

The Islamic University Journal of Applied Sciences (IUJAS)

Refereed periodical scientific journal

An Open Access Journal

The articles are published under the Creative Commons Attribution 4.0 International license (CC BY 4.0)
(http://creativecommons.org/licenses/by/4.0/)



Print ISSN: 1658-7936
Online ISSN: 1658-7944

Volue: VIII Issue: I Year: 2026


بسم الله الرحمن الرحيم

Islamic University Journal of Applied Sciences (IUJAS)

The Islamic University Journal of Applied Sciences (IUJAS)

Issued By

Islamic University of Madinah, Madinah, Saudi Arabia

July 2026

Paper version

Filed at the King Fahd National Library No. 8742/1439 on 17/09/1439
AHInternational serial number of periodicals (ISSN) 1658-7936

Online version

Filed at the King Fahd National Library No. 8742/1439 on 17/09/1439
AHInternational Serial Number of Periodicals (e-ISSN) 1658-7944

The Journal's Website

https://journals.iu.edu.sa/jesc

(The views expressed in the published papers reflect the views of the researchersonly, and do not necessarily reflect the opinion of the journal)

Publication Rules at the Journal (*)

❖ General rules:

- Report original scientific research (the main results and conclusions must not have beenpublished or submitted elsewhere).
- Fit with the topics of the journal.
- Report novel results, innovative work and show a new scientific contribution.
- Not to bear similarity of more than 25% of a previously published work of the sameauthor(s).
- Follow the rules, regulation and authentic research methodologies.
- Fulfill the required items and the format of the journal provided in appendix below relatedto the guide for author.
- Opinions expressed in published articles commit the authors themselves only and notnecessarily the opinion of the journal.

❖ For all articles:

- The exclusive right to publish and distribute and article, and to grant rights to others,including commercial purposes.
- For open access articles, IU will apply the relevant third-party user license where IUpublishes the article on its online platforms.
- The right to provide the article in all forms and media so the article can be used on thelatest technology even after publication.
- The authority to enforce the rights in the article, on behalf of an author, against thirdparties, for example in the case of plagiarism or copyright infringement.

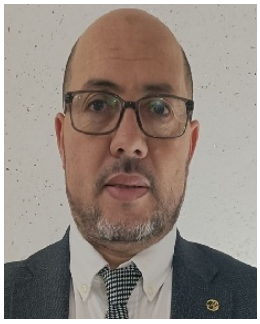
(*) These general rules are explained in details along with other rules for Author's guide in the journal's website: https://journals.iu.edu.sa/jesc

Editorial Board

	**Editor-in-Chief:** Mohamed Benghanem Professor in Physics, Faculty of Science, Islamic University of Madinah, KSA **Orcid:** https://orcid.org/0000-0002-2527-8741
	**Managing Editor:** Ahmad B. Alkhodre Professor, Computer science, The National Institute of Applied Sciences (INSA), Lyon, France **Orcid:** https://orcid.org/0000-0001-6168-3552

Editorial Board Members

	Aly Ramadan Seadawy Professor, Mathematics, Taibah University, Madinah, Saudi Arabia **Orcid:** https://orcid.org/0000-0002-7412-4773
	Reda Abdelmonsef A. Ibrahim Professor, Biology, Kafrelsheikh University, Egypt **Orcid:** https://orcid.org/0000-0001-6472-5666

	**Adel Mellit** Professor in Electrical Enginnering & Renewable Energy systems, ICTP, Trieste, Italy **Orcid:** https://orcid.org/0000-0001-5458-3502
	**Imran Shakir** Professor of Material Science & Engineering, California University, Los Angeles, USA **Orcid:** https://orcid.org/0000-0003-3385-4363
	**Mussa. A. Said** Professor of Chemistry, School of Chemical Sciences, University of East Anglia, Norwich, UK **Orcid:** https://orcid.org/0000-0003-3073-5449
	**Fazal Noor** Professor, Computer science and engineering, Concordia University Montreal. Canada **Orcid:** https://orcid.org/0000-0002-0096-3435
	**Hesham Rushdey El-Saeedi** Professor in Chemistry, Uppsala University, Sweden **Orcid:** https://orcid.org/0000-0002-2519-6690

	**Syed Modassir Hussain** Professor of Applied Mathematics, OP Jindal University, Raigah, Chattisggarh, INDIA **Orcid:** https://orcid.org/0000-0003-1847-4776
	**Basem Rashid Alamri** Associate Professor, Electrical Engineering, Taif University, Saudi Arabia Orcid: https://orcid.org/0000-0002-8667-0042
	**Saad Talal Alharbi** Professor in Computer Science, Human Computer Interaction, Faculty of Computers, Taibah University, Saudi Arabia Orcid: https://orcid.org/0000-0003-0913-8631
	**Muhammad Imran** Professor of Information Technology, Federation University, Australia Orcid: https://orcid.org/0000-0002-6946-2591
	**Yazed Alsaawy** Professor, Computer and information systems, Islamic University of Madinah. Saudi Arabia **Orcid:** https://orcid.org/0000-0001-5031-3388

	**Abdul Qadir Bhatti** Professor, Civil Engineering, Pacific Earthquake Engineering Research (PEER) Centre, UC Berkely USA **Orcid:** https://orcid.org/0000-0001-5433-7803
	**Shamsuddin Ahmed** Professor, Industrial Engineering, Graduate School of Business University of the South Pacific Suva, Fiji. **Orcid:** https://orcid.org/0000-0001-9169-3516
	**Ameer Azam** Professor of Applied Physics, Aligarth Muslim University, Aligarth, India Orcid: https://orcid.org/0000-0003-2431-1682

Consulting Members

	**Mohamed A. Shenashen** Professor of Material Science, National Institute for Materials Science (NIMS), Japan **Orcid:** https://orcid.org/0000-0003-1592-5877
	**Abdel-karrem Nasser Alomari** Professor of Mathematics, Faculty of Science, Yarmouk University. 211-63 Irbid, Jordan **Orcid:** https://orcid.org/0000-0001-5374-0916
	**José Campos E Matos** Assoc Professor, Department of Civil Engineering Minho University, Portugal **Orcid:** https://orcid.org/0000-0002-1536-2149
	**Hasan Al-Jabbouli** Associate Professor of Computer Science, New Jersey City University: Jersey City, NJ, USA **Orcid:** https://orcid.org/0000-0001-9880-8099
	**Marwan Hassani** Associate Professor of Computer Science, Eindhoven University of Technology: Eindhoven, Netherland **Orcid:** https://orcid.org/0000-0002-4027-4351
	**Tarek Mahmoud Omara** Associate Professor of Mathematics, Applied Statistics, Faculty of Commerce, Kafrelsheikh University, Egypt **Orcid:** https://orcid.org/0000-0003-3674-1743

Editorial Secretary

	**Ahmad Ziad Al-Zuhaily** Assistant Editor, Islamic University of Madinah. Saudi Arabia
	**Abdulrahman Saeed Odeh** Assistant Editor, Islamic University of Madinah. Saudi Arabia

Table of Contents

Article Number	Article Title	Page
1	Comparative Analysis of Deep Recurrent Neural Network Models for Ultra Short-Term Wind Power Forecasting	**1**
2	Magnetic Ferrite/TiO2 Nanocomposites for Enhanced Photocatalytic Water and Air Remediation	**22**
3	Mapping research trends in generative artificial intelligence: a bibliometric and network analysis	**46**
4	Hydrodynamic Behavior of Offset Jets Under Various Situations	**62**
5	Enhancing Cybersecurity in IoT-Based Maternal Health Monitoring Systems Using Machine Learning Algorithms	**78**
6	Structural evaluation and sustainable practices in the construction industry within the framework of saudi vision 2030	**110**

7	Simulation and Experimental Assessment of Various MPPT Methods For Photovoltaic Systems	**139**
8	Numerical Prediction of Nonlinear Response and Stability of Short and Long Hydrodynamic Journal Bearing	**153**
9	Theoretical investigation using DFT Calculations, Vibrational Assignments, Infrared Spectrum and Normal Coordinate Analysis of (2-Chloro-4-fluorophenyl) boronic Acid	**168**
10	Explainable Deep Learning Models for Citrus Leaf Disease Identification and Agricultural Decision Support	**189**
11	Numerical study of air flow and temperature distribution in a room equipped with a cooling ceiling	**218**
12	OCSA: An Optimized Cuckoo Search Algorithm with Guided Exploration and Refined Local Search	**230**
13	Silicon and Silicon Carbide Technologies for Nuclear Medicine: Performance, AI-Augmented Imaging, and Future Prospects	**249**
14	Machine Learning Approaches for Indoor Temperature Prediction in Northern Saudi Arabia	**269**

15	Enhancing Arabic Dialectical Speech Recognition with Large Language Models	**280**
16	Approximation of a Common Solution to Split Equilibrium and Fixed Point Problems for a Nonexpansive Semigroup a Modified Viscosity Method	**299**
17	Autonomous Drones for Precision Pest Control in Desert Agriculture: Technologies, Applications, and Future Prospects	**319**

Islamic University Journal of Applied Sciences (IUJAS)
https://journals.iu.edu.sa/jesc

Volume VIII, Issue I, July 2026, Pages 01-21

Comparative Analysis of Deep Recurrent Neural Network Models for Ultra Short-Term Wind Power Forecasting

Adel Mellit [1,2] *

[1] Faculty of Science and Technology, University of Jijel, 18000, Jijel, Algeria

[2] Abdus Salam International Center of Theoretical Physics (ICTP), Trieste, Italy

*Corresponding author: (Adel Mellit), *Email Address*: adel_mellit@univ-jijel.dz & a.mellit2@ictp.it

Abstract

The rapid growth of global wind power capacity has made accurate forecasting crucial for grid stability, economical allocation and efficient operation of wind farms. In this paper, we investigate the performance of various deep recurrent neural network (DRNN) models, including GRU, LSTM, BiGRU, and BiLSTM, deeper LTSM and LSTM-GRU for ultra short-term wind power forecasting. Furthermore, we incorporate a self-attention mechanism into LSTM (SA-LSTM) and GRU (SA-GRU) models to capture complex temporal dependencies and enhance forecasting accuracy. Extensive experiments on real-world wind power datasets demonstrate that the SA-LSTM and SA-GRU modes slightly outperform conventional architectures with an accuracy of 97%, while all evaluated models exhibit strong capabilities in capturing short-term fluctuations in wind power. These results emphasize the potential of advanced deep learning models to support real-time wind farm control, energy management and grid integration.

Keywords: Wind power; Wind speed; Forecasting; Deep Learning; Self-attention mechanism.

https://doi.org/10.63070/jesc.2026.014
Received 28 March 2026; Revised 24 April 2026; Accepted 28 April 2026.
Available online 02 May 2026.
Published by Islamic University of Madinah on behalf of *Islamic University Journal of Applied Sciences.*
This is an open access article published under the Creative Commons Attribution 4.0 International license (CC BY 4.0):https://creativecommons.org/licenses/by/4.0/

1. Introduction

Over the past decade, global wind power capacity has expanded dramatically, becoming one of the largest sources of renewable electricity generation. According to recent statistics, total installed wind power capacity worldwide has exceeded 1,100 gigawatts, with the vast majority coming from onshore installations and a growing share from offshore farms [1]. In 2024 alone, a record amount of new wind power capacity was added, reflecting the accelerated pace of deployment in key regions such as China, Europe, and the United States, and underscoring the pivotal role of wind power in reducing carbon emissions in the electricity sector and achieving global climate goals.

Ultra short-term wind forecasting, which typically covers time periods ranging from a few seconds to one hour (e.g., a 10-minute forecast), is critical for the real-time operation and stability of modern power systems, including microgrids [2]. Given the intermittent and rapidly changing nature of wind, accurate short-term forecasts enable system operators to maintain a balance between power generation and demand, ensuring grid frequency and voltage stability. Furthermore, these forecasts support effective participation in daily electricity markets by reducing imbalance costs and improving bidding strategies.

At the wind farm level, ultra short-term forecasting facilitates the implementation of advanced control strategies, including turbine tuning and power smoothing, which improves energy extraction and reduces mechanical stress [3]. Moreover, it plays a crucial role in improving energy storage management and anticipating sudden fluctuations in production, ultimately enabling greater reliance on renewable energy sources while reducing dependence on conventional backup power generation [4].

Over the past decade, various AI-based approaches have been developed to address the challenges of wind energy forecasting, with a primary focus on designing models that achieve the highest possible accuracy [5]. For example, in [6] a three-stage model to improve the accuracy of one-hour wind power forecasts was developed. The average absolute error was reduced from 0.333 to 0.278, with the SWD-ANFIS (Stationary Wavelet Decomposition Adaptive Neuro-fussy inference system) and (Artificial Neural Networks (ANN) models outperforming each other in the early stages. The ensemble predictions closely match actual data, though the reliance of the model on historical data and multiple modeling steps may limit real-time use. Another study [7] enhanced the National Center for Atmospheric Research system by combining numerical weather prediction and machine learning to improve short-term forecasts, quantify uncertainty, and predict extreme events such as icing. Quantile-

based data control and blended ramp forecasting improve accuracy, though the complexity of the system may limit real-time implementation.

In [8], the study showed that ANFIS provides the most accurate 1-hour-ahead wind power forecasts, with ANN and RBFN-OLS also performing well, while Radial Basis Function Network (RBFN)-Hybrid and RBFN-SGD underperform. Although all AI-based methods outperform persistence, the reliance on historical data and seasonal modeling may limit adaptability to sudden, unanticipated wind fluctuations. The proposed Self-Attention Temporal Convolutional Network (SATCN)-LSTM model in [9] effectively combines self-attention temporal convolution with Long short-term Memory (LSTM) to improve ultra-short-term wind power forecasting, outperforming LSTM, TCN, Convolutional Neural Network (CNN)-LSTM, and TCN-LSTM with up to 17.56% reduction in RMSE. While the authors in [10] demonstrated that the GRU model outperforms ARIMA in short-term wind power prediction, achieving high accuracy with root mean squared error (RMSE) of 0.047, mean absolute error (MAE) of 0.03, and a coefficient of 0.89. However, the effectiveness of the model depends on the quality and location-specific nature of the data, potentially limiting its generalization to other wind farms or regions.

In [11], a Bayesian bidirectional gated recurrent unit (BiGRU) model with encoder-decoder architecture and feature-temporal attention was proposed. But, the complexity of the model and dependance on advanced modeling and uncertainty quantification limit its real-time implementation and widespread adoption. To improve the forecasting accuracy another model based on dual-channel TCN-Informer architecture was proposed in [12]. The results proven that the model can effectively improves 15-minute and 1-hour wind power forecasts by decomposing sequences, classifying mode complexity, and applying specialized models to each group, achieving higher accuracy than traditional methods. Another work in [13] where the authors developed a proposed hybrid framework, by combining data analysis, model prediction and real-time error correction, which greatly improves the accuracy of very short-term wind energy prediction for both single-step and multi-step forecasts.

In [14], the study demonstrated that an AI-driven, preprocessing-enhanced framework using SCADA data significantly improves short-term wind power forecasting, with XGBoost achieving the highest accuracy (R^2=0.9890, RMSE=0.0385) by addressing nonlinearities and ensuring physically plausible predictions. The authors in [15] proposed a KSC-TFT framework, combining shape-based turbine clustering with Temporal Fusion Transformer (TFF) modeling, significantly improves ultra-short-term wind power forecasts by capturing turbine-level heterogeneity, reducing NMAE and NRMSE by over 50%, and enhancing probabilistic forecast accuracy [16]. It was shown in that FS-ISABO-VMD-IMSGNet framework significantly enhances ultra-short-term wind power forecasting, achieving an

accuracy of 28.2%. However, the multi-stage model design and complex optimization increase computational demands, which could limit its applicability in real-time operations.

In [17], the authors propose a combination forecasting method integrating Bidirectional LSTM (BiLSTM) with time attention and kernel density estimation to improve ultra short-term wind power predictions under extreme weather, achieving average improvements of 10.04% in PICP and 20.47% in PINAW. Another work in where the authors proposed a hybrid optimization of data, features, and model architecture, using DBSCAN for anomaly correction, correlation-based feature selection, and an improved informer with window attention and Fourier transform. The method achieves higher accuracy, better stability, and faster execution than state-of-the-art models, but its multi-stage process and reliance on complex preprocessing may increase computational demands, potentially limiting real-time deployment.

The main objective of this work is to investigate the performance of various deep learning (DL) models for ultra short-term wind power (WP) forecasting. A configuration is proposed that integrates a self-attention mechanism with LSTM and GRU networks (SA-LSTM and SA-GRU) to enhance forecasting accuracy. To assess the effectiveness of the proposed approach, a comprehensive comparative study is conducted among GRU, LSTM, BiGRU, BiLSTM, deeper LSTM, LSTM-GRU, SA-LSTM, and SA-GRU models, accompanied by a detailed analysis and discussion of their forecasting performance.

The remainder of this research is divided as follows: Section 2 presents the dataset used in developing the models, along with its analysis and preprocessing. Section 3 provides a detailed description of the proposed methodology, including the integration of attentional mechanisms with LSTM and GRU models. Section 4 presents the experimental results and evaluates the performance of the models studied, including both traditional DL-based iterative constructs and those enhanced with attentional mechanisms. Finally, the conclusions and future projections are presented in the last section.

2. Dataset analysis and processing

The dataset used in this study consists of wind power time series with a 10-minute forecasting horizon. The data were collected from the SCADA system of a wind turbine installed in southern Italy, spanning the period from 23/11/2005 to 28/12/2026, and comprising a total of 14,080 samples. As an example, Figure 1a illustrates the measured wind speed and the corresponding wind power generation.

The persistence of significant correlations across large lags indicates (See Fig.1b) non-stationary characteristics, which motivates the use of attention-based RNNs capable of capturing long-range dependencies.

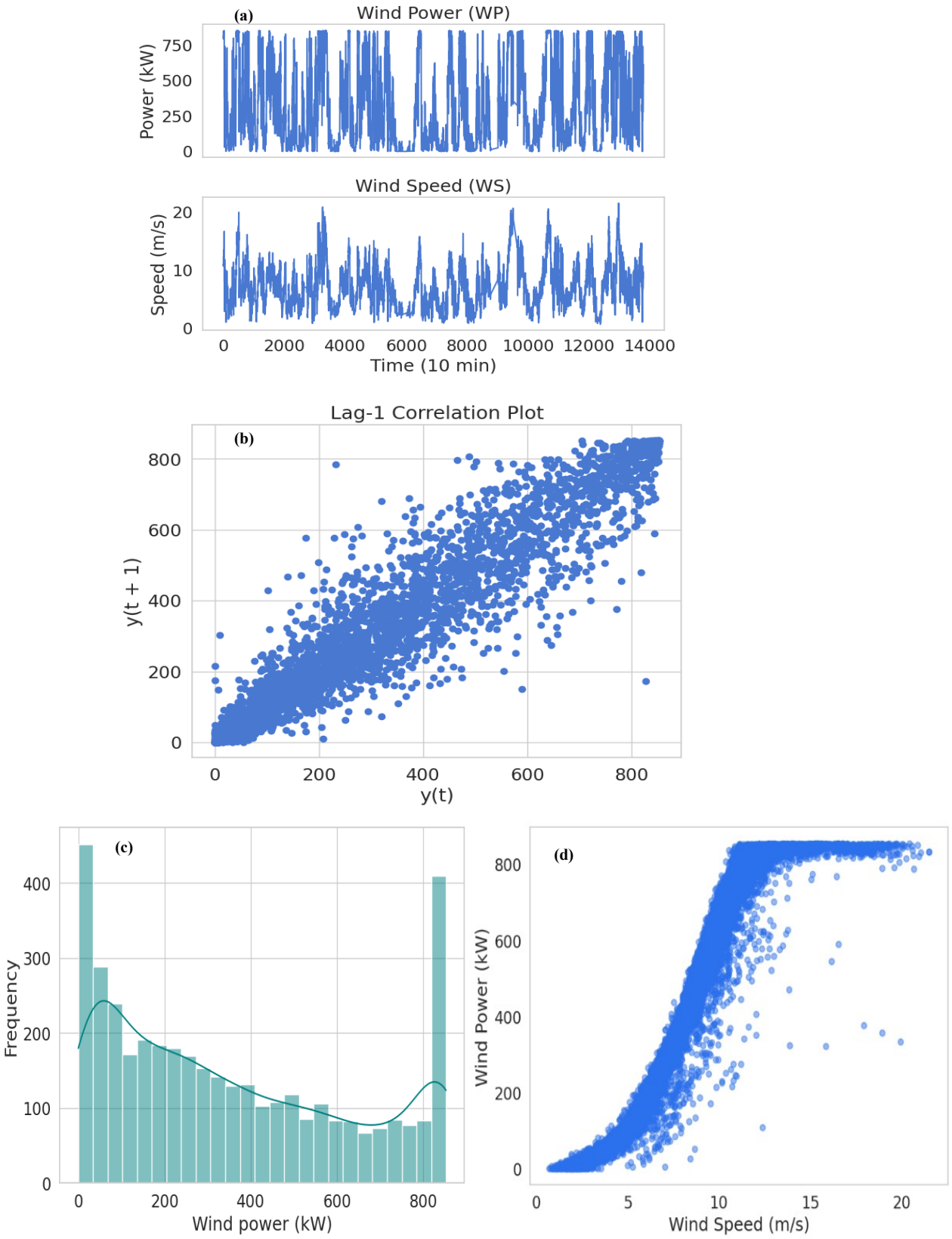


Fig. 1 a) Variation of wind power and wind speed over a period of 14000 samples, b) Lag-1 correlation plot ($y_{t+1}=f(y_t)$), c) Probability density function of WP variation and d) scatter curve of WP and WS.

The probability density function (See Fig.1c) for wind power exhibits a highly skewed multi-mode distribution, a characteristic feature of real-world wind power data. At low power levels (around 0-100 kW), a distinct peak is observed, indicating that the system often operates at low output. This behavior is generally consistent with low wind speed conditions or startup, where the turbine generates minimal power. As power increases (100-600 kW), the distribution gradually decreases, forming a long tail. This suggests that average power output occurs at a slower rate, reflecting the natural variability of wind speed over time. Interestingly, a secondary peak appears near the peak power range (around 800-850 kW). This indicates that the turbine often operates close to its rated power, most likely during periods of strong and stable wind conditions. This behavior is typically observed at high-efficiency wind farms. Figure 3d shows the scatter curve of WP and WS variation.

In this study, the historical WP data were preprocessed using Min-Max (See Eq.1) normalization to scale all values within the range [0, 1], ensuring that each feature contributes proportionally to the learning process and preventing dominance by larger values. The normalization is performed using the equation:

$$WP_{norm} = \frac{WP - WP_{min}}{WP_{max} - WP_{min}} \quad (1)$$

Where WP_{norm} is the normalized value, WP_{min} and WP_{max} are minimum and maximum values of WP respectively. After normalization, the dataset was split into training and testing subsets, with 80% used to train the forecasting model and 20% reserved for testing and validation. Figure 2 illustrates the entire normalized wind power dataset along with the training and testing partitions. After removing outliers and negative values, the dataset size became 11,946. The first 80% of the data points (9,556), shown in blue, were used for model training, capturing the underlying temporal patterns of wind power generation. The remaining 20% (2390), shown in red color, was reserved for testing and validation to evaluate the predictive performance of the model on unseen data. This visualization emphasizes the temporal continuity of the dataset and the importance of maintaining separate datasets for unbiased model evaluation.

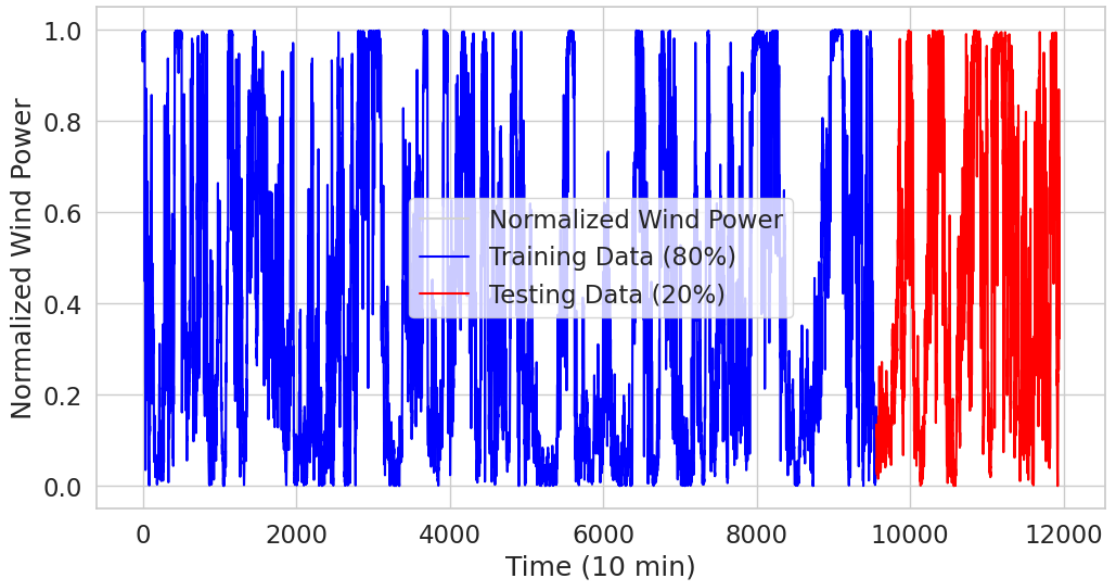


Fig. 2. Normalized WP dataset with 80% training and 20% testing split.

3. Methodology

3.1 Forecasting of wind power (WP)

Wind energy forecasting aims to predict future wind power values based on historical data. This process can be expressed mathematically as follows [19].

$$(\mathrm{WP}_{t+1},..,\mathrm{WP}_{t+k}) = f_{DNN}(\mathrm{WP}_{t-n},\mathrm{WP}_{t-n+1},\ldots,\mathrm{WP}_{t}) \quad (2)$$

f_{DNN} is a deep neural network (DNN), which will be used to estimate the future values of WP based on the fine-tuned parameters. In our case, k=1, so we will estimate one step ahead of WP (Next 10 min) and $t \in \{1, \ldots, n\}$, n is the size of the sequence.

3.2 Deep RNN-based models forecasting of wind power

3.2.1 Deep recurrent neural networks (DRNNs)

The deep recurrent neural network (DRNNs) models examined include LSTM, GRU, BiGRU, BiLSTM, and LSTM-GRU models, as well as a deeper LSTM architecture. The predictive performance of these models depends heavily on the optimization of key hyperparameters, such as the number of layers, hidden units, and the length of the input sequence (search window), which controls

the amount of historical data used. Training parameters, including the learning rate, optimizer, batch size, and number of training cycles, affect convergence and stability. Furthermore, regulation techniques (such as dropout and L1/L2 penalties) are essential to prevent overshoot, particularly in deep and bidirectional models. Overall, fine-tuning of the hyperparameters is critical to ensuring the accuracy and reliability of very short-term wind energy forecasts.

a) Long Short-Term Memory

LSTM [20,21] is adopted in this study due to its effectiveness in modeling sequential data. Unlike conventional RNNs, LSTM introduces additional structural components, specifically, a memory cell and three gating mechanisms: input, output, and forget gates. These gates regulate the flow of information, enabling the network to selectively retain or discard information over time. This architecture effectively mitigates the vanishing gradient problem commonly encountered in standard RNNs, thereby allowing the model to capture long-term dependencies within time-series data. By combining one or more LSTM layers with a dense output layer, the resulting DRNN becomes particularly well-suited for time-series forecasting tasks, as illustrated in Fig. 3.

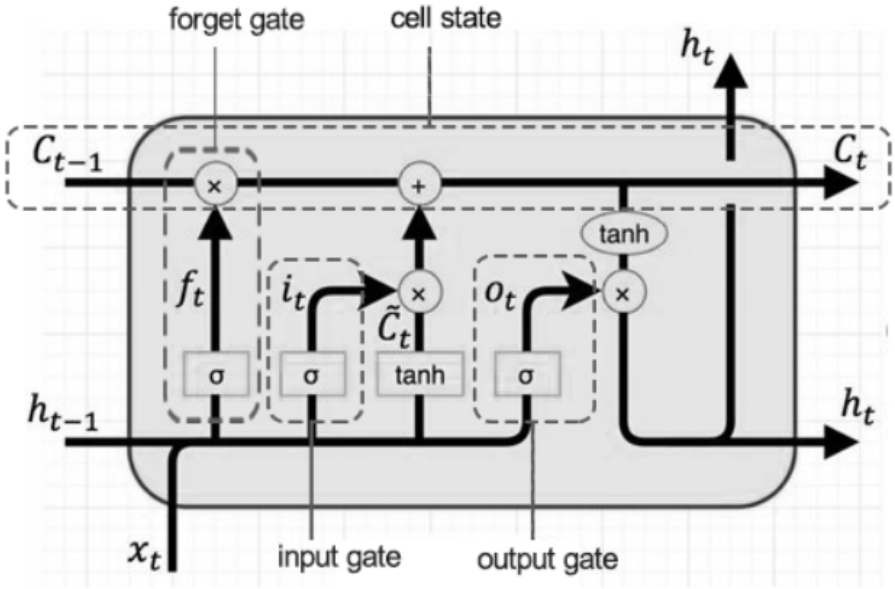


Fig. 3. Basic architecture of an LSTM deep neural network

Equations (3-8) mathematically describe the operation of the LSTM and how the information is processed from the input to the output passing through the respective gates:

$$i_t = \sigma\left(x_t U^i + h_{t-1} W^i\right) \quad (3)$$

$$f_t = \sigma(x_t U^f + h_{t-1} W^f) \quad (4)$$

$$o_t = \sigma(x_t U^0 + h_{t-1} W^0) \quad (5)$$

$$\hat{C}_t = tanh(x_t U^g + h_{t-1} W^g) \quad (6)$$

$$C_t = \sigma\left(f_t * C_{t-1} \quad + i_t * \hat{C}_t\right) \quad (7)$$

$$h_t = tanh(C_t) * o_t \quad (8)$$

Where i_t is the input gate activation, x_t is the current input, h_{t-1} is the hidden state from previous step, U_t is the weight matrix for the input xt Wt is the weight matrix for the previous hidden state h_{t-1}, σ is the sigmoid activation function, f_t is the forget gate, C_t is the cell candidate, o_t is the output gate, h_t is the hidden state.

b) Gate recurrent unit

The GRU [21,22] is a simplified variant of the LSTM architecture, designed to address the vanishing gradient problem encountered in training recurrent networks (see Fig. 4). Unlike LSTM, GRU employs two gating mechanisms—namely, the update gate and the reset gate, which regulate the flow of information within the network. These gates determine how much past information should be retained and how much new input should be incorporated into the current state. This streamlined structure reduces computational complexity while maintaining the ability to capture temporal dependencies. The mathematical formulation of the GRU is provided in Eqs. (9)–(12).

$$z_t = \sigma\left(x_t U^z + h_{t-1} W^z\right) \quad (9)$$

$$r_t = \sigma(x_t U^r + h_{t-1} W^r) \quad (10)$$

$$\hat{h}_t = tanh(x_t U^h + (r_t * h_{t-1}) W^h) \quad (11)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t \hat{h}_t \quad (12)$$

Here *r* is a reset gate, and *z* is an update gate

.

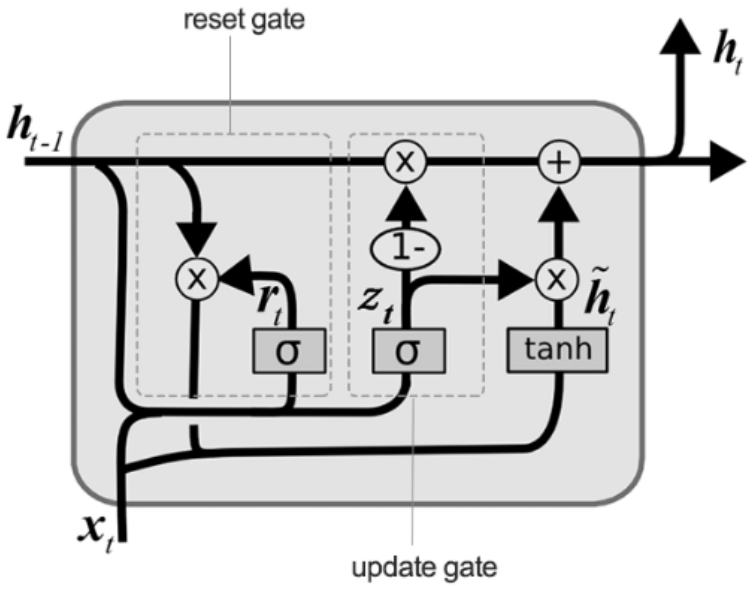


Fig. 4. Basic architecture of a GRU deep neural network.

c)BiLSTM and BiGRU

Bidirectional recurrent models, such as BiLSTM and BiGRU, extend conventional architectures by processing input sequences in both forward and backward directions (see Fig. 5) [23,24]. This dual processing enables the model to simultaneously capture dependencies from past (previous time steps) and future (subsequent time steps) contexts, leading to a more comprehensive representation of temporal dynamics. By combining the hidden states from both directionsm, typically through concatenation or summation, bidirectional models can better learn complex patterns and contextual relationships within time-series data. This capability often results in improved prediction accuracy, particularly in scenarios where the full sequence is available and future context provides valuable information.

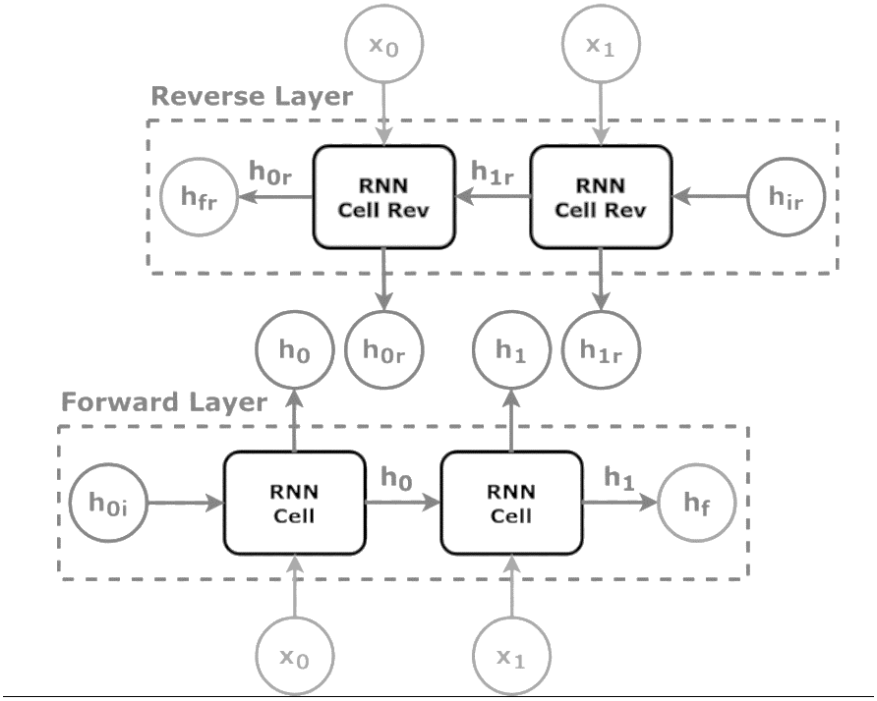


Fig. 5. Basic architecture Bidirectional RNN (BiGRU and BiLSTM).

They can increase the amount of information available to the network, improving the context available to the algorithm. The forward RNN processes in one direction:

$$f_\theta(x_0, h_0) = (y_0, h_1), f_\theta(x_1, h_1) = (y_1, h_2), \ldots \quad (13)$$

The backward RNN processes in the opposite direction:

$$f'_\theta(x_N, h'_0) = (y'_N, h'_{N-1}), f'_\theta(x_{N-1}, h'_{N-1}) = (y'_{N-1}, h'_{N-2}), \ldots \quad (14)$$

Where f_θ and f'_θ are the activation functions for forward and backword RNN.

d) stacked and deeper RNNs (GRU and LSTM)

Deep, stacked RNNs, including LSTM and GRU architectures, refer to models consisting of multiple sequentially ordered recurrent layers, where the output of each layer serves as the input for the next. This deep structure enables the network to learn hierarchical time representations, with lower layers capturing short-term patterns, while upper layers model more complex and abstract dependencies over longer timescales [26]. By increasing network depth, RNNs enhance the ability of the model to

represent nonlinear dynamics and improve prediction accuracy. However, deeper architectures introduce additional challenges, such as increased computational complexity and a higher risk of overshoot, necessitating fine-tuning of hyperparameters and the use of regulation techniques like dropout and early termination.

3.2.2 Self-attention mechanism in DRNNs

Self-attention is an advanced mechanism that enhances recurrent neural networks (RNNs), including LSTM, GRU, BiLSTM, and BiGRU, by enabling them to dynamically focus on the most relevant parts of an input sequence when generating predictions. In conventional LSTM and GRU architectures, the hidden state at each time step summarizes past information; however, their ability to capture long-term dependencies may be limited due to the progressive attenuation of earlier information.

To overcome this limitation, self-attention computes a context vector as a weighted sum of all hidden states in the sequence, allowing the model to directly access both recent and distant information. The attention mechanism can be formulated as follows [27]:

$$e_t = v^T tanh(W_h h_t + b) \tag{15}$$

$$\alpha_t = \frac{exp(e_t)}{\sum_{k=1}^{T} exp(e_k)} \tag{16}$$

$$c = \sum_{t=1}^{T} \alpha_t \, h_t \tag{17}$$

where e_t is the attention score, α_t denotes the normalized attention weight obtained via the Softmax function, and c is the resulting context vector.

This mechanism allows the model to assign higher importance to the most informative time steps while reducing the influence of less relevant ones, thereby improving the modeling of long-term dependencies [28]. Furthermore, when combined with bidirectional architectures such as BiLSTM and BiGRU, self-attention becomes even more effective by exploiting both past and future contextual information.

Consequently, self-attention significantly enhances the model's ability to capture complex temporal patterns and salient features, leading to improved performance in time-series forecasting, classification, and sequence generation tasks [28]. As an illustrative example, the following outlines the key steps for integrating self-attention into an LSTM (or GRU) model for time-series forecasting.

As an illustrative example, the following outlines the key steps for applying self-attention to an LSTM (or GRU) model in time series forecasting.

Step #1: Dataset

Given a time series:

$$X=\{x_1, x_2, \ldots x_K\} \tag{18}$$

Create sliding windows of length *n* as input and the next value as target:

$$X_{input}=[x_{t-n}, \ldots x_{t-1}],\ y_t=x_t \tag{19}$$

Step #2: LSTM layer or GRU

For each timestep *t*, LSTM computes hidden state h_t and cell state c_t:

$$f_t = \alpha\left(W_f.[h_{t-1}, x_t] + b_f\right) \tag{20}$$

$$i_t = \alpha(W_i.[h_{t-1}, x_t] + b_i) \tag{21}$$

$$o_t = \alpha(W_o.[h_{t-1}, x_t] + b_o) \tag{22}$$

$$C_t = f_t C_{t-1} + i_t.th(W_c.[h_{t-1}, x_t] + b_c) \tag{23}$$

$$h_t = O_t.\beta(C_t) \tag{24}$$

Where σ is the sigmoid function, x_t is the input, f_t is the forget gate, i_t is the input gate, o_t is the output gate, h_t is the output, c_t is the cell state, c_{t-1} is the cell update, β, α are the activation functions, W_o, W_i, W_f and W_c are the weights, and b_f, b_i, b_o and b_c are the bias.

Step #3: Compute attention weights

For a sequence of LSTM hidden states $H=[h_1,h_2,\ldots,h_n]$ and the current step hidden state h_t, compute attention score $e_{t,i}$ for each timestep *i*:

$$e_{t,i} = h_t W_a h_i \tag{25}$$

Normalize scores with SoftMax to get attention weights $\alpha_{t,i}$:

$$\alpha_{t,i} = \frac{exp(e_{t,i})}{\sum_{j=1}^{n} exp(e_{t,j})} \tag{26}$$

Step #4: Compute the context vector and combine it with the hidden state.

Weighted sum of all hidden states:

$$C_t = \sum_{j=1}^{n} \alpha_{t,i} h_i \tag{27}$$

Concatenate context vector with current hidden state:

$$\tilde{h}_t = tanh\left(W_c\left[\alpha_{t,i}; h_i\right]\right) \tag{28}$$

Step #5: Output prediction and update weights

Feed h_t to a dense layer for forecasting

$$\hat{y}_t = W_0\tilde{h}_t + b_0 \quad (29)$$

Loss function (e.g., MSE)

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2 \quad (30)$$

3.3.3 Fine-tuned parameters

All the iterative models developed in this study, including LSTM, GRU, BiLSTM, BiGRU, deeper and stacked RNNs and attention mechanism-based variants (SA-LSTM and SA-GRU), were finely tuned to improve prediction accuracy. Key hyperparameters, such as hidden units, number of layers, sequence length, dropout rates, learning rate, batch size, and attention mechanism size, were optimized [29,30]. Early stop was implemented to prevent overshoot, while attention mechanisms were tuned to focus on the most critical time steps. This tuning ensured an effective balance between temporal dependency incorporation, generalization, and computational efficiency.

4. Results and discussion

To evaluate the performance of the investigated RNN-based models (LSTM, GRU, BiLSTM, and BiGRU), including their attention-enhanced variants SA-LSTM and SA-GRU, the error metrics shown in Appendix are used: Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), the coefficient of determination (R^2) (See the Appendix section).

Figure 6 a illustrates the evolution of the loss function (MSE), showing that a low final value of 0.003 was achieved. As an illustrative example, Figure 6 b presents a comparison between the measured and forecasted WP obtained using the developed SA-GRU model. A strong agreement between the measured and forecasted values can be observed. Furthermore, Figure 6c depicts the scatter plots, revealing a high correlation between the measured and predicted data, with an R^2 of 97.17 %.

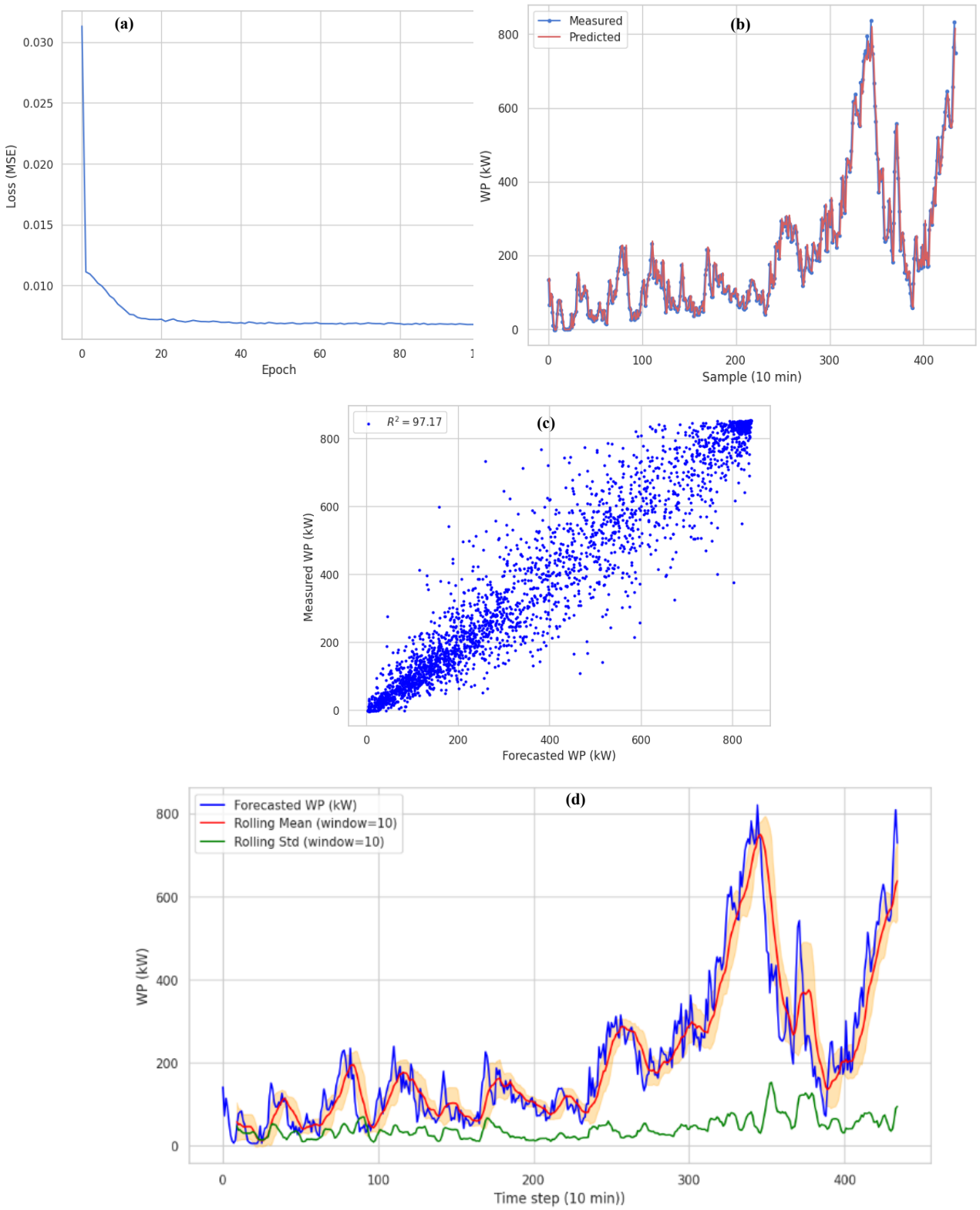


Fig. 6 a) Loss function, b) Measured vs. forecasted PW for a period of 450 samples, c) Scatter plot between measured and forecasted PW, and d) Forecasted PW with rolling statistics.

Figure 6 d illustrates the predicted WP over 450 time steps along with its rolling statistics. The blue curve represents the raw forecasted WP, showing high variability with sharp peaks and valleys characteristic of wind intermittency. The red curve depicts the rolling mean (window=10), which smooths short-term fluctuