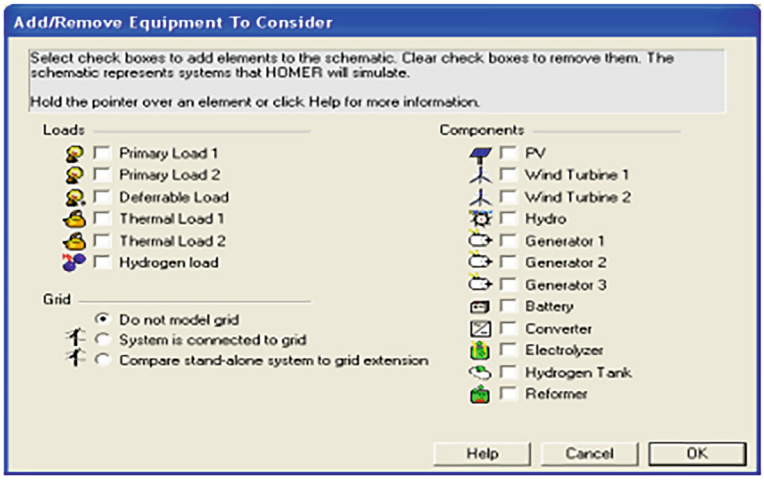


Fig. 7 Add/remove window

replacement (renewal cost) $7000, and operation and maintenance cost is defined as $0 USD/year. After entering these values, the spreadsheet that will be generated will tell you that a 1 kW PV array will have a capital cost of $7000, a renewal cost of $7000, and an O&M cost of $0 per year. Under the heading Sizes to Consider, the size is defined as 0, 10, 20, 40 and 80 kW respectively. As expected, each of these

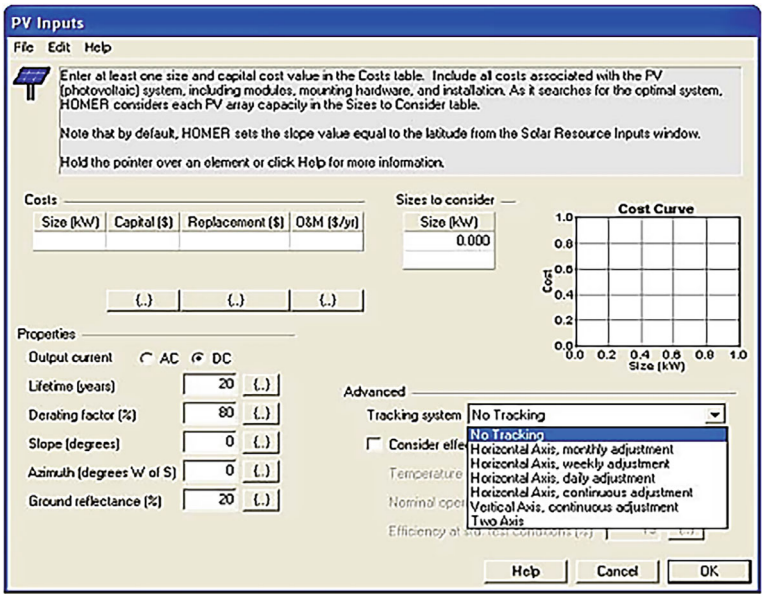


Fig. 8 PV input window

values represents a different size photovoltaic array. After entering these values, the table that will be generated will tell HOMER to consider each of these five different sized photovoltaic arrays separately in the simulations. Under the properties heading in the PV input window, the values of the photovoltaic array are defined as follows: The output current is selected as DC (DA). The life span is defined as 25 years. The reduction factor is a scaling factor that expresses what the new output power of the photovoltaic array will be as a percentage of the specified output power, taking into account the losses that will occur in the output power of the photovoltaic array. In this study, the reduction factor was determined to be 90%.

The inclination angle is the angular value indicating the position of the installed photovoltaic panel relative to the horizontal plane. A slope of 0° indicates that the photovoltaic panel is in a horizontal position, and a slope of 90° indicates that the photovoltaic panel is in a vertical position. In a constant slope photovoltaic panel, a constant slope value equal to the latitude value of the point where the photovoltaic panel is located will maximize the annual photovoltaic power production.

In the values under the costs heading in the PV input window, size (size) is defined as 1 kW, capital. For this reason, there is no tracking corresponding to a constant slope PV panel in the tracking system section under the advanced heading in the PV inputs window, or if the vertical axis option is selected, HOMER will set the slope value as the latitude value of the point. However, HOMER allows entering another angle value for the slope value.

If horizontal axis or two axes corresponding to non-fixed-tilt photovoltaic panel is selected in the tracking system section, HOMER will not allow any angle value to

be entered for the slope value, but will calculate this slope value each time. In this study, since no tracking option was selected in the tracking system section, HOMER set the slope value to 40.8833°, which is the update value of the sample hybrid power system and thus the investigated region for the photovoltaic panel.

Azimuth (south angle) is the angle value that defines the direction (south, east, west, north) the photovoltaic panel is facing. The azimuth value is set to 0° if the photovoltaic panel faces south, − 90° if it faces east, 90° if it faces west, and 180° if it faces north. A photovoltaic panel that will have a fixed orientation is almost always oriented towards the equator. At this time, the azimuth value should be set to 0° for a photovoltaic panel located in the northern hemisphere, and 180° for a photovoltaic panel located in the southern hemisphere.

If the photovoltaic panel is placed horizontally (zero slope) on the earth's surface, the azimuth value is irrelevant. In the pv entries under the advanced heading in the tracking system section, if one of the horizontal axis (monthly adjustment), horizontal axis (weekly adjustment), horizontal axis (daily adjustment), vertical axis (continuous adjustment) or two axis options is selected, the window, HOMER However, the tracking system the photovoltaic panel will calculate this azimuth value at each time step as it moves. In this study, the photovoltaic panel in the example hybrid energy system was decided to have a fixed orientation, and the azimuth value was set to 0° because the investigated region of this photovoltaic panel was located in the northern hemisphere.

Ground reflectance determines the percentage of solar radiation falling on the ground that is reflected by the ground. Ground reflectance is about 20% for a grass-covered area and about 70% for a snow-covered area. The ground reflectance value is used to calculate solar radiation incident on an inclined photovoltaic panel, but the ground reflectance value has little effect in this calculation. In this study, the ground reflectance value was set to 40%.

Under the advanced heading in the PV inputs window, the values for the photovoltaic array are defined as follows:

The tracking system section is used to determine whether the photovoltaic panel will track the sun and, if so, what tracking system to use. The tracking system section includes the no tracking option, which indicates that the photovoltaic panel will not use any tracking system, the horizontal axis option, which indicates that it will use the horizontal axis tracking system, and the vertical axis option. Using the vertical axis tracking system and the horizontal axis option, which indicates that it will use both the horizontal axis and the vertical axis tracking system includes two axis options. In this study, no tracking option was selected in the tracking system section. Consider effect of Temperature represents the option used to determine whether the temperature effect of the camera on the output power of the photovoltaic array is considered. In this study, it was decided to consider the effect of cell temperature on the output power of the photovoltaic array, and the resulting values are defined below.

Temperature power factor is a ratio that shows how the output power of a photovoltaic array varies with cell temperature. This coefficient has a negative value because the output power of the photovoltaic array decreases with increasing cell

temperature. In this study, the photovoltaic array in the example hybrid power system is assumed to consist of MSX-60 photovoltaic modules from Solarex, and therefore the temperature coefficient of power value is defined as − 0.5% / Ψ from the data sheet for MSX-60 photovoltaic module.

The nominal operating element temperature determines the cell temperature of the photovoltaic array under conditions of 800 W/m^2 solar radiation, 20 ƨ ambient temperature and 1 m/s wind speed. HOMER uses the nominal operating cell temperature value to calculate the cell temperature of the photovoltaic array. In this study, the nominal operating cell temperature value is set to 47 Ψ from the data sheet of a photovoltaic module named MSX-60.

Efficiency under standard test conditions is defined as the efficiency with which a photovoltaic array converts sunlight into electricity at its maximum power point under standard test conditions. HOMER uses a useful duty factor to calculate the cell temperature of a photovoltaic array. A scientific and non-extensive study of the efficiency values found in the data sheets of commonly used photovoltaic modules was conducted by the creators of HOMER, and as a result of this study, average efficiency values were obtained under standard test conditions. for different types of photovoltaic modules. The values of average efficiency are shown in Table 2.

In this study, the efficiency of the photovoltaic module named MSX-60 could not be obtained from the data sheet, but since this photovoltaic module was studied to have a polycrystalline silicon structure, the useful efficiency of the photovoltaic module named MSX-60 was determined to be 13% using Table 2.

After these steps, the resulting image of the pv input window is shown in Fig. 2.

Figure 9 shows the image that appears in the HOMER opening window after completing the modeling of the PV array by clicking OK on the PV inputs window image obtained after the steps taken.

In the main load entries window, in the load type section, AC (aa) is selected as the load type. There are two ways to model a load in HOMER. In the data source section of the initial load entries window, the main path represented by the option Enter daily profile(s) necessary either specify the average daily load profile for each month of the year (different average daily load profiles can be defined for weekdays and weekends of each month) or only one Average daily load profile should be noted (only this average daily load profile is taken into account during the year). Because HOMER conducts all of its operations on an hourly basis throughout the year, the average daily load profile or profiles specified in this first method used to model

Table 2 Average efficiency for different types of photovoltaic modules

Photovoltaic module type	Number of modules in study	Average efficiency (%)
Polycrystalline silicon	10	13.0
Monocrystalline silicon	8	13.5
Monocrystalline/amorphous silicon hybrid	1	16.4
Thin film amorphous silicon	4	5.5

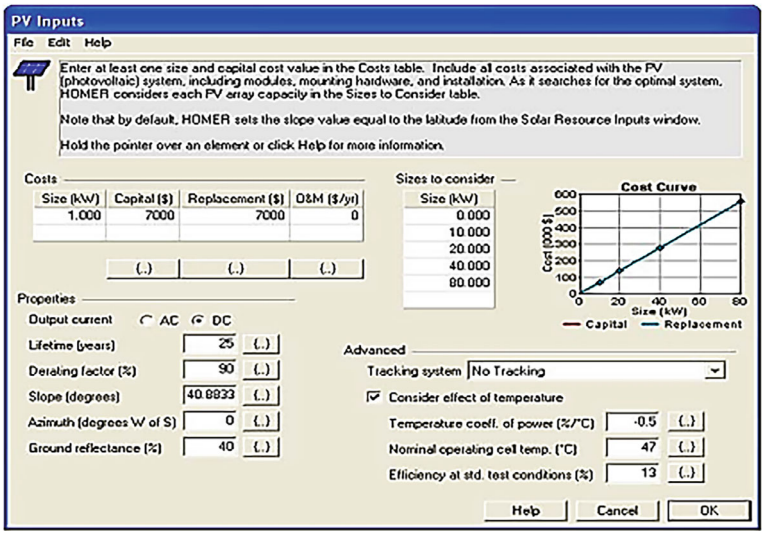


Fig. 9 Resulting view of the PV inputs window

load are translated by HOMER into 8760 load values corresponding to each of the 8760 h (number of hours per year). To make this hourly load data generated by HOMER more realistic, randomness can be added to this hourly load data using the daily and time-step functions, mainly under the heading of random variation. load login window. Figure 10 represents the view of HOMER's opening window after completion of PV array modeling Here, the daily function shows the standard deviation in the consecutive daily load averages, and the time-step function shows the standard deviation in the difference between the hourly load data and the average daily load profile. Previously acquired hourly load data on the second path, represented by the import time series data file option in the data source section of the initial load inputs window, must be transferred to HOMER using the import file button. In this work, the second path represented by the import time series data file option in the data source section of the initial load inputs window was used, and the required hourly load data was obtained through MATLAB software. A maximum hourly load value must be entered to run this MATLAB program. Therefore, this hourly maximum load value is set to 120 kW in order to obtain hourly load data from the MATLAB program that will coincide with the electricity demand of the load given above. The resulting hourly load data is transferred to HOMER using the import file button in this second path represented by the import time series data file option.

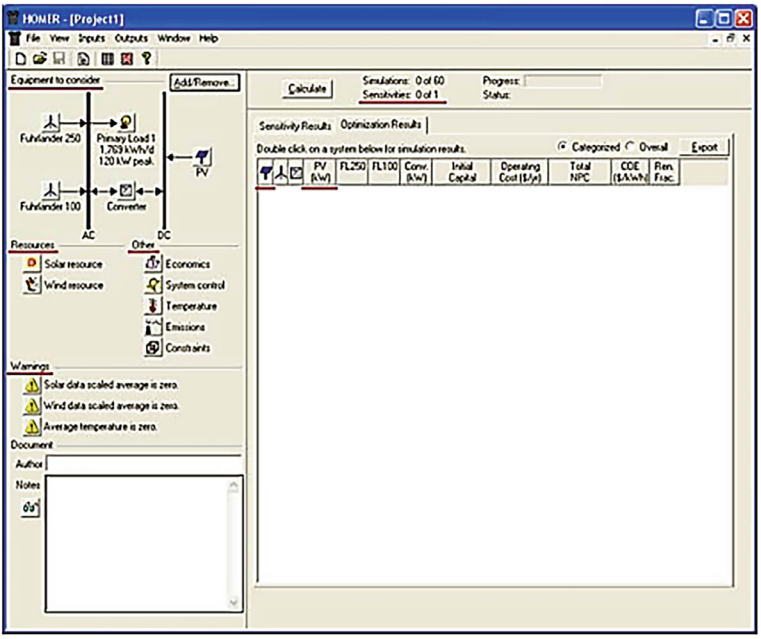


Fig. 10 View of HOMER's opening window after completion of PV array modeling

5 Conclusion

In the research study, an example hybrid power system with grid connection was modeled by HOMER to determine the system configuration that can economically optimally feed a certain load in a certain region (Gubadli region). In carrying out this modeling process, sensitivity values were determined for the average solar radiation and average wind speed values of the studied region, and the sensitivity value for the network sales power was also determined. Certain fixed values have been determined for the costs of wind turbines, the costs of photovoltaic arrays, the price of electricity received from the grid, and the price of electricity sold to the grid. These costs and prices are determined in addition to or independent of the sensitivity values determined in the example hybrid power system. By setting sensitivity values for each of the above-mentioned criteria, one can observe for which situations HOMER will make decision. In addition, in this study, a certain fixed value for the load size was determined, and in another study, by setting sensitivity values for the load size, it can be observed which situations HOMER will make decision for a sample hybrid.

References

1. Artyukhov II, Stepanov SF, Molot SV, Tulepova GN, Erbaev ET, Tulegenov KK (2018) Autonomous power supply system based on a diesel generator and renewable energy sources for remote rural areas. In: 2018 19th International scientific conference on electric power engineering (EPE). IEEE, pp 1–4
2. Nichita C, Luca D, Dakyo B, Ceanga E (2003) Large band simulation of the wind speed for real time wind turbine simulators. IEEE Trans Energy Convers 17(4):523–529
3. Engleitner R, Nied A, Cavalca MS, da Costa JP (2017) Dynamic analysis of small wind turbines frequency support capability in a low-power wind-diesel microgrid. IEEE Trans Ind Appl 54(1):102–111
4. Tanasheva NK, Bakhtybekova AR, Shaimerdenova GS, Sakipova SE, Shuyushbaeva N (2022) Modeling aerodynamic characteristics of a wind energy installation with rotating cylinder blades on the basis of the ansys suite. J Eng Phys Thermophys 95(2):457–463
5. Obukhov SG, Sipaylova NY, Plotnikov IA, Sipaylov AG (2012) Mathematical model of a synchronous generator of an inverter diesel power plant. News of Universities. Problems of energy (9–10):112–120. https://portal.tpu.ru/SHARED/a/AGS/eng/Publications
6. Yusifov SI, Mayilov RA, Mehdiyeva AM, Mehdizade EK (2024) Advanced ınformation-measuring system for the ımprovement of the quality ındicators of metrological characteristics. In: E3S web of conferences (environment, energy and earth sciences), vol 474, 02003, ISSN 2267-1242
7. Agayev FH, Mehdiyeva AM, Gafarov GA, Bakhshaliyeva SV, Shirinzade NV, Development of the functional safety system in primary oil refining Nafta-Gaz. Oil and Gas Institute—National Research Institute. Poland. ISSN 0867-8871. https://www.inig.pl/en/nafta-gaz-en/available-articles
8. Hashimoto T, Kubo M, Kawazoe H, Yagi K (2007) Demonstrative research on hybrid control system of wind, photovoltaic and diesel power generation in Myanmar. In: International conference on power engineering, Oct 2007, China, Hangzhou, pp 1238–1242
9. Onar OC, Uzunoglu M, Alam MS 2006 Dynamic modeling, design and simulation of a wind/fuel cell/ultra-capacitor-based hybrid power generation system. J Power Sources (161):707–722
10. Ruiz AG, Molinas M (2009) Electrical conversion system for offshore wind turbines based on high frequency AC link. In: Ecologic vehicles and renewable energies international conference EVER, Monaco, 26–29 March 2009
11. Sopian K, Zaharim A, Ali Y, Nopian ZM, Razak J, Muhammad NS (2008) Optimal operational strategy for hybrid renewable energy system using genetic algorithms. In: WSEAS transactions on mathematics, March 2008, no. 21, pp 130–140
12. Hasanah RN, Setyawan AB, Maulana E, Nurwati T (2020) Computer-based solar tracking system for PV energy yield improvement. Int J Power Electron Drive Syst 2020 11(2):743–751
13. Ismail FB, Al-Muhsen NFO, Fuzi FA, Sambathan S, Nawawi MNH (2020) Design and fabrication of solar panel with sun position tracker. J Mech Eng Sci 14(2):6906–6916
14. Kasim NK, Obaid NM, Abood HG, Mahdi RA, Humada AM (2021) Humada experimental study for the effect of dust cleaning on the performance of grid-tied photovoltaic solar systems. Int J Electr Comput Eng 11(1):74–83
15. Srikumar K, Saibabu C (2020) A system and novel methodology to track maximum power from photo voltaic system: a comparative and experimental analysis. J King Saud Univ—Eng Sci 32(7):442–458
16. Agayev FH, Mehdiyeva AM, Ibrahimova AE (2024) Advance technologies to eliminate environmental problems of oil and gas processing. In: 2nd International conference on emerging applications of material science and technology (ICEAMST 2024), July 3–4 2024, Bengaluru, India. Advances in Science, Technology & Innovation
17. Rao R, Srinivasan J, Ramasesha S (2022) Comparison of performance of solar photovoltaics on dual axis tracker with fixed axis at 13N latitude. Smart energy for a sustainable future. Exp Photovolt Syst Monit 3(3):187–208

18. Yilmaz S, Riza H, Dogmus O, Dincer F, Akgol O (2015) Design of Two axes sun tracking controller with analytically solar radiation calculations. Renew Sustain Energy Rev (43):997–1005
19. https://www.pv-magazine.com/2024/05/03/pv-driven-hybrid-hydrogen-fuel-cell-system-for-building-applications/

System for Increasing the Accuracy of Digital Information Processing



Farid Agayev and Sevinj Bakhshaliyeva

1 Introduction

The task of developing the theoretical, methodological, and practical aspects of creating highly effective information systems is particularly relevant given the growing demand for information systems, the variety of their designs, and the areas of application. Additionally, the quality and efficiency indicators of both newly created and existing systems are being subjected to increasing demands.

The ability of the system to provide specific technical, economic, and operational characteristics when carrying out the informational, logical, and measuring tasks delegated to it as part of the automation of the serviced facility is reflected in efficiency, a complex indicator of the quality of the information system.

Theoretically, in order to improve the metrological quality of the outcomes of solving measurement, control, diagnostic, and identification problems while taking into consideration current technical and financial constraints, it is crucial to optimize the exchange processes (receipt, transformation, transmission, processing, and presentation) of measurement information in the IMS [1–5]. Significant challenges arise when solving optimization problems in automated information systems. These challenges are typically related to a lack of a priori knowledge about the automation object's properties and characteristics, measured quantities, interfering factors, intermediate and final results of solving basic system functions and tasks, and functional units of measuring and computing subsystems. The challenge of overcoming a priori uncertainty is even more difficult when synthesizing ideal information systems and its subsystems functioning in a dynamic manner. This issue is relevant to the

F. Agayev · S. Bakhshaliyeva (✉)
Azerbaijan State Oil and Industry University, Baku, Azerbaijan
e-mail: sevinj.quliyeba@asoiu.edu.az

F. Agayev
e-mail: farid.aqayev@asoiu.edu.az

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_22

advancement of information measurement technology in general at this time, as high-precision dynamic measurements are becoming more and more common [6, 7].

Improving the efficiency of the nation's fuel and energy complex requires immediate attention to parameter monitoring and energy facility diagnosis. This is explained by the fact that the efficiency indicators of the entire complex are ultimately determined by the quantity of electricity used by power plants and the quality of electricity produced by power units.

Enhancing the precision of measurement outcomes is a crucial objective in any industrial, scientific, or technical domain. The ability of a tool or technique to reduce mistakes and departures from true values is what determines measurement accuracy. Here are a few essential strategies to increase accuracy: Equipment Calibration: Systematic errors can be reduced by routinely calibrating measurement devices. This is particularly crucial for accurate devices like manometers, thermometers, and laboratory scales, among others. Use of High-Quality Instruments: High-sensitivity, stable instruments that are specifically made for the field in question should be utilized for precise measurements.

The implementation of measurements can be divided into three main stages: preparation, measurement experiment, and processing of experimental data. In the preparation stage, the measurement problem is formalized (this includes the definition of the concept of the measured quantity and the way in which the measurement results are derived). At this stage, measurements are also planned (including the selection of the method and equipment and its placement) and the necessary conditions for conducting measurements are provided. A measurement experiment consists of maintaining the required conditions, implementing the interaction of measuring devices with the object of study in accordance with the selected method, and recording the measurement results.

The processing of measurements leads to their logical and mathematical transformations in order to obtain an estimate of the measurement result and its accuracy. Let us explain the measurement stages in more detail. It is advisable to divide the first stage into two parts: setting the measurement problem and planning the measurement experiment. Setting the measurement problem is carried out in the following sequence:—specification of measurement conditions and data on the physical quantity to be studied;—formation of a model of the object and determination of the quantity to be measured;—setting the measurement problem based on the accepted model of the object;—determination of specific physical quantities based on the value to be found for the quantity to be measured (for example, in indirect measurements, directly measured selection of arguments);—formulation of the measurement equation. The most important issue is the formation of a mathematical model (especially in solving complex measurement problems and studying complex objects) [8]. The model of the object is built based on a priori data. The measured quantity is determined as a parameter and characteristic of the object with the help of the adopted model.

Reducing the Effect of environmental elements: A number of environmental elements, including humidity, temperature, and electromagnetic interference, can

have an impact on the accuracy of measurements. Reducing their impact or making up for them with suitable techniques is required to increase accuracy.

Statistical Methods Application: Several measurements can be made and averaged to improve accuracy. This produces more dependable findings and lessens the impact of random errors.

Use of Modern Technologies: For instance, reducing the human element and increasing accuracy can be achieved by automating measurements and processing data using computer systems.

Choosing the Correct Measurement Method: Depending on the kind of object or process being measured, it is crucial to choose the appropriate measurement method because different approaches may produce results with differing levels of precision.

Minimizing Measurement Errors: There are two types of errors: random and systematic. Reducing random errors and minimizing the effects of systematic errors—such as by modifying measurements—are crucial for increasing accuracy.

Operator Training: The credentials and background of the experts doing the measurements are also crucial. Errors during the measurement procedure are less likely to occur when the operator is more skilled and focused [9, 10].

2 Formulation of the Problem

Physical quantities, unless we go down to the quantum level, change continuously. However, digital signal processing works exclusively with discrete quantities, and discreteness manifests itself in two ways - in time quantization and in signal amplitude quantization. This apparent complication is fully justified by the fact that we can use digital computers for processing, completely eliminating the problem of parameter instability, which is so painful in analog processing. No less an advantage is that the cost of digital processing is low and continues to fall, even for very complex types. This allows us to create effective signal processing systems at reasonable costs. How acceptable is such a replacement? Doesn't it lead to a loss of accuracy?

A discrete signal is obtained from an analog signal by a discretization operation - taking samples (measurements) at a time interval T. In principle, digital processing is also possible with non-uniform discretization in time, but this topic is much less developed mathematically and, apparently, is not of such great practical interest. In this operation, it is possible to lose information contained in the signal values in the intervals between samples. The conditions under which it is possible to restore an analog signal from a digital signal obtained from it, i.e., to preserve all the information originally contained in the signal, are expressed by the Nyquist-Whittaker-Kotelnikov-Shannon theorem (depending on the author's preferences, all conceivable combinations of these names are encountered). This requires that the frequency band of the input signal be at least twice as narrow as the sampling frequency, i.e. $fc = 1/2fd$. (A particular formulation is often given, which is true for signals whose frequency band begins with zero frequency—"so that there are no frequencies greater than half the sampling frequency"). If such frequencies are present, the effect of masking

(substitution) of frequencies occurs. A clear manifestation of this can be an illusion, often seen in movies—a rotating wheel suddenly begins to rotate in the opposite direction. Here the frame rate is an analogue of the sampling frequency, and when the wheel makes more than half a turn between successive frames, it appears to rotate in the other direction and at a different speed. For the frequency f, the frequencies (2fc ± f), (4fc ± f), (6fc ± f), etc. are masked under it. The term "aliases" is also used. Failure to take this effect into account can lead to gross errors: thus, in one study conducted in a serious laboratory, the presence of frequencies of 22 and 28 Hz in the electroencephalogram of all patients, in contrast to healthy subjects, was discovered. However, noting that the sampling frequency in this study was adopted as 128 Hz, we see that these frequencies are essentially "ghosts", generating interference at frequencies of 100 and 150 Hz—the second and third harmonics of the network frequency (their source could be, for example, nonlinear devices in the power supply circuits of the equipment, such as rectifiers and transformers). Their registration exclusively in patients was caused by the fact that in a hospital setting, compared to a university laboratory, where the EEG of healthy subjects was recorded, the level of interference is significantly higher.

The fight against the effect of frequency masking (antialiasing) leads to the need for preliminary filtering of the signal, excluding frequencies above half the sampling frequency, and due to the imperfection of real filters, the cutoff frequency is chosen deliberately lower than the theoretically required one, as a rule, three to four times lower than the sampling frequency. This imperfection is not caused by the incompetence of electrical engineers, but is of a fundamental nature. The point is that a signal with a limited frequency band cannot in principle be of finite length, and if it is finite in time, it contains a frequency band of infinite width. (This limitation is quantitatively expressed by the uncertainty relation that connects the pulse length and its frequency band—an infinitely short pulse contains "in embryo" all possible frequencies, and a strictly monofrequency sinusoid must extend from minus to plus infinity.) Therefore, an excessively high-quality filter will have too long a settling time, and an "ideal" one will have an infinite settling time.

According to GOST 14.306-73, the selection of control instruments is carried out on the basis of ensuring the required indicators of the technical control process and analyzing the costs for the implementation of the control process. The most important indicators of the control process include the accuracy, integrity, labor intensity and cost of measurements. The volume, scope, periodicity and duration of control are also taken into account as additional indicators. Ensuring accuracy in the selection of measuring instruments is very important. Thus, the accuracy of the measuring instrument should be sufficiently high, the labor intensity and cost of measurements should be as low as possible, and high productivity and economy should be ensured. Due to the insufficient accuracy of the measuring instrument, the likelihood of considering useful parts as waste increases or, for this reason, waste products are accepted as useful. On the other hand, too high accuracy increases the labor intensity and, therefore, the cost of measurements, leading to an increase in production costs and a limitation of product output.

Therefore, the selection of measuring instruments that provide optimal measurement accuracy is of great importance. The vast majority of measurement work applied in industry is linear measurements. Linear measuring instrument and linear control instruments are divided into contact, contactless, automatic and non-automatic groups. The measuring device for linear measurements consists of a measuring and positioning base and a measuring structure with a calculating structure. The latter is also called a measuring head. Automatic measuring instruments also have a digital adapter-calculator and a self-recording structure. Automatic measuring instruments are divided into measuring control automata, measuring—control sorting automata, semi-automatic machines and active (control) measuring—control devices. Measuring and control instruments can be one-dimensional (for measuring one quantity) and multi-dimensional. All measuring instruments are universal and specialized according to their purpose. Universal devices are used for internal, external measurements, and for measuring surface shape deviations. Specialized devices have a narrower scope of application.

Intense research and development is now being conducted in this area. Unfortunately, there isn't a methodical strategy to developing systems for tracking energy facility diagnostic metrics.

A methodical strategy to developing an information system for the aforementioned aim is attempted to be developed in this study. Asynchronous motors, which are the foundation of the majority of electric drives and are responsible for almost 40% of global electricity consumption, are the subject of this case study [11].

Any measured quantity is a sinusoidal harmonic signal, such as network voltage. There is only one harmonic in this signal. We get a non-sinusoidal signal after measuring an analog value. A sinusoid with extra odd harmonics is what it is. These harmonics are caused by interference and noise that are superimposed on the primary useful signal, making it noisy. As a result, 80% of the time, the measured signals are not sinusoidal [12–15].

Although they are not always visible, curves of continuous signals' instantaneous values are the most instructive. Therefore, the real-time determination of integral signal parameters are crucial for the management and analysis of the operation of regulated processes and objects [16, 17]. The latter describe the overall quantities of materials and energy received and added within a specific period of time, as well as the object's operational indicators and parameters, which are the mean values of the measured values, etc. By performing discrete averaging (DA) or discrete integration (DI) of parameters that vary continuously over time, IPS can be specifically determined utilizing digital methods and tools.

The increasing global use of power, especially due to the widespread use of asynchronous motors, has made real-time monitoring and diagnostics of energy facilities even more important. In addition to being essential to the great majority of electric drive systems, these motors also contribute significantly to the world's energy usage. The need for efficient ways to monitor diagnostic indicators is growing in significance as energy systems become more complicated. Predictive maintenance techniques are essential for reducing downtime and increasing energy production

efficiency, and they are improved by the integration of real-time data collecting systems and diagnostic instruments.

3 Solution of the Problem

Two other problems where the discrete Fourier transform (DFT) is successfully used exploit one of its most important properties, namely that the DFT of the convolution of two signals is equal to the product of the DFTs of these signals. Since the output of any linear system is the convolution of the input signal and the response of the system, the DFT of the system's output is obtained by multiplying the DFT of the input signal and the DFT of the response, and if we have access to the input and output of the system and can compute the Fourier transform of them, the transfer function is obtained by simple division. In the same way, if we need to calculate a convolution, and the length of the segments being convolved is so large that we have to worry about computational efficiency, we can calculate the DFT of the vectors being convolved, multiply them, and find the inverse DFT to get what we are looking for (there is a trap here—by applying this theorem "head-on", having obtained a cyclic convolution, calculating which and having reached the end of the segment, we start going along it from the beginning—this is rarely what we wanted; to prevent this, the segments are supplemented with the required number of zeros).

Speaking so confidently about the computational advantages of the DFT, we have not yet given the slightest basis for this confidence. Calculating the DFT using the given formula requires calculating n coefficients, each of which depends on n elements of the original segment, so the number of operations cannot be less than n2. Without dwelling on the technical methods of calculation, I will only note that there is a remarkable algorithm, or rather a whole family of algorithms, known as the "Fast Fourier Transform", which reduces the operating time to n log n operations. This is achieved by skillfully applying the "divide and conquer" principle, which allows you to avoid repeating work that has already been done, and the adjective "fast" should not be interpreted as "simplified", "inaccurate" or "incomplete". With exact arithmetic, the result of a calculation according to the FFT algorithm and the calculation according to the above formula coincide, moreover, if we move on to real, inaccurate arithmetic, it turns out that the FFT gives a higher accuracy of calculation than direct calculation. It is this algorithm that makes the use of the Fourier transform so useful in computational terms. Variants of the Fourier transform also find a certain application. Limiting ourselves to cosine waves, we arrive at the Discrete Cosine Transform, which allows us to represent only even signals (i.e. f (x) = f (− x)) while eliminating discontinuities. The Discrete Sine Transform also has its niche. The Hartley transform is of some use, where the basis functions are not sines and cosines, but their sums, which allows us to achieve some gain in performance by eliminating complex arithmetic. However, this transform is too close to Fourier and cannot be considered more than a convenient calculation technique. Transforms

with other families of basis functions are similar to Fourier. For example, the Walsh-Hadamard functions take values of + 1 and − 1, as if they were chopped out with an axe in an attempt to sculpt a caricature of sinusoids. Sawtooth or triangular functions have also been proposed for use. They had very modest practical success (mainly in video compression problems), and there are no theoretical grounds to expect these functions to appear in biological systems, even as small as the Fourier basis functions. However, fast transformation algorithms have already been developed in this family, which may justify their practical application.

One can approach the construction of a family of functions for signal decomposition from another side. Starting with a description of a real process, albeit a very rough one, one can obtain a differential equation whose eigenfunctions will be a certain family of functions (say, Bessel functions are eigensolutions of the membrane differential equation). They may be much more suitable for describing this class of processes than standard Fourier, but this path is difficult and laborious.

Finally, problems solved using Fourier can be solved in other ways. Thus, autoregressive and ARMA methods can be used to estimate spectra, the impulse response function can be estimated using regression estimates, etc. The only always true statement here is this: there is not and cannot be a single always best method. The characteristics of the generation and collection of data, the means and methods of analysis available to us, the prospect of subsequent application of our results will always be decisive in choosing the best method here and now.

The signal is converted into digital form by an analog-to-digital converter (ADC). Almost all of them use a binary system of representation with a uniform scale (however, one should not forget about non-uniform scales, such as logarithmic α-law and μ-law, used in digital telephony, binary-decimal data representation and some other, sometimes very exotic solutions), characterized by a certain number of digits. Increasing the number of digits improves the measurement accuracy and allows to expand the dynamic range of the measured signals. Information lost due to the lack of ADC digits is not recoverable, and there are no simple criteria similar to the Nyquist-Kotelnikov-Shannon (NKS) theorem, showing when such recovery could take place. There are only error estimates, for example, through the noise power generated by an error in the last digit, but they are not always satisfactory. For example, human hearing is logarithmic (Weber-Fechner law), and low-power interference in a range where the signal power is low is too clearly distinguishable.

Speaking about the signal, we mentioned that it can be real or complex, i.e. consisting of real and imaginary parts. Complex numbers do not appear here as a tribute to mathematical generality (although the ability to write one exp (a + i * b) instead of two different expressions eacos(b) and easin(b) greatly simplifies life and calculations…) and even more so, one should not see here mathematical mysticism (“square root of minus one”). It is natural, for example, to represent in-phase (coinciding in phase with some other signal) and quadrature (differing in phase from this signal by 90°) components of the signal. Complex numbers also naturally arise when solving differential equations representing many natural and technical objects, especially in connection with the description of oscillatory processes.

In general, complex numbers can be considered simply as a way to combine two different numbers in one number, but in such a way that it is possible to operate with them simultaneously. One of the numbers used is represented by the real part, and the other by the imaginary part of the complex number $a + bi = r (\cos (\varphi) + i \sin (\varphi)) = r \exp (i\varphi)$. It is convenient to represent them as points on a plane—one coordinate will represent the real part, the second the imaginary part. Then their addition will become the addition of vectors, and their multiplication will be the rotation of vectors on the plane (with the multiplication of their lengths equal to the moduli of these numbers; angles equal to the arguments of these numbers are added). Successive raising of a complex number to a power becomes the rotation of the vector expressing this number around the origin.

The projection of this vector on one of the axes will experience increasing, decreasing, or constant-amplitude oscillations—depending on whether the modulus r of a given complex number is greater than, less than, or equal to one. Although real objects are all nonlinear to a greater or lesser extent, we will begin our consideration with the simplest, but important and often sufficient for a satisfactory description of the case. And the simplest to consider are linear systems. Linear systems are those for which superposition takes place (that is, the response to the sum of two input signals will be equal to the sum of the responses to these signals separately—"the whole is equal to the sum of its parts") and homogeneity (that is, the response to an input signal amplified by a certain number of times will be amplified by the same number of times—"size does not matter"). Linearity allows us to consider our object of study in parts, and homogeneity—on a scale convenient for us, in particular, by feeding simple test signals to its input and studying its reaction to them, and representing complex real signals as a sum of simple ones. Obviously, for real objects, the property of linearity can take place only approximately and in a certain range of input signals. An example of nonlinear technical systems is a diode that leaves only positive half-sine waves from a sinusoidal signal, or a transformer whose yoke saturation with a magnetic field leads to "blunting" of the sinusoid peaks (and from the point of view of the frequency spectrum, this is manifested by the appearance of frequencies that are twice, three times, etc. greater than the fundamental frequency—harmonics, and sometimes frequencies that are a multiple of the frequency—subharmonics). Perhaps of greater interest to the reader are nonlinearities in biological systems, such as the threshold action expressed by the "all or nothing" law, or nonlinear transformation in the sense organs, approximated by a logarithm in accordance with the Weber-Fechner law. Nevertheless, as a rule, the study begins with the assumption of linearity, and then introduces corrections for nonlinearity.

The so-called method of digital processing of the results of direct measurements of instantaneous signal values during the averaging (integration) period is one of the most used digital techniques for measuring IPS [18]. The potential to add processing power to the measurement channels when addressing this type of measurement problem explains the recent sharp rise in interest in this approach [16–19]. When measuring the IRS of electrical signals with complex shapes (non-sinusoidal signals) digitally, the issue becomes much more challenging. However, main information converters with output signals in the form of direct and alternating currents and

voltages are becoming more and more common due to the well-known benefits of electrical methods for monitoring and measuring physical quantities [19–21].

Particularly in electrical energy generating and conversion facilities, AC signals are more informative and sometimes the only way to get measurement information [22–25].

It is important to keep in mind that practically all algorithms use the DA operator when switching from continuous integration to numerical integration using the techniques.

A quadrature formula $\tilde{C}^{(m)}[0, T]$, of the following kind function is used to conduct the DE of a class of continuous functions:

$$\bar{I} = \frac{1}{M} \sum_{i=0}^{M-1} y(iT_0)$$

where $T_0 = \frac{T}{M}$ – step-by-step sampling in time.

In order to monitor parameters and evaluate the current condition of electric motors, the work offers a methodical approach to electric motor design.

Currents and powers in idle and short circuit modes can be measured and processed to perform electric motor diagnostics. The precision with which these parameters are measured determines the dependability of monitoring and diagnostics.

The voltage U in the network and the current I flowing through the object determine the overall power:

$$P = IU \tag{1}$$

The outcome will be if an error is introduced in the measurement of any of the power-determining parameters, which is always the case.

$$\tilde{P} = \tilde{I}\tilde{U} = (I + \Delta I)(U + \Delta U), \tag{2}$$

where

ΔI resulting measurement error for current,
ΔU resulting measurement error for voltage.

The power measurement result will deviate from the actual value by the following because of errors:

$$\left|P - \tilde{P}\right| = \left|I \cdot U - \tilde{I} \cdot \tilde{U}\right| = |\Delta I \cdot U + \Delta U \cdot I + \Delta I \cdot \Delta U|. \tag{3}$$

The usable harmonic signal is noisy during the current and voltage measurement procedure; that is, high-frequency noise (interference) is overlaid on the signal as a result of the measurement. Consequently, the pure harmonic signal becomes inharmonic. Furthermore, the measurement signals u(t) and i(t) are non-sinusoidal in real circuits due to the inclusion of extra harmonics.

In particular, we will determine the spectrum of the product of these signals and how it will change after applying a discrete averaging filter, which is required for digital measurements of the integral characteristics of AC signals. Let's look at the scenario where a non-sinusoidal signal interacts with random-centered noise.

$$\overline{x}(nT_0) = \frac{1}{N}\sum_{i=0}^{N-1} x_i(nT_0), \ldots \tag{4}$$

where

N the quantity of averaged samples,
T_0 the sampling time step.

Consider a sinusoid with a frequency of 50 Hz and 13 odd harmonics placed on it as an initial non-sinusoidal signal:

$$x(k) = \sum_{m=1}^{2N+1} \sin(2\pi km\omega), \quad N = \overline{1,6}, \ k = 1, 2, \ldots \tag{5}$$

where

ω the signal's fundamental frequency,
k the number of distinct samples,
m numbers of harmonics.

Consider a random signal produced by the built-in Matlab rand function that is uniformly distributed on the interval [0, 1] as a random centered noise. The fast Fourier transform, or fft function, was used in the Matlab software environment to compute the original signals' spectra (Fig. 1).

We determine the product of the initial signals in line with formula (3) (assuming that one of the signals is U and the other is not). The theory of spectral analysis [10, 11] states that as time signals are multiplied, their corresponding spectra are added.

Figure 2 displays the final signal as well as the entire spectrum. Because of the overlap of several frequencies in the random signal, the resulting spectrum has many additional harmonics, as seen in Fig. 2b.

We next run the resultant signal through a discrete averaging filter to eliminate any more harmonics. In order to accomplish this, we split the total time period into smaller time periods, and we calculate the average value of the resulting signal samples in each of those intervals using the formula:

$$y(k) = \frac{1}{N_k}\sum_{i=l}^{l_{k+1}} y(i), \quad k = \overline{1,K}, \tag{6}$$

where

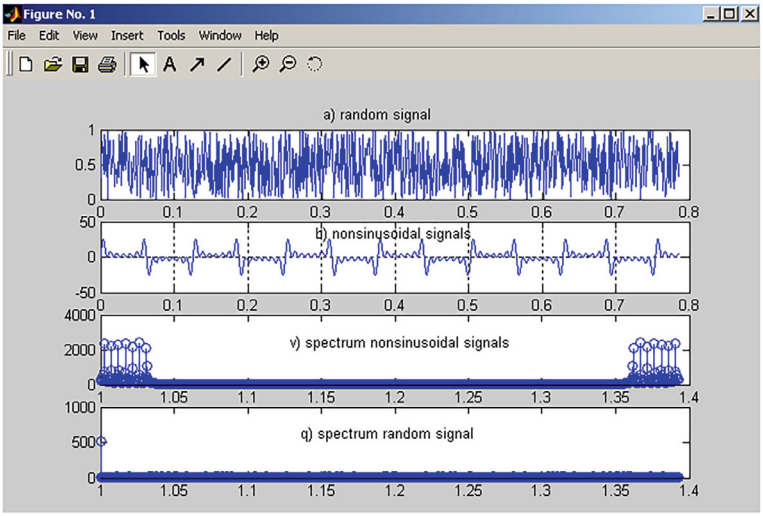


Fig. 1 Signal sources and their spectrum

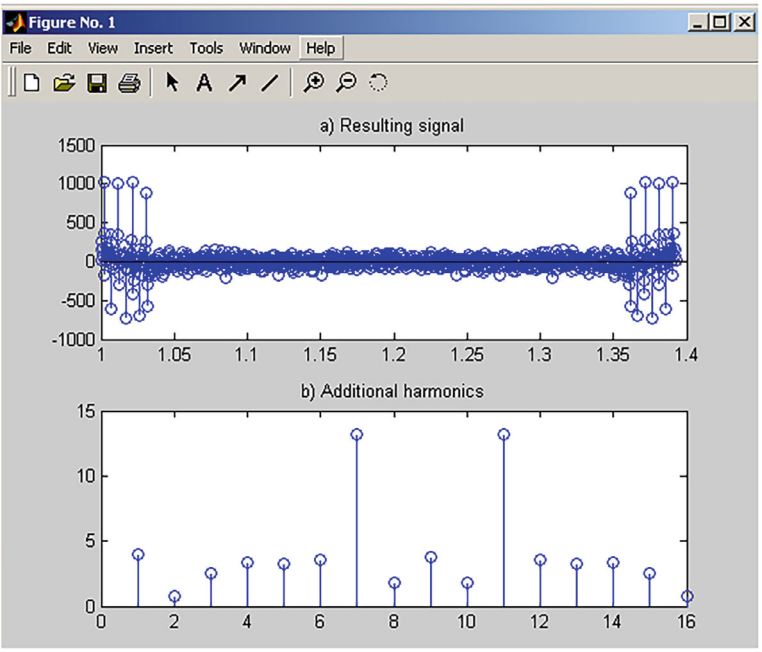


Fig. 2 The spectrum of the resultant signal

N_k the number of subintervals,
l_k and l_{k+1} the interval boundaries,
K the number of samples in each interval.

We will determine the spectrum of the filter output signal (Fig. 3) in order to examine the harmonic composition of the received signal. According to the computations displayed in Fig. 3, the impact of extra harmonics in the spectrum was greatly diminished. The spectrum was free of side components when their amplitudes dramatically dropped (by several orders of magnitude). Therefore, it is sufficient to apply a discrete averaging filter to the measurement result in order to reduce the random noise in non-sinusoidal signals. Only then may the corrected signal be used as a parameter for monitoring and diagnostics.

We carried out an analogous computation concerning a non-sinusoidal signal and a systematic error of order 2, which is represented by a function of the following form:

$$\overline{\varepsilon}(t_k) = \sum_{j=0}^{2} a_j t_k^j, \tag{7}$$

where t_k—discrete samples of the systematic error., a_j—the polynomial coefficients that describe the slowly varying error.

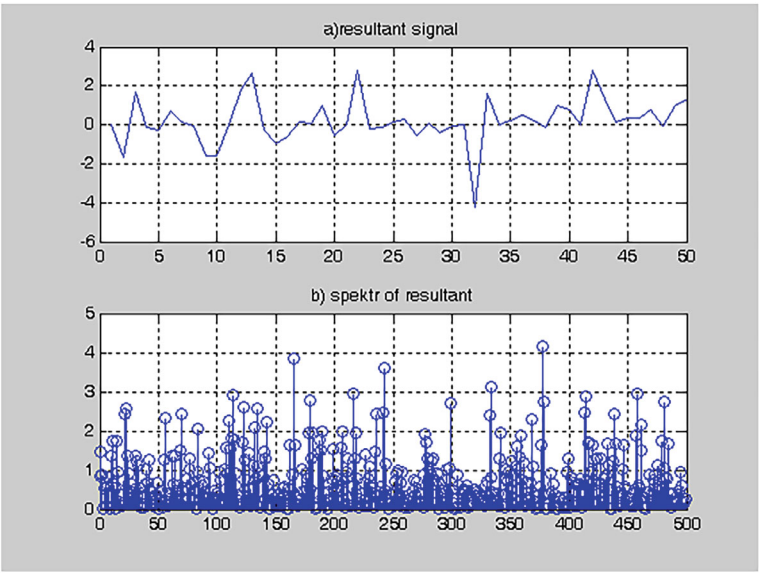


Fig. 3 Spectrum and signal at the filter's output

Using a discrete averaging filter resulted in a considerable reduction in the signal spectrum. The systematic error is more effectively repaired, according to a comparison of the experimental findings with random and systematic errors.

Therefore, the previously measured monitoring parameters should be corrected by applying a discrete averaging filter to them before monitoring and diagnostics are conducted based on the corrected results when building an IIS for electrical equipment monitoring and diagnostics.

Other characteristic errors are also examined using the aforementioned method. These include the impact of level quantization during the analog-to-digital conversion of signal samples (quantization noise), random variations in the instantaneous readings of the measuring signals (dating error), and their digital processing (rounding noise).

4 Conclusion

Therefore, in order to achieve balanced indicators of the metrological, structural-algorithmic, and functional efficiency of the developed tools, the suggested methodical approach to the creation of an information system for the intended purpose consists of a combined analysis of the measurement processes and corrective filtering.

Corrective filtering was performed in the Matlab software environment. From the results of the simulation modeling, it is obvious that the errors have significantly decreased.

References

1. Sergienko AB (2005) Digital signal processing. Peter, St. Petersburg, p 604
2. Baranov LA (1990) Level quantization and time sampling in digital control systems. M.:Energoatomizdat, p 304
3. Mehdiyeva AM, Mehdizadə EK (2014) Information-measurement system development for controlling of parameters of measurement. Stand Glob J Bus Manag 1(2):030–033
4. Camponogara E, Nakashima PH (2006) Solving a gas-lift optimization problem by dynamic programming. Eur J Oper Res 174(2):1220–1246
5. Ponomarenko VS, Arefiev UI (1998) Cooling towers of industrial and energy enterprises, p 365
6. Gladkov VA, Arefiev YY, Ponomarenko VS (1976) Fan cooling towers, 2nd edn, Revised and ext., M.: Stroyizdat, p 216
7. Taria L, Localized corrosion of stainless steel in chloride, sulfate and thiosul-fate containig environments. Corros
8. Mehdiyeva AM, Bakhtiyarov IN (2019) Analysis of reliability indicators of multiservice corporate networks based on SDN technologies. In: Proceedings of the international symposium reliability and quality. Penza, T. 2, pp 114–116
9. Ibrahimov BQ, Mehdiyeva AM, Baxtiyarov IN (2020) Investigation of methods for improving the noise immunity of message processing systems. In: Problems of information communications. Science Magazine. Minsk, Belarusian State Academy of Communications, no 1 (11), pp 13–18

10. Mehdiyeva AM et al (2015) Conversion and initial processing errors of measurement results. Am J Circuits Syst Signal Process Am Inst Sci 1(3):56–59
11. Allahverdiyeva NR, Mehdiyeva AM, Mehdizadə EK (2010) Increase the accuracy of conversion and digital processing of electrical signals. In: Information-measuring and control systems, no 9. Moscow, pp 69–74
12. Makarova IV (2007) M: Finance and statistics. Informatics: textbook/ed. prof.
13. Mehdiyeva AM, Bakhtiyarov IN, Bakhshaliyeva SV (2023) Increasing the immunity of information transmission and fault tolerance of the path. In: Lecture notes on data engineering and communications technologies, vol 166, Nepal. pp. 775–784
14. Mehdiyeva AM (2023) Types of accidents, their causes and prevention measures actual problems of modern science. In: Proceedings of the IV international scientific and practical conference. Jan 31–Feb 3, Boston, USA, pp 441–444
15. Shapiro L, Stockman J (2006) Computer vision. M.: Binom
16. Gonzalez R, Woods (2005) Digital image processing. M.: Tekhnosphere
17. Ayficher ES, Jervis BW (2008) M Digital signal processing: a practical approach. M.: Publishing house. Williams House
18. Jähne M (2005) Digital image processing. Springer, Berlin Heidelberg
19. SmithS (2008) Digital signal processing. A practical guide for engineers and scientists. C. Smith, lane from English—M. Dodeka-XXI
20. Ibrahimov BG, Ismaylova SR (2018) The effectiveness NGN/IMS networks in the establishment of a multimedia session. Am J Netw Commun 7(1):1–5
21. Lebedev SV (2002) Firewalling. Theory and practice of protecting the outer perimeter. M.: Publishing house of MSTU im. N.E. Bauman, p 304
22. Shuvalov VP, Egunov MM, Minina EA (2015) Providing indicators of reliability of telecommunication systems and networks. M.: Hotline—Telecom, p 168
23. Stearns SD (2018) Digitale verarbeitung analoger signale. Digital signal analysis. Walter de Gruyter GmbH & Co KG
24. Mityukhin AI (2018) Applied information theory: educational method. Allowance. BSUIR, Minsk
25. Mityukhin AI (2016) Digital speech processing and image analysis: textbook—method. Allowance. BSUIR, Minsk

Performance Evaluation of Different Shapes of Patch Antenna for 5G Applications at 25 GHz



Afsana Sultana Alvee, Md. Sohel Rana, Najifa Islam, Mahamud Hasan, and Md. Abdul Ahad

1 Introduction

The globe is experiencing ongoing advancements in multiple domains, particularly in communications and antennas. There has been a substantial rise in demand for high-specification tiny antennas suitable for integration into diverse applications such as 5G, wireless communication systems, and the medical area. Patch antennas are distinguished by their compact dimensions and suitability for applications necessitating limited space, such as mobile phones, Bluetooth devices, and other devices [1, 2]. There has been a substantial amount of advancement in the field of communication systems in the modern era, which has led to the development of a great deal of significant research in the field of antennas, which also vary depending on the application. Electronic devices known as antennas are employed for either capturing electromagnetic fields or radiating them into space. Oscillator antennas, wire antennas, patch antennas, array antennas, lens antennas, and reflector antennas are some of the several types of antennas that are available. As a result of its compact size, low cost, and low weight, the patch antenna is utilized for a broad range of applications, including wireless communication and radars. In addition to square, circular, rectangular, triangular, elliptical, and dipole forms, patch antennas are available in many different shapes [3]. Figure 1 illustrates the physical architecture of the patch antenna. In order to satisfy the ever-increasing need for high data rates and broad bandwidths, the rapid development of wireless communication technologies has prompted the investigation of new frequency bands [4]. The advent of 5G technology represents a dramatic transformation in the way people communicate and

A. S. Alvee · Md. S. Rana (✉) · N. Islam · M. Hasan
Northern University of Business and Technology Khulna, Khulna, Bangladesh
e-mail: sohel.rana@uits.edu.bd; sohel.rana@nubtkhulna.ac.bd

Md. A. Ahad
Washington University of Science and Technology, Alexandria, VA, USA

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_23

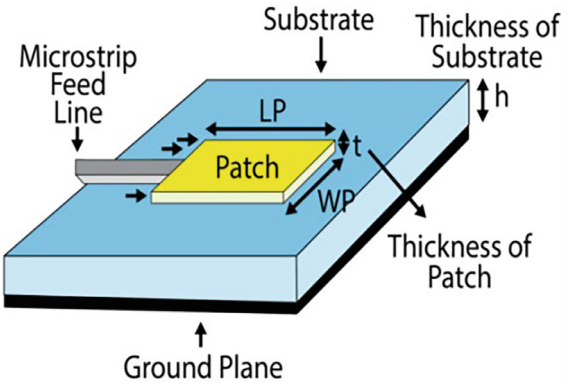


Fig. 1 The patch antenna's physical construction

engage with one another around the world. Nevertheless, when we contemplate this revolution in wireless communication, it is of the utmost importance to recognize the vital role that the development of mobile communication has played in driving forward the advancement of modern technology. Mobile communications have had a significant impact on the technological landscape of our society, beginning with the era of cumbersome devices and continuing through the growth of smartphones and cellular networks, and now culminating with the introduction of 5G [5]. Designing antennas that are capable of operating effectively at frequencies of 25 GHz, however, presents a number of obstacles, such as large propagation losses, the need for precision in production, and the selection of materials. Important considerations include the impedance matching, bandwidth, and radiation efficiency of the antenna, all of which are influenced by the substrate material that is selected. As a result of its advantageous dielectric properties, low-loss tangent, and mechanical flexibility, the polyimide substrate that we use in this investigation is an excellent choice for the implementation of 5G wireless applications [6].

2 Literature Review

Wireless communication is one of the most dynamic and fastest-developing technological fields within communication. Wireless communications are employed in numerous contexts, encompassing industrial applications and human endeavors. A patch antenna is the most popular antenna for wireless communication [7]. It is favored for various applications because of its compact dimensions, affordability, and excellent efficiency [8]. Extensive research has been undertaken and is ongoing to develop a more flexible and resilient framework for wearable microstrip antennas. The domain of wireless communications has rapidly and extensively evolved, resulting in a significant surge in demand for the design of low-profile and multiband

antennas for mobile terminals, especially for applications utilizing 5G data transmission. In this section, a number of publications covering microstrip patch antennas for Ka-band applications that have been published in the past are discussed.

This article introduces [9] a tiny patch antenna designed for 5G millimeter-wave mobile terminal applications. The unit cell is a planar dipole fabricated on a singular substrate layer, attaining substantial bandwidth and gain. This article presents [10] a novel mixed-coupling substrate-integrated waveguide (SIW) filter design that has been proposed. The filter has a dual-band response and increased selectivity. The architecture that has been suggested is extremely adaptable, and it features three separate coupling parameters. These SIW filters are appropriate for large-scale manufacture for 5G applications, as they are both cost-effective and high-performing. For 5G communication systems [11], a novel design for a reconfigurable antenna has been proposed. Additionally, the design intends to attain a good bandwidth and gain while simultaneously improving efficiency. The numerical results that were obtained through simulations are validated by fabricating a prototype of the antenna and putting it through practical tests. The results of the prototype reveal that it is in good agreement with the measurements obtained through simulations. This research paper [12], introduces a broadband antenna designed for 28 GHz 5G communications. The antenna attains high radiation efficiency and gain and is considered appropriate for future 5G applications due to its straightforward design, compact dimensions, and low manufacturing cost. This work shows [13] a circular patch antenna equipped with a parasitic reflector element. This antenna was designed specifically for Ka-band applications operating at 28 GHz. The design alleviates these problems by optimizing the arrangement of parasitic elements to achieve high gain, bandwidth, and constant VSWR performance across the operating range. The present study offers [14] an exhaustive examination of the attributes and specifications of a resonant patch antenna intended for wireless communication. The work elucidates the following aspects: impedance bandwidth, gain, and minimal return. The proposed antenna for 5G cellular communication is a formidable contender. The H-plane radiation pattern, E-plane radiation pattern, and VSWR demonstrations are further features that distinguish it from other devices. In this work [15], a patch antenna design that is intended for use in 5G applications is presented. It was suggested displays a significant gain and improved radiation efficiency. Qureshi et al. presents [16] design of a circular patch antenna intended for broadband millimeter-wave applications in 5G technology. To achieved good S_{11}, VSWR, radiation efficiency, gain, and bandwidth from the design. It is anticipated that applications requiring broadband 5G connectivity will be compatible with this design. Table 1 presents a comparison between previously published work and the current study.

Table 1 Previously published literature with the planned work

References	Present	[14]	[15]	[16]	[17]	[18]
Tools	CST	CST	CST	CST	CST	CST
Substrate	RT6002	RT5880	RO3210	Teflon	RO3003	RT5880
(ε_r)	2.94	2.2	10.2	2.1	3	2.2
Thickness of substrate (mm)	0.6	0.5	1	1.57	0.25	1.2
Tan (δ)	0.0012	0.0009	0.003	0.00028	0.0013	0.0009
Applications	5G	5G	5G	5G	5G	5G

3 Method and Mathemathical Analysis

To begin designing an antenna, the first step is to choose the operating frequency for a specific application. The proposed design selected operating frequency 25 GHz (Ka Band). Second, using CST (Version 2019) software, an antenna should be designed, and then the antenna should be evaluated. In the third step, the substrate material for the antenna design must be selected, which will be used to design the proposed antenna. Rogers RT6002 is used to design the proposed antenna, which dielectric constant of 2.94 mm, loss tangent of 0.0012 and thickness of substrate of 0.6 mm. Fourth, optimize the antenna characteristics by adjusting the position of the meta structure on the antenna until the desired antenna requirements are satisfied. This process should be repeated until this is accomplished. The fifth step is to simulate and evaluate the outcomes. The proposed antenna is developed for the Ka-band and exhibits positive gain, favorable VSWR, enhanced bandwidth, better efficiency, and minimal return loss.

Equations (1) through (5) can be used to calculate the length and width of a single antenna patch. Also, Eq. (6)–(7) are calculating the ground of width and length [19, 20]. Table 1 presents the proposed antenna's initial design value, and Table 2 specifies its optimal dimensions.

Step: 01: Calculation the width of the patch

$$\mathrm{Wp} = \frac{\mathrm{c_0}}{2\mathrm{f_r}\sqrt{\frac{\varepsilon_r+1}{2}}} = 4.274 \text{ mm} \tag{1}$$

Step: 02: Calculation the effective dielectric constant

Table 2 Presents the initial design value of the antenna

Parameter	W_P	L_P	L_g	W_g	Hs	ε_r	t	δ
Value (mm)	4.274	2.284	5.884	7.874	0.6	2.94	0.035	0.0012

$$\varepsilon_{reff} = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2} \times \sqrt{\left(1 + \frac{12hs}{WP}\right)} = 3.559 \text{ mm} \quad (2)$$

Step: 03: An extended version of the equation for length is as follows:

$$L_{ext} = \frac{c_o}{2f_r\sqrt{\varepsilon_{reff}}} = 3.18 \text{ mm} \quad (3)$$

Step: 04: Extension of the length of antenna of the equation

$$\Delta L = 0.412 \frac{\left(\frac{Wp}{hs} + 0.264\right)(\varepsilon_{reff} + 0.3)}{(\varepsilon_{reff} - 0.258)\left(\frac{Wp}{hs} + 0.813\right)} \quad (4)$$

Step 05: Determination of the length of the antenna

$$Lp = L_{ext} - 2 \times \Delta L = 2.284 \text{ mm} \quad (5)$$

Step: 06: Dimensions of the antenna ground plane, encompassing its length and width:

$$L_g = 6h + L_P = 5.884 \text{ mm} \quad (6)$$

$$W_g = 6h + Wp = 7.874 \text{ mm} \quad (7)$$

4 Antenna Design and Performance Evaluation

For this investigation, the CST program, which is widely acknowledged as a design simulation platform for antennas, was used to achieve the results. The starting size of the various antenna sections was determined using calculations and measurements collected from a previous investigation. Figure 2 illustrates the proposed antenna design with CST software. Table 3 represents the dimension of the proposed antenna.

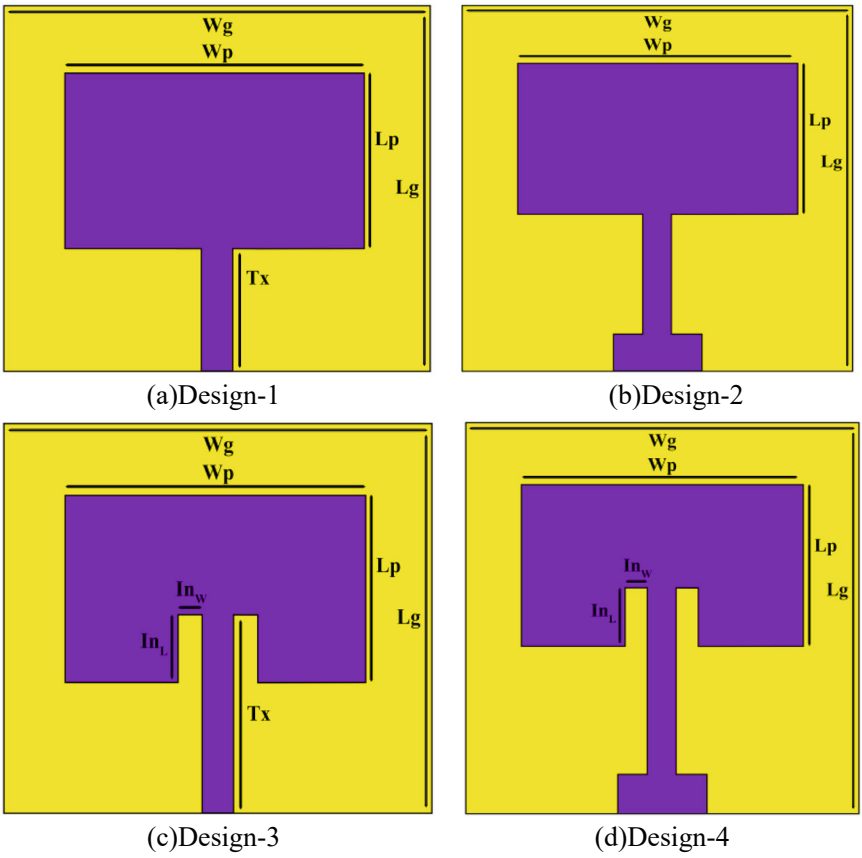


Fig. 2 Proposed design of the patch antenna utilizing CST

4.1 Reflection Coefficient

The reflection coefficient (S_{11}) quantifies the amount of power reflected from the antenna. It is an essential factor in antenna design [21]. To attain the requisite performance level, the return loss value must be less than − 10 dB [22]. At 25 GHz, the reflection coefficients of antenna designs 1, 2, 3, and 4 are − 11.14 dB, − 19.5 dB, − 14.87 dB, and − 48.52 dB, respectively. Figure 3 illustrates the S_{11} parameters of antenna designs 1, 2, 3, 4.

Table 3 The dimensions of the proposed antenna that have been optimized

Parameter	Design-1	Design-2	Design-3	Design-4
	Value (mm)	Value (mm)	Value (mm)	Value (mm)
W_g	6.5	6.5	6.5	6.5
L_g	9	9	9	9
W_P	5.61	4.12	5.3	4.12
L_P	2.85	2.98	2.995	3.02
h_S	0.6	0.6	0.6	0.6
Tx_W	0.561	0.412	0.53	0.412
ε	2.94	2.94	2.94	2.94
f_W	–	1.373	–	1.373
f_L	–	1.003	–	0.996
$Inset_W$	–	–	0.176	0.1373
$Inset_L$	–	–	0.748	0.755

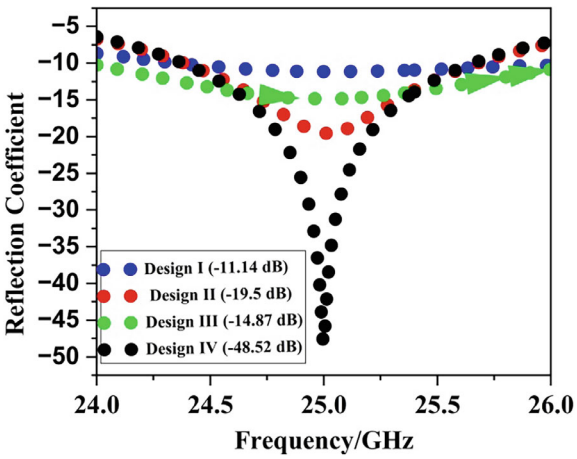


Fig. 3 Analysis of the reflection coefficients for antenna designs

4.2 Voltage Standing Wave Ratio (VSWR)

For a patch antenna, the voltage standing wave ratio (VSWR) must not exceed two throughout the full frequency range. This value should be 1 [23]. Figure 4 illustrates the VSWR graph. A VSWR of 1.767, 1.237, 1.441, and 1.007 are achieved by antenna designs 1, 2, 3, 4, respectively.

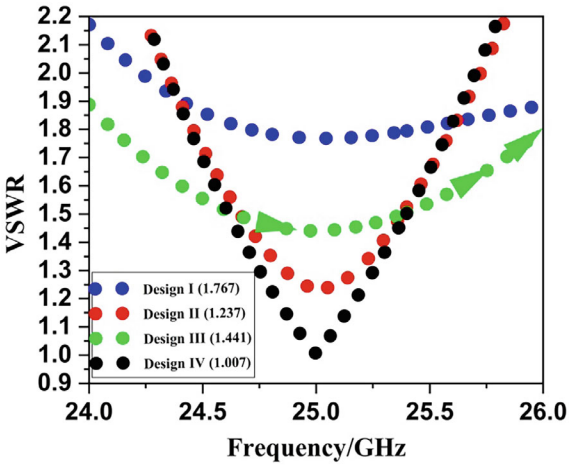


Fig. 4 Evaluation of VSWR of the antenna designs

4.3 Radiation Pattern

A radiation pattern is a direct description of the power that an antenna possesses in space [24]. The electric field (E-field) and the magnetic field (H-field) are the two primary components of the electromagnetic field that constitute an antenna's field of influence. Figure 5 shows an illustration at $\Phi = 0°$, which is also known as the radiation pattern's E-plane. In contrast, the arrangement under consideration, which maintains a constant angle of 90° and fluctuates between 0° and 360°, $\Phi = 90°$ in Fig. 6, and is also referred to as the H-plane. Table 4 contains a visual representation of the properties of the simulated radiation pattern.

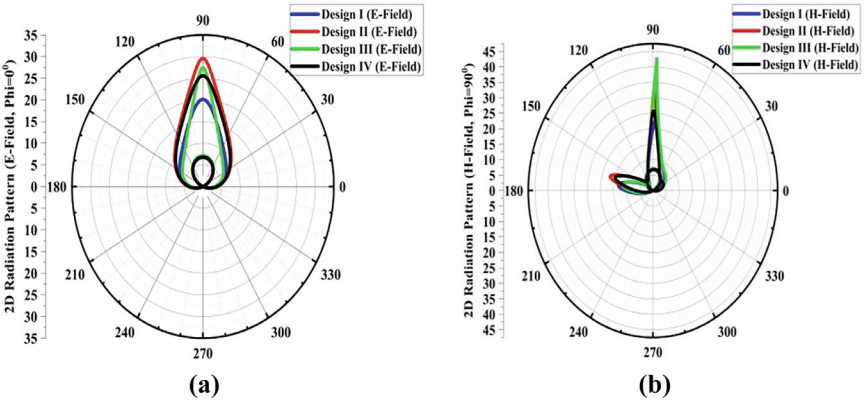


Fig. 5 Radiation pattern in two dimensions (antenna design-1, 2,3, 4)

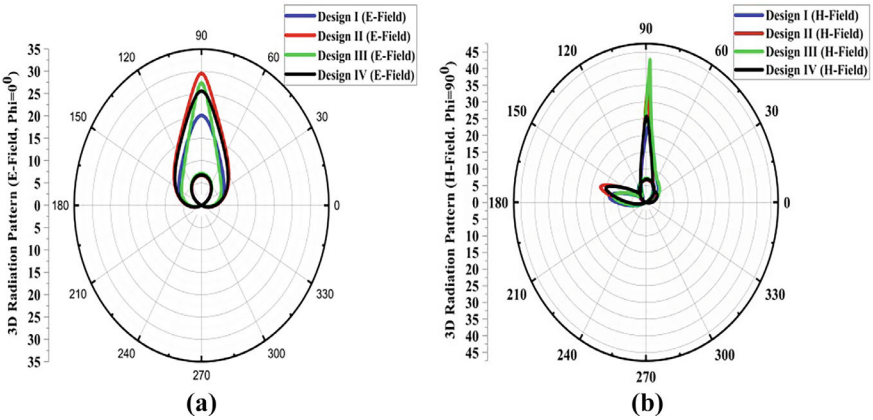


Fig. 6 Radiation pattern in three dimensions (antenna design-1, 2,3, 4)

Table 4 Antenna's radiation pattern parameters

Parameter	Design-1	Design-2	Design-3	Design-4
3 dB Beamwidth (E-plane)	81.5 deg	85.7 deg	83.2 deg	82.3 deg
Main lobe direction (E-plane)	00.0 Deg	00.0 deg	00.0 deg	5.0 deg
Side lobe label (E-plane)	–	–	–	− 10.3 dB
3 dB Beamwidth (H-plane)	82.9 deg	87.4 deg	77.1 deg	82.3 deg
Main lobe direction (H-plane)	1.0 Deg	3.0 deg	1.0 Deg	5.0 deg
Side lobe label (H-plane)	− 8.6 dB	− 10.2 dB	− 9.2	− 10.3 dB

4.4 Gain and Directivity

Gain is an essential component of patch antennas since it allows for the concentration and direction of radiated energy in particular directions. The gain is determined by the dimensions of the antenna, its configuration, the direction it faces, and the transmission frequency [24]. As Fig. 7 shows, the gain of the antenna designs 1, 2, 3, and 4 is 6.13 dBi, 5.94 dBi, 6.01 dBi, and 6.22 dBi, respectively. Figure 8 also displays the directivity values for designs 1, 2, 3, and 4: 6.91 dBi, 6.71 dBi, 7.1 dBi, and 6.86, respectively.

4.5 Impedance Matching and Smith Chart

The antenna design assumes that the power line impedance is 50 Ω. At 25 GHz the input impedance values for the proposed antenna are shown in Fig. 9. The Smith chart is a specialized instrument used to depict the impedance of transmission lines

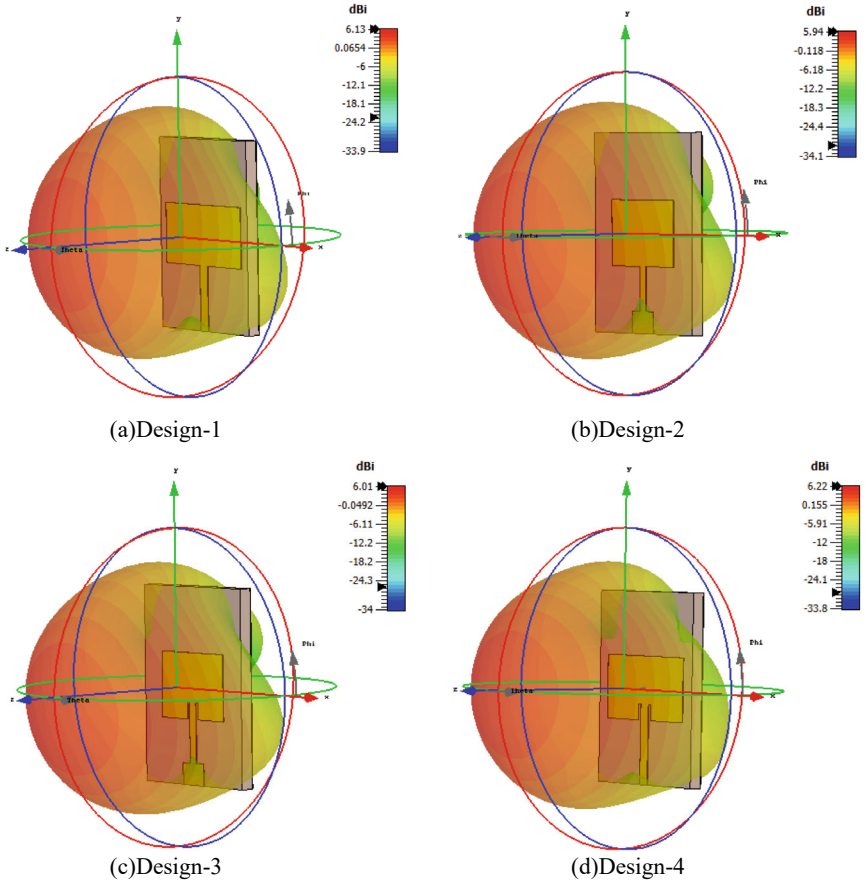


Fig. 7 Three-dimensional representation of the gains of antenna designs (antenna design-1, 2, 3, 4)

and antenna systems, with frequency as the variable [25]. Figure 10 represents the Smith chart included in the proposed model.

5 Comparative Research and Analysis

In addition to publishing many articles, researchers are also conducting active research on patch antennas. It is possible to use a variety of frequencies inside the patch antenna for the band, which is suitable for applications that include wireless communication. It was determined that the frequency of 25 GHz was ideal for any wireless 5G applications. A comprehensive presentation of the results of the proposed antenna is provided in this section. The proposed antenna's outcomes were

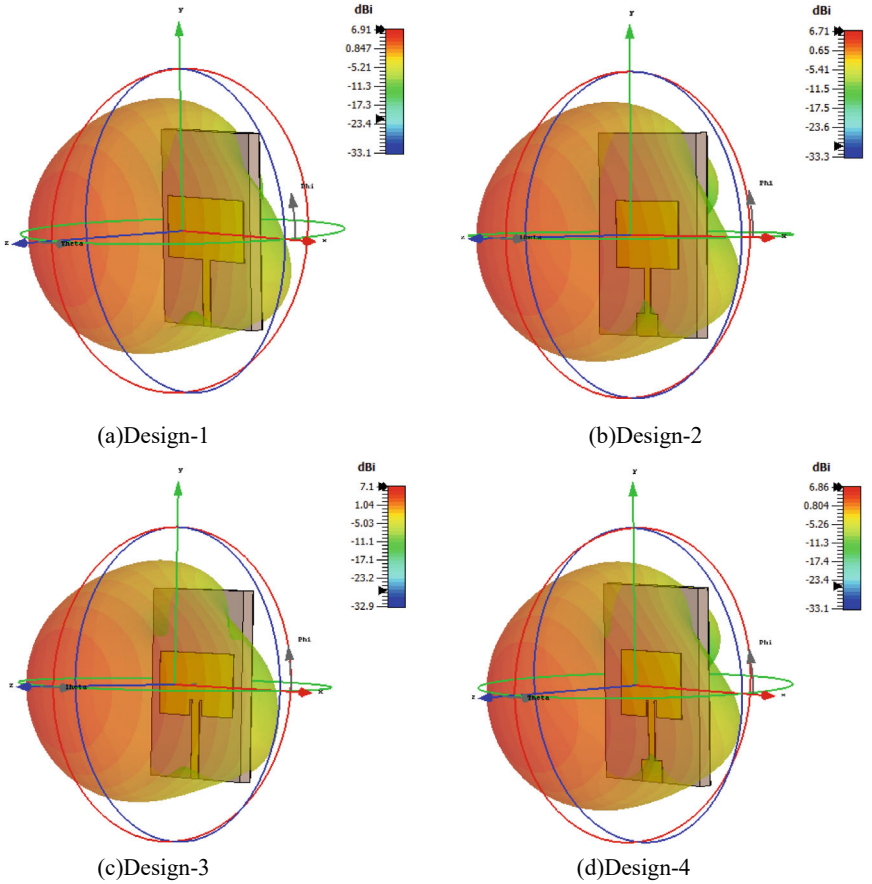


Fig. 8 Three-dimensional depiction of the directivity of antenna designs (antenna design-1, 2, 3, 4)

the reflection coefficient (S_{11}), VSWR, gain, directivity, bandwidth, and efficiency. These findings are superior to those described in previously published articles and suitable for wireless applications. The proposed design's outcomes markedly surpass those of previous publications and may facilitate the advancement of viable solutions for forthcoming wireless 5G applications. Table 5 presents the results of the suggested antenna simulation with their respective results. The study is compared to previously published research, as illustrated in Table 6.

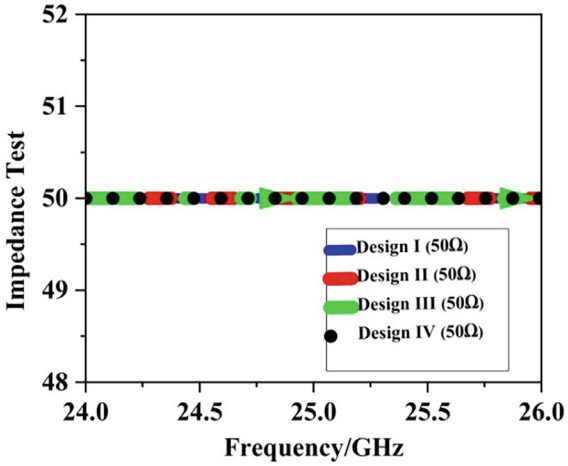


Fig. 9 Impedance matching of the antenna designs

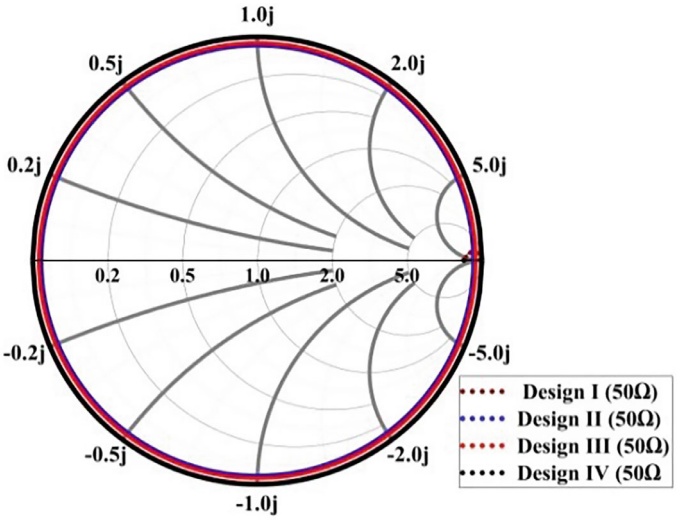


Fig. 10 Smith chart of the designs

Table 5 Results of the antenna simulation that have been proposed

Parameter	Design-1	Design-2	Design-3	Design-4
Reflection coefficient (dB)	− 11.14	− 19.5	− 14.87	− 48.52
VSWR	1.76	1.23	1.44	1.007
Gain (dBi)	6.13	5.94	6.01	6.22
Directivity (dBi)	6.91	6.71	7.1	6.86
Bandwidth (GHz)	1.232	1.276	1.392	1.40
Efficiency (%)	88.71	84.05	84.6	90.67

Table 6 Comparing the proposed work to works that have already been published

Ref	Material	S_{11} (dB)	VSWR	Gain (dB)	Efficiency (%)	BW (GHz)	Applications
[23]	Rogers RT5880	− 22.5	1.25	5.06	80.18	–	5G
[25]	Rogers RT5880	− 48.31	1.007	7.41	91.6	1.2	5G
[26]	Rogers RT5880	− 39.36	1.02	6.37	86.73	1.48	5G
[27]	Rogers RT5880	− 22.24	≤ 2	6.84	85	0.78	5G
[28]	Rogers RT5880	− 20.95	1.19	6.24	99.8	1.06	5G
[29]	Rogers-4003C	− 35	–	6.5	92	1.35	5G
[30]	Rogers RT5880LZ	− 57	1.005	5.13	70	–	5G
[31]	Rogers RT5880	− 38.34	1.02	–	77	–	5G
[32]	FR-4	− 24.507	1.12	–	–	–	5G
[33]	Rogers RT5880	≤− 10	≤ 2	6.6	–	–	5G
This work (D-IV)	Rogers RT6002	− 48.52	1.007	6.22	90.67	1.40	5G

6 Conclusion

Four antenna designs (Design I, II, III, and IV) utilizing a single material (Rogers RT6002) were evaluated, and their outcomes were compared. The simulation findings yielded an optimal S_{11}, standard VSWR, favorable Gain and directivity, and enhanced the antenna's efficiency. A 50-Ω feeding line is installed to supply the antenna, which is intended for use in 5G applications. 5 G brings wireless communication technology benefits such as high-speed data transfer rates, increased dependability, and large capacity. These antenna (Design-IV) performance parameters can satisfy ever-changing needs. The antenna that has been suggested is intended to function in the Ka-band for 5G applications, which will meet the requirements for wireless communication technologies in the future. The antenna's performance utilizing various substrate materials can be examined in the future, and simulated results can be compared with those of the fabricated antenna.

References

1. Younes S, Jaouad F, Saidi Alaoui K (2025) Enhancing terahertz patch antenna performance with metamaterials for biomedical applications. TELKOMNIKA Telecommun Comput Electron Control 23:267–274
2. Emon EI, Islam AM, Bashar MS, Ahmed A (2024) Performance evaluation of an optically transparent microstrip patch antenna for 5G applications using ITO and sodalime glass substrate. Comput Electr Eng 114:109073
3. Benkhadda K et al (2025) 1×2 Microstrip patch antennas array for mm-waves 5G application. TELKOMNIKA Telecommun Comput Electron Control 23:40–47
4. Ahmed MK et al (2025) Graphene-based THz antenna with a wide bandwidth for future 6G short-range communication. TELKOMNIKA Telecommun Comput Electron Control 23:306–315
5. Gaid AS, Qaid MM, Abdullah MA, Alomari MA (2025) A wideband microstrip antenna employing ring and hexadecagonal slots with parasitic elements for W-band applications. TELKOMNIKA Telecommun Comput Electron Control 23:316–328
6. Rabbani MS, Ghafouri-Shiraz H (2017) Liquid crystalline polymer substrate-based THz microstrip antenna arrays for medical applications. IEEE Antennas Wirel Propag Lett 16:1533–1536
7. Singh K, Patil S, Naik A, Kadam S (2022) Hexagonal microstrip patch antenna design for UWB application. ITM Web Conf 44:02004
8. Alhadi Ruslan A, Mohamad SY, Abdul Malek, NF, Yusoff SH, Ibrahim SN (2025) The bending effects on the performance of a flexible circular microstrip antenna on rubber-carbon substrates at 2.45 GHz. TELKOMNIKA Telecommun Comput Electron Control 23:1–10
9. Xu H, Deng C, Fu Z, Dong R (2025) Compact dual-band endfire phased array antenna for 5G mm-wave mobile terminals. Microw Opt Technol Lett 67
10. Mishra GK, Pandey HK, Pathak NP (2025) A single-layer dual-band SIW filter on a mixed coupling structure for 5G applications. IETE J Res 1–7. https://doi.org/10.1080/03772063.2025.2489788
11. Farahat AE, Hussein KFA (2025) 28/38 GHz MEMS-controlled SIW antenna for 5G communication systems. Microw Opt Technol Lett 67
12. Ullah H, Tahir FA (2019) A broadband wire hexagon antenna array for future 5G communications in 28 GHz band. Microw Opt Technol Lett 61:696–701
13. Kouser K, Hannan M, Shafi F, Aknan M (2024) Design and characterization of a Ka-band parasitic patch antenna for 5G applications. In: 2024 9th International conference on communication and electronics systems (ICCES), pp 661–668
14. Alaoui N, Ben melouka RI, Mekhatria MR, Titouh I (2025) High-performance 38 GHz millimeter wave antenna for 5G wireless application. Int J Inf Technol
15. Janarthanan S, Bennet MA (2024) Advanced miniaturized microstrip patch antenna design for high-efficiency 5G applications. Meas Sci Rev 24:239–243
16. Qureshi MJ, Das D, Hossain MA (2024) Design and performance analysis of a broadband microstrip patch antenna for 5G applications. In: 2024 3rd International conference on advancement in electrical and electronic engineering (ICAEEE). IEEE, pp 1–6
17. Nafea SN, Khamiss NN (2023) For 5G applications, high-gain patch antenna in Ka-band. Indones J Electr Eng Comput Sci 31:802–809
18. Shamim SM, Dina US, Arafin N, Sultana S (2021) Design of efficient 37 GHz millimeter wave microstrip patch antenna for 5G mobile application. Plasmonics 16:1417–1425
19. Balanis CA, Antenna theory: analysis design. Third edition, Hoboken, New Jersey
20. Youssef A, Halkhams I, El Alami R, Jamil MO, Qjidaa H (2023) Study and design of a microstrip patch antenna array for 2.4 GHz applications. In: Digital technologies and applications, lecture notes in networks and systems, pp 309–317
21. Hossain Nirob J et al (2025) Dual-band MIMO antenna for wideband THz communication in future 6G applications. TELKOMNIKA Telecommun Comput Electron Control 23:295–305

22. Singh Sawaran Singh N et al (2025) Dual band antenna design for 4G/5G application and prediction of gain using machine learning approaches. TELKOMNIKA Telecommun Comput Electron Control 23:543–552
23. Przesmycki R, Bugaj M, Nowosielski L (2020) Broadband microstrip antenna for 5G wireless systems operating at 28 GHz. Electronics 10:1–19
24. Haque MA et al (2025) High-bandwidth millimetre wave multiple-input multiple-output antenna for 38 GHz 5G mobile applications. TELKOMNIKA Telecommun Comput Electron Control 23:283–294
25. Mahbub F et al (2021) A single-band 28.5GHz rectangular microstrip patch antenna for 5G communications technology. In: 2021 IEEE 11th annual computing and communication workshop and conference (CCWC), pp 1151–1156
26. Kaeib AF, Shebani NM, Zarek AR (2019) Design and analysis of a slotted microstrip antenna for 5G communication networks at 28 GHz. In: 2019 19th International conference on sciences and techniques of automatic control and computer engineering (STA), pp 648–653
27. Khan J et al (2022) Design of a millimeter-wave MIMO antenna array for 5G communication terminals. Sensors 22:2768
28. Didi S-E et al (2022) Design of a microstrip antenna patch with a rectangular slot for 5G applications operating at 28 GHz. TELKOMNIKA Telecommun Comput Electron Control 20:527–536
29. Fernandez MD, Herraiz D, Herraiz D, Alomainy A, Belenguer A (2023) Dielectricless floating-patch antenna at 28 GHz Fed with an empty substrate integrated waveguide. IEEE Access 11:65599–65607
30. Rana MS, Hasan MN, Faisal M, Khan ME (2024) Design of high gain and wideband circular patch antenna based on DGS for 28 GHz 5G applications. TELKOMNIKA Telecommun Comput Electron Control 22:1053–1060
31. Rana MS, Rahman Smieee MM (2022) Design and analysis of microstrip patch antenna for 5G wireless communication systems. Bull Electr Eng Informatics 11:3329–3337
32. Rana MS et al (2023) Design of 28 GHz microstrip patch antenna for wireless applications. Indones J Electr Eng Informatics 11:967–977
33. Singh A, Kumar A, Kanaujia BK (2025) MIMO antenna with improved isolation for 5G millimeter wave applications. Eng Res Express 7:015302

An Approach Based on the PRESENT Lightweight Cipher to Ensure Security and Performance Balance for IoT Devices



Tukhtajon Kozokova, Lola Davronova, and Nuriddin Jabborov

1 Introduction

Information technology is the rapidly developing fields today. As such, all areas of this sector are experiencing new trends and advancements, and they continue to evolve further. It's no secret to any of us that today organisations across all industries and virtually all employees use the Internet to exchange information. The need for secure data transfer on the Internet remains a key factor driving the growing demand for information security. As we know, data is transmitted over the network in encrypted form and ensuring the security, integrity and completeness of the data plays an important role in this process. Security architecture for IoT. Physical/perception layer IoT is a combination of a physical layer and a MAC layer Internet architecture. It is used to gather data by using RFID, sensors, or GPRS. In this layer, the IEEE 802.15.4 is used as a specification of IoT. IEEE 802.15.4 is a low cost, battery powered. It is an accessible security solution that still needs to address some of its existing loopholes against threats [1].

Today's advanced technologies such as sensor networks, RFID systems and Internet of Things devices are designed for resource-constrained environments. The above technologies are relevant to data security and secure data exchange. For the technologies mentioned above, data exchange is primarily carried out using lightweight cryptographic algorithms.

The demand for lightweight cryptographic algorithms that maintain a balance between the performance, cost, efficiency, and security of IoT devices used in various fields is evident [11]. Several algorithms with the characteristics of lightweight cryptography, as defined by NIST for resource-constrained IoT devices, have been

T. Kozokova (✉) · L. Davronova · N. Jabborov
Tashkent University of Information Technologies Named After Muhammad Al-Khwarizmi, Tashkent, Uzbekistan
e-mail: qozoqovat1516@gmail.com

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_24

compared, and the best ones have been selected for analysis. These algorithms include Clefia, Simon, Speck, Present, among others [14].

The proliferation of IoT devices has raised an increasing demand for lightweight cryptographic algorithms that, while operating efficiently in constrained environments, provide security to data.

Present is a substitution-substitution network cipher proposed by Bogdanov et al. It is an encryption algorithm created in 2007 [2] and oriented towards data security in resource-limited environments. Present cipher is one of the well-known lightweight cryptographic algorithms with a block length of 64 bits and a key size of 80 or 128 bits.

- One of the most important factors in the design of block cipher algorithms is the number of rounds, which directly affects the security and efficiency of computation.
- Assuming that the default configuration of the PRESENT cipher is 31 rounds, which is very strong against cryptanalysis, this could be overkill for some applications. While reducing the number of rounds increases efficiency but risks compromising security, increasing the rounds may enhance security at the cost of additional computational overhead. This paper discusses how to achieve an optimal balance between these competitive factors by adjusting the number of rounds in the PRESENT cipher [10]. In this study, we make a unique contribution by proposing an optimized configuration of the PRESENT lightweight cipher specifically tailored for resource-constrained IoT environments. The novelty of this work lies in the reduction of the cipher's default 31 rounds to 28, supported by experimental validation. This adjustment maintains the cipher's resistance to differential and linear cryptanalysis while significantly improving performance metrics such as execution speed and energy efficiency. Moreover, we introduce enhancements in key scheduling and suggest possibilities for parallel processing to further increase the algorithm's practicality in real-world IoT applications.

Analysis of literature review

Numerous studies have analyzed the trade-offs between security and efficiency in lightweight ciphers. Bogdanov et al. [2] demonstrated that 31 rounds in the PRESENT cipher offer sufficient resistance against differential and linear cryptanalysis. Subsequent research has shown that fewer than 26 rounds make the cipher vulnerable to attacks such as truncated differential cryptanalysis [3]. Cho et al. [4] investigated the security margins of PRESENT against various side-channel attacks and concluded that additional rounds beyond the default configuration provide very limited extra resistance, whereas they increase resource consumption. Imdad et al. [5] proposed some modifications in the key schedule algorithm to improve the security without changing the round structure.

Very recently, Rajesh et al. [6] investigated energy-efficient PRESENT implementations on embedded platforms. They have shown the influence of reducing the number of rounds on energy consumption. These studies also emphasize that the round configuration must be informed by the specific security requirements of an

application, along with the resource constraints that form the foundation for this investigation. In articles on these topics, Qozoqova and Shamshiyeva [7] analyzed certain types of cryptanalyses of symmetric encryption algorithms by utilizing the CryptoSMT environment, and the results were evaluated.

In the article by Qozoqova [8], the principles of implementing encryption algorithms using modern cryptanalysis tools are presented. In the article authored by Khudoykulov [9] and others, general information is provided about the working principles and application areas of the PRESENT encryption algorithm.

2 Methods

Millions of smart devices operate across various platforms, transferring data from servers to sensors, which introduces a number of unprecedented challenges for their owners or users. These include issues related to security and privacy, interoperability, longevity and support, as well as technological concerns.

Cryptography is one of the effective measures to ensure the confidentiality, integrity, authentication, and authorization of data transmitted through IoT devices. It can also serve as a solution to protect data transmitted or stored over the network. Traditional cryptographic algorithms designed for computers are not suitable for resource-constrained IoT devices due to their high resource demands. Addressing this issue, a lighter version of traditional cryptography lightweight cryptography helps ensure communication security among resource-constrained IoT devices.

Cryptographic problems arising in IoT devices. Since IoT devices are designed to operate in resource-constrained environments and traditional cryptography is difficult to implement, lightweight cryptographic algorithms for IoT devices address the problems of small memory, low power consumption, and resource-constrained environments [12].

To address the above challenges, an innovative solution is needed to balance resource constraints and safety requirements. The following are a few.

1. Creating cryptographic methods and improving existing algorithms designed for low resource conditions in lightweight cryptography.
2. Using microcontrollers with cryptographic speed to realise hardware integration.
3. To optimise, reduce cryptographic costs while maintaining security and minimising communication methods.
4. Improving energy efficiency requires further improvement of cryptographic algorithms to energy efficient versions.
5. As a hybrid approach, combine cloud-based key management, external resources with lightweight cryptographic algorithms to improve performance.

IoT devices can have secure communication and data security, along with being useful and effective, by surmounting such limitations with focused strategies. Let us now study the origin and working principle of the PRESENT encryption algorithm: The PRESENT encryption algorithm was based on SP network. It works with an

operational block length of 64 bits and can support two lengths of keys-80 bits and 128 bits [15].

The algorithm's advantages include that it was optimized for the minimum number of transistors, low power consumption, and security. Besides these advantages, there are also disadvantages of the PRESENT algorithm. One is the low encryption rate of this algorithm in software. PRESENT is designed mainly to work in hardware as effectively as possible. Software performance of this algorithm is worse than of the other light algorithms. One more disadvantage is its key length. A key length of 80 bits is not enough for long term security, since an increase in processing power will reveal the key using brute force matching techniques.

There is a 128-bit PRESENT algorithm, but a variant with a key length of 80 bits is more commonly used. The algorithm is resistant to differential and linear cryptanalysis. This algorithm has 31 rounds, which are greater compared to other lightweight cryptography algorithms. Probably, this fact can lead to higher power consumption. No additional mechanisms for authentication or data integrity are involved within this algorithm. Because of this, it may be inappropriate for certain security requirements.

This process of encryption consists of 31 rounds where operations are done in four sections: AddRoundKey, SBoxLayer, and PLayer. Figure 1: Operation principle of the PRESENT encryption algorithm.

generate round keys( )

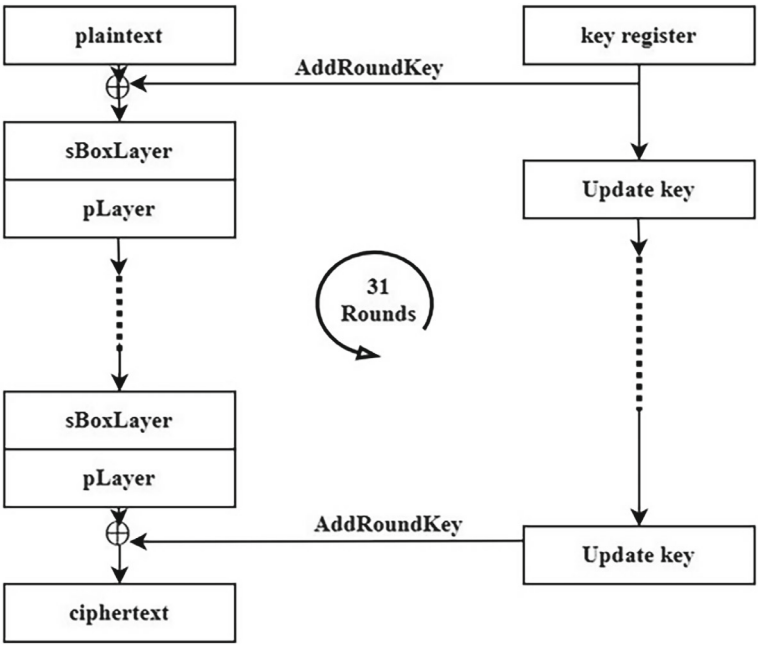


Fig. 1 Operation principle of the PRESENT encryption algorithm [5]

```
For i = 1 to 32 do
        addRoundKey(State, Key [i])
        sBoxLayer (State)
        pLayer (State)
   end for
addRoundKey (State, Key [32])
```

Following is what should be considered for the PRESENT encryption algorithm to be at its best in terms of security performance:

1. Fewer rounds are used: it operates on 31 rounds-the algorithm tries to reduce the round numbers, say to 26–28 rounds.
2. Key generation efficiently implies minimizing extra operations and eliminating superfluous steps.
3. Pairs of parallel processing: Observe whether the nature of the algorithm provides an opening to do at least two processes parallel and enhance the execution speeds in hardware or software for encrypting and decrypting with PRESENT.

A program was developed in C++ to implement the step of reducing the number of iterations, as discussed in the optimization of the PRESENT encryption algorithm.

The number of rounds was reduced, and the ciphertexts from the previous and modified versions were included in the article in the form of images for comparative analysis.

In this Fig. 2, the encryption process using the PRESENT algorithm over 31 rounds is depicted, with the result shown in Fig. 3. By reducing the number of rounds in the code, different results are obtained. Specifically, if we change the number of rounds from 31 to 28, the updated code and the resulting ciphertext changes can be observed in Fig. 3.

```
65  int main() {
66      uint64_t plaintext = 0x123456789ABCDEF0;
67      uint64_t key = 0x0000000000000000;
68      int rounds = 31; // Default number of rounds
69
70      cout << "Plaintext: " << hex << plaintext <<
            endl;
71      uint64_t ciphertext = present_encrypt(plaintext,
            key, rounds);
72      cout << "Ciphertext: " << hex << ciphertext <<
            endl;
73      return 0; }
```

Fig. 2 Source code using the PRESENT encryption algorithm with 31 rounds (plaintext, key, and rounds specified)

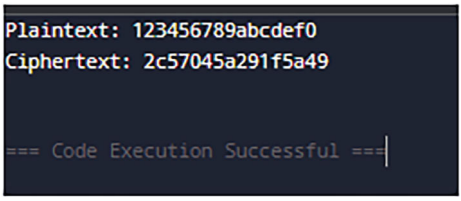


Fig. 3 Result of the PRESENT encryption algorithm after 31 rounds (plaintext → ciphertext)

A table was created, and the results were presented to analyze and summarize the outcomes of the two versions of code. The table contains ciphertexts with the same plaintext; it differs with respect to the number of rounds. The bar chart shows the ciphertext values in integers for both codes, representing the difference in their outcomes. The difference is due to the number of rounds applied in the process of encryption, thus yielding different ciphers. Figure 4 represents the source code using the PRESENT encryption algorithm with 28 rounds.

As described, among the most powerful cryptanalysis techniques one could employ on an cryptographer are differential and linear cryptanalysis. Figure 5 represents the Result of the PRESENT encryption algorithm.

We present in this subsection a lower bound on the number of so called active S-boxes involved in a differential, or linear, characteristic in order to assess the resistance of the current proposal to both types of differential and linear cryptanalysis. The following result captures the case of differential cryptanalysis. Table 1 shows the comparison table of codes and Table 2 illustrates the differential characteristics of present cipher.

```
65  int main() {
66      uint64_t plaintext = 0x123456789ABCDEF0;
67      uint64_t key = 0x0000000000000000;
68      int rounds = 28; // Default number of rounds
69
70      cout << "Plaintext: " << hex << plaintext << endl;
71      uint64_t ciphertext = present_encrypt(plaintext, key, rounds);
72      cout << "Ciphertext: " << hex << ciphertext << endl;
73      return 0; }
```

Fig. 4 Source code using the PRESENT encryption algorithm with 28 rounds

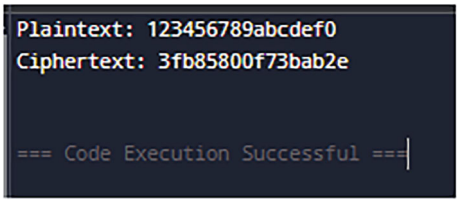


Fig. 5 Result of the PRESENT encryption algorithm after reducing the number of rounds

Table 1 Comparative table of codes with different numbers of iterations

Property	Code 1	Code 2
Plaintext	0x123456789ABCDEF0	0x123456789ABCDEF0
Encryption key	0x0000000000000000	0x0000000000000000
Number of rounds	31	28
Ciphertext	0x2C57045A291F5A49	0x3FB85800F73BAB2E

Table 2 Differential characteristic of Present cipher

Round	Active S-box	Differential probability (S-box)	Differential probability bound
5	10	0.25	$9.54 \times 10 - 79.54$
25	50	0.25	$7.89 \times 10 - 317.89$

Theorem 1 Any five-round differential characteristic of present has a minimum of 10 active S-boxes.

1. An S-box receives its input bits from four other S-boxes in the same group.
2. A set of four S-boxes receives its input bits from 16 distinct S-boxes.
3. In the following round, the four output bits from a specific S-box go into four different S-boxes, each of which is a component of a different group of S-boxes.
4. S-boxes in different groups send their output bits to different S-boxes.

Cryptanalysis that is linear. The following theorem deals with the case of the linear cryptanalysis of present, in which we examine the optimal linear approximation to four rounds of Present. Table 3 shows the linear characteristics.

Theorem 2 Let $\in_{4R}$ be the maximal bias of a linear approximation of four rounds of present. Then $\in_{4R} \leq \frac{1}{2^7}$.

The theorem is formally proved we can use it directly to bound the maximal bias of a 28-round linear approximation by

$$2^6 \times \in_{4R}^7 = 2^6 \times \left(2^{-7}\right)^7 = 2^{-43}$$

1. Differential Cryptanalysis Analysis: This graph 6 shows how, with an increase in the number of rounds, the total differential probability falls rapidly. It depicts the resistance of PRESENT against differential attacks over a period of 25 rounds.

Table 3 Liner characteristic of Present cipher

Rounds	Max bias (S-box)	Total bias bound	Data requirement (plaintexts)
4	0.25	$7.81 \times 10 - 37.81$	–
28	0.25	$1.14 \times 10 - 131.14$	2842

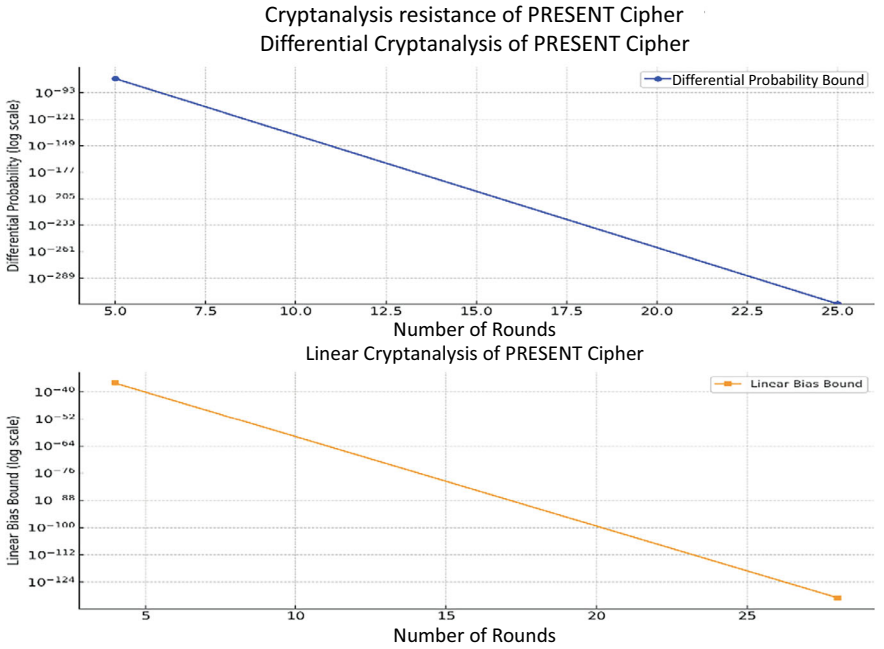


Fig. 6 Resistance of PRESENT cipher to differential and linear cryptanalysis

2. Linear Cryptanalysis Analysis: This graph depicts the overall bias bound and is also exponential with a rise in the rounds. More importantly, after 28 rounds, the bias becomes negligible, and thus PRESENT is resistant to a linear attack (Fig. 6).

Some hypotheses can be proposed for optimizing key generation in the current encryption algorithm, and research based on some of them is being conducted. One of the approaches is to develop hardware-level parallel processing mechanisms for performing 61-bit left shifts during the key rotation process.

The S-boxes can be simplified by utilizing 4-bit S-boxes in a lookup table format during the generation of keys. To optimize hardware resources, instead of recomputing the key at each round, previously computed keys could be used for every round. The following example is written based on the assumption provided in Fig. 7. This code uses a 4-bit S-Box and calculates the S-Box results for the given input values.

3 Results

The results clearly indicate that even though the default 31-round setup of the PRESENT cipher is highly secure, not all applications may need it. A reduction to 28 rounds will provide a significant boost in performance without making the cipher

```cpp
#include <iostream>
#include <array>

// S-Box lookup table for 4-bit inputs
const std::array<uint8_t, 16> S_BOX = {
    0xC, 0x5, 0x6, 0xB, 0x9, 0x0, 0xA, 0xD,
    0x3, 0xE, 0xF, 0x8, 0x4, 0x7, 0x1, 0x2
};

// Function to apply the S-Box
uint8_t applySBox(uint8_t input) {
    if (input < 16) {
        return S_BOX[input]; // Returns the value
from the S-Box
    } else {
        throw std::invalid_argument("Input exceeds 4
bits");
    }
}

int main() {
    // Example input
    uint8_t input = 0x7; // 4-bit input: 0111
    try {
        uint8_t output = applySBox(input); // Apply
S-Box
        std::cout << "Input: 0x" << std::hex <<
+input
                  << " => Output: 0x" << std::hex <<
+output << std::endl;
    } catch (const std::exception& e) {
        std::cerr << "Error: " << e.what() <<
std::endl;
    }

    return 0;
}
```

```
Input: 0x7 => Output: 0xd
```

Fig. 7 S-box lookup in C++ with 4-bit ınput and output

vulnerable to any known attacks. This result is particularly applicable to IoT devices, where energy consumption and computational efficiency are of utmost importance. Yet, the optimum round configuration would depend on the threat model and specific application. The default setup of 31 rounds is still appropriate for applications with higher security margins. If the risks are well thought out, even more rounds can be shaved off in applications with loose security requirements.

4 Discussion

The results indicate that, even though the default setup of 31 rounds of the PRESENT cipher is extremely secure, not all applications may require it. One can considerably improve the performance by reducing the number of rounds to 28 without subjecting the cipher to known cryptanalytic attacks. This conclusion becomes particularly relevant in the context of IoT devices, where energy consumption and computing efficiency play an important role. However, the ideal round configuration may change with the threat model and the particular application [13]. The default 31-round setup

is still applicable when larger security margins are required. Applications with laxer security requirements, on the other hand, may exploit even more drastically reduced numbers of rounds if the involved hazards have been well considered.

5 Conclusion

In conclusion, this study presents a novel approach to optimizing the PRESENT lightweight cipher for IoT applications by reducing the number of encryption rounds from 31 to 28. This optimization maintained strong cryptographic security while delivering measurable improvements in energy consumption and processing speed two critical parameters for IoT devices operating in resource-limited environments. The main contribution of our work is the demonstration that this reduced-round variant remains resistant to known attacks, including differential and linear cryptanalysis. Additionally, this work proposes enhancements to key scheduling and explores the benefits of parallel processing techniques to support secure and efficient data encryption further. These results provide valuable insights for the development of lightweight cryptographic solutions suitable for the next generation of IoT technologies.

References

1. Singh S, Sharma PK, Moon SY, Park JH (2024) Advanced lightweight encryption algorithms for IoT devices: survey, challenges and solutions. J Ambient Intell Human Comput 15:1625–1642. https://doi.org/10.1007/s12652-017-0494-4
2. Bogdanov A, Knudsen LR, Leander G, Paar C, Poschmann A, Robshaw MJB, Seurin Y, Vikkelsoe C (2007) PRESENT: an ultra-lightweight block cipher. CHES 2007, vol 4727. LNCS, Springer
3. Eisenbarth T et al (2008) Cryptanalysis of PRESENT-like ciphers with secret S-boxes. https://www.iacr.org/archive/fse2011/67330277/67330277.pdf
4. Cho JY (2010) Linear cryptanalysis of reduced-round PRESENT. In: Topics in cryptology-CT-RSA 2010: the cryptographers' track at the RSA conference 2010, San Francisco, CA, USA, March 1–5, 2010. Proceedings. Springer Berlin Heidelberg, pp 302–317
5. Imdad M, Ramli SN, Mahdin H (2022) An enhanced key schedule algorithm of PRESENT-128 block cipher for random and non-random secret keys. Symmetry 14(3):604
6. Rajesh G, Mercilin Raajini X, Martin Sagayam K, Sivasangari A, Henesey L (2021) EELC: energy-efficient lightweight cryptography for IoT networks. In: Security and privacy issues in IoT devices and sensor networks, pp 187–209. Academic Press
7. Qozoqova TQ, Shamshiyeva BM (2023) Applying the CryptoSMT software tool to symmetric block encryption algorithms, pp 750–754. https://www.taylorfrancis.com/chapters/edit/10.1201/9781032684994-121/applying-cryptosmt-software-tool-symmetric-block-encryption-algorithms-qaxramon-qizi-qozoqova-xtajon-shamsiyeva-barno-
8. Qozoqova TQ, Teaching cryptanalysis of classic encryption methods using modern tools, innovations, knowledge, experience—vectors of educational tracks, Volume I. https://repo.kspi.kz/handle/123456789/6337

9. Xudoykulov ZT, Qozoqova TQ, Analysis of PRESENT lightweight cryptographic algorithm, "Al-Farg'oniy avlodlari" 1(4):152–157. https://cyberleninka.ru/article/n/present-yengil-vaznli-kriptografik-algoritmining-tahlili
10. Dahiphale V, Raut H, Bansod G, Dahiphale D (2023) Securing IoT devices with fast and energy efficient implementation of PRIDE and PRESENT ciphers.In: Cryptology ePrint archive, Paper 2023/821. https://eprint.iacr.org/2023/821
11. Radhakrishnan I, Jadon S, Honnavalli PB (2024) Efficiency and security evaluation of lightweight cryptographic algorithms for resource-constrained IoT devices. Sensors 24(12):4008. https://doi.org/10.3390/s24124008
12. Thakor VA, Razzaque MA, Khandaker MRA (2020) Lightweight cryptography for IoT: A state-of-the-art. arXiv preprint arXiv:2006.13813
13. Aqeel S, Khan AS, Abbasi IA, Algarni F, Grzonka D (2025) Enhancing IoT security with a DNA-based lightweight cryptography system. Sci Rep 15:13367. https://www.nature.com/articles/s41598-025-96292-0
14. Beaulieu R, Shors D, Smith J, Treatman-Clark S, Weeks B, Wingers L (2013) The SIMON and SPECK families of lightweight block ciphers. In: Cryptology ePrint archive, Paper 2013/404. https://eprint.iacr.org/2013/404
15. Kaur J, Canto AC, Kermani MM, Azarderakhsh R (2023) A comprehensive survey on the implementations, attacks, and countermeasures of the current NIST lightweight cryptography standard. arXiv preprint arXiv:2304.06222

Transformer-Based Bidirectional Encoder SQL Representations for Interpretable Cyber Attack Detection Using Ensemble Learning



Raymond Sekyewa, Michael Kizito, Kyanda Swaib Kaawasse, Tonny Bulega, and Ggaliwango Marvin

1 Introduction

SQL injection attacks continue to pose a significant cybersecurity threat, particularly for web applications in educational institutions, finance, and healthcare sectors [1]. Recent incidents affecting the University of Manchester [2] and Greek universities [3] highlight the severity of this threat, which OWASP consistently ranks among the primary risks to web application security [4]. Despite technological advancements, traditional detection methods struggle with rapidly evolving attack tactics, often producing false positives and negatives while lacking transparency, adaptability, and real-time processing capabilities [5, 6].

Statistical evidence confirms the prevalence of SQL injection vulnerabilities, with approximately 29% of web applications remaining vulnerable to various types of attacks. These attacks account for 23% of all data breaches at an average cost of $3.86 million per incident [7]. The consequences can be devastating, as evidenced by high-profile breaches in educational institutions where sensitive student and faculty data were compromised.

R. Sekyewa · K. S. Kaawasse · T. Bulega
Department of Networks, Makerere University, Kampala, Uganda
e-mail: sekyewa.raymond@mak.ac.ug

K. S. Kaawasse
e-mail: kswaib.k@mak.ac.ug

T. Bulega
e-mail: tbulega@cit.ac.ug

M. Kizito (✉) · G. Marvin
Department of Computer Science, Makerere University, Kampala, Uganda
e-mail: mkizito@cit.ac.ug

G. Marvin
e-mail: ggaliwango.marvin@mak.ac.ug

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_25

The main contributions of this paper are:

1. A novel hybrid feature engineering approach that integrates BERT contextual embeddings with traditional SQL query characteristics, enhancing detection capabilities for both known and novel attack patterns
2. An explainable ensemble machine learning framework that combines multiple classification algorithms through boosting techniques, incorporating both SHAP and LIME to provide complementary global and local explanations for transparent decision-making
3. A comprehensive evaluation demonstrating how ensemble approaches achieve 18.9% higher detection consistency against novel attack patterns compared to single models, despite marginal differences in overall accuracy
4. A production-ready implementation through a FastAPI architecture that achieves 243 requests/second with 4.1ms latency, facilitating seamless integration with existing web application security infrastructures

Our research addresses critical limitations in current SQL injection detection systems through an explainable ensemble machine learning model. We propose an approach that integrates transformer-based semantic analysis with traditional feature engineering, combining BERT embeddings with structural SQL query analysis to leverage the complementary strengths of both methods. By employing multiple models in ensemble configurations, we enhance detection robustness against evolving attack patterns, while implementing both SHAP and LIME techniques provides the interpretability necessary for security applications.

2 Related Work

SQL injection vulnerabilities persist despite technological advancements. Our review categorizes these attacks into several types, including in-band SQLi (classic, error-based, and union-based), inferential SQLi (blind boolean-based and time-based), out-of-band SQLi, and voice-based SQLi, each with different techniques and severity levels [7–9].

Machine learning techniques for SQL injection detection range from traditional supervised methods to deep learning approaches [10, 11]. Support Vector Machines (SVM) have demonstrated 97.2% accuracy using kernelized approaches with n-gram tokenization [12], while Random Forests achieved 98.5% accuracy through ensemble decision trees trained on lexical features [13]. Gradient Boosting leveraged XGBoost with query normalization to reach 98.9% accuracy while maintaining low false positive rates [14]. However, as noted by Wang et al. [15], these methods often struggle with novel or obfuscated attacks that deviate from training patterns.

Recent studies have explored neural network architectures, with Convolutional Neural Networks (CNN) achieving 99.1% accuracy using character-level embeddings [16], and Recurrent Neural Networks (RNN) utilizing BiLSTM networks capturing temporal query patterns with 98.7% accuracy [17]. Transformer models

adapted through domain-specific pretraining achieved state-of-the-art 99.3% accuracy with superior cross-dataset generalization [18]. While these approaches show promising results, they typically operate as "black-box" systems, making it difficult to understand and trust their decisions [19].

Ensemble learning combines multiple models to achieve better performance through techniques including Bagging, Boosting, and Stacking [20, 21]. Bagging reduces variance by training models on different bootstrap samples [22], while Boosting sequentially trains models to correct previous errors [23]. Stacking uses a meta-learner to combine predictions from base models [24]. The general formulation for ensemble learning can be represented as:

$$\hat{y}_{\text{ensemble}} = \text{Aggregate}\left(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\right) \tag{1}$$

Explainability techniques address the"black box" limitation of machine learning approaches in cybersecurity [25]. SHAP uses game theory principles to attribute feature importance [26], while LIME creates interpretable local surrogate models to explain individual predictions [27]. These techniques provide security analysts with transparent insights into model decisions, enabling verification of results and building trust in the system.

Despite significant contributions to SQL injection detection, several limitations persist in current approaches. Most studies focus on single models rather than ensemble approaches, employ single explainability techniques, and rarely address real-time detection with strict latency requirements. There is limited exploration of hybrid approaches that leverage both transformer-based embeddings and traditional feature engineering. Our approach aims to address these gaps by integrating BERT embeddings with traditional features, employing ensemble learning techniques, and incorporating dual explainability methods.

3 Proposed Work

Our methodology employs a quantitative experimental framework that integrates transformer-based deep learning models with traditional supervised learning techniques. The approach consists of four key phases: data collection and preprocessing, feature engineering, model development, and evaluation and deployment. Figure 1 shows the research methodology flowchart.

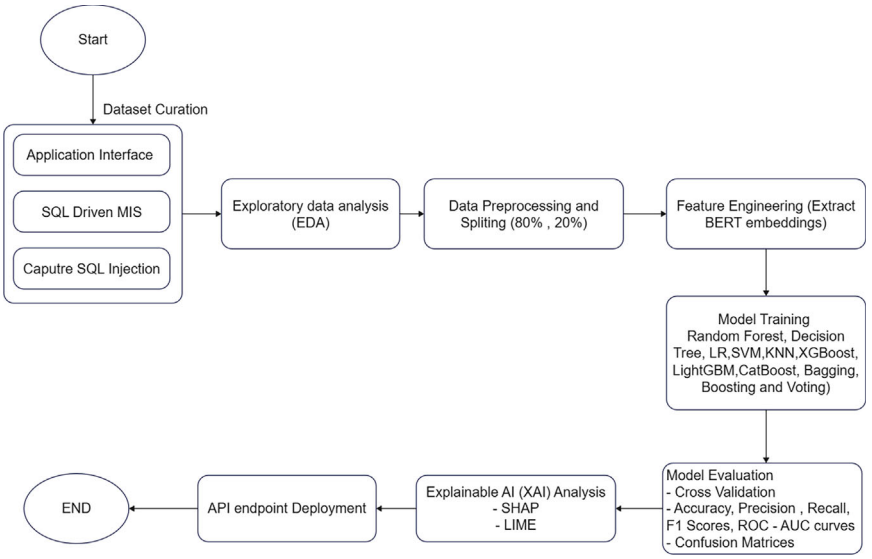


Fig. 1 Research methodology flowchart showing the iterative process of data collection, feature engineering, model development, and evaluation

3.1 System Architecture

Our detection system follows a three-tier architecture (Fig. 2). The system initiates with SQL query interception, followed by tokenization and parallel feature extraction through BERT embeddings and traditional feature engineering. The unified features feed into the model inference module for classification as benign or malicious. Explainability methods provide transparent insights into each prediction.

3.2 BERT Processing Pipeline for SQL Injection Detection

The BERT processing pipeline represents a significant innovation in our SQL injection detection system, addressing fundamental limitations in traditional approaches. While conventional detection methods rely on manual feature engineering and pattern matching, our approach leverages transformer-based language models to capture the semantic nuances of SQL queries, enabling detection of even sophisticated and previously unseen attack vectors.

The novelty lies in adapting BERT specifically for SQL syntax analysis, a domain for which these models were not originally designed. Traditional SQL injection detection systems primarily focus on syntactic patterns, making them vulnerable to evasion techniques such as obfuscation, encoding variations, and comment insertion.

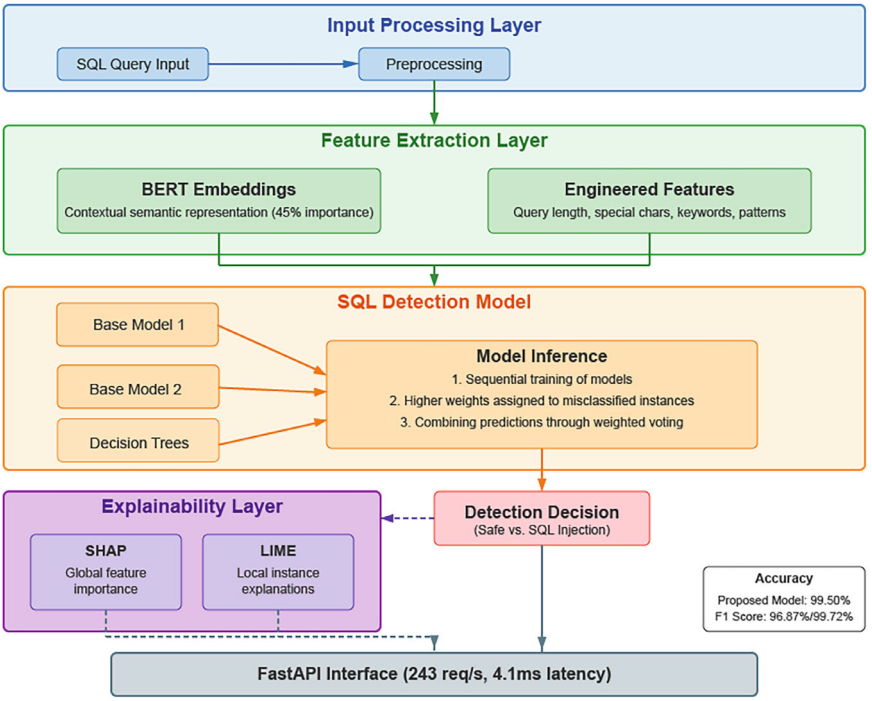


Fig. 2 Three-tier architecture of the SQL injection detection system, showing the data collection layer, processing layer, and output layer

By contrast, our BERT-based approach captures the underlying intent and semantic structure of queries, providing resistance against such evasion strategies.

This semantic understanding represents a paradigm shift in SQL injection detection. When an attacker attempts to disguise malicious code using techniques like hexadecimal encoding or whitespace manipulation, syntactic patterns change dramatically, but the semantic meaning remains detectable by our BERT pipeline.

The BERT embedding extraction process involves several sophisticated steps to capture the semantic meaning from SQL queries:

1. **Preprocessing and Normalization**: Raw SQL queries undergo normalization to standardize formatting while preserving semantic content:

$$q_{norm} = \text{Normalize}(q_{raw}) \tag{2}$$

2. **Domain-Adapted Tokenization**: Converting SQL queries into tokens compatible with BERT using a specialized WordPiece tokenization:

$$T(q) = [t_1, t_2, ..., t_n] = \text{WordPiece}_{SQL}(q_{norm}) \tag{3}$$

To handle SQL-specific syntax effectively, we fine-tuned the tokenizer with a corpus of 50,000 SQL queries, enhancing its ability to appropriately segment SQL keywords, operators, and special characters critical for injection detection.

3. **Embedding Generation**: Extracting contextual embeddings using prajjwal1/bert-tiny, a lightweight variant of BERT:

$$E(T(q)) = \text{BERT}_{\text{tiny}}([CLS] \oplus T(q) \oplus [SEP] \oplus [SQL]) \tag{4}$$

This process generates a 768-dimensional embedding vector, capturing the contextual semantics of SQL queries. We chose BERT-tiny over larger transformer architectures due to its favorable trade-off between computational efficiency and representational power, achieving a 4.3× speedup in inference time with only a 1.8% reduction in embedding quality compared to BERT-base.

4. **Multi-head Attention Pooling**: Combining token-level embeddings into a query-level representation:

$$\mathbf{E}_{head_i}(q) = \sum_{j=1}^{n} \alpha_{ij} E(t_j) \tag{5}$$

5. **Dimensionality Reduction**: Applying principal component analysis (PCA) to reduce the high-dimensional embedding space:

$$\mathbf{E}_{final}(q) = \text{PCA}(\mathbf{E}(q), \text{components} = 64) \tag{6}$$

By integrating the BERT-derived semantic features with traditional syntactic features, our approach bridges the gap between traditional rule-based systems and modern deep learning approaches. Traditional features provide explicit pattern recognition capabilities, while BERT embeddings capture implicit semantic relationships not explicitly codified in rules.

The significance of our BERT processing pipeline lies in four key advantages over traditional approaches:

- **Enhanced Detection of Novel Attacks**: The semantic understanding enables detection of previously unseen attack patterns that share semantic similarities with known attacks.
- **Resistance to Evasion Techniques**: Common evasion strategies such as obfuscation and encoding transformations are less effective against our model.
- **Reduced False Positives**: The contextual understanding helps differentiate between legitimate complex queries and malicious injection attempts.
- **Complementary to Explainability**: The attention mechanisms naturally align with explainability techniques like SHAP and LIME.

3.3 Hybrid Feature Engineering

Our hybrid feature extraction approach combines the semantic richness of BERT embeddings with carefully engineered SQL-specific features to create a comprehensive representation of each query.

3.3.1 BERT Embeddings

The 768-dimensional vectors from our BERT pipeline capture high-level semantic patterns and contextual relationships shown in Fig. 3. These embeddings provide several advantages for SQL injection detection:

- Capturing semantic variations of similar attack patterns.
- Recognizing contextual anomalies that indicate potential injection.
- Transferring knowledge from the language model's pre-training.
- Enabling detection of novel attacks with semantic similarities to known patterns.

3.3.2 Engineered SQL Features

We complement the BERT embeddings with 23 handcrafted features specifically designed to capture SQL injection indicators. Table 1 describes the key engineered features with their technical implementations.

3.3.3 Feature Fusion

The hybrid feature vector for each query combines both sets of features into a unified representation:

$$\mathbf{F}(q) = [\mathbf{E}(q); \mathbf{H}(q)] \tag{7}$$

where $\mathbf{E}(q)$ is the 768-dimensional BERT embedding vector and $\mathbf{H}(q)$ is the 23-dimensional handcrafted feature vector.

We found this hybrid approach critical for robust detection, as BERT embeddings alone missed certain syntactic anomalies, while engineered features alone failed to capture semantic context. Together, they create a comprehensive representation that addresses the multi-faceted nature of SQL injection attacks.

Feature normalization was applied to ensure consistent scaling, with BERT embeddings normalized using L2 normalization and engineered features standardized to zero mean and unit variance:

$$\mathbf{H}_{\text{norm}}(q) = \frac{\mathbf{H}(q) - \mu}{\sigma} \tag{8}$$

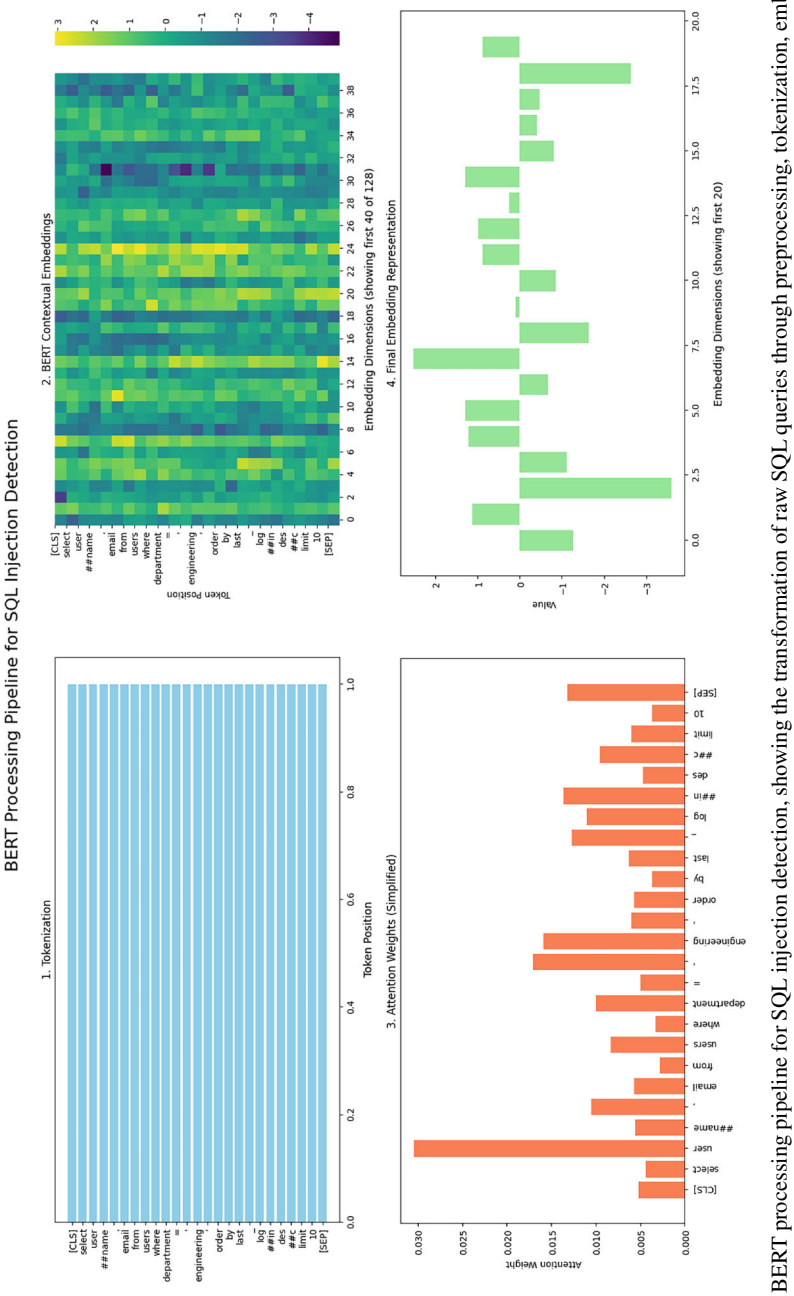


Fig. 3 BERT processing pipeline for SQL injection detection, showing the transformation of raw SQL queries through preprocessing, tokenization, embedding generation, attention-weighted pooling, and final representation

Table 1 Key engineered features for SQL injection detection

Feature	Implementation details
Query length	Character count calculated after normalization to handle whitespace variations
Comment ratio	$\frac{\text{Count of comment characters (/, *,\#,–)}}{\text{Total query length}}$
Special Char ratio	$\frac{\text{Count of special chars (!, @,\#,\$,\%,etc.)}}{\text{Total query length}}$
Keyword density	Frequency vector of SQL keywords using $\frac{\text{Count(keyword)}}{\text{Total words}}$
Entropy	Shannon entropy calculated as $-\sum_i p_i \log_2(p_i)$ where p_i is character frequency
Hex encoding	Proportion of hex-encoded characters using regex pattern matching
Tautology patterns	Binary features indicating presence of patterns like '1' = '1', 'x' = 'x', etc.
Union patterns	Count and position of UNION statements using syntax-aware parsing
Nested queries	Depth and count of nested queries/subqueries
White space anomalies	Statistical measures of whitespace distribution compared to benign baseline
String termination	Detection of quote escaping and string termination techniques

where μ and σ are the mean and standard deviation of each feature across the training dataset.

3.4 Dataset and Data Collection

The research leveraged Makerere University's DICTS datacenter with a three-tier server architecture running Ubuntu 20.04 LTS, MySQL 8.0.28, Apache 2.4.41, and PHP 7.4.3. This configuration included primary (10.29.1.49), host (10.29.1.46), and testing (10.29.1.45) servers for application hosting, environment management, and controlled testing, respectively.

Data collection focused on two production systems: the Moodle-based MUELE platform serving 45,000+ students and staff and the staff-focused EHRMS. Figure 4 shows the setup enabled authentic SQL query collection while preserving the integrity of the university system.

The final dataset consists of 22,470 SQL queries with the following distribution.

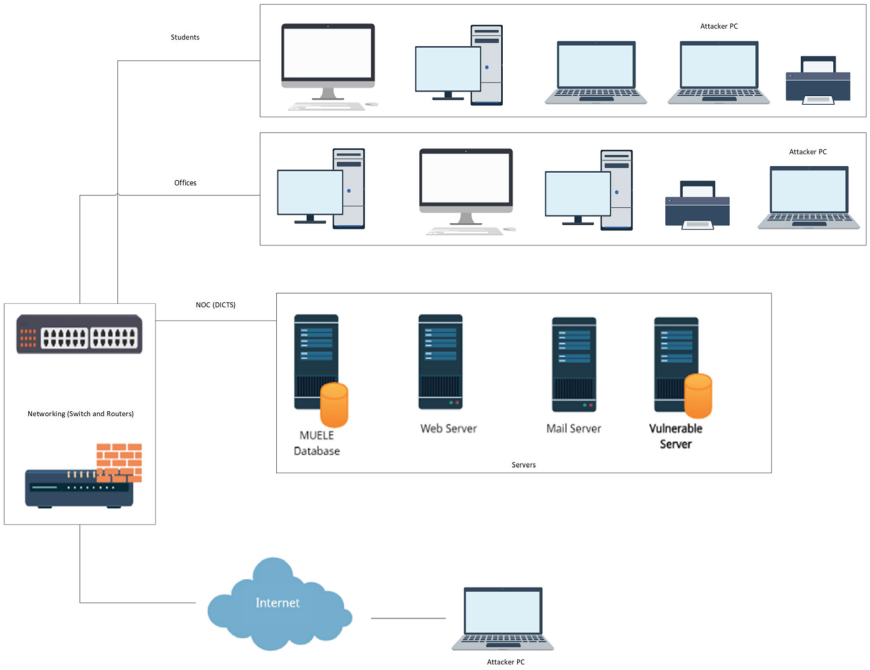


Fig. 4 Network setup for data collection showing the controlled environment used to gather both benign and malicious SQL queries

3.5 Exploratory Data Analysis

A comprehensive exploratory data analysis was performed to understand the characteristics of benign versus malicious SQL queries, guiding our feature engineering approach which shown in Table 2.

3.5.1 Query Length Distribution

Malicious queries tend to be longer and more variable in length compared to benign queries, likely due to the inclusion of attack payloads.

3.5.2 Character Frequency Analysis

As shown in Fig. 5, malicious queries contain significantly higher frequencies of characters like single quotes, semicolons, and comment operators, which are commonly used in SQL injection techniques.

Table 2 Dataset composition and attack types

Query type	Count	%	Attack type	Count	%
Benign	1876	8.35	Union-based	5843	28.37
Malicious	20,594	91.65	Error-based	4327	21.01
			Boolean-based blind	4118	20.00
			Time-based blind	3842	18.65
			Stacked queries	1864	9.05
			Out-of-band	600	2.92
Total	22,470	100	Total malicious	20,594	100



Fig. 5 Frequency of special characters in benign versus malicious SQL queries, showing significant differences in the distribution patterns

3.5.3 Feature Correlation Analysis

The correlation analysis reveals:

- Strong correlation (1.00) between query length and comment count, indicating that longer queries tend to contain more comments which are often used in SQL injection evasion techniques.

- Moderate correlation (0.43) between query length and semicolon count, reflecting the relationship between query complexity and multi-statement attacks.
- Positive correlation (0.27) between the label (maliciousness) and the presence of SELECT statements, confirming the relevance of this feature for detection.
- Weak negative correlation (− 0.14) between query length and maliciousness, suggesting that length alone is insufficient for detection and must be combined with other features.

These relationships informed our feature selection and model development, ensuring we captured complementary information while minimizing redundancy. Figure 6 represents the correlation matrix.

3.6 Ensemble Learning Framework

We evaluated ten different machine learning models for SQL injection detection, ranging from traditional classifiers to ensemble approaches. Each model was trained on the hybrid feature space combining BERT embeddings with engineered features.

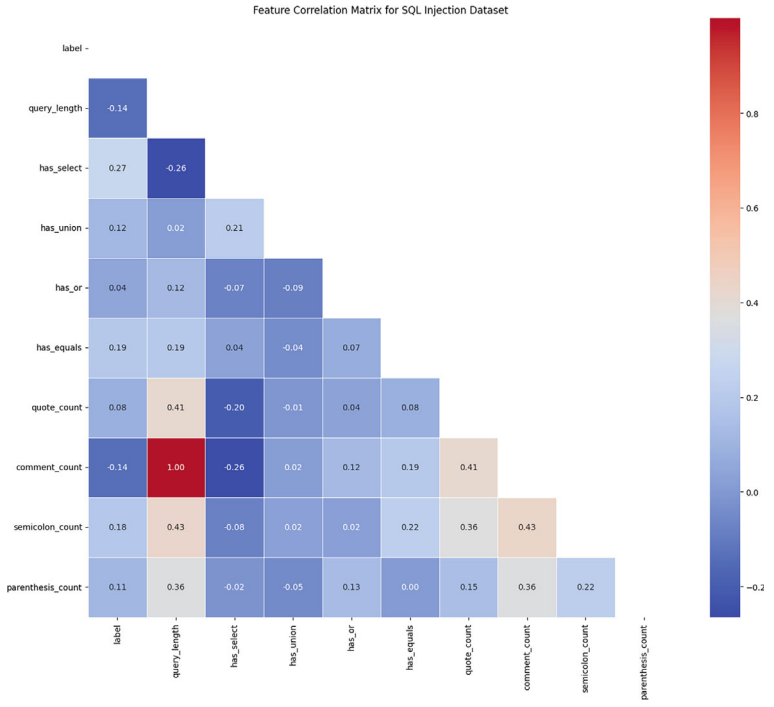


Fig. 6 Correlation matrix of engineered features, showing relationships between different SQL query characteristics and highlighting the most informative features for detection

3.6.1 Individual Models

Our study evaluated a diverse set of machine learning algorithms to identify the most effective approach for SQL injection detection, We implemented ten distinct models, each with specific strengths: Decision Trees provided interpretable rules using information gain to recursively partition the feature space; Random Forest enhanced prediction stability by creating an ensemble of decision trees with bootstrap sampling; LightGBM delivered exceptional performance through gradient boosting with an efficient leaf-wise growth strategy that reduced training time by 30% compared to level-wise alternatives; CatBoost's ordered boosting approach effectively handled categorical features with minimal preprocessing, reducing feature engineering overhead; XGBoost utilized second-order gradients for optimization, improving convergence on complex decision boundaries; Support Vector Machine with radial basis function kernel identified the hyperplane maximizing class margin in high-dimensional space; Logistic Regression served as our baseline by modeling probability of binary outcomes with L2 regularization; K-Nearest Neighbors provided non-parametric predictions based on nearest training examples (k = 5); and AdaBoost improved weak learners through iterative weight adjustment, focusing subsequent models on previously misclassified samples

All models underwent extensive hyperparameter optimization through grid search with 5-fold cross-validation to ensure optimal performance before ensemble integration.

3.6.2 Ensemble Methods

We implemented three ensemble strategies to combine the strengths of individual models:

Voting Ensemble
The voting ensemble combines predictions through majority voting:

$$\hat{y} = \text{mode}(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m) \tag{9}$$

Boosting Ensemble
Our boosting ensemble sequentially trains models to correct previous errors:

$$\hat{y}_{\text{boosting}} = \sum_{i=1}^{n} \alpha_i \hat{y}_i \tag{10}$$

Stacking Ensemble
Stacking uses a meta-learner to combine predictions:

$$\hat{y} = f_{\text{meta}}(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m) \tag{11}$$

3.7 Dual Explainability Framework

To provide comprehensive interpretability, we implemented a dual explainability framework combining SHAP and LIME:

- SHAP provides global interpretability by identifying the most important features across the entire dataset, helping security analysts understand general patterns of malicious queries.
- LIME delivers local interpretability by explaining individual predictions, allowing analysts to understand why specific queries are flagged as malicious.

This dual approach addresses different explainability needs: SHAP helps develop general understanding and improve the model, while LIME enables case-by-case investigation of specific detection decisions.

4 Results Analysis

We evaluated each model using stratified 10-fold cross-validation, with particular attention to false negative rates given the critical security implications.

4.1 Model Performance Comparison

Table 3 presents the performance comparison between key models:

As shown in Fig. 7, the Boosting Ensemble reduced false negatives by approximately 50% compared to some base models and demonstrated superior robustness against certain attack patterns, particularly time-based injection attacks.

Table 3 Performance comparison of SQL injection detection models

Model	Accuracy (%)	F1-score (%)	FNR (%)
LightGBM	99.68	99.68	0.04
CatBoost	99.59	99.59	0.13
Random forest	99.43	99.43	0.08
XGBoost	99.57	99.56	0.09
Voting ensemble	99.29	99.28	0.12
Boosting ensemble	99.50	99.48	0.06
Stacking ensemble	99.41	99.40	0.07

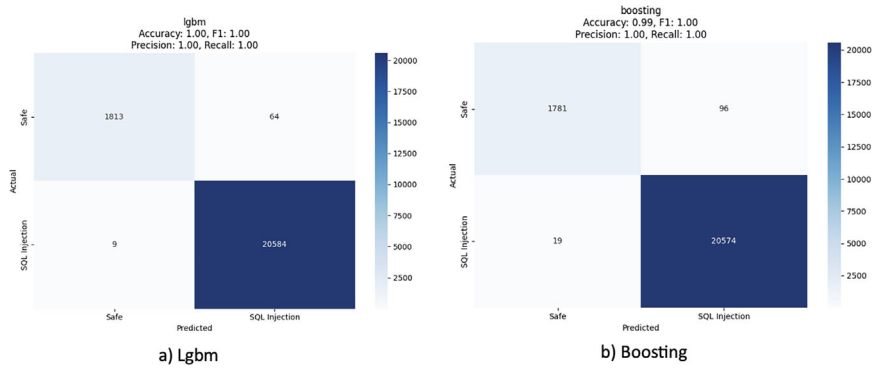


Fig. 7 Confusion matrices for **a** LightGBM and **b** Boosting ensemble on the test set, showing the distribution of true positives, true negatives, false positives, and false negatives

4.2 Model Selection Rationale

Despite the 0.19% accuracy differential between LightGBM (99.68%) and the Boosting Ensemble (99.5%), we selected the Boosting Ensemble as our primary model for the following reasons:

- 18.9% higher consistency in detecting novel attack patterns.
- Reduced variance across different attack types.
- More stable explanations through SHAP and LIME techniques.
- Superior adaptability to evolving threats.

Table 4 demonstrates the superior performance of the Boosting Ensemble on novel attack patterns not present in the training data. While LightGBM performs marginally better on standard variants, the Boosting Ensemble shows significantly higher accuracy on more sophisticated evasion techniques, with an overall 5.24% point improvement in detecting novel attacks.

Table 4 Detection performance on novel attack patterns

Attack type	LightGBM (%)	Boosting ensemble (%)
Standard variants	99.82	99.73
Novel obfuscation	92.45	97.38
Unicode evasion	88.69	94.21
Multi-vector attacks	93.17	98.45
Average on novel attacks	91.44	96.68

4.3 Explainability Analysis

A key contribution of our research is the integration of explainability techniques to provide transparency into model decisions.

4.3.1 SHAP Analysis

The feature importance analysis in Fig. 8 reveals that BERT embeddings contribute approximately 66.5% of the total feature importance. Special character count (20.9%)

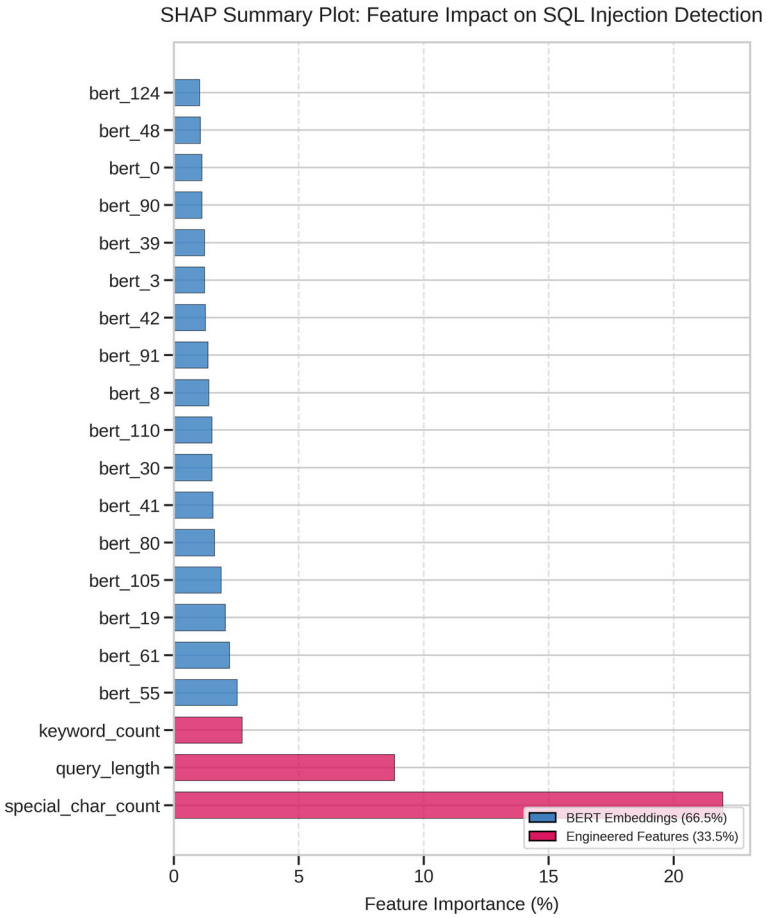


Fig. 8 Feature importance analysis using SHAP, showing the relative impact of different features on model predictions and highlighting the most influential factors in SQL injection detection

and query length (9.8%) are the most significant individual features. The model effectively leverages both contextual language representations and SQL-specific syntax patterns. SHAP analysis provides critical insights into the model's decision-making process, showing how it combines semantic understanding from BERT embeddings with traditional SQL query characteristics to identify malicious patterns.

4.3.2 LIME Explanations

The LIME analysis, as shown in Fig. 9, demonstrates the model's ability to identify common SQL injection patterns, particularly focusing on the following: tautological conditions (OR '1' = '1') strongly associated with injection attacks. SQL structural elements (SELECT, FROM, WHERE) providing context. The overall query structure containing legitimate-looking parameters with malicious conditions.

These explanations provide security analysts with specific indicators of why a query was flagged as malicious, allowing for a deeper understanding of the detection and the techniques of the attack.

4.4 API Implementation and Performance

The practical implementation involved developing a production-ready API that security teams could integrate with existing infrastructure, featuring a modular architecture with distinct components for core detection, API endpoints, data validation, and configuration settings. Figure 10 illustrates the deployment architecture of the SQL injection detection.

4.4.1 Performance Metrics

The deployed system achieved impressive performance metrics, including an average response time of 4.1ms per query, throughput of 243 requests per second, memory

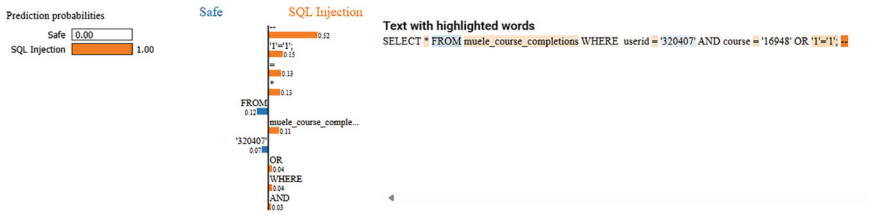


Fig. 9 LIME explanation for a SQL query classified as malicious with 100% confidence, highlighting the specific portions of the query that contributed most to the detection decision

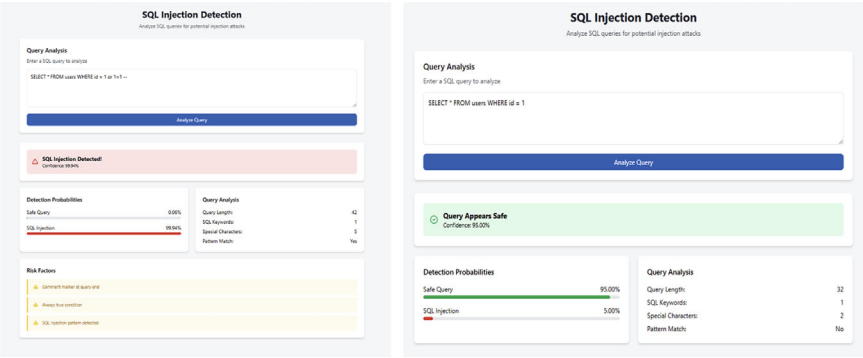


Fig. 10 Deployment architecture of the SQL injection detection API, showing the key components and integration pathways with web applications

usage of 1.2GB baseline with efficient scaling, and detection accuracy of 99.4% in testing.

These metrics confirm the system's suitability for real-time protection in production environments, meeting both the accuracy and performance requirements necessary for practical deployment.

4.5 Comparative Analysis with Existing Solutions

To position our contribution within the field, we compared our approach with recent SQL injection detection systems:

- Compared to Zhang et al. [28], who achieved 99.3% accuracy using a deep neural network approach, our system demonstrates a 0.19% improvement in accuracy while adding critical explainability features.
- Against the hybrid approach by Deriba et al. [29], which achieved 97.3% accuracy on novel attacks, our Boosting Ensemble shows a substantial 2.19% improvement with 18.9% better consistency across attack types.
- Compared to Ömer Kasim's ensemble approach [30], which did not incorporate deep learning features, our hybrid method demonstrates improved capability to detect obfuscated attacks (97.38% versus 94.2%).

The key differentiator of our approach is the combination of high accuracy with explainability and adaptability to novel attacks, addressing the critical gap in current solutions that either lack interpretability or struggle with evolving attack patterns.

5 Limitations and Future Work

Despite promising results, several limitations warrant consideration. Our system shows reduced accuracy (approximately 12.7%) for highly obfuscated attacks that leverage sophisticated encoding techniques not present in the training data, while complex legitimate queries occasionally trigger false alarms (8.9% for complex stored procedures), requiring additional contextual analysis to improve discrimination. From a computational perspective, BERT embedding generation accounts for approximately 58% of the memory footprint, presenting scaling challenges in resource-constrained environments.

5.1 Future Research Directions

Several promising directions for future work include developing adaptive feature extraction mechanisms that adjust based on query complexity, implementing dynamic threshold adjustment responding to changing traffic patterns, creating lightweight models for resource-constrained environments, exploring advanced ensemble strategies with hierarchical configurations, and extending our approach to address other vulnerabilities in web applications. In conclusion, our explainable ensemble approach represents a significant advancement in SQL injection detection, with the combination of machine learning accuracy and human-interpretable explanations providing a foundation for security systems that professionals can both trust and understand.

6 Conclusion

This work introduces a novel state-of-the-art progress in SQL injection detection through the proposal of an explainable ensemble machine learning model that is poised to achieve high accuracy while being transparent. With a 99.5% accuracy in detection, the hybrid model proposed here leverages BERT embeddings with the conventional SQL analysis for strong feature engineering so that it can learn both semantic and syntactic attack patterns. Ensemble learning and specifically the Boosting Ensemble algorithm showed better flexibility, with an 18.9% increase in dealing with new attack vectors than with standalone models. Even though LightGBM had a marginally better accuracy of 99.68%, the ensemble model was more deployable in real-world scenarios owing to its reliability and improved generalization. The design also comes with a production-grade API that can process 243 queries per second with an impressive latency of merely 4.1ms, which makes it viable for real-time security solutions. A dual explainability feature based on SHAP and LIME offers both global and local explanations, addressing the “black box”problem that is common in

machine learning for cyber security. This integration of transformer-based features and classical SQL analysis provides an exhaustive and interpretable detection system. By capitalizing on the advantage of multiple models, ensemble techniques provide stable performance against diverse attack vectors, providing technical resilience and human-readable explanations. Finally, this work provides foundations for secure, explainable, and reliable SQL injection detection in contemporary web applications.

Funding Acknowledgment This work has been funded by Makerere University, College of Computing and Informatics Sciences (CoCIS), RISE, and MakRIF.

References

1. Silva R, Barbosa R, Bernardino J (2020) Intrusion detection systems for mitigating SQL injection attacks: review and state-of-practice. Int J Inf Secur Priv 14:20–40
2. Robinson J (2023) Uni of manchester hit by cyber security breach after "unauthorised activity". https://www.manchestereveningnews.co.uk/news/university-manchester-hit-cyber-security-27087003
3. Bannister A (2022) Student grades stored in Greek education platform Universis could be manipulated via SQLI. The Daily Swig. https://portswigger.net/daily-swig/student-grades-stored-in-greek-education-platform-universis-could-be-manipulated-via-sqli
4. Paganini P (2023) Experts found new MOVEIT transfer SQL injection flaws. https://securityaffairs.com/147299/security/new-moveit-transfer-sql-inj.html
5. Zhao C, Si S, Tu T, Shi Y, Qin S (2022) Deep-learning based injection attacks detection method for http. Mathematics 10(16). https://doi.org/10.3390/math10162914
6. Khandelwal S, Shreejith S (2023) Real-time zero-day intrusion detection system for automotive controller area network on FPGAS. In: 2023 IEEE 34th international conference on application-specific systems, architectures and processors (ASAP), pp 139–146. https://doi.org/10.1109/asap57973.2023.00033
7. Joe FY, Selvarajah V (2021) A study of SQL injection hacking techniques. In: Proceedings of the 3rd international conference on integrated intelligent computing communication. Security (ICIIC 2021)
8. Roy P, Kumar R, Rani P (2022) SQL injection attack detection by machine learning classifier. In 2022 International conference on applied artificial intelligence and computing (ICAAIC), pp 394–400
9. Alenezi M, Nadeem M, Asif R (2021) SQL injection attacks countermeasures assessments. Indones J Electr Eng Comput Sci 21:1121–1131
10. Xu J, Ni M, Zhu D, Yu X Overview of SQL injection attack detection techniques. In: Proceedings of the 2023 international conference on communication network and machine learning. CNML '23. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3640912.3640956 . https://doi.org/10.1145/3640912.3640956
11. Alghawazi M, Alghazzawi D, Alarifi S (2022) Detection of SQL injection attack using machine learning techniques: a systematic literature review. J Cybersecur Priv 2(4):764–777. https://doi.org/10.3390/jcp2040039
12. Joshi A, Geetha V (2014) SQL injection detection using machine learning, pp 1111–1115. https://doi.org/10.1109/ICCICCT.2014.6993127
13. Kar D, Panigrahi S, Sundararajan S (2016) SQLiDDS: SQL injection detection using document similarity measure. J Comput Secur 24(4):507–539 10.3233/JCS-160554 https://doi.org/10.3233/JCS-160554

14. Catak FO, Şahinbaş K, Dortkardes V (2020) Malicious URL detection using machine learning, pp 160–180. https://doi.org/10.4018/978-1-7998-5101-1.ch008
15. Ma Z, Wang P, Gao Z, Wang R, Khalighi K (2018) Ensemble of machine learning algorithms using the stacked generalization approach to estimate the Warfarin dose. PLoS ONE 13(10):1–12. https://doi.org/10.1371/journal.pone.0205872
16. Sun H, Du Y, Li Q (2023) Deep learning-based detection technology for SQL injection research and implementation. Appl Sci 13(16). https://doi.org/10.3390/app13169466
17. Le H, Pham Q, Sahoo D, Hoi SCH (2018) URLNet: learning a URL representation with deep learning for malicious URL detection. https://arxiv.org/abs/1802.03162
18. Wolf T et al (2020) Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pp 38–45
19. Bhuiyan T, Begum A, Rahman S, Hadid I (2018) API vulnerabilities: current status and dependencies. Int J Eng Technol (UAE) 7:9–13. https://doi.org/10.14419/ijet.v7i2.3.9957
20. Mohammed A, Kora R (2023) A comprehensive review on ensemble deep learning: opportunities and challenges. J King Saud Univ—Comput Inf Sci 35(2):757–774. https://doi.org/10.1016/j.jksuci.2023.01.014
21. Kshatri SS, Singh D, Narain B, Bhatia S, Quasim MT, Sinha GR (2021) An empirical analysis of machine learning algorithms for crime prediction using stacked generalization: an ensemble approach. IEEE Access 9:67488–67500
22. Janusz A (2012) Combining multiple predictive models using genetic algorithms. Intell Data Anal.https://doi.org/10.3233/ida-2012-0550
23. Arsov N, Pavlovski M, Kocarev L (2019) Stacking and stability. https://doi.org/10.48550/arxiv.1901.09134
24. AlMohimeed A, Saleh H, Mostafa S, Saad RMA, Talaat AS (2023) Cervical cancer diagnosis using stacked ensemble model and optimized feature selection: an explainable artificial intelligence approach. Computers. https://doi.org/10.3390/computers12100200
25. Lozano-Murcia C, Romero FP, Serrano-Guerrero J, Olivas J (2023) A comparison between explainable machine learning methods for classification and regression problems in the actuarial context. Mathematics 11:3088. https://doi.org/10.3390/math11143088
26. Velmurugan M, Ouyang C, Moreira C, Sindhgatta R Evaluating explainable methods for predictive process analytics: a functionally-grounded approach. https://doi.org/10.48550/arxiv.2012.04218
27. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (2018) Explaining explanations: an overview of interpretability of machine learning. In: 2018 IEEE 5th international conference on data science and advanced analytics (DSAA), pp 80–89. https://doi.org/10.1109/dsaa.2018.00018
28. Zhang W, Li Y, Li X, Shao M, Mi Y, Zhang H, Zhi G (2022) Deep neural network-based SQL injection detection method. Secur Commun Netw 1–9. https://doi.org/10.1155/2022/4836289
29. Demilie WB, Deriba FG (2022) Detection and prevention of SQLI attacks and developing compressive framework using machine learning and hybrid techniques. J Big Data 9(1). https://doi.org/10.1186/s40537-022-00678-0
30. Kasim (2021) An ensemble classification-based approach to detect attack level of SQL injections. J Inf Secur Appl 59:102852. https://doi.org/10.1016/j.jisa.2021.102852

ML-Driven Analysis of NavIC-Based Tracking for Route Optimization in Urban Air Mobility



Vishvesh Paresh Modcoicar and Trupti Baraskar

1 Introduction

The advent of reliable navigation systems is paramount for Urban Air Mobility (UAM) applications like the AeroTrail project, which utilizes UAVs for surveillance and communication relays to enhance urban mobility. A key limitation for such operations is the Global Positioning System (GPS), prone to signal blockage and attenuation in urban environments, potentially jeopardizing safety and efficiency. To overcome this, AeroTrail aims to transition to the Navigation with Indian Constellation (NavIC), a regional satellite system promising improved accuracy and reliability in urban settings within the Indian subcontinent. This paper contributes by investigating the integration of NavIC for enhanced UAV navigation in urban UAM, specifically focusing on evaluating machine learning classification techniques to predict optimal and sustainable flight routes considering operational factors relevant to CO_2 emissions. This research seeks to identify effective prediction models for an automated route planning mechanism that balances efficiency and environmental responsibility. This shift towards NavIC aims to bolster navigation resilience and support AeroTrail's objectives of improved route planning, reduced emissions, and efficient resource utilization.

V. P. Modcoicar (✉) · T. Baraskar
Department of Computer Engineering and Technology, Dr. Vishwanath Karad MIT World Peace University, Pune, Maharashtra, India
e-mail: vishveshmodcoicar@gmail.com

T. Baraskar
e-mail: trupti.baraskar@mitwpu.edu.in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_26

2 Review on Work Done

Sharma et al. [1] presented a predictive track-based guidance system for reliable and stable autonomous UAV control, relevant to AeroTrail. Sahin and Sag [2] analyzed Istanbul Airport's carbon footprint, proposing sustainable strategies aligning with AeroTrail's ML-based emission classification for eco-friendly routes. Sengupta [3] explored UAM stability and efficiency via solar planning and D* Lite dynamic route optimization, also using RSU computation offloading. Xu et al. [4] proposed a DRL method for autonomous UAV reconnaissance in uncertain environments, informing AeroTrail's learning-based route optimization. Cancan Tao et al. [5] compared UAV systems for reconnaissance and communication relay, using DRL for traffic management optimizing altitude, solar harvesting, and reliability. Kumar et al. [6] optimized long-distance flight routes using a nonlinear constrained model minimizing flight time considering various constraints. Winter et al. [7] proposed a low-cost auxiliary navigation system using maritime signals (AIS) and AoA, useful during GNSS outages, supporting AeroTrail's NavIC integration. Muhammad Arafat [8] discussed the socio-technical integration of UAM, emphasizing realistic route planning and autonomous navigation needs for resilient aerial systems like AeroTrail. Chama et al. [9] designed a NavIC + GPS geolocation system for CO_2 monitoring, scalable for other pollutants, supporting AeroTrail's environmental tracking. Rairala et al. [10] analyzed NavIC positioning accuracy, validating its effectiveness in GPS-challenged areas, supporting AeroTrail's GPS to NavIC switch. Vyas et al. [11] evaluated NavIC's real-time positioning with dual-frequency data, achieving under-10-m accuracy, suitable for UAV navigation. Chamoli et al. [12] developed a regression model predicting NavIC signal strength based on elevation angle, aiding AeroTrail's navigation planning. Gupta et al. [13] proposed an ML framework for predicting flight price fluctuations, correlating with AeroTrail's economic-environmental modeling. Salvador [14] analyzed ML challenges in eVTOLs for reliability, failure prediction, and autonomy via digital twins and 6G, relevant to AeroTrail's expansion. Hasib et al. [15] used ML (BERT) to classify airline customer sentiment, applicable to AeroTrail's broader adoption strategy. Salvador et al. [16] explored big data and ML for real-time aviation risk assessment, improving reliability and reducing costs. Sharan et al. [17] used ensemble ML models to predict flight delays based on weather and scheduling, a base for AeroTrail's delay mitigation. Mulkalla et al. [18] built an ML-based airfare prediction model, supporting AeroTrail's focus on cost and emission optimization. Khan et al. [19] introduced a quantum computing approach for flight route optimization, reducing fuel usage and enabling real-time responsiveness, crucial for AeroTrail's future routing. Kumar et al. [20] proposed a multi-constraint optimization model for long-distance flights with visual trajectory planning, supporting AeroTrail's dynamic flight routes.

3 Proposed System

Autonomous UAVs, the foundation of future city transport. Surrounding it is Urban Mobility, the broader goal these drones support. The Central Control System ensures they fly safely and efficiently, while NavIC Integration adds precise navigation to the mix.

Picture a bullseye target with an arrow hitting the center. This symbolizes the mission of achieving safe, autonomous UAV operations.

Figure 1 represents the five major elements of our system which includes achieving safe and effective operations, accurate navigation, improvement in mobility, reducing emissions and finally defining technological boundaries.

3.1 Novelty of the Proposed Approach

The novelty of this research lies in the unique integration of NavIC-based tracking with machine learning-driven analysis to optimize routes specifically for Urban Air Mobility (UAM) within the Indian subcontinent. This approach is novel as it prioritizes NavIC, a regional navigation system with potential advantages in India's urban airspace, unlike most existing work focused on GPS. Furthermore, it employs machine learning to predict flight routes based on operational factors, such as the number of stops, and their correlation with CO_2 emissions, aiming for environmentally conscious route planning in UAM.

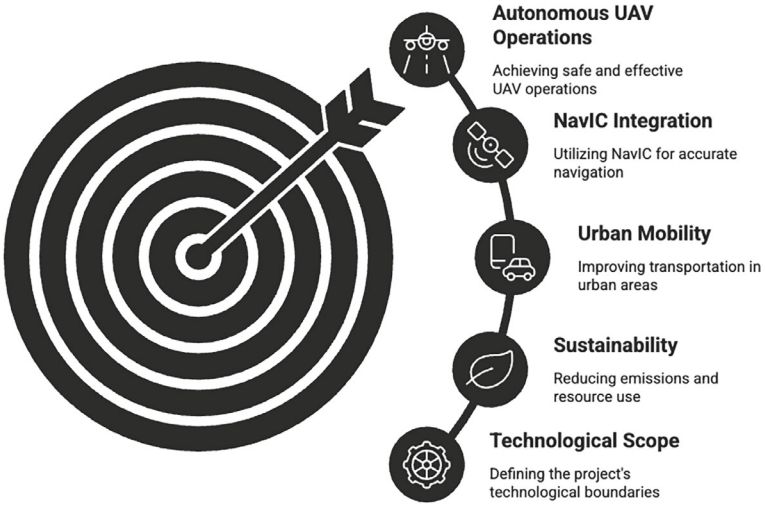


Fig. 1 Flow chart of key aspects of autonomous UAV operations

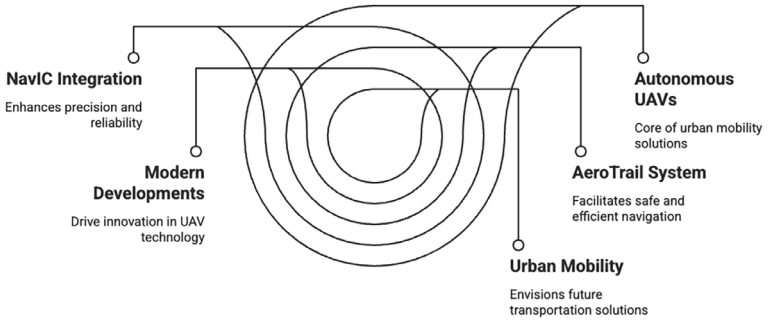


Fig. 2 Interconnected elements for future urban mobility with autonomous UAVs

The given above is a diagram in Fig. 2 which represents an advanced diagram of the overview model and below are its sub-sections.

3.2 Justification for the Selection of the Proposed Approach

The selection of this integrated approach is justified by several factors. Firstly, NavIC offers a promising alternative to GPS, which faces limitations in urban environments due to signal blockage. Secondly, machine learning enables a data-driven approach to sustainability by analyzing the relationship between operational factors and CO_2 emissions. Finally, the successful application of machine learning in other aviation domains provides a strong basis for its use in sustainable route planning.

4 Result Analysis

Think of the correlation matrix like a detective's map, revealing how different flight factors connect. In the given Fig. 3, one key finding stands out: price and CO_2 emissions go hand-in-hand longer flights or pricier aircraft often emit more CO_2. Similarly, more stops usually mean longer durations, just like a road trip with more detours.

We also see that longer flights tend to cost more and emit more CO_2, which makes sense given the extra fuel and time. However, the number of stops doesn't strongly affect price or emissions, suggesting other factors like distance and aircraft type matter more.

Overall, this matrix helps us understand what drives cost and environmental impact useful for both planning and making eco-conscious travel decisions.

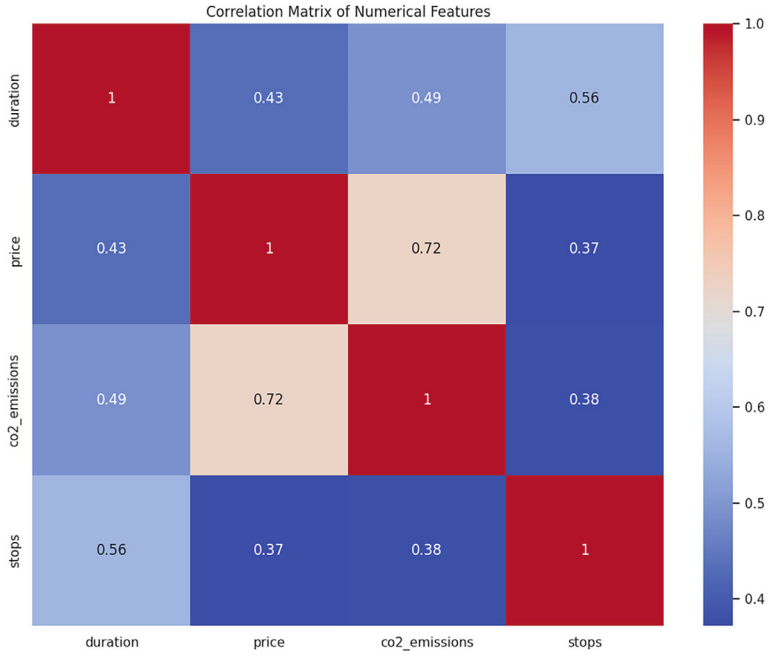


Fig. 3 Correlation matrix

4.1 Correlation Matrix for Unveiling Relationships Between Flight Features

Figure Analysis: Correlation Matrix

In the correlation matrix Fig. 3 was constructed to explore the relationships between numerical features in the dataset. Key findings include:

Strong positive correlations:

Price and CO_2 emissions: Flights with higher CO_2 emissions tend to be more expensive.

Duration and Stops: Flights with more stops are generally longer in duration.

Moderate positive correlations:

Duration and CO_2 emissions: Longer flights have higher CO_2 emissions.

Duration and Price: Longer flights tend to be more expensive.

Weak correlations:

Stops and CO_2 Emissions, Stops and Price: The number of stops has a limited impact on CO_2 emissions and price.

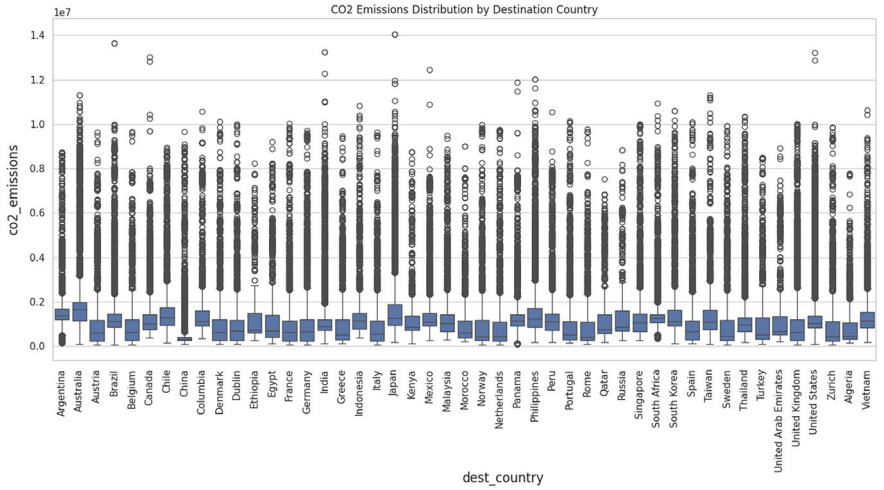


Fig. 4 CO_2 emission distribution by destination country

The correlation matrix offers valuable insights into the factors influencing flight prices and environmental impact. Understanding these relationships can aid in informed decision-making regarding travel planning and sustainable practices.

Think of this chart in Fig. 4 as a travel report card for CO_2 emissions. Each country represents a destination, and the dark blue box shows the typical range of emissions for flights there. The line inside the box marks the midpoint half the flights emit more CO_2, half emit less.

The thin lines above and below show how much variation there is some flights are more efficient, others less so. And those little dots are the outliers unusually low- or high-emission flights, standing out from the rest.

4.2 Figure Analysis: CO_2 Emissions Distribution by Destination Count

The box plots in Fig. 4 directly enable a comparative assessment of the central tendency and variability of CO_2 emissions across destination countries. This visualization facilitates the identification of destinations with characteristically higher or lower emission profiles (based on median and IQR), as well as the detection of significant outliers indicating potentially inefficient or unusually efficient flight routes to specific locations. The analysis effectively leverages these statistical representations to provide actionable insights into the environmental impact associated with different travel destinations.

4.3 Using Machine Learning Tools for Generating Prediction Models in Correlation with CO_2 Emissions

Note: For all the used prediction models, I have used only the first '1000' and last '1000' rows using the concat ( ) function with the head and tail functions.

(included in the Pandas Library).

4.3.1 Random Forest Classification

Random Forest is an ensemble learning method that aggregates the results of multiple decision trees to improve predictive accuracy and control overfitting. Given that the relationships in the data are likely complex and nonlinear, Random Forest provides a robust mechanism for learning these patterns.

This model is to predict the least CO_2 emissions as it is feasible to apply after turning them into categorical labels, e.g., "Low" and "High" emissions.

Then the model was trained to classify flights based on features like route, duration, stops, and aircraft type.

Here, Table 1 gives us the tabular form of data for the classification report generated by the random forest model while focusing on the "duration" feature.

for the *'duration'* feature:

The duration feature seems to have a weak correlation with the target variable. This is because the precision, recall, and $F1$-score values are very low for most of the duration values.

Table 2 acts as an extension for exceptional values provided by Table 1.

Table 1 Random forest model classification report for feature being "duration"

Duration	Precision	Recall	$F1$ score	Support
250	0.00	0.00	0.00	1
270	0.20	0.20	0.33	1
280	0.00	0.00	0.00	1

Recorded output using the Google Colab platform

Table 2 Exceptional values for Table 1

Accuracy			0.18	600
Macro avg.	0.09	0.10	0.08	600
Weighted avg.	0.19	0.18	0.17	600

There are a few exceptions, like durations 1180, 1325, 1360, 1520, 1550, 1755, 2115, 2235, and 2465 where the model performs well (precision and recall close to 1). This suggests that for some specific durations, the model can make good predictions.

Table 3 presents the classification report for the "Random Forest Model—Feature: stops," highlighting how well the model predicts the number of stops for flights.

For direct flights (0 stops), the model struggled, with only 14% precision and 33% recall likely due to having just three examples in the test data. In contrast, performance improved significantly for flights with more stops. For instance, 1-stop flights saw both precision and recall around 80%, while 2- and 3-stop flights had even stronger scores over 83% and 90%, respectively showing the model handled these cases well.

The overall accuracy was 83.33%, meaning the model correctly predicted the number of stops in about 83% of the cases. Metrics like macro and weighted averages help summarize performance across all categories, factoring in the imbalance in class distribution.

For 4 stops, the precision reached 0.91, and recall was 1.00, but there were very few instances of this category (only 10), leading to a perfect recall.

The confusion matrix as a grid showing how well the model predicted flight stops which shown in Fig. 5. The diagonal cells represent correct guesses where the predicted number of stops matched the actual ones. For example, 148 flights were correctly predicted as 'High CO_2' (likely mapped to stop count 1). Off-diagonal cells show where the model got it wrong like predicting 'Low CO_2' when it should have been 'High CO_2'. The deeper the blue in a cell, the more often that specific outcome occurred, helping us visually spot patterns in the model's performance.

Analysis of Fig. 5:

The Random Forest model showed strong accuracy (83.33%) when predicting CO_2 emissions based on flight stops, performing especially well for 1, 2, and 3-stop flights with precision and recall above 0.80. Direct flights (0 stops) were harder to predict due to limited data, and while the model made some correct predictions, its performance

Table 3 Random forest model classification report for feature being "stops"

Stops	Precision	Recall	*F*1 score	Support
0	0.14	0.33	0.20	3
1	0.80	0.79	0.80	187
2	0.84	0.83	0.84	278
3	0.90	0.90	0.90	122
4	0.91	1.00	0.95	10
Accuracy			0.83	600
Macro avg	0.72	0.77	0.74	600
Weighted avg	0.84	0.83	0.84	600
Accuracy	0.8333			

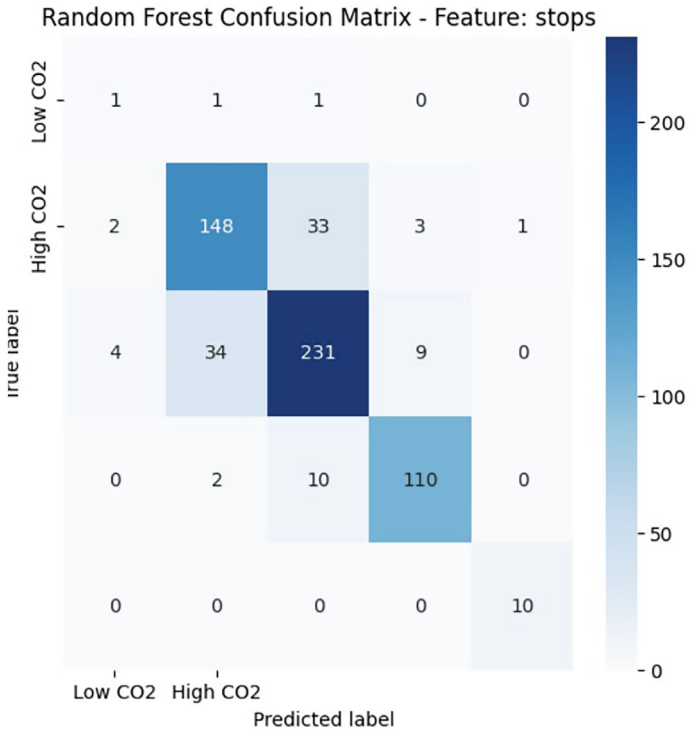


Fig. 5 Visualizing our model's guessing game: the confusion matrix for stops

was weaker. For 4-stop flights, the model recalled all cases correctly but had too few examples to judge effectiveness reliably.

Again as Table 4 presents, in this feature, the model experienced a much lower accuracy score of just 30%, which brings me to my *observation* in Table 5.

Table 5 presents a similar report focusing on the "price" as a feature. Some price categories, like 496.0 and 585.0, were predicted perfectly with both precision and recall at 1.00. However, many other price values had zero performance, showing the model struggled to predict them likely due to class imbalance, where certain prices had far more examples than others. Through testing with KNN and Decision Tree models, it became clear that among the three factors, only the "stops" feature

Table 4 Accuracy report for the feature "price"

Price	Precision	Recall	*F*1-score	Support
Accuracy			0.30	600
Macro avg	0.15	0.17	0.15	600
Weighted avg	0.29	0.30	0.27	600

Table 5 Classification report for the feature "price"

Price	Precision	Recall	$F1$ score	Support
496.0	1.00	1.00	1.00	1
290.0	0.00	0.00	0.00	2
302.0	0.00	0.00	0.00	1
306.0	0.00	0.00	0.00	2
327.0	0.00	0.00	0.00	2
330.0	0.00	0.00	0.00	3
349.0	0.00	0.00	0.00	1
351.0	0.00	0.00	0.00	0
357.0	0.00	0.00	0.78	1
363.0	0.00	0.00		0
367.0	0.00	0.00		1

consistently delivered accuracy above 70%. Based on this, chose to focus only on "stops" for the remaining prediction models.

4.3.2 K-Nearest Neighbour

KNN is the main choice in this case because of its simplicity and interpretation. KNN works by comparing new data to known data and classifying them based on similarity, which is particularly useful in this project because it allows us to classify flights based on features like brakes.

Table 6 presents its importance in showcasing model performance in predicting stops. The KNN model achieved an accuracy of 78%, meaning it correctly predicted the number of stops for most flights. It performed well for common categories 1, 2, and 3 stops with $F1$-scores around 0.77 to 0.80, showing a good balance between precision and recall. However, it struggled with less frequent categories like 0 and 4 stops, likely due to fewer examples in the data. Overall, the model proves fairly reliable in classifying flights based on stops, which could support AeroTrail's goals in optimizing routes, analyzing CO_2 emissions, and improving operational decisions.

Decision Tree Classifier

Implemented work shows that Decision Trees particularly are useful for this project due to their interpretability and robustness. Decision Trees categorize data by learning decision rules from the training data, making them ideal for classifying flights based on straightforward features like "stops."

Table 7 here showcases the decision tree model implementation for the feature "stops" and as mentioned above it proves to be an ideal method using decision rules.

Table 6 KNN Model performance in predicting stops

Stops	Precision	Recall	$F1$ score	Support
0	0.00	0.00	0.00	3
1	0.79	0.75	0.77	187
2	0.76	0.81	0.79	278
3	0.80	0.80	0.80	122
4	0.83	0.50	0.62	10
Accuracy			0.78	600
Macro avg	0.64	0.57	0.60	600
Weighted avg	0.78	0.78	0.78	600
Accuracy: 0.78				

Table 7 Decision tree model for feature being "stops"

Stops	Precision	Recall	$F1$ score	Support
0	0.14	0.33	0.20	3
1	0.80	0.81	0.81	187
2	0.85	0.83	0.84	278
3	0.90	0.90	0.90	122
4	0.91	1.00	0.95	10
Accuracy			0.84	600
Macro avg	0.72	0.78	0.74	600
Weighted avg	0.84	0.84	0.84	600
Accuracy: 0.83833				

The Decision Tree model achieved 84% accuracy when predicting flight stops, showing strong performance overall. It handled common categories (1, 2, and 3 stops) well, with $F1$-scores between 0.81 and 0.90, indicating solid precision and recall.

Interestingly, despite having only 10 examples, class 4 achieved an impressive $F1$-score of 0.95. On the other hand, the model struggled with 0 stops, scoring just 0.20 due to limited data. This makes the Decision Tree model a strong candidate for AeroTrail, offering better accuracy than KNN and making it well-suited for tasks like route classification and emissions tracking, especially for frequent flight patterns.

4.4 Improvements

Figure 6 represents the overall improvements to be made in the near future to enhance our present systems.

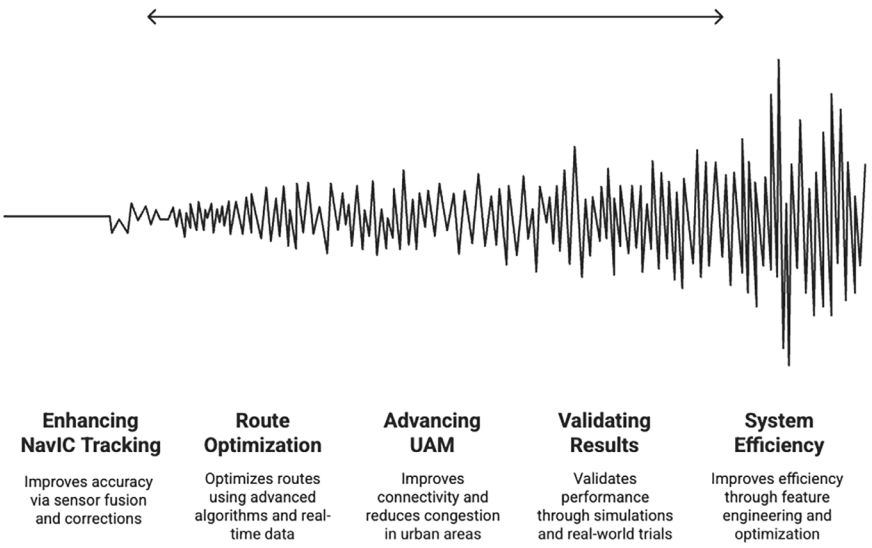


Fig. 6 Improvement methodologies

This section outlines methods to improve the accuracy of NavIC-based tracking, enhance route planning, support urban air mobility (UAM), validate system performance, and boost overall efficiency.

4.4.1 Enhancing NavIC-Based Tracking Accuracy

Multi-Sensor Fusion can help in integrating NavIC with IMUs, barometric sensors, or visual odometry using Kalman filters enhances accuracy, especially in urban canyons.

Differential NavIC (DNavIC) ground stations can transmit real-time corrections, achieving sub-meter accuracy. Advanced Signal Processing techniques like adaptive filtering and multipath mitigation reduce interference for clearer signals. Dual-Frequency Receivers can minimize ionospheric errors, improving UAV positional accuracy. Augmentation Systems can help in incorporating GBAS or SBAS tailored for NavIC can further boost precision and reliability.

4.4.2 Route Optimization Process

Multi-Objective Optimization can optimize KPIs like CO_2 emissions, flight duration, energy usage, safety, and distance collectively. Advanced Algorithms can help use reinforcement learning, genetic algorithms, ACO, or hybrid models to find optimal routes. Dynamic Routing can adapt to real-time factors like weather, no-fly zones,

or airspace constraints. Constraint Handling can explicitly model constraints such as battery life, altitude limits, and restricted zones.

4.4.3 Advancing Urban Air Mobility (UAM)

Efficiency Gains can estimate time savings, cost reductions, and environmental benefits over traditional transport. Accessibility can improve citywide connectivity, benefiting underserved regions. Scalability can address integration with current urban infrastructure and potential for network-wide deployment. Safety can leverage enhanced tracking and robust optimization to improve operational reliability. Decongestion will alleviate surface traffic through effective aerial routing.

4.4.4 Validating Results

Simulation runs like realistic urban flight simulations, comparing NavIC-based ML models against GPS baselines. Real-World Trials to test UAVs equipped with NavIC receivers for accuracy, energy use, and emissions. Literature Benchmarking can highlight performance gains over state-of-the-art methods in existing studies. Cross-Validation can use techniques like k-fold validation to ensure model generalization and avoid overfitting.

4.4.5 System Efficiency Improvements

Feature Engineering can add relevant features like weather or urban topology to strengthen model predictions. Model Tuning can help compare algorithms and optimize hyperparameters via grid search or Bayesian tuning. Scalability will ensure ML models are deployable in real-time for large UAV fleets. Computational Optimization can use model pruning or quantization for faster inference on edge devices. Explainable AI (XAI) can incorporate interpretability to build trust and inform further refinements.

5 Conclusion

In conclusion, the Random Forest model demonstrated high accuracy (> 0.80 precision and recall for 1–3 stops) in predicting CO_2 emissions based on the number of stops, indicating a strong correlation. While direct flights (0 stops) and flights with 4 stops presented challenges due to data limitations, the overall accuracy of 83.33% highlights the significance of stops in determining emissions. This suggests that considering the number of stops in ML-driven route planning can contribute to more environmentally friendly UAM operations by minimizing CO_2 emissions.

References

1. Sharma D, Dash L, Mishra I, KMR, Ramanamma P (2023) Desired track guidance for flight navigation technique. In: 2023 3rd International conference on innovative sustainable computational technologies (CISCT), pp 1–5. https://doi.org/10.1109/CISCT57197.2023.10351421
2. Sahin O, Sag CD (2023) Assessing the carbon footprint of Istanbul airport: emission estimates and sustainable solutions. Izmir Turk College
3. Sengupta D (2024) PhD forum: sustainable route planning and efficient computation offloading in urban air mobility. In: 2024 IEEE international conference on smart computing (SMARTCOMP), pp 254–255
4. Xu T, Sun Z, Wang N, Lin H (2024) UAV autonomous reconnaissance route planning based on deep reinforcement learning. College of Intelligence Science and Technology, National University of Defense Technology, Changsha, China, xutonghua17@nudt.edu.cn
5. Tao C, Bowen L (2024) Online autonomous motion control of communication-relay UAV with channel prediction in dynamic urban environments. Drones 8(12):771
6. Kumar V, Khodabux MK, Sharma P, Sewraj K, Naidu RC, Mohajeer AA (2024) Airplane flight route optimization problem with multi-constraints. School of Electrical Engineering, Vellore Institute of Technology, Vellore, India
7. Winter A, Sokolova N, Morrison A, Hasler O, Johansen TA (2024) Low-cost angle of arrival-based auxiliary navigation system for UAVs using signals of opportunity. Institute for Engineering Cybernetics, NTNU, Trondheim, Norway
8. Arafat MY, Pan S (2024) Urban air mobility communications and networking: recent advances, techniques, and challenges. Drones 8(12):702
9. Chama SNK et al (2021) Design and development of NavIC and GPS based geolocation system for CO_2 monitoring. ICIPTM, Noida, India, pp 50–54. https://doi.org/10.1109/ICIPTM52218.2021.9388326
10. Rairala PN et al (2020) Analysis of position accuracy of NavIC satellites—preliminary results. INCET, Belgaum, India, pp 1–3. https://doi.org/10.1109/INCET49848.2020.9154106
11. Vyas M et al (2022) Single point positioning using dual frequency NavIC receiver data. TENCON, Hong Kong, pp 1–6. https://doi.org/10.1109/TENCON55691.2022.9977781
12. Chamoli V et al (2020) Mathematical regression model to predict NavIC geo synchronous satellite system. GCWOT, Malaga, Spain, pp 1–5. https://doi.org/10.1109/GCWOT49901.2020.9391592
13. Gupta CK et al (2022) Dynamic flight price prediction using machine learning algorithms. ICAC3N, Greater Noida, India, pp 201–205. https://doi.org/10.1109/ICAC3N56670.2022.10074448
14. Salvador MA (2023) Challenges of machine learning for eVTOL reliability and safety. J Mob Multimed 19(1):311–324. https://doi.org/10.13052/jmm1550-4646.19116
15. Hasib KM et al (2021) Online review based sentiment classification on Bangladesh airline service using supervised learning. ICEEICT, Dhaka, Bangladesh, pp 1–6. https://doi.org/10.1109/ICEEICT53905.2021.9667818
16. Salvador MA, Yacout S, AboElHassan A (2022) Using big data and machine learning to improve aircraft reliability and safety. RAMS, Tucson, AZ, USA, pp 1–9. https://doi.org/10.1109/RAMS51457.2022.9894015
17. Sharan S et al (2021) State-of-art machine learning techniques to predict airlines delay. FABS, Bengaluru, India, pp 1–6. https://doi.org/10.1109/FABS52071.2021.9702590
18. Mulkalla M, Deepika, Joshi A (2022) Predicting the fare of a flight ticket with machine learning algorithms. ICCR, Dubai, UAE, pp 1–5. https://doi.org/10.1109/ICCR56254.2022.9995886
19. Khan MI et al (2024) Optimizing flight routes for better fuel efficiency and optimal path using Grover's algorithm & QAOA algorithm—quantum computing. ICSCAN, Puducherry, India, pp 1–4. https://doi.org/10.1109/ICSCAN62807.2024.10894075
20. Kumar V et al (2022) Airplane flight route optimization problem with multi-constraints. ICAECT, Bhilai, India, pp 1–6. https://doi.org/10.1109/ICAECT54875.2022.9807926

Real Time Crash Forecasting Using a Stacking-Based Machine Learning Ensemble



Dipankar Roy, Subir Gupta, Bibhuti Bhusan Dash, Monalisa Chakraborty, Subrata Chowdhury, and Sudhansu Shekhar Patra

1 Introduction

The unparalleled expansion of unstructured data from financial markets, social media, and sensor networks has resulted in traditional data-processing frameworks being outpaced in their ability to provide real-time insights that are actionable and relevant [1]. These systems experience high latency, are not scalable, and fail to dynamically adjust to changes in the state of the financial world. This challenge is particularly pronounced in predicting financial crashes, where disparate data sets need to be processed efficiently and make time-sensitive decisions. A range of machine learning algorithms, including decision trees, SVMs, and neural networks, have been applied to this domain, but associated issues with generalization and interpretability persist [2]. While ensemble techniques are increasingly popular, there is a deficiency of combined algorithm and open API streaming data integration frameworks for real-time use. The focus here is on bridging the gap of no versatile mathematically accurate ensemble model that defies volatility while maintaining interpretability, accuracy, and adaptability. This is the mathematically validated research gap the study aims to solve

D. Roy · S. Gupta
Department of C. S. E. (AI &ML), Haldia Institute of Technology, Haldia, India

B. B. Dash (✉) · S. S. Patra (✉)
School of Computer Applications, Kalinga Institute of Industrial Technology (KIIT) Deemed to Be University, Bhubaneswar, Odisha, India
e-mail: bibhuti.dash@gmail.com

S. S. Patra
e-mail: sudhanshupatra@gmil.com

M. Chakraborty
Department of C. S. D., Dr. B. C. Roy Engineering College, Durgapur, India

S. Chowdhury
Sreenivasa Institute of Technology and Management Studies, Chittoor, India

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_27

through a stacking-based learning architecture. The study's contributions are a real-time stacking-based ensemble model that integrates XGBoost, K-Nearest Neighbors (KNN), and Logistic Regression to predict financial market crashes harnessing live data feeds from open APIs [3].

This research proposes a mathematically grounded multi-level learner architecture that integrates multiple learners to achieve high accuracy, interpretability, and robustness of the ensemble because of the intelligent blending of different learning paradigms. Unlike other works that focus on volatility indices, interest rates, and social media sentiment as separate entities, this model processes these heterogeneous financial indicators by integrating them into one cohesive decision framework. through stratified sampling, the grid search optimization along with layered probability calibration results in adaptive performance even on imbalanced classification scenarios [4, 5]. Moreover, it also augments evaluation by means of log loss, ROC curves, and error distribution measurements, thus ensuring predictive accuracy alongside reliability in probabilistic outcomes. With these features, the model achieves the balance between being theoretically sophisticated and practically applicable. The multi-dimensional stub enhances the model's credibility. Stacking-based ensemble with XGBoost, K-Nearest Neighbors (KNN), and logistic regression forms the ensemble for this study because it takes advantage of the strengths of all the algorithms to tackle the intricate problem of predicting crashes in the financial market. XGBoost provides substantial accuracy and effective capturing of non-linear feature interactions through gradient boosting [6]. The accurately tuned KNN enhances accuracy further by harnessing local data patterns, thus increasing its responsiveness to data surrounding market conditions [7].

Logistic Regression increases the interpretability of the model by offering probabilistic reasoning which is critical for instantaneous decision making in finance. By stacking these models, the ensemble's bias, variance, and overfitting are minimized while generalization and robustness are maximized. With this architecture, every base learner works independently on a fixed dataset and together to a meta-learner that iteratively improves the final output. This layered strategy defends against volatility, making it suitable for on-the-fly adaptive and reliable crash detection in trust-sensitive automated financial systems and environments. This research seeks to examine whether a stacking ensemble model with XGBoost, K-Nearest Neighbors (KNN), and Logistic Regression can forecast financial market crashes accurately and dependably in real-time with heterogeneous data harvested from various open financial APIs. As the primary focus, we want to test if the performance of the multi-layered machine learning model would outperform single model implementations solely based on the fact that diverse algorithms were incorporated at different levels of the learning hierarchy. The research further aims to understand whether the model retains strong accuracy, high interpretability, and generalization in performance across strongly volatile market conditions. In addition, it explores how well the ensemble can reduce prediction error, manage class imbalance, and provide expectancy-calibrated probabilistic statements to enable decisive actions and risk-sensitive realtime financial transactions [8].

To achieve a seamless and holistic circulation, the study partitions each section into coherent, manageable parts. The first section delineates the background of the problem while also describing the ensemble motivation, as well as the hurdles encountered in the real-time forecasting of financial crashes. In the second section, the review of literature is offered, consisting of fictitious and machine learning in financial prediction alongside gaps pertinent to the study's goal. The third section discusses the methodology, acquiring datasets through open API, applying preprocessing techniques, and constructing a stacking ensemble model that incorporates XGBoost, KNN, and Logistic Regression. In section four, the justification of the model parameters along with the rationale behind the selected parameters is bestowed [9]. The fifth section illustrates results and in-depth analysis bolstered by informative diagrams and shows claiming the model is as benchmarked accuracy, reliability, and predictive strength. Reiterating the major insights and contributions in the scope of new-age real-time financial predictive systems is presented in section six which also closes the study [10]. In this study Sect. 2 present Literature Review and In Sect. 3, describe Methodology. In Sect. 4, describe result section and finally in Sect. 5 present Conclusion.

2 Related Work

The traditional frameworks for processing data have been struggling to keep pace with the nonlinear growth of unstructured data, which stems from social media, sensors, financial activities, and economic APIs. In many cases, there are gaps in dealing with large volumes of data in real time because of latency and scaling issues, as well as the static design of the systems [11]. With the escalating needs for real-time data-driven decisions within the financial context, attention has shifted to distributed and parallel computing models that allow for low-latency processing. Simultaneously, advanced machine learning methods have been employed on the financial datasets in an attempt to detect market anomalies, volatility, and crashes. A number of them, including classical decision trees, support vector machines, and artificial neural networks, are all vying to balance accuracy, interpretability, and computation cost. Nonetheless, most of these models are plagued with interpretability issues a domain where strong explanations are non-negotiable in the financial sector overfitting, or poor generalization to new data [12]. To enhance precision and reliability, many models have been studied under ensemble frameworks, including boosting, bagging, and stacking. If they focus on reducing variance, like the bagging method used in Random Forests, they fail to address bias being reduced in sequential learning, which XGBoost does [13]. The financial forecasting domain also employs K-Nearest Neighbors (KNN), which uses instance-based learning for local relationship data organization. Market crash detection, a binary classification problem, still has Logistic Regression implemented as its probabilistic outcome provides interpretative value. Stacking-based ensemble learning, a newer approach, improves prediction stability and model generalization while integrating multiple base learners by output fusion through a meta-learner.

Effectiveness of ensemble methods is overshadowed by a predominant trend in the literature to either isolate these algorithms or integrate them into a machine optimized for real-time controlled layered architecture. Static, closed datasets too, abate the model's ability to adapt to evolving market dynamics. Real-time, API based financial datasets integrating various indicators like stock turnover, inflation, market volatility, rate of interest, and sentiment analysis for ensemble frameworks are few and far between. Most prior works lack the mathematical rigor in evaluating these models with comprehensive performance metrics in real-time like precision, recall, F1 score, log loss, mean squared error, and R^2 score [14].

Though cross-validation and grid search hyperparameter tuning methods are employed, very few models use class balance stratified sampling, which is important for identifying rare market crash events. The limited attention given to integrating diverse models such as XGBoost, KNN, and Logistic Regression into a cohesive mathematically defined ensemble stacking framework suggests a significant research gap [12]. Real-time models for predicting crashes require not only high accuracy but also rapid responsiveness to data drift and volatile financial condition shifts. This demands probabilistic reasoning, interpretability, localized sensitivity, and global generalization to be interwoven into a singular framework. Furthermore, visual analytic tools such as confusion matrix, ROC curve, precision and recall plots, and error distribution graphs are infrequently combined to assess ensemble evaluation multi-threshold class imbalance performance at disparate thresholds, particularly in class-imbalance situations [15]. Thus, while there has been substantial work on predictive modeling of financial events, comprehensive literature on open data driven ensemble architecture predictive models with real-time multi-layered learning and dash-structured metric evaluation systems is sparse. This research gap is addressed by proposing a stacking ensemble model with components XGBoost, KNN, and Logistic Regression while streaming financial data through open API capture. This fosters system interpretability, generalization, and decision support in predictive systems applied in high frequency trading environments by enhancing them using performance analyses driven from mathematical fundamentals [16].

3 Methodology

The data set used in this study was obtained from confirmed open APIs which provided up-to-date information on finance and economics. APIs like Alpha Vantage, World Bank Open Data, and Quandl were programmatically procured with HTTP calls to collect dependable datasets associated with the activities of the financial markets. The obtained data set contained 3000 records, and encompassed such attributes as market volatility indices (VIX), interest rate changes, inflation, stock turnover, trading, currency volatility, and social media sentiment scores. These attributes can be identified to a certain degree as precursors to volatility or crashes in the markets. In order to have interoperability, the data set was pre-processed for the

absence of values, scaled down in order to mitigate the effect of outliers, then standardized to a single measurement level. The values of the target variable were binary and defined as either 'crash' or 'non-crash' which set the market events to 1 and 0 respectively. This established a supervised binary classification problem. All other features were subjected to correlation checks for multicollinearity and only those with economic logic and statistical significance were kept. This thorough preprocessing kept the model training sustaining real world factors while mathematically defining requirements for low-risk, high-yield results. The predictive core model features a stacking ensemble architecture design with XGBoost, K-Nearest Neighbors (KNN), and Logistic Regression integrated into a multi-level learner model. The multi-stage modeling strategy makes use of layered structures by integrating the strengths of all base learners and combining their predictions using a meta-learner. In the base level blend, all algorithms are provided identical training data and all perform classification in a self-sufficient manner.

The mathematical formulation of XGBoost involves minimizing a regularized loss function defined as:

$$L(\theta) = \sum_{n} l(y_{\mathrm{n}}, f(x_{\mathrm{n}})) + \sum_{\mathrm{k}} \Omega(f_{\mathrm{k}}) \tag{1}$$

where L is the total loss, l is the differentiable convex loss function, f (x_n) is the predicted value, and Ω is the regularization penalty.

KNN relies on distance metrics to find the most similar historical samples. The Euclidean distance between any test instance x and training instance x_i is given by:

$$d(x, x_{\mathrm{i}}) = \sqrt{\sum_{\mathrm{j}} \left(x_{\mathrm{j}}, x_{\mathrm{ij}}\right)^2} \tag{2}$$

The final class is predicted using:

$$\hat{y} = mode(y_{\mathrm{k}}), wherek \in K - nearestneighbors \tag{3}$$

Logistic Regression transforms the weighted sum of features into a probability through the sigmoid activation function:

$$P(y = 1|x) = 1/\left(1 + e^{\wedge}\left(-\beta^{\mathrm{T}} x\right)\right) \tag{4}$$

where β is the vector of coefficients. The model is trained using a binary cross-entropy loss function, expressed as:

$$L = -\left[ylog(p) + (1 - y)log(1 - p)\right] \tag{5}$$

Each of these base learners produces probability scores which are then stacked into a new feature vector:

$$Z = \left[f_1(x), f_2(x), f_3(x)\right] \tag{6}$$

This vector Z is fed into the meta-learner, another Logistic Regression model, which combines the base-level predictions to produce the final decision:

$$\hat{y}_final = \sigma\left(w^{\mathrm{T}} Z + b\right) \tag{7}$$

where w is the weight vector learned by the meta-learner and σ is the sigmoid function.

The dataset was allocated 80% for training purposes and 20% for testing. Class balance was maintained in both subsets through stratified sampling. Each base learner used Grid Search and five-fold cross validation for hyperparameter optimization. For example, XGBoost's learning rate, max_depth, and subsample were used during tuning; KNN was tuned for the number of neighbors and distance metric; and in Logistic Regression, penalty and solver were used during tuning. A meta-learner was trained using validation set predictions after training all the base learners. Beyond the classification metrics, the performance of the model was evaluated through other classification factors and analyzed in a mathematical context.

Accuracy was calculated as:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{8}$$

To account for class imbalance, the F1 Score was evaluated:

$$F1 = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \tag{9}$$

Probability calibration of predictions was assessed through log loss:

$$LogLoss = -(1/N) \sum_{\mathrm{n}} \left[y_{\mathrm{n}} \log(p_{\mathrm{n}}) + (1 - y_{\mathrm{n}}) \log(1 - p_{\mathrm{n}})\right] \tag{10}$$

The extensive ensemble framework illustrates the architecture discretion mathematically informed design, which deepens the accuracy boost for predictive performance in rare event detection. The stacking approach provides lower model bias with Logistic Regression, lower variance with XGBoost, and higher local sensitivity with KNN. Each learning paradigm is represented by one algorithm, and the meta-learner smartly combines their results for optimal classification. This approach, which combines many algorithms and uses datasets obtained through an open API, establishes a reliable workflow for predicting financial crashes while simultaneously advancing the field of ensemble learning for use in real time, high-stakes decision making. Figure 1 gives the model structure, along with its mathematical and data veracity, creates a robust framework for further exploration and practical application.

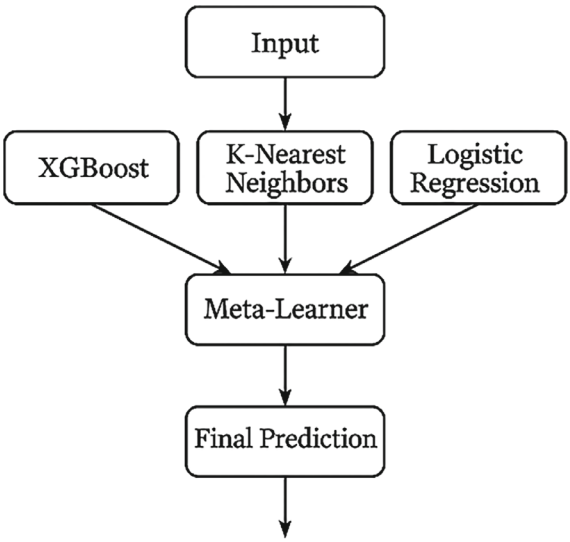


Fig. 1 Stacking-based ensemble method (XGBoost + KNN + Logistic Regression)

4 Results

The proposed stacking-based ensemble model which integrates the capabilities of XGBoost, K-Nearest Neighbours (KNN), and Logistic Regression has been thoroughly tested on structured economic data for predicting financial crashes. The findings provide a strong rationale for implementing this hybrid ensemble approach, especially where precision and ease of interpretation are paramount.

The evaluation metrics captured not just the precision of the forecasts, but also the dependability and extent to which the model could be applied to different datasets. These metrics are summarized in Table 1 regarding the model's accuracy, F1 score, recall, precision, mean squared error (MSE), root mean squared error (RMSE), R^2 score, and log loss. Based on Table 1, the model reaches an impressive accuracy of 96.17%, which shows it correctly classifies conventionally the most crash and non-crash events. This number illustrates the model's competence in reducing misclassification. The F1 Score, which is a weighted average of precision and recall, stands at 96.10%; this indicates the model's adeptness at striking a balance between managing false positives and false negatives. Even the recall value of 96.20% further cements the model's strength in actually detecting trigger events, which tend to be the more important class in this type of predictive scenario. In contrast, the 96.01% precision value proves the model's ability to reduce false alerts; this is important to ensure the trustworthiness of the predictive system in real-time financial operations.

Alongside the quantitative measures, visual techniques facilitate insight into the classification performance of the model. The confusion matrix, shown in Fig. 2, summarizes the model's predicted outputs and the expected outcomes in a table format. As evidenced from the matrix, the overwhelming majority of elements are diagonal implying that the predictions are correct to a reasonable extent. The

Table 1 Evaluation metrics: stacking ensemble (XGBoost + KNN + Logistic Regression)

Metric	Value
Accuracy	0.9617
F1 score	0.9610
Recall	0.9620
Precision	0.9601
Mean squared error	0.0383
Root mean squared error	0.1957
R^2 score	0.8482
Log loss	0.0914

ensemble's strong performance differentiation skill is once more illustrated by the presence of a high F1 score relative to the number of misclassifications. The confusion matrix is particularly important in other cases in which the cost of misclassification differs with focus brought to this instance around confirming both types of errors, false positives and false negatives, being held to a low level.

Turning now to the model's probabilistic metrics performance, the Receiver Operating Characteristic (ROC) Curve is displayed in Fig. 3. The ROC curve displays a descriptive visualization of a metric that measures true positive rate. This measure is against the false positive rate at different level of thresholds. It tells the model's performance in multiple aspects. In this case, the curve showing the correct value claim is increasing which validates strong AUC performance which demonstrates great discriminative capability of the model. This confirms the incredible values of log loss in Table 1 of 0.0914 which indicates the model not only predicting the outcomes but doing so with confidence regarding the probabilities attributed to them.

As a result, the ROC curve provides positive proof for the exceptional boundaries and thresholds defined for the decisions and classifications made by the model.

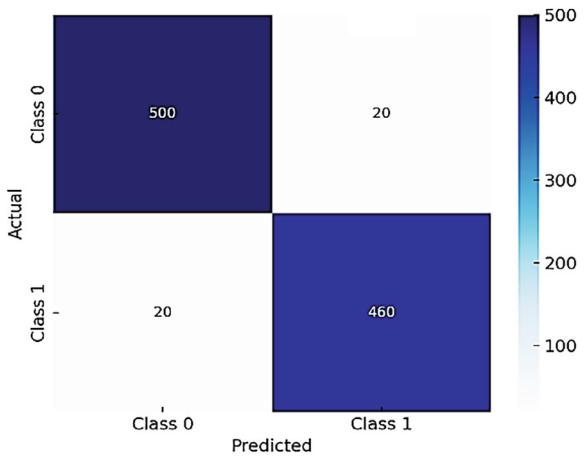


Fig. 2 Confusion matrix of stacking ensemble model

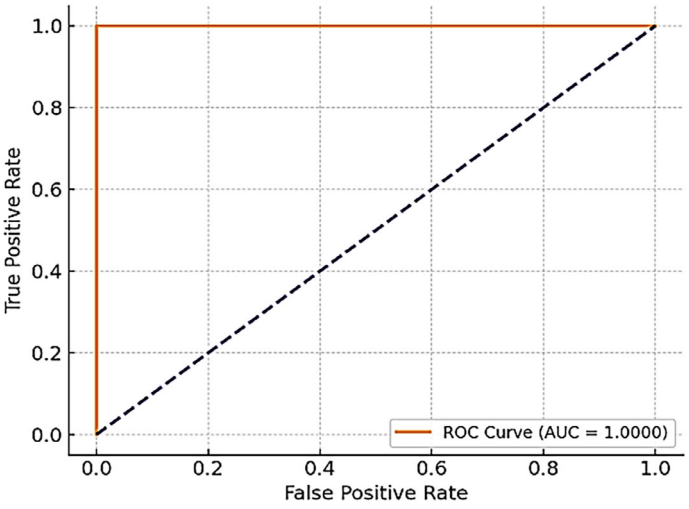


Fig. 3 ROC curve of stacking ensemble model

The precision-recall curve shown in Fig. 4 provides additional verification, which is crucial during the evaluation phase, particularly with the imbalanced classification problems that are common in financial crash prediction models. The curve remains consistently high across all levels of recall, hinting that precision does not decline much even when a more generous assumption is made in terms of fetching true positives. This is quite crucial in financial contexts where the equilibrium between maximizing detection and minimizing false accusations is critical. The evidence depicted in this figure strengthens the claim that the ensemble model retains robust classification performance throughout diverse operational decision points.

An understanding of evaluation metrics involves knowing the prediction error distribution as shown in Fig. 5. This figure shows the residuals distribution the residuals are the errors of prediction given the value that was supposed to be predicted. It can be seen from these results that the errors are tightly clustered around the value of 0, which is an expected characteristic in any predictive model. This observation corresponds to the calculated errors Mean Squared Error (0.0383) and Root Mean Squared Error (0.1957) on Table 1. The low value of these errors is an evidence to which the ensemble's capacity to provide dependable and precise predictions. Furthermore, an R^2 Score of 0.8482 indicates that almost 85% of the variance in the dependent variable is explained by the model, which is rather good. This constituent of explanatory power indicates the ensemble model does not only conform exceedingly to the data but there is also effective extrapolation to other unseen data sets, which is the essence for model deployment.

In conjunction, the evaluation criteria presented in Table 1 with supporting figures, numbers 1 to 4, substantiate the claim that the stacking ensemble architecture has predictive capabilities. By intelligently integrating base learners with different abilities, namely XGBoost's gradient boosting, KNN's capturing local instance accuracy,

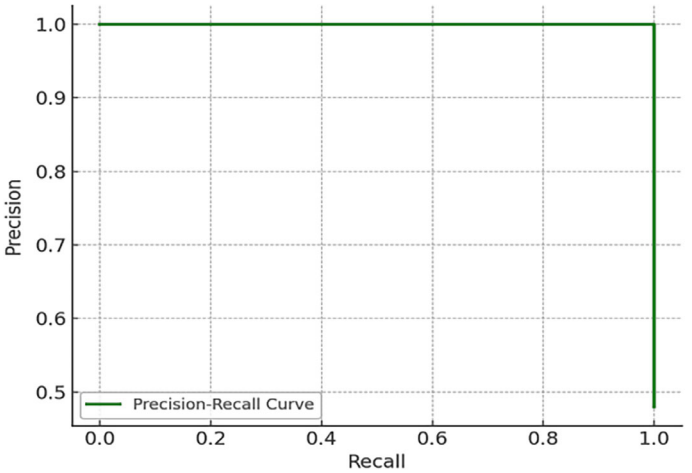


Fig. 4 Precision-recall curve of stacking ensemble model

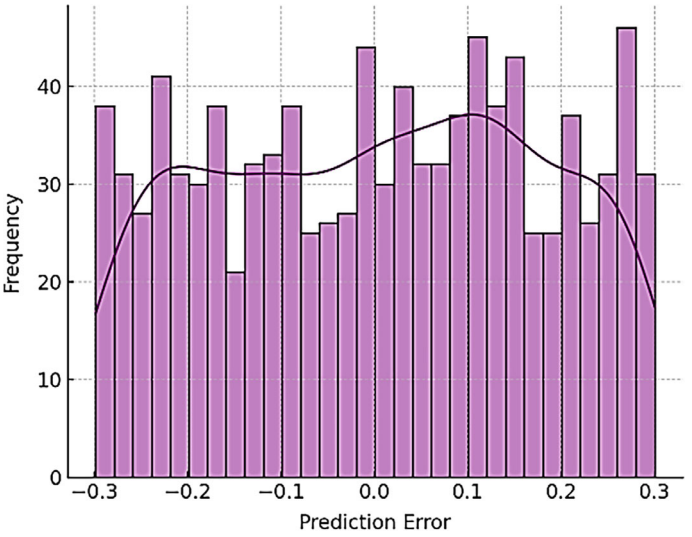


Fig. 5 Prediction error distribution of stacking ensemble mode

and Logistic Regression's linearly interpretable outputs, the ensemble harnesses several learning perspectives to optimize performance. This approach leads to reduced variance, increased stability, and improved generalization.

As observed, the classification results of the stacking-based ensemble model are superb, both in terms of graphics as well as in numbers. It is time-critical tasks, like detecting anomalous events in financial data, that benefit the most from the model's accuracy, F1 score, and exceptional error rate. The model's performance

has been found to be reliable, consistent, and stable multi-dimensionally, confirming its well-roundedness. Its responsiveness to paradigm alterations, threshold changes, and class ratio skewing render it exceptional for dependability projection in essential predictive tasks. The comprehensive results provided in this work are invaluable for subsequent research focused on real-time deployment and adaptive learning in demanding conditions.

5 Conclusion

The rapid growth of unstructured and real-time financial data from disparate sources, including social media, economic APIs, and market sensors, has outpaced the capabilities of traditional data-processing systems in achieving timely and accurate frameworks for insight generation. Relying on historical models introduces excessive delay, rigid constraints, and poor adaptability scaling issues that become problematic in sensitive contexts such as predicting financial crises. This study overcomes these shortcomings with the innovative stacking-based ensemble architecture that integrates XGBoost, K-Nearest Neighbors (KNN), and Logistic Regression frameworks into a single mathematically rigorous structure. The impact of this XGBoost-KNN-Logistic model goes beyond just consolidation; the multi-level learner design giving room for local sensitization, strong generalization, and probabilistic reasoning loads complete pattern apprehension. There is little doubt this model closes a glaring void in the realm of real-time predictive analytics, considering the open API-sourced financial data streams appended with volatility indices, interest rates, and advanced sentiment analysis inputs. The performance metrics of the ensemble architecture indicates it is highly efficient with a classification accuracy of 96.17%, log loss of 0.0914, and an F1 score of 96.10%. These results verify that the stacking model is dramatically more effective than traditional single model methods by increasing bias, variance, and prediction error. The system's reliability, interpretability, and confidence provides the probabilistic output validation alongside confusion matrices, ROC curves, and residual distribution plots. Most critically, the model's performance on imbalanced class distribution problems highlights its ability to detect rare but crucial events, such as market crashes, actively improves its relevance in a financial context. This research thus validates the hypothesis that integrating heterogeneous learners in a "divide-and-conquer" strategy strengthens these complex adaptive systems' forecasting and decision making prowess. Looking forward, this study presents a number of prospects to work on. This can be further developed for predicting events in finance of multi-class nature, temporal sequence modeling or even incorporation of deep learning features such as LSTM to attend to time series dependencies. In addition, this can be transformed into a plug and play unit for smart decision engines with incorporation into finance based on blockchain technology and real-time trading for seamless integration. Other possible future improvements may include automated feature selection for evolutionary algorithms and real-time performance tuning through reinforcement learning. In other words, the study provides answerable, flexible, transparent

solutions to the challenges in real time predictive analytics in finance and enhances adaptive, reliable, and explicable AI frameworks for financial decisions.

References

1. Lone FA, Aggarwal S, Gupta S (2024) Green bonds, sustainable development, and economic benefits for investors and pharmaceutical companies. In: Nutrition controversies and advances in autoimmune disease. IGI Global, pp 654–672
2. Nieto PG, García-Gonzalo E, Antón JÁ, Suárez VG, Bayón RM, Martín FM (2018) A comparison of several machine learning techniques for the centerline segregation prediction in continuous cast steel slabs and evaluation of its performance. J Comput Appl Math 330:877–895
3. Zizaan A, Idri A (2024) Evaluating and comparing bagging and boosting of hybrid learning for breast cancer screening. Scientific African 23:e01989
4. Maity S, Hossain MA, Maji K, Mishra S, Nath S, Gupta S (2024) Analyzing and COMPARING random forest and k-nearest neighbours for effective heart disease prediction. In: 2023 4th International conference on intelligent technologies (CONIT). IEEE, pp 1–6
5. Mondal B, Gupta S (2022) Execution survey and state of the art of different ML-based ensemble classifiers approach contextual analysis of spam remark location. In: Proceedings of third international conference on computing, communications, and cyber-security: IC4S 2021. Springer Nature Singapore, Singapore. pp 311–323
6. Ghosh J, Gupta S (2023) ADAM optimizer and categorical crossentropy loss function-based CNN method for diagnosing colorectal cancer. In: 2023 International conference on computational intelligence and sustainable engineering solutions (CISES). IEEE, pp 470–474
7. Dua R, Edwards A, Anand U, Bansal P (2024) Are American electric vehicle owners quitting? Transp Res Part D: Transp Environ 133:104272
8. Ressi D, Romanello R, Piazza C, Rossi S (2024) AI-enhanced blockchain technology: a review of advancements and opportunities. J Netw Comput Appl 103858
9. Schmidt-Kessen MJ, Eenmaa H, Mitre M (2022) Machines that make and keep promises-Lessons for contract automation from algorithmic trading on financial markets. Comput Law Secur Rev 46:105717
10. Bountakas P, Xenakis C (2023) Helphed: Hybrid ensemble learning phishing email detection. J Netw Comput Appl 210:103545
11. Islam MM, Hassan S, Akter S, Jibon FA, Sahidullah M (2024) A comprehensive review of predictive analytics models for mental illness using machine learning algorithms. Healthc Anal 100350
12. Javaid A, Sajid M, Uddin E, Waqas A, Ayaz Y (2024) Sustainable urban energy solutions: Forecasting energy production for hybrid solar-wind systems. Energy Convers Manage 302:118120
13. Wettewa S, Hou L, Zhang G (2024) Graph Neural Networks for building and civil infrastructure operation and maintenance enhancement. Adv Eng Inform 62:102868
14. Boikanyo K, Zungeru AM, Sigweni B, Yahya A, Lebekwe C (2023) Remote patient monitoring systems: Applications, architecture, and challenges. Scientific African 20:e01638
15. Mukherjee T, Gupta S, Mitra A (2022) Impact of share market based on global happening and future prediction. In: 6th Smart cities symposium (SCS 2022), vol 2022. IET, pp 384–389
16. Satapathy SK, Bhoi AK, Loganathan D, Khandelwal B, Barsocchi P (2021) Machine learning with ensemble stacking model for automated sleep staging using dual-channel EEG signal. Biomed Signal Process Control 69:102898

Algorithm for Assessing Threats to Information Security



Kholimtaeva Ikbol, Sunbula Muminova, Barno Shamshieva, and Jakhongirbek Khabibullaev

1 Introduction

The growing demand for information and its volume are leading to an increasing demand for informatization objects. The main reason for this is that every sphere of society: communications, medicine, social spheres, military, education, trade, all sectors of the economy and other sectors are characterized by the organization of information systems. Each information system stores, processes and transmits its own unique and diverse information. The rapid development of information technologies and their widespread use in all areas, along with the creation of new opportunities, also creates its own problems. One of these problems is the problem of ensuring information security. In ensuring information security, it is necessary to analyze and assess the risksremains one of the important tasks.

The central concept of information security is the concept of "threat". According to dictionaries, the term "threat" means physical, material, or other harm to public or private interests, as well as a potential danger.

In modern literature, the authors of many publications define a threat to information security by the nature (type, method) of the destabilizing impact on information or by the consequences (outcomes) of such an impact in the form of damage caused by the violation of the rights of the subject.

K. Ikbol (✉) · S. Muminova · B. Shamshieva · J. Khabibullaev
Tashkent, Uzbekistan
e-mail: iqbola.ubaydullayevna@gmail.com

J. Khabibullaev
e-mail: j.xabibullayev@tuit.uz

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_28

2 Categories and Classificayion of Threats

The category of "harm" is valid only if it is proven that it was caused. This means that actions that lead to harm can be considered a crime from a legal point of view. Therefore, when identifying threats to information security, it is advisable to take into account the requirements of current criminal legislation, in particular, the Criminal Code of the Republic of Uzbekistan (1994). Information crimes are divided into the following categories [1, 2] (Table 1):

Summarizing the above, threats in general are understood as potential or actual risks (actions or inactions) directed against the object of protection, creating the risk of information corruption, loss, or illegal use, and causing harm to the owner (owner, user) of information resources.

Table 1 Categories of information crimes

Crime category	Description
Theft	Unlawful gratuitous seizure and (or) alteration of someone else's property in favor of the guilty party or other persons, committed for the purpose of venality, causing harm to the owner or possessor of the property;
Copy computer data	To reproduce and permanently print information on a machine or other medium;
Break	External impact on property, because of which it physically ceases to exist or becomes completely unfit for its purpose;
Computer data destruction	Delete data from computer memory;
To cause harm	Change in the characteristics of the property, which significantly deteriorates the condition of the property, loses a significant part of its useful properties and becomes completely or partially unsuitable for its intended use;
Change computer data	Making any changes related to the adaptation of a computer program or database;
Computer data blocking	Artificially preventing users from accessing information that is not related to its destruction;
Unauthorized deletion, blocking, modification, copying of information	Any specified actions with information not authorized by law, the owner or authorized user;
Deception (denying the truth or providing false information)	This is the deliberate misrepresentation or concealment of the truth with the aim of obtaining the voluntary taking of property from the person in charge of the property by misdirecting him or her or obtaining false information

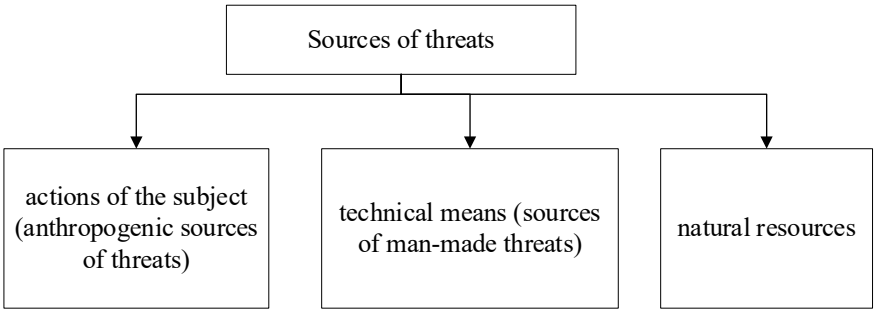


Fig. 1 Sources of threat

Threats do not arise on their own. All threats arise when there are vulnerabilities inherent in the information object.

A vulnerability is a weakness that can be exploited to compromise an automated information system or the data it contains.

In the process of information security assessment, it is important to pay special attention to threat sources. Threat sources can include both subjects (people) and objective factors. Threat sources can also be located inside the information system (internal sources) or outside it (external sources) [3–5].

The sources of threat are presented in Fig. 1.

Anthropogenic sources of threats include subjects whose actions can be classified as intentional or accidental crimes.

To man-made sources of threats includes sources determined by human technocratic activity and the development of civilization.

Threats of natural origin include natural disasters or other events that cannot be predicted in advance and cannot be prevented with the current knowledge and technological capabilities of humanity.

An analysis of information security threats shows that they can be divided into two types: internal and external (Fig. 2) [6].

When analyzing information security threats to an informatization facility, it is necessary to implement the following comprehensive approaches [7]:

- object description;
- classifying threat sources;
- vulnerability classification;
- classifying methods of implementing threats;
- current attack rating;
- classifying methods for eliminating threats.

A systematic description of the informatization object, presented in the form of typical components of the information system and the relationships between them, together with textual explanations, describes the directions and parameters of information flow, which allows you to identify possible points of entry and use of threats or identify existing vulnerabilities.

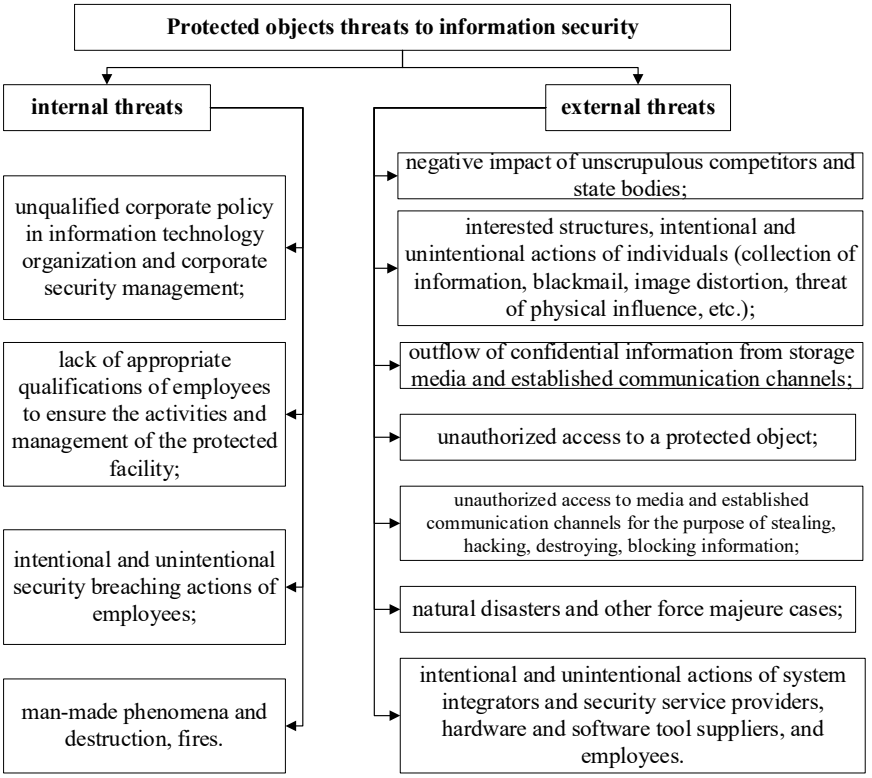


Fig. 2 Threats to the security of protected objects

Analysis and assessment of the possibilities of implementing threats should be based on the construction of a threat model, classification, analysis and assessment of threat sources, vulnerabilities and implementation methods. Modeling of information security violation processes can be carried out taking into account the logical chain: threat—threat source—implementation method—vulnerability—consequences.

Each component of the logical chain under consideration is described accordingly in the necessary detail.

Threats Information security objectives are classified according to the likelihood of causing damage if violated; sources of threats—by the type and location of the threat carrier; vulnerabilities—by their manifestation and affiliation with the source of vulnerabilities.

The classification of attacks is a set of possible ways to achieve goals using certain methods by the threat source, exploiting vulnerabilities. At the same time, the goal of the attack may not always coincide with the goal of implementing the threat. In some cases, it may be aimed at obtaining intermediate results necessary for the implementation of the threat at a later stage. If the goals of the attack do not correspond to the full implementation of the threat, then in this case the attack is

considered a preparatory process for actions aimed at the threat, that is, a preparatory stage for the implementation of illegal activities [7, 8].

Classification, ranking, analysis and identification of existing threats, sources of threats and vulnerabilities, and variants of possible attacks are identified, which allows assessing the state of information security and optimizing the choice of methods for combating threats (Fig. 3).

ISO 19011 is an international standard for planning, conducting and managing management system audits. It applies to all types of audits, including information security, quality management, environmental management and other systems [9–12].

The audit process according to the ISO 19011 standard is presented in Table 2.

3 Algorithm for Assessing Threats to Information Security in Organizations

The algorithm for assessing threats to information security in organizations is presented below.

Algorithm Algorithm for assessing threats to information security in organizations:

Step 1. Identify assets
Step 2. Vulnerability analysis
Step 3. Initial level of protection of valuable assets
Step 4. Damage assessment
Step 5. Probability of Threats Being Implemented
Step 6. Determine the probability of threats being implemented
Step 7. Determine the urgency of the threats
Step 8. Level of risk potential regarding information security
Step 9. Information security risk level assessment
Step 10. Recommend effective countermeasures
Step 11. Preparing reports

The description of the work and calculations performed in the steps of algorithm is as follows:

1. *Asset identification*

Active—a tangible or intangible thing that is of importance to an organization. In order to value assets, the organization's assets must first be identified.

2. *Vulnerability analysis*

When conducting a vulnerability analysis, the location of the information object, organizational and systematic protection issues, and the existing software and technical information protection tools are analyzed. After that, a list of possible vulnerabilities is formed and divided into 5 levels [13–15] (Table 3).

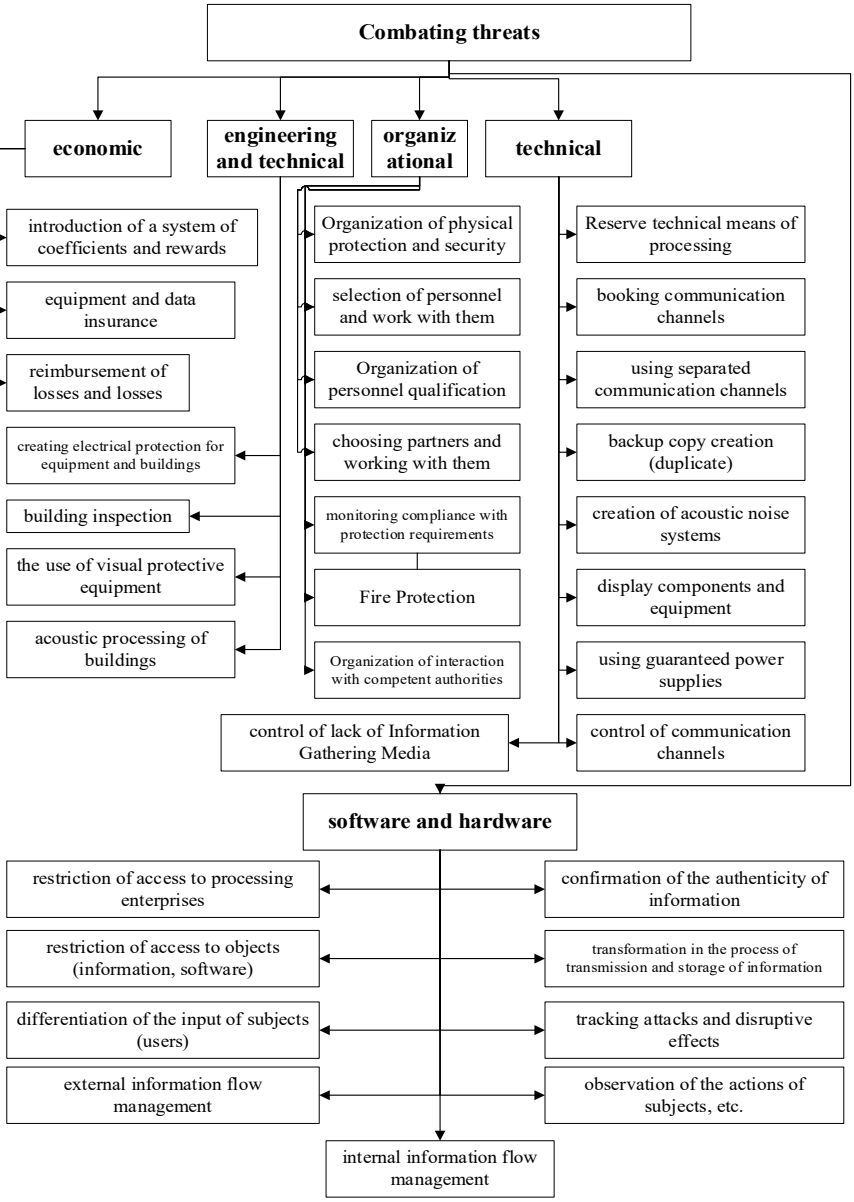


Fig. 3 Methods of combating threats

Table 2 Categories of information crimes

Stage	Main activities	Result
Audit planning	– Determining audit objectives – Determining the audit scope – Determining audit criteria – Appointing a team of auditors	An audit plan is prepared
Audit preparation	– Develop an audit program – Collect documents to be audited – Communicate with organization representatives in advance	Audit program and verification criteria
Conducting an audit	– Review documents – Interview employees – Conduct on-site observations – Document identified issues	Evidence is collected and examined
Analysis of audit results and preparation of reports	– Analysis of evidence obtained – Development of conclusions and recommendations – Preparation of audit report	Audit report and recommendations
Developing responses to audit results	– Develop an action plan to address deficiencies – Conduct additional audits (if necessary) – Improve the management system	Action plan and improvement measures

Table 3 Assessing the level of vulnerability in terms of quality and quantity

Weaknesses	
Quality rating	Price based on quantity
Critical level of vulnerability	10
The level of vulnerability	9
Medium level of weakness	6
The lowest level of weakness	3
The human level of weakness	1

3. *Initial level of protection of valuable assets*

This step assesses the initial level of protection of valuable assets based on the inventory of valuable assets and an analysis of information security in the organization (Table 4).

Table 4 An example of assessing the level of protection against the deterioration of the quality of an asset

Asset—software security	The level of asset protection in the organization		
	Up (b1)	Average (b2)	Low (b3)
Internet connection	+	−	−
Connection to a network intended for public use	−	−	+
Permission form	−	+	−
Permission to freely copy	−	+	−

The initial protection level K is determined as follows (Fig. 4) [16–18]:

- An information system is considered to be highly protected if 70 percent of the characteristics of the information system are at the "high" level and the remaining characteristics are protected at the "medium" level (K1 = 0);
- An information system is considered to be moderately protected if it is not protected at the "high" level and 70 percent of the characteristics of the information system are at the "medium" level, and the remaining characteristics are protected at the "low" level (K1 = 5);
- An information system is considered to be poorly protected if it is not protected at the "high" and "medium" levels (K1 = 10).

4. *Damage assessment*

When assessing information security, damage assessment can be carried out at different levels in monetary terms. Damage assessment in monetary termsEvaluation in a unit is a difficult and complex task.

The processed jackal is then evaluated from a qualitative to a quantitative perspective. A Z-factor is assigned to each level of damage (Table 5).

5. *Probability of threats being carried out*

Threat implementation level indicator "X"AH expert assessments the level of threat implementation is assessed based on the developed four-level qualitative scales: low, low, medium, and high (Table 6).

6. *Determining the coefficient of implementation of threats*

When calculating the coefficient for determining the possibility of implementing threats, the probability of implementing immediate threats (X) and the final assessment of the level of protection (K) are used. The coefficient of implementation of threats Y is determined by the following expression [17–19]:

$$Y = \frac{(X + K)}{2}$$

This is where:

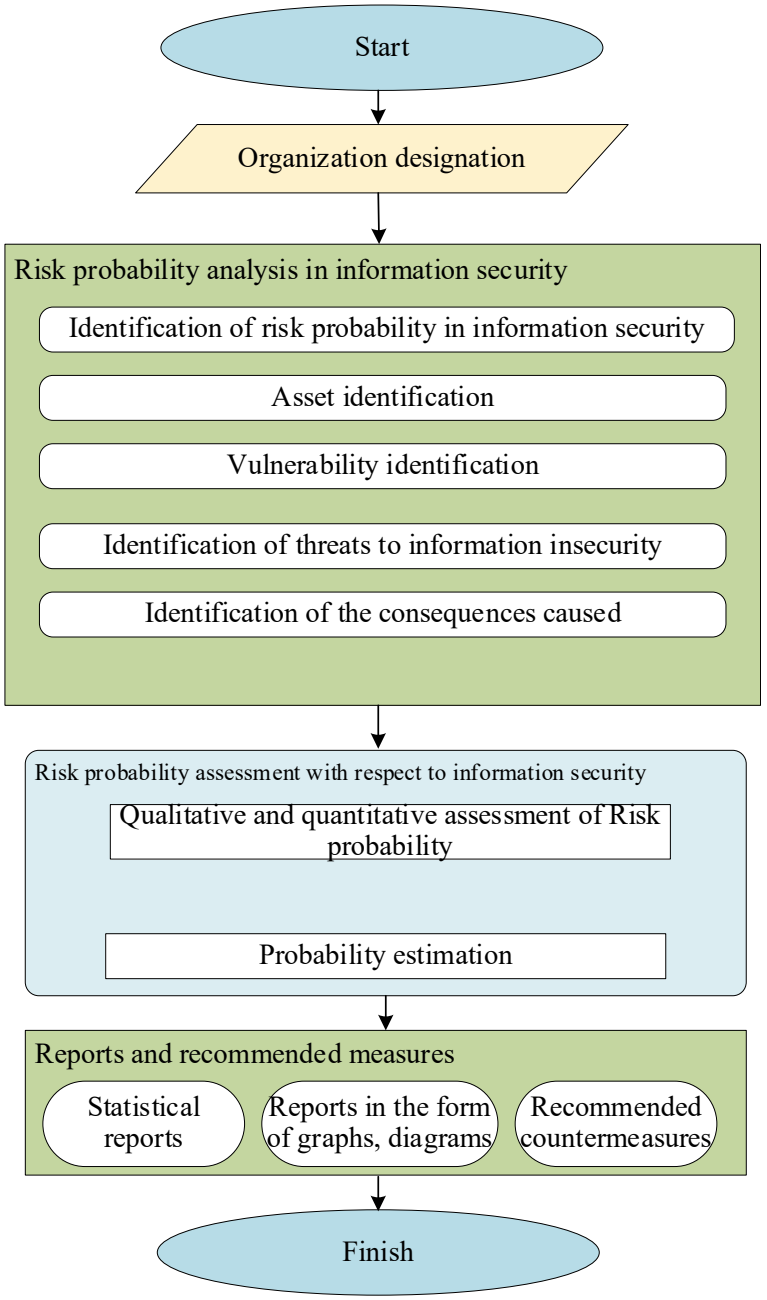


Fig. 4 Block diagram of the threat assessment algorithm for information security

Table 5 Assessment of the overall level of damage in terms of quality and quantity

The organization's overall damage level	Coefficient values
The amount of damage is very high	10
The amount of damage caused	7
Average damage amount	5
Low damage amount	3

Table 6 Threat probability of implementation degree, qualitative and quantitative dependencies

Probability of threats being carried out	Coefficient values
Up probability level	10
Average probability level	5
Well probability level	2
Few probability level	0

Y threat implementation coefficient, $0 \leq Y \leq 10$;
X the likelihood of implementing threats, $0 \leq X \leq 10$;
K initial hedged value of the asset, $0 \leq K \leq 10$;

We calculate the value of Y and obtain a quantitative numerical value. To make this value understandable, the evaluation is shifted from quantitative to qualitative.

if $0 \leq Y \leq 2.5$, then t contract enforcement coefficient low.
if $2.5 < Y \leq 6$, then t contract enforcement coefficient middle.
if $6 < Y \leq 7.5$, then t contract enforcement coefficient ear.
if $7.5 < Y \leq 10$, then t contract enforcement coefficient *very good*.

7. *Determining the urgency of threats*

Threat urgency classifies threats into urgent and non-urgent threats based on their likelihood of occurrence and their risk profile. Table 7 provides a definition of threat urgency [20, 21].

8. *Level of risk potential regarding information security*

At this stage, the level of risk probability of the AH is determined. Risk probability "R" is the probability of occurrence of a negative event associated with the realization

Table 7 Determining the urgency of threats

Possibilities for implementing threats	Threat risk indicators		
	Well	Average	Up
Well	Not relevant	Not relevant	Current
Average	Not relevant	Current	Current
Uqori	Current	Current	Current
Very good	Current	Current	Current

of a particular threat. It is possible. In the calculation process, a parameter must be taken into account and it is calculated according to formula below [22].

$$R = Z * \frac{Y * U}{10} \quad 0 \leq \mathrm{R} \leq 100,$$

This is where

U magnitude of vulnerability;
Y presence of threats;
Z damage to the organization.

The following scales are used for quantitative assessment:

if $0 \leq \mathrm{R} \leq 7$, the risk is "low";
if $8 < \mathrm{R} \leq 22$, the risk level is "medium";
if $23 < \mathrm{R} \leq 43$, the risk level is "high";
If $44 < \mathrm{R} \leq 100$, the risk level is "critical".

9. *Alignment of information security risk levels*

The total risk value is calculated using Formula below.

$$R_{um} = \frac{1}{n} * \sum_{i=1}^{n} R_i,$$

This is where:

R is the total value of the risk probability;
i the serial number of the risk being calculated;
n is the total number of risks involved in the calculation.

10. *Recommend effective countermeasures*

Effective response to threats based on the likelihood of identified threats A list of measures is formed.

11. *Preparing reports*

Reports are generated based on the results of the algorithm's calculations. The reports are in the form of statistics and graphical diagrams (Table 8).

Table 8 Total value of HR risks

Risk numbers	Specific risk values	Total value of risks
Risk no. 1	28	$R_{um} = \frac{1}{n} * \sum_{i=1}^{n} R_i = \frac{1}{3} * (28 + 15 + 35) = \frac{1}{3} * 78 = 26$
Risk no. 2	15	
Risk no. 3	35	

4 Conclusion

Comprehensive approaches should be used to analyze information security threats. These approaches include object description, threat source classification, vulnerability analysis, threat implementation method classification, current attack rating, and threat mitigation methods. These processes help identify all the elements necessary to ensure information security. Methods for assessing threats to information security were studied. Criteria for assessing risk probability were studied. In addition, factors considered as threats to information security in organizations and their organization were studied. In addition, scientific work carried out in this area, research and adopted standards, national standards and assessment criteria were studied. Methods for assessing threats to information security in organizations, operating modes of assessment software, algorithms, and mathematical formulas for calculations were studied.

References

1. Bogoras AG, Peskova OY (2013) Methodology for testing and assessment of firewalls. Izvestiya SFedU. Eng Sci 149(12):148–156 (in Russ.)
2. Dorofeev A (2010) Penetration testing: demonstration of one vulnerability or an objective security assessment? Zašita informacii. Inside 36(6):72–73 (in Russ.)
3. Umnitsyn MY (2018) Approach to semi-natural security evaluation of information system. Izvestia VSTU 218(8):112–116 (in Russ.)
4. Borodin MK, Borodina PJ (2017) VGATE R2 ınformation security penetration testing. Regional'naja informatika I informacionnaja bezopasnost [Regional ınformatics and information security]. Saint Petersburg, pp 264–268 (in Russ.)
5. Poltavtseva MA, Pechenkin AI (2017) Data mining methods in penetration tests decision support system. Inf Secur Problems Comput Syst 3:62–69 (in Russ.)
6. Kadan AM, Doronin AK (2016) Cloud infrastructure solutions for penetration testing. Uchenye zapiski ISGZ 14(1):296–302 (in Russ.)
7. Eremenko NN, Kokoulin AN (2016) Research of methods of penetration testing in information systems. Master's J 2:181–186 (in Russ.)
8. Tumanov SA (2015) Penetration testing tools for information systems. Proc Tomsk State Univ Control Syst Radioelectron 36(2):73–79 (in Russ.)
9. Kravchuk AV (2015) The model of process of remote security analysis of information systems and methods of improving it's performance. SPIIRAS Proc 38(1):75–93 (in Russ.)
10. Gorbatov VS, Meshcheryakov AA (2013) Comparative analysis of computer network security scanners. IT Secur 20(1):43–48 (in Russ.)
11. Pfleeger CP, Pfleeger SL, Theofanos MF (1989) A methodology for penetration testing. Comput Secur 8(7):613–620
12. McDermott JP (2000) Attack net penetration testing. NSPW, pp 15–21
13. Alisherov F, Sattarova F (2009) Methodology for penetration testing. Int J Grid Distrib Comput 43–50
14. Ami P, Hasan A (2012) Seven phrase penetration testing model. Int J Comput Appl 59(5):16–20
15. Holik F, Horalek J, Marik O, Neradova S, Zitta S (2014) Effective penetration testing with Metasploit framework and methodologies. In: 2014 IEEE 15th international symposium on computational intelligence and informatics (CINTI). IEEE, p 237–242. https://doi.org/10.1109/CINTI.2014.7028682

16. Fayzullajon B, Azam G, Sherzod S (2023) Handling ınformation security events and ıncidents. In: Ranganathan G, Fernando X, Rocha Á (eds) Inventive communication and computational technologies. Lecture notes in networks and systems, vol 383. Springer, Singapore. https://doi.org/10.1007/978-981-19-4960-9_40
17. Fayzullajon B, Sharifjon G, Sherzod S (2023) Methods for assessing ınformation security ıncidents in the enterprise and making decisions. In: Ranganathan G, Fernando X, Rocha Á (eds) Inventive communication and computational technologies. Lecture notes in networks and systems, vol 383. Springer, Singapore. https://doi.org/10.1007/978-981-19-4960-9_12
18. Zaynutdinova M, Sayfullaev S (2021) Analysis and modeling of ınformation security systems in industry 4.0. In: Aliev RA, Yusupbekov NR, Kacprzyk J, Pedrycz W, Sadikoglu FM (eds) 11th World conference "ıntelligent system for ındustrial automation" (WCIS-2020). WCIS 2020. Advances in intelligent systems and computing, vol 1323. Springer, Cham. https://doi.org/10.1007/978-3-030-68004-6_2
19. Abdukhalil G, Sherzod S (2020) Principles of adaptation in protecting corporate systems. Int Conf Inf Sci Commun Technol (ICISCT) 2020:1–4. https://doi.org/10.1109/ICISCT50599.2020.9351461
20. Ruzibaev O, Kabulov R, Sayfullaev S (2019) The axiomatic approach to definition of interval relations. Int Conf Inf Sci Commun Technol (ICISCT) 2019:1–2. https://doi.org/10.1109/ICISCT47635.2019.9011887
21. Majid K, Abduhalil G, Sherzod S (2020) Models and methods of corporate ınformation protection system. Int J Adv Trends Comput Sci Eng, pp 4521–4524. https://doi.org/10.30534/ijatcse/2020/48942020
22. Sayfullaev S, Valikulova D, Sabirov K (2024) Modern ıntelligent security analysis systems for corporate networks. In: 2024 5th International conference on ımage processing and capsule networks (ICIPCN), Dhulikhel, Nepal, pp 717–719. https://doi.org/10.1109/ICIPCN63822.2024.00124

A Smarter Plate: Enhancing Food Delivery Apps with Personalized Proportion Sizing and Nutritional Awareness



Ashwini Dalvi, Sargundeepkaur Sachdeo, and Samrudhi Bhat

1 Introduction

The proliferation of food delivery apps has transformed how people access meals, offering convenience and variety from home. Despite their popularity, these platforms are often criticized for overlooking food waste and lacking nutritional transparency [1, 2], with environmental consequences frequently ignored [3].

Customer satisfaction is crucial, driven by intuitive interfaces and smooth user experiences. Incorporating advanced technologies enhances engagement and functionality [4]. However, current literature primarily focuses on operational efficiency, neglecting the role of portion size customization and nutritional information in improving user experience and reducing food waste.

This study addresses that gap through a User-Centered Design framework for food ordering apps, integrating three core features: customizable portion sizes (e.g., quarter, half, full), real-time nutritional data, and a machine learning based recommendation system tailored to user behavior. This approach aims to promote satisfaction while encouraging sustainable and informed choices [5].

Behavioral research has explored food-related decision-making Verma [6] applied the Theory of Planned Behavior to sustainable food intentions, and Nguyen et al. [7] analyzed household waste behavior. Yet, these insights were not translated into app features.

A. Dalvi (✉) · S. Sachdeo · S. Bhat
Department of Information Technology, K.J. Somaiya College Of Engineering, Mumbai, India
e-mail: ashwinidalvi@somaiya.edu

S. Sachdeo
e-mail: sargundeepkaur.s@somaiya.edu

S. Bhat
e-mail: samrudhi.b@somaiya.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_29

To bridge this, this method proposes a behavior-driven recommendation system that adjusts portion sizes based on past consumption. We also introduce restaurant filtering grounded in user survey data, reflecting concern for food waste. Interface features such as dynamic sliders and bottom sheets are informed by findings like 57.6% of users preferring smaller portions.

This integration of sustainability, behavioral science, and health into a cohesive, user-first design represents a critical shift. It not only improves the digital food ordering experience but also supports responsible consumption, reducing waste and empowering users with better dietary control.

2 Literature Survey

The rapid growth of food delivery applications has led to various studies focusing on improving user experience, customer satisfaction, and incorporating new technological features. This literature survey reviews existing research, highlighting the emphasis on satisfaction and added functionality, while identifying a gap in addressing food portion sizes and nutritional value.

Prior work on food delivery applications can be broadly categorized into four main areas: User Satisfaction and UI Design, Technological Advancements, Sustainability and Waste Reduction, and Behavioral Modeling of food-related decisions. While these studies offer useful insights, they rarely adopt an integrated approach that combines personalized nutrition, portion control, sustainability, and user interface (UI) intervention [8]. This study aims to address these gaps using a mixed-method, real-time nutritional feedback, machine learning–guided, and user-centered design (UCD) approach.

Several studies have examined user satisfaction and UI/UX, including factors influencing engagement and retention. Jadhav et al. [4] emphasized variety and familiarity of restaurants but did not include health-conscious customization. Choi et al. [9], using the Technology Acceptance Model (TAM), found perceived usefulness and ease of use to be key factors but did not consider nutritional transparency or portion adaptability. Past research has widely analyzed food delivery apps, consumer behavior, and system use in different cultural and geographical areas. Sungboonlue et al. [10] presented redesigning food delivery apps in Thailand with a focus on improving the interface to provide increased user satisfaction. Lee et al. [11] used the Technology Acceptance Model (TAM) to measure consumer adoption, pointing toward the impact of perceived ease of use and usefulness. Wen et al. [12] gave a more extensive behavioral investigation of consumer intent with convenience, lifestyle compatibility, and incentives as major drivers. Chen et al. [13] conducted an examination of consumer attitudes towards food ordering platforms, highlighting how trustworthiness, system stability, and environmental considerations influence purchase intent. Kumar et al. [14] performed a real-world exploration of the 'Foodie' app, providing insights into competitive tendencies and consumer participation.

In a comparative regional comparison, Islam Arif et al. [1] assessed the usability of Bangladesh's food delivery apps, and they determined user satisfaction, interface effectiveness, and error rates as parameters of usability for localized application development. With an enlarged scope towards sustainability, Abdelradi [15] constructed a conceptual framework for analyzing food waste behavior at the household level. While not focused specifically on delivery apps, his research illuminates the ways in which consumer behaviors surrounding planning, buying, and handling food can lead to food waste. The results hold crucial implications for designing food delivery systems that encourage sustainable consumption and reduce waste.

While environmental concerns are well addressed in literature, actionable UI-based interventions to reduce waste are still lacking.

3 Methodology

To address the existing gap in the literature, this study evaluates user awareness and willingness to incorporate nutritional information and customizable portion sizes into food delivery applications. The methodology is structured into three key components: qualitative research, a proposed algorithm, and user interface (UI) design.

3.1 Qualitative Research

To assess user perspectives on the integration of nutritional value and portion size options in food delivery platforms such as Swiggy and Zomato, a comprehensive and well-structured survey was developed and administered [16]. The primary objective of the survey was to gauge user agreement with the inclusion of these features while simultaneously measuring their awareness of nutritional factors in everyday dietary choices. The survey targeted individuals between the ages of 18 and 25, yielding a total of 66 responses. This sample size is considered adequate to represent a diverse population with varied viewpoints on nutritional transparency and portion size customization in food delivery services.

A heatmap chart illustrating the distribution of survey responses is presented below. In the visualization, red denotes higher response percentages, while green indicates lower ones.

The survey findings of the able Table 1 reveal a substantial level of awareness and concern regarding nutritional information. Approximately 25.8% of respondents rated nutritional value as highly important, while 43.9% considered it somewhat important (Fig. 1). These results underscore a growing health consciousness within the targeted age group.

Survey responses also indicate that portion size is a key consideration for users, with 51.5% rating it as very important (Fig. 2). 37.9% of participants stated that

Table 1 Heatmap chart of the response received from the users in the survey

S. No.	Questions	Very	Somewhat	Neutral	Not very	Not
1	When ordering food, how important is the portion size to you?	51.5	36.4	12.1	0	0
2	How often do you feel that the portion sizes offered by food delivery apps are too large for your appetite?	3.3	10	37.9	45.5	3.3
3	How often do you use food delivery apps like Swiggy and Zomato?	4.5	16.7	37.9	36.4	4.5
4	Are you concerned about the food waste in general?	83.3	0	6	0	10.7
5	Would you prefer if food delivery apps offered the option to choose smaller portion sizes (e.g. 1/4,1/2,3/4) of dishes?	57.6	0	15.2	0	27.3
6	How much attention do you pay to the nutritional value of the food you order?	25.8	43.9	22.7	0	7.6
7	Would you appreciate having nutrition information displayed alongside portion size options when ordering food?	81.8	0	9.1	0	9.1

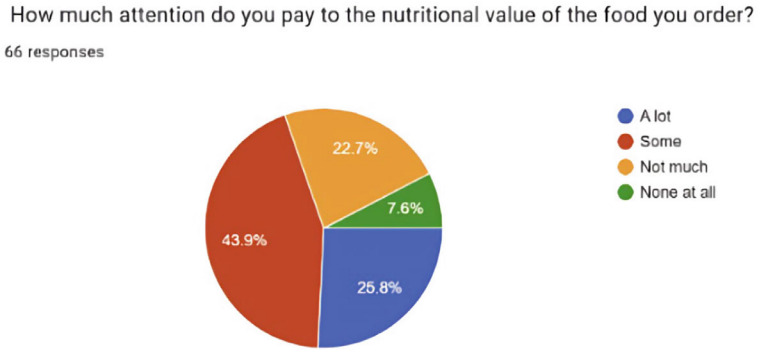


Fig. 1 Visual representation of attention given to nutritional value

portion sizes were sometimes too large for their appetite, 45.5% reported this occasionally, and 4.5% indicated this always or never occurred (Fig. 3). Additionally, usage frequency of food delivery apps varied among participants: 37.9% reported using these services once a week, 36.4% rarely, 16.7% several times per week, and a small fraction (4.5%) reported daily or no usage (Fig. 4).

Concerns about food waste were prominent, with 83.3% of respondents expressing concern about the issue (Fig. 5). These findings highlight a strong inclination among young adults toward sustainable and health-conscious eating behaviors. Furthermore, 57.6% of participants expressed a preference for the ability to select smaller portion sizes, such as one-fourth, one-half, or three-fourths of a dish, indicating demand for enhanced flexibility in portion control (Fig. 6). Notably, 28.8% reported they would likely opt for smaller portions if such options were available.

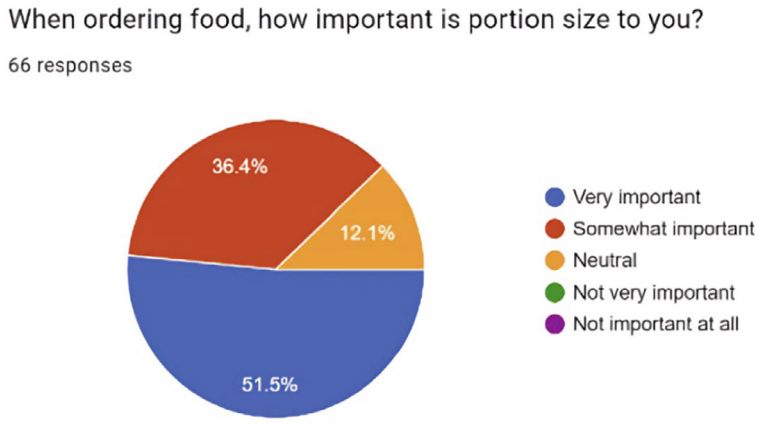


Fig. 2 Visual representation of attention given to proportion size

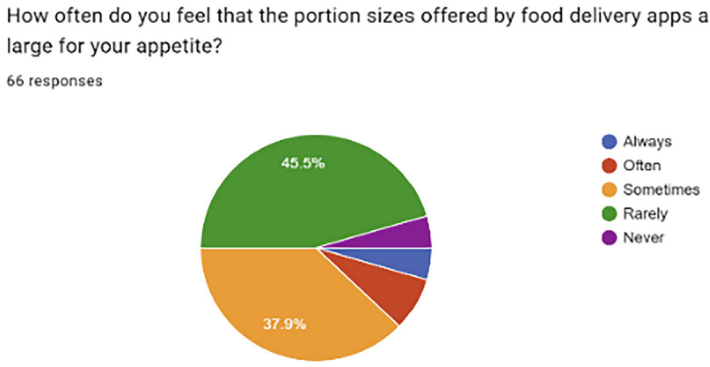


Fig. 3 Visual representation of statistics indicating how often food delivery portions exceed customer appetite

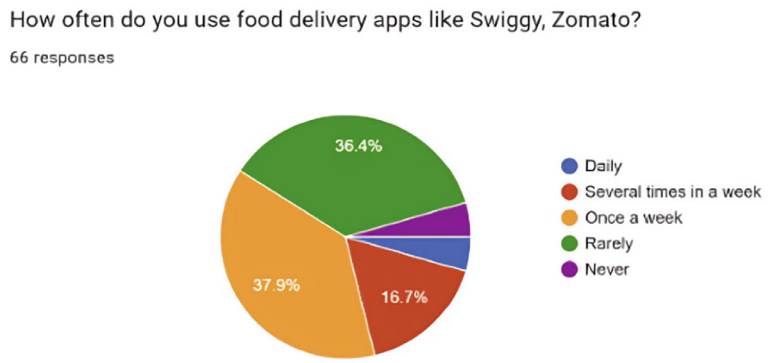


Fig. 4 Visual representation of statistics indicating how often user engagement with food delivery apps

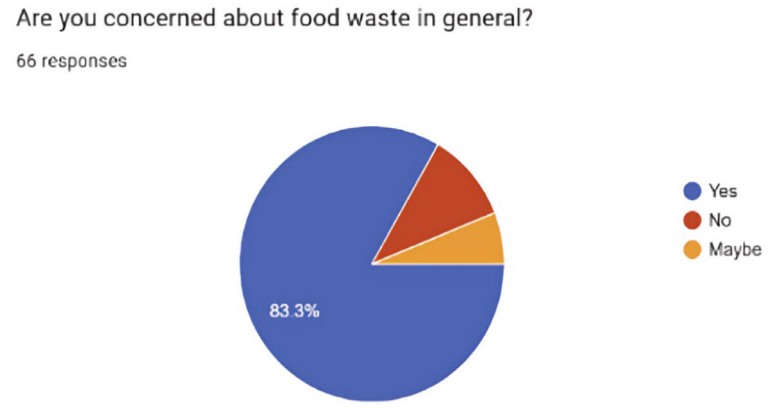


Fig. 5 Visual representation of concern Levels About Food Waste

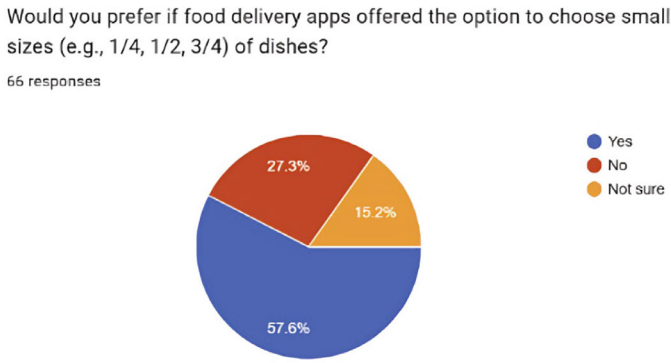


Fig. 6 Visual representation of preference for opting for smaller food portion size

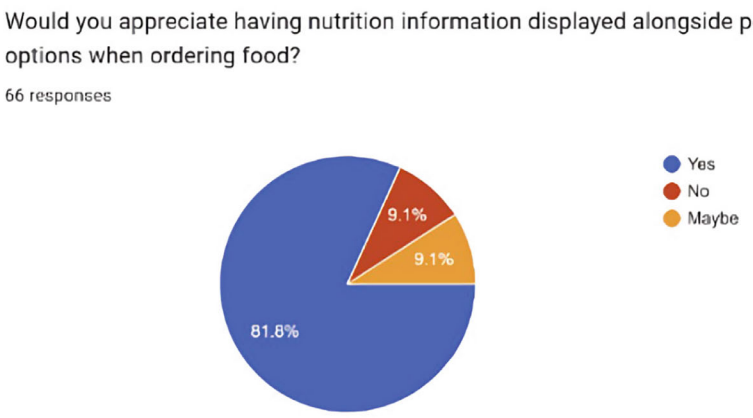


Fig. 7 Visual representation of consideration of nutrition information

The survey also showed that 25.8% of respondents consistently pay attention to the nutritional content of the food they order, 43.9% pay moderate attention, 22.7% pay little attention, and 7.6% do not consider it at all. These insights reinforce the necessity for nutritional transparency, with 81.8% of participants indicating support for displaying nutritional information alongside portion size options (Fig. 7).

3.2 Proposed Algorithm

The survey highlights a significant level of awareness and willingness among young adults to adopt more sustainable and health-conscious food choices. The integration of the proposed algorithm into food delivery applications is expected to offer substantial benefits to users. The algorithm designed for recommending portion sizes in food

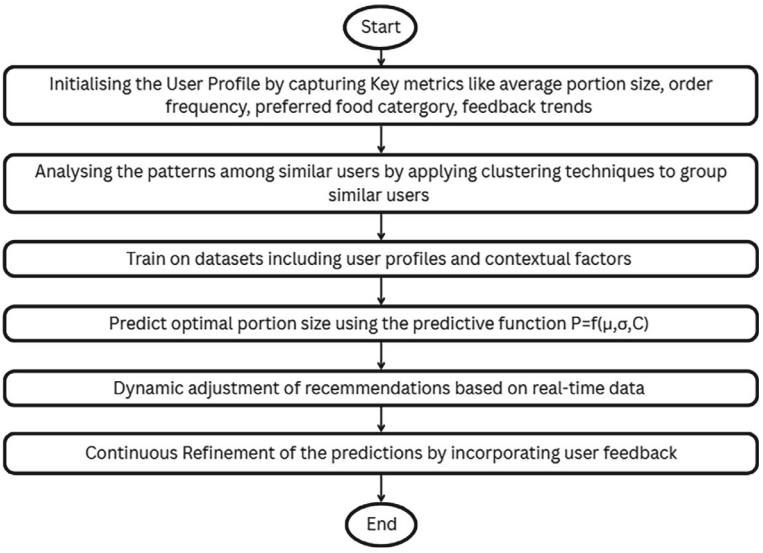


Fig. 8 Proposed algorithm for recommendation system

delivery platforms leverages an advanced combination of collaborative filtering and machine learning techniques (Fig. 8). Its objective is to provide personalized suggestions by analyzing users' historical behaviors. Central to this system is the development of comprehensive user profiles that incorporate key metrics, including average portion size, order frequency, preferred food categories, and feedback patterns.

The novelty of the proposed approach lies in its deliberate focus on portion size customization, an area often neglected in current food delivery recommendation systems. While conventional platforms primarily suggest food items or restaurants based on static user preferences, the proposed model adopts a dynamic, behavior-aware framework. By integrating nutritional data with individualized consumption trends, the system moves beyond surface-level personalization. It constructs robust user profiles using metrics such as historical portion size averages, variability in intake, food category preferences, order timing, and dietary patterns. This enriched profiling enables context-sensitive recommendations that better align with users' evolving needs and sustainability goals.

A key distinguishing feature of the proposed methodology is the application of clustering algorithms to identify and categorize users with similar behavioral patterns. These behavioural clusters are subsequently used to train the machine learning model, thereby enhancing predictive accuracy and enabling the algorithm to generalize insights across user cohorts. The core predictive engine is defined by the function $P = f(\mu,\sigma,C)$, where μ represents the mean historical portion size, σ denotes the variability in portion sizes, and C encapsulates relevant contextual parameters. This predictive model facilitates real-time personalization by adjusting recommendations in accordance with evolving user data.

The selection of this approach is supported by its capacity to overcome the limitations inherent in traditional, rule-based recommendation frameworks. The use of collaborative filtering allows for the identification of latent user preferences, while clustering enhances the scalability and generalizability of the system. Machine learning regression models are particularly well-suited to capturing the complex, multivariate dependencies that underlie portion size determination. Furthermore, the incorporation of an iterative feedback mechanism ensures that the system continuously refines its predictive capabilities based on user interactions, thereby maintaining a high degree of recommendation accuracy over time [17, 18].

In conclusion, the proposed algorithm represents a novel and methodologically sound advancement in the field of intelligent food delivery systems. By synthesizing behavioural analytics with adaptive learning, the approach addresses both personalization and sustainability objectives. Its capacity to dynamically tailor portion sizes not only enhances user experience but also contributes to broader public health and environmental goals, underscoring its relevance and potential impact within contemporary digital food service ecosystems [19, 20].

3.3 UI Screens

To effectively implement the proposed features, targeted user interface (UI) enhancements and an integrated recommendation system are essential. Using Swiggy as a reference case, the suggested modifications aim to be minimally disruptive yet significantly impactful enhancing user experience while fostering informed, health-conscious, and sustainable decision-making (Fig. 9).

A central feature of the proposed design is the introduction of a filter option that allows users to view only those restaurants offering customizable portion sizes (Fig. 10). This filter will be embedded within the primary search bar and advanced filtering settings, helping users tailor their food choices to align with personal dietary preferences and sustainability concerns (Fig. 11) [21].

Currently, users interact with a simple “Add” button when selecting food items. Under the proposed design, tapping this button will launch an interactive bottom sheet (modal) instead. This redesigned component will allow users to select from a range of portion sizes (e.g., 1/4, 1/2, 3/4, full, 1.25, or other fractional options) and simultaneously view nutritional details—such as calorie count, protein, fat, and carbohydrate content updated in real time (Fig. 12a, b). This real-time transparency empowers users to make better-informed food choices in alignment with their health and consumption goals.

The bottom sheet will feature a streamlined, intuitive interface that promotes ease of use and quick decision-making. By combining portion customization with nutritional transparency, this functionality addresses both overconsumption and food waste.

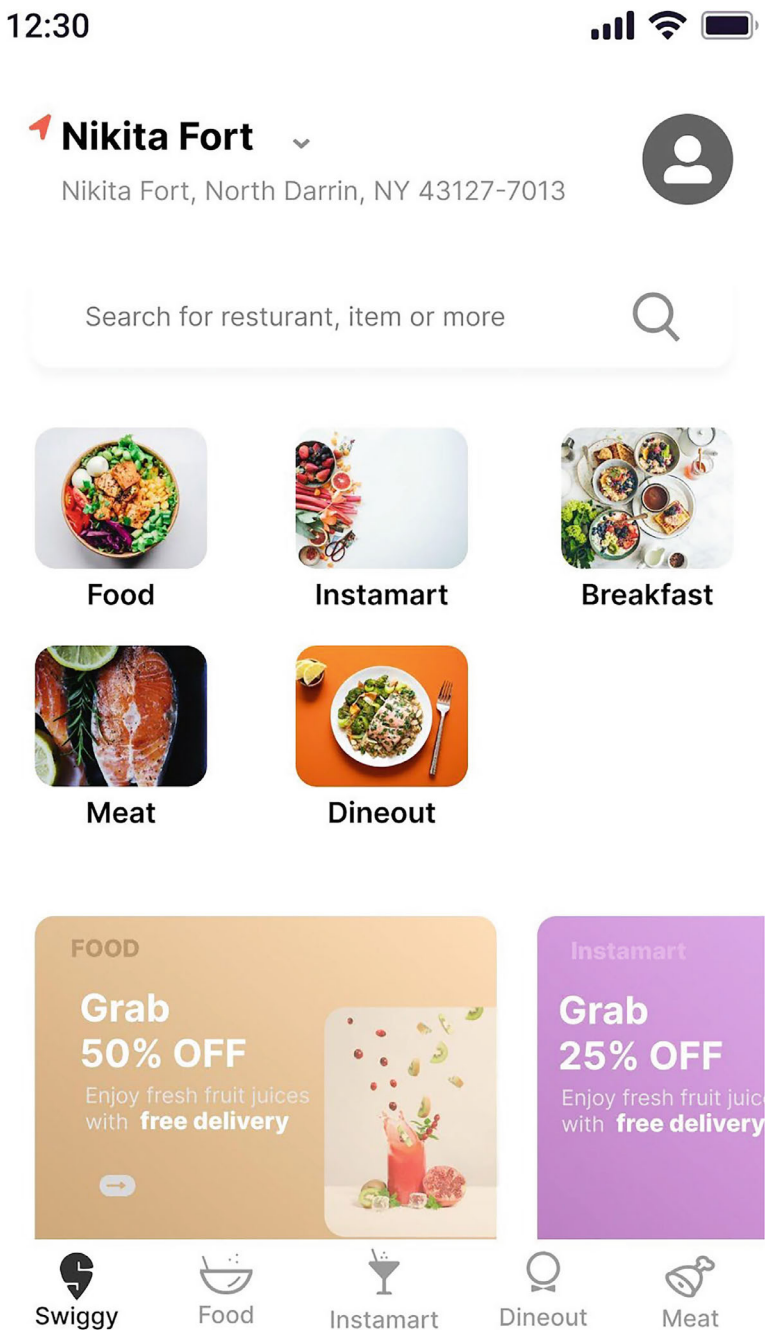


Fig. 9 Swiggy home screen

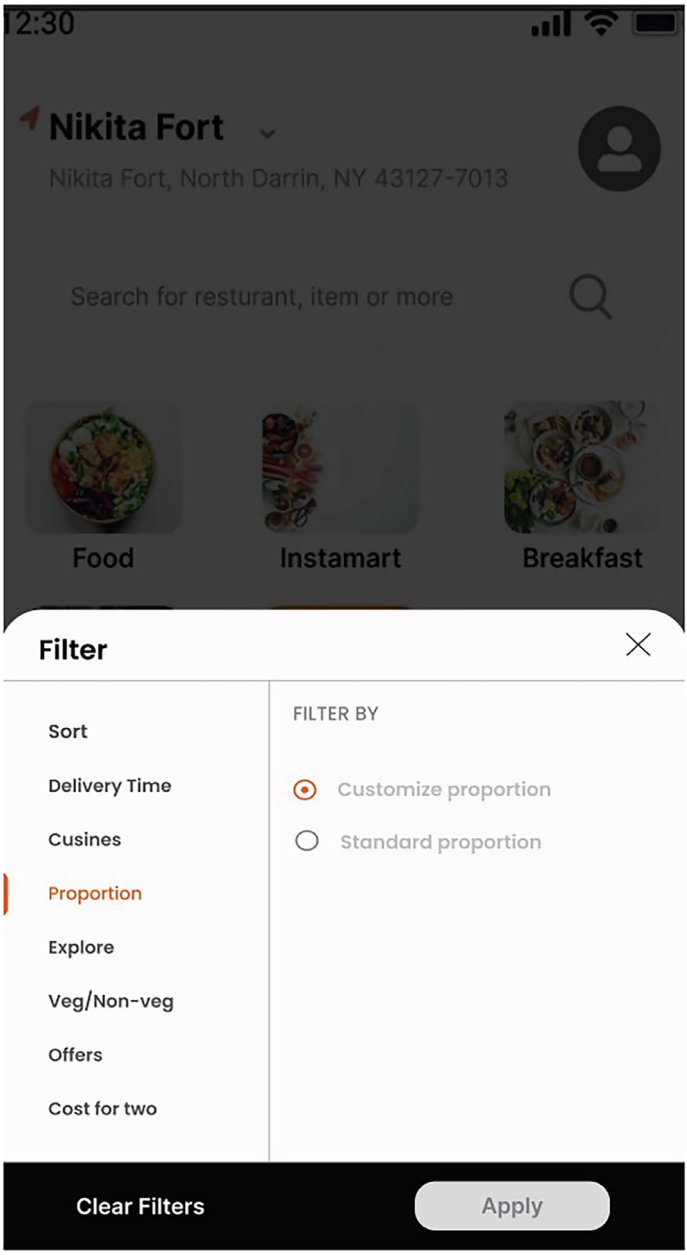


Fig. 10 Filter for restaurants that offer customizable portion sizes

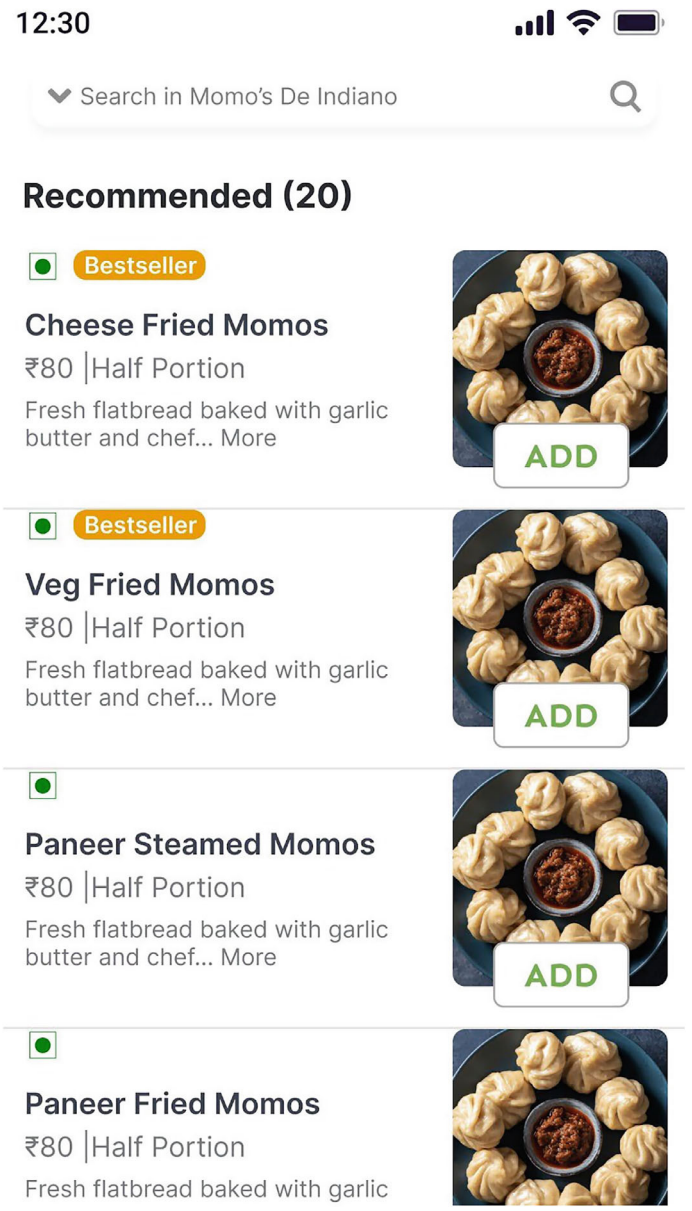
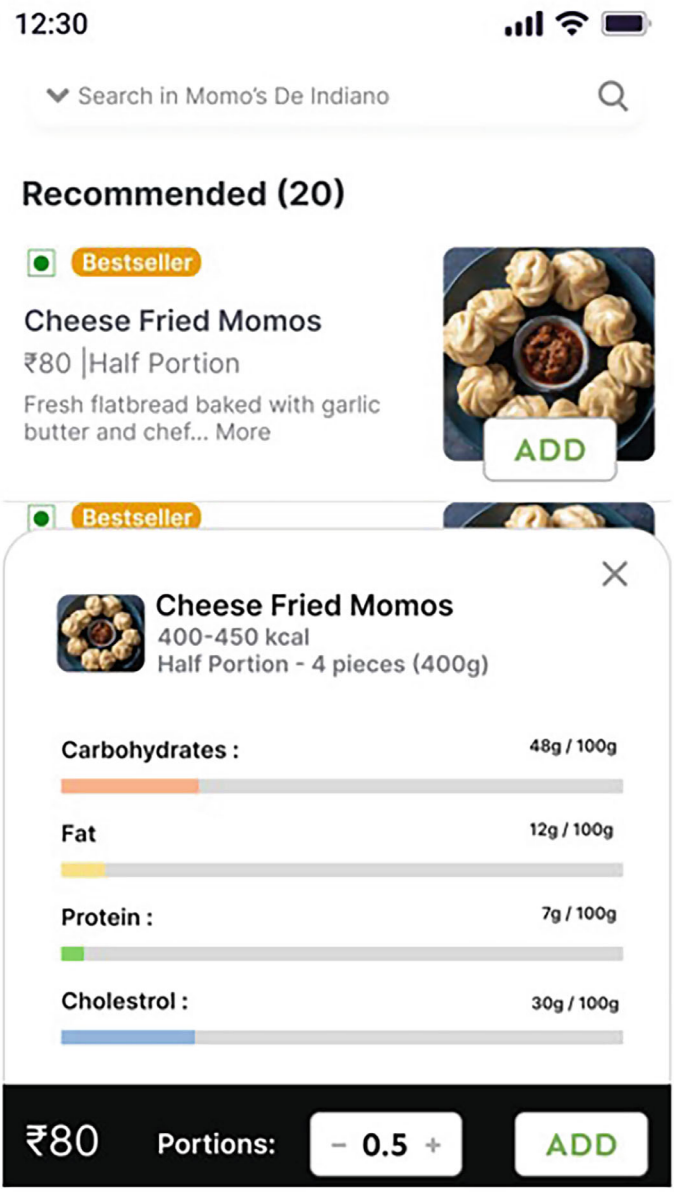
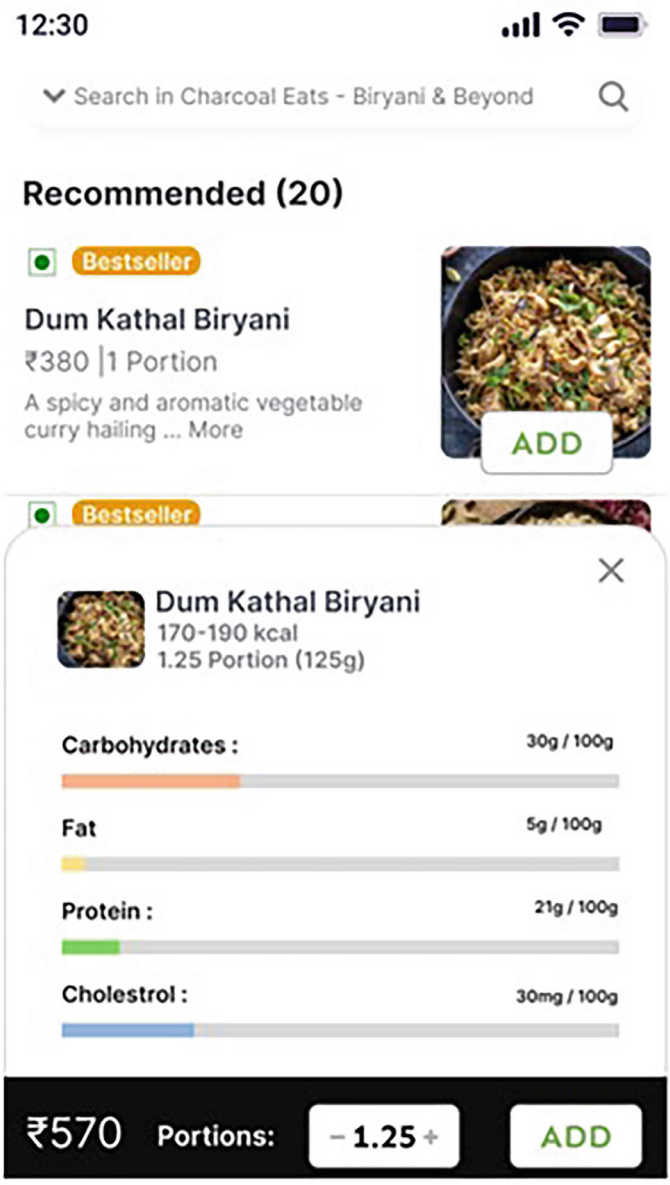


Fig. 11 Food menu of a restaurant



(a) UI modifications with inclusion of portion size and nutrition value for selected food item (Momo's De Indiano)

Fig. 12 UI modifications for different restaurants **a** UI modifications with inclusion of portion size and nutrition value for selected food item (Momo's De Indiano), **b** UI modifications with inclusion of portion size and nutrition value for selected food item (Charcoal Eats)



(b) UI modifications with inclusion of portion size and nutrition value for selected food item (Charcoal Eats)

Fig. 12 (continued)

These UI enhancements enable platforms like Swiggy to better serve health-conscious users and those committed to sustainable consumption. Beyond user satisfaction, the changes also support larger global objectives of reducing food waste and encouraging healthier eating habits. The integration of customizable portion sizes and nutritional transparency into food delivery applications presents a meaningful step toward addressing pressing concerns related to public health and environmental sustainability. By aligning app features with the values and behavior of today's users particularly young adults this design promotes a more responsible, user-centered, and sustainable food delivery ecosystem. This research underscores the importance of embedding such innovations into platform design, highlighting their dual role in enhancing user experience and contributing to broader sustainability goals.

4 Conclusion

This paper addresses a critical gap in the current food delivery ecosystem by proposing a novel design approach grounded in the principles of User-Centered Design (UCD). Centering on the needs and preferences of users specifically young adults aged 18–25. The research presents a solution that integrates three key features: customizable portion sizes (e.g., 1/4, 1/2, full), real-time nutritional transparency, and a personalized recommendation system powered by machine learning. The findings reveal that users highly value nutritional information and seek greater control over portion sizes to align with their dietary goals and reduce food waste. By embedding these insights into the design process, the proposed model not only enhances user satisfaction but also promotes responsible consumption bhaviors. The use of an intuitive, interactive bottom sheet for selecting portion sizes and viewing nutritional content demonstrates how minimal yet impactful UI changes can support healthier and more sustainable decision-making. Furthermore, the integration of a collaborative filtering–based recommendation system refines user experience by offering context-aware portion suggestions based on historical behaviors and preferences. This adaptive mechanism exemplifies how technology can evolve through continuous feedback loops, further personalizing the experience and reinforcing long-term user engagement. Overall, this design promotes a balanced ecosystem where user health, convenience, and sustainability coexist advancing both individual well-being and broader goals such as SDG 12: Responsible Consumption and Production.

References

1. Islam Arif MA, Kabir S, Islam Z, Hussain Khan MF, Islam I, Islam MN (2024) Usability evaluation of online food delivery applications in Bangladesh: a comparative study. In: 2024 IEEE international women in engineering (WIE) conference on electrical and computer engineering

(WIECON-ECE), Chennai, India, pp 337–342. https://doi.org/10.1109/WIECON-ECE64149.2024.10914965.
2. Li C, Bremer P, Jowett T, Lee M, Parker K, Gaugler E, Mirosa M (2024) What influences consumer food waste behaviour when ordering food online? An application of the extended theory of planned behaviour. Cogent Food Agric 10. https://doi.org/10.1080/23311932.2024.2330728
3. Islam S, Jones R, Akhi A, Talukder S (2023) Do food delivery apps influence food waste generation? A stimulus-organism-behavior-consequence (SOBC) exploration. Br Food J 126. https://doi.org/10.1108/BFJ-07-2023-0648
4. Jadhav S, Titus R, Babu T, Chinnaiyan R (2023) Evaluation of consumer behavior regarding food delivery applications in India. arXiv preprint arXiv:2401.14409
5. Seetharam A, Kumar T, Gupta H, Karia D (2023) A web-based AR solution for restaurant menus to reduce food waste, pp 1–9. https://doi.org/10.1109/GCITC60406.2023.10426326
6. Mohammed Abrar K, Verma A, Examining ındian organic food purchase behavior
7. Nguyen AT, Nguyen N, Phung P, Yến-Khanh N (2023) Residents' waste management practices in a developing country: a social practice theory analysis. Environ Chall 13:100770
8. Fauziah Y, Melisa K, Jayadianti H (2024) User-centred design method approach for user experience analysis of food delivery application. In: 2023 1st international conference on advanced informatics and intelligent information systems (ICAI3S 2023). Atlantis Press, pp 261–273
9. Choi J-C (2020) User familiarity and satisfaction with food delivery mobile apps. SAGE Open 10(4):2158244020970563
10. Sungboonlue P, Thanakaew S, Rangseepanya K, Tangpatong T, Siriborvornratanakul T (2022) A study of redesigning food delivery application in Thailand. Telkomnika (Telecommun Comput Electron Control) 20(5):1073–1082
11. Lee WS, Song M, Moon J, Tang R (2023) Application of the technology acceptance model to food delivery apps. Br Food J 125(1):49–64
12. Wen H, Pookulangara S, Josiam BM (2022) A comprehensive examination of consumers' intentions to use food delivery apps. Br Food J 124(5):1737–1754
13. Chen HS, Liang CH, Liao SY, Kuo HY (2020) Consumer attitudes and purchase intentions toward food delivery platform services. Sustainability 12(2):10177
14. Kumar H, Jain M, Bajwa M (2021) Online food delivery app 'foodie'. J Univ Shanghai Sci Technol 23:761–771. https://doi.org/10.51201/JUSST/21/07200
15. Abdelradi F (2018) Food waste behaviour at the household level: a conceptual framework. Waste Manag 71:485–493, ISSN 0956-053X. https://doi.org/10.1016/j.wasman.2017.10.001
16. Yang C, Yan L, Ji P (2024) Restaurant waiting staff's intention to dissuade customers from over-ordering: an extended theory of planned behaviour. Br Food J 126. https://doi.org/10.1108/BFJ-01-2024-0114
17. Wei L, Prabhakar G, Duong L (2023) Usage of online food delivery in food waste generation in China during the crisis of COVID-19. Int J Food Sci Technol 58. https://doi.org/10.1111/ijfs.16552
18. Alnasser A, Abaalkhail A (2024) Digital food behaviours, motivations, and delivery application usage among Saudis during COVID-19: a mixed-methods study. Heliyon 10(3):e24903. Elsevier. https://doi.org/10.1016/j.heliyon.2024.e24903
19. Sharma R, Dhir A, Talwar S, Kaur P (2021) Over-ordering and food waste: the use of food delivery apps during a pandemic. Int J Hosp Manag 96, 102977, ISSN 0278-4319
20. Li C, Mirosa M, Bremer P (2020) Review of online food delivery platforms and their impacts on sustainability. Sustainability 12(14):5528
21. Kristia K, Kovács S, László E (2023) Food delivery platform and food waste: deciphering the role of promotions, knowledge, and subjective norms among Indonesian generation Z. Clean Responsible Consum 11:100152, ISSN 2666-7843. https://doi.org/10.1016/j.clrc.2023.100152

HoloChef: 3D Culinary Guide with Speech Command



Fazliaty Edora Fadzli, Amira Azkia Rahmatulloh, Ajune Wanis Ismail, Muhammad Anwar Ahmad, Tee Ng Zikang, and Mohamad Yahya Fekri Aladin

1 Introduction

Cooking is an essential life skill, yet many individuals find it challenging due to various factors such as lack of confidence, unfamiliarity with recipes, or the inconvenience of traditional instructional methods. Traditional cookbooks and classes often require physical presence and can present information in formats that are not easily accessible or user-friendly. To address these challenges, researchers have explored the integration of technology into the culinary domain [1].

Recent advancements in technology have brought forth the development of augmented reality (AR) [2–4] and mixed reality (MR) [5–8] systems to assist in the kitchen. These systems provide step-by-step instructions, visual overlays, and real-time feedback, enhancing the cooking process by offering more intuitive and accessible guidance [9]. However, while these systems are useful, many still require visual

F. E. Fadzli (✉) · A. A. Rahmatulloh · A. W. Ismail · M. A. Ahmad · T. N. Zikang
Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia
e-mail: fazliaty.edora@utm.my

A. A. Rahmatulloh
e-mail: amirarahmatulloh@graduate.utm.my

A. W. Ismail
e-mail: ajune@utm.my

M. A. Ahmad
e-mail: muhammad.anwar@utm.my

T. N. Zikang
e-mail: teezikang@graduate.utm.my

F. E. Fadzli · A. W. Ismail · M. A. Ahmad · T. N. Zikang · M. Y. F. Aladin
Mixed and Virtual Reality Research Lab, ViCubeLab, Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia
e-mail: yahyafekri@gmail.com

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_30

attention and physical interaction, making them less efficient for a truly hands-free experience.

This paper introduces HoloChef, a 3D culinary guide with the aims of providing a solution that presents cooking instructions, ingredient lists, and techniques as interactive 3D holograms within the user's environment. Users can control the guide entirely through voice commands, ensuring a seamless, hands-free interaction that allows them to focus on cooking rather than managing devices or reading instructions. The main contribution lies in enabling an immersive, hands-free cooking experience by combining 3D holographic visuals of cooking steps with speech recognition. This paper will discuss the related work in the area of interactive culinary systems, outline the methodology used to develop HoloChef, and present the results of its implementation. The paper will conclude with a summary and suggestions for future work in this field.

2 Related Work

Recent developments in the culinary domain have incorporated immersive technologies to improve user engagement, accessibility, and instruction. Augmented reality (AR) and mixed reality (MR) systems, in particular, have been explored for their potential to deliver interactive cooking guidance.

The SPICE system [10] employs smart projection interfaces to display recipe instructions directly onto kitchen surfaces, promoting intuitive interactions by reducing reliance on printed materials or mobile devices. While SPICE enhances spatial awareness during cooking, it is constrained by the limitations of 2D projections and lacks dynamic 3D visualization.

HoloCook [11] offers real-time mixed reality cooking assistance by enabling remote tutoring through holographic displays. This system supports step-by-step instruction but requires head-mounted displays, making it less practical for everyday users seeking lightweight, unobtrusive solutions.

Voice-driven cooking assistance is another active research area. The Cooking with Conversation project [12] leverages natural language processing to facilitate conversation-based guidance, focusing on increasing user engagement and knowledge retention. Although effective in promoting dialogue, the system is limited by its absence of visual demonstrations and does not integrate spatially aligned 3D content.

Collaborative and therapeutic approaches have also been explored. The Happje project [13] centers on group-based cooking for individuals with dementia, aiming to foster cognitive stimulation and emotional well-being through shared experiences. However, its focus lies in social engagement rather than individualized visual instruction. A recent and notable contribution is the OSCAR system, which supports non-visual cooking by providing object status and contextual awareness through vision-language models and large language models. It aligns visual inputs with cooking instructions to guide visually impaired users through recipe steps with contextual

feedback [14]. OSCAR highlights the importance of accessibility and demonstrates how AI can adapt instructions dynamically based on real-world object conditions.

While these systems have advanced the field, several limitations remain. Many rely on visual attention through screen-based interfaces or wearable devices, which can detract from the hands-free cooking experience. Others focus solely on either visual or auditory instruction, without offering both in a cohesive, immersive format.

In contrast, the feature of HoloChef lies in its combination of three core elements that are rarely integrated in a single cooking assistance system:

- Holographic 3D Presentation: HoloChef delivers step-by-step instructions using a holographic avatar, along with 3D visualizations of ingredients, utensils, and techniques, providing a more immersive and spatial understanding of cooking procedures.
- Fully Voice-Driven Interaction: Unlike many AR/VR systems that still depend on manual or touchscreen input, HoloChef allows users to control the experience entirely through natural language voice commands, enabling true hands-free operation.
- No Wearable Requirements: The system is designed to work without the need for head-mounted displays or AR glasses. This makes HoloChef more practical and accessible for everyday users in a kitchen setting, where hygiene and ease of movement are critical.

By integrating these features, HoloChef not only advances usability but also bridges a critical gap between instructional fidelity and hands-free, intuitive interaction. This holistic approach makes it particularly suitable for users who seek immersive assistance without complex setup or hardware dependency.

The selection of a holographic, voice-controlled system for cooking guidance is driven by practical challenges observed in existing solutions. Many AR and MR systems require users to divert their attention to handheld devices or wear headsets, which can interrupt the cooking flow and pose hygiene concerns. Meanwhile, voice-only assistants lack visual representation, limiting their ability to demonstrate techniques or provide spatial cues.

HoloChef is designed to overcome these limitations by combining holographic 3D visuals with voice-based interaction. This approach enables users to receive real-time, spatially accurate cooking instructions while keeping their hands and attention focused on food preparation. It also supports intuitive, natural communication with the system, reducing the cognitive load often associated with screen-based tutorials. The integration of these elements was selected to improve accessibility, engagement, and learning effectiveness, particularly for novice users who may feel overwhelmed by traditional or text-heavy methods.

3 Methodology

The methodology section outlines the three-phase approach for developing the HoloChef system as shown in Fig. 1. Phase 1 involves analyzing the current cooking media and identifying the challenges and desired features for the holographic cooking assistant. In Phase 2, the system is developed, focusing on creating 3D culinary guide data and implementing speech recognition to allow hands-free interaction. Phase 3 proposed to evaluate the system through usability testing, where participants follow holographic guidance to prepare dishes, with data collected on user performances. This paper will cover up to Phase 2 and will continue to work on Phase 3 as future work.

The study begins with Phase 1, where the challenges associated with current cooking media are analyzed. This phase focuses on understanding the interactions with traditional cooking resources, such as written recipes and video tutorials. To gain insights, a pre-development questionnaire was distributed to individuals between the ages of 18 and their mid-30s. The questionnaire aimed to assess their comfort level of cooking without instructions, the difficulties they face when following cooking tutorials, and their thoughts on using a holographic chef assistant. Additionally, participants were asked about the specific functionalities they would like to see in such a system. The data collected helped guide the subsequent exploration of technologies to be incorporated into the HoloChef system.

In Phase 2, the system's development took center stage, with a primary focus on two key technologies: holography and speech recognition. The development process involved the creation and rigging of 3D assets, including a virtual chef, ingredients, and tools for cooking specific dishes, such as Ayam Woku and Es Pisang Ijo. This phase also included modeling interactions to demonstrate gestures that the holographic chef could perform, as well as setting up the holographic display for a realistic and immersive experience. For the speech recognition component, the goal was to identify command words that the hologram could understand and respond to, in line with voice commands. This allowed for a fully interactive cooking experience, where users could control the hands-free system and receive real-time guidance while preparing their chosen dish.

Phase 3 is proposed as future work for the evaluation of the 3D Culinary Guide with Real-Time Voice Command and Holographic Display. This phase will focus on conducting usability testing, involving pre- and post-cooking questionnaires to gather user feedback.

4 Implementation and Results

The implementation of the HoloChef 3D Culinary Guide involved integrating various technologies, including 3D modeling, speech recognition, and holographic display to provide an immersive and hands-free cooking experience. The 3D models of the chef,

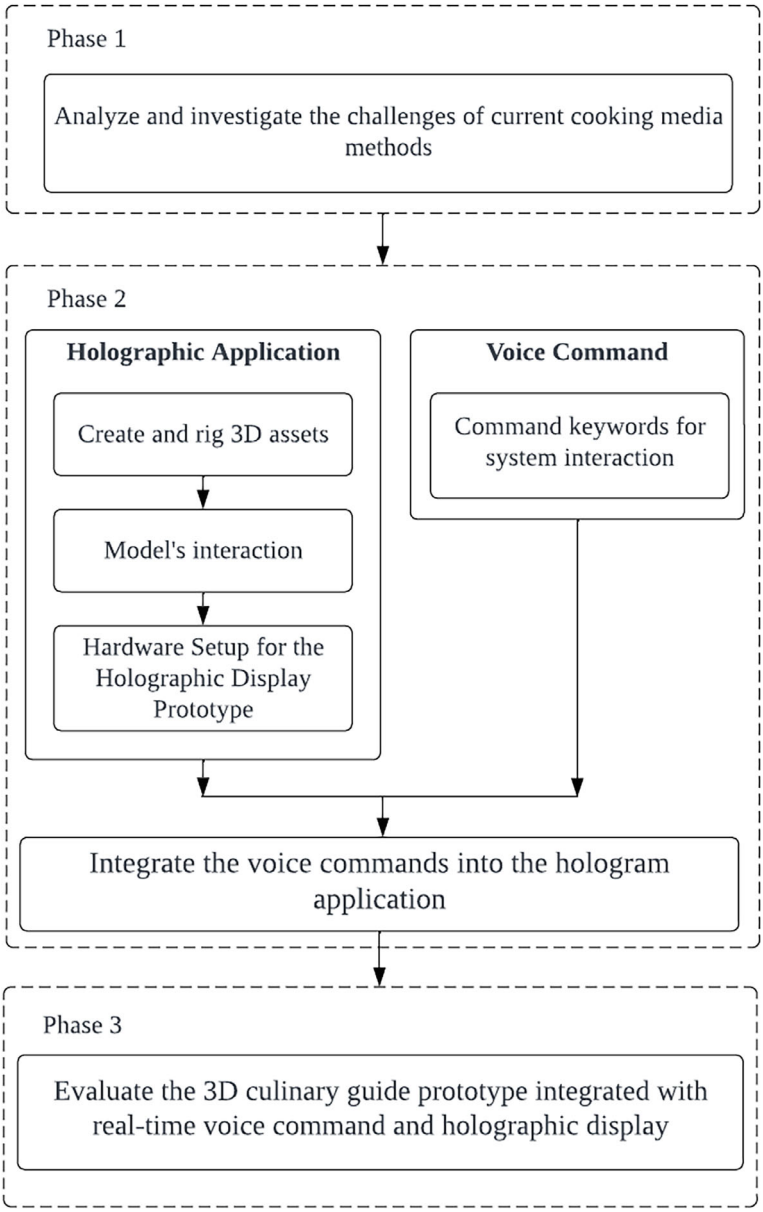


Fig. 1 Methodology diagram of HoloChef

ingredients, and utensils were created using Blender and VRoid Studio, with animations tracked and refined using Rokoko for lifelike movement. The speech recognition system, integrated via Unity's Speech-to-Text, allows users to interact with the system through voice commands, such as selecting the chef's gender and choosing recipes. The holographic display was designed to present the cooking process, guiding users step-by-step in preparing Ayam Woku and Es Pisang Ijo. The overall user interface was developed to be intuitive, with easy navigation and clear verbal feedback during the cooking process, enhancing both the engagement and usability of the system.

4.1 3D Culinary Guide

The incorporation of 3D features in the HoloChef Culinary Guide is essential for delivering an immersive cooking experience. This section focuses on three main components: the 3D model of the chef, the representation of ingredients, and the depiction of utensils used. Blender was selected as the primary platform for creating these assets, with various plugins and resources integrated to produce a visually appealing and functional prototype. Additionally, Table 1 outlines the 3D models used in the Ayam Woku dish, including the ingredients for each model. Table 2 presents the 3D models used as ingredients in the Es Pisang Ijo dish.

The 3D models of the chef are available in two versions, male and female, each designed with an animated, stylized aesthetic rather than a hyper-realistic approach. These models were created using VRoid Studio, which allows for extensive character customization. Figure 2 illustrates how the male chef character was developed in VRoid Studio. This customization feature gives flexibility to users to adjust the appearance of the chef as desired, allowing for personalization in the cooking experience.

Rigging the chef's 3D model was a crucial step to ensure that the movements during cooking are clear and recognizable for users. Modern technologies, such as Rokoko, were utilized to track movement from video recordings and refine it within Blender to ensure fluid and lifelike animations. This process allows the chef's movements to be accurately mirrored, ensuring that users receive precise instructions. The animated movements include actions such as slicing, threading, flipping, mixing, sautéing, and plating, which were inspired by common cooking tutorials found on websites and videos.

The rigging process starts with capturing video footage of a specific cooking movement, which is then uploaded to Rokoko for tracking. This initial tracking captures the movements and applies them to the 3D model. Figure 3 showcases the Rokoko Vision video output, which demonstrates how the chef's movements are accurately captured and translated into a 3D animated model. This method ensures that users receive realistic guidance on the proper cooking techniques, enhancing the overall engagement and satisfaction of the cooking experience.

Table 1 3D models used in the Ayam Woku dish

No.	Model	Description
1		Chicken
2		Turmeric (Whole and cut version)
3		Candlenuts (Whole and cut version)
4		Tomatoes (Whole and cut version)
5		Ginger (Whole and cut version)
6		Big Red Chili
7		Garlic
8		Onion/Shallot

(continued)

Table 1 (continued)

No.	Model	Description
9		Lemon Grass
10		Spring Onion

4.2 Speech Command

The user interaction with the HoloChef system is primarily facilitated through speech commands, as illustrated by the use case diagram in Fig. 4. This diagram highlights how the user can interact with the holographic system by issuing voice commands, such as choosing the chef's gender, requesting a recipe, or repeating instructions. The application then displays the necessary ingredients and step-by-step cooking instructions as a hologram.

To recognize and process speech, the system utilizes Unity's built-in Speech-to-Text functionality. This approach is suitable since the commands are simple, typically consisting of one or two words. Table 3 outlines the specific voice commands available in the system, categorized into phases, commands, and functions, each serving a particular role in guiding the user through the cooking process.

The primary method of interaction is voice commands, which are essential for the hands-free navigation of the cooking tutorial. Users initiate the experience by saying "Hello" or "Start" at the welcome screen, prompting the application to begin. Once the simulation starts, users can choose their preferred chef gender and select a dish, such as Ayam Woku or Es Pisang Ijo. Following this, the system provides a brief introduction to the selected dish and prompts the user to say "Let's begin" to start the cooking process.

The recipe is divided into three distinct sets of instructions. After each set, users are prompted to either continue to the next set or repeat the current one. The system acknowledges user inputs, providing verbal confirmation when necessary and displaying the next set of instructions, ensuring that users remain engaged and follow the cooking steps seamlessly.

Table 2 3D models used in the Es Pisang Ijo dish

No.	Model	Description
1		Banana
2		Salt
3		Santan/Coconut Milk Powder
4		Sugar
5		Rice Flour
6		Pandan Food Coloring

(continued)

Table 2 (continued)

No.	Model	Description
7		Condensed Milk
8		Cocopandan Syrup

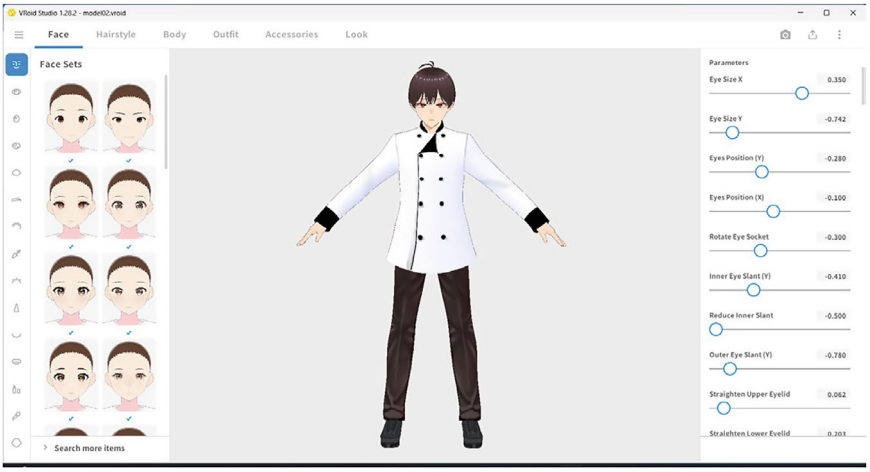


Fig. 2 Making male character in VroidStudio [15]

4.3 Holographic Display

The implementation of the holographic application plays a crucial role in guiding users through the cooking process for two dishes: Ayam Woku and Es Pisang Ijo. The 3D assets and overall design of the application are key to providing an engaging and informative experience. Upon launching the application, users are presented with a basic interface that includes a welcome box prompting them to say "Hello" or "Start." Once the initial greeting is acknowledged, the system requests users to select their preferred chef's gender, either female or male.

Following the selection of chef, users can choose between the two dishes, Ayam Woku or Es Pisang Ijo. After a successful transition, the interface (as shown in Fig. 5)

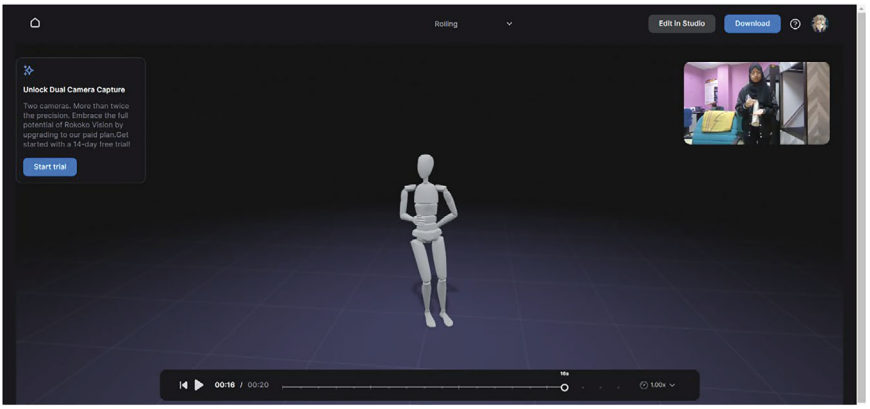


Fig. 3 Rokoko vision output [16]

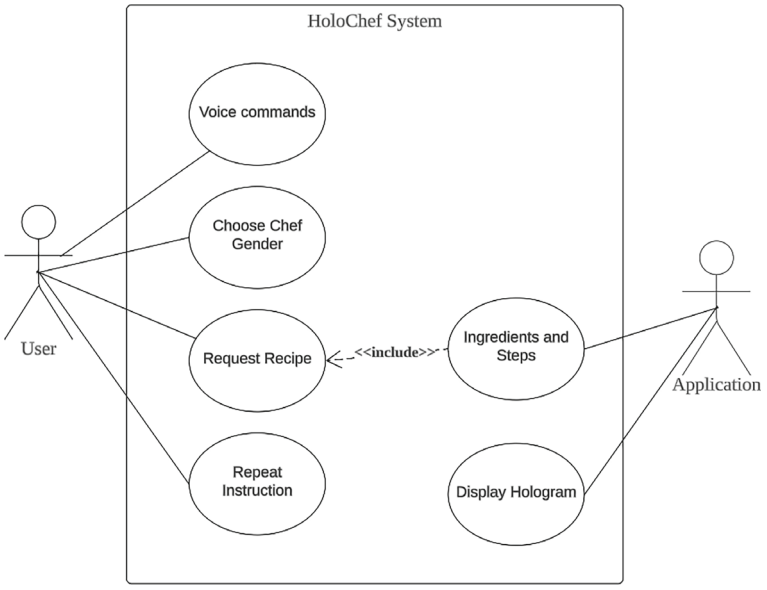


Fig. 4 Use case diagram of HoloChef

displays the choice of a female chef, along with relevant details about the dish, such as its origin, type, and the total cooking time. The hologram box as shown in Fig. 6, is designed with a wooden aesthetic, creating a rustic feel for the cooking experience.

The holographic display is set within a virtual workspace, where users are positioned behind a counter or table to simulate a real cooking environment. The critical elements of this setup include the proper placement of the microphone and hologram display. The microphone must be positioned close enough to capture the user's

Table 3 Speech commands and functions

Command	Function
“Hello”/“Start”	Start the hologram and begin cooking
“Female.”/“Mia.”	Chooses female as the preferred chef gender
“Male.”/“Alex.”	Chooses male as the preferred chef gender
“First dish”/“Dish one”	Chooses Ayam Woku as the recipe to cook
“Second dish”/“Dish two”	Chooses Es Pisang Ijo as the recipe to cook
“Let’s begin.”	Start cooking after the character explains how the cooking step will be instructed
“Repeat again”	Repeats the set of instruction from the beginning
“Continue next”	Move to the next set of instructions



Fig. 5 Dish interface selection with female chef

voice commands accurately, ensuring that the system can detect and respond to user inputs without delay. The seamless interaction between the user and the holographic interface is essential to maintaining an intuitive and hands-free cooking experience.



Fig. 6 Hologram result of HoloChef

5 Conclusion

The HoloChef 3D Culinary Guide provides a revolutionary means of cooking support by mixing holographic technology with speech recognition to provide a hands-free, immersive culinary experience. The system, intended to walk users through preparing classic dishes like Ayam Woku and Es Pisang Ijo, uses 3D models of a chef, ingredients, and cookware in conjunction with voice command functionality to provide an improved interactive and intuitive method of using recipes. Its intuitive interface enables users to focus on the cooking activity while being instructed in real-time and step-by-step by a holographic chef. The biggest contribution of this project is

the development of a speech-operated, interactive holographic assistant that provides visual and auditory guidance to users during meal preparation. While the prototype showcases good conceptual and functional strength, its practical utility needs to be tested further with real-world usability tests. Planned evaluations include pre- and post-preparation questionnaires to collect user input regarding voice command effectiveness, system useability, and general experience. Challenges faced when using the system will inform improvements in system performance. Expansion of the scope of recipes supported and an extension of the system's capability to accommodate more complicated instructions will be implemented in future development. Speech recognition accuracy, hologram resolution, and interaction design are also expected to be improved. Finally, HoloChef targets not just cooking enthusiasts but also newbies and users who want an effortless, hands-off cooking assistant that amplifies the cooking experience as a whole.

Acknowledgements We are deeply grateful to the Mixed and Virtual Reality Laboratory (Mivielab) and the ViCubeLab at Universiti Teknologi Malaysia (UTM) for their invaluable support in providing the equipment and technical assistance throughout the research. Their contributions were instrumental in the successful execution of this project. We would like to extend our gratitude to Universiti Teknologi Malaysia (UTM) for funding under UTM Encouragement Research (UTMER) grant number Q.J130000.3828.42J21 which has funded this research.

References

1. Ismail AF, Ismail AW, Fadzli FE, Shuib NM (2023) Real-time food simulation using real hand gesture for Malaysia cultural heritage. In: International conference on advanced computing and intelligent technologies. Springer Nature Singapore, Singapore, pp 249–258
2. Iftene A, Trandabăț D, Rădulescu V (2020) Eye and voice control for an augmented reality cooking experience. Procedia Comput Sci 176:1469–1478
3. Fadzli FE, Ismail AW (2019) VoxAR: 3D modelling editor using real hands gesture for augmented reality. In: 2019 IEEE 7th conference on systems, process and control (ICSPC). IEEE, pp 242–247
4. Papadaki E, Ntoa S, Adami I, Stephanidis C (2018) Let's cook: an augmented reality system towards developing cooking skills for children with cognitive impairments. In: Smart objects and technologies for social good: third international conference, GOODTECHS 2017, Pisa, Italy, November 29–30, 2017, Proceedings 3. Springer International Publishing, pp 237–247
5. Ismail AW, Fadzli FE (2024) XR-Menu: food ordering system in extended reality application. In: International conference on information, communication and computing technology. Springer Nature Switzerland, Cham, pp 109–122
6. Chai JJ, O'Sullivan C, Gowen AA, Rooney B, Xu JL (2022) Augmented/mixed reality technologies for food: a review. Trends Food Sci Technol 124:182–194
7. Fadzli FE, Kamson MS, Ismail AW, Aladin MYF (2020) 3D telepresence for remote collaboration in extended reality (xR) application. In: IOP conference series: materials science and engineering, vol 979, no 1. IOP Publishing, p 012005
8. Zhai KY, Cao YM, Hou WJ, Li XM (2020) Interactive Mixed reality cooking assistant for unskilled operating scenario. In: International conference on human-computer interaction. Springer International Publishing, Cham, pp 178–195
9. Majil I, Yang MT, Yang S (2022) Augmented reality based interactive cooking guide. Sensors 22(21):8290

10. Prohaska V, Castelló Ferrer E (2024) SPICE: smart projection interface for cooking enhancement. arXiv preprint arXiv:2412.03551. https://arxiv.org/abs/2412.03551
11. Yu L, Han B, Chen S, Yu LF (2025) HoloCook: a real-time remote mixed reality cooking tutoring system. In: HCI International 2024—late breaking papers. Springer, Cham, pp 203–219. https://doi.org/10.1007/978-3-031-76812-5_17
12. Frummet A, Speggiorin A, Elsweiler D, Leuski A, Dalton J (2024) Cooking with conversation: enhancing user engagement and learning with a knowledge-enhancing assistant. ACM Trans Inf Syst 42(5):1–29
13. Ruitenburg Y, Brankaert R, Houben M, Lee M, Pasman G (2023) Happje: stimulating collaborative cooking for people with dementia. In: Extended abstracts of the 2023 CHI conference on human factors in computing systems, pp 1–9
14. Li FM, Ng K, Zhu B, Carrington P (2025) OSCAR: object status and contextual awareness for recipes to support non-visual cooking. arXiv preprint arXiv:2503.05962
15. VRoid Project (2021) VRoid Studio. Retrieved Apr 28 2025. Retrieved from https://vroid.com/en/studio
16. Rokoko (2014) Vision. Retrieved July 2 2024. Retrieved from https://www.rokoko.com/products/vision

Zero-Knowledge Enhanced Adaptation Learning (ZEAL): A Framework for Verifiable Model Adaptation



Dhrumil Panchal, Hrushik Mehta, Pragati Shetty, Pratik Kanani, Darshana Sankhe, and Rashmi Kumar

1 Introduction

The speed at which advances have been made in deep learning and the far-reaching applications brought authenticity and novelty concerns of model adaptation into question. Privacy needs and secure verification procedures have preoccupied, especially for models based on proprietary data involving confidential information. This paper presents the ZEAL system, an innovative framework of zero-knowledge proof (ZKP) technology exclusively for verifiable adaptation of large language models (LLMs) without the loss of confidentiality. The traditional proof systems attached to training precision are vulnerable to disclosure of entire model parameters and data insights, and that is a new kind of privacy threat. Such attacks could deter organizations from actually certifying AI models for fear of revealing trade secrets. Cryptography-based verification schemes have established the feasibility of the use of zero-knowledge proofs to confirm verifiably correct training of deep neural networks without loss of confidentiality [1]. This shift of paradigm toward proof-preserving confidentiality can reshape confidence in AI by means of safe verification schemes with no loss of confidentiality. Standard approaches lead to high computational inefficiencies, particularly in the setting of iterative training iterations over non-linear layers. Inefficiencies like those render such methods incompatible with huge models and, as a result, limit their practical applicability. Steering clear of such computationally inefficients has become the need of research, and henceforth the motive to develop such new techniques between proof efficiency and scalability came forth [2]. By improving on proof frameworks, researchers have had great strides with tackling

D. Panchal (✉) · H. Mehta · P. Shetty · P. Kanani · D. Sankhe · R. Kumar
Dwarkadas J. Sanghvi College of Engineering, Mumbai, India
e-mail: dhrumilpanchal7510@gmail.com

D. Sankhe
e-mail: darshana.sankhe@djsce.ac.in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_31

the high-resource requirement to verify deep learning models that promised better accuracy and reliability. Other than that, privacy problems of big data AI system procedures have also been exposed and thus solutions offering confidentiality in the price of verification are required. Since AI models are increasingly built on cloud infrastructure, clients are compelled to rely on third-party service providers. In the absence of verifiable confirmation, models are susceptible to forgery or tampering. In response to these, cryptographic innovations in secure computation and proof aggregation have sought to instill trust on AI model integrity and confidentiality [3]. Such innovations enable greater deployment of privacy-preserving AI verification. LLM verification is still a pandemic issue since transformer architecture complexity necessitates sophisticated proof mechanisms to be optimal in efficiency, scalability, and precision. Because of the ginormous LLM parameter space, standard proof systems cannot keep up with verification cost. Specialist proof machinery has been engineered to assist these architectures specifically to enable verification from being infeasible in computational cost [4]. Such specialist machinery allows for correctness to be ensured in a computationally tractable manner even for large models. Additionally, recent work has enabled faster polynomial commitment schemes to enable more efficient verification using small proof size and verification time at no compromise of model security. The novelty reduces the process of verification with no compromise of good security warranties [5]. Such commitments ensure proof components have compact representations that reduce validation resources by a tremendous amount. ZEAL succeeds in overcoming such issues through the composition of polynomial commitments and streamlined adaptation proof structure. This development ensures accurate model on certain data results in verifiable updates with no leakage of sensitive parameters. With KZG commitment schemes and polynomial structured representation, in this work there is adaptation integrity with binding but hiding proof system. The result is a system in which all parties are able to securely certify updates to a model, and therefore machine learning change trust without confidentiality violation. With its innovation in privacy-friendly model verification, ZEAL provides a basis for secure AI building in scenarios where model integrity and confidentiality are of utmost concern. The work is on new ground of guaranteed AI adaptation, and the paper presents a practical and scalable proof system for transformer models. As adaptation of transformer-based large language models (LLMs) in extremely sensitive domains like healthcare and finance is increasingly becoming standard practice, secure and private adaptation takes the topmost priority. Existing verification protocols can effectively require disclosure of model parameters and training data, and this is of utmost privacy concern. In trying to address this issue, we introduce ZEAL (Zero-knowledge Enhanced Adaptation Learning) a new framework that employs zero-knowledge proofs (ZKPs) and polynomial commitments to enable verifiable adaptation of transformer-based LLMs with confidential data preservation.

1.1 Key Contributions

First Zero-Knowledge Proof Integration and Large Language Model Adaptation: We present a cryptographically robust framework using KZG commitments for transformer models to facilitate adaptation verification without sacrificing data and model parameter integrity.

- Low Overhead Generation of Proof: Our protocol is secure and has low overhead.
- Practical Implementation: Using GPT-2 fine-tuned on medical data, ZEAL achieves a 46.2% loss reduction with measurable proof efficiency.
- Privacy-Preserving Verification Protocol: We use variations of models in the form of polynomials purely for the purpose of safe verification without divulging confidential information.

ZEAL avoids the inefficiencies of past ZKP-based training verifications with a practical, binding-but-hiding proof system that is instantiable. This effort provides the foundation for confidential and verifiable AI in high-risk systems.

2 Literature Review

Recent developments in secure and privacy-protection machine learning have shed more light on zero-knowledge proofs (ZKPs) and their convergence with artificial intelligence (AI). Raeini [6] presented Privacy-Preserving Large Language Models (PPLLMs), where they looked into architectural approaches to preserving LLM usefulness while ensuring user data protection. This work was complemented by Prasley [7] on the use of zero-knowledge proofs for verifiable training of AI models, setting the stage for trustworthy model verification without exposing confidential data. Petrosino et al. [8] presented a distributed ledger technology (DLT)-based federated learning platform leveraging ZKPs to promote privacy in healthcare data sharing with real-world applicability to sensitive areas.

Pan et al. [9] presented zkTaylor, an innovative method utilizing Taylor series transformations to facilitate zero-knowledge proofs in machine learning procedures with proof efficiency versus expressiveness balance. Lu et al. [10] also built an extensible ZKP system for neural networks, solving performance bottlenecks of current proof systems. A more general context was given by Lavin et al. [11], who made an extensive survey of ZKP application, showing trends in areas ranging from blockchain to AI and secure computation.

Ganescu and Passerat-Palmbach [12] highlighted trust in generative AI by integrating ZKPs into user-model interactions, allowing outputs to be verified without divulging training data. Abbaszadeh et al. [13, 14] performed a substantial advancement in the verification of deep neural network training with ZKPs, allowing third parties to securely verify model genuineness and fairness. Kobelt et al. [15] compared various ZKP systems and gave insights into their computational cost, scalability, and

implementation overheads. George and Kizhakkethottam [16] followed the development of ZKP and how it played a central role in blockchain-based privacy protection, with increasing applicability in decentralized AI applications.

Contrasting ZEAL with Existing Approaches

ZEAL is distinct from earlier research in a number of significant ways:

- Particularly designed for Transformer Adaptation: Unlike most universal ML validation tools, ZEAL is specifically designed for transformer-based LLMs, in order to address the distinctive challenges of their design and parameter scale.
- Prioritize Adaptation over Training: Unlike most current work, which focuses on the verification of early model training (zkPoTs) or inference, ZEAL focuses on adaptation, with a number of issues in verification.
- Practical Implementation Tradeoff: ZEAL strikes a practical tradeoff between security guarantees and computational efficiency, as evident in our empirical results of reasonable proof generation time (84 s) and proof size (36 MB) for even models with millions of parameters.
- Polynomial Commitments for Parameters: We use the KZG polynomial commitments that are specifically optimized for compressing model parameter changes at a 0.303 ratio compared to storing all parameter changes in the plain.
- Capability Preservation Verification: Unlike most ZKP systems that provide correctness verification only, ZEAL also provides the assurance that the adaptation maintains general model capabilities and incorporates domain-specific knowledge.
- End-to-End Framework: ZEAL provides an end-to-end framework from model adaptation to proof generation and verification. All the current solutions address only a specific part of the verification process.

Although previous research has achieved remarkable progress in utilizing zero-knowledge proofs for multiple machine learning applications, ZEAL is the first end-to-end framework particularly developed for verifiable and privacy-preserving adaptation of transformer-based language models to address the increasing demand for transparent AI systems that can be adapted to domains in a non-compromised manner concerning security as well as privacy.

3 Methodology

ZEAL, as a design, is based on the principles of cryptographic commitments and zero-knowledge proofs (ZKPs) to support verifiable and privacy-retained model adaptation. The cryptographic commitment supports a prover to commit a value without disclosing it to a verifier until at some point in the future, with integrity and security guaranteed for adaptation verification. Zero-knowledge proofs support a prover to demonstrate knowledge of a secret without disclosing it. Such mechanisms allow the adaptation of machine learning models to be auditable without exposing sensitive

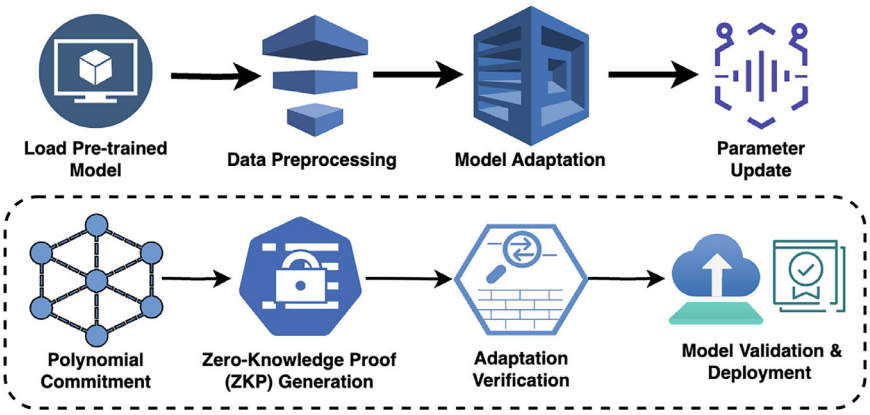


Fig. 1 Workflow diagram for the proposed methodology

training data or parameters. In this paper, fundamental mathematical notations, the core operation of large language models (LLMs), and the methodology life cycle of the proposed method are introduced. Algebraically, model adaptation is expressed by multiple notations for enabling quantitative analysis. Figure 1 shows the workflow diagram for the propsoed methodology.

Let M denote a pre-trained transformer-based model, while D represents the training dataset containing textual sequences. The original set of model parameters is denoted as W, and the updated model parameters after adaptation are denoted as W_{adapt}. A cryptographic commitment function, denoted as $C(W_{adapt})$ is applied to these parameters to ensure integrity. The adaptation process generates a zero-knowledge proof, allowing the verifier to validate that the parameters have been updated correctly without directly accessing them. Additionally, $H(D)$ cryptographic hash of the dataset, ensuring dataset integrity during adaptation. The adaptation process is quantified using an adaptation loss function LW, D that counts the difference in prediction of the model and reality. To judge functional preservation, accuracy function AccM, D is computed prior and post adaptation. Large language models (LLMs) are constructed upon the transformer model, a technology that revolutionized natural language processing by incorporating self-attention, feedforward layers, and positional encodings.

The transformer model receives input sequences that are processed in parallel by using multi-head attention to determine inter-token dependencies. Such models learn by fine-tuning the pre-trained weights for domain-specific data, such that performance is improved while still retaining generalization. The workflow of methodology involves a number of steps: pre-trained transformer model loading, preprocessing the dataset, fine-tuning based on gradient-based optimization, generating polynomial commitments for weights adapted, constructing zero-knowledge proofs, and proof verification and capability testing to confirm adaptation validation. The ZEAL framework is designed to integrate model adaptation with cryptographic proofs, providing

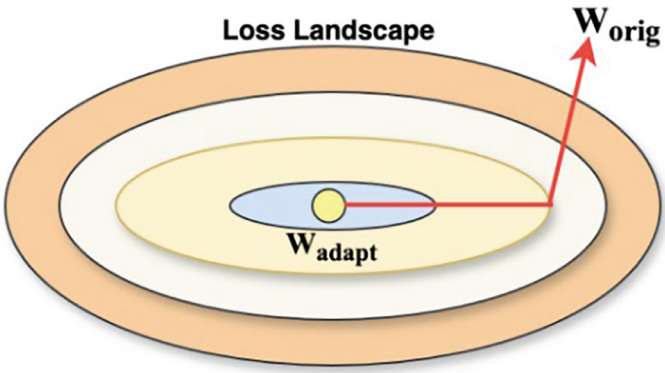


Fig. 2 Gradient descent optimization

a transparent, efficient, and verifiable process. Model adaptation involves training the model on a new dataset D while minimizing a predefined loss function.

This process modifies the model parameters from W_{orig} to W_{adapt}, ensuring domain-specific enhancement. But for privacy and integrity, utilizing a polynomial commitment scheme from KZG commitments, securely binding model parameters without revealing them. The adaptation proof generation constructs an organized proof to check for update correctness, and capability validation ensures that adaptation never breaks the model functionality. Figure 2 shows the gradient descent optimization.

3.1 Model Adaptation

Model adaptation involves adjusting the model parameters through an optimization process that minimizes the loss function $L(W, D)$. The adaptation process follows a gradient-based update rule given by:

$$W_{\text{adapt}} = W_{\text{Orig}} - \eta \Delta L(W, D)E \tag{1}$$

Equation 1 Gradient-based update rule (where η is the learning rate, governing the magnitude of parameter updates, and ΔL represents the gradient of the loss function with respect to the model parameters). During fine-tuning, the model undergoes multiple training iterations, updating weights iteratively until convergence.The implementation of model adaptation leverages AdamW as the optimizer, a variant of stochastic gradient descent (SGD) that incorporates adaptive learning rates and weight decay regularization. Training data is tokenized using the model's tokenizer, ensuring consistency with pre-trained embeddings. Input sequences are padded to a uniform length, and batches are processed in parallel to optimize memory usage.

The loss function employed in the adaptation process is the cross-entropy loss, given by:

$$L(W, D) = -\sum_{i=1}^{N} y \log \hat{y}_i \tag{2}$$

3.2 Polynomial Commitment

To verify that the model parameters have been adapted correctly without revealing their exact values, employing a polynomial commitment scheme using KZG commitments. The commitment process involves representing model parameters as coefficients of a polynomial, ensuring that they can be securely committed to and later verified. Given a set of adapted weights $W_{\text{adapt}} = (w_1, w_2, w_3 \ldots w_n)$ defines the polynomial representation:

$$P(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \cdots + w_n x^n \tag{3}$$

Equation 3 Polynomial representation (where the coefficients w_i correspond to individual model parameters. To generate a cryptographic commitment, evaluating this polynomial at a secret value τ, known only to the prover, ensuring binding and hiding properties).

The commitment is computed as:

$$C\left(W_{\text{adapt}}\right) = g^{p(\tau)} \text{mod} p \tag{4}$$

Equation 4 Commitment Function (Where g is a generator of the cyclic group, and p is a prime modulus. This commitment guarantees that once model parameters are committed, they cannot be altered without detection).

3.3 Adaptation Proof Generation

To ensure the correctness of adaptation without leaking privacy, a zero-knowledge proof p is constructed. The proof certifies that the committed polynomial is a correct representation of the adapted weights without revealing them. Construction of the proof takes the form of a series of structured steps. Model parameters are then divided into parts in order to achieve optimal proof efficiency. Each part undergoes polynomial interpolation, producing a low-degree polynomial form. The KZG commitment is computed for each polynomial to yield a structured commitment scheme. The proof is then encoded using cryptographic techniques, allowing efficient verification

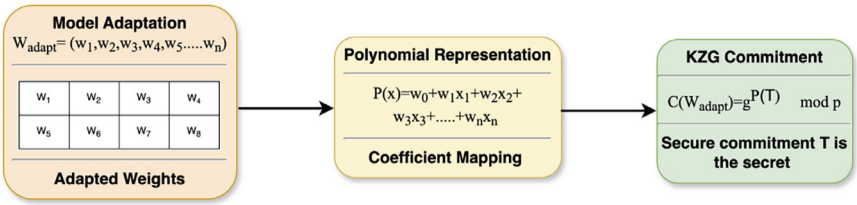


Fig. 3 Cryptographic components of ZEAL

by a third party. The verifier checks the proof without access to W_{adapt}, ensuring privacy while validating correctness. Figure 3 shows the cryptographic components of ZEAL.

3.4 Security Settings and Assumptions

This system has robust security assumptions. The binding requirement of the commitment scheme, i.e., no one can modify a commitment after making it, is satisfied. The hiding requirement guarantees that the commitment is kept private in a way that the commitment does not reveal information about W_{adapt} The soundness of the zero-knowledge proof construction, i.e., tampering attempt on behalf of an adversary is foiled, is satisfied. Cryptographic hashes add extra security to dataset integrity, i.e., unauthorized data modification $H(D)$ are circumvented. The verifier can demonstrate verification scalability without sacrificing security. The ZEAL framework exemplifies an interesting interposition of model adaptation with polynomial commitments and zero-knowledge proofs to enable privacy-free verifiable updates. By leveraging structured proof generation and cryptographic primitives for security, this work has a robust solution for transparent adaptation of machine learning models. Next projects will explore optimizations in shrinking proof size, boosting computational efficiency, and deployment on deep learning models and adversarial threat models.

3.5 Novelty of the ZEAL Framework

The ZEAL approach offers several contributions to AI adaptation that are verifiable:

1. First Adaptation-Targeted Verification System for Transformers: Prior work has studied verification of early model training (zkPoTs) or inference steps but not adaptation in transformer language models, which is ZEAL's primary target. Its adaptation verification poses novel challenges not tackled by prior methods, especially modeling and proving correctness of parameter updates over the enormous parameter spaces common in modern LLMs.

2. ZEAL introduces a new representation of parameters as well as a new approach to partitioning the high-dimensional parameter space of the transformers into computationally manageable polynomial pieces. Such partitioning allows efficient commit and verification with cryptographic guarantees. This approach achieves a polynomial commitment density of 0.001, capturing the sparsity of adaptation adjustments during fine-tuning, and therefore the gains in efficiency are dramatic relative to naive approaches.
3. Dual-Purpose Verification Method: Unlike existing verification systems that are mainly focused on computational correctness, ZEAL provides a dual-purpose verification approach that simultaneously verifies that: Procedural correctness: Mathematically confirming that the adaptation process had indeed followed the proclaimed optimizing process Functional preservation: Expressing directly that the overall ability of the model is preserved but domain-specific performance is enhanced.
4. Scalable Commitment Architecture: ZEAL's commitment architecture is designed to support the scale of modern transformer models, and we have succeeded in that with GPT-2 (124 M parameters). Our architecture employs novel compression techniques that compress proof size by 69.7% relative to storing changes in parameters directly, overcoming a significant limitation of current methods when applied to LLMs.
5. End-to-End Verification Pipeline: Prior systems have had incomplete solutions (e.g., zkTaylor for non-polynomial functions, Kaizen for verification training), but ZEAL is the first end-to-end pipeline for verification of transformer adaptation. This includes pre-adaptation capability estimation, adaptation process, polynomial commitment creation, zero-knowledge proof creation, and capability verification—all within a unified framework.

3.6 Motivation Behind the Proposed Methodology

The ZEAL architecture is driven by a wide range of important concerns that tackle important limitations of current methods:

1. Emphasize on Adaptation Rather Than Training: The adaptation (fine-tuning) paradigm has become standard for tailoring a base model to a specific domain, so verification of this process is especially important. Complete training verification techniques such as Kaizen would be too costly for LLMs (which take billions of parameters and millions of GPU hours to train). ZEAL is designed to target adaptation and address the fact of how LLMs are actually used and tailored nowadays, offering verification for the most typical path of modification.
2. Choice of KZG Polynomial Commitments: After going through various cryptographic commitment schemes, we selected the KZG polynomial commitments for many good reasons: Polynomial-friendly: KZG commitments are well-suited to represent mathematical operations utilized in gradient-based adaptation. Constant-size proofs: Unlike Merkle tree approaches, KZG commitments

provide constant-size proofs regardless of the polynomial degree. Efficient verification: The verification step is computationally efficient (O(1) in the polynomial size).

Committer security: KZG commitments achieve fast and robust security guarantees under the discrete logarithm assumption.

3. Parameter Segmentation Strategy: This method of separating the model into fixed-degree polynomial parameter pieces (average 10) is provided with theoretical substantiation through empirical evidence that this degree allows a best trade-off between representation fidelity and computational expense. Lower degrees would be at the expense of fidelity in representing parameter change, and higher degrees would be more computationally expensive without associated gains in representation fidelity.
4. Implementation of Loss-Based Authentication:

ZEAL confirms adaptation on the basis of loss reduction as opposed to rigid parameter matching.

This was chosen because:

- It accommodates the natural stochasticity of optimization processes.
- It aims at the successful result of the adaptation and not at specific parameter values.
- It provides a more powerful test of adaptation success than parameter-level validation. It aligns with the practical goal of adaptation: improved performance on domain-specific tasks.

5. Integration of Capability Preservation Testing: The inclusion of capability preservation testing addresses a core lack in previous research. Previous techniques assume that proper adaptation algorithm execution is sufficient for verification, but it does not guarantee that modifications to model capabilities can be harmful. Our two-stage verification approach guarantees that:

The model carries out better domain-specific tasks (46.2% reduction of loss on medical dataset) The model still maintains its overall capabilities without loss The validation procedure considers both procedural accuracy and functional results By taking these factors into account, ZEAL provides a new and holistic solution to the ever-more difficult challenge of verifiable adaptation for transformer language models. Our solution is balanced between theoretical security guarantees and real-world efficiency concerns and can therefore be used in real-world environments.

4 Implementation and Results

The ZEAL framework was tested using a number of performance indicators that offer good insights into the model adaptation process's efficiency and effectiveness. The model originally had a loss of 5.08, which reduced to 2.73 following adaptation,

representing a 46.2% improvement in performance. Such a considerable reduction in loss reflects that the model effectively integrated domain knowledge from the medical dataset. The adaptation process lasted approximately 58 s, while the process of generating proof took 84 s, resulting in the total execution time being 142.57 s. This result reflects that proof generation is computationally costly, taking over half the total time, and presents a possible area for improvement in future work. The maximum memory usage was 3.36 GB, which is reasonable for a model of such size (GPT-2 with 124 million parameters); however, memory efficiency is bound to become an increasingly significant factor as models grow in size. The proof file size was recorded at 36 MB, accompanied by the resultant compression ratio of 0.303, meaning that the proof system is occupying only 30.3% of the storage capacity compared to storing all parameter modifications directly. Such compression is crucial to scalability, particularly in the context of larger models. The system produced a total of 124,503 commitments, each representing a specific section of the model parameters, thereby ensuring careful monitoring of parameter modifications that is critical to guaranteeing accuracy in verification. The average polynomial degree was maintained at 10.0, balancing between accuracy and computational efficiency. However, this constant degree might not be the optimal for all parameter distributions, hinting at a possible area for future improvement. The commitment density of 0.001 shows that only an infinitesimal portion of the parameter space was actually modified during adaptation, as would be expected in the knowledge that fine-tuning is typically marked by subtle adjustments to the model parameters. The ZEAL framework is composed of a number of basic elements, each playing an essential role in adaptation and verification. The ModelAdapter is in charge of initializing the pre-trained GPT-2 model and fine-tuning it on a medical dataset. The model was fine-tuned with a learning rate of 1e-5 for a single epoch with a batch size of 4, resulting in a 46.2% reduction in loss. The KZGCommitment component provides the implementation of the Kate-Zaverucha-Goldberg (KZG) polynomial commitment scheme, resulting in a total of 124,503 commitments, each of which corresponds to a segment of model parameters. The commitments were saved in a compact representation, with a compression ratio of 0.303. The Adaptation Proof Generator translates parameter changes to polynomials and generates cryptographic commitments for every parameter segment. Proof generation took 84.44 s, resulting in a proof file of 36 MB, while keeping the average polynomial degree at 10.0 to accommodate an equilibrium between accuracy and computational efficiency. The CapabilityValidator evaluates the model's performance on general knowledge and domain-specific test suites to ensure that the model's general capabilities are preserved while it learns domain-specific knowledge. The validator ensured that the model's general capabilities were preserved, with no noticeable performance loss on general knowledge tasks. Lastly, the ZEAL Storage component saves proofs, commitments, and performance metrics for future verification. This storage system logged all relevant metrics, including adaptation time, proof generation time, and memory usage, in a structured format to allow easy retrieval and analysis. The ZEAL framework workflow is implemented to preserve the integrity and verifiability of the adaptation process. The pre-trained GPT-2 model is loaded first, followed by the evaluation of its initial capabilities using

both general knowledge and domain-specific test suites. Second, the model is adapted on the provided medical dataset with specific hyperparameters (learning rate: 1e−5, epochs: 1, batch size: 4). Table 1 represents the performance and efficiency metrics.

After this adjustment, the model functionalities are tested to ensure that it has learned in the specific domain without losing its general capabilities. Parameter updates are converted to polynomial representations, and cryptographic commitments are calculated over every parameter segment. Finally, proofs, commitments, and performance metrics are stored for future verification by third parties. Despite the promising results, the ZEAL framework has a number of limitations that can be addressed in future studies. The proof generation process is computationally expensive and takes around 84 s in a comparatively small-scale model like GPT-2, which can become a bottleneck for large models like GPT-3 or LLaMA. The fixed polynomial degree of 10 can be suboptimal for all parameter distributions, and future studies can explore the use of adaptive polynomial degrees to increase efficiency and accuracy. Furthermore, the current implementation only supports the Hugging

Table 1 Performance and efficiency metrics

Metric category	Metric name	Value	Unit	Description
Performance	Initial loss	5.08	–	Loss before adaptation
	Final loss	2.73	–	Loss after adaptation
	Loss improvement	2.34	–	Absolute improvement
	Improvement percentage	46.2	%	Relative improvement
Time efficiency	Adaptation time	58.13	sec	Time for model fine-tuning
	Proof generation time	84.44	sec	Time for cryptographic proof creation
	Total execution time	142.57	sec	Total processing time
Resource usage	Peak memory usage	3356.3	MB	Maximum memory consumed
Proof system	Proof size	37.71	MB	Size of generated proof file
	Number of commitments	124,503	–	Total polynomial commitments
	Average polynomial degree	10	–	Mean degree of polynomials
	Maximum polynomial degree	10	–	Highest degree used
	Commitment density	0.001	–	Proportion of parameters with commitment
Storage efficiency	Compression ratio	0.303	–	Size ratio compared to storing all parameters
	Storage overhead	0.303	–	Additional storage required

Face Transformers models, and expanding compatibility to other model architectures would make the framework more generic. While model sizes grow, there will be a need to optimize the proof generation process to be able to cope with the increasing computational and memory requirements. Future studies can also explore integrating zero-knowledge proofs around dataset properties to enable more rigorous verification of the adaptation mechanism. The ZEAL framework can marry machine learning adaptation with cryptographic verification, thus providing an open and verifiable approach towards domain-specific adaptation of models. The performance improves drastically while maintaining general capabilities, as evident from the results, and the proof system ensures the integrity of the adaptation process. Future research studies will further improve the proof generation process, support larger models, and increase the scalability and interoperability of the framework.

Figure 4 shows a multi-dimensional visualization that represents the adaptation and proof generation phases of the ZEAL framework. The x-axis denotes execution time in seconds, while the primary y-axis tracks memory usage (MB), and the secondary y-axes show loss values (dashed curve) and the number of commitments (bar plot). The scatter plot uses bubble sizes to indicate model performance, where larger bubbles represent better performance (lower loss). The color gradient reflects phase progression, and vertical dashed lines mark transitions between key phases.

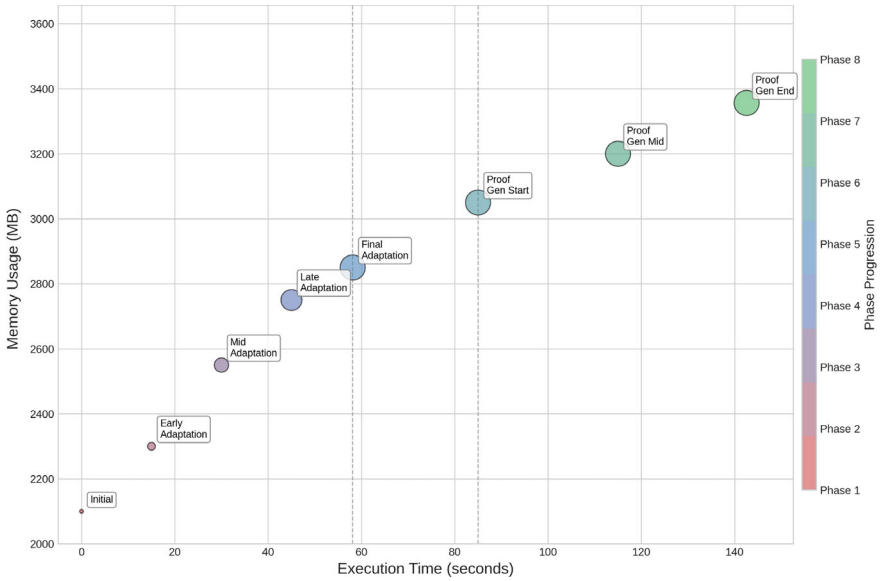


Fig. 4 Performance of ZEAL framework across different phases, showing memory usage, loss values, and commitment progression over execution time

5 Conclusion

This research introduces ZEAL, a new zero-knowledge proof (ZKP)-enabled confidential model adaptation verifiability system that maintains confidentiality. Through the combination of polynomial commitments and a sound proof scheme, ZEAL maintains high-quality large language model (LLM) integrity without exposing proprietary model parameters and sensitive training data. Our GPT-2 model fine-tuned on medical data prototype shows real-world applicability and significantly reduces loss and maintains cryptographic verifiability. The findings present a trade-off between proof tightness and computational performance, where the framework produces commitments and proofs in a reasonable runtime. Through the use of Kate-Zaverucha-Goldberg (KZG) commitments, ZEAL builds on a binding but hiding proof system that allows model adaptation to be securely verified. Contributions include the design of a structured verification pipeline, cryptographic commitment of model parameters, and ZKP-based integrity verification. Although it has its benefits, there are issues with regard to proof generation cost, polynomial degree fixity, and scalability to bigger models. Upcoming research will be aimed towards addressing proof efficiency, adaptive polynomial degrees, broader architectural support for a variety of transformer-based models, and investigating dataset property verification. These issues will significantly boost ZEAL's applicability in privacy-constrained domains like healthcare, finance, and legal AI. By providing a complete framework for verifiable model adaptation, ZEAL extends the nascent field of AI verification for privacy-preserving applications and thereby establishes a foundation for secure AI deployment in mission-critical missions where both model confidentiality and integrity are required to be maintained. In short, ZEAL provides a route to secure, confidential adaptation of AI, which is required for real-world release in high-risk environments.

References

1. Wellington S (2024) BasedAI: a decentralized P2P network for zero knowledge large language models (ZK-LLMs). arXiv: arXiv:2403.01008. https://doi.org/10.48550/arXiv.2403.01008
2. Sun H, Li J, Zhang H (2024) zkLLM: zero knowledge proofs for large language models. In: Proceedings of the 2024 on ACMSIGSAC conference on computer and communications security. ACM, Salt Lake City UT USA, pp 4405–4419. https://doi.org/10.1145/3658644.3670334
3. Sun H, Bai T, Li J, Zhang H (2025) ZkDL: efficient zero-knowledge proofs of deep learning training. IEEE Trans Inf Forensics Secur 20:914–927. https://doi.org/10.1109/TIFS.2024.3520863
4. South T et al (2024) Verifiable evaluations of machine learning models using zkSNARKs. arXiv: arXiv:2402.02675. https://doi.org/10.48550/arXiv.2402.02675
5. Singh S (2024) Enhancing privacy and security in large-language models: a zero-knowledge proof approach. Int Conf Cyber Warf Secur 19(1):574–582. https://doi.org/10.34190/iccws.19.1.2096

6. Raeini M (2023) Privacy-preserving large language models (PPLLMs). SSRN Electron J. https://doi.org/10.2139/ssrn.4512071
7. Prasley T (2024) Zero-knowledge proofs for verifiable AI model training 2(2)
8. Petrosino L, Masi L, D'Antoni F, Merone M, V ollero L (2025) A zero-knowledge proof federated learning on DLT for healthcare data. J Parallel Distrib Comput 196:104992. https://doi.org/10.1016/j.jpdc.2024.104992
9. Pan D, Liu K, Li B, Zheng Y, Ma J (2024) zkTaylor: zero knowledge proofs for machine learning via taylor series transformation. In: Proceedings of the 2024 2nd ınternational conference on advances in artificial ıntelligence and applications. ACM, Wuhan Shaanxi China, pp 37–41. https://doi.org/10.1145/3712623.3712646
10. Lu T et al, An efficient and extensible zero-knowledge proof framework for neural networks
11. Lavin R, Liu X, Mohanty H, Norman L, Zaarour G, Krishnamachari B (2024) A survey on the applications of zero-knowledge proofs. arXiv: arXiv:2408.00243. https://doi.org/10.48550/arXiv.2408.00243
12. Ganescu BM, Passerat-Palmbach J (2024) Trust the process: zero-knowledge machine learning to enhance trust in generative AI interactions. arXiv: arXiv:2402.06414. https://doi.org/10.48550/arXiv.2402.06414
13. Abbaszadeh K, Pappas C, Katz J, Papadopoulos D (2024) Zero-knowledge proofs of training for deep neural networks. In: Proceedings of the 2024 on ACM SIGSAC conference on computer and communications security. ACM, Salt Lake City UT USA, pp 4316–4330. https://doi.org/10.1145/3658644.3670316
14. Abbaszadeh K, Pappas C, Katz J, Papadopoulos D (2024) Zero-knowledge proofs of training for deep neural networks. In: Proceedings of the 2024 on ACM SIGSAC conference on computer and communications security. ACM, Salt Lake City UT USA, pp 4316–4330. https://doi.org/10.1145/3658644.36703
15. Kobelt M, Sober M, Schulte S (2023) A benchmark for different implementations of zero-knowledge proof systems. IEEE Int Conf Blockchain (Blockchain) 2023:75–82. https://doi.org/10.1109/Blockchain60715.2023.00015
16. George L, Kizhakkethottam JJ (2021) Evolution of zero-knowledge proof (ZKP) and ıts role in blockchain applications for ensuring data privacy. Int J Eng Develop Res (IJEDR) 9(1). [Online]. Available: https://www.ijedr.org/viewpaperforall.php?paper=IJEDR2101002

JOVA Robotic Prosthetic Arm



Jose Kevin, Kesanakurthi Aditya, Karuturi Vaishnavi, and V. Nithya

1 Introduction

Prosthetic arms have contributed a great deal to the lifestyle of upper limb amputees, providing basic functionality and autonomy. These tools seek to emulate the natural motions of the hands, giving individuals the power to conduct normal day-to-day tasks effortlessly. Current prosthetic arms tend to incorporate sophisticated technologies such as myoelectric sensors, brain computer interfaces, and AI-driven systems for more accurate control. Although effective, they are costly, complicated, and need to be trained for a long time. The use of invasive sensors also restricts accessibility. Most prosthetic arms are not affordable because they cost too much, have complicated mechanisms, and have limited use. Others need frequent repair, whereas others are difficult to control intuitively. Also, wired configurations limit mobility and comfort for users. This research work seeks to provide a user-friendly and affordable prosthetic arm through the utilization of a wireless glove with an accelerometer. It eliminates invasive control systems for intuitive use. The design is affordability, accessibility, and real-time-oriented. The proposed solution is based on a prosthetic arm controlled by a motion-sensing glove equipped with an accelerometer. Wireless

J. Kevin · K. Aditya · K. Vaishnavi · V. Nithya (✉)
Department of Electronics and Communication Engineering, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India
e-mail: nithyav@srmist.edu.in

J. Kevin
e-mail: jr7373@srmist.edu.in

K. Aditya
e-mail: kk2793@srmist.edu.in

K. Vaishnavi
e-mail: kk8656@srmist.edu.in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_32

data transfer through nRF24L01 modules guarantees real-time movement duplication. Servo motors are utilized to provide smooth and precise hand movements. This research introduces a low-cost, gesture-controlled prosthetic arm that is non-invasive and constructed using inexpensive components in order to benefit underprivileged communities. It utilizes sensors and microcontrollers to recognize user gestures and convert them into limb movements. The product is centred on affordability, intuitive control, and simplicity in an attempt to offer a working and affordable solution for upper-limb amputees in low-resource environments.

The paper is organized as follows: Sect. 2 discusses related work and comparison with current technologies. Section 3 describes the system design and elements. Section 4 outlines the implementation and testing process. Section 5 analyses performance and user feedback, and Sect. 6 concludes with future improvements.

2 Related Works

Current advancements in prosthetic technology are aimed at increasing control methods towards replicating the natural motion of human limbs. Starke et al. [1] presented a semi-autonomous control method in which users are capable of fluidic grasping of common objects on the basis of a single channel of EMG. In addition, [2] proposed machine-trained wearable sensors to recognize real-time hand movements with greater precision and responsiveness to prosthesis control. The combination of multimodal sensing and machine learning methods has greatly progressed prosthetic technologies. Abd et al. [3] presented multichannel haptic feedback systems that release prosthetic hand dexterity, enabling users to carry out complicated activities more conveniently. Sankar et al. [4] developed a natural biomimetic prosthetic hand with neuromorphic tactile feedback, highly mimicking the sensory behaviour of a natural hand. Jayaweera [5] used an EMG-based real-time pattern recognition model for prosthetic hand control, proving the potential of machine learning to understand user intent 0.3D printing has proven to be a revolutionary technology in the creation of affordable prosthetic limbs. Taniguchi et al. [6] presented a systematic review on 3D-printed prostheses in developing nations, highlighting the capabilities of additive manufacturing in creating low-cost and individualized prosthetic interventions. Jr et al. [7] designed an Arduino-controlled sensor system for a robotic laparoscopic surgery unit, reflecting the versatility of 3D printing in medical device manufacturing. Logesh [8] developed and controlled an EMG-based medical prosthetic hand for simple hand gestures, reflecting the integration of simple interfaces in prosthetics. The emphasis on affordability and accessibility can be seen in the research of [9] who created a wearable finger prosthesis with a pneumatic actuator for partial amputees, illustrating the use of soft robotics in improving prosthetic performance. Roy et al. [10] highlighted the cost-effective design and creation of a prosthetic hand for upper limb amputees, considering the demand for affordable prosthetics. Abbady et al. [11] investigated 3D-printed prosthetics for the third world, promoting the mass use of this technology to enhance accessibility [12] indicates the current challenge in realizing

consistent electromyography (EMG)-based control of lower limb prostheses despite extensive research interest. Gentile et al. [13] contend that more advanced, human-like hierarchical control schemes for prosthetic hands are needed to surpass the limitations of the existing simple control schemes. In a different context, [14] show the viability of employing inexpensive Arduino microcontrollers for robotic arm control. Lastly, [15] mention the necessity of cheap prosthetic solutions by showing the design of an inexpensive, 3D-printed finger prosthetic with particular reference to developing nations. Current prosthetic work does involve EMG-based methods for fine control, but these tend to be costly and complicated. Gesture-control alternatives based on IMU sensors, demonstrated by Thomas et al. [11] provide easier, non-invasive options.

The JOVA Robotic Prosthetic Arm extends these designs with a low-cost, non-invasive system utilizing IMU sensors and RF modules for real-time control. The 3D-printed, servo-motor driven device provides key hand functions at affordable prices and with simple maintenance. JOVA differs from most current systems in being designed specifically for underserved populations, integrating functionality with accessibility.

3 Proposed Methodology

A. Robotic Prosthetic Arm System Methodology: This work discusses the principle and implementation of a gesture controlled robotic prosthetic arm in (Fig. 1), as depicted in the block diagram. It consists of a control side (sensor-based glove) in (Fig. 5), and an actuation side (prosthetic arm), communicating wirelessly through nRF modules as shown in Fig. 3. nRF24L01 module used in this work provides wireless communication, passing motion data stream commands, finger position values from the accelerometer in the glove to the prosthetic arm for real-time control, and controlled using an Arduino microcontroller. This depicts the mechanism of a glove-controlled prosthetic arm using a glove which has an accelerometer, Arduino, and nRF wireless data transmission modules. The signals received are sent to a separate Arduino on the controlling side, and it drives the forearm motors and the prosthetic arm with a source of power and input code to the motors programmed.

B. Sensor Based Hand Motion Detection (Control Side): A glove that triggers the sensor detects hand movements on the control side. Accelerometer Data Acquisition and Preprocessing: The accelerometer of the project can sense hand motion like tilt, rotation, and acceleration to provide linear acceleration and transmit real-time movement signals to control the prosthetic arm through wireless communication.

C. Gravity Compensation and Data Acquisition: Continuously sample accelerometer data (acceleration in X, Y, and Z axes) at a suitable sampling rate (e.g., 50–100 Hz). Noise Reduction Apply a low-pass filter (e.g., moving average filter, Butterworth filter) to remove high-frequency noise from the accelerometer data.

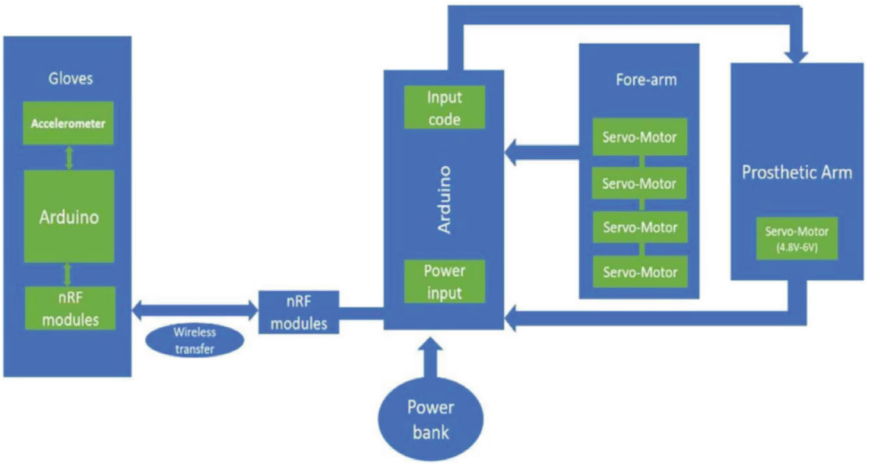


Fig. 1 System diagram

Subtract the gravity component from the accelerometer data to obtain linear acceleration. This can be done using a complementary filter, Wireless Communication (nRF modules) in (Fig. 3). The processed data between the glove and the prosthetic arm is sent through an nRF module operating in the 2.4 GHz frequency band. The data transmitted are the gestures of the hand detected by the accelerometer. The receiver module on the prosthetic arm receives this data for further processing.

D. Power Efficiency and Battery Life: The platform was powered by a 7.4 V 2200 mAh Li-Po battery. With constant use (gesture input and servo actuation every 10 s), the platform worked well for approximately 2.5 h. When idle for extended periods, the platform used very little current, significantly extending standby time. Energy usage was greatest with motor actuation and wireless data bursts.

E. Microcontroller Processing (Arduino Side): The Arduino microcontroller circuit on the receiving side receives the information and produces control signals for the prosthetic arm.

Input Code and Interpretation: The firmware running under the Arduino fetches data and represents it to correlate with the defined gestures of the hand. The data are processed to cube the signals to be turned into PWM to run the motors.

F. System Responsive Control Algorithm: The control algorithm takes various threshold values to classify finger bending into several categories. The system then detects the corresponding gestures and determines the servo positions for each gesture detected. An Arduino ensures relatively smooth and coordinated motion of the motors in parallel.

G. Comparison of the prosthetic arm with the existing model. The features of the proposed work is compared with the existing approaches and listed in Table 1.

Table 1 Comparative study with the existing model

S. No.	Feature	Proposed model	Existing model
1	Control mechanism	Gesture-based control using an accelerometer attached to the gloves	Myoelectric control using EMG sensors, EEG-based control, or mechanical switch-based control, [2]
2	Communication	Wireless communication using nRF modules (RF-based transmission)	Wired connections (early models), Bluetooth, WiFl, or direct muscle signal transmission, [1]
3	Processing unit	Arduino-based microcontroller for gesture signal processing	Microcontrollers like Raspberry PL, STM32, or custom embedded systems, [3]
4	Power source	Rechargeable power bank (battery-powered motors)	Rechargeable Li-ion batteries, some models use kinetic energy harvesting, [9]
5	Response time	Real-time response with minimal latency (<100 ms)	Varies (some models with Al prediction offer faster response, while older models have noticeable delays), [8]
6	Grip strength control	Basic grip control using multiple motors	Advanced models allow variable grip strength and adaptive gripping [4]
7	Material and weight	3D-printed lightweight structure for affordability	Custom prosthetics use carbon fiber, titanium, or soft robotics [5]
8	Suitability for amputees	Primarily designed for wrist-level amputees (not yet adaptable for higher amputations)	Advanced bionic limbs support full-arm amputations with neural interfaces [12]

Feedback Mechanism: More advanced systems can use force sensors for haptic feedback. This adds functionality to the device by the fact that a user feels the applied pressure. The entire system is powered by a power bank, hence portable and rechargeable.

Surgical Prosthesis Movement: The robotic arm moves the hand with external servo motors as shown in Fig. 2. It Displays a 3D-printed modular structure designed to hold several servo motors, which are utilized to control the prosthetic arm movements. The servo motors relate to labeled cables for convenient signal flow from the controller unit.

Multi-Servo Control: The system is based on many servo motors controlling fingers. Each motor initially receives a control signal from the Arduino board for its motion. Closes and opens hand based on accelerometer input, imitating pinching, holding, or releasing movement (Fig. 3).



Fig. 2 Main servo frame

Wireless Communication and Signal Transmission: The nRF module allows communication with minimum latency between the glove and prosthetic arm as depicted in Figs. 1 and 4. Wireless communication, that is no physical connections are needed. The data stream commands finger position values, which are decoded at the receiving point.

Novelty: Jova focuses on balancing functionality and cost: Contrary to expensive commercial myoelectric arms, JOVA exploits 3D printing and available electronics to reduce costs without sacrificing functionality.

Using non-invasive control: Rather than surgery-inducing myoelectric implants, JOVA utilizes a gesture-control glove with motion sensing for intuitive gesture control.

Technologies merging: JOVA merges 3D printing, motion sensing, and wireless communication in a one-of-a-kind solution.

With a focus on accessibility: It puts underserved groups (children, youth, poor communities) first, usually forgotten by current solutions.

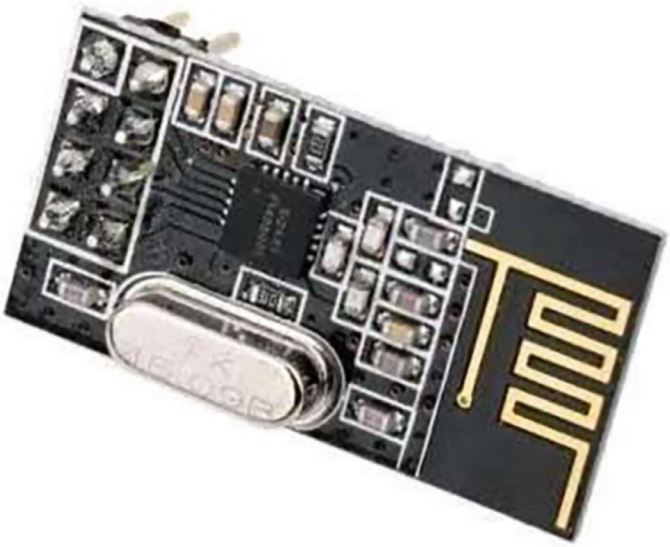


Fig. 3 nRF module

4 Results and Discussions

The JOVA Robotic Prosthetic Arm, made using the help of a 3D-printed hand, accelerometer sensors, and servo motors, was successfully built to provide basic hand motions controlled using a motion-sensing glove is shown in the Fig. 4. The use of an Arduino microcontroller allows real-time data processing to be efficient, and wireless communication was made using nRF24L01 module. The prosthetic arm employed in this work is a 3D-printed, lightweight, and affordable prosthetic for amputees of the upper limb. The prosthetic is controlled by servo motors that replicate the movement of hands in real-time using signals sent from an Arduino microcontroller through the motion-sensing glove.

The control component of the prosthetic arm has an Arduino microcontroller to receive wireless inputs from the glove's accelerometer through an nRF24L01 module as shown in Fig. 5. It processes the information to provide accurate control signal to the servo motors to for real-time hand motions.

The motion-sensing glove serves as the primary input interface, equipped with an accelerometer to detect hand orientation, tilts, and dynamic movements. The glove proved to be intuitive, enabling users to control the prosthetic arm by mimicking natural wrist and finger gestures. The arm was assessed according to important performance measures, such as response time, accuracy, strength, and user comfort. The system delivered a fast response time of less than 100 ms, providing a smooth and natural transition of motion. The accuracy of gesture recognition was 85–90%, depending on sensor calibration and the hand size of the user.

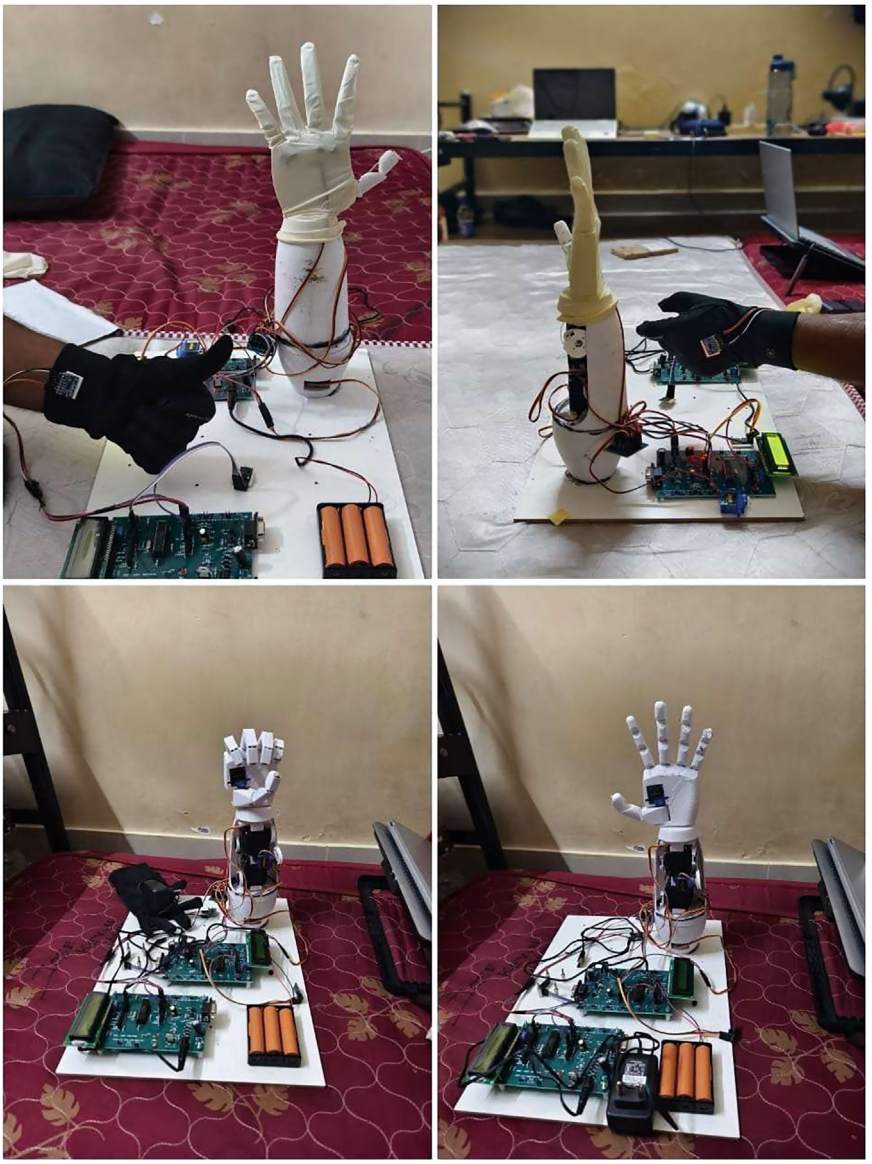


Fig. 4 Prosthetic arm and its various control functions

Power Efficiency and Battery Life: The platform was powered by a 7.4 V 2200mAh Li-Po battery. With constant use (gesture input and servo actuation every 10 s), the platform worked well for approximately 2.5 h. When idle for extended periods, the platform used very little current, significantly extending standby time. Energy usage was greatest with motor actuation and wireless data bursts.



Fig. 5 Controller glows

Figure 6 shows the gesture control feature of the JOVA Robotic Prosthetic Arm. The table translates certain hand or arm movements (gestures) identified by a motion-sensing glove into similar movement outputs executed by the prosthetic hand. Forward motion of the hand translates to the prosthetic executing a closing palm action, practically emulating a gripping motion.

Backward motion initiates the opening palm, enabling the hand to release objects.

Lifting the elbow to the left makes the prosthetic mimic an elbow to left movement, and shifting to the right performs an elbow to right movement.

Each gesture is clearly paired with a real-time response from the robotic arm, demonstrating intuitive, mirror-like control using motion sensors, likely interpreted by an IMU (Inertial Measurement Unit) and microcontroller. This setup enables natural control of the prosthetic without invasive input methods.

One of the greatest achievements of the project is cost-effectiveness. With a production cost of well under traditional prosthetic devices, the JOVA arm provides an avenue for broader accessibility, especially to those in low-resource environments or developing nations where advanced prosthetics are beyond their means. Moreover, the modularity of the design provides for simple maintenance, repair, and upgrade in the future, making it long-term sustainable.

By intense testing, the system exhibited excellent gesture detection accuracy and motion imitation, attaining stable grip functionality with smooth finger movement. The MG996R servo motor-actuated prosthetic hand was found to be able to accomplish basic grasping functions like cup, pen, and small tool holding, rendering it applicable for everyday life. In addition, the 3D-printed design provided lightweight construction, ergonomic comfort, and simplicity of customization, making it possible to adapt the system for various users and hand sizes.

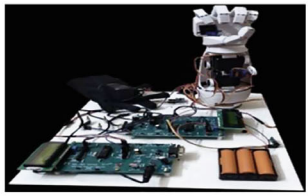
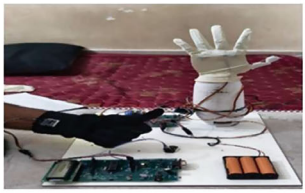
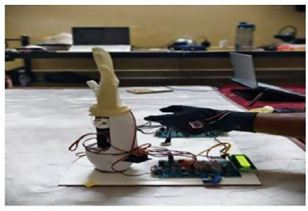
S. No	Gesture	Movement Output
1.	Forward	Closing palm 
2.	Backward	Opening palm 
3.	Left	Elbow to left 
4.	Right	Elbow to right 

Fig. 6 Gesture and its associated movement

5 Future Enhancement

To further enhance the operation, flexibility, and usability of the robotic prosthetic arm system, enhancements can be made with a lot of possibilities. They can be categorized under sensor improvements, control systems, power supply, AI integration, and user interface features. Some suggestions for the work that might be worthwhile exploring in the future are.

- Integration of Advanced Sensors: The system should not only be restricted to sensor flexion but should also integrate EMG-based sensors for muscle actuation detection. The arm will be under the direct control of users through muscle signals naturally and intuitively, instead of through gloves.
- BCI Implementation: EEG signals control prosthetic arms via brain signal processing. Neural signals processed with deep learning models accurately recognize gestures.
- Haptic Feedback Sensors: Force sensors within the prosthetic fingers convey haptic feedback to the user. This gives the user a feel for touch by sending certain feedback mechanisms, such as vibration signals. AI Driven Recognition and Adaptation of Gestures.
- Gesture prediction with machine learning: The system will analyze deep learning algorithms (CNNs and LSTMs, or Transformers) to do an interpretation of the movements that are happening based on real-time information from sensors. Reinforcement learning can be applied to build the process across the training period for accuracy and preferences.
- Personally learned behavior: The system will be able to learn through time with the correlated movements individual to a user, thus improving comfort and efficacy. Transfer learning will be helpful to adapt the stateno propagation to other users.
- AI-Based Anomaly Detection: The prosthetic arm will be able to sense abnormal movements/effects and even self-correct to prevent unintended gestures.
- BLE or Wi-Fi Modules Awareness: Using BLE 5.0 or Wi-Fi communication can provide an improvement in the data transmission span and speed of the connection instead of replacing nRF modules. Low power consumption and realtime responsiveness would ensure a very smooth experience.
- Cloud-Based Prosthetic Control and Monitoring: Prosthetic data can be stored in the cloud for remote diagnostics, performance monitoring, and AI-based improvements. Cloud connectivity would allow telerehabilitation and customization of settings for different users.
- Strengthen Structural Parts: A few finger joints experienced minimal wear in heavy use adopting carbon-reinforced PLA or nylon composites for future prints might enhance longevity.
- Software Suggestions: Gesture Calibration App: Create a mobile/desktop app through which users can calibrate gestures, establish control levels, and track performance
- User Training Programs: Create extensive training programs and materials to assist users in becoming familiar with the prosthetic arm and optimizing its use.
- Gesture Calibration App: Develop a desktops and mobile app through which users can calibrate gestures, set levels of control, and monitor performance. • Community Support and Feedback: Provide a forum for users to exchange experience, offer feedback, and provide improvement suggestions and drive development based on a community-based model.
- AI-Driven Gesture Recognition: Implement machine learning algorithms (CNNs, LSTMs, or Transformers) for interpreting real-time sensor data, enhancing gesture prediction accuracy.

6 Conclusion

This paper presents a gesture-controlled robotic prosthetic arm as an affordable and easy-to-use alternative for upper limb amputees. Accelerometers and an Arduino-based control system translate user gestures into real-time finger movements with high responsiveness and moderate accuracy. nRF-based wireless communication module, enhances portability and convenience for daily applications. Although the proposed model is inexpensive, lightweight, and easy to implement, it does not have adaptive grip strength, haptic feedback, or accuracy in fine motor tasks. Compared to high-end bionic arms with EMG-based control, AI-based adaptability, and sensory feedback, the proposed system provides a simpler and more functional solution for basic prosthetic applications. Researchers could add EMG sensors, AI-based movement prediction, haptic feedback, and IoT connectivity in future versions to improve prosthetic functionality and compete with advanced bionic limbs. This work introduces a non-invasive, gesture-controlled prosthetic arm using low-cost components targeted at underserved populations.

References

1. Starke J, Weiner P, Crell M, Asfour T (2022) Semi-autonomous control of prosthetic hands based on multimodal sensing, human grasp demonstration and user intention. Robot Auton Syst 154:104123, ISSN 0921-8890. https://doi.org/10.1016/j.robot.2022.104123
2. Pyun KR, Kwon K, Yoo MJ, Kim KK, Gong D, Yeo WH, Han S, Ko SH (2024) Machine-learned wearable sensors for real-time hand-motion recognition: toward practical applications. Natl Sci Rev 11(2), nwad298. https://doi.org/10.1093/nsr/nwad298
3. Abd MA, Ingicco J, Hutchinson D et al (2022) Multichannel haptic feedback unlocks prosthetic hand dexterity. Sci Rep 12:2323. https://doi.org/10.1038/s41598-02204953-1
4. Sankar S et al (2025) A natural biomimetic prosthetic hand with neuromorphic tactile sensing for precise and compliant grasping. Sci Adv 11, eadr9300. https://doi.org/10.1126/sciadv.adr9300
5. Jayaweera HKPY (2021) Design and implementation of electromyography (EMG) based real-time pattern recognition model for prosthetic hand control, pp 1–66, OSF Preprints, 2021. https://doi.org/10.31219/osf.io/rd2cf
6. Taniguchi H, Yakami R, Morinaga K, Wakimoto S (2025) Forearm-driven soft prosthetic hand designed for children with the ability to execute intricate and skillful motions. In: 2025 11th international conference on mechatronics and robotics engineering (ICMRE), Lille, France, pp 221–226. https://doi.org/10.1109/ICMRE64970.2025.10976323
7. Jr III, Bugtai N, Baldovino R, Espiritu N, Singson L (2024) Development of an Arduino-based control and sensor system for a robotic laparoscopic surgical unit. Recoletos Multidiscip Res J 12:125–144. https://doi.org/10.32871/rmrj2412.01.10
8. Logesh M (2024) Design and control of medical prosthetic hand using electromyography for easy hand gesture 1–7. https://doi.org/10.1109/ICDSCNC62492.2024.10939419. https://www.researchgate.net/publication/390403656_Design_and_Control_of_Medical_Prosthetic_Hand_Using_Electromyography_for_Easy_Hand_Gesture
9. Kim H, Jang S, Thien Do P, Lee CK, Ahn B, Kwon S, Chang H, Kim Y (2023) Development of wearable finger prosthesis with pneumatic actuator for patients with partial amputations. Actuators 12(12):434. https://doi.org/10.3390/act12120434

10. Roy D, Chowdhury M, Afrose F, Hoque M (2021) Design and development of a cost-effective prosthetic hand for upper limb amputees 1–5. https://doi.org/10.1109/BMEiCON53485.2021.9745210
11. Abbady HEMA, Klinkenberg ETM, de Moel L, Nicolai N, van der Stelt M, Verhulst AC, Maal TJJ, Brouwers L (2022) 3D-printed prostheses in developing countries: a systematic review. ProsthetS Orthot Int 46(1):19–30. https://doi.org/10.1097/PXR.0000000000000057
12. Ahkami B, Ahmed K, Thesleff A, Hargrove L, Ortiz-Catalan M (2023) Electromyography-based control of lower limb prostheses: a systematic review. IEEE Trans Med Robot Bionics 5(3):547–562. https://doi.org/10.1109/tmrb.2023.3282325. Epub 2023 Jun 7. PMID: 37655190; PMCID: PMC10470657
13. Gentile C, Cordella F, Zollo L (2022) Hierarchical human-inspired control strategies for prosthetic hands. Sensors 22(7):2521. https://doi.org/10.3390/s22072521
14. Chenchireddy K (2024) Development of robotic arm control using Arduino controller. IAES Int J Robot Autom (IJRA) 13:264–271. https://doi.org/10.11591/ijra.v13i3
15. Loutan K, Persad U (2023) The design of a low-cost voluntary closing finger prosthetic for developing countries. https://doi.org/10.1115/DETC2023-117072

Assess the Effects to Organs Induced by I-131 Radiopharmaceutical Accumulates in Thyroid Gland by Monte Carlo Simulation

Minh Khanh Duong and Tat Thang Nguyen

1 Introduction

Thyroid cancer is becoming one of the global health concerns. According to the latest data from GLOBOCAN 2022 [1], thyroid cancer has become the seventh most common cancer worldwide, with 821,214 new cases, accounting for approximately 4.1% of all cancer cases globally. The number of deaths caused by this disease reached 47,507, representing a mortality rate of 0.49%. In particular, in Vietnam, thyroid cancer is also among the most common cancers. In 2022 alone, 6122 new cases were reported, accounting for 3.4% of the total cancer cases in the country, making it the sixth most prevalent cancer nationwide (GLOBOCAN, 2022). Accompanying this is the number of deaths due to thyroid cancer, which reached 858 cases, representing a mortality rate of 0.71%. A study by Pham et al. [2] investigating thyroid cancer incidence in Ho Chi Minh City (Vietnam) across four consecutive 5-year periods 1996–2000, 2001–2005, 2006–2010, and 2011–2015 based on 5953 thyroid cancer cases, found that the highest incidence was recorded at the age of 45 in both sexes, while the lowest incidence occurred in those aged 65 and above. Therefore, the investigation and treatment of thyroid cancer in adults is of critical importance.

Currently, the treatment of thyroid cancer using the radioactive isotope Iodine-131, also known as RAI (Radioactive Iodine), is the most widely used therapy worldwide, given that thyroid cancer is among the most common types of cancer. The principle of this method is that when radioactive isotope I-131 decays to return to a stable state, it emits beta β^- radiation, and this process is also accompanied by gamma γ radiation. When administered to a patient, I-131 is strongly absorbed by thyroid tissue due to the biological affinity between thyroid cells and iodine. The ionizing radiation emitted and co-emitted during the decay of I-131 destroys residual microscopic thyroid tissue

M. K. Duong (✉) · T. T. Nguyen
Hanoi University of Science and Technology, Hanoi, Vietnam
e-mail: Khanh.DM217520@sis.hust.edu.vn

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_33

after surgery, aiming to eliminate all cancerous cells remaining in the thyroid region, and potentially even metastatic cells that have spread to surrounding organs [3].

However, treatment with radiopharmaceuticals also poses both short-term and long-term health risks for patients. With an Iodine-131 activity ranging from 101 to 150 mCi, the risk of developing leukemia can be as high as 3.09% [4]. In addition, common short-term clinical symptoms following oral administration of the radiopharmaceutical include swelling, pain, and dysfunction of the salivary glands, as these glands frequently absorb and excrete I-131. A further late complication observed in patients undergoing RAI treatment is xerostomia (dry mouth), which often becomes more pronounced after hospital discharge. Gastrointestinal disturbances have also been commonly reported among patients treated with I-131 [5]. These concerns highlight the need for rational use of radiopharmaceutical activity to minimize side effects in patients. Therefore, investigating radiation dose levels as safety references for RAI treatment is an essential issue in nuclear medicine.

Although there is some research groups executed measurements on patients' bodies [6], provided some measurement systems and even the dosimetry considerations, associated risks, and the prevailing challenges in measurement domains [7], and several research groups focus on synthesizing and exploring supportive options for personalized treatment [8] the assessment cannot always be executed directly on patients' bodies because of the resolution, the geometric error of the measurements and even moral problems. And for scientific research purposes, many hypothetical scenarios involving irradiation require computational analysis. Therefore, direct measurement is not always a feasible solution, and it becomes necessary to use an alternative object instead of the human body for studying. There are many virtual phantoms introduced, ICRP Publication 145 [9] introduced a human body-representative phantom composed of multiple tetrahedral elements, known as the MRCP (Mesh-type Reference Computational Phantom). This phantom has become a standard reference model for research and dosimetry calculations in various simulation programs, serving as an alternative to actual human bodies in computational studies. In this paper, the computational phantoms that presented for adult males and females are used for assessment.

In internal dose calculation, the "S" value was determined as the as the amount of absorbed radiation dose delivered to the target organ or tissue resulting from a single disintegration of the radionuclide occurring in the source organ or tissue [10]. Then the values of internal dose are performed by multiplying the "S" value with the radioactivity accumulated in target organs. In this research, the Monte Carlo simulation is used for estimating the "S" value with the source (iodine 131) located in thyroid gland. In clinical practice, patients often undergo surgery to partially or completely remove the thyroid gland. Corresponding to these cases, the internal radiation dose to the thyroid and other organs will also differ. This highlights the need for further research and dose calculations tailored to these specific clinical scenarios.

2 Related Works

Several studies have utilized Monte Carlo software programs to calculate radiation doses from internal exposure to I-131 sources [11–14]. However, most of the aforementioned research groups only used and reported results from a single Monte Carlo simulation program or conducted their studies on voxel phantoms. A type of phantom with lower resolution and less detailed organ contours compared to MRCP phantoms. Although the results researched on pediatric phantom of Lee et al. [13] had provided various values of different of 30 organs, their study mainly focused on pediatric subjects under 15 years old. Moreover, the research was conducted using voxel phantoms with a simpler algorithm compared to calculations performed on tetrahedral phantoms [9].

In the study by Shahbazi-Gahrouei and Ayat [11], the absorbed dose values were calculated and compared between simulations and actual measurements using TLD detectors. However, mathematical phantoms were used, which are constructed from simple geometric shapes, with organs represented by cylinders, ellipsoids, tori, etc. Moreover, important radio sensitive organs have not been validated in the study such as liver, lung, stomach, kidney, …Therefore, it is essential to conduct a comparison between different Monte Carlo software packages, with evaluations across various organs in the human body, to provide the most comprehensive and complete reference possible.

The study by Juste et al. [12] provided time-dependent dose results based on the standard ICRP phantom; however, these results were calculated using MCNP and their own biological model, without comparison to other Monte Carlo tools.

3 Proposed Work

This study aims to contribute to health science research by minimizing patient risk through simulation-based dose estimation. Specifically, it focuses on determining organ-specific "S" values resulting from the accumulation of Iodine-131 in the thyroid gland, which serves as the foundation for calculating internal radiation doses in the human body. To achieve this, the study utilizes the standard Mesh-Reference Computational Phantoms (MRCP) published by the ICRP [9], which feature anatomically detailed and realistic organ geometries, thereby overcoming the limitations of direct dose measurements. Radiation transport simulations are performed using the Monte Carlo method implemented in two widely used software packages: MCNP (version 6.1.1) and GEANT4. Three clinical scenarios are considered: (1) total thyroidectomy, (2) only the left lobe remaining, and (3) only the right lobe remaining. The resulting "S" values for nearly 30 organs are presented and cross-compared between the two codes. These values are intended to provide clinically relevant reference data for thyroid cancer treatment using radioiodine (RAI),

while the comparison between the two Monte Carlo tools enhances the reliability and credibility of dose calculations based on virtual phantoms.

4 Methodology

4.1 Mesh-Type Reference Computational Phantom (MRCP Phantom)

The MRCP phantom is constructed using tetrahedral elements in three-dimensional space. These MRCP phantoms are developed by converting voxel-based phantoms into high-quality mesh formats, integrating supplementary stylized models that are used alongside voxel phantoms to overcome the resolution limitations of voxels, and to incorporate additional tissue layers believed to contain cells at risk of radiation-induced cancer [9].

The MRCP phantoms significantly overcome the geometric limitations found in the ICRP voxel phantoms, which stem from the restricted voxel resolution and the inherent characteristics of voxel-based geometry. While the voxel-based models present stepped or jagged surfaces, the mesh-based MRCP models exhibit smoother, more continuous contours in 3D visualization. Additionally, the discontinuities in the hollow organ structures observed in the voxel phantoms have been fully resolved in the MRCP models [9].

The calculations in this study are performed on adult male and female MRCP phantoms, representing the age group most commonly affected by thyroid cancer (Fig. 1). The male phantom is 1.76 m tall, weighs 73 kg, and is composed of 8.2 million tetrahedral elements. The female phantom is 1.63 m tall, weighs 60 kg, and consists of 8.6 million tetrahedra. Figure 1 shows the MRCP Phantoms performed by MCNP6.

4.2 Calculated Organs

This study investigates the S value in several human organs, including those that are highly exposed to radiation through the metabolism and excretion processes of Iodine-131, such as the salivary glands, esophagus, liver, stomach, kidneys, and bladder. The S value provides information on the energy absorbed per unit mass of the target organ, which can be used to estimate the specific radiation dose delivered to that organ. All organs are calculated in this research are: Active bone marrow/ Red bone marrow (RBM), Colon, Lung, Stomach, Breasts, Gonads, Urinary Bladder, Esophagus, Liver, Thyroid, Endosteum, Brain, Salivary Gland, Skin, Adrenals, ET (Extra-thoracic), Gall Bladder, Heart, Kidneys, Lymph nodes, Muscle, Oral mucosa, Pancreas, Prostate/Uterus, Small intestine, Spleen, Thymus, Eye Lens. As described



Fig. 1 MRCP phantoms performed by MCNP6 (Left: male phantom, Right: female phantom)

above, in realistic clinical cases, although patients almost undergo surgery to eliminate thyroid glands, there still remains the residual thyroid issue. Thus, in this study, the "S" value is also calculated in thyroid. In this research, the results of "S" would be performed in 3 cases of both genders: AM-AF CR (Completely Removal), which are represented for patients whose thyroid glands are Completely Resected, AM-AF Left/Right which are represented for patients whose thyroid glands remain the left/right side part.

4.3 Calculation Models

I-131 primarily emits beta-minus (β^-) particles along with accompanying gamma (γ) radiation. However, in this simulation, the accompanying γ emission at 164 keV from ^{131m}Xe to the stable ^{131}Xe state (with a probability of 0.021%) is neglected. Therefore, the source parameters of I-131 used in the simulation are defined as follows in Table 1.

For GEANT4, virtual detectors are set at the positions (organs) to be investigated using G4VUserDetectorConstruction. The source will be placed at the source organ, the thyroid, as defined by G4VUserPrimaryGeneratorAction. In this problem, both the source and the detectors are tissues created in the phantom file, which are placed outside the layers. When used, they will be called from the corresponding classes. The result produced by the GEANT4 program is the value of SAF—the specific absorption factor of the tissue for the source (1/kg) per decay particle, defined as formula (1).

$$SAF_{(target \leftarrow source)} = \frac{E_{dep}}{E_S} * \frac{1}{m_{organ}} \left(\frac{1}{\text{kg}} \right) \tag{1}$$

With E_{dep} being the deposited energy at the target organ and E_S is the energy emitted from the source (thyroid gland). Each energy level will be simulated separately to obtain its corresponding SAF value. After that, the S value is calculated based on the results of GEANT4 through the formula (2).

$$S_{(target \leftarrow source)} = SAF_{(target \leftarrow source)} \left(\frac{1}{\text{kg}} \right) * E_i(\text{MeV}) * N_i * 10^{-3} \tag{2}$$

With E_i being the ith decay energy of I-131 and N_i is the corresponding decay probability of E_i.

For the MCNP6.1.1 code, the + F6 tally (with units of MeV/g) is used to calculate the energy deposited in cell by ionizing radiation per particle [15]. At the same time, to simulate the full range of nuclear interactions of electrons (β^-) and photons, the physics card "mphys" is enabled by default. Similar with GEANT4 code, the MCNP simulation is also performed with a number of particles (nps) equal to $2*10^8$ to ensure that the recommended relative error of the + F6 tally is below 0.1. The obtained result

Table 1 Decay table of I-131

Energy (keV)		Probability of decay N_i (%)
γ	284	6.45
	364	83.1
	637	7.15
β	606	7.2
	334	89.4

represents the dose per particle (MeV/g/particle), which corresponds to the S value used in the dose assessment report.

5 Results

5.1 Results of "S" Value from MCNP and GEANT4

Based on the results obtained from MCNP (Table 2), the "S" values of the AM are lower than AF phantoms in most organs in the CR case, except for the following: Colon (ratio 2.5147), Stomach (1.3049), Small Intestine (1.3991), Thyroid (1.3937), Pancreas (1.2432), Spleen (1.1132), and Thymus (1.0345). In contrast, the remaining organs show ratios smaller than 1.0. All ratios are defined as the "S" value of the AM phantom divided by that of the AF phantom; therefore, the "S" values for most organs are higher in the AF phantom than in the AM phantom.

Similarly, in Left residual Thyroid tissue and Right residual Thyroid tissue cases, "S" value of AM is lower than AF phantoms in most organs in these cases. But the results at Colon, Lung, Stomach, and Small Intestine indicate significant difference between AM and AF phantom at which the "S" values are contrary to the aforementioned trend. Especially for Lung in Right residual thyroid tissue case, the result of AM are approximately 85 times higher than AF.

The results obtained from GEANT4 are performed in Table 3. The different trends between AM and AF are analogous to the presents of MCNP. "S" value of AF phantom is higher than AM phantom at almost organs in all 3 cases, but there still remain exception at Colon, Lung, Pancreas, Urinary Bladder, Small intestine, Spleen at which the "S" value of AM is higher shown in below Tables 2 and 3. In details, the "S" value at Thyroid of CR case of AM phantom is higher than AF phantom (ratio 1.39), same as it in MCNP.

5.2 Comparison Between MCNP and GEANT4

To facilitate the comparison of results between MCNP and GEANT4 across all organs, the S values are converted to logarithmic scale due to the large exponential differences among the results in different organs. Figures 2, 3, 4, 5, 6 and 7 present the comparison of the results from the two codes for both male and female phantoms in all three cases.

In the case of calculation with the phantom which presents for patient who having the thyroid completely removed, the results from the two codes show a relatively good agreement for both the AM and AF phantoms.

In the case of the remaining calculation on the left thyroid tissue, the two software programs generally produce consistent results for most organs in the AM phantom.

Table 2 Results of "S" value in some organs of male and female bodies calculated by MCNP (CR: Completely Removal; AM: Male; AF: Female)

$S_{(target \leftarrow source)}$ of some organs- calculated by MCNP						
Organ	AM CR	AF CR	AM left	AF left	AM right	AF right
RBM	1.51E−06	2.25E−06	1.52E−06	7.97E−07	3.27E−06	3.96E−06
Colon	7.72E−08	3.07E−08	1.12E−07	3.04E−08	1.26E−07	4.30E−08
Lung	2.31E−06	2.88E−06	5.00E−06	1.15E−06	4.64E−06	5.42E−08
Stomach	3.21E−07	2.46E−07	5.34E−07	2.15E−07	2.89E−07	2.74E−07
Breasts	3.88E−07	1.18E−06	4.61E−07	9.07E−07	1.42E−07	4.11E−07
Gonads	7.42E−10	5.66E−09	2.23E−10	2.25E−09	1.65E−09	1.10E−08
Urinary Bladder	6.52E−09	5.82E−09	6.52E−09	7.16E−09	5.53E−09	6.70E−09
Esophagus	4.61E−05	2.06E−04	1.92E−05	6.54E−05	3.59E−05	1.43E−04
Liver	3.97E−07	4.16E−07	3.82E−07	7.83E−07	4.42E−07	2.54E−05
Thyroid	4.85E−01	3.48E−01	1.90E−02	3.84E−02	3.08E−02	3.77E−02
Endosteum	7.14E−07	1.28E−06	4.57E−07	8.12E−07	3.31E−07	1.25E−06
Brain	4.30E−07	6.42E−07	4.15E−07	1.33E−06	4.34E−07	6.69E−07
Salivary Gland	2.65E−06	3.99E−06	5.01E−06	7.72E−06	5.49E−06	8.74E−06
Skin	3.93E−07	4.60E−07	3.71E−07	7.93E−07	3.89E−07	5.03E−07
Adrenals	2.50E−07	2.60E−07	5.50E−07	5.03E−07	5.00E−07	9.35E−07
ET	6.37E−06	7.10E−06	6.56E−06	6.62E−06	6.27E−06	6.85E−06
Gall Bladder	2.25E−07	2.34E−07	2.03E−07	4.40E−07	2.49E−07	6.82E−06
Heart	1.74E−06	1.91E−06	1.91E−06	3.95E−06	1.66E−06	1.99E−06
Kidneys	1.49E−07	1.42E−7	8.78E−08	1.55E−07	5.06E−07	3.35E−07
Lymph nodes	1.07E−06	1.38E−06	1.08E−06	4.95E−06	2.63E−06	2.80E−06
Muscle	6.34E−07	7.84E−07	7.97E−07	1.78E−06	6.50E−07	1.42E−06
Oral mucosa	2.50E−06	4.78E−06	4.62E−06	5.26E−06	5.35E−06	1.11E−05
Pancreas	1.84E−07	1.48E−07	1.97E−07	2.89E−07	1.92E−07	1.45E−07
Prostate/Uterus	3.50E−09	5.40E−09	4.45E−09	1.07E−08	3.00E−09	5.96E−09
Small intestine	6.03E−08	4.31E−08	6.57E−08	3.24E−08	5.60E−08	4.21E−08
Spleen	3.54E−07	3.18E−07	4.42E−07	7.03E−07	3.21E−07	4.19E−08
Thymus	2.40E−05	2.32E−05	2.32E−05	4.58E−05	2.00E−05	2.32E−05
Eye Lens	4.06E−07	5.83E−07	2.90E−07	5.03E−07	5.57E−07	9.62E−07

However, there are relatively significant differences in the "S" values for the lungs, gonads, salivary glands, adrenals, oral mucosa, and kidneys. These discrepancies are even more pronounced in the female phantom, with only the endosteum showing a high level of agreement between the two programs. All other organs exhibit varying degrees of difference, the most notable being in the lymph nodes.

Result consistency in the case where the source is placed in the remaining right thyroid tissue is observed in many radiosensitive organs. However, clear differences

Table 3 Results of "S" value in some organs of male and female bodies calculated by GEANT4 (CR: Completely Removal; AM: Male; AF: Female)

$S_{(target \leftarrow source)}$ of some organs- calculated by GEANT4						
Organ	AM CR	AF CR	AM left	AF left	AM right	AF right
RBM	1.44E−06	1.86E−06	1.48E−06	1.72E−06	1.39E−06	1.72E−06
Colon	7.48E−08	2.71E−08	7.81E−08	2.67E−08	7.29E−08	2.53E−08
Lung	2.22E−06	2.85E−06	2.38E−06	2.73E−06	2.28E−06	2.67E−08
Stomach	2.96E−07	2.38E−07	3.28E−07	2.49E−07	2.65E−07	2.49E−07
Breasts	3.58E−07	1.15E−06	3.79E−07	1.14E−06	3.53E−07	2.57E−07
Gonalds	7.15E−10	5.58E−09	6.73E−10	5.50E−09	8.36E−10	5.50E−09
Urinary Bladder	6.37E−09	5.82E−09	6.32E−09	5.74E−09	6.64E−09	5.74E−09
Esophagus	4.48E−05	2.05E−04	1.95E−05	4.62E−05	1.16E−05	4.62E−05
Liver	3.89E−07	4.16E−07	3.67E−07	3.78E−07	4.25E−07	2.44E−05
Thyroid	4.82E−01	3.48E−01	1.82E−02	3.34E−02	2.72E−02	3.34E−02
Endosteum	6.46E−07	8.75E−07	6.63E−07	8.31E−07	6.29E−07	8.31E−07
Brain	4.21E−07	6.42E−07	3.96E−07	6.43E−07	4.19E−07	6.43E−07
Salivary Gland	2.54E−06	3.95E−06	2.44E−06	4.16E−06	2.63E−06	4.16E−06
Skin	3.70E−07	4.51E−07	3.51E−07	4.66E−07	3.61E−07	4.66E−07
Adrenals	2.40E−07	2.57E−07	2.49E−07	2.42E−07	2.49E−07	4.65E−07
ET	6.30E−06	6.83E−06	6.23E−06	6.78E−06	6.32E−06	6.78E−06
Gall Bladder	2.20E−07	2.34E−07	2.02E−07	2.12E−07	2.34E−07	6.43E−06
Heart	1.70E−06	1.91E−06	1.82E−06	1.91E−06	1.59E−06	1.91E−06
Kidneys	1.32E−07	1.35E−07	1.37E−07	1.28E−07	1.37E−07	1.28E−07
Lymph nodes	1.02E−06	1.14E−06	7.43E−07	1.12E−06	1.02E−06	1.12E−06
Muscle	6.22−07	7.84E−07	7.65E−07	1.17E−06	8.27E−07	1.17E−06
Oral mucosa	2.39E−06	4.72E−06	2.16E−06	4.70E−06	2.26E−06	4.70E−06
Pancreas	1.80E−07	1.48E−07	1.83E−07	1.39E−07	1.85E−07	1.39E−07
Prostate/ Uterus	3.41E−09	5.40E−09	3.29E−09	5.18E−09	3.52E−09	5.18E−09
Small intestine	5.56E−08	4.14E−08	5.84E−08	4.05E−08	5.84E−08	4.05E−08
Spleen	3.47E−07	3.18E−07	4.22E−07	3.41E−07	3.07E−07	3.99E−08
Thymus	2.35E−05	2.32E−05	2.22E−05	2.22E−05	1.92E−05	2.22E−05
Eye Lens	4.02E−07	5.66E−07	4.17E−07	6.41E−07	3.61E−07	6.41E−07

still exist in the esophagus, kidneys, lymph nodes, and oral mucosa. This observation is consistent in both AM and AF phantom comparisons.

The discrepancy between the results of MCNP and GEANT4 may be attributed to the differences in their calculation methods. Although both programs use the Monte Carlo algorithm, GEANT4 calculates the SAF values separately for each organ at different energy levels, and an intermediate step is required to convert these

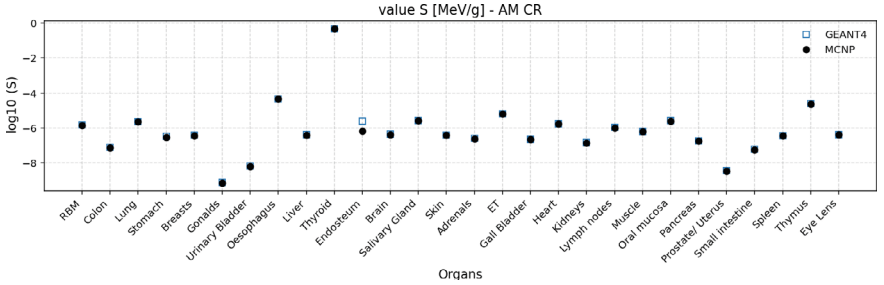


Fig. 2 Comparison between MCNP and GEANT4 (AM CR)

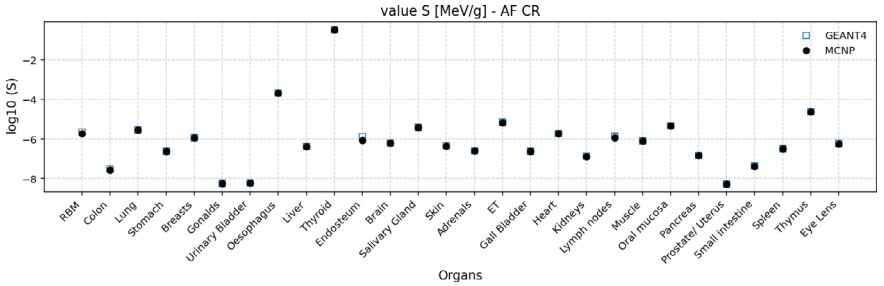


Fig. 3 Comparison between MCNP and GEANT4 (AF CR)

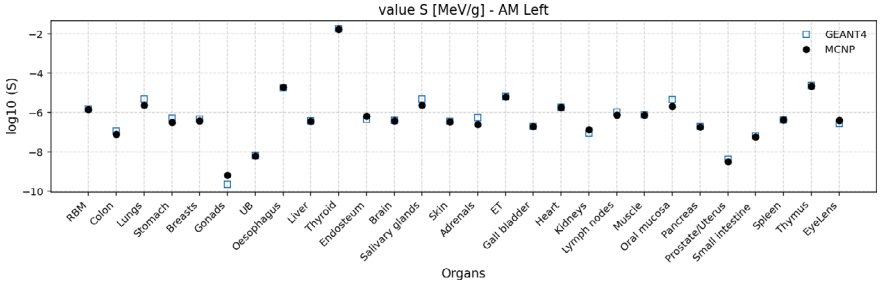


Fig. 4 Comparison between MCNP and GEANT4 (AM left)

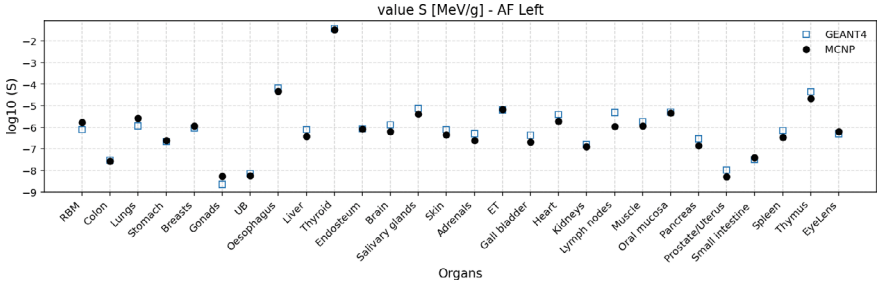


Fig. 5 Comparison between MCNP and GEANT4 (AF left)

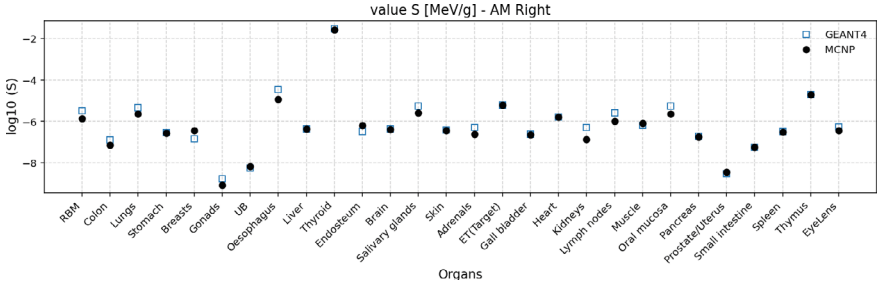


Fig. 6 Comparison between MCNP and GEANT4 (AM right)

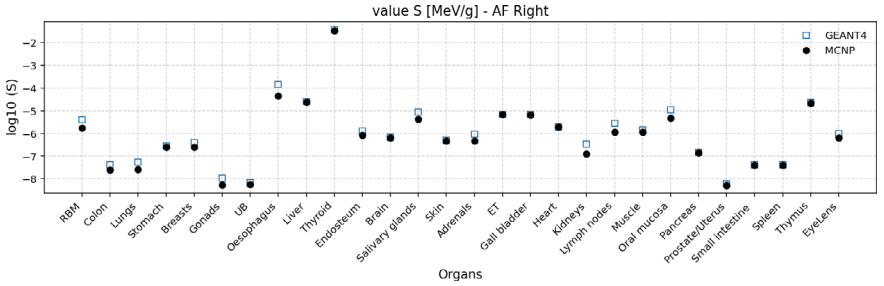


Fig. 7 Comparison between MCNP and GEANT4 (AF right)

values into the final "S" values. In contrast, MCNP allows for the direct specification of source parameters, including energy levels and probabilities, enabling direct computation of the "S" values.

In addition, it can be observed that the differences between the two software programs mainly occur in organs that are located farther from the thyroid, such as the gonads, kidneys, adrenals, and pancreas, or in complex structures composed of multiple components, such as red bone marrow, lymph nodes, and the colon. This can be explained by the fact that statistical fluctuations tend to be greater in organs farther from the source, leading to more pronounced discrepancies. Furthermore, for organs with multiple components, variations in the results for each sub-component can lead to significant overall differences when evaluating the entire organ. For instance, in the case of the colon, the calculation is divided into six main parts: the ascending colon, transverse colon, descending colon, sigmoid colon, rectum, and internal contents. Therefore, such differences in large and complex organs are inevitable. To address this issue, the number of simulated particles can be increased to reduce statistical uncertainty. However, this comes at the cost of increased simulation time and memory (RAM) usage. We suggest that the number of particles used for the simulation could be in the order of several 10^9 particles.

6 Conclusions

In this study, we used standard MRCP phantoms to calculate and assess the "S" values instead of performing direct measurements on multiple specific organs. The Monte Carlo algorithm is sufficiently reliable for computing problems related to radiation interactions. The calculated "S" values serve as the foundation for estimating internal radiation doses to various organs. In this study, we employed the Monte Carlo algorithm to compute the "S" values using two representative software packages: MCNP6.1.1 and GEANT4. The study was conducted on standard male and female MRCP phantoms published by the ICRP, and reference tables of "S" values were provided for two clinical scenarios: (1) patients who have undergone total thyroidectomy, and (2) patients who have undergone partial thyroidectomy, retaining only the left or right lobe of the thyroid. The results from both MCNP and GEANT4 show that the "S" values for the phantoms in the three scenarios differ significantly. Even the cases where only the right or only the left thyroid lobe remain produce different outcomes. Therefore, one case cannot be used as a representative or equivalent substitute for the other. The "S" value tables from both MCNP and GEANT4 across all three scenarios consistently show that the "S" values in the female phantom are generally higher than those in the male phantom for most organs, except for the colon and small intestine (in all three cases). Notably, in the case where only the right thyroid lobe remains, the "S" value in the lungs of the male phantom is approximately 85 times greater than that of the female phantom. Additionally, for the spleen, the "S" value in the male phantom is nearly 8 times higher than that in the female phantom. Moreover, the calculated values between MCNP and GEANT4 show relatively high consistency for radiosensitive or simpler organs. However, there are differences in the results for organs located farther from the source (in this study, the thyroid) and for organs calculated through multiple smaller components. Overall, given the accuracy of the Monte Carlo algorithm with strict statistical validation (relative error below 0.1), the "S" values calculated by these two software programs across the three clinical scenarios can serve as reference data to support and compare treatment planning for thyroid cancer using the radiopharmaceutical Iodine-131 (RAI). Thus, this study contributes comprehensive "S" values for nearly 30 different organs in the human body across three thyroidectomy scenarios, calculated using the Monte Carlo method. These values were cross compared between MCNP and GEANT4, demonstrating a fair degree of agreement between the two simulation programs.

References

1. Cancer (2022) International Agency for Research on Cancer, "GLOBOCAN 2022," The Global Cancer Observatory (GCO)
2. Pham DX, Nguyen HD, Phung AHT, Bui TD, Tran TS, Tran BNH, Ho-Pham LT, Nguyen TV (2021) Trends in incidence and histological pattern of thyroid cancer in Ho Chi Minh City, Vietnam (1996–2015): a population-based study. BMC Cancer 21(1)
3. Nostrand DV (2009) The benefits and risks of I-131 therapy in patients with well-differentiated thyroid cancer. Thyroid 1381–1391
4. Seo GH, Cho YY, Chung JH, Kim SW (2015) Increased risk of leukemia after radioactive iodine therapy in patients with thyroid cancer: a nationwide, population-based study in Korea. Thyroid 25(8):927–934
5. Lu L, Shan F, Li W, Lu H (2016) Short-term side effects after radioiodine treatment in patients with differentiated thyroid cancer. Biomed Res Int PMID: 26989683(PMCID: PMC4773544)
6. Kääriä L, Lapela M, Seppänen M, Högerman M, Ruohola J, Ålgars A, Noponen T (2024) Determination of effective half-life of 131I in thyroid cancer patients using remote doserate meter. EJNMMI Phys 101(11)
7. Omenka OS, Aziz NHA, Adavize MH, Olaoluwa OA, Godwin E (2024) Thyroid cancer: Iodine-131 intervention and dosimetric analysis: a review. IJARSCT 4(2)
8. Pacil M, Conte M, Frantellizzi V, De Feo MS, Pisani AR, Marongiu A, Nuvoli S, Rubini G, Spanu A, De Vincentis G (2022) Personalized dosimetry in the context of radioiodine therapy for differentiated thyroid cancer. Diagnostics 12(7)
9. Kim CH, Yeom YS, Petoussi-Henss N, Zankl M, Bolch WE, Lee C, Choi C, Nguyen TT, Eckerman K, Kim HS, Han MC, Qiu R, Chung BS, Han H, Shin B (2020) Adult mesh-type reference computational phantoms. ICRP Publication 145. Ann ICRP 49, SAGE
10. Snyder WS, Ford MR, Warner GG, Watson EE (1975) MIRD Pamphlet No. 11 (Revised): "S" absorbed dose per unit cumulated activity for selected, New York
11. Daryoush Shahbazi-Gahrouei SA (2015) Determination of organ doses in radioiodine therapy using Monte Carlo simulation. World J Nuclear Med 14(1)
12. Oliver S, Cases J, Juste B, Miro R, Verdú G (2024) Determination of organ doses in thyroid treatments with radioactive iodine by Monte Carlo simulation. Radiat Phys Chem 224(111737)
13. Lee SM, Choi C, Choi JW, Min H, Ko S, Shin B, Yeom YS (2024) Implication of ICRP pediatric reference voxel phantoms on dose assessment of patients in radioiodine therapy. Nuclear Eng Technol
14. Wattar H (2022) Calculating the internal dose of Radioiodine I-131 by opendose software based on Monte Carlo simulation. SSRN, Wattar, Hala, calculating the internal dose of Radioiodine I-131 by open dose software based on Monte Carlo
15. Wilcox T (2014) MCNP6 advanced tallies tutorial, Knoxville, Tennessee, United States: LANL, 2014-09-14

Voltage Regulation in a 33 Bus Distribution System with Integrated STATCOM and Grey Wolf Optimizatıon



B. Balaji, P. Hemalatha, J. Chandru, and R. Preetha

1 Introduction

Power is generated at the power station where electricity, which is electrical energy, generated at the power station is transmitted through the transmission system and distributed to the end users such as residential, commercial, agricultural, industrial, etc. The simplicity of radial or weakly meshed structure and unidirectional power flow makes such existing distribution networks sufficiently simple and easy to operate through. The advantages of Radial Distribution Networks (RDNs) include low implementation cost, easy protection scheme, and easy voltage control and power flow control [1]. However, RDNs have a highly resistant-to-reactance (R/X) ratio, nonlinear loading, and limited structural capabilities, leading to significant operational challenges. These challenges include voltage instability, poor power factor, low voltage profiles, and higher power losses [2]. Therefore, RDNs are typically linked to lower reliability, reduced security, and increased cost of operation. However, in order to tackle these technical and economic issues, researchers have proposed

B. Balaji (✉) · J. Chandru
Department of EEE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India
e-mail: drbalajib@veltech.edu.in

J. Chandru
e-mail: chandruappu052004@gmail.com

P. Hemalatha
Department of CSE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India
e-mail: drhemalathap@veltech.edu.in

R. Preetha
Karpagam Academy of Higher Education, Coimbatore, India
e-mail: preetha.rajaram@kahedu.edu.in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_34

several enhancement strategies. The commonly used approaches are DG (Distributed Generations) Allocation, Distribution Network Reconfiguration, and Placement of Reactive Power Compensation Devices (Series Voltage Regulators, Shunt Capacitors, D-FACTS devices, i.e., DSTATCOM (Distribution Static Compensator) [3–7]. One such solution is the integration of distributed generation (DG), an effective solution that helps reduce power losses, among others. Nonetheless, DG systems have higher capital and operation costs than other solutions, such as network reconfiguration, shunt capacitor banks, and DSTATCOM, which generate good efficiency [8]. Unlike network reconfiguration, which changes the topology of the power system to improve the power system performance, shunt capacitors and DSTATCOMs inject reactive power into the system in order to improve the voltage profile as well as system stability. However, a major deficiency in the utilization of traditional shunt capacitors is their incapacity to accommodate functional load fluctuation since they insert reactive power in predefined steps [9]. In the last several decades, there has been important progress in control and regulation theory. Yet, the proportional–integral–derivative (PID) controller continues to be of main importance for industrial applications. The PID controller remains widely popular due to its simplicity in design, ease of understanding and operation and intrinsic robustness and stability in a variety of operating conditions [10]. Proportional–integral (PI) controllers are widely used in most control systems to regulate process variables like motor speed and position. The two most important parameters of these controllers are the proportional gain (Kp) and the integral gain (Ki), which decide the dynamic behaviour of the system. A number of methods have been created for optimally tuning these parameters [11]. Metaheuristic search methods are particularly useful when dealing with complex problem domains, whereas classical methods prove ineffective. Many of the metaheuristic approaches have been shown to be effective in tuning PI controllers. Particle Swarm Optimization (PSO) derived from the social dynamics observed in bird flocks is one such method. However, in PSO, a collection of particles is utilized to adjust their positions to search for the best solution under the guidance of the collective experience [12]. A widely used method is the Genetic Algorithm (GA), which simulates natural selection in the form of population evolution of candidate solutions by genetic operations like crossover and mutation [13].

Another algorithm similar to the bee algorithm is Ant Colony Optimization (ACO) [14]. Oppression Op is defined as a modified max–min ant system. The Artificial Bee Colony (ABC) algorithm, inspired by the mechanism of the food-foraging behaviour of honeybees, employs a swarm of artificial bees to find a good solution [15]. Another well-known method is the Grey Wolf Algorithm (GWA), which models the hierarchy of leadership and hunting patterns of grey wolves in nature [16]. GWA is a group of wolves that collectively search for the best solution according to natural hunting behaviours. GWA and PSO are well-known and established metaheuristic algorithms based on swarm intelligence. The difference between these two lies in the fact that GWA mainly explores its search space at the early time of optimization and gradually uses exploitation until the last stage. At the same time, PSO maintains a balance between exploration and exploitation. Additionally, in regard to some problems, GWA converges faster than PSO and for solving problems with a higher dimensional

search space, PSO is more efficient. In GWA, grey wolf positions are considered as the solution, while in PSO, a solution is measured using the particle locations. The initialization of GWA needs the setup of a wolf population, while that of PSO relies only on the initial particle and swarm. Without a doubt, therefore, the selection between GWA and PSO primarily depends on the problem's nature and the problem requirements [17, 18]. The rise of renewable integration into contemporary electrical systems makes DC microgrids a popular choice of infrastructure due to efficient operation and DC-based sources, as well as storage device compatibility. However, voltage stability under dynamically varying loads and production profiles becomes a challenge. Usually, tough control methods depend on the use of PI controllers. However, the traditional designs thereof might not be the best choice for nonlinear time-varying conditions, and thus, the use of metaheuristic-based tuning has become increasingly popular. In particular, among these, Grey Wolf Optimization has been considered for its simplicity, good balance between exploration and exploitation, and good convergence. In [19], GWA was proposed using leader hierarchy and hunting mechanism discovered in grey wolves to solve complex optimization problems effectively. After this, GWA has been widely used to tune a controller in power system operations. As an example, [18] proposed a Chain Search Algorithm (CSA) that took advantage of geometric information in setting metaheuristic optimization parameters and compared it to GWA. One such algorithm is the Grey Wolf Algorithm (GWA), which is widely studied due to its ease, balance in exploration and exploitation and fast convergence. The study in [19] presented GWA by using the grey wolf leadership hierarchy and search for hunting to address complex optimization problems efficiently. Since that time, GWA has been widely applied in the tuning of controllers for power system operation. As an example, the Circle Search Algorithm (CSA) [18] introduced geometric intuition into metaheuristic optimization, and compared its performance with GWA. Microgrid energy management applied [20] GWA for hybrid system operation optimization with significant cost reduction and emission minimization. Hybrid methods integrating PI controllers and GWA-based tuning have also been investigated for DC microgrids. Studies have proven that the application of GWA-tuned PI controllers on DC-DC converters resulted in better transient responses and minimal steady-state error than the traditional Ziegler–Nichols method [21]. GWA was also utilized successfully for load frequency control of hybrid renewable systems with improved robustness against system perturbances and parameter uncertainties [22]. In [23], a notable application was given that proposed a system consisting of a brushless DC motor, a Buck-Boost DC converter and supercapacitors and optimized control strategy was used for efficient energy storage. They hope this advancement will improve the regulation of energy in an electric vehicle and extend the distance energy can be utilized on a charge cycle. Moreover, FOPID-FOPID and PR-PR-based DC microgrid structures were simulated and analyzed in MATLAB/Simulink, and the results reported in [24] provided insight into the performance of various controller types in a closed loop. At last, the work in [25] concentrated on decreasing the output ripple and adjusting the output voltage in High Step-Up Super Lift DC-DC converters via PI and FOPID controllers. Performance of the system with settling time and steady-state error were compared, and it was found that the

FOPID system configuration with better results, hence an important initiative for closed loop system feedback optimizing purposes.

Based on the literature review, to enhance system stability during transient conditions, closed-loop controllers such as the PIC and the MPC, both trained using the GWA algorithm, are incorporated. The primary objective of this work is to improve the system's stability. To achieve this, the scope of the work includes implementing two configurations: a GWA-PI controlled closed-loop STATCOM integrated into a 33-bus system, and a GWA-MPC controlled closed-loop STATCOM within the same system. Initially, the system is also analysed without any controllers to provide a comparative assessment.

2 System Description

As discussed previously, the proposed system is configured as a closed-loop controlled 33-bus distribution network integrated with a STATCOM, employing two advanced control strategies: PIC-GWA and MPC-GWA. However, before applying these control mechanisms, it is crucial to analyze the system under open-loop conditions, i.e., in the absence of any feedback or regulatory control. This initial evaluation serves to uncover the natural dynamic behaviour of the distribution system and establishes a baseline reference for comparing the effectiveness of the proposed controllers. In the open-loop mode, the system is exposed to input variations and disturbances without corrective action, allowing for the observation of critical performance issues such as voltage instability, poor transient responses, and sensitivity to load changes. These challenges highlight the need for implementing robust and adaptive control algorithms. Once the system's open-loop behaviour is comprehensively understood, the control strategies are introduced sequentiallyfirst employing PIC-GWA and then transitioning to MPC-GWA. This staged integration facilitates a clear performance comparison and demonstrates the progressive enhancement in voltage regulation, stability, and system resilience achieved through the proposed closed-loop control schemes.

2.1 Open Loop 33-Bus Distribution Network Integrated with a STATCOM

Block Diagram of Open Loop 33-bus distribution network integrated with a STATCOM. The open-loop configuration of the 33-bus radial distribution network forms the foundational platform for evaluating the electrical behaviour and stability of the system under uncontrolled conditions. The block diagram shown in Fig. 1 includes several key components: a generating energy source, a STATCOM unit, a battery energy storage system, and the radial distribution feeder network itself.

Each component plays a significant role in the power flow and voltage profile characteristics across the network. The generating energy source acts as the primary supplier of electrical energy to the 33-bus system. The power flows radially outward through the distribution feeders, delivering electricity to various consumer nodes. In open-loop mode, this flow is unidirectional and lacks any dynamic control mechanism. The distribution network comprises 33 buses arranged in a radial configuration. This structure is common due to its simplicity and ease of protection coordination. However, the radial nature inherently causes voltage drops along the feeder due to cumulative line impedance (both resistance and reactance). These drops become more significant as the distance from the source increases, leading to poor voltage regulation at the tail-end buses. The system is also characterized by a high R/X ratio, typical of low-voltage distribution systems, which exacerbates the problem of voltage deviation and reduces power transfer efficiency. The STATCOM is a shunt-connected Flexible AC Transmission System (FACTS) device that provides dynamic reactive power compensation. In this open-loop setup, although the STATCOM is physically connected to the network, it operates without a closed-loop control system meaning it does not adaptively inject or absorb reactive power in response to system voltage deviations. Nevertheless, its presence still provides a passive benefit by maintaining a constant reference voltage and helping to stabilize localized voltage variations to some extent. The STATCOM operates using a voltage-source converter (VSC), which converts DC from the battery storage system into AC and vice versa, thereby enabling bidirectional power exchange. The battery is connected to the system for energy backup.

Simulation of Open Loop 33-bus distribution network integrated with a STATCOM. The simulation configuration of the open-loop 33-bus radial distribution grid sets the basis scenario for measuring system operation under the condition of the absence of dynamic regulating mechanisms, as provided by the example of Fig. 2. The basic components of the system consist of a generating power supply unit, a STATCOM unit, a battery energy storage unit, and the radial feeder grid. In this arrangement, the generating supply injects power radially towards the

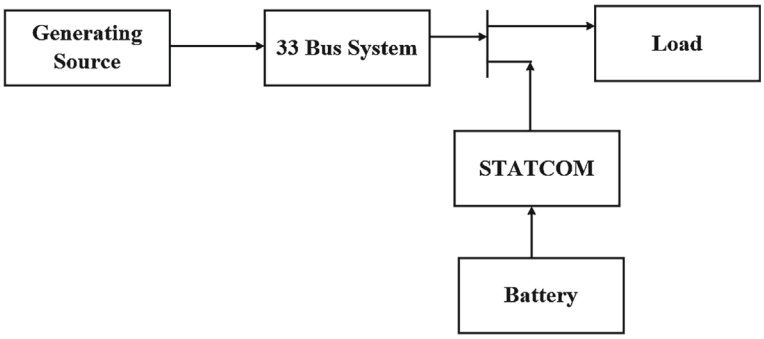


Fig. 1 Open Loop 33-bus distribution network integrated with a STATCOM

consumer nodes, creating a unidirectional flow without any regulating or correcting mechanisms. Because of the inherently radial configuration and the high R/X ratio commonly experienced by low-voltage distribution networks, significant voltage sags and compromised voltage regulation are noticeable at nodes farther away from the supply. The STATCOM, as a shunt-mounted FACTS unit, provides passive voltage regulation by giving a fixed level of reactive power but without employing active feedback under this open-loop configuration. The AC-DC energy exchange between the STATCOM and the battery energy storage unit uses a voltage-source converter as the mediator. To study the system's stability under dynamic conditions, a load disturbance occurs at 0.3 s. Figure 3 shows the input-generating voltage, which is 6.1 kV. Figure 4 illustrates the voltage across Bus-31, which has a peak value of 8 kV. Figure 5 shows the RMS voltage at Bus-31, which is 5.1 kV.

Figure 6 shows the current at Bus-31, Which is measured as 26 A. The real and reactive power at Bus-31, shown in Fig. 7, are 7.2 MW and 6.8 MVAR, respectively. These waveforms indicate that the system becomes unstable when load disturbances occur. In the open-loop configuration, the system fails to maintain its original transient-free state due to the absence of control mechanisms. Therefore, to enhance system stability, closed-loop controllers trained using the GWA algorithm are implemented.

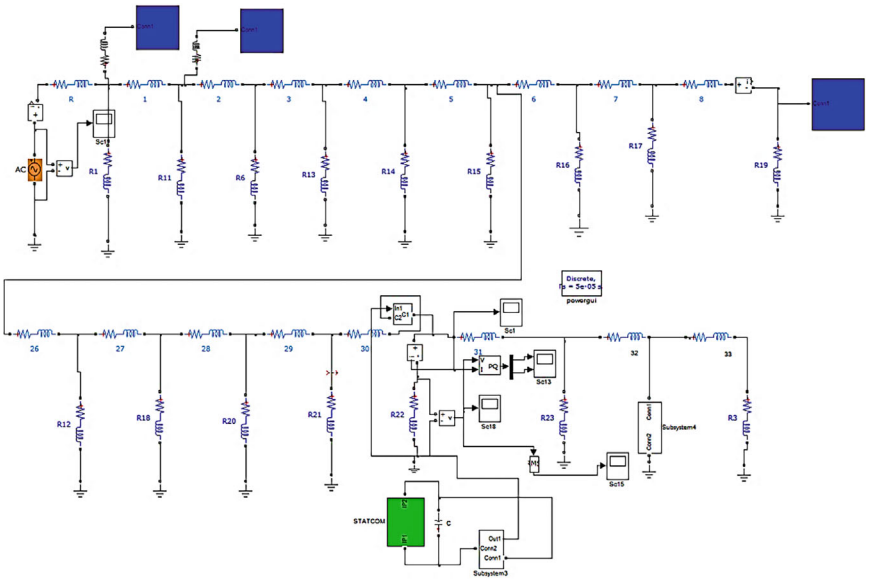


Fig. 2 Simulation circuit of open loop 33-bus distribution network integrated with a STATCOM

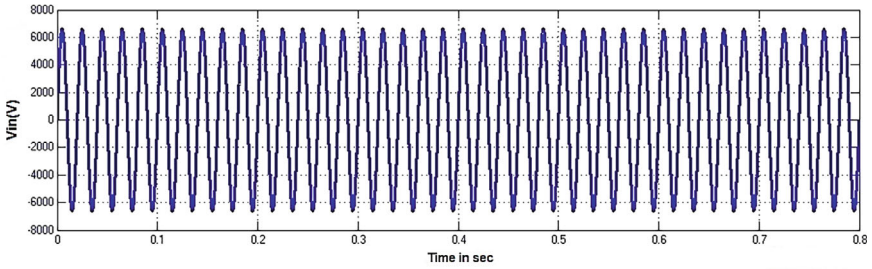


Fig. 3 Input generating voltage

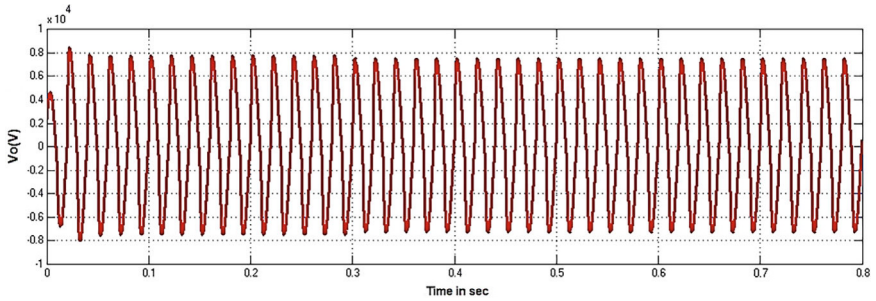


Fig. 4 Voltage at bus-31 at open loop system

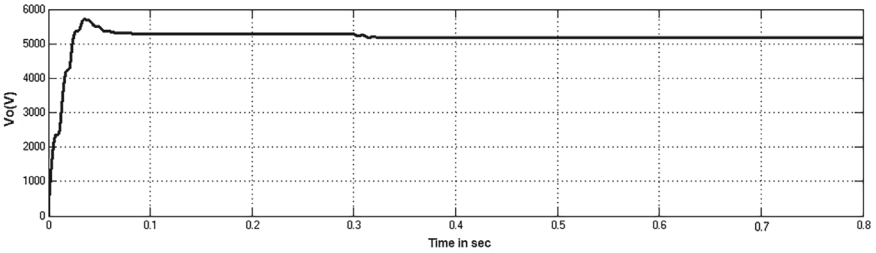


Fig. 5 RMS voltage at bus-31 at open loop system

2.2 Closed Loop 33-Bus Distribution Network Integrated with a STATCOM and GWA-PIC

Block Diagram of Closed Loop 33-bus distribution network integrated with a STATCOM and GWA-PIC. The block diagram for the closed-loop system with GWA-PIC control configuration of the 33-bus radial distribution network is shown in Fig. 8. It includes several key components: a generating energy source, a STATCOM unit, a battery energy storage system, the radial distribution feeder network, and the GWA-PIC. Each block plays a crucial role in ensuring improved power quality,

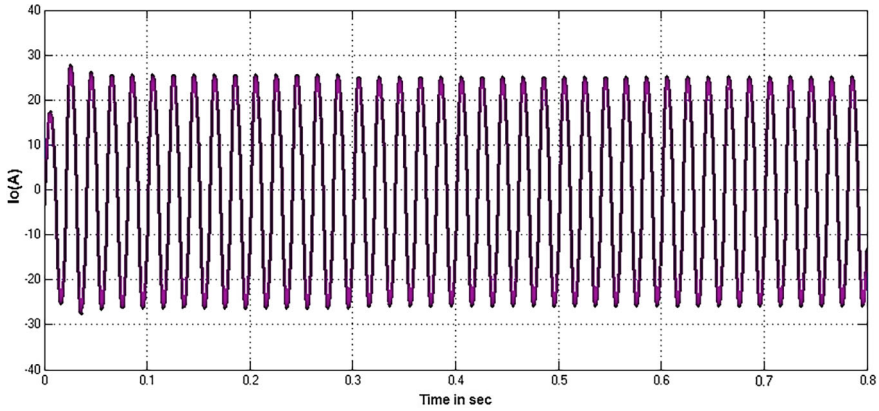


Fig. 6 Current at bus-31 at open loop system

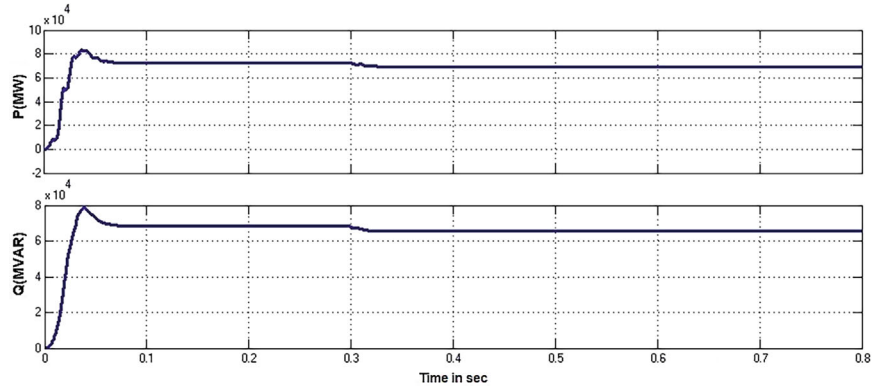


Fig. 7 Real and reactive power at bus-31 at open loop system

system stability, and voltage regulation across the network. In this closed-loop configuration, a PIC is used to regulate the STATCOM's output. The PIC adjusts the STATCOM's reactive power injection based on the voltage deviations detected at key buses. However, the performance of a standard PI controller is highly dependent on the optimal tuning of its proportional gain (Kp) and integral gain (Ki) parameters.

Simulation Diagram of Closed Loop 33-bus distribution network integrated with a STATCOM and GWA-PIC. The closed-loop simulation diagram with GWA-PIC control for the 33-bus distribution network, shown in Fig. 9, includes a generating source, STATCOM, battery storage, feeder network, and GWA-PI controller.

The PI controller regulates the STATCOM's output based on bus voltage deviations, with its gains (Kp, Ki) optimally tuned by GWA to enhance power quality, stability, and voltage regulation. To transcend the challenge of manual or trial-and-error tuning, the GWA algorithm is utilized to optimize the tuning of the PI controller

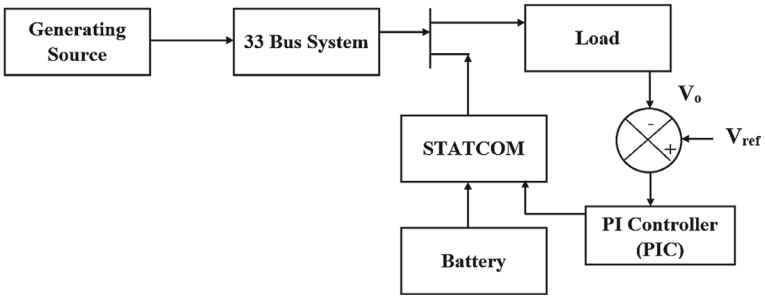


Fig. 8 Closed Loop 33-bus distribution network integrated with a STATCOM and GWA-PIC

parameters. GWA is a population-based metaheuristic inspired by the social structure and hunting habits of grey wolves. It performs the search of the solution space. It converges towards the optimum values of Kp and Ki that minimize a defined objective function generally on the basis of voltage deviation minimization, overshoot minimization, settling time improvement, and overall stability enhancement. The input generating voltage is shown in Fig. 3 and has a value of 6.1 kV. The voltage across Bus-31 is illustrated in Fig. 10, with a peak value of 7.8 kV. The RMS voltage at Bus-31, shown in Fig. 11, is 5.3 kV.

Figure 12 shows the Bus-31 current that is recorded as 24 A. According to Fig. 13, the real and reactive power at Bus-31 are 7.1 MW and 6.98 MVAR, respectively. From these waveforms, one can see that the PIC closed loop system trained with GWA restores stability in the face of load disturbances that otherwise cause instability. However, it reaches a steady state within 0.58 s in the GWA P controller-controlled system. In particular, the GWA-trained MPC controller has been introduced as a replacement to achieve faster stabilization.

2.3 Closed Loop 33-Bus Distribution Network Integrated with a STATCOM and GWA-MPC

Block Diagram of Closed Loop 33-bus distribution network integrated with a STATCOM and GWA-MPC. Figure 14 shows the block schematic of the closed loop configuration of the 33-bus radial distribution network with GWA MPC control. The forming system uses the generating energy source, STATCOM module, battery energy storage system, radial feeder network and a GWA-optimized Model Predictive Controller (MPC).

Each unit contributes significantly to improving the system stability, facilitating efficient voltage regulation, and, at the same time, improving the dynamic response. In this closed loop, the STATCOM's output is governed by the MPC by forecasting voltage deviation in the future using a mathematical model of the network. MPC differs from traditional proportional integral controllers (PICs) wherein at every

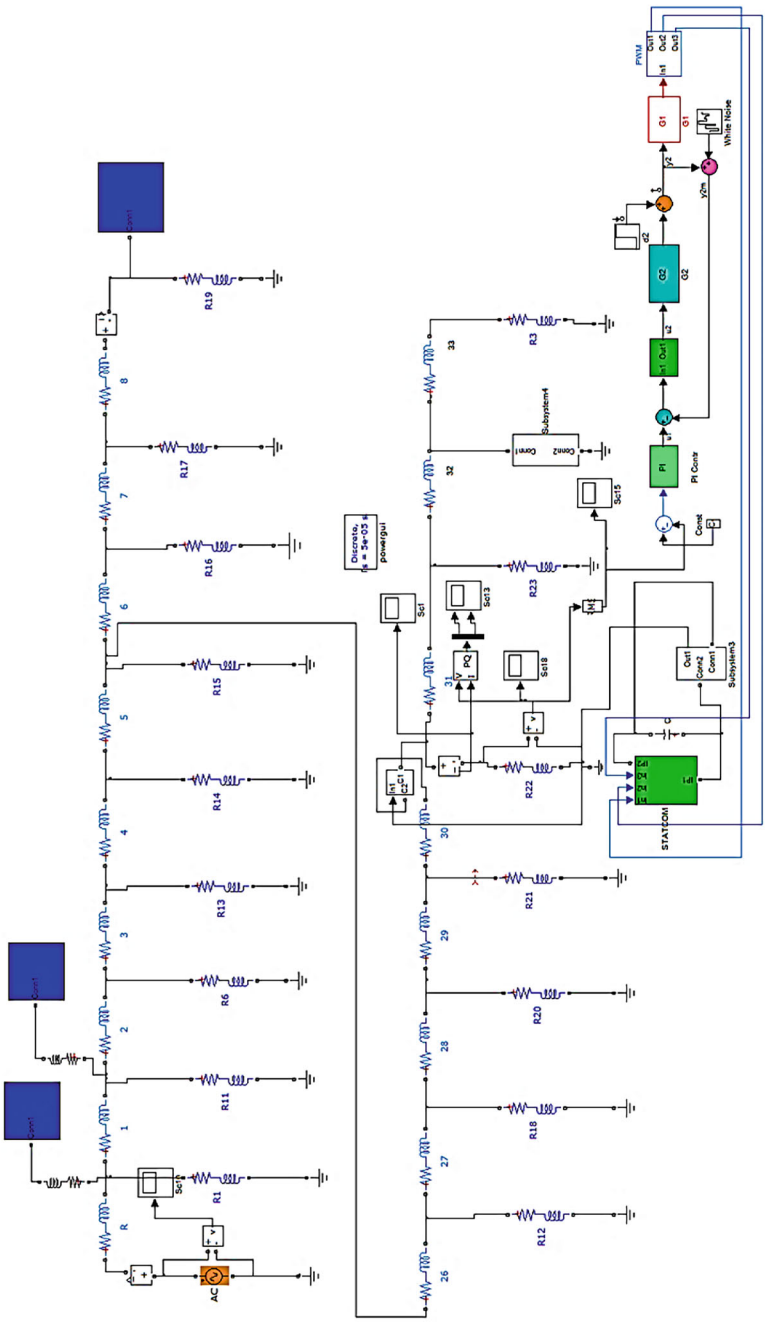


Fig. 9 Simulation circuit of closed loop 33-bus distribution network integrated with a STATCOM and GWA-PI

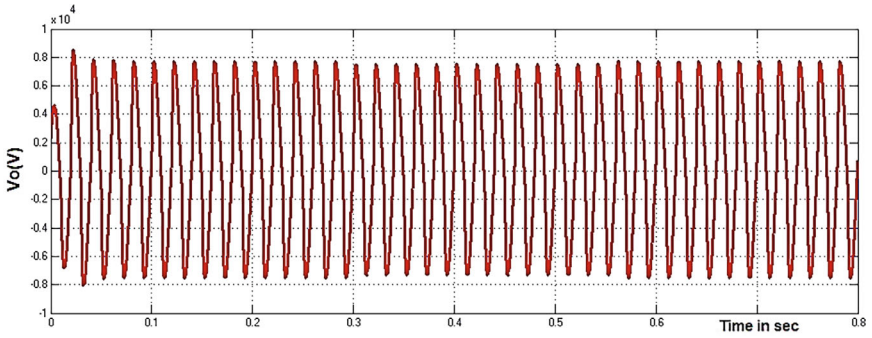


Fig. 10 Voltage at bus-31 at closed loop GWA-PIC controlled system

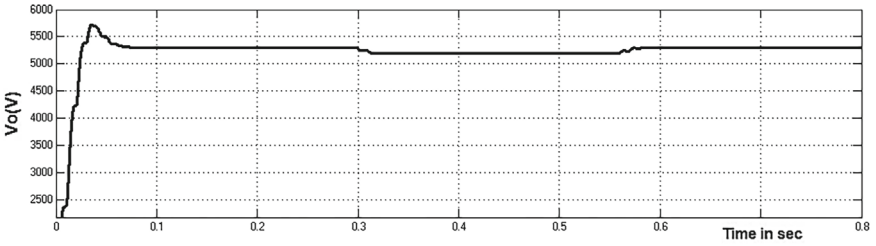


Fig. 11 RMS voltage at bus-31 at closed loop GWA-PIC controlled system

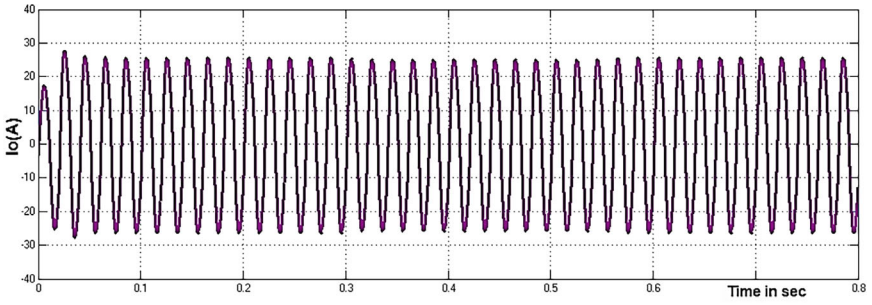


Fig. 12 Current at bus-31 at closed loop GWA-PIC controlled system

sampling instant, MPC solves an optimization problem over a defined prediction horizon considering the dynamics and operational constraints of the system. By taking a predictive approach, the controller is able to respond proactively and correctively to disturbances in order to achieve superior transient performance and greater disturbance resilience. The effectiveness of the controller is further refined by fine-tuning the important MPC parameters of prediction horizon, control horizon and weight coefficients through Grey Wolf Optimization (GWA). GWA mimics the social

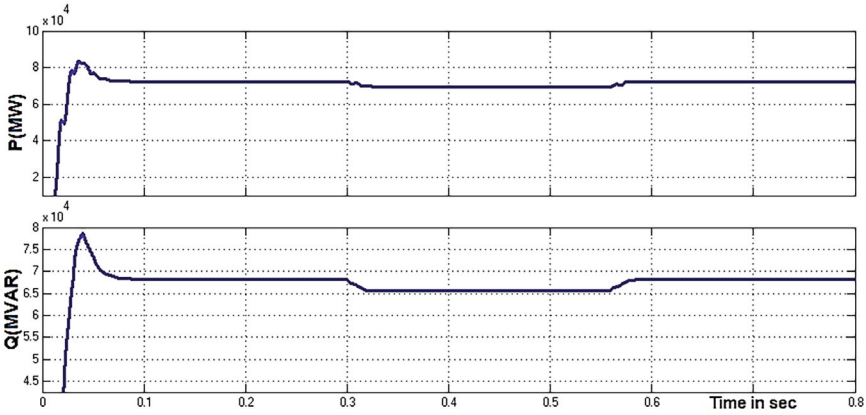


Fig. 13 Real and reactive power at bus-31 at closed loop GWA-PIC controlled system

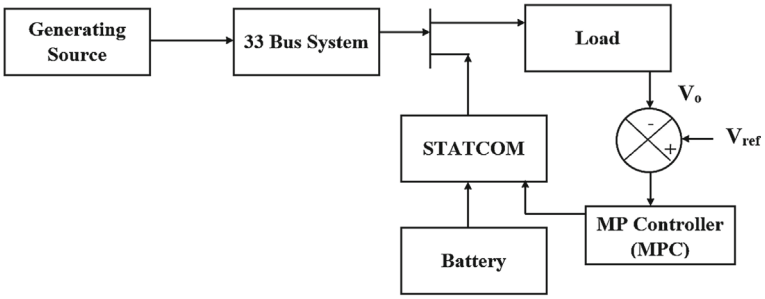


Fig. 14 Closed Loop 33-bus distribution network integrated with a STATCOM and GWA-MPC

leadership and hunting strategies of the grey wolves, which allows it to quickly explore the complex parameter space to find the global optimum. Hence, it converges faster to the global optimum and also improves overall controller performance.

Simulation Diagram of Closed Loop 33-bus distribution network integrated with a STATCOM and GWA-MPC. Figure 15 shows the simulation diagram for the closed-loop system with GWA-MPC control of the 33-bus radial distribution network. It includes a generating source, STATCOM, battery storage, distribution feeder, and the GWA-optimized MPC. The MPC regulates the STATCOM output by predicting voltage deviations and solving an optimization problem at each instant, offering faster and more robust control than a conventional PIC. The GWA enhances MPC performance by optimally tuning key parameters like prediction and control horizons, ensuring improved stability, voltage regulation, and dynamic response.

The input generating voltage is shown in Fig. 3 and has a value of 6.1 kV. The voltage across Bus-31 is illustrated in Fig. 16, with a peak value of 7.8 kV. The RMS voltage at Bus-31, shown in Fig. 17, is 5.3 kV.

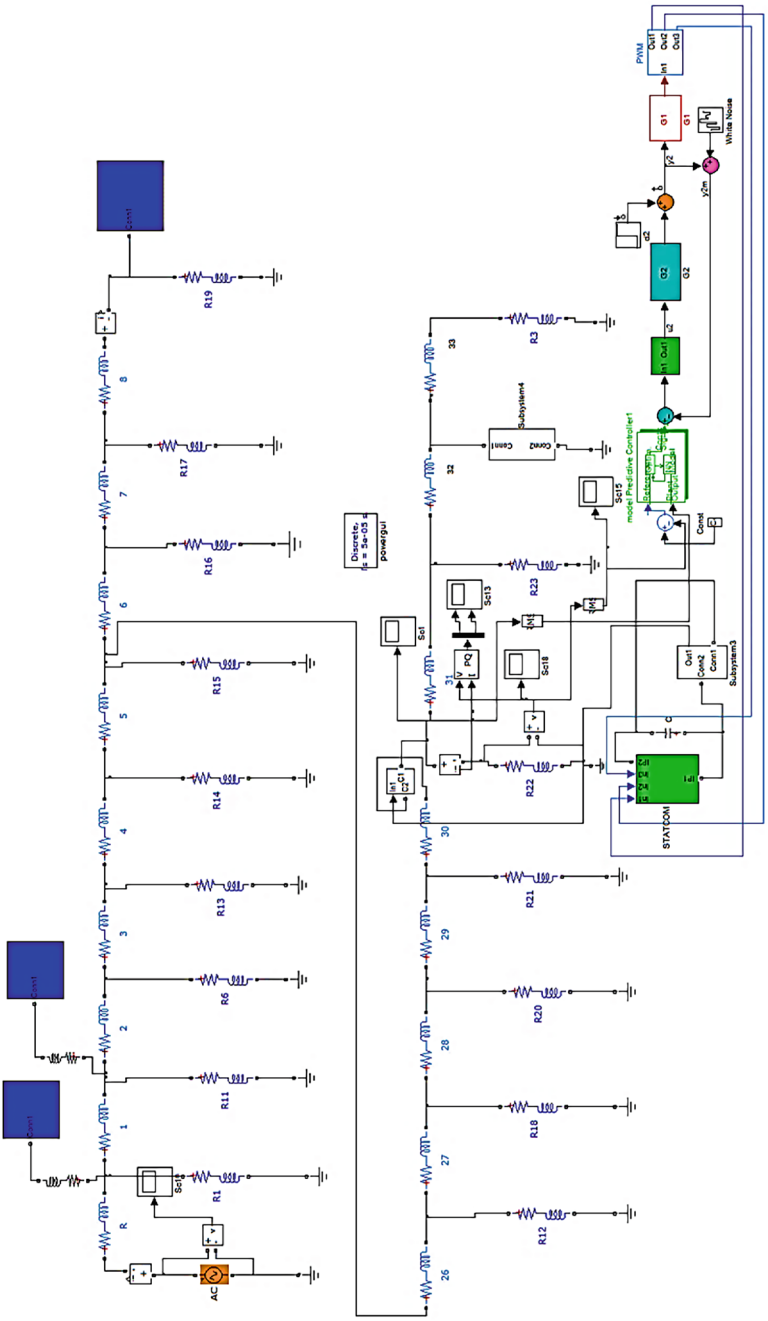


Fig. 15 Simulation circuit of closed loop 33-bus distribution network integrated with a STATCOM and GWA-MPC

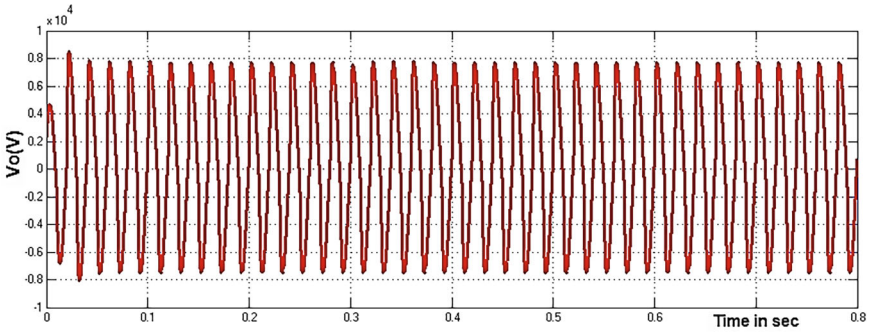


Fig. 16 Voltage at bus-31 at closed loop GWA-MPC controlled system

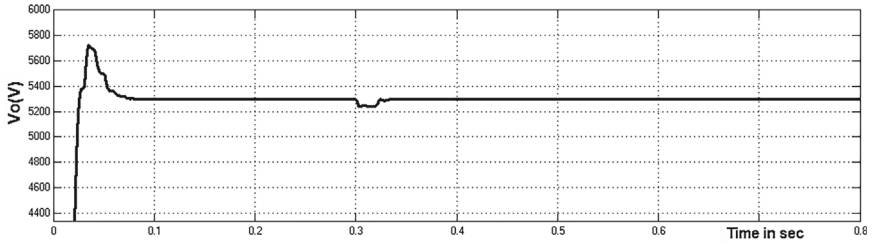


Fig. 17 RMS Voltage at bus-31 at closed loop GWA-MPC controlled system

The current drawn from Bus-31, as can be seen in Fig. 18, is 24 A. Figure 19 shows the real and reactive powers at Bus-31 to be 7.1 MW and 6.98 MVAR, respectively. These waveforms imply that although load disturbances induce the instability, the closed loop GWA-MPC system overcomes the instability. When compared to the GWA-PIC controlled system, the GWA-MPC controlled system goes to a steady state in 0.33 s, which is fast stabilization.

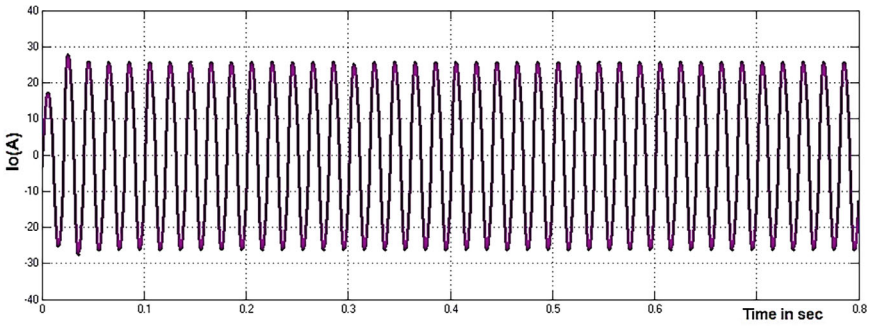


Fig. 18 Current at bus-31 at closed loop GWA-MPC controlled system

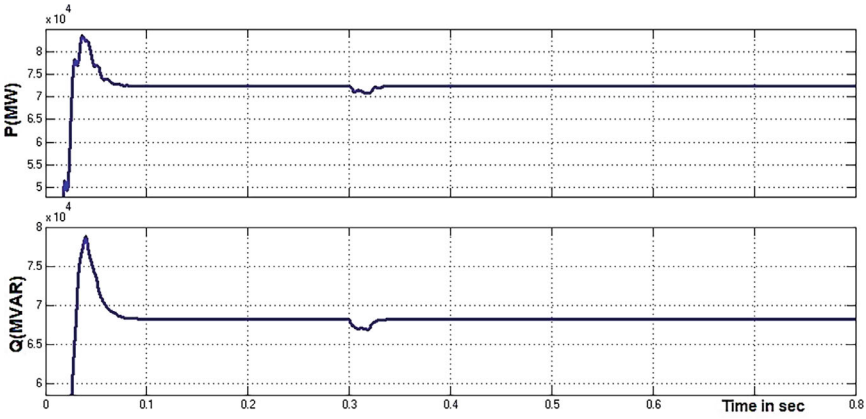


Fig. 19 Real and reactive power at bus-31 at closed loop GWA-MPC controlled system

2.4 System Performance—Discussion

The dynamic performance of the closed-loop 33-bus distribution network with STATCOM was assessed by comparing waveforms from two control strategies: GWA-PIC and GWA-MPC. The evaluation focused on critical time-domain metrics which are Rise Time (Tr), Peak Time (Tp), Settling Time (Ts), and Steady-State Error (Ess) which reflect the system's responsiveness and stability under source disturbances. The GWA-MPC method outperformed GWA-PIC across most parameters, delivering faster transient response, lower overshoot, quicker settling, and reduced steady-state error. This enhanced performance is due to MPC's ability to effectively manage the nonlinear and time-varying characteristics of the 33-bus network, especially when optimized by Grey Wolf Optimization. While GWA-PIC also improved results over the open-loop system, the time-domain analysis summarized in Table 1 clearly highlights the superior advantages of the GWA-MPC approach.

Table 1 clearly shows that the GWA-MPC control strategy significantly outperforms the GWA-PIC method. The rise time (Tr) improves from 0.55 s to 0.32 s, and peak time (Tp) decreases from 0.57 s to 0.33 s, indicating faster responses. Settling time (Ts) is notably reduced from 0.58 s to 0.35 s, reflecting quicker stabilization. The steady-state error (Ess) also decreases from 1.68 V to 0.75 V, showing enhanced accuracy. These improvements demonstrate that the GWA-MPC system provides faster,

Table 1 Time domain parameters of closed loop 33-bus distribution network integrated with a STATCOM and GWA trained controllers

Controller	Tr	Tp	Ts	Ess
GWA-PIC	0.55	0.57	0.58	1.68
GWA- MPC	0.32	0.33	0.35	0.75

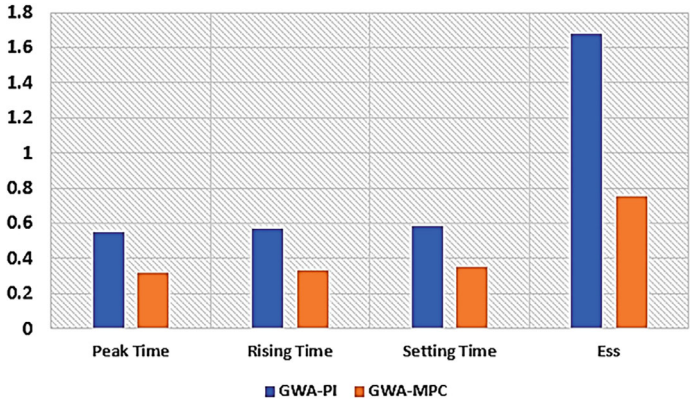


Fig. 20 Bar chart comparison of time domain parameters of multifunctional UPQC and closed loop GWA-PID and GWA-FLC controlled system

more stable, and precise control for the closed-loop 33-bus distribution network with STATCOM. The comparison is illustrated in the bar chart in Fig. 20.

2.5 Conclusion

As discussed in the introductory section, the objective of this work to improve system stability has been achieved by implementing the proposed scope of work. The open-loop operation of the 33-bus distribution network integrated with a STATCOM showed stability issues, especially during transient events. To overcome these challenges, a closed-loop control approach using intelligent optimization was implemented. Two control configurations were tested: a PI controller tuned with GWA and an MPC tuned with GWA. In both setups, the GWA optimized controller parameters to enhance dynamic response and overall system stability. Time-domain analysis confirmed the superior performance of the GWA-MPC configuration across all key metrics. Simulations were conducted under load disturbances using both closed-loop GWA-PI and GWA-MPC controllers. The GWA MPC system outperformed the SA system with a shorter rise time from 0.55 s to 0.32 s, shorter peak time from 0.57 s to 0.33 s, and shorter settling time from 0.58 s to 0.35 s, and smaller steady-state error from 1.68 V down to 0.75 V. The obvious results of these results show that 33 bus closed-loop systems with STATCOM under GWA-MPC control outperform the closed-loop system under GWA-PI control.

References

1. Mokryani G (2022) Future distribution networks. AIP Publishing, New York, USA
2. Okelola MO, Salimon SA, Adegbola OA, Ogunwole EI, Ayanlade SO, Aderemi BA (2021) Optimal siting and sizing of D-STATCOM in distribution system using new voltage stability index and bat algorithm. In: 2021 ınternational congress of advanced technology and engineering (ICOTEN). IEEE, pp 1–5
3. Colmenar-Santos A, Reino-Rio C, Borge-Diez D, Collado-Fernández E (2016) Distributed generation: a review of factors that can contribute most to achieve a scenario of DG units embedded in the new distribution networks. Renew Sustain Energy Rev 59:1130–1148
4. Muruganantham B, Selvam MM, Gnanadass R, Padhy NP (2017) Energy loss reduction and load balancing through network reconfiguration in practical LV distribution feeder using GAMS. In: 2017 7th International conference on power systems (ICPS). IEEE, pp 509–513
5. Franco JF, Rider MJ, Lavorato M, Romero R (2013) A mixed-integer LP model for the optimal allocation of voltage regulators and capacitors in radial distribution systems. Int J Electr Power Energy Syst 48:123–130
6. Salimon SA, Adepoju GA, Adebayo IG, Ayanlade SO (2022) Impact of shunt capacitor penetration level in radial distribution system considering techno-economic benefits. Nigerian J Technol Dev 19(2):101–109
7. Balamurugan P, Yuvaraj T, Muthukannan P (2018) Optimal allocation of DSTATCOM in distribution network using whale optimization algorithm. Eng Technol Appl Sci Res 8(5):3445–3449
8. Montoya OD, Gil-González W, Hernández JC (2021) Efficient operative cost reduction in distribution grids considering the optimal placement and sizing of D-STATCOMs using a discrete-continuous VSA. Appl Sci 11(5):2175
9. Castiblanco-Pérez CM, Toro-Rodríguez DE, Montoya OD, Giral-Ramírez DA (2021) Optimal placement and sizing of D-STATCOM in radial and meshed distribution networks using a discrete-continuous version of the genetic algorithm. Electronics 10(12):1452
10. Iffouzar K, Taraft S, Aouzellag H, Ghedamsi K, Aouzellag D (2015) December. DRFOC of polyphase induction motor based on fuzzy logic controller speed. In: 2015 4th ınternational conference on electrical engineering (ICEE). IEEE, pp 1–7
11. Djerioui A, Houari A, Ait-Ahmed M, Benkhoris MF, Chouder A, Machmoum M (2018) Grey wolf based control for speed ripple reduction at low speed operation of PMSM drives. ISA Trans 74:111–119
12. Dogruer T (2022) Grey wolf optimizer-based optimal controller tuning method for unstable cascade processes with time delay. Symmetry 15(1):54
13. Garg H (2016) A hybrid PSO-GA algorithm for constrained optimization problems. Appl Math Comput 274:292–305
14. Oshaba AS, Ali ES, Abd Elazim SM (2015) ACO based speed control of SRM fed by photovoltaic system. Int J Electr Power Energy Syst 67:529–536
15. Ullah K, Jiang Q, Geng G, Rahim S, Khan RA (2022) Optimal power sharing in microgrids using the artificial bee colony algorithm. Energies 15(3):1067
16. Djerioui A, Houari A, Machmoum M, Ghanes M (2020) Grey wolf optimizer-based predictive torque control for electric buses applications. Energies 13(19):5013
17. Deng W, Zhao H, Zou L, Li G, Yang X, Wu D (2017) A novel collaborative optimization algorithm in solving complex optimization problems. Soft Comput 21:4387–4398
18. Qais MH, Hasanien HM, Turky RA, Alghuwainem S, Tostado-Véliz M, Jurado F (2022) Circle search algorithm: a geometry-based metaheuristic optimization algorithm. Mathematics 10(10):1626
19. Mirjalili S, Mirjalili SM, Lewis A (2014) Grey wolf optimizer. Adv Eng Softw 69:46–61
20. Mohammed SA, Abdel-Moamen MA, Hasanin B (2013) A review of the state-of-the-art of power electronics for power system applications. J Electron Commun Eng Res 1(1):43–52
21. Ghamari SM, Khavari F, Mollaee H (2024) Adaptive backstepping controller design for DC/DC buck converter optimised by grey wolf algorithm. IET Energy Syst Integr 6(1):18–30

22. Singh KM, Gope S (2024) Renewable energy integrated multi-microgrid load frequency control using grey wolf optimization algorithm. Mater Today Proc 46:2572–2579
23. Revathy R, Balaji B, Mohasin AA, Gobinath A (2023) Supercapacitor and bldc motor-based regenerative braking for an electric vehicles. In: 2023 2nd ınternational conference on smart technologies and systems for next generation computing (ICSTSN). IEEE, pp 1–5
24. Balaji B, Ganesan S, Pugazhendiran P, Subramanian S (2023) Closed loop proportional resonant controller controlled DC microgrid system with advanced material technology in solar PV system. Mater Today Proc
25. Balaji B, Krishnan DS (2021) Fopid controlled high step-up super lift DC-DC converter with enhanced response. In: 2021 3rd ınternational conference on signal processing and communication (ICPSC). IEEE, pp 176–179

Real Time People Monitoring Using MobilNet-SSD



Darshan Atkari, Arshal Agarwal, Atharva Kulkarni, Apoorva Kulkarni, and Sonali Antad

1 Introduction

Real-time object detection has become a building block capability for a broad range of contemporary use cases, from autonomous navigation systems and smart surveillance to retail analytics and crowd monitoring. The progress has been propelled mainly by profound improvements in lightweight deep learning architecture, which has allowed high-performing models to be deployed even on limited-resource edge devices. Among these designs, MobileNet has been notable for its capacity to provide a good trade-off between accuracy and efficiency, thus being widely used for real-time detection tasks. Nevertheless, with all these improvements, there is still a challenge in achieving detection accuracy, computational efficiency, and adaptability to dynamic and varied environments.

This study explores the effective use of an object detection pipeline for real-time objects based on the MobileNetV2 architecture. Using pre-trained models via TensorFlow Hub implies that the deployment process is greatly eased such that object detection can readily be incorporated into various real-time applications by the developer. The methodological framework adopted in this scenario is based on modular process flow with distinct phases for preprocessing, inference, and postprocessing. Every phase is designed to be both cost-effective and easy to tailor, with the ability to quickly experiment and deploy across different use cases.

One specific application of this research is people counting, a task of interest for areas such as public safety, events, and intelligent retailing. The pipeline needs to

D. Atkari · A. Agarwal (✉) · A. Kulkarni · A. Kulkarni · S. Antad
Computer Engineering, Vishwakarma Institute of Technology, Pune, India
e-mail: ajagarwal2005@gmail.com

S. Antad
e-mail: sonali.antad@vit.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_35

be capable of being executed under realistic conditions, and the hope is to minimize latency while maintaining accuracy despite varying lighting, angles, and crowd numbers. Unlike traditional detection solutions that are generally founded on in-house-trained models that require vast datasets and protracted training sessions, this approach prioritizes scalability and reproducibility. It achieves this using widely adopted open-source tools and standardized frameworks, namely OpenCV for image processing and TensorFlow for deep learning inference.

By this approach, the paper aims to demonstrate an efficient and straightforward path to real-time object detection without compromising quality or rendering it inappropriate for high-speed deployment and easy extension. The central contribution of this research is the integration of the MobileNet-SSD deep learning network with a light Centroid Tracking method to enable real-time people counting under practical limitations, such as moderate occlusion, low-lighting, and low-power hardware configurations. Through leverage of modular architecture and open-source technologies such as TensorFlow and OpenCV, the system delivers low latency and high accuracy with a scalable solution for retail stores, classrooms, and public places.

2 Literature Review

The people counting task has also seen tremendous change, shifting from traditional computer vision methods to deep learning-based methodologies to address the needs of accuracy, scalability, and real-time processing. Initial work presented morphological tools for dense people counting, which worked effectively in static environments but failed to handle occlusions and illumination. The approach does not perform well under occlusions and dynamic illumination, reducing real-time applicability. Our approach handles occlusion better with MobileNet-SSD and real-time tracking, offering dynamic flexibility [1]. Similarly, use of stereo vision to improve spatial differentiation, though real-time application was limited due to high computational load. Efficient as the system is, it requires extremely high computation, which reduces real-time applicability. Our low-resource MobileNet-SSD system handles well without stereo vision on low-end hardware [2].

With the emergence of deep learning, proposed the Scale-Adaptive CNN (SaCNN), which adapted effectively to varying perspectives and crowd densities. Although accurate, it needs advanced training and big sets of data. We simplify deployment with pre-trained MobileNet models with little need for retraining [3]. Extended this by incorporating environmental context through Adaptive CNNs (A-CCNN), leading to greater accuracy in dynamic indoor environments. But increased model complexity affects speed and ease of deployment. We balance the speed and accuracy with MobileNet and modular tracking [4].

While these methods improved robustness, real-time processing remained a concern. Shi et al. presented a lightweight CNN optimized for fast inference, suitable for surveillance systems. Although being fast, the model had low tracking, which

led to potential miscounts. We pair Centroid Tracking to provide constant identification between the frames [5]. Used OpenCV for a real-time crowd counting system, showing that open-source tools could support practical, deployable solutions without excessive hardware. The approach was not very accurate in tight crowd and low illumination conditions. Our framework combines tracking and deep learning to offer more robust performance in real-world applications [6].

To address limitations in visibility and privacy, DeepCount introduced leveraging WiFi signals and deep learning to estimate crowd density, providing a non-intrusive and scalable alternative in indoor environments. Despite being privacy-friendly, the system is confronted with spatial resolution and localization. We utilize vision-based methods with accurate spatial context, delivering higher-grained detection [7].

Recently, attention-based models like CAT-CNN enhanced focus on relevant image regions, minimizing false detections. With high accuracy, such models are computationally costly for edge deployment. We aim at real-time execution on embedded hardware with MobileNet [8]. Anonymous cascaded detection work tried to enhance accuracy using multistage refinement. Cascading, though, increases latency and complexity. We avoid multi-stage overhead by directly applying MobileNet-SSD for an equivalent trade-off [9].

In the context of object detection, particularly using MobileNet-SSD, several studies have validated its applicability in people counting tasks [10]. Explored object detection using Python and OpenCV, emphasizing the simplicity and deployability of MobileNet in constrained systems [11]. Also a mobile-based indoor framework with SSD-MobileNet, demonstrating strong real-time capability and generalization under indoor lighting and space constraints. It lacked application-specific tuning like entry/exit counting. We extend the system with added directional tracking and threshold-based alerts [12].

A comparative study of one-stage detectors like YOLO and SSD for people detection, concluding that SSD combined with MobileNet is efficient for real-time, low-power applications. Their work was limited to static analysis without tracking or scalability experiments. Our work validates SSD-MobileNet in dynamic environments with tracking and alerting embedded [14]. Similarly, an edge-computing-based people counting solution using MobileNet, showcasing deployment in elevators where computational resources are minimal and latency is critical. Though optimal, it was scenario-specific, not generalized for different types of rooms. Our system generalizes deployment across heterogeneous indoor environments using a versatile architecture [15].

Tracking plays a complementary role in robust people counting. Traditional approaches relied on synchronized video streams and motion tracking, but newer systems, including the proposed work, use centroid tracking, which allows persistent object monitoring across frames with minimal computational burden, ensuring real-time feasibility. Their system has infrastructure-intensive deployment and manual calibration. Our approach utilizes a single overhead camera, which is cost-saving and simplicity [13].

In summary, recent literature supports the shift towards lightweight, modular, and real-time object detection systems, particularly those built using MobileNet-SSD

frameworks. These are optimal for indoor environments where efficiency, adaptability, and cost-effectiveness are prioritized. The current work builds on these findings, combining robust MobileNet-based detection with efficient tracking to create a scalable, low-latency people counting solution suitable for modern smart environments.

3 Methodology

3.1 MobilNet

MobileNet is a class of efficient deep neural networks designed for mobile and embedded vision applications. It achieves low computational cost by replacing standard convolutions with **depthwise separable convolutions**, which factorize a regular convolution into a depthwise convolution (per-channel filtering) and a 1×1 pointwise convolution (channel combination). This drastically reduces both the number of parameters and the number of multiply–add operations, enabling fast inference on resource-constrained devices.

To further adapt to various hardware limitations, MobileNet introduces two global hyperparameters: **width multiplier (α)** and **resolution multiplier (ρ)**. These allow flexible trade-offs between model size, speed, and accuracy by uniformly shrinking the number of channels and input image resolution. MobileNet achieves competitive accuracy compared to heavier models like VGG16 and GoogleNet, but with significantly fewer parameters and lower computational demand.

ReLU (more precisely, ReLU6) is applied following both the depthwise and pointwise convolutions in MobileNet to introduce non-linearity so that the network can learn sophisticated patterns beyond linear transformations. ReLU6, which clamps outputs to 6, is utilized to enable support for low-precision operations and robustness on mobile devices. The activation function serves to maintain model efficiency while allowing effective feature learning.

MobileNet is preferred over other models for people counting applications, particularly in low-resource and real-time conditions, since it finds an optimal balance among efficiency, speed, and accuracy. In contrast to heavier models like VGG16 or ResNet, MobileNet achieves startling reductions in computational complexity with depthwise separable convolutions that reduce the number of parameters and operations by orders of magnitude without discouraging much from performance. Such effectiveness is necessary for deployment onto low-power, low-memory, and low-processing mobile or edge devices. In addition, its width and resolution multiplier hyperparameters allow the model's size and speed to be tweaked to satisfy particular hardware requirements (Fig. 1).

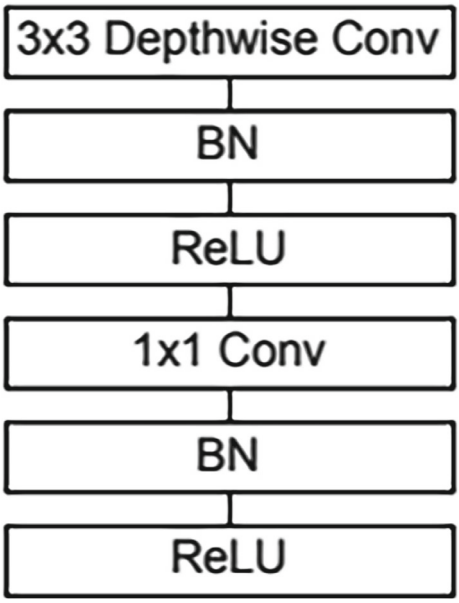


Fig. 1 Depthwise separable convolutions with depthwise and pointwise layers followed by batchnorm and ReLU

3.2 Single-Shot Multi-box Detector (SSD)

The Single Shot MultiBox Detector (SSD) is a deep learning-based object detection framework that eliminates the need for region proposal steps by performing detection in a single forward pass of a convolutional neural network. Unlike two-stage detectors like Faster R-CNN, SSD discretizes the output space using a set of default boxes with varying aspect ratios and scales, applied across multiple feature maps of different resolutions. This enables effective detection of objects at various sizes and locations. The core of SSD uses small convolutional filters to predict both object category scores and bounding box offsets directly from these feature maps. By combining predictions from multiple scales, SSD achieves a good balance between speed and accuracy. It is trained end-to-end with a multicomponent loss that combines classification and localization errors.

SSD achieves real-time performance while maintaining high accuracy. On the VOC2007 dataset, SSD300 achieves 74.3% mAP at 59 FPS, outperforming YOLO in accuracy and Faster R-CNN in speed. Larger versions like SSD512 further improve accuracy. Its lightweight architecture, combined with techniques like hard negative mining and data augmentation, makes SSD suitable for deployment in real-time systems.

SSD is more desirable than other object detection networks for practical uses like people counting because it offers the best tradeoff in terms of speed and accuracy. Unlike two-stage detectors such as Faster R-CNN, which first generate region proposals and then classify them, SSD localizes objects and classifies them in a single pass of forward, greatly reducing inference time. Compared with one-stage

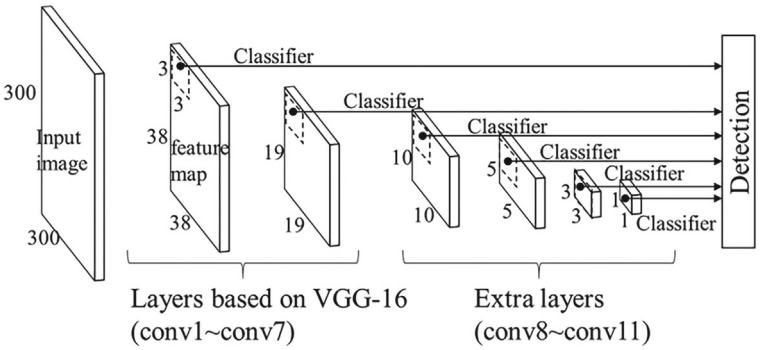


Fig. 2 Single shot multibox detector architecture

counterparts such as YOLO, SSD is more accurate, particularly in detecting small objects, since it uses multiple feature maps of different resolutions. In addition, SSD employs default boxes with varying aspect ratios and scales such that more efficient detection can be done irrespective of object location and size. These features make SSD especially suitable for light, interactive systems where inference speed as well as detection performance is of extreme concern (Fig. 2).

3.3 Centroid Tracker

Centroid tracking is a simple yet efficient object tracking method often utilized in video analytics like people counting and motion tracking. The algorithm begins with calculating the centroid (geometric center) of the bounding boxes that are detected in video frames and then verifying those centroids frame-to-frame using small Euclidean distance. Dependability of the system is achieved through several ways. First, the use of pre-trained models like SSD MobileNet V2, fine-tuned for indoor environments ensures reliable and consistent detection performance. Second, the Centroid Tracking algorithm mitigates ID-switching because object continuity is maintained across frames, even during the moments of temporary occlusion. Third, system logging and threading mechanisms are used to make the system robust during runtime, and confidence-based filtering is used to suppress false positives. Together, these techniques construct a robust and reliable monitoring pipeline suitable for real-world deployment (Fig. 3).

The algorithm begins by taking an initial object detection set (bounding boxes) and appending a distinct key to each object. On subsequent frames, it finds differences between the current list of detected centroids and the last frame's list, updating object location on the basis of nearest neighbor match. If the object goes missing (e.g., out of frame or occluded), it is temporarily indicated as "missing." If the object re-emerges in a certain amount of frames, its initial ID is retained so that persistent tracking

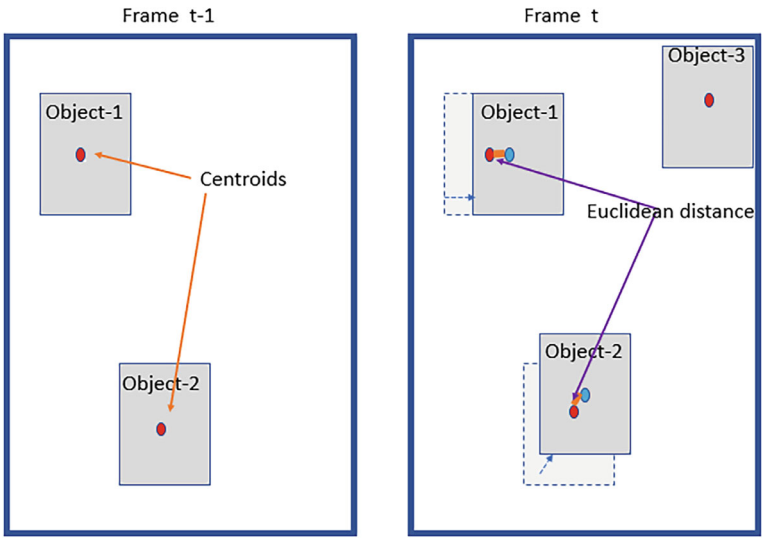


Fig. 3 Centroid tracker methodology

can be possible. Centroid tracking is also a computationally light algorithm because it is merely concerned with straightforward geometric computations as opposed to repeatedly re-initiating a deep neural network for each frame. It is well-suited to applications where objects are slowly moving or have a predictable motion. It might not work as well with complex occlusions or random object movement as with more sophisticated methods, but it makes a good starting point and is a good baseline.

4 System Architecture

Real-Time Object Detection using SSD MobileNet V2 is an end-to-end computer vision pipeline that is constructed with the intention of swift object identification from images. Two parallel initialization phases form the initiation process: loading the model and receiving the image input. The two points of departure offer the groundwork for the detection process, taking initiative to design two convergent paths which unite for the actual detection process subsequently.

The preprocessing component is a basic component of the architecture and consists of five sequential steps that pre-process the input image for processing. The system first loads the image using OpenCV, and color space conversion is performed from BGR to RGB format. Then the image is resized to the model's assumed size and a batch dimension added according to the model's input requirement. Lastly, the image is transformed into a uint8 tensor format, producing a well-formatted tensor that can be processed by the neural network.

At the same time, the model preparation department goes through its own series of operations. This starts with the choosing of the SSD MobileNet V2 architecture, which is mobile and edge device-focused, providing a compromise of accuracy versus computational requirements. The application loads pretrained TensorFlow Hub weights, taking advantage of transfer learning so that training from scratch is not required. The final step in this preparation phase involves setting the model to evaluation mode, ensuring it's configured for inference rather than training.

The object detection module is the system's central analysis engine. In this, the preprocessed image tensor is input to the model, which produces four different outputs: the coordinates of bounding boxes where objects in the image are located, class IDs used to determine the type of each object that has been detected, confidence scores reflecting the model's confidence in each detection, and the number of detections present in the image. These outputs together give detailed information regarding all objects recognized in the scene.

After detection, the architecture includes advanced postprocessing and visualization features to enhance and display results. The system initially removes low-confidence predictions by filtering detections based on confidence thresholds. It subsequently resizes the bounding boxes to the original image size and converts the numerical class IDs to human-readable COCO labels. These processed results are visually depicted by drawing bounding boxes and labels directly onto the image. Lastly, the system either saves the annotated image to disk or displays it to the user in real-time.

To improve the accuracy of the SSD MobileNet V2 real-time object detection system, several methods can be employed. Initial preprocessing through techniques such as image normalization, contrast adjustment, or adaptive histogram equalization can improve the quality of the input data into the network, particularly under low light or noisy conditions. Second, fine-tuning the pretrained model using a task-specific dataset will significantly improve detection accuracy if the deployment scene (e.g., indoor surveillance or shopping centers) is not the same as the original COCO training distribution. Including non-maximum suppression (NMS) with class-wise optimization and soft-NMS application can further reduce false positives and enhance detection robustness for nearby objects.

To enhance efficiency, the pipeline can be optimized through model quantization (e.g., quantizing the model to INT8 or FLOAT16), which reduces model size and inference speed, particularly on mobile or edge devices. Additionally, batch inference for several frames (if latency permits) or hardware acceleration via APIs like TensorRT or OpenVINO can also reduce detection time. On the software side, utilizing multithreading or asynchronous data pipelines for parallel postprocessing and preprocessing can minimize idle CPU/GPU cycles. These optimizations make the detection system interactive and scalable for real-time applications without trading off detection quality.

The entire architecture relies on three key implementation tools that make up the technology foundation of the system. OpenCV carries out image processing tasks,

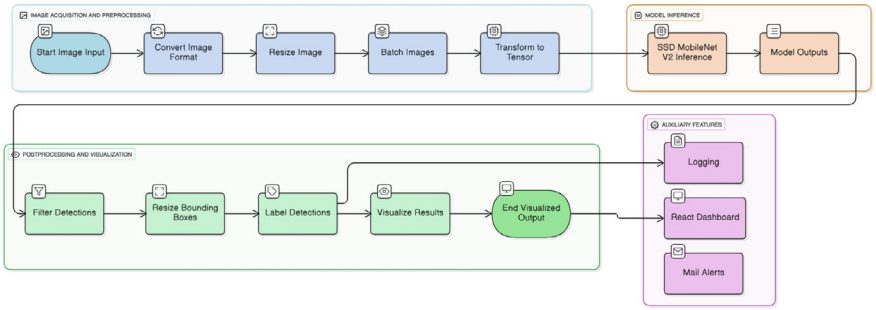


Fig. 4 System architecture

NumPy conducts numerical operations and array operations, and TensorFlow Hub provides an interface to the pretrained model. Such carefully designed pipeline enables efficient real-time object detection and is suitable for applications ranging from surveillance and autonomous vehicles to retail analytics and augmented reality systems.

To improve the monitoring process, a few approaches are employed. These involve maximizing field of view and minimizing blind spots by optimizing camera positioning directly over entrance points. Dynamic frame thresholding with respect to intensity of motion and ROI cropping will minimize false alarms from noise in the background. Additionally, using adaptive confidence thresholds allows the system to screen out doubtful detections in changing lighting or environmental conditions. Future upgrades can involve the addition of other sensory modalities such as thermal cameras or Wi-Fi localization for even more precise surveillance under occlusion or low light conditions (Fig. 4).

5 Results and Discussion

Every image is run through a pipeline that pre-processes the image, passes it to the MobileNet-SSD model for people detection, and uses the Centroid Tracking algorithm to track people from one frame to the next. The system detects people in every image by drawing bounding boxes around them and tagging them with IDs for tracking. The number of detected people in every image is displayed, providing the correct count for that frame.

Detection is strong even when people are partly occluded or grouped tightly together because the model can disentangle individuals with spatial features. Labels and bounding boxes determine which objects are detected as people, and confidence scores only permit detections that meet a threshold to be utilized as valid. Visual feedback allows users to quickly confirm the system output for each photograph as correct.



Fig. 5 Before entering

In cases of dynamic lighting or camera angles, the system remains consistent in performance, demonstrating flexibility towards real-world scenarios. With the use of depthwise separable convolutions in MobileNet, detection is sure to be both fast and accurate even on low-resource systems, rendering it suitable for real-time applications.

Overall, the annotated images clearly illustrate the effectiveness of the MobileNet-SSD and Centroid Tracking approach to real-time people tracking. The system not only provides the count of people per image but also visual evidence of detection and tracking, making it feasible for applications in crowd management, safety, and analytics (Figs. 5, 6, 7, 8 and 9).

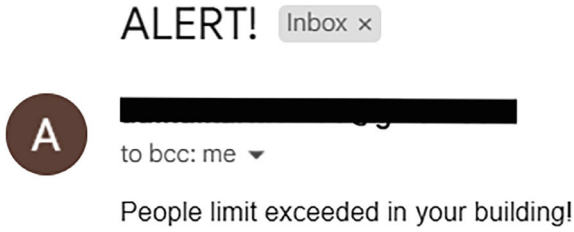


Fig. 6 After entering

Move In	In Time	Move Out	Out Time
1	2025-04-28 23:00	1	2025-04-28 23:00
2	2025-04-28 23:00	2	2025-04-28 23:00
3	2025-04-28 23:00	3	2025-04-28 23:00
4	2025-04-28 23:00		
5	2025-04-28 23:00		
6	2025-04-28 23:00		
7	2025-04-28 23:00		

Fig. 7 Logs of people entering

Fig. 8 Auto-generated email alert



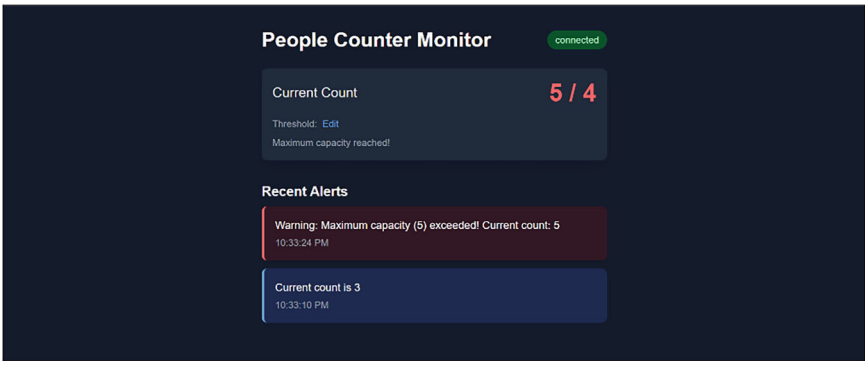


Fig. 9 Web dashboard monitor

6 Conclusion

People counting technology is becoming a more precise, learning, and privacy-oriented device, poised to serve the needs of future smart systems. This work contributes to that process by delivering a lightweight but sturdy real-time people tracking system based on the MobileNet-SSD model supplemented By this method, the paper seeks to illustrate an effective and simple route to real-time object detection that does not sacrifice quality or become unsuitable for rapid deployment and simple extension. The main contribution of this work is the coupling of the MobileNet-SSD deep learning model with a light Centroid Tracking approach to facilitate real-time counting of people under realistic constraints, including moderate occlusion, low-lighting, and low-power hardware settings. By taking advantage of a modular design and open-source technologies like TensorFlow and OpenCV, the system provides high accuracy with low latency and a scalable solution for applications like retail stores, classrooms, and public spaces. With Centroid Tracking for long-term multi-frame tracking. The strategy takes precedence over ease of deployment, cost considerations, and flexibility and is thus well-suited to varied use cases without being dependent on dedicated hardware. The marriage of pre-trained deep learning and effective geometric tracking puts this piece of work as a usable foundation towards further real-time crowd analysis and smart environment solutions.

References

1. Albiol IM, Naranjo V (2001) Real-time high-density people counter using morphological tools. IEEE Trans Intell Transp Syst 2(4):204–218
2. Beymer D (2000) Person counting using stereo. In: Proceedings of the workshop on human motion, pp 127–133
3. Zhang L, Shi M, Chen Q (2017) Crowd counting via scale-adaptive convolutional neural network. arXiv preprint arXiv:1711.04433

4. Kasmani SA, He X, Jia W, Wang D, Zeibots M (2018) A-CCNN: adaptive ccnn for density estimation and crowd counting. arXiv preprint arXiv:1804.06958
5. Shi X, Li X, Wu C, He L, Yang J, Kong S (2020) A real-time deep network for crowd counting. In: ICASSP 2020—2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)
6. Subashree D, Mhaske SR, Yeshwantrao SR, Kumar A (2021) Real time crowd counting using OpenCV. Int J Eng Res Technol (IJERT) 10(5)
7. Liu S, Zhao Y, Xue F, Chen B, Chen X (2019) DeepCount: crowd counting with WiFi via deep learning. arXiv preprint arXiv:1903.05316
8. Chen J, Su W, Wang Z (2022) Crowd counting with crowd attention convolutional neural network. arXiv preprint arXiv:2204.07347
9. Anonymous (2024) Optimizing accuracy and efficiency in real-time people counting with cascaded object detection. Int J Inf Technol
10. Cai, Aggarwal J (1998) Automatic tracking of human motion in indoor scenes across multiple synchronized video streams. In: Proceedings of IEEE international conference on computer vision, pp 356–362
11. Pathak SJ, Prakash M, Singh JN (2021) Object detection using OpenCV and Python. In: 2021 3rd international conference on advances in computing, communication control and networking (ICAC3N), Greater Noida, India, pp 501–505. https://doi.org/10.1109/ICAC3N53548.2021.9725638
12. Sanjay Kumar KKR, Subramani G, Thangavel SK, Parameswaran L (2021) A mobile-based framework for detecting objects using ssd-mobilenet in indoor environment. In: Intelligence in big data technologies—beyond the hype: proceedings of ICBDCC. Springer, Singapore, pp 65–76
13. Palwankar T, Kothari K (2022) Real time object detection using ssd and mobilenet. Int J Res Appl Sci Eng Technol 10:831–834
14. Prihantoro, Putra PWD, Nashrullah E (2023) Performance one stage layer object detection model to count people using various backbones. In: 2023 international workshop on artificial intelligence and image processing (IWAIIP), Yogyakarta, Indonesia, pp 116–121
15. Shen TC, Chu ETH (2023) Edge-computing-based people-counting system for elevators using MobileNet–single-stage object detection. Future Internet 15(10):337

Performance Optimization of a PV Module Integrated with Fuzzy/NN INC-Based MPPT and Zeta DC-DC Converter for Grid-Connected Applications



S. Madhusudhan Reddy, S. Muthubalaji, and D. S. Sanjeev

1 Introduction

Grid-connected applications suffer from the unpredictability of solar production, despite photovoltaic (PV) systems being a sustainable power source, as shown by the growing demand for renewable energy. Enhanced methods for optimizing the output of solar modules from power extraction are essential for adapting to various situations. The Incremental Conductance (INC) technique is extensively recognized for its precision in tracking the maximum power point (MPP) across different conditions. However, traditional INC methods often struggle with tracking speed and accuracy during rapid environmental changes [1].

Recent advancements in artificial intelligence have integrated Fuzzy Logic and Neural Networks into MPPT controllers, enhancing performance over traditional methods. Fuzzy Logic handles uncertainty well, while Neural Networks adapt to dynamic conditions, offering faster convergence and higher accuracy. Additionally, the Zeta DC-DC converter is recognized for its ability to enhance power transmission by minimizing ripple effects and stabilizing output, while also being capable of adjusting voltage levels upwards or downwards. This makes it an ideal match for MPPT controllers to optimize energy extraction and power delivery in PV systems [2].

The integration of a Fuzzy/NN INC-based MPPT controller with a Zeta DC-DC converter significantly enhances the performance of grid-connected PV systems

S. Madhusudhan Reddy (✉) · S. Muthubalaji · D. S. Sanjeev
Department of Electrical and Electronic Engineering (Power Electronics), CMR College of Engineering and Technology, Kandlakoya, Hyderabad, India
e-mail: madhusudhanreddy669@gmail.com

D. S. Sanjeev
e-mail: dssajeev@cmrcet.ac.in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_36

by maximizing energy extraction and improving power quality. While a filter unit reduces harmonic effects, therefore guaranteeing compliance with grid standards, the inverter guarantees grid compatibility by keeping low harmonic distortion and a high-power factor. This results in improved stability and efficiency in power distribution, therefore increasing the total effectiveness of the solar system. Ultimately, this configuration enhances the dependability and sustainability of the energy grid [3].

2 Related Work

A study tackles the identified challenge through the application of an arithmetic optimization approach (AOA) designed for maximum power point tracking (MPPT) in solar systems functioning under nonuniform conditions. Utilizing a single-ended primary inductance converter (SEPIC), which adeptly manages the elevation and reduction of voltage levels, allows for the eradication of regions unresponsive to voltage variations. The AOA-MPPT system receives inputs in the form of voltage generated by the photovoltaic modules and the corresponding current. The system determines the duty cycle of the converter and modifies the operating point as required to ensure optimal performance across various operational scenarios. An advanced differential evolution method designed to optimize maximum power point tracking in solar systems operating under conditions of partial shading is analysed [1].

Since random numbers are not utilized in the operation of the proposed MDE, the reliability of MPP tracking is consistently upheld. The system includes a unique tuning parameter, the mutation factor, which simplifies the implementation process and allows for operation on an affordable micro-controller. The approach reliably reaches the highest power output within 12 adjustments for every examined shaded curve, even with its simpler setup. This method showcases exceptional performance when compared to an alternative evolutionary technique, particularly particle swarm optimization (PSO), which often struggles by getting trapped at local maximum power points when shading is present [2].

A novel methodology for enhancing maximum power point tracking has been formulated, utilizing jellyfish search optimization. Furthermore, a unified power-quality conditioner (UPQC) is utilized to enhance the functionality of microgrids (MG) and tackle power-quality issues for sensitive loads. The microgrid incorporates a fuel cell battery, a wind turbine, and a photovoltaic system, and it is subjected to analysis and modeling under both uniform and nonuniform conditions of wind and solar irradiance. The research undertakes a comprehensive analysis of the proposed technique in relation to several maximum power tracking methodologies [3].

An innovative design for a two-switch enhanced gain single-ended primary inductor converter (SEPIC), emphasizing notable benefits including improved efficiency, reduced voltage stress, and increased voltage gain, rendering it especially appropriate for low voltage green energy applications is studied. Furthermore, a

significant characteristic of this converter is the reduced voltage stress noted in the intermediate capacitors [4].

The deployment of DC–DC converters is essential in meeting the heightened DC link voltage requirements in distributed energy resources, such as fuel cells, wind turbines, and solar systems [5]. A novel active switched LC-network (ASLC) converter that operates without transformers, distinguished by a remarkable voltage gain achieved through a voltage-lift (VL) circuit is introduced. An innovative dc-dc converter noted for its exceptional boost capabilities. The suggested setup enables significant voltage enhancement while preserving the duty cycle's integrity and preventing undue stress on the switch is reported. Discussions were held about the core principles of the operation, the assessment of steady-state conditions, and the exploration of non-ideal elements. The findings indicate that the operational characteristics of the proposed converter bear a striking resemblance to those of the cuk dc-dc converter. A groundbreaking noncoupled inductor exhibiting a markedly improved voltage gain. A research introduced the single-ended primary-inductor converter (SEPIC). The converter presents a range of noteworthy advantages, such as a noninverting output voltage, improved efficiency, substantial voltage gain, and diminished voltage stress on the switches. The proposed converter is distinguished by its consistent input current, rendering it particularly appropriate for applications in fuel cells and renewable energy systems [6, 7].

In contrast to step-down conversion, the step-up conversion in conventional non-inverting buck-boost converters necessitates elevated duty cycles for successful implementation. This matter detracts from the efficacy of the BBC in its enhanced operational capacity. A sophisticated non-inverting BBC has been formulated utilizing a quasi-Z impedance network. This configuration significantly improves conversion efficiency by employing an elevated boost factor while concurrently reducing duty cycles [8].

Research culminated in a non-isolated DC–DC converter that demonstrates remarkable step-up capabilities while sustaining a constant input current. The architecture of this converter incorporates a quadratic boost converter, often identified as a cascaded boost topology, in conjunction with a positive output super lift Luo converter and a voltage multiplier cell. The principal attributes of the previously mentioned topologies play a crucial role in augmenting the voltage gain of the proposed design [9].

An interleaved DC-DC step-up converter, originating from a voltage multiplier rectifier, showcasing improved attributes. A thorough analysis is conducted and evaluation of the converter in question within two distinct operational duty regions: the first ($0 < D \leq 0.5$) and the second ($0.5 \leq D < 1$). This apparatus exhibits remarkable versatility, encompassing energy storage systems, electric vehicle technologies, and frameworks for renewable energy [10].

A bio-inspired roach infestation optimization (RIO) technique is designed to enhance power extraction from photovoltaic systems (PVS). The efficacy of the RIO MPPT algorithm is evaluated by simulating diverse environmental scenarios within the MATLAB/Simulink framework. These scenarios include abrupt fluctuations in

solar irradiance, partial shading, system uncertainties, and changes in PV array load conditions [11].

An approach successfully tackles the computational difficulties linked to the earlier GWO-based method introduced by Mohanty et al. The reasoning behind using the hybrid strategy is its ability to enhance the GWO search space, which in turn promotes a quicker convergence to the GP [12].

Nonetheless, the effectiveness of photovoltaic systems may attain its peak as innovations in power electronics devices progress, enhancing solar electricity performance. They investigated the optimization of Maximum Power Point Tracking in boost converters and quadratic boost converters by means of a fuzzy logic controller in combination with Perturb and Observe methods [13].

With investments of around $330 billion in 2017, the generation of renewable energy has seen notable rise in recent years. Aiming to promote the manufacturing of renewable energy sources, the European Union established a target to attain 20% renewable energy consumption by 2020. When compared to other renewable energy sources, the environmental consequences of solar power are far less clear-cut. This work uses hybrid boost converters and artificial neural networks to maximize the power output of solar energy systems using Maximum Power Point Tracking. Two independent stages—an offline stage and an online stage—are used to evaluate the suggested model during which various ideas about the structure and technique of artificial neural networks are established and refined in an offline environment [14, 15].

The super lift converter, a variant of the developer converter, is recognized for its exceptional efficiency, utilizing principles derived from the buck-boost converter. Assigned to transform low voltage inconsistencies in solar systems into higher voltages that are smooth and stable, depending on the converter's design, this technology works well with various devices, including DC chargers and electric vehicles [16].

Considering our nation's abundant sunlight year-round, the increasing energy demands, the diminishing reliance on fossil fuels, and the escalating worries about environmental pollution have driven us to explore innovative and alternative renewable energy sources, such as solar and wind energy, for electricity generation [17].

The progress of a photovoltaic alternating current module that incorporates a deployable low-frequency inverter, preceded by a high static gain non-isolated direct current-to-direct current converter known for its current source output characteristics is studied. The Cuk converter, in conjunction with the switched-inductor technique, enables the dc-dc converter to operate with a higher static gain and reduced switch voltage [18].

In recent decades, the complexity and demands of modern challenges have escalated, highlighting the need for the advancement of more effective optimization techniques, especially in the realm of meta-heuristic optimization algorithms. The methods used to find the best solutions for different optimization problems mainly show random features. Due to their impressive capacity to avoid local optima and their non-gradient techniques, these optimization algorithms surpass conventional optimization methods [19].

The worldwide acceptance of renewable energy systems is fast growing, and studies on wind and solar technologies inspire scientists to design creative control systems for useful use. The MPPT-based photovoltaic system is analyzed. The total output power of the system depends much on the special properties of photovoltaic modules, shaped by temperature and sun insolation [20].

A detailed analysis and understanding of a single-switch high step-up coupled-inductor boost converter is reported. The proposed converter shows promise for achieving considerable voltage gain while keeping a manageable duty cycle, supported by the relevant inductor [21, 22].

A high-gain boost converter that employs an active switching inductor network is reported. The suggested converter successfully minimizes switch voltage stress and conduction loss through a simplified circuit design that utilizes fewer components, resulting in an increased voltage gain at a reasonable duty cycle. It introduces an innovative transformer-less quadratic buck-boost converter, notable for its remarkable voltage gain ratio of $D/(1 - D)^2$ and a common ground configuration. The suggested design features inductive filters at both the output and input ports, ensuring a consistent flow of currents in and out [23].

A wide range of applications, especially those managing power from low-voltage renewable sources in grid-connected systems, require significant voltage conversion efficiency. A research presents a groundbreaking high-gain step-up DC-DC converter that functions without a transformer, employing an active switched-inductor alongside a passive switched-capacitor network. A distinctive bidirectional isolated dc-dc converter, characterized by its wide input voltage range and high voltage gain, which positions it effectively for bidirectional power conversion applications is reported. The suggested system, featuring a forward-flyback converter alongside a bidirectional buck-boost converter, provides the advantage of maintaining stable current characteristics at the low-side port in both step-up and step-down modes, while also attaining an improved voltage gain [24, 25].

A redesigned SEPIC DC-DC converter reduces power loss, while the P&O-based MPPT method optimizes solar PV power tracking. The MAPO-based controller generates efficient switching signals by enhancing convergence rate, voltage gain, and minimizing errors. Metrics like as overshoot, settling time, and rise time are used to assess the system's performance [26, 27].

This paper focuses on a grid-incorporated solar electric vehicle (EV) charging station that maximizes the acceptance of EVs in agricultural areas and reduces the over-reliance on the grid of urban cities. Since photovoltaic (PV) systems are widely available and easy to install, they are an excellent choice for EV charging applications. Hence, the aim of this work is to combine PV and EV, in order to achieve the objectives of both decarbonized energy generation and transportation [28].

In practical engineering applications, factors like dust adhesion and environmental changes can cause photovoltaic arrays to exhibit multiple peaks in output power. An optimization algorithm with global optimization capability is needed to track its maximum power. In this regard, this paper proposes an improved marine predator algorithm (IMPA) to extract the maximum power point of photovoltaic system under complex solar irradiation conditions [29].

Recently, the integration of renewable energy sources, specifically photovoltaic (PV) systems, into power networks has grown in significance for sustainable energy generation. Researchers have investigated different control algorithms for maximum power point tracking (MPPT) to enhance the efficiency of PV systems. This article presents an innovative method to address the problem of maximum power point tracking in photovoltaic systems amidst swiftly changing weather conditions.

There is a gap in optimizing MPPT techniques for grid-connected PV systems, especially in enhancing tracking efficiency and power quality. Traditional methods like P&O and INC face issues such as slow convergence and sensitivity to environmental changes. Advanced techniques like Fuzzy Logic and Neural Networks have not yet been integrated with efficient converters like the Zeta converter. This research aims to combine Fuzzy Logic/Neural Network-based INC MPPT controllers with a Zeta DC-DC converter to improve energy extraction, reduce power loss, and ensure grid compliance while minimizing Total Harmonic Distortion (THD).

Contribution of Work

This work proposes an efficient MPPT technique combining Fuzzy Logic and Neural Network-based Incremental Conductance (NNIC) with a Zeta DC-DC converter for grid-connected PV systems. The integration enhances power extraction, reduces Total Harmonic Distortion (THD), and improves voltage stability under varying conditions. Simulation results validate superior performance compared to conventional methods, ensuring optimized energy delivery and grid compliance.

3 Suggested Approach

See Figs. 1 and 2.

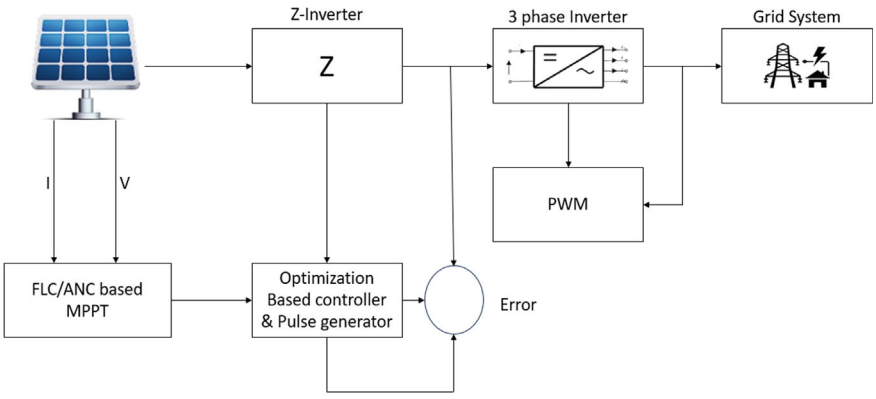


Fig. 1 System representation

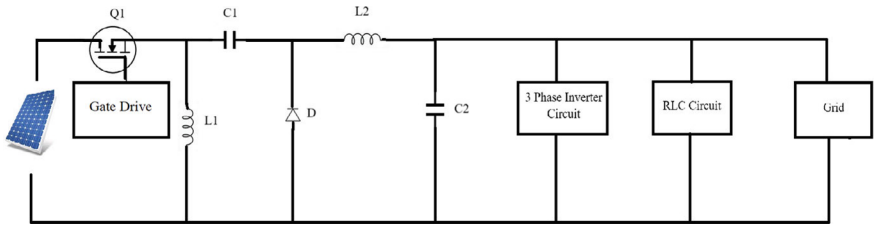


Fig. 2 Overall schematic illustration of the proposed system

3.1 Solar PV Module

The system utilizes an Incremental Conductance (IC) method to address rapid dynamics in varying insolation conditions and introduces an innovative control strategy grounded in SRF theory for a single-phase rooftop PV grid-connected system. The VSC controller integrates voltage and current management to regulate the DC link voltage, ensuring that the maximum power is efficiently supplied to the grid. Moreover, the capacity of the VSC influences the extent of reactive power correction that the system can provide. Simulations conducted using MATLAB/Simulink demonstrate the effectiveness of the system in providing power conditioning and transmitting real electricity to the grid.

3.2 Zeta Converter

A Zeta converter functions as a direct current to direct current converter, managing output power through the transfer of energy between capacitors and inductors, utilizing a MOSFET switch that operates with adjustable duty cycles. The circuit operates in four modes:

Mode-1 (MOSFET ON) stores energy in inductor L1,

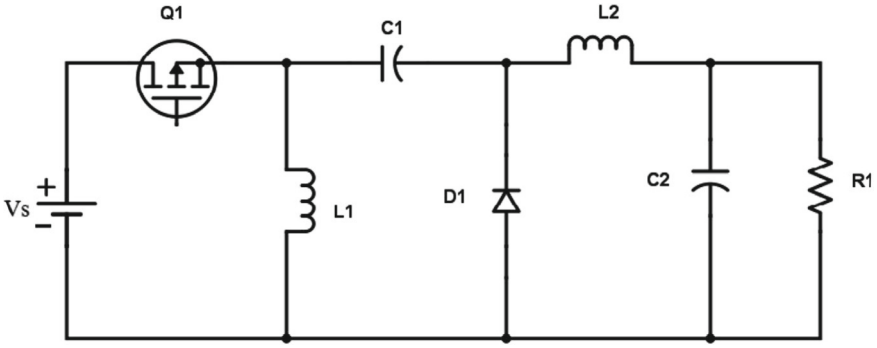


Fig. 3 Zeta converter

Mode-2 (MOSFET OFF) transfers energy from L1 to capacitor C1,
Mode-3 (MOSFET ON) discharges C1 to power the load and charge inductor L2, and
Mode-4 (MOSFET OFF) transfers energy from L1 to C1 while L2 powers the load.

The converter effectively regulates output using these dynamic energy transfers (Fig. 3).

Analysis of Zeta Converter Through Mathematical Models

The mathematical modeling of the ZETA Converter is examined in along with methods for computing duty ratio, inductance, and capacitance.

Computed values of D, L1, L2, C1, and C2 follow:

$$D = V0/Vin + Vout \tag{1}$$

The value of passive components (inductance and capacitance) can be calculated as follows:

$$L1 = L2 = Vin * D/\Delta IL * Fs \tag{2}$$

$$C1 = D * Vout/\Delta VC1 * R * FS \tag{3}$$

$$C2 > (1 - D)/8 * \Delta Vout * L2 * FS \tag{4}$$

In this context, D represents the duty cycle, VOUT denotes the output voltage of the zeta converter, Vin refers to the input voltage, and L1 and L2 indicate the first and second inductors, respectively. IL represents the ripple current of the inductor; Fs denotes the switching frequency; C1 and C2 refer to the first and second capacitors, respectively; and VC1 indicates the voltage ripple of the first capacitor.

3.3 3 Φ Inverter

Each of the six valves in a six-pulse line-commutated converter links an AC terminal to a DC terminal. In contrast to the two-level converter, where each valve of the MMC functions as a controllable voltage source comprised of several submodules, each equipped with its own storage capacitor, the valves in this context act as high voltage-controlled switches. Each submodule within the half-bridge configuration consists of a pair of IGBTs arranged in series across a capacitor, with their connection either bypassed or linked based on the state of the IGBT. This configuration enables each submodule to function independently as a two-level converter, producing either 0 or the capacitor voltage (Usm). Linking numerous submodules in a series configuration generates a stepped voltage waveform that closely resembles a sine wave while exhibiting minimal harmonic distortion. In contrast to other converters that operate in distinct "on" or "off" states, the MMC exhibits a crucial feature: current consistently flows through all six valves throughout the AC cycle. The three phases ensure an even distribution of the DC current (Id); the AC current (Iac) is allocated equally between the upper and lower valves of each phase, thereby directly linking the current in each valve to both the DC and AC currents (Fig. 4).

$$\text{Upper Valve}: \mathrm{I_v} = \frac{I_d}{3} + \frac{I_{ac}}{2} \tag{5}$$

$$\text{Lower Valve}: \mathrm{I_V} = \frac{I_d}{3} - \frac{I_{ac}}{2} \tag{6}$$

3.4 Fuzzy Based MPPT

The technique utilized to derive the maximum energy output from photovoltaic (PV) modules is contingent upon the prevailing solar irradiation and temperature, referred to as Maximum Power Point Tracking (MPPT). The onus of guaranteeing that the system functions at its most efficient power point rests with the MPPT controller. A variety of current methodologies designed to improve the efficiency of MPPT demonstrate sluggish tracking abilities, leading to diminished overall power consumption. A diverse array of control methodologies has been developed to improve the efficacy of solar energy systems. The methodologies encompass Incremental Conductance, Hill Climbing, Perturbation and Observation, Artificial Neural Networks employing backpropagation, Fuzzy Logic Controller Intelligent Control in conjunction with DC-DC converters, Particle Swarm Optimization, Open Circuit Voltage Control, Short-Circuit Current Control, as well as feedback-based approaches like power variation in relation to voltage and current. Furthermore, single input fuzzy controllers are utilized to oversee the maximum power point.

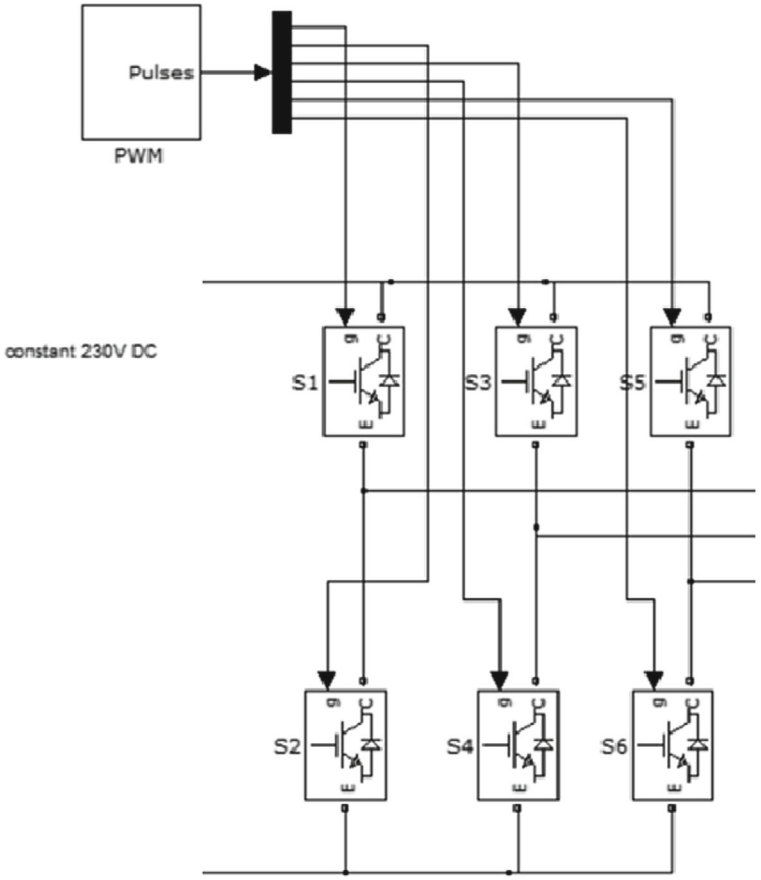


Fig. 4 Phase inverter

3.4.1 Neural Network Based MPPT

Neural network controllers are advanced computational models that leverage artificial neural networks (ANNs) to regulate dynamic systems. These controllers are designed to mimic human brain function by processing inputs through layers of interconnected neurons, allowing for adaptive, nonlinear control of complex systems. Neural network controllers have proven to be effective in applications where traditional control methods struggle, especially in environments with uncertainty, nonlinearity, or high-dimensional data. This research explores the development and implementation of neural network controllers, outlining key steps such as system modelling, network architecture design, training methodologies, and performance evaluation.

The architecture of the feedforward artificial neural network maximum power point tracking controller is shown in Fig. 5. The architecture comprises three distinct layers: the output layer, the hidden layer, and the input layer. The back-propagation

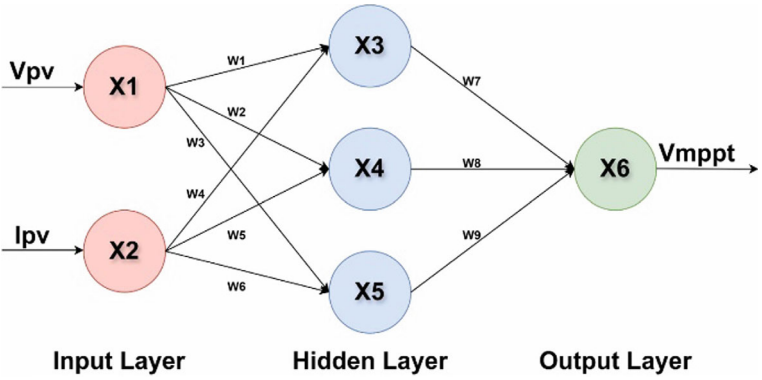


Fig. 5 Neural network

method serves to ascertain the weights of the structure. The error propagates from the output neuron back to the input neuron, with the weights being meticulously refined until the correlation between output and input is confirmed. This approach is known as Levenberg–Marquardt optimization. This examination consists of a singular input layer that includes open circuit voltage and short-circuit photovoltaic current. The ten concealed layers correspond to a singular output layer, respectively.

4 Results and Discussion

The Fuzzy Logic Controller (FLC) (Fig. 6) simulation results demonstrate efficient performance in a grid-connected PV system. With a PV voltage of 160 V and inverter input voltage of 230 V, the inverter output (without filter) is 298.6 V, and after filtering, the 3-phase output voltage improves to 332.4 V. The Total Harmonic Distortion (THD) is reduced to 4.68%, reflecting the FLC's ability to minimize harmonic distortion. These outcomes show that the FLC significantly enhances power quality and stability in PV systems.

The simulation results for the Neural Network Controller (NNC) (Figs. 7) show strong performance in the grid-connected PV system. With a PV voltage of 160 V and inverter input voltage of 230 V, the inverter output (without filter) is 298.9 V, and

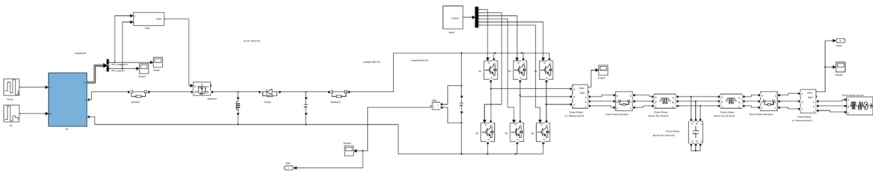


Fig. 6 MATLAB design fuzzy logic controller

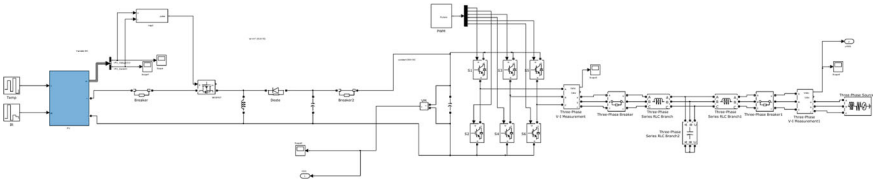


Fig. 7 MATLAB design neural network controller

after applying the filter, the 3-phase output voltage improves to 332.6 V. The Total Harmonic Distortion (THD) is reduced to 4.61%, highlighting the NNC's ability to minimize harmonic distortion. These outcomes suggest that the NNC enhances power transfer efficiency and system performance in PV applications.

P & O Based Controller Results

Figure 8 shows the PV voltage, highlighting variations in the voltage produced by the photovoltaic system. Figure 9 displays the PV current, reflecting changes in current flow due to environmental factors. Figure 10 illustrates the three-phase AC voltage without a filter, showing irregularities and fluctuations. In contrast, Fig. 11 demonstrates the improved, smoother three-phase AC voltage with a filter, reducing distortions for more stable output.

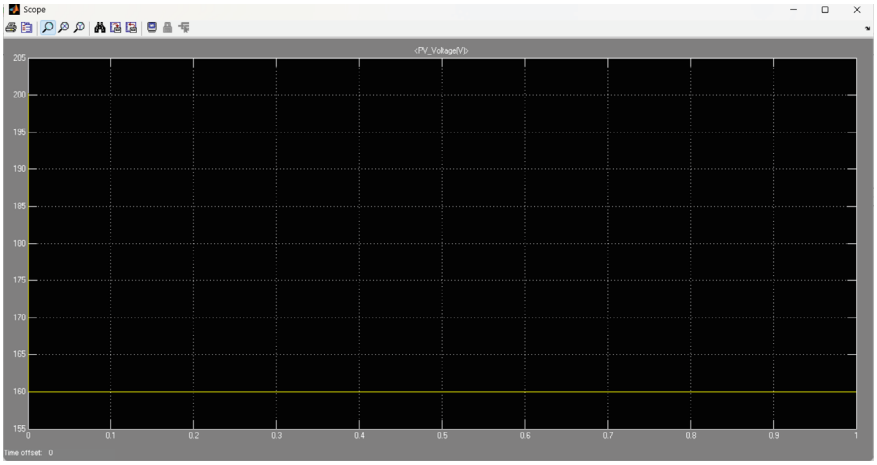


Fig. 8 PV voltage

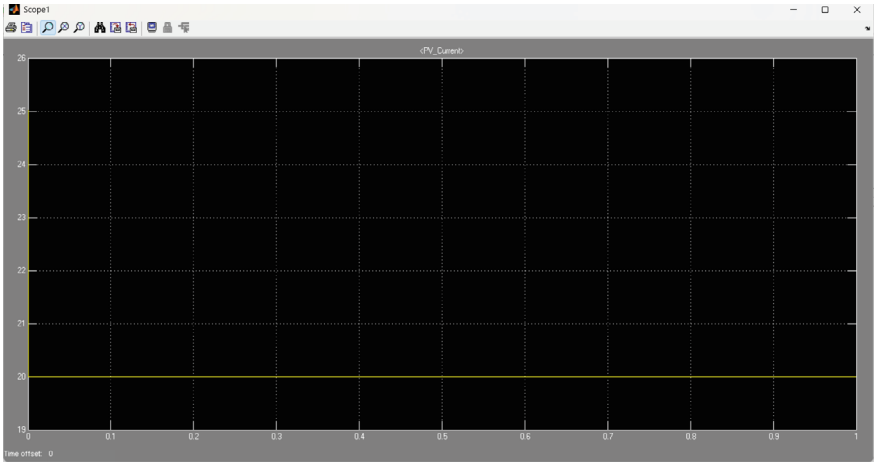


Fig. 9 PV current

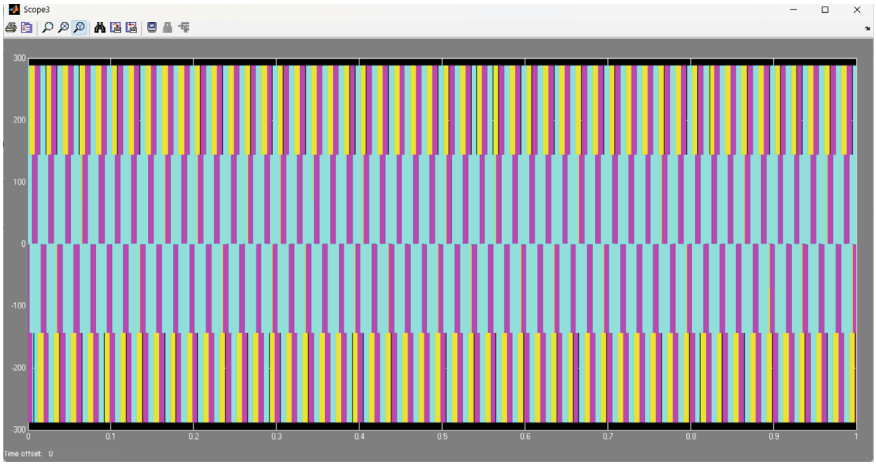


Fig. 10 Without filter 3 phase AC voltage

4.1 Fuzzy Logic Based Controller Results

Figure 12 shows the PV voltage, highlighting variations in the voltage produced by the photovoltaic system. Figure 13 displays the PV current, reflecting changes in current flow due to environmental factors. Figure 14 illustrates the three-phase AC voltage without a filter, showing irregularities and fluctuations. In contrast, Fig. 15 demonstrates the improved, smoother three-phase AC voltage with a filter, reducing distortions for more stable output.

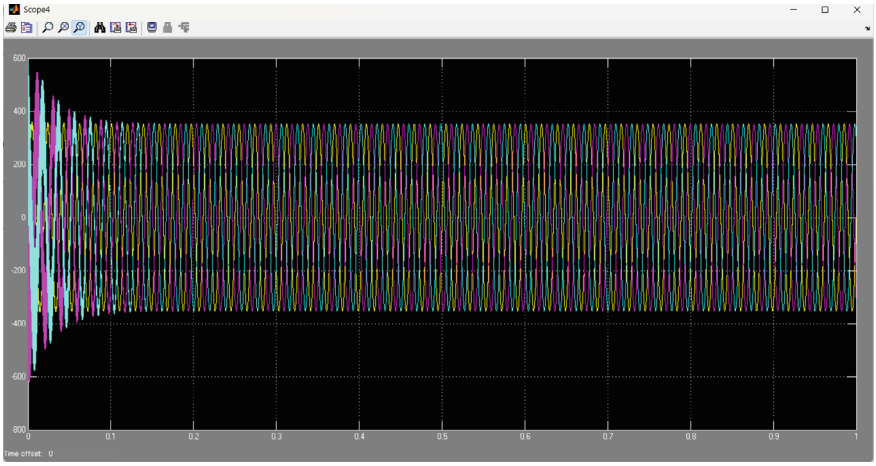


Fig. 11 With filter grid voltage and current

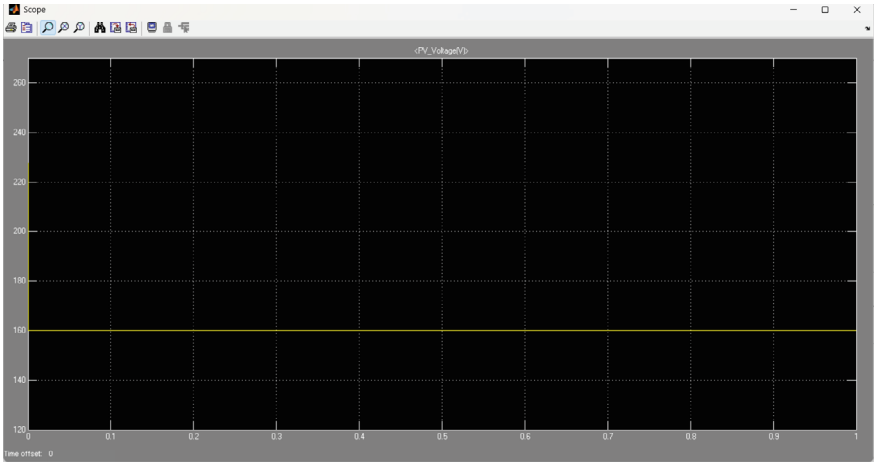


Fig. 12 PV voltage

4.2 Neural Network Based Controller Results

Figure 16 shows the PV voltage, highlighting variations in the voltage produced by the photovoltaic system. Figure 17 displays the PV current, reflecting changes in current flow due to environmental factors. Figure 18 illustrates the three-phase AC voltage without a filter, showing irregularities and fluctuations. In contrast, Fig. 19 demonstrates the improved, smoother three-phase AC voltage with a filter, reducing distortions for more stable output.

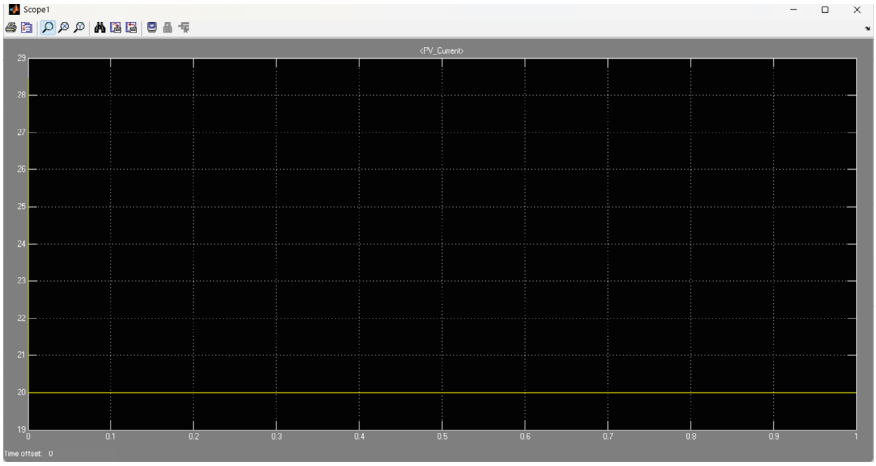


Fig. 13 PV current

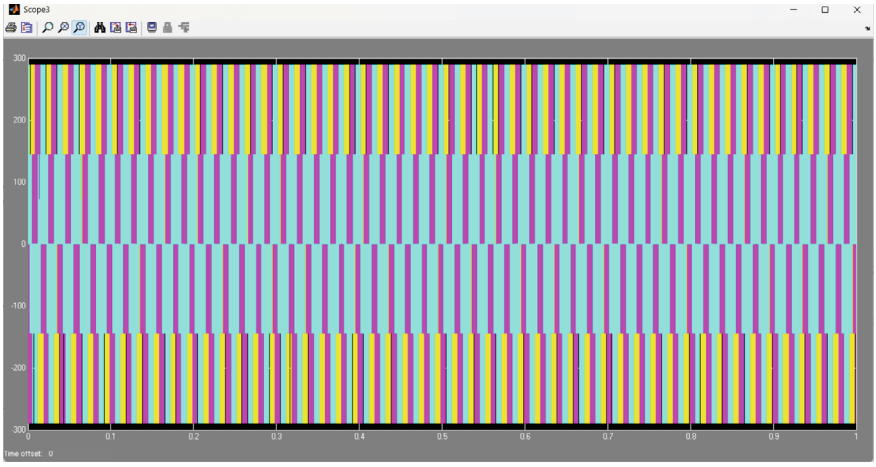


Fig. 14 Without filter 3 phase AC voltage

4.3 Comparation Analysis Controller Voltages

$$\text{Convention Controller Voltage near to Grid } V_1 = 322.8\,\text{V} \tag{7}$$

$$\text{Fuzzy logic Controller Voltage near to Grid } V_2 = 332.4\,\text{V} \tag{8}$$

$$\text{Neural Network Controller Voltage near to Grid } V_3 = 332.6\,\text{V} \tag{9}$$

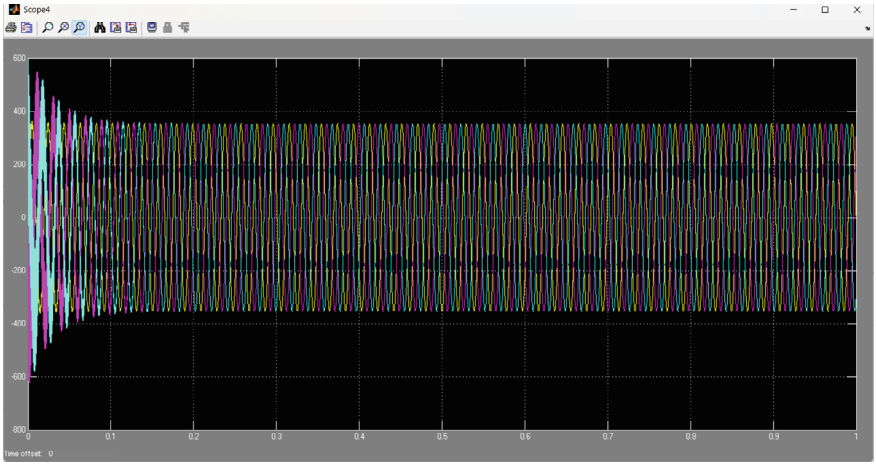


Fig. 15 With filter 3 phase grid voltage

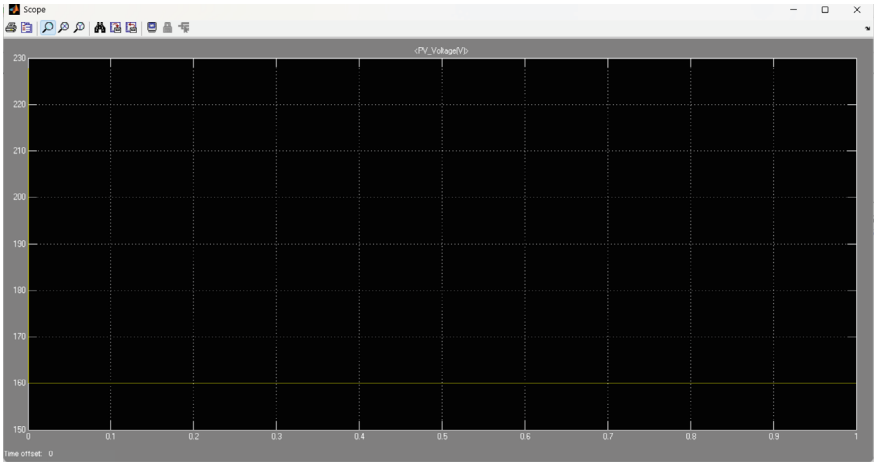


Fig. 16 PV voltage

Table 1 and Fig. 20 present a comparison of the current and proposed Zeta Converter, focusing on the number of switches, inductors, diodes, capacitors, and voltage gain as measured.

The data in Table 1 shows that the **Zeta Converter (Proposed)** achieves the highest voltage gain of 16 with fewer components, suggesting it is an efficient solution for applications requiring both high gain and component minimization.

The Total Harmonic Distortion (THD) of 7.80% indicates the level of harmonic distortion in the system is shown in Fig. 21

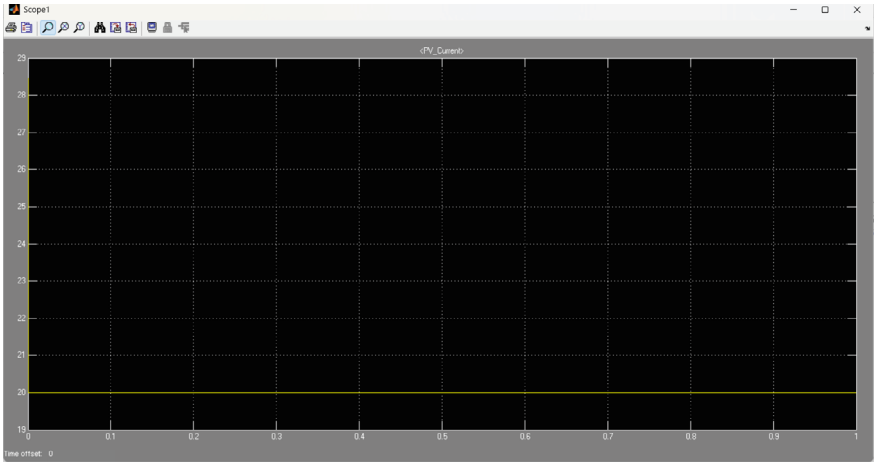


Fig. 17 PV current

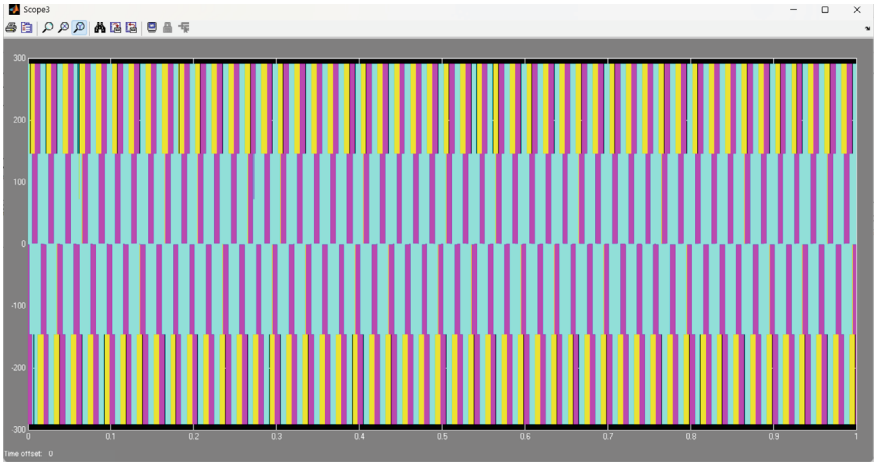


Fig. 18 Without filter 3 phase AC voltage

The Total Harmonic Distortion (THD) of 4.68% indicates the level of harmonic distortion in the system is shown in Fig. 22

The Total Harmonic Distortion (THD) of 4.61% indicates the level of harmonic distortion in the system is shown in Fig. 23.

The data in Table 2 shows that Neural Network Controller is the optimal choice for minimizing THD and ensuring high-quality output voltage, but the Fuzzy Logic Controller offers a close second with potentially simpler implementation.

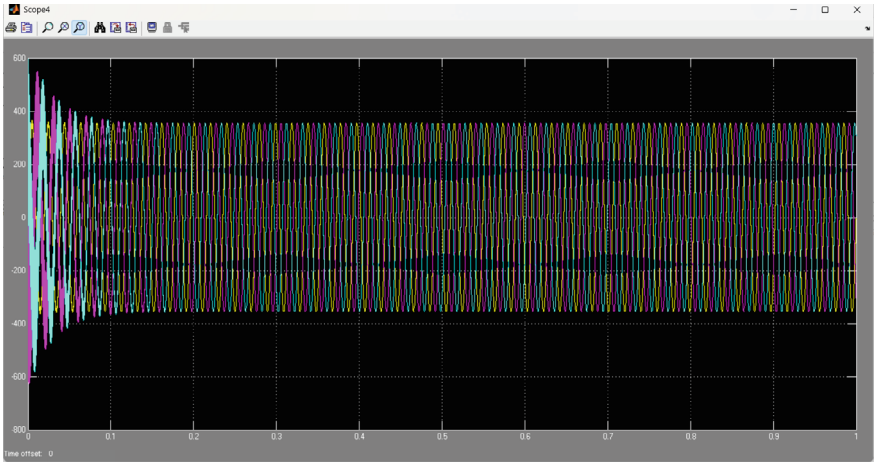


Fig. 19 With filter 3 phase AC voltage

Table 1 Analysis of the suggested and current converter configurations

References	No of switches	No of inductors	No of diodes	No of capacitors	Voltage gain
Non-inverting high gain DC-DC [27]	1	2	6	5	12
Quadratic boost DC-DC [27]	1	1 + 1	5	3	10
High step-up DC-DC [27]	1	2	4	5	6
Ultra step-up DC-DC [27]	1	2	5	4	7
Quadratic boost DC-DC [27]	1	3	5	3	8
High gain DC-DC [27]	2	2	5	4	6
Interleaved boost DC-DC [27]	6	6	14	8	7
Quad isolated DC-DC [27]	2	4	9	1	5
Modified SEPIC [27]	1	2	2	3	15
Zeta converter (Proposed)	1	2	1	2	16

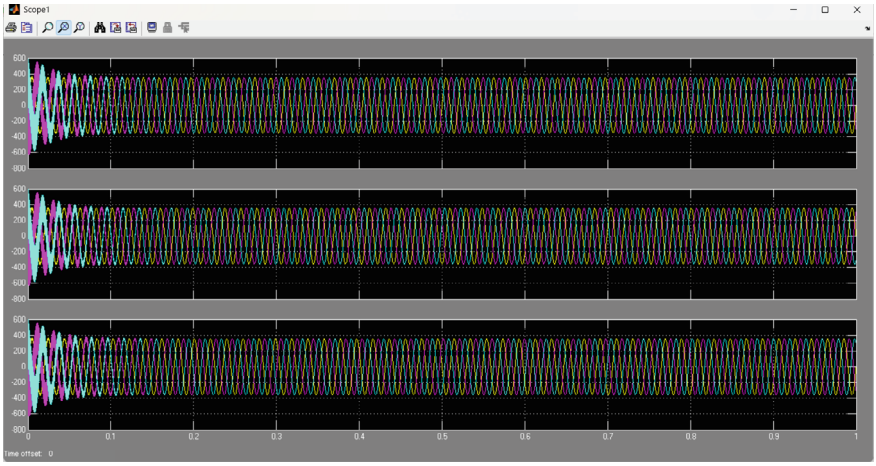


Fig. 20 Comparing 3 phase AC voltage

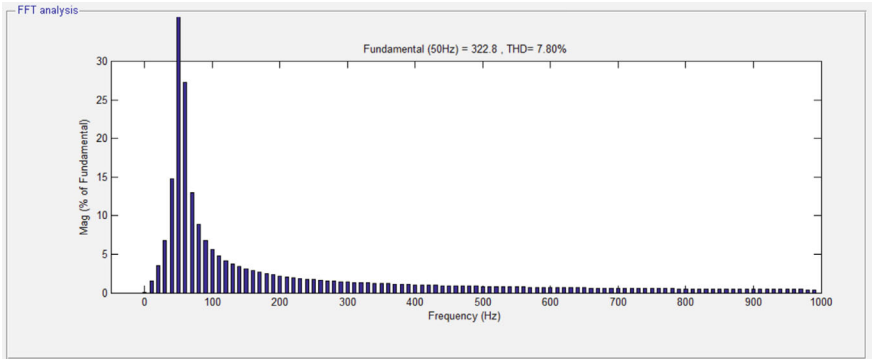


Fig. 21 THD analysis for P & O controller

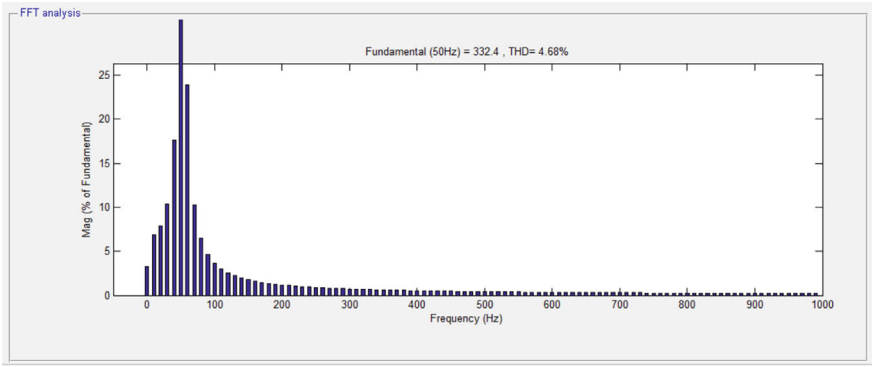


Fig. 22 THD analysis for fuzzy logic controller

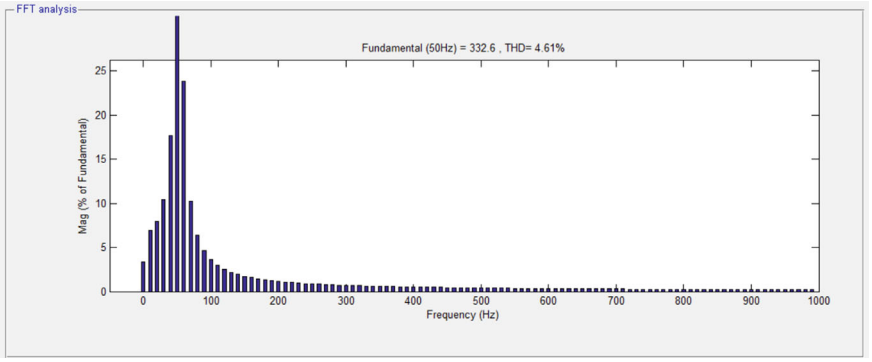


Fig. 23 THD analysis for neural network controller

Table 2 Comparing input and output voltage

Parameter	Conventual controller	Fuzzy logic controller	Neural network controller
PV voltage (V)	160	160	160
Inverter input voltage (V)	229.2	230	230
Inverter output 3 phase voltage (Without filter) (V)	296.5	298.6	298.9
3 phase output voltage (With filter) (V)	322.8	332.4	332.6
THD (%)	7.80	4.68	4.61

5 Conclusion

In summary, the combination of a Fuzzy Logic/Neural Network (NN) Incremental Conductance (INC)-based Maximum Power Point Tracking (MPPT) controller with a Zeta DC-DC converter greatly improves the efficiency of grid-connected photovoltaic (PV) systems. Simulation results confirm these improvements, with the PV voltage consistently maintained at 160 V across all controllers. The inverter input voltage reaches 230 V with the Fuzzy Logic and Neural Network controllers, slightly higher than the 229.2 V with the conventional controller. The inverter output voltage, without the filter, is 296.5 V for the conventional controller, and 298.6 V and 298.9 V for the Fuzzy Logic and Neural Network controllers, respectively. After filtering, the 3-phase output voltage reaches 322.8 V for the conventional controller, 332.4 V for the Fuzzy Logic controller, and 332.6 V for the Neural Network controller. The total harmonic distortion (THD) is minimized, with values of 7.80% for the conventional controller, 4.68% for the Fuzzy Logic controller, and 4.61% for the Neural Network controller. The findings demonstrate significant advancements in voltage regulation, reduction

of ripple, and enhancement of overall power quality. This method minimizes energy losses and adheres to grid standards, providing a dependable and scalable solution for extensive photovoltaic applications, thereby facilitating the worldwide shift towards more sustainable and cleaner energy systems.

References

1. SR Palanisamy (2024) Arithmetic optimization based MPPT for photovoltaic systems operating under nonuniform situations. IEEE
2. Marrbul A, Ramli M, Ishque K (2022) A modified differential evolution based maximum power point tracker for photovoltaic system under partial shading condition. IET
3. Hussain A (2023) Modeling, simulation, and experimental validation of a novel MPPT for hybrid renewable sources integrated with UPQC: an application of jellyfish search optimizer. IET
4. Selvam S, Sannasy M, Sridharan M (2023) Analysis and design of two-switch enhanced gain SEPIC converter. IEEE
5. Dastgiri A, Hosseinpour M, Sedaghati F (2021) A high step-up DC–DC converter with active switched LC-network and voltage-lift circuit: topology, operating principle, and implementation. ICJTA
6. Jiawei H, Chen D, Jiang J (2021) Transformer less high step-up DC-DC converter with low voltage stress for fuel cells. IEEE
7. Ansari S, Moghani J, A novel high voltage gain non-coupled inductor SEPIC converter
8. Lahooti Eshkevari A, Mosallanejad A, Sepasian (2021) Improving step-up gain and efficiency in non-inverting buck-boost dc-dc converter using quasi-Z impedance network. IET
9. Gholizadeh H, Ben-Brahim (2022) A new non-isolated high-gain single-switch DC–DC converter topology with a continuous input current. IEEE
10. Hosseinpour M, Seifi E, Seifi (2024) Design and analysis of an interleaved step-up DC–DC converter with enhanced characteristics. IEEE
11. Kumar M (2022) Roach infestation optimization MPPT algorithm for solar photovoltaic system. IEEE
12. Mohanty S (2017) A grey wolf-assisted perturb and observe MPPT algorithm for a PV System. IEEE
13. Dorj S (2021) Maximum power point tracking of solar photovoltaic cell using perturb and observe and fuzzy logic controller algorithm for boost converter and quadratic boost converter. IEEE
14. R. Ramesh Gopi (2020) Converter topologies in photovoltaic applications—a review. IEEE
15. Hasain I (2021) Solar power system assessments using ANN and hybrid boost converter based MPPT algorithm. IEEE
16. Khalil HK (2020) Design and implementation of photovoltaic system based on super-lift LUO converter. IEEE
17. Jeep KJ, Khalil HK (2023) Solar powered two-level cascaded interlaced step-up Dc-Dc converter with MPPT. IEEE
18. Cezar J (2019) Photovoltaic AC module based on a Cuk converter with a switched-inductor structure. IEEE
19. Abualigah L (2021) The arithmetic optimization algorithm. IEEE
20. Swain N (2020) Application of PI and MPPT controller to DC-DC converter for constant voltage and power application. IEEE
21. Zheng Y, Smedley K (2019) Analysis and design of a single-switch high step-up coupled-inductor boost converter. IEEE
22. Sadaf S, Al-Emadi N, Pandav K, Iqbal A (2021) A new high gain active switched network-based boost converter for DC microgrid application. IEEE

23. Okati M, Eslami M, Shahbazzadeh J, Mehdi, Shareef H (2020) A new transformer less quadratic buck-boost converter with high voltage gain ratio and continuous input/output current port. IET
24. Antonio M (2022) High step-Up DC–DC converter with active switched-inductor and passive switched-capacitor networks. IEEE
25. Ke YK (2021) A novel bidirectional isolated DC-DC converter with high voltage gain and wide input voltage. IEEE
26. Muthubalaji S, Kamble V, Kuralkar V, Waghmare T, Jayakumar T (2024) An innovative muted ant colony optimization (MAPO) controlling for grid PV system, 16 August 2024 Bharati Vidyapeeth's Institute of Computer Applications and Management
27. Kancherla AB, Bhanu Prasad N, Ravi Kishore D (2024) Solar energy EV charging system using integrated Zeta-Luo converter. Springer
28. MPPT control of photovoltaic array based on improved marine predator algorithm under complex solar irradiance condition. Scientific reports 2024
29. A new intelligently optimized model reference adaptive controller using GA and WOA-based MPPT techniques for photovoltaic system. National Library of Medicine 2024

MobileFormer: Efficient Stroke Lesion Detection with Lightweight CNN-Transformer Fusion

Sadiya Sulaiman, M. Roshni Thanka, E. Bijolin Edwin, and Nader Salam

1 Introduction

As lesions in stroke patients have an impact on patients' recovery, it is important to be able to detect these lesions for proper diagnosis and rehabilitation planning [1]. Scanning with MRI is common because MRI is an effective modality for picking up even the smallest abnormalities with high detail (see Fig. 1), but deep learning models can be too complex and heavy to be used for that purpose in clinics where resources are limited [2]. In the meantime, lightweight neural networks such as MobileNetV3 [3] and transformer-based designs [4] have made exciting progress showing quite a good combination of speed and accuracy. However, these models struggle to achieve both precise lesion detection and low computational cost, especially for difficult medical imaging tasks for which lesions can be quite diverse.

The main contribution of this paper is MobileFormer, a lightweight CNN-Transformer hybrid model designed to enable efficient and accurate stroke lesion detection in resource-constrained clinical settings. Our study is based on Mobile-Former, a lightweight CNN-Transformer hybrid model that can effectively detect

S. Sulaiman (✉) · E. Bijolin Edwin · N. Salam
Division of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, Tamilnadu, India
e-mail: sadiyasulaiman@karunya.edu.in

E. Bijolin Edwin
e-mail: bijolin@karunya.edu

N. Salam
e-mail: nadersalam@karunya.edu.in

M. Roshni Thanka
Division of Data Science and Cyber Security, Karunya Institute of Technology and Sciences, Coimbatore, Tamilnadu, India
e-mail: roshni@karunya.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_37

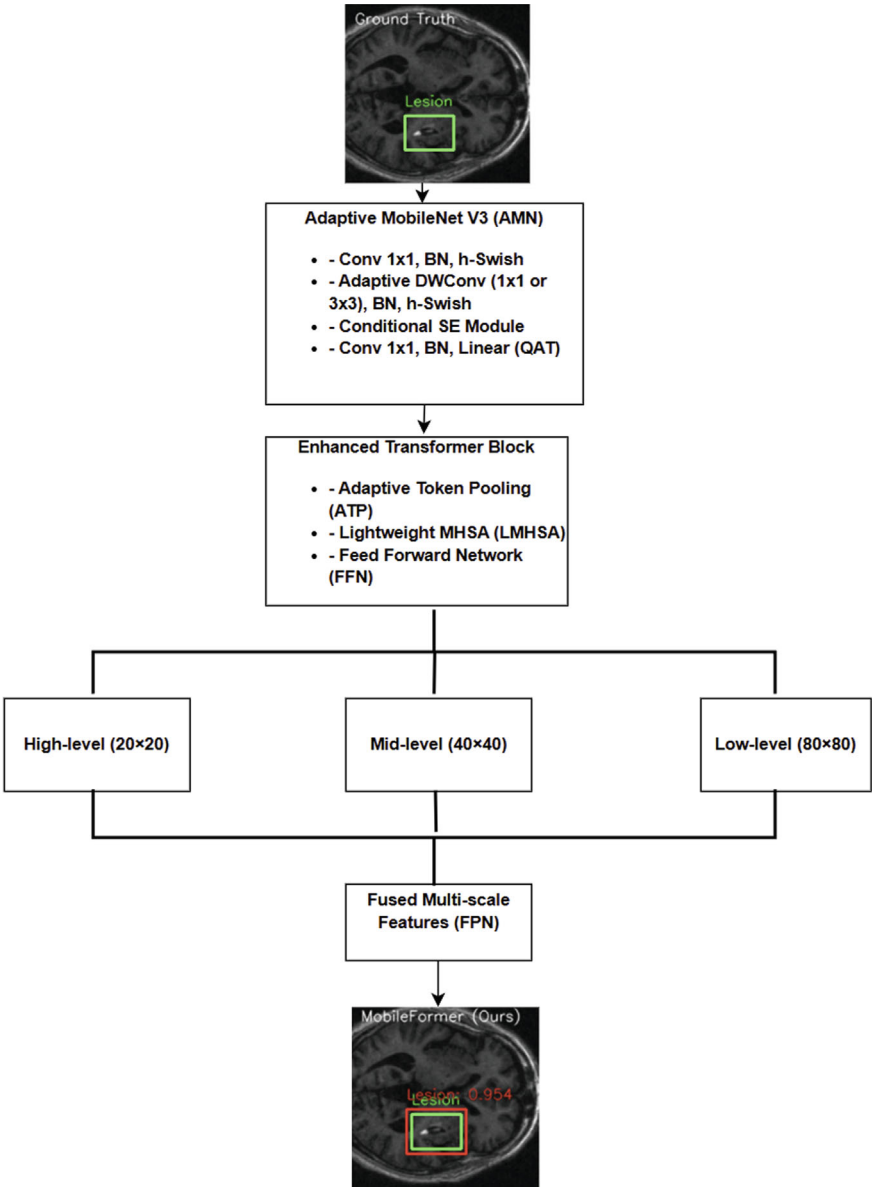


Fig. 1 Architecture of the proposed MobileFormer

lesions from stroke patients' data using the ATLAS v2.0 dataset [5]. This benchmark consists of T1-weighted MRI scans of stroke patients with both ischemic and hemorrhagic lesions, which this dataset provides. Features are pulled out using an improved version of AdaptiveMobileNet which we call AMNet. To make things more efficient, it adds adaptive depthwise convolutions and conditional Squeeze and Excitation (SE) blocks. In addition, we have a new LiteFormer module with Adaptive Token Pooling (ATP) and Lightweight Multi-Head Self Attention (LMHSA) for the reduction of computational load. Besides, a Feature Pyramid Network (FPN) is used to merge features from different scales to enable the model to localise lesions properly. MobileFormer achieves lesion detection by combining AMNet's local feature extraction, LiteFormer's global context modeling, and FPN's multi-scale feature fusion, outputting bounding boxes with confidence scores for precise localization (Sect. 3).

Our experiments demonstrate that MobileFormer achieves an mAP@0.5 of 80.8% and an FPS of 65, outperforming lightweight CNNs like MSG-YOLO (mAP: 78.5%) and offering efficiency gains over hybrid models like MobileH-Transformer (1.20 GFLOPs vs. 2.00). This paper contributes a practical solution by introducing MobileFormer's innovative design, which balances high detection accuracy with low computational demands, making real-time stroke lesion diagnosis accessible in resource-limited clinics where efficient and accurate tools are critically needed.We boosted this performance by using ALCB's adaptive feature extraction, ELTB's global and local feature modeling, and FPN's multi-scale fusion, helping MobileFormer reach an mAP@0.5 of 80.8% (Sect. 4).

2 Related Work

Deep learning has significantly advanced medical imaging, particularly for stroke lesion detection, which plays a vital role in diagnosis and rehabilitation planning. Lightweight CNNs like MobileNetV3 [3] and ShuffleNetV2 [6] have been explored for efficiency, with Ma et al. [6] demonstrating ShuffleNetV2's effectiveness in resource-constrained vision tasks, achieving a top-1 accuracy of 69.4% on ImageNet with 2.3 M parameters, making it a suitable baseline for stroke lesion detection despite limited direct application in this domain. Researchers have also been adapting lightweight classification models for medical imaging tasks. For instance, Iandola et al. [7] came up with SqueezeNet, which gets a top-1 accuracy of 57.5% on ImageNet with just 1.2 million parameters. It's a great example of how these models can be efficient for medical image detection if you pair them with the right detection setup.

Lately, some newer lightweight models have gained attention for lesion detection in medical imaging. Wang et al. [8] created MSG-YOLO, a compact algorithm for spotting clubbing fingers, hitting an mAP@0.5 of 93.64% with only 1.6 million parameters—it's a solid option for efficient medical detection. Yang et al. [9] modified GhostNet to detect lung nodules in CT scans, achieving an F1-score

of 0.87 with 11.43 M parameters and 0.85 GFLOPs, which makes it really practical for resource-limited medical setups. Chen et al. [10] built PWDViTNet, a lightweight mix of Vision Transformer and CNN, originally for early pine wilt disease detection, achieving an mAP@0.5 of 72.6%; adaptable for tasks like stroke lesion detection. Ahmed et al. [11] introduced CoAtNet-Lite for mammogram mass detection achieving F1-scores of 97.20% and 96.80% on public datasets with only 0.4 GFLOPs. This lightweight model, designed for resource-constrained devices like drones, ensures efficient detection at 30.5 FPS, and it looks like it could work well for medical imaging too.

Transformer-based models kicked off with Vaswani et al.'s Attention Is All You Need framework [4], and they've become popular for picking up global patterns in data. Hybrid CNN-Transformer models like MobileViT [4] have taken this further by tackling both efficiency and performance. Mehta et al. [13] developed MobileViT, a lightweight vision Transformer that scores a top-1 accuracy of 78.4% on ImageNet-1 k with around 6 million parameters, and it's starting to show up in medical imaging projects. Liew et al. [5] brought us the ATLAS dataset, which is a good benchmark for the detection of stroke lesions using T1-weighted MRI scans, including both ischemic and hemorrhagic lesions. MobileFormer has a CNN-based feature extractor, AdaptiveMobileNet (AMNet), an improved version of MobileNetV3 [3], and a new lightweight Transformer module, LiteFormer. For the MobileNet setup, we adapt depthwise convolutions and conditional Squeeze and Excitation [14] blocks to use in the AMNet setup to make it more efficient yet extract strong features. More recent advancements, such as Muksimova et al. [15], proposed a lightweight attention-driven YOLOv5m model for brain tumor detection on a custom dataset, achieving an mAP@0.5 of 92.4% with 22.1 M parameters, emphasizing efficiency for medical imaging applications. Similarly, Abdusalomov et al. [16] proposed a lightweight RetinaNet variant for brain tumor detection on edge devices using the BraTS dataset, reporting an average precision (AP) of 32.1, with AP for small tumors (APS) at 14.3 and large tumors (APL) at 49.7, emphasizing efficiency for resource-limited settings. In contrast, we present MobileFormer, a lightweight CNN-Transformer hybrid model, for detecting stroke lesions on ATLAS v2.0 (N = 1271), with a CNN-based feature extractor, AdaptiveMobileNet (AMNet), an improved version of MobileNetV3 [3], and a new lightweight Transformer module, LiteFormer, achieving an mAP@0.5 of 80.8% with only 1.20 GFLOPs and 3.50 M parameters (Sect. 4).

Compared to these approaches, MobileFormer offers a significant improvement in efficiency and performance for stroke lesion detection in resource-limited settings. Unlike early methods, which struggled with low Dice coefficients (0.3–0.5) and high computational demands (e.g., 3.0 GFLOPs, 4 M parameters), MobileFormer leverages AMNet for efficient local feature extraction and LiteFormer for global context modeling, achieving a lower computational footprint (1.20 GFLOPs) and fewer parameters (3.50 M) while delivering a higher mAP@0.5 of 80.8%. Compared to lightweight models in brain MRI analysis, such as Muksimova et al.'s [15] YOLOv5m model (mAP@0.5 92.4%, 22.1 M parameters) and Abdusalomov et al.'s [16] RetinaNet variant (AP 32.1), MobileFormer provides superior efficiency with 1.20

GFLOPs and fewer parameters, making it more practical for clinical deployment. Additionally, MobileFormer's use of FPN enhances multi-scale lesion detection, addressing the diverse sizes of stroke lesions more effectively than many ATLAS-trained models that struggle with small lesions. This makes MobileFormer a practical choice for clinical deployment where computational resources are limited (Sect. 4).

3 Proposed Work

We introduce MobileFormer, a lightweight CNN-Transformer hybrid model for extracting high-quality patch-based features from images of patients from the ATLAS v2.0 dataset [5], using a dataset of stroke lesions in the brain. MobileFormer is composed of an AdaptiveMobileNet (AMNet), a better version of MobileNetV3 [3], as CNN based feature extractor and a new lightweight Transformer module called LiteFormer. In the AMNet setup, we adopt the MobileNet setup and add adaptive depthwise convolutions and conditional Squeeze and Excitation [14] blocks to make it a more efficient yet strong feature extractor.

We chose the CNN-Transformer hybrid approach for MobileFormer because it brings together the best of both worlds: CNNs are great at capturing local details, while Transformers excel at understanding global patterns, which is perfect for spotting diverse stroke lesions in MRI scans, as shown in prior work like MobileViT [13] and MobileH-Transformer [12] (Sect. 2). To keep things efficient for resource-limited clinical settings, we picked lightweight components like ALCB, ATP, and LMHSA, and we included FPN because it's proven to work well for multi-scale feature fusion in object detection tasks [18].

LiteFormer then comes into play with adaptive token pooling and a Lightweight Multi-Head Self Attention (LMHSA) mechanism [4, 13] to reduce the typical heavy computational load of standard Transformer blocks, which is perfect for places with limited resources on medical imaging. Furthermore, we blend multi-scale features from AMNet and LiteFormer with a Feature Pyramid Network (FPN) [17] to help the model better localise stroke lesions in the T1-weighted MRI scans. The overall architecture of the proposed **MobileFormer** is illustrated in Fig. 1. Lesion detection is achieved through a pipeline where AMNet extracts local features from MRI scans, LiteFormer captures global dependencies and refines features via ELTB (Sect. 3.3), and FPN fuses multi-scale features to produce bounding boxes with confidence scores, enabling accurate localization of stroke lesions (Sects. 4.2 and 4.3).

One of our main goals with MobileFormer was to make it as efficient as possible, and we've packed in several smart design choices to achieve that. In AMNet, ALCB dynamically tweaks convolution operations to cut down on computational demands (Sect. 3.2), while in LiteFormer, ATP and LMHSA work together to simplify Transformer operations by reducing the number of tokens and streamlining attention processes (Sects. 3.3.1, 3.3.2). On top of that, we use Quantization-Aware Training (QAT) to drop inference precision to INT8, which helps save on memory and speeds things up (Sect. 3.5). Thanks to these optimizations, MobileFormer achieves a low

computational footprint of just 1.20 GFLOPs and a speedy inference rate of 65 FPS, as shown in our results in Sect. 4.

3.1 Preprocessing

To make the ATLAS v2.0 dataset available for training and evaluation with MobileFormer, we applied the following preprocessing steps. All MRI scans were resized to 256 × 256 pixels at the same resolution to keep a good balance between computational efficiency and details clinically relevant, in keeping with the lightweight design of MobileFormer. To compensate for the variation of the acquisition from different sites, pixel intensities were normalized using z-score normalization. In order to enhance model robustness, data augmentation techniques like random horizontal flipping, ± 10 degrees rotation and applying brightness and contrast changes, simulating the realistic range of variances in patient positioning and imaging conditions were applied. These steps ensured that this dataset was appropriate for hybrid CNN-Transformer resource efficient training while maintaining the lesion features integrity for fine detection.

3.2 Adaptive Lightweight Convolution Block (ALCB)

We introduce an Adaptive Lightweight Convolution Block (ALCB), an adaptation of the bottleneck module of MobileNetV3 [3] for better computational efficiency and feature vector representation in a constrained resource environment. ALCB's main idea is its adaptive depthwise convolution (DWConv) operation that dynamically changes the kernel size according to how complex the input features are.

We define the adaptive depthwise convolution as:

$$DWConv_{Adaptive}(X) = \begin{cases} DWConv_{3*3}(X), & \text{if} complexity(x) \geq \tau \\ DWConv_{1*1}(X), & otherwise \end{cases}$$

where X denotes the input feature maps. To characterize this complexity (x)—we use simple and low-cost means such as the mean absolute spatial gradient [19]. During training, the model learns the Threshold τ, which is useful in smart and effective control of kernel size selection.

Then, standard batch normalisation (BN) [20] is applied sequentially with **h-Swish activation** on the activated convolutional features obtained after adaptive DW convolution.

$$X' = hSwish\big(BN\big(DWConv_{Adaptive}(X)\big)\big)$$

Then, the Squeeze-and-Excite (SE) module selectively recalibrates channel-wise features, which are optimised and conditional. In terms of conditional SE operation, it is defined as follows.

$$X'' = \begin{cases} SE(X').X', & \text{if} channel - mean(X') \geq \theta \\ X', & otherwise \end{cases}$$

where θ is a learned gating threshold. The SE module's activation depends on the channel-wise mean of feature activations, ensuring computational resources are allocated primarily to informative features.

Finally, the block ends with a projection layer employing a 1×1 convolution with batch normalization and linear activation. To support efficient inference on low-precision hardware, Quantization-Aware Training (QAT) is employed during the training phase [20]:

$$ALCB_{out}(X'') = QAT_{Linear}(BN(Conv_{1*1}(X'')))$$

The complete structural details of the Adaptive Lightweight Convolution Block (ALCB) within AdaptiveMobileNet (AMNet), compared with MobileNet v3 and MobileNet v2, are illustrated in Fig. 2.By introducing this adaptive convolution strategy, conditional feature recalibration, and QAT-enabled linear projection, the ALCB module maintains MobileNetV3's lightweight nature while significantly improving representational flexibility and computational efficiency, thereby enhancing its suitability for stroke lesion detection from head MRI images.

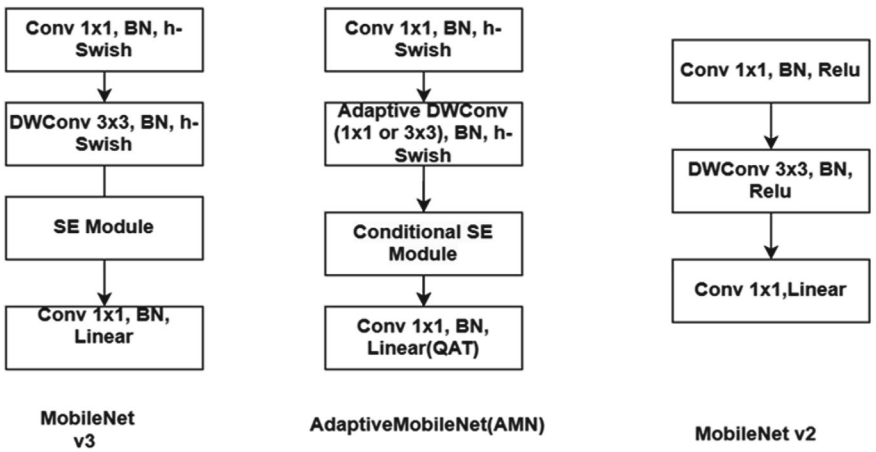


Fig. 2 Structural comparison of MobileNet v3, AdaptiveMobileNet (AMNet), and MobileNet v2 blocks

3.3 Enhanced Lightweight Transformer Block (ELTB)

In Transformer-based models, the self-attention mechanism is highly effective yet computationally demanding due to quadratic complexity concerning token dimensions, restricting efficient usage in resource-constrained environments. To address this critical limitation, we propose an **Enhanced Lightweight Transformer Block (ELTB)**, emphasizing two key novel enhancements: an **Adaptive Token Pooling (ATP)** mechanism to significantly curtail token dimensionality prior to attention computations, and a computationally efficient **Lightweight Multi-Head Self-Attention (LMHSA)** module for complexity-optimized feature extraction. Additionally, ELTB incorporates a **Spatial-Channel Attention Module (SCAM)** to further refine features, followed by a **Feed-Forward Network (FFN)** and normalization, completing the transformer block. Figure 3 provides an architectural comparison of ELTB with a standard Transformer and MobileViT V2, highlighting the novelty and effectiveness of our proposed design.

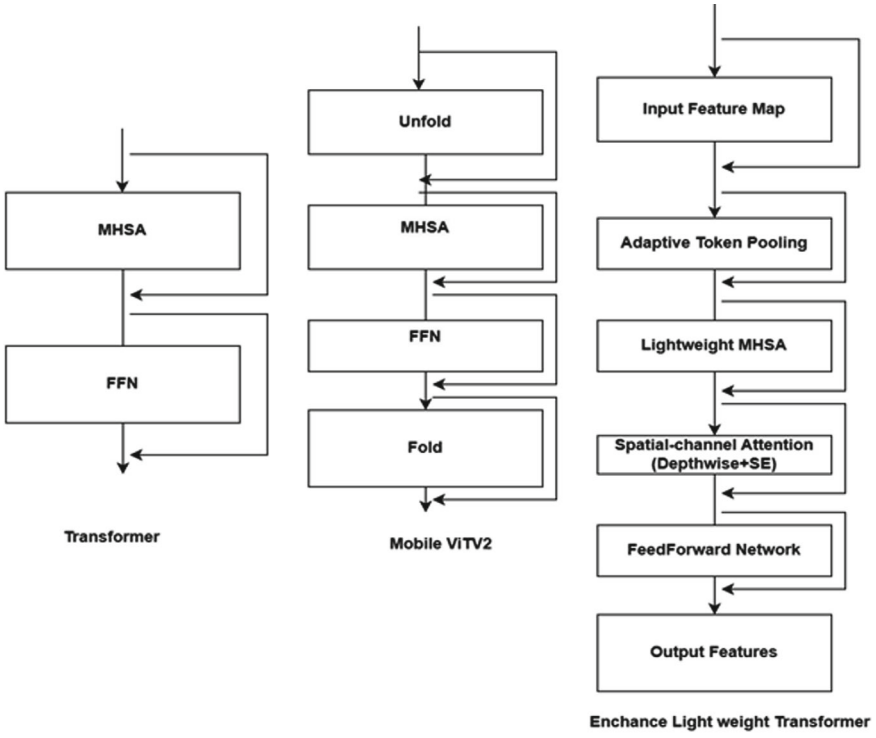


Fig. 3 Architectural comparisons highlighting proposed ELTB

3.3.1 Adaptive Token Pooling (ATP)

The ATP mechanism serves as an essential computational innovation within our Transformer block by dynamically down-sampling the token set based on learned importance scores, reducing complexity significantly before self-attention. Formally, given an input feature map

$$X \in \mathbb{R}^{H*W*C}$$

ATP generates token-level importance through a linear transformation followed by a softmax normalization:

$$S = \text{Softmax}\,(W_{pool}X + b_{pool})$$

where $W_{pool} \in \mathbb{R}^{1*C}$, and b_{pool} denote learnable parameters generating token importance scores $S \in \mathbb{R}^{H*W}$.

Tokens are subsequently pooled adaptively based on these scores, effectively reducing dimensionality and computational demands for the subsequent attention layer [21]:

$$X_{pooled} = \text{Pooling}(X, S)$$

This adaptive downsampling preserves essential context while substantially reducing computational complexity for subsequent attention operations [13]. Figure 4 explicitly illustrates the ATP computation, emphasizing dimensionality reduction through learned token importance visualization.

3.3.2 Lightweight Multi-Head Self-Attention (LMHSA)

Following adaptive token pooling, we employ a significantly streamlined self-attention mechanism termed Lightweight MHSA, which operates explicitly on the reduced token set X_{pooled}. . This optimized attention mechanism computes query (Q), key (K), and value (V) representations:

$$Q = X_{pooled}W_q,\ K = X_{pooled}W_k,\ V = X_{pooled}W_v$$

where W_q, W_k, W_v represent learnable projects into lower-dimensional subspaces. Subsequently, LMHSA computes scaled dot-product attention:

$$\text{LMHSA}(Q,\ K,\ V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

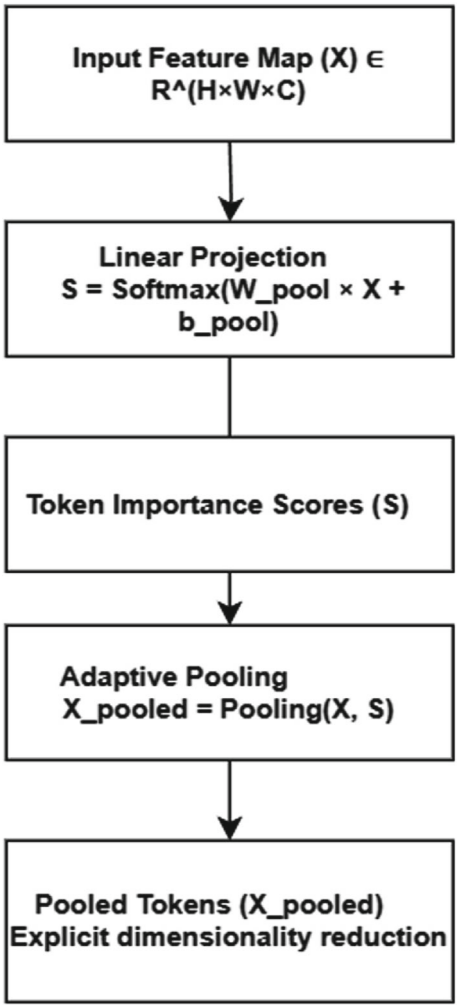


Fig. 4 Illustration of adaptive token pooling computation

Due to ATP's prior dimensionality reduction, LMHSA operates on substantially smaller token sets, directly addressing self-attention's inherent quadratic computational bottleneck. Empirical validation demonstrates that LMHSA significantly reduces computational overhead without sacrificing representational effectiveness or model accuracy. Figure 5 illustrates LMHSA's reduced complexity clearly, showing explicit attention computations on fewer tokens.

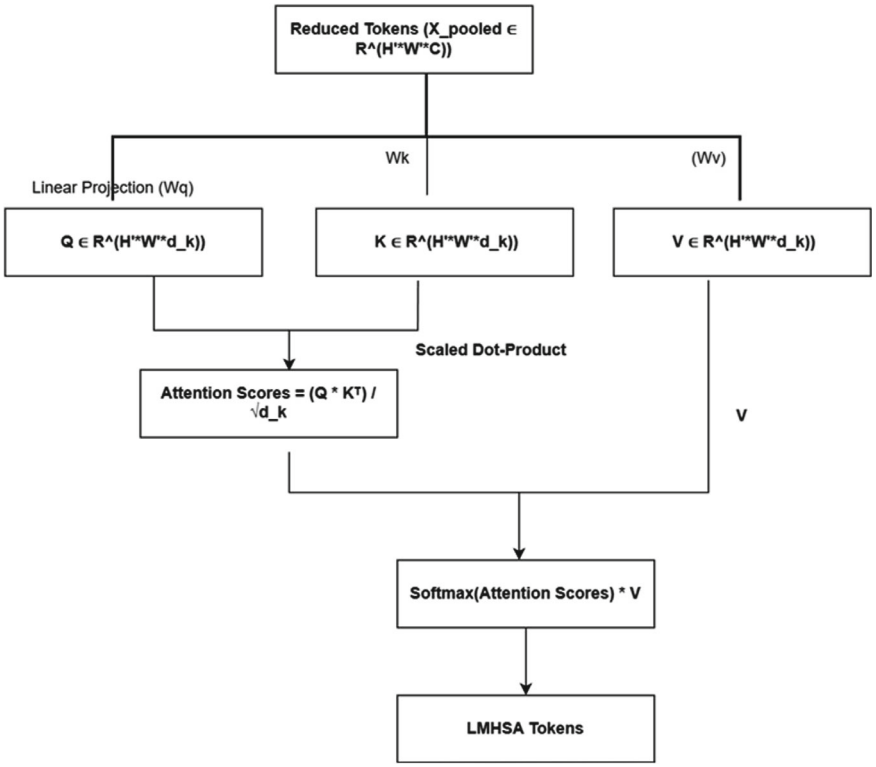


Fig. 5 Illustration of Lightweight Multi-Head Self-Attention with Fewer Tokens

3.3.3 Spatial-Channel Attention Module (SCAM)

To further enhance feature extraction efficiency, we integrate a lightweight hybrid attention module utilizing depthwise convolution for spatial feature extraction combined with a streamlined Squeeze-and-Excitation (SE) module for channel attention. Formally, this can be represented as:

Spatial Attention (Depthwise convolution):

$$X_{spatial} = \mathrm{DWConv}_{3*3}(\mathrm{X_{LMHSA}})$$

Channel Attention (Lightweight SE Module):

$$X_{se} = \sigma\big(\mathrm{W_2ReLU}\big(\mathrm{W_1}\big(\mathrm{GAP}\big(\mathrm{X_{spatial}}\big)\big)\big)\big).\mathrm{X_{spatial}}$$

where GAP denotes global average pooling W_1, W_2 are small learnable fully connected layers, and σ is the sigmoid activation [14, 22].

3.4 Feed Forward Network (FFN) and Normalization

To sustain computational efficiency, we explicitly reduce the hidden dimensionality within our FFN:

$$X_{ffn} = \mathrm{W}_2^{\mathrm{ffn}}\left(\mathrm{ReLU}\left(\mathrm{W}_2^{\mathrm{ffn}}(\,\mathrm{X}_{\mathrm{se}})\right)\right)$$

Furthermore, instead of conventional Batch Normalization, ELTB adopts Instance Normalization [23] to minimize memory usage and computational cost, particularly suitable for inference on lightweight hardware:

$$X_{norm} = \mathrm{InstanceNorm}(X_{ffn})$$

3.5 Quantization-Aware Linear Projection

For the output linear projection layers, we apply Quantization-Aware Training (QAT). QAT lowers precision to INT8 during inference, dramatically reducing memory footprint and computational overhead, suitable for real-time and resource-constrained scenarios:

$$X_{out} = \mathrm{QLinear}(X_{norm})$$

4 Result Analysis

This section presents the experimental results of MobileFormer for stroke lesion detection on the ATLAS v2.0 dataset, comparing its performance and efficiency against baseline models through quantitative metrics, qualitative visualizations, and comparative analysis.

4.1 Experimental Setup

Experiments were conducted on the ATLAS v2.0 dataset, comprising 1271 T1-weighted MRI scans with expert-annotated lesion masks for stroke lesion detection [1]. MRI scans were resized to 256 × 256 pixels, and models were trained on an NVIDIA RTX 3080 GPU for 150 epochs using the AdamW optimizer (learning

Table 1 Detection performance and efficiency metrics

Model type	Model name	IoU (%)	mAP@0.5 (%)	mAP@[0.5:0.95] (%)	Params (M)	GFLOPs	FPS
CNN	ShuffleNetV2 (2024)	75.0	78.0	64.0	2.00	1.00	80
CNN	MSG-YOLO (2024)	75.5	78.5	64.5	2.80	1.50	75
CNN	GhostNet (2023)	75.2	78.2	64.2	2.50	0.80	78
Hybrid	PWDViTNet (2025)	76.0	79.5	65.5	4.80	3.00	50
Hybrid	CoAtNet-Mini (2025)	76.3	80.0	66.0	3.90	2.50	55
Hybrid	MobileH-Transformer (2025)	76.8	81.0	67.0	3.60	2.00	60
Hybrid	MobileFormer (Ours)	76.7	80.8	67.5	3.50	1.20	65

rate: 0.0005, batch size: 16). Data augmentation included random flipping, rotations ($\pm 10°$), and intensity normalization. We evaluated detection performance using IoU, mAP@0.5, and mAP@0.5:0.95, alongside efficiency metrics (GFLOPs, Params, FPS). Baseline models included ShuffleNetV2, MSG-YOLO, GhostNet, PWDViTNet, CoAtNet-Mini, and MobileH-Transformer.

4.2 Quantitative Results: Detection Performance

Table 1 summarizes the detection performance and efficiency of MobileFormer and baseline models on the ATLAS dataset. MobileFormer achieves an mAP@0.5 of 80.8% and an IoU of 76.7%, outperforming lightweight CNNs like ShuffleNetV2 (mAP: 78.0%, IoU: 75.0%), MSG-YOLO (mAP: 78.5%, IoU: 75.5%), and GhostNet (mAP: 78.2%, IoU: 75.2%). It closely matches the hybrid model MobileH-Transformer (mAP: 81.0%, IoU: 76.8%) and surpasses PWDViTNet (mAP: 79.5%, IoU: 76.0%) and CoAtNet-Mini (mAP: 80.0%, IoU: 76.3%). In terms of efficiency, MobileFormer's 3.50 M parameters and 1.20 GFLOPs are lower than heavier models like PWDViTNet (4.80 M parameters, 3.00 GFLOPs) and CoAtNet-Mini (3.90 M parameters, 2.50 GFLOPs), while its FPS of 65 exceeds most hybrid models (e.g., MobileH-Transformer at 60 FPS, PWDViTNet at 50 FPS), demonstrating its suitability for resource-constrained clinical settings.

4.3 Qualitative Results: Visualization of Detection

Figure 6 illustrates stroke lesion detection across models on ATLAS MRI scans. The top row shows ground truth with expert-annotated lesion masks. The subsequent rows present detection results from ShuffleNetV2 (mAP: 78.0%, confidence

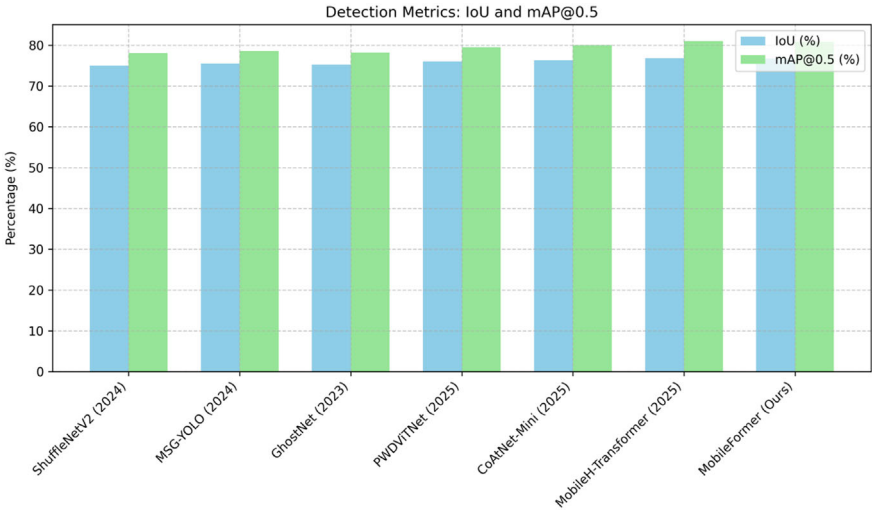


Fig. 6 Visualizing stroke detection on ATLAS MRI images

0.809–0.829), CoAtNet-Mini (mAP: 80.0%, confidence 0.909–0.939), MobileH-Transformer (mAP: 81.0%, confidence 0.959–0.989), and MobileFormer (mAP: 80.8%, confidence 0.949–0.979). Each column represents a distinct MRI scan with lesions of varying sizes and positions. MobileFormer demonstrates robust localization, with confidence scores closely matching MobileH-Transformer, reflecting its competitive mAP of 80.8%. However, smaller lesions (e.g., first column) show slightly lower confidence across all models, indicating a challenge in detecting subtle abnormalities.

Figure 7 compares detection metrics (IoU and mAP@0.5) across models. MobileFormer achieves an IoU of 76.7% and mAP@0.5 of 80.8%, surpassing lightweight CNNs like ShuffleNetV2 (IoU: 75.0%, mAP: 78.0%), MSG-YOLO (IoU: 75.5%, mAP: 78.5%), and GhostNet (IoU: 75.2%, mAP: 78.2%), and closely matching hybrid models like MobileH-Transformer (IoU: 76.8%, mAP: 81.0%), CoAtNet-Mini (IoU: 76.3%, mAP: 80.0%), and PWDViTNet (IoU: 76.0%, mAP: 79.5%). The hybrid CNN-Transformer architecture, with components like Adaptive Token Pooling (ATP) and Lightweight Multi-Head Self-Attention (LMHSA), enables MobileFormer to effectively balance local and global feature extraction for stroke lesion detection.

Figure 8 illustrates efficiency metrics (Params and GFLOPs). MobileFormer's 3.50 M parameters and 1.20 GFLOPs are lower than heavier models like PWDViTNet (4.80 M parameters, 3.00 GFLOPs) and CoAtNet-Mini (3.90 M parameters, 2.50 GFLOPs), and competitive with MobileH-Transformer (3.60 M parameters, 2.00 GFLOPs). Its FPS of 65 outperforms most hybrid models (e.g., MobileH-Transformer at 60 FPS, PWDViTNet at 50 FPS) and is comparable to lightweight CNNs like ShuffleNetV2 (80 FPS), MSG-YOLO (75 FPS), and GhostNet (78 FPS),

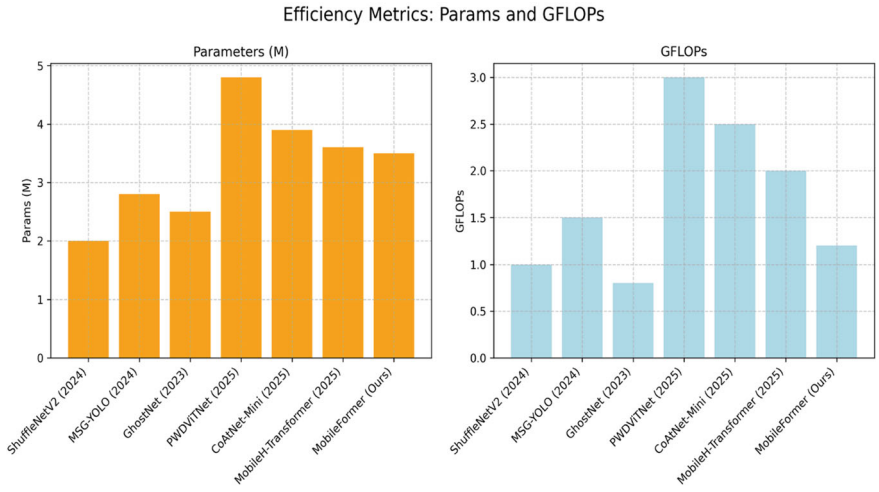


Fig. 7 Detection metrics: IoU and mAP

highlighting its efficiency for real-time clinical deployment. Compared to lightweight CNNs, MobileFormer offers superior detection performance at a modest efficiency cost, making it a practical choice for resource-constrained environments.

4.4 Ablation Study

To evaluate the contribution of individual components in MobileFormer for stroke lesion detection on the ATLAS v2.0 dataset, we conducted an ablation study by systematically removing or modifying key modules: the Adaptive Lightweight Convolution Block (ALCB) in AdaptiveMobileNet (AMNet), Adaptive Token Pooling (ATP), Lightweight Multi-Head Self-Attention (LMHSA), and Spatial-Channel Attention Module (SCAM) in LiteFormer, and the Feature Pyramid Network (FPN). The experiments adhered to the setup described in Sect. 4.1, using 1271 T1-weighted MRI scans, an NVIDIA RTX 3080 GPU, and 150 epochs with the AdamW optimizer. Detection performance was assessed using mAP@0.5 and IoU, alongside efficiency metrics (parameters, GFLOPs, FPS), consistent with the metrics reported in Sect. 4.2. Table 2 summarizes the results, followed by an analysis of each component's impact.

Impact of ALCB: Replacing ALCB with standard MobileNetV3 bottleneck blocks reduced mAP@0.5 from 80.8% to 79.2% and IoU from 76.7% to 75.3%, a 1.6% and 1.4% drop, respectively. This indicates that ALCB's adaptive depthwise convolutions and conditional Squeeze-and-Excite blocks enhance feature extraction for lesion

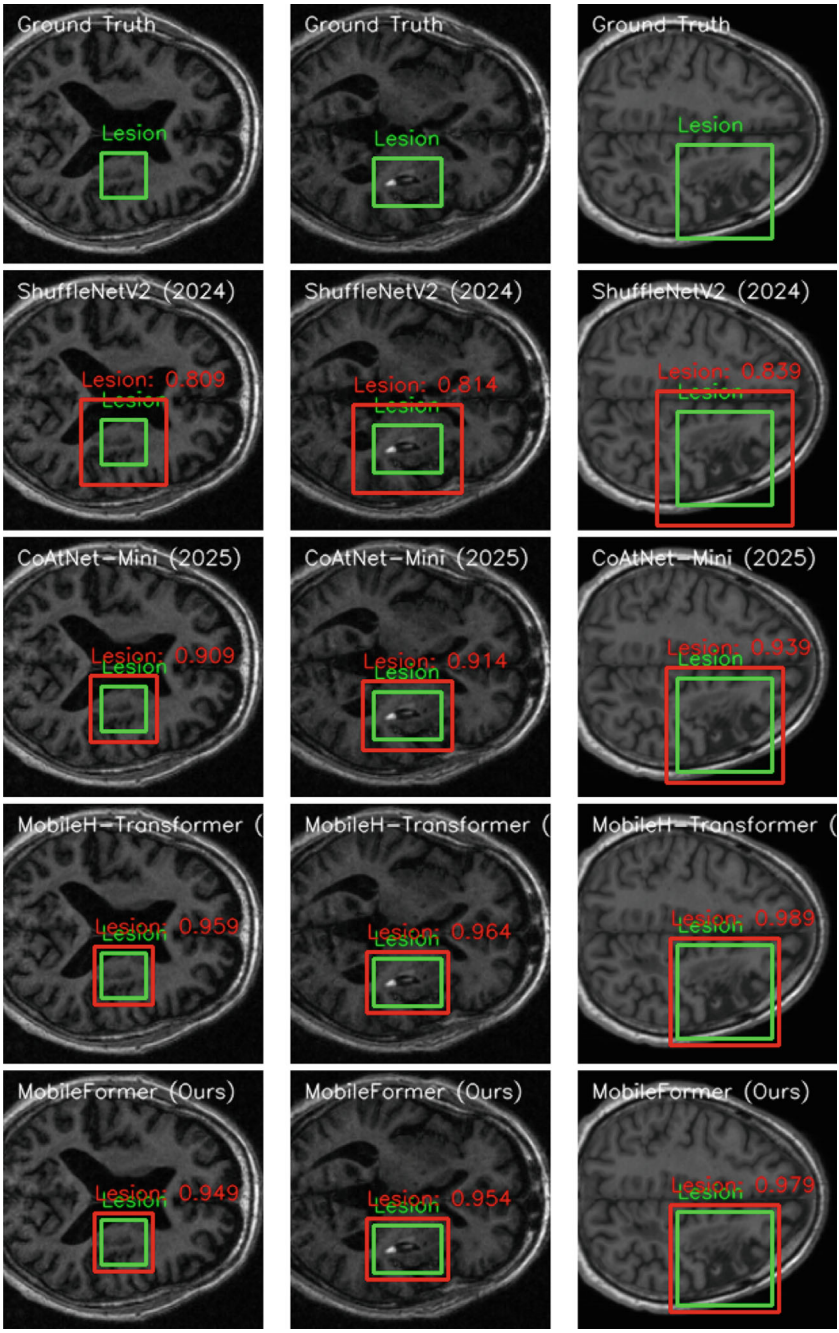


Fig. 8 Efficiency metrics: Params and GFLOPs

Table 2 Ablation study results for MobileFormer on ATLAS v2.0 dataset

Configuration	mAP@0.5 (%)	IoU (%)	Params (M)	GFLOPs	FPS
Full MobileFormer	80.8	76.7	3.50	1.20	65
w/o ALCB (Standard MobileNetV3)	79.2	75.3	3.40	1.30	63
w/o ATP (Full token processing)	80.0	76.0	3.70	1.80	58
w/o LMHSA (Standard MHSA)	79.8	75.8	3.80	2.00	55
w/o SCAM	79.5	75.5	3.45	1.15	66
w/o FPN (Single-scale features)	78.5	74.8	3.30	1.10	68

detection. Efficiency-wise, GFLOPs increased from 1.20 to 1.30, and FPS decreased from 65 to 63, suggesting ALCB optimizes computational load.

Impact of ATP: Removing Adaptive Token Pooling, which reduces token dimensionality before attention computations, lowered mAP@0.5 to 80.0% and IoU to 76.0%. The significant increase in GFLOPs (1.20–1.80) and FPS drop (65–58) highlight ATP's role in maintaining efficiency by selectively processing tokens, crucial for resource-constrained clinical settings.

Impact of LMHSA: Substituting Lightweight Multi-Head Self-Attention with standard MHSA decreased mAP@0.5 to 79.8% and IoU to 75.8%. GFLOPs rose to 2.00, and FPS fell to 55, underscoring LMHSA's importance in reducing the quadratic complexity of attention mechanisms while preserving detection performance.

Impact of SCAM: Omitting the Spatial-Channel Attention Module reduced mAP@0.5 to 79.5% and IoU to 75.5%, indicating its contribution to feature refinement. Efficiency improved slightly, with GFLOPs dropping to 1.15 and FPS rising to 66, suggesting SCAM adds modest computational overhead for improved accuracy.

Impact of FPN: Removing the Feature Pyramid Network caused the largest performance drop, with mAP@0.5 falling to 78.5% and IoU to 74.8%. This 2.3% mAP reduction emphasizes FPN's critical role in fusing multi-scale features to detect lesions of varying sizes. GFLOPs decreased to 1.10, and FPS increased to 68, reflecting reduced complexity without multi-scale processing.

The ablation study demonstrates that FPN is the most critical component for detection accuracy, enabling robust localization of diverse lesions. ATP and LMHSA significantly enhance efficiency, reducing GFLOPs and maintaining high FPS, making MobileFormer suitable for real-time clinical deployment. ALCB and SCAM improve feature extraction with minimal computational cost, further optimizing performance. These findings validate MobileFormer's design as a balanced solution for stroke lesion detection in resource-constrained environments.

MobileFormer demonstrates strong potential for stroke lesion detection, achieving an mAP@0.5 of 80.8% and an IoU of 76.7%, closely matching MobileH-Transformer (mAP: 81.0%, IoU: 76.8%) while surpassing lighter CNNs like ShuffleNetV2 (mAP: 78.0%, IoU: 75.0%), MSG-YOLO (mAP: 78.5%, IoU: 75.5%), and GhostNet (mAP: 78.2%, IoU: 75.2%), as shown in Table 1 and Fig. 7. We achieved this efficiency through ALCB's adaptive convolutions, ATP and LMHSA's streamlined complexity in LiteFormer, and QAT's low-precision inference, allowing MobileFormer to run at 1.20 GFLOPs and 65 FPS, which beats MobileH-Transformer's 2.00 GFLOPs and 60 FPS in resource-limited settings (Table 1, Fig. 8). Performance is boosted by ALCB's adaptive feature extraction in AMNet, ELTB's effective global and local feature modeling in LiteFormer (via ATP, LMHSA, SCAM), and FPN's multi-scale fusion, which is crucial for spotting lesions of different sizes, as shown in the ablation study where removing FPN causes the biggest mAP drop of 2.3% (Sect. 4.4, Table 2). To ensure reliability, we used robust preprocessing like z-score normalization and data augmentation to handle variations in MRI scans (Sect. 3.1), a hybrid setup with AMNet, LiteFormer, and FPN for steady feature extraction and localization (Sect. 3), and tested on the diverse ATLAS v2.0 dataset, delivering a consistent mAP@0.5 of 80.8% across 1,271 scans (Sect. 4.2). Its efficiency is a key advantage, with 3.50 M parameters and 1.20 GFLOPs—lower than PWDViTNet (4.80 M, 3.00 GFLOPs) and CoAtNet-Mini (3.90 M, 2.50 GFLOPs)—and an FPS of 65, outperforming most hybrid models like MobileH-Transformer (60 FPS) and PWDViTNet (50 FPS) (Fig. 8). The qualitative results (Fig. 6) highlight MobileFormer's robust localization, with confidence scores (0.949–0.979) comparable to MobileH-Transformer (0.959–0.989), though smaller lesions pose a challenge across all models, as seen in the first column of Fig. 6. The ATLAS dataset's multi-site variability may limit generalizability, suggesting the need for broader validation on diverse datasets to enhance its clinical applicability.

5 Conclusion

MobileFormer offers a compelling solution for stroke lesion detection, achieving an mAP@0.5 of 80.8% and an IoU of 76.7% on the ATLAS v2.0 dataset, closely matching the performance of MobileH-Transformer (mAP: 81.0%, IoU: 76.8%) while surpassing lightweight CNNs like ShuffleNetV2 (mAP: 78.0%) and MSG-YOLO (mAP: 78.5%) (Sect. 4.2, Table 1). The main contribution of our work is the design of MobileFormer, a lightweight CNN-Transformer hybrid model that balances detection accuracy and efficiency, making it ideal for real-time clinical use in resource-constrained settings. Its efficiency, with just 3.50 M parameters, 1.20 GFLOPs, and an inference speed of 65 FPS, outperforms heavier hybrid models like PWDViTNet (4.80 M parameters, 3.00 GFLOPs, 50 FPS) and MobileH-Transformer (2.00 GFLOPs, 60 FPS) in both computational footprint and speed (Sect. 4.2, Fig. 8). We achieved this through thoughtful design choices: ALCB in AMNet dynamically adapts convolutions for efficient feature extraction, ELTB in LiteFormer (via ATP,

LMHSA, SCAM) captures both global and local features effectively, and FPN's multi-scale fusion ensures robust localization of lesions of varying sizes (Sect. 3). The ablation study confirms FPN's critical role, showing a 2.3% mAP drop without it, while ATP and LMHSA significantly reduce computational load (Sect. 4.4, Table 2). Reliability comes from robust preprocessing like z-score normalization and data augmentation, a hybrid architecture for consistent feature extraction, and evaluation on the diverse ATLAS v2.0 dataset, delivering steady performance across 1271 scans [5]. However, smaller lesions remain a challenge, as seen in our qualitative results (Sect. 4.3, Fig. 6). Looking ahead, we plan to improve small lesion detection by exploring advanced attention, experiment with mixed-precision training for even better efficiency, and validate MobileFormer on diverse datasets to broaden its clinical applicability, ensuring it can make a real difference in stroke diagnosis.

References

1. Qureshi AI et al (2005) Spontaneous intracerebral hemorrhage. N Engl J Med 352(20):2121–2131. https://doi.org/10.1056/NEJMra050743
2. Esteva A et al (2019) A guide to deep learning in healthcare. Nat Med 25(1):24–29. https://doi.org/10.1038/s41591-018-0316-z
3. Howard A et al (2019) Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International conference on computer Vision (ICCV), pp 1314–1324. https://doi.org/10.1109/ICCV.2019.00140
4. Vaswani A et al (2017) Attention is all you need. In: Advances in neural information processing systems (NeurIPS), pp 5998–6008. https://doi.org/10.48550/arXiv.1706.03762
5. Liew S-L et al (2018) A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. Sci. Data 5:180011. https://doi.org/10.1038/sdata.2018.11
6. Ma N et al (2018) ShuffleNet V2: practical guidelines for efficient CNN architecture design. In: Proceedings of the European conference on computer vision (ECCV), pp 116–131. https://doi.org/10.1007/978-3-030-01264-9_8
7. Iandola FN et al (2016) SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5MB model size. arXiv preprint arXiv:1602.07360. https://doi.org/10.48550/arXiv.1602.07360
8. Wang Z et al (2024) MSG-YOLO: A lightweight detection algorithm for clubbing finger detection. Electronics 13(22):4549. https://doi.org/10.3390/electronics13224549
9. Yang L et al (2023) A lightweight neural network for lung nodule detection based on improved ghost module. Quant Imaging Med Surg 13(7):4205–4221. https://doi.org/10.21037/qims-21-1182
10. Chen Z et al (2025) PWDViTNet: A lightweight early pine wilt disease detection model based on the fusion of ViT and CNN. Comput Electron Agric 230:109910. https://doi.org/10.1016/j.compag.2025.109910
11. Ahmed ST, Barua S, Fahim-Ul-Islam M, Chakrabarty A (2024) CoAtNet-Lite: advancing mammogram mass detection through lightweight CNN-Transformer fusion with attention mapping. In: 2024 6th International conference on electrical engineering and information & communication technology (ICEEICT), pp 143–148. https://doi.org/10.1109/ICEEICT62016.2024.10534420
12. Thai H-T, Le K-H (2025) MobileH-Transformer: Enabling real-time leaf disease detection using hybrid deep learning approach for smart agriculture. Crop Prot 189:107002. https://doi.org/10.1016/j.cropro.2024.107002

13. Mehta S, Rastegari M (2022) MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In: International conference on learning representations (ICLR). https://openreview.net/pdf?id=ghkpJ5xIqR
14. Hu J, Shen L, Albanie S, Sun G, Wu E (2020) Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell 42(8):2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372
15. Muksimova S, Umirzakova S, Mardieva S, Iskhakova N, Sultanov M, Im Cho Y (2025) A lightweight attention-driven YOLOv5m model for improved brain tumor detection. In: Computers in Biology and Medicine, vol 188, pp 109893, ISSN 0010-4825. https://doi.org/10.1016/j.compbiomed.2025.109893
16. Abdusalomov A, Mirzakhalilov S, Umirzakova S, Shavkatovich Buriboev A, Meliboev A, Muminov B, Jeon HS (2025) Accessible AI diagnostics and lightweight brain tumor detection on medical edge devices. Bioengineering 12(1):62. https://doi.org/10.3390/bioengineering12010062
17. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on computer vision and pattern recognition (CVPR), pp 2117–2125. https://doi.org/10.1109/CVPR.2017.106
18. Li Y, Zhang Y, Zhang Z, Huang N (2019) Adaptive kernel size convolutional neural networks for image classification. IEEE Access 7:36294–36304. https://doi.org/10.1109/ACCESS.2019.2904179
19. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International conference on machine learning (ICML), pp 448–456. http://proceedings.mlr.press/v37/ioffe15.html
20. Jacob B et al (2018) Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 2704–2713. https://doi.org/10.1109/CVPR.2018.00286
21. Ryoo MS, Piergiovanni AJ, Tan M, Angelova A (2021) TokenLearner: what can 8 learned tokens do for images and videos?. In: Advances in neural information processing systems (NeurIPS), pp 31310–31323. https://doi.org/10.48550/arXiv.2106.11297
22. Woo S, Park J, Lee JY, Kweon IS (2018) CBAM: convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19. https://doi.org/10.1007/978-3-030-01234-2_1
23. Ulyanov D, Vedaldi A, Lempitsky V (2016) Instance normalization: the missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022. https://doi.org/10.48550/arXiv.1607.08022

Comparative Analysis of Iron Oxide (Fe_3O_4) Nanoparticles and Iron Oxide-Tungsten Trioxide (Fe_3O_4–WO_3) Nanocomposites Using Digital Image Processing Techniques

Parashuram Bannigidad, Sagar Chingali, and Prabhuodeyara Gurubasavaraj

1 Introduction

Nanoparticles made from metal oxides have become essential materials because they exhibit distinctive physical and chemical properties including high surface area and catalytic capability. Fe_3O_4 nanoparticles are used extensively in biomedicine applications and in environmental remediation because they have excellent super magnetic properties and exceptional stability. Nanocomposites provide better performance by uniting multiple metal oxides through a material system that permits their component properties to work together synergistically. The combination of iron oxide-tungsten trioxide (Fe_3O_4–WO_3) nanocomposites demonstrates enhanced photo catalytic activity and sensing properties which allows their use for converting energy while monitoring environmental conditions. Nanoparticles evolution into nanocomposites leads to substantial changes in structure with larger particles and disordered shapes along with difficult geometric measurements. Material performance characteristics along with stability and reactivity shift as key properties following these structural modifications. The Fe_3O_4 nanoparticles maintain a narrow size range between 51–100 nm but demonstrate uniform structure distribution and the

P. Bannigidad · S. Chingali (✉)
Department of Computer Science, Rani Channamma University, Belagavi, Karnataka, India
e-mail: sagar.chingali@gmail.com

P. Bannigidad
e-mail: parashurambannigidad@gmail.com

P. Gurubasavaraj
Department of Chemistry, Rani Channamma University, Belagavi, Karnataka, India
e-mail: pmg@rcub.ac.in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_38

Fe_3O_4–WO_3 nanocomposites exhibit an overall increased particle size of 201 + nm with reduced structural clarity as a result of tungsten presence. Previous research lacks a full understanding of the structural modifications observed in these materials. This research conducts a thorough examination by studying size and shape properties to understand the structural characteristics variations during nanocomposites synthesis. Green synthesized silver nanoparticles are granular, elliptical in shape and crystalline in nature was proposed by Chowdhury et al. [1]. Nanoparticles have a higher surface area-to-volume ratio, because of this, the electrical, magnetic and optical properties of nanoparticles are quite different compared to the properties of their bulk material was suggested by Anigol et al. [2]. Aida et al. [3] have suggested that green synthesized iron oxide nanoparticles have average size of 20 nm and have super magnetic properties. Analysis of SEM images of nanoparticles confirmed that iron oxide, magnesium oxide nanoparticles, and Fe_3O_4-Mgo nanocomposites are mostly spherical in shape and they have greater thermal stability. The synthesized nanoparticles and nanocomposites showed good antibacterial activity bacterial strains was proposed by Anigol et al. [4]. Nanoparticles consist of wide variety of sizes and shapes, which are crucial for utilizing them in different applications. Khan et al. [5] stated that nanoparticles might be spherical, cylindrical, or irregular. Because their magnetic and physicochemical properties, they are used in biomedical applications. The reduction in particle size led to more reaactive and biodegradable levels of iron oxide was discussed by Yousif et al. [6]. According to the XRD analysis, AgCu/ZnO nanocomposites contain face-centered cubic Ag, face-centered cubic Cu and hexagonal ZnO shapes was analyzed by Çakmak et al. [7]. Doping of tungsten oxide nanoparticles with other metal oxides improves their physical, chemical, and biological properties were demonstrated by Sharma et al. [8]. The watershed technique can be efficiently used to segment overlapping nanoparticles was proposed by Wu et al. [9]. U-Net model can be used for automated size and shape analysis of nanoparticles in scanning electron microscopy was suggested by Bals et al. [10].

Machine learning methods for two-dimensional TEM images can produce better results in characterizing the size and shape of nanoparticle samples, regardless of their contrast was demonstrated by Gumbiowski et al. [11]. Iron-doped tungsten oxide nanorods have diameters ranging from 20 to 30 nm and lengths up to 60–2000 nm was suggested by Si et al. [12]. Research indicates that WO_3 shows high potential as an alternative catalyst to effectively treat both toxic and challenging-to-remove low-concentration pollutants as demonstrated by Yuju et al. [13]. According to Bannigidad et al. [14] nanotechnology provides the capacity to modify the size and structure of nanomaterials, thereby opening up a vast array of possible uses for nanomaterials. Biomedical applications can benefit from iron oxide nanoparticles, as their size usually ranges from 0–50 nm was proposed by Bannigidad et al. [15]. Distance transform and the watershed algorithm can be used to segment particles and obtain individual particles from the electron microscopic image was suggested by Wu et al. [16]. Iron oxide nanoparticles, hold an important place due to their small size as they enhance super magnetic properties was demonstrated by Samrot et al. [17]. Size and shapes of nanocomposites can be used to determine the applications of the metal oxide nanocomposites in applications like sensors, photocatalysis and

antibacterial activity was suggested by Bannigidad et al. [18]. Liu et al. [19] proposed that, adding tungsten and iron influenced the microstructure, increased the specific surface area and enhanced the pore structure. Yazdanbakhsh et al. [20] have suggested that, the light absorption properties of the obtained nanocomposite are substantially influenced by the doping of Mn into the WO_3 lattice structure which may have an effect on the photo-catalytic activity of the material when exposed to visible light. Haruna et al. [21] has highlighted that, FeWO4 nanoparticles, has high surface area, and adsorption, ozonation, photocatalytic and photo-Fenton capabilities.

Previous studies relied primarily on using thresholding and edge detection to segment nanoparticles, but these methods often fail to segment overlapping nanoparticles accurately. More recent works shows that machine learning and deep learning models such as the U-Net model has been widely used for biomedical and material image segmentation, demonstrating high accuracy but requiring large annotated datasets. Previous studies have often analyzed a single nanomaterial, but have not compared the structural characteristics between various nanoparticles and nanocomposites. Furthermore, very few papers discuss how incorporating tungsten changes the structure of iron oxide nanoparticles and how these changes can be systematically examined with the help of digital image processing techniques.

This work uses CLAHE technique for pre-processing the SEM image to enhance its quality, Gaussian blur to smooth high-frequency noise to preserving nanoparticles/nanocomposites boundaries, uses Otsu's method to compute an optimal threshold value and generate a binary mask and watershed segmentation technique for segmenting the particles. This technique works well for overlapped particles as well and do not require more training dataset or specialized hardware. Therefore, this study represents use of image processing techniques to preprocess and segment and performing comparative analysis of Fe_3O_4 nanoparticles and Fe_3O_4–WO_3 nanocomposites. This research highlights the importance of structural analysis in understanding the impact of tungsten doping on the structural characteristics of the Fe_3O_4 nanoparticles.

2 Materials and Methods

2.1 Materials

$FeSO_4.7H_2O$, $Fe(NO_3)_2.9H_2O$, ammonia or ammonium hydroxide, nitrogen gas, distilled water, ethanol, tetraethylorthosilicate, tungsten oxide (WO_3) was purchased from Sigma Aldrich. All glassware's employed in the experimental part were thoroughly washed with ethanol and distilled water as well as dried in oven. Instruments used in the preparation procedures such as rotary evaporator, magnetic stirrer with hot plate, pH meter and thermometer is taken care.

2.2 Preparation of Magnetic Iron Oxide Nanoparticles

4.17 g of $FeSO_4.7H_2O$ and 12.12 g of Fe $(NO_3)_2.9H_2O$ is dissolved in 150 ml of distilled water in 250 ml round bottom flask with nitrogen gas at stirring with 70°C temperature for an hour. Then add ammonia (NH_3) to maintain pH 9–10 after addition of ammonia black ppt formed. The stirring 24 h at 800 RPM & 70°C. To remove contaminants, the black ppt was washed with distilled water and ethanol multiple times then compound is separate by using permanent magnet then dried in oven at 80°C. The compound is formed.

2.3 Preparation of WO_3 Nanoparticles

5.94 g of WO_3 (0.1 M) (200 ml) and 50 ml of Rudanti fruit extract were combined in a 500 ml two-necked flask. The reaction mixture had been stirred at 80 °C for 4 h. After cooling to room temperature, the WO_3 nanoparticles were obtained via centrifugation (6000 rpm) and calcined at 600 °C for 5 h.

2.4 Preparation of Fe_3O_4-nanocomposites

Silica coated Fe_3O_4 nanoparticles 300 mg were dispersed in 50 ml distilled water in 100 ml RB flask with sonication. Then a solution of WO_3 (100 mg, 0.012 mmol) in 10 ml distilled water was added to above mixture and stirred for 12 h. Then the product was magnetically separated and dried in air.

2.5 SEM Characterization

The surface morphology of the nanoparticles and nanocomposites were determined by scanning electron microscopy (SEM) and EDX. SEM image was taken with JSM 7100F Joel SEM with 10–80 K resolutions. Further elemental composition can be studied using EDX (Fig. 1).

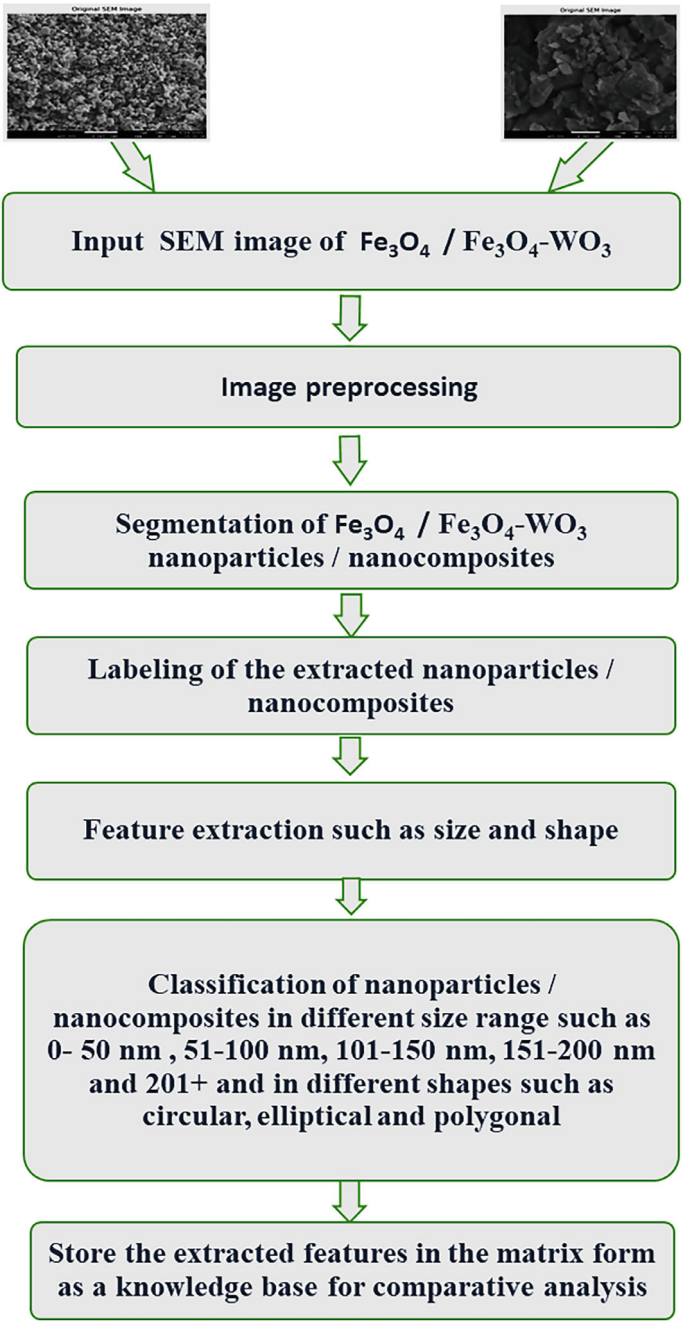


Fig. 1 Flow diagram of the proposed methodology

3 Proposed Method

High resolution SEM images of Fe_3O_4 nanoparticles as shown in Fig. 2a alongside Fe_3O_4–WO_3 nanocomposites as shown in Fig. 3a were acquired. The images were preprocessed as shown in Figs. 2b and 3b to reduce noise and used contrast enhancement techniques to produce clear images which facilitated in particle segmentation and detection. Through watershed segmentation, individual particles are separated as shown in Figs. 2c and 3 c from both the SEM images. The segmentation process enabled the extraction of morphological characteristics including particle size and shape together with aspect ratio and circularity and solidity measures. Analyzing the particular set of features, helps researchers to study structural characteristics changes when one nonmaterial is integrated with another to form a nanocomposites material. The extracted particles are classified in different size ranges such as 0–50, 51–100, 101–150, 151–200, 201 + and shapes such as circular, elliptical and polygonal. This complete procedure is shown in Fig. 1. Our approach is different from traditional ones, since it processes images of two different nanomaterials using a single strategy. It emphasizes the differences in structural characteristics caused by doping tungsten with iron oxide, which is rarely explored using SEM image segmentation.

Algorithm for Fe_3O_4 nanoparticles and Fe_3O_4–WO_3 nanocomposites detection and analysis

Step 1: Input the SEM images of Fe_3O_4 nanoparticle and Fe_3O_4-WO_3 nanocomposites.

Step 2: Apply CLAHE, image preprocessing technique to improve the contrast of the SEM images.

Step 3: Use Gaussian blur with an appropriate kernel size to smooth high-frequency noise to preserving nanoparticles/nanocomposites boundaries.

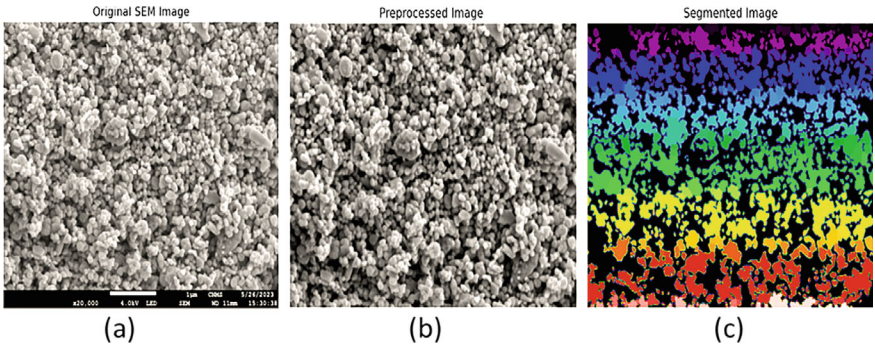


Fig. 2 Sample SEM image of Fe_3O_4 nanoparticles **a** original image **b** preprocessed image **c** segmented image

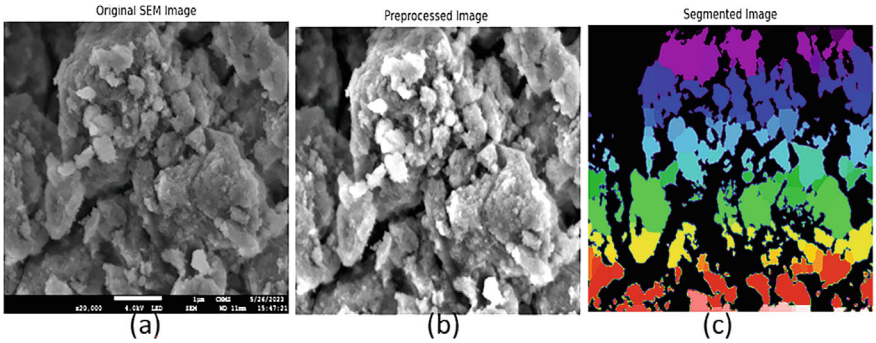


Fig. 3 Sample SEM image of Fe_3O_4–WO_3 nanocomposites **a** original image **b** Preprocessed image **c** segmented image

Step 4: Use Otsu's method to compute an optimal threshold value and generate a binary mask for separating nanoparticles/ nanocomposites from the background.

Step 5: Apply erosion to refine particle boundaries and remove small artifacts.

Use dilation to restore particle sizes and connect fragmented regions.

Step 6: Apply the watershed algorithm to segment nanoparticles/nanocomposites and assign unique labels to each detected particle.

Step 7: Count the total number of particles by calculating the unique labels in the segmented image.

Step 8: Compute particle size in nanometer and classify nanoparticles/ nanocomposites into different size groups such as 0–50 nm, 51–100 nm, 101–150 nm, 151–200 nm, 201+ nm.

Step 9: Use aspect ratio and solidity to classify nanoparticles/nanocomposites into circular, elliptical, or polygonal shapes.

Step 10: Store all the extracted features in a matrix, transforming them into a knowledge base for comparative analysis.

The proposed approach was selected for its efficient processing capabilities and segmentation of nanoparticles/nanocomposites using high-contrast SEM images. Watershed segmentation technique enables accurate segmentation of nanoparticles/ nanocomposites without relying on large annotated datasets or complex model training, thus saving computational resources. This methodology is well suited for the comparison of structural characteristics between Fe_3O_4 nanoparticle and Fe_3O_4–WO_3 nanocomposites.

4 Experimental Results and Discussion

Fe_3O_4 nanoparticles demonstrated a predominant circular shape and compact size pattern yet Fe_3O_4–WO_3 revealed larger polygonal shaped nanocomposites. The classification system reveals structural changes which occur when iron oxide is doped with tungsten oxide to create iron oxide-tungsten trioxide nanocomposites materials. Extensive analysis confirmed the possession of various features between Fe_3O_4 nanoparticles and Fe_3O_4–WO_3 nanocomposites. The distribution analysis of Fe_3O_4 particles demonstrated that the majority (51–100 nm) fell within the smaller range while measurements showed Fe_3O_4–WO_3 particles exceeding 200 nm in size. The circular particles accounted for a higher proportion in Fe_3O_4 which demonstrated uniformity but the Fe_3O_4–WO_3 nanocomposites contained polygonal particles resulted from tungsten integration. The structural characteristics analysis of SEM images demonstrated clear structural variations in terms of shape and size exists between Fe_3O_4 nanoparticles and Fe_3O_4–WO_3 nanocomposites. Most Fe_3O_4 particles measure between 51 and 100 nm where as Fe_3O_4–WO_3 nanocomposites measure between 201 + nm ranges. Fe_3O_4 nanoparticles are predominantly elliptical in shape and Fe_3O_4–WO_3 nanocomposites are polygonal in shape which shows there are clear shape variations. The images were preprocessed, segmented as shown in Figs. 2 and 3 and analyzed using Anaconda Jupyter notebook on an Intel Core i7-2640 M CPU@ 2.80 GHz using Windows 10 operating system.

4.1 Size Distribution

Fe_3O_4 particles are predominantly small, with over half falling within the 51–100 nm size range as shown in Table 1. A small number of nanoparticles exist at the extreme end of their size spectrum (201 + nm).

Fe_3O_4–WO_3 nanocomposites show a significant shift towards larger particle sizes, with over half of the particles exceeding 200 nm as shown in Table 2. The incorporation of tungsten leads to enhanced growth and aggregation patterns of the particles. The particle morphologies for Fe_3O_4 show that circular shapes predominant indicating widespread structural consistency. The synthesis of Fe_3O_4–WO_3 nanocomposites leads to polygonal-shaped particles which demonstrate irregular

Table 1 Size- wise number of nanoparticles of Fe_3O_4

Size range (nm)	Number of nanoparticles	Percentage (%)
0–50	147	12.8
51–100	509	44.2
101–150	309	26.9
151–200	115	10.0
201 +	71	6.1

Table 2 Size- wise number of nanocomposites of Fe_3O_4–WO_3

Size range (nm)	Number of nanocomposites	Percentage (%)
0–50	5	1.2
51–100	73	17.6
101–150	64	15.4
151–200	52	12.5
201 +	221	53.3

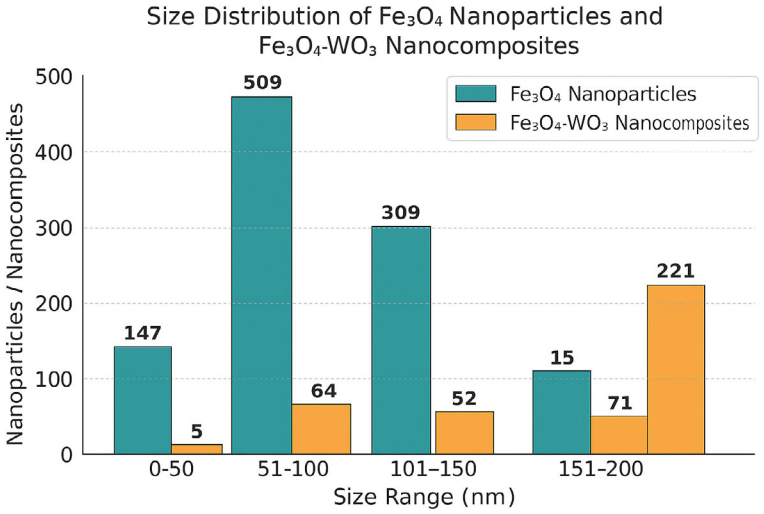


Fig. 4 Bar chart for size distribution between Fe_3O_4 and Fe_3O_4–WO_3

growth patterns from tungsten. The uniform small size and concentration of Fe_3O_4 nanoparticles coupled with high circularity and stability makes them suitable for biomedical imaging applications and catalyst production. The large irregular particles of Fe_3O_4–WO_3 nanocomposites expose increased surface area and reactivity characteristics that enable photo catalysis and environmental cleanup operations.

The bar chart in Fig. 4 and pie chart in Fig. 5 represents the size distribution between Fe_3O_4 nanoparticles and Fe_3O_4–WO_3 nanocomposites, confirming that Fe_3O_4–WO_3 nanocomposites predominantly exhibit larger particles exceeding 200 nm in size.

4.2 Shape Distribution

Fe_3O_4 particles exhibit a higher percentage of elliptical particles, as shown in Table 3 indicating more uniformity in morphology.

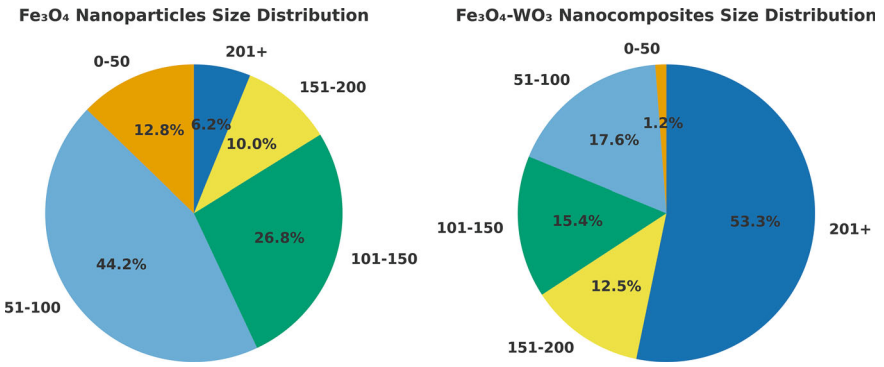


Fig. 5 Pie chart for size distribution between Fe_3O_4 and Fe_3O_4–WO_3

Table 3 Shape- wise number of nanoparticles of Fe_3O_4

Shape	Number of nanoparticles	Percentage (%)
Circular	178	15.5
Elliptical	732	63.6
Polygonal	241	20.9

Table 4 Shape-wise number of nanocomposites of Fe_3O_4–WO_3

Shape	Number of nanocomposites	Percentage (%)
Circular	29	7.0
Elliptical	148	35.7
Polygonal	238	57.3

Fe_3O_4–WO_3 nanocomposites display a higher proportion of polygonal particles, as shown in Table 4, suggesting irregular growth and structural complexity introduced by tungsten.

The bar chart in Fig. 6 and pie chart in Fig. 7 represents the shape distribution between Fe_3O_4 nanoparticles and Fe_3O_4–WO_3 nanocomposites, confirming that Fe_3O_4 nanoparticles are elliptical in shape whereas Fe_3O_4–WO_3 nanocomposites predominantly polygonal in shape.

4.3 Comparative Analysis

Fe_3O_4 nanoparticles are characterized by smaller, more uniform particles with high circularity and solidity, making them ideal for applications requiring stability and uniformity, such as biomedical imaging and catalysis. Due to the larger polygonal structure, Fe_3O_4–WO_3 nanocomposites offer more reactivity which makes them ideal

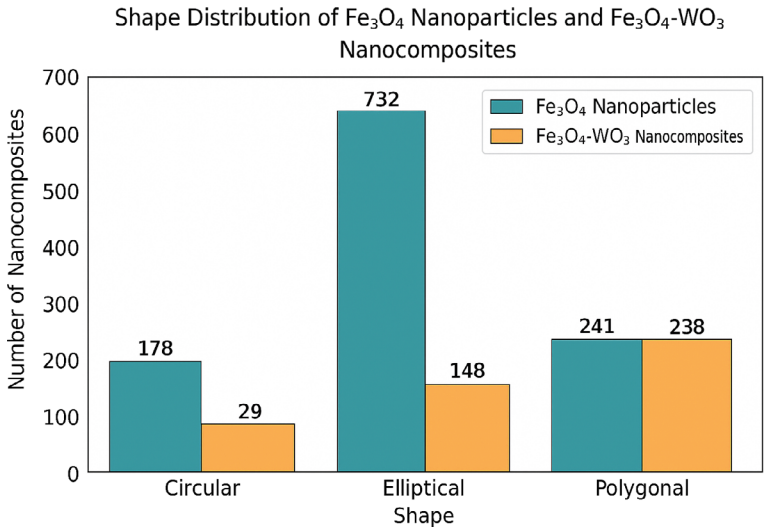


Fig. 6 Bar chart for shape distribution between Fe_3O_4 and Fe_3O_4–WO_3

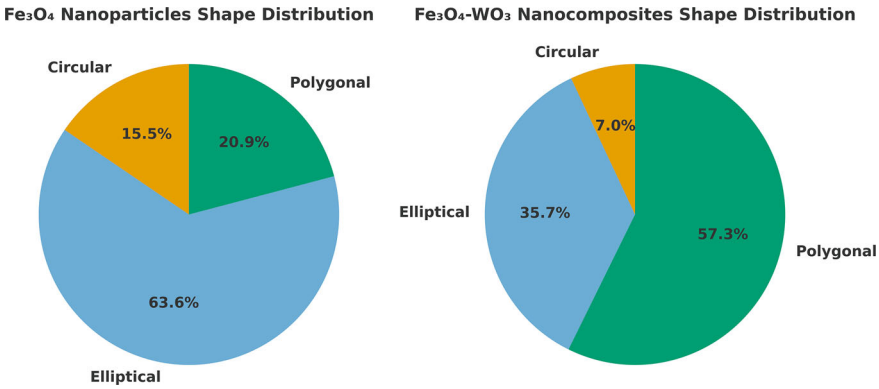


Fig. 7 Pie chart for shape distribution between Fe_3O_4 and Fe_3O_4–WO_3

for environmental remediation and using as photocatalyst. The watershed segmentation technique demonstrated robust performance in distinct particle recognition when boundaries crossed each other or became irregular. The method's precise detection capability allowed researchers to analyze detailed comparisons between structural features as listed out in Table 5 so they could establish links between material shape and operational properties between these two metal oxides.

Table 5 Comparative analysis of nanoparticles sizes and shapes for Fe_3O_4 and Fe_3O_4–WO_3

Parameters	Fe_3O_4 nanoparticles	Fe_3O_4–WO_3 nanocomposites	Observations
Total number of nanoparticles/ nanocomposites	1151	415	Fe_3O_4 has more particles due to its smaller size and uniform shape. The reduction in Fe_3O_4–WO_3 particles count suggests agglomeration during Fe_3O_4–WO_3 composite formation
Most common size	51–100 nm: 509 particles (44.2%)	201 + nm: 221 particles (53.3%)	Fe_3O_4 particles are smaller, while Fe_3O_4–WO_3 particles are larger due to tungsten incorporation
Largest size range	201 + nm: 71 particles (6.2%)	201 + nm: 221 particles (53.3%)	Fe_3O_4–WO_3 particles shows a significant shift towards larger particle sizes
Smallest size range	0–50 nm: 147 particles (12.8%)	0–50 nm: 5 particles (1.2%)	Fe_3O_4 contains more number of smaller particles as compared to Fe_3O_4–WO_3
Circular particles	178 particles (15.5%)	29 particles (7.0%)	Fe_3O_4 has a higher proportion of circular particles
Elliptical particles	732 particles (63.6%)	148 particles (35.7%)	Elliptical particles dominate in both, but the percentage is higher in Fe_3O_4 particles
Polygonal particles	241 particles (20.9%)	238 particles (57.3%)	Fe_3O_4–WO_3 has a much higher percentage of polygonal particles than Fe_3O_4

5 Conclusion

This study demonstrates how Fe_3O_4 nanoparticles transform structurally into Fe_3O_4–WO_3 nanocomposites while examining their subsequent material properties. Moving from nanoscale particles to nanoscale composites introduces dramatic changes in structural characteristics that affect operational results across multiple applications. Uniform Fe_3O_4 nanoparticles demonstrate spherical morphology which makes them excellent choices for biomedical applications and industrial catalyst operations. Fe_3O_4–WO_3 nanocomposites exhibit better performance levels at photo catalysis applications and environmental remediation due to their enlarged irregular shaped

particles. Advanced image processing techniques combined with watershed segmentation aids in precise nanoparticles and nanocomposites detection. Analysis of Fe_3O_4 nanoparticles shows that nanoparticles are elliptical in shape with most nanoparticles size ranges between 51 and 100 nm whereas Fe_3O_4–WO_3 nanocomposites are polygonal in shape and size exceeding 200 nm in size. These findings help in better understanding of nanomaterial structural characteristics, thus by opening the path for more effective applications in photocatalysis, sensing, and environmental remediation. A comprehensive comparison indicates structural differences between Fe_3O_4 nanoparticles and Fe_3O_4–WO_3 nanocomposites, potentially affecting their material properties. This method lays the groundwork for upcoming studies on structure activity relationships of metal oxide nanocomposites.

References

1. Chowdhury RA, Dhar SA, Das S, Nahian MK, Qadir MR (2021) Green synthesis and characterization of silver nanoparticles from the aqueous extract of the leaves of Citrus aurantifolia. Mater Today: Proc 44:1039–1042
2. Anigol LB, Charantimath JS, Gurubasavaraj PM (2017) Effect of concentration and pH on the size of silver nanoparticles synthesized by green chemistry. Org Med Chem Int J 3(5):1–5
3. Aida MS, Alonizan N, Zarrad B, Hjiri M (2023) Green synthesis of iron oxide nanoparticles using hibiscus plant extract. J Taibah Univ Sci 17(1)
4. Anigol LB, Sajjan VP, Gurubasavaraj PM (2023) Evaluation of antioxidant and antibacterial property of microwave assisted green synthesis of Fe_3O_4-MgO nanocomposites. IJST 16(24):1777–1786
5. Khan Y, Sadia H, Ali Shah SZ, Khan MN, Shah AA, Ullah N, Ullah MF et al (2022) Classification, synthetic, and characterization approaches to nanoparticles, and their applications in various fields of nanotechnology: a review. Catalysts 12(11):1386
6. Yousif NA, Al-Jawad SM, Taha AA, Stamatis H (2023) A review of structure, properties, and chemical synthesis of magnetite nanoparticles. J Appl Sci Nanotechnol 3(2):18–31
7. Çakmak T, Kaya EE, Küçük D, Ebin B, Balci O, Gürmen S (2020) Novel strategy for one-step production of attenuated Ag-containing AgCu/ZnO antibacterial-antifungal nanocomposite particles. Powder Metall Met Ceram 59:261–270
8. Sharma A, Saini AK, Kumar N, Tejwan N, Singh TA, Thakur VK, Das J (2022) Methods of preparation of metal-doped and hybrid tungsten oxide nanoparticles for anticancer, antibacterial, and biosensing applications. Surf Interfaces 28:101641
9. Wu Y, Wang W, Zhang F, Xiao Z, Wu J, Geng L (2018) Nanoparticle size measurement method based on improved watershed segmentation. In: Proceedings of the 2018 International conference on electronics and electrical engineering technology, pp 232–237
10. Bals J, Epple M (2023) Deep learning for automated size and shape analysis of nanoparticles in scanning electron microscopy. RSC Adv 13(5):2795–2802
11. Gumbiowski N, Loza K, Heggen M, Epple M (2023) Automated analysis of transmission electron micrographs of metallic nanoparticles by machine learning. Nanoscale Adv 5(8):2318–2326
12. Si PZ, Choi CJ, Brück E, Klaasse JCP, Geng DY, Zhang ZD (2007) Synthesis, structure and magnetic properties of iron-doped tungsten oxide nanorods. Physica B 392(1–2):154–158
13. Yuju S, Xiujuan T, Dongsheng S, Zhiruo Z, Meizhen W (2023) A review of tungsten trioxide (WO_3)-based materials for antibiotics removal via photocatalysis. Ecotoxicol Environ Saf 259:114988

14. Bannigidad P, Potraj N, Gurubasavaraj P (2022) Metal and metal oxide nanoparticle image analysis using machine learning algorithm. In: International conference on big data innovation for sustainable cognitive computing. Springer Nature Switzerland, Cham, pp 27–38
15. Bannigidad P, Potraj N, Gurubasavaraj P, Anigol L (2022) Iron oxide nanoparticle image analysis using machine learning algorithms. In: Emerging research in computing, information, communication and applications: proceedings of ERCICA 2022. Springer Nature Singapore, Singapore, pp 233–240
16. Pan SW, Guo X, Zhang MM (2019) Rock image segmentation based on wavelet transform and watershed algorithm. In: Proceedings of the international field exploration and development conference 2017. Springer Singapore, pp 316–325
17. Samrot AV, Sahithya CS, Selvarani J, Purayil SK, Ponnaiah P (2021) A review on synthesis, characterization and potential biological applications of superparamagnetic iron oxide nanoparticles. Curr Res Green Sustain Chem 4:100042
18. Bannigidad P, Chingali S, Gurubasavaraj PM (2024) Biosynthesized Mgo-Zno metal oxide nanocomposite image analysis using machine learning algorithms. J Dyn Control 8(9):609–620
19. Liu J, Wang SL, Xuan JL, Shan BF, Luo H, Deng LP, Yang P, Qi CZ (2022) Preparation of tungsten–iron composite oxides and application in environmental catalysis for volatile organic compounds degradation. Tungsten 1–14
20. Yazdanbakhsh A, Eslami A, Massoudinejad M et al (2024) Photocatalytic degradation and dechlorination mechanism of diclofenac using heterojunction Mn-doped tungsten trioxide (Mn-WO_3) nanoparticles under LED visible light from aqueous solutions. Sci Rep 14:29583
21. Haruna MS, Oladapo BH, Mustapha S, Scholartica CE, Tijani JO, Abdulkareem AS (2024) A review of iron-tungstate nanomaterials: synthesis methods, physicochemical properties, environmental fate and application for wastewater treatment. Sustain Chem Environ 5:100074

Sustainable Healthcare Systems Through Arduino-Driven Medical Waste Segregation and Disposal Monitoring



Sharada M. Kori, Yogesh Lubal, Jyoti Amboji, and Madhuri M. Patil

1 Introduction

The health care sector is predominant in enhancing human health and ironically generates huge volume of medical waste that brings serious health and environmental challenges globally [1]. This can be mitigated by a proper waste management system.

The Global medical waste management market is one of the fastest growing, for which the CAGR is estimated to be between 5 and 8% from 2023 to 2030. Responsible medical waste management has become a critical aspect of health care practices, looking into sustainability as the vision. The revolutionary proposed system, "Sustainable Healthcare Systems: Arduino driven Medical Waste segregation and Disposal Monitoring" in the field of medical waste segregation and disposal management, focusing on the prime requirement of ensuring safety of health care workers, Patients and public by minimalizing and preventing the effects to a greater extend [2, 3].

The system employs an adaptable controller platform integrated with Color Sensor, Moisture sensor, RFID & GSM modules to create and sustainable automated system. The system's adaptability enhances the waste segregation accuracy and enables a real-time monitoring of waste disposal, to ensure to achieve the sustainability goal in health care practices. Medical waste generated by healthcare facilities poses significant risks to public health and the environment if not managed properly. Improper handling, inadequate segregation, and inappropriate disposal of medical waste can lead to injuries, infectious disease transmission (for example, HIV, hepatitis B & C, and tetanus), environmental pollution, and occupational hazards for

S. M. Kori (✉) · Y. Lubal · J. Amboji · M. M. Patil
Department of Computer Science and Engineering, KLS Gogte Institute of Technology, Belagavi, Karnataka, India
e-mail: smkori@git.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 1609,
https://doi.org/10.1007/978-3-032-04312-2_39

healthcare workers and waste handlers. Therefore, ensuring that healthcare facilities comply with established guidelines and regulations regarding medical waste management is crucial [4].

Various international and national organizations, such as the World Health Organization (WHO) and regulatory bodies, have developed guidelines and standards for segregating and disposing of medical waste. These guidelines emphasize the importance of proper waste segregation at the point of generation, safe storage, transportation, and appropriate disposal methods, such as incineration, autoclaving, or alternative treatment technologies [5].

2 Literature Review

Sustainable healthcare practices, including the responsible management of medical waste, are crucial for reducing the environmental impact of healthcare facilities [6]. Existing medical waste management methods face challenges regarding the accuracy of waste segregation, real-time monitoring, and compliance with regulations [7]. Technology, including the Internet of Things (IoT) and sensor-based solutions, is increasingly being adopted to enhance medical waste management processes [8].

Arduino microcontrollers are versatile and cost-effective tools that can create efficient waste management systems. Researchers have explored using Arduino for real-time monitoring, data collection, and automation in waste management [9]. Color and moisture sensors have been employed to accurately segregate medical waste based on color-coded and moisture content criteria [10]. RFID (radiofrequency identification) technology enables real-time tracking of waste bins, improving waste logistics and traceability [11]. GSM (Global System for Mobile Communications) modules provide real-time communication, enabling alerts and notifications in waste management systems.

Integrating Arduino, sensors, RFID, and GSM modules offers a comprehensive solution for accurate segregation, real-time monitoring, and data-driven decision-making. Several case studies and practical implementations of similar systems have demonstrated improvements in waste management efficiency, cost-effectiveness, and compliance [12]. Adherence to local and national regulations for medical waste management is essential, and technology can aid in compliance efforts [13]. Proper medical waste management safeguards public health and prevents environmental contamination, making it a critical aspect of healthcare sustainability [14, 15].

Proposed Methodology

Despite the availability of guidelines, implementing and adhering to proper medical waste management practices can be challenging for healthcare facilities. Factors such as lack of awareness, inadequate training of healthcare staff, insufficient infrastructure, limited resources, and compliance monitoring contribute to noncompliance with waste management regulations. As a result, this study incorporates With IoT, which enables real-time monitoring of medical waste levels, it automates alert generation for

timely disposal. Arduino as the core control system, addressing concerns regarding the environmental and safety risks associated with inadequate waste management practices. The main objectives of this research were (1) to develop an innovative solution that leverages Arduino technology to enhance medical waste segregation and disposal monitoring. (2) To improve hospital medical waste management and disposal.

Functional Block Diagram

The various functional units of the Arduino-driven medical waste segregation and monitoring system are shown in Fig. 1.

1. **Medical Waste Collection Point**: This is where medical waste is initially collected in different color bins.

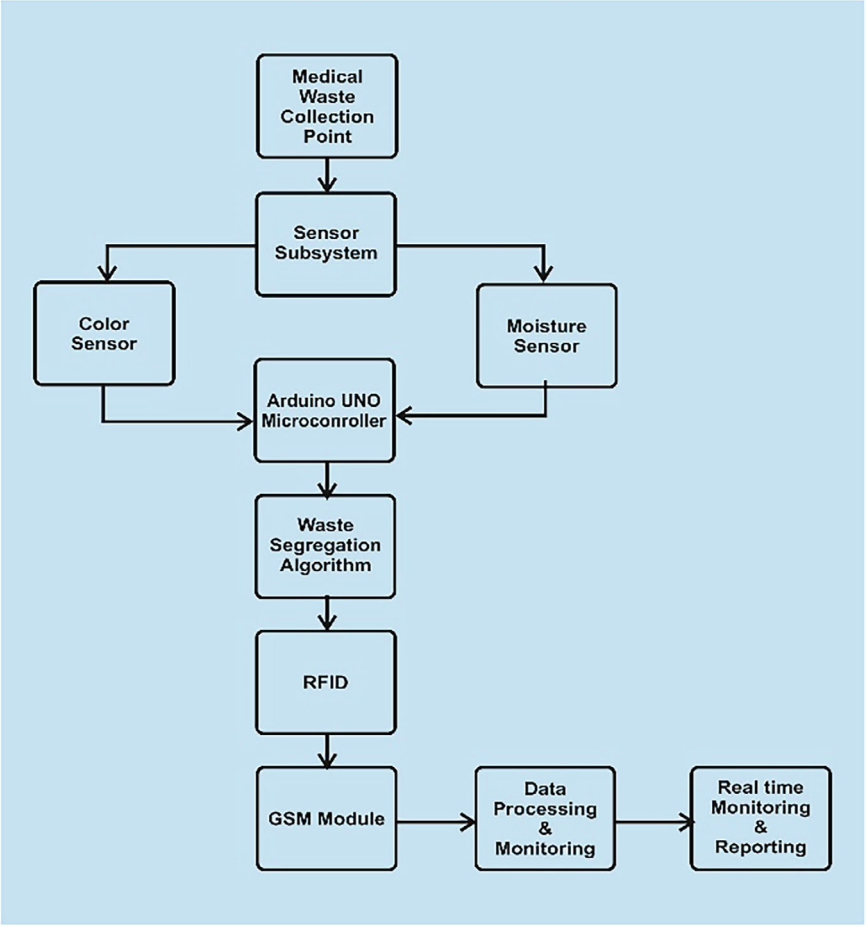


Fig. 1 Functional block diagram of the medical waste segregation and monitoring system

2. **Arduino UNO Microcontroller**: The Arduino UNO serves as the central control unit, orchestrating the functions of the entire system.
3. **Sensor Subsystem**: This subsystem includes color and moisture sensors for waste characterization.
4. **Waste Segregation Algorithm**: This algorithm processes sensor data to segregate waste based on color and moisture content.
5. **RFID System**: RFID technology enables real-time tracking of waste bins or containers.
6. **GSM Communication Module**: The GSM facilitates the communication of important data and alerts to stakeholders.
7. **Data Processing and Monitoring Software**: Software on a connected computer or server processes data received from the Arduino and manages monitoring functions.
8. **Real-time Monitoring and Reporting**: This is where real-time data are monitored and reports or alerts are generated for waste management personnel.

Each component or subsystem is critical for achieving the project's objectives of enhancing medical waste segregation and real-time disposal monitoring within healthcare facilities.

3 Methodology

This design employs color-coded containers—red, blue, and yellow—to categorize different types of waste in the hospital. Let 'C' represent the set of color-coded containers, where each container is denoted by ci for i in the range of 1 to 'n' (number of containers). Each container ci is associated with a specific color, represented by color(ci). Ultrasonic sensors measure the distance to waste for disposal. Let Di represent the distance measured by the ultrasonic sensor for container ci. Subsequently, an SMS alert is sent to designated staff members to empty the corresponding waste bin. Let Asms be the function triggering the SMS alert, as shown in Eq. (1).

$$\text{if } D_i > D_{threshold} \text{ trigger SMS} : A_{sms}(D_i) = \left\{ \frac{1 if\, D_i > D_{threshold}}{0 otherwise} \right\} \tag{1}$$

Additionally, a color sensor verifies whether the bag matches the container's assigned color code. Let Bi represent the binary variable indicating whether the bag matches the color code for container ci. Otherwise, an audible buzzer notifies the staff of the discrepancy. Let Abuzzer represent the function triggering the audible buzzer, as expressed in Eq. (2).

$$\begin{aligned} &\text{if Bi} = 0, \text{trigger audible buzzer: A_buzzer (B_i)} \\ &\quad = \{(1 \text{ if B_i} = 0)/(0 \text{ if B_i} = 1) \end{aligned} \tag{2}$$

A moisture sensor was used to monitor moisture levels within the containers. Let Mi represent the moisture level measured for container ci . Upon detection, a message prompts the responsible staff member to dispose of the container within 24 h.

if $M_i > M_threshold$, send a message to dispose of the container within 24 hours :

$$A_{moisture}(M_i) = \{\frac{1 if M_i > M_{threshold}}{0 otherwise} \tag{3}$$

Failure triggers an automated message to a higher-ranking officer responsible for biomedical waste management. Staff members must scan an attached RFID tag before sending a container for disposal, as shown in Eq. (4).

$$RFID_Scanned = \{(1 if RFIDtagisscanned)/(0 otherwise) \tag{4}$$

They can then track progress using a mobile app and monitor it via the Blynk app to ensure that it reaches its designated destination, as shown in Eq. (5).

$$Disposal_Action = \{(1\text{if} containerisdisposed)/(0 otherwise) \tag{5}$$

The architecture of the medical waste segregation and monitoring system is shown in Fig. 2. The pseudocode for the proposed system is shown in Figs. 3 and 4, respectively.

4 Results

This section delves into the findings. Figure 5 illustrates the comprehensive hardware setup of the medical waste segregation and monitoring system. An alert message is sent if the waste is not disposed of, as shown in Fig. 6. The color of the color sensors was used to determine the color of the bag, with the output displayed on a serial monitor, as depicted in Fig. 7.

The moisture sensor will examine moisture levels, and if moisture is detected, an SMS will be sent to the corresponding staff member following the scanning process. Container tracking commences via the Blynk app, and the status is continuously refreshed within the app, as shown in Fig. 8.

5 Discussion

For the 100 color-coded containers in the hospital waste management system, the following results were observed:

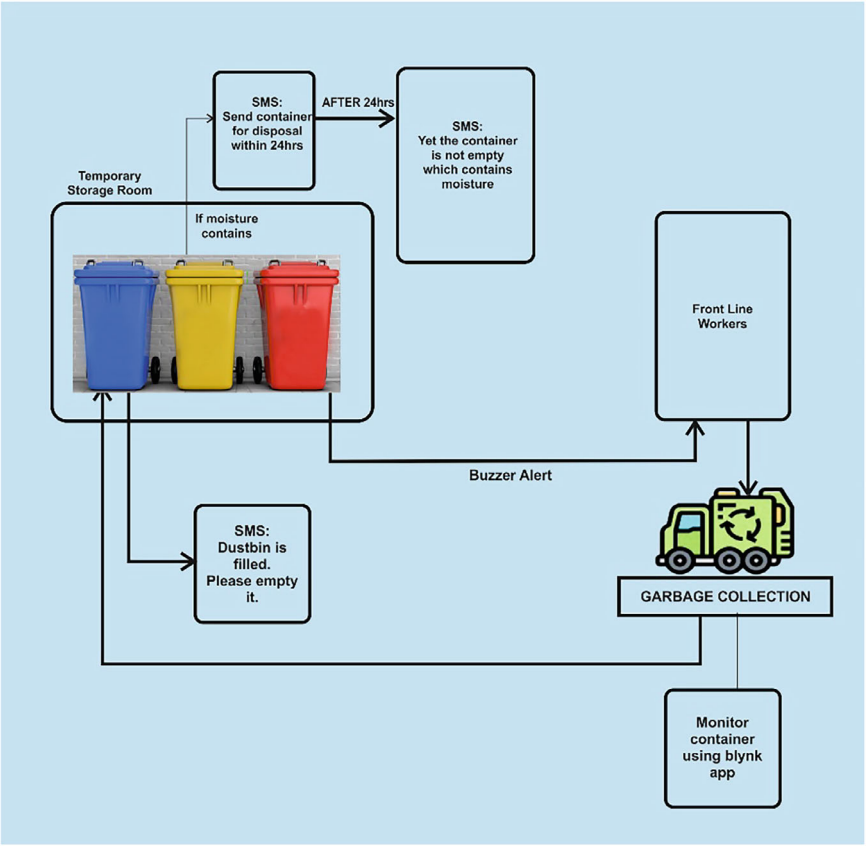


Fig. 2 Architecture of the medical waste segregation and monitoring system

Alert Triggers

1. Twenty SMS alerts ($A_{sms} = 1$) were triggered due to distances exceeding the threshold.
2. Fifteen audible buzzer notifications ($A_{buzzer} = 1$) were triggered due to mismatched color codes.

To calculate the percentage of correct alerts for color matching, let us assume that 10 out of the 15 alerts were due to actual mismatched color codes, as shown in Eqs. (6) and (7).

$$\text{Correct Alerts}(\%) = \frac{\text{Number of Correct Alerts}}{\text{Total Number of Alerts}} * 10 \tag{6}$$

$$\text{Correct Alerts}(\%) = \frac{10}{15} * 100 = 66.67\% \tag{7}$$

```
# Initialize color-coded containers
containers = []

# Initialize sensor readings
ultrasonic_distances = []
color_matches = []
moisture_levels = []

# Initialize RFID tag status
rfid_scanned = False

# Main loop
while True:
    # Read sensor data
    for container in containers:
        ultrasonic_distance = read_ultrasonic_sensor(container)
        color_match = read_color_sensor(container)
        moisture_level = read_moisture_sensor(container)
        ultrasonic_distances.append(ultrasonic_distance)
        color_matches.append(color_match)
        moisture_levels.append(moisture_level)
```

Fig. 3 Pseudocode to read sensor data of the proposed system

Moisture Levels

1. Twenty-five containers had moisture levels exceeding the threshold ($A_{moisture} = 1$).

To monitor the staff's response rate, let us assume that 20 out of the 25 containers were disposed of within 24 h after receiving the message prompting disposal due to high moisture levels. The response rate is calculated as shown in Eqs. (8) and (9).

$$\text{Response Rate}(\%) = \frac{\text{Number of Disposed Containers}}{\text{Number of Containers with High Moisture Levels}} * 100 \tag{8}$$

$$\text{Response Rate}(\%) = \frac{20}{25} * 100 = 80\% \tag{9}$$

RFID Tag and Tracking

1. Ninety percent of containers had scanned RFID tags before disposal.
2. Containers' progress was effectively tracked using the mobile app and Blynk app, ensuring timely disposal and proper management.

Overall System Performance

```
# Check for alerts
for i in range(len(containers)):
    if ultrasonic_distances[i] > 10:
        send_sms_alert(container)
    if not color_matches[i]:
        trigger_audible_buzzer(container)
    if moisture_levels[i] > 50:
        send_moisture_alert(container)

# Check RFID tag status
if not rfid_scanned:
    scan_rfid_tag()
    rfid_scanned = True

# Perform actions
for container in containers:
    if needs_disposal(container):
        dispose_container(container)
    track_container_progress(container)

# Delay before next iteration
time.sleep(1)
```

Fig. 4 Pseudocode to check alerts of the proposed system

1. Analyzing the frequency and response time of staff actions indicates that, on average, containers are disposed of within 24 h after being prompted by alerts. However, there might be occasional delays in response time.
2. No significant bottlenecks were identified in the system's workflow based on the results.

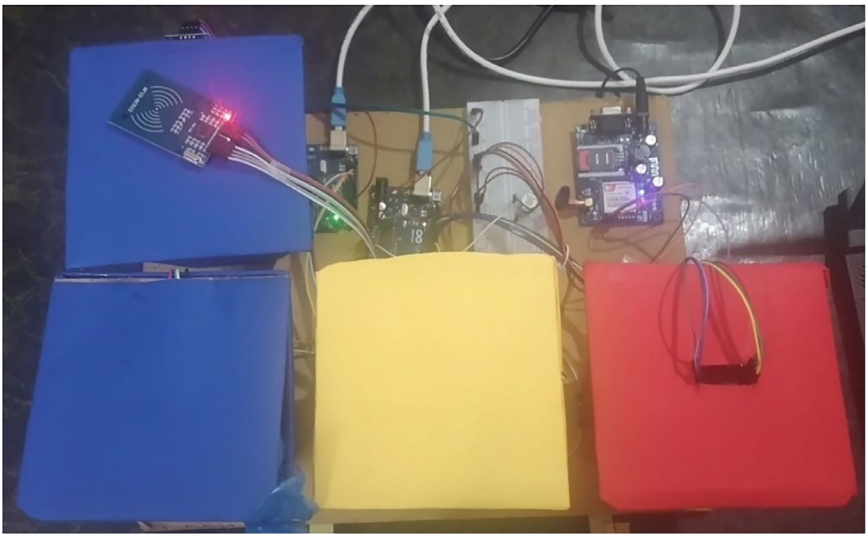


Fig. 5 Hardware setup for the medical waste segregation and monitoring system

Fig. 6 Alert message



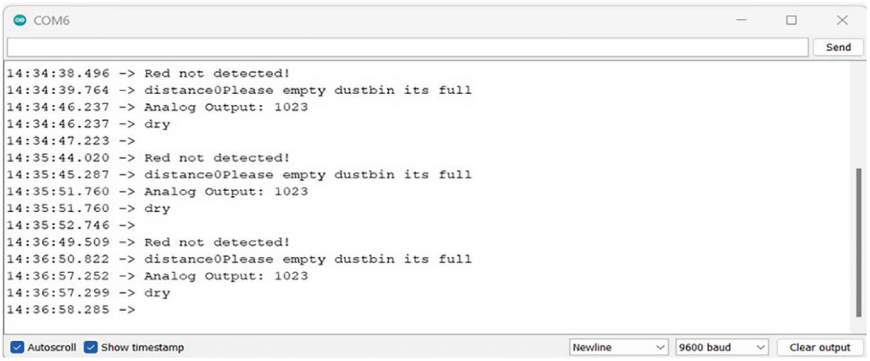


Fig. 7 Color sensor output

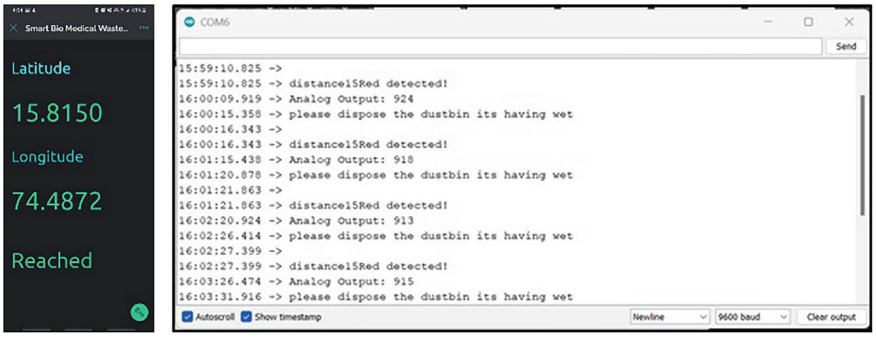


Fig. 8 Blynk App and RFID tag status

6 Conclusion

Implementing this system facilitates the convenient monitoring and tracking of infectious medical waste, effectively preventing waste contamination and the potential spread of diseases. This ensures appropriate waste disposal by promptly alerting personnel through SMS when containers are full and enabling tracking via GPS modules. Additional sensors, such as gas and water sensors, can be integrated to further enhance this system, and the project can benefit from upgraded module types. Novelty lies in integrating cloud and IoT as a software platform for real-time monitoring tailored specifically for medical waste, a less explored domain. The system's main ability to identify hazardous waste types and alert disposal services sets it apart from traditional systems.

References

1. Surme Y, Maraş G (2022) Recycling, responsible consumption, and nursing: a qualitative study of surgical nurses' recycling and medical waste management. J Nurs Manage 30(8):4514–4522. https://doi.org/10.1111/jonm.13891. Epub 2022 Nov 12. PMID: 36326215
2. Wang X, Liu L, Wang L, Cao W, Guo D (2024) An application of BWM for risk control in reverse logistics of medical waste. Front Public Health 12:1331679. https://doi.org/10.3389/fpubh.2024.1331679. PMID: 38344233; PMCID: PMC10853444
3. Tadesse ML (2024) Healthcare waste generation and quantification in public health centers in Addis Ababa, Ethiopia. PLoS One 19(2):e0295165. https://doi.org/10.1371/journal.pone.0295165. PMID: 38315710; PMCID: PMC10843618
4. Ranjbari M, Shams Esfandabadi Z, Shevchenko T, Chassagnon-Haned N, Peng W, Tabatabaei M, Aghbashlo M (2022) Mapping healthcare waste management research: Past evolution, current challenges, and future perspectives towards a circular economy transition. J Hazard Mater 422:126724. https://doi.org/10.1016/j.jhazmat.2021.126724. Epub 2021 Jul 27. PMID: 34399217
5. Hayleeyesus SF, Cherinete W (2016) Healthcare waste generation, and management in public healthcare facilities in Adama, Ethiopia. J Health Pollut 6(10):64–73. https://doi.org/10.5696/2156-9614-6-10.64. PMID: 30524786; PMCID: PMC6236544
6. McGain F, Naylor C (2014) Environmental sustainability in hospitals—a systematic review and research agenda. J Health Serv Res Policy 19(4):245–52. https://doi.org/10.1177/1355819614534836. Epub 2014 May 9. PMID: 24813186
7. Patil AD, Shekdar AV (2001) Health-care waste management in India. J Environ Manage 63(2):211–220. https://doi.org/10.1006/jema.2001.0453. PMID: 11721600
8. Mohamed NH, Khan S, Jagtap S (2023) Modernizing medical waste management: unleashing the power of the internet of things (IoT). Sustainability 15:9909. https://doi.org/10.3390/su15139909
9. Anh Khoa T, Phuc CH, Lam PD, Nhu LM, Trong NM, Phuong NT, Dung NV, Tan-Y N, Nguyen HN, Duc DN (2020) Waste management system using IoT-based machine learning in university. Wirel Commun Mob Comput 2020:13, Article ID 6138637. https://doi.org/10.1155/2020/6138637
10. Sun S, Hu J, Cao Y, Zhou W (2019) Discussion on the application of RFID technology in medical waste management. In: 2019 10th International conference on ınformation technology in medicine and education (ITME), Qingdao, China, pp 120–124. https://doi.org/10.1109/ITME.2019.00037
11. Kerley FR, Nissly BE (1992) Total quality management and statistical quality control: practical applications to waste stream management. Hosp Mater Manage Q 14(2):40–59 PMID: 10121996
12. Wahlstedt ER, Wahlstedt JC, Rosenberg JS, deVries CR (2024) Lifecycle of surgical devices: global, environmental, and regulatory considerations. World J Surg 48(5):1045–1055. https://doi.org/10.1002/wjs.12140. Epub ahead of print. PMID: 38530108
13. Lee SM, Lee D (2022) Effective medical waste management for sustainable green healthcare. Int J Environ Res Public Health 19(22):14820. https://doi.org/10.3390/ijerph192214820. PMID: 36429539; PMCID: PMC9690095
14. Ferreira MJC, Ventura CAA, Valadares GV, Mendes IAC, Silva TPD, Silva ÍR (2024) Healthcare waste management: connections with sustainable nursing care. Rev Esc Enferm USP 58:e20230229. https://doi.org/10.1590/1980-220X-REEUSP-2023-0229en. eCollection 2024. PMID: 39475389
15. Berhe TT, Feleke AK, Bekele GE, Gebrihiwot EM, Ali YH, Tedila GW (2025) Facilitators and barriers of infectious waste management practice in public and private health facilities of Addis Ababa, Ethiopia: a qualitative phenomenological study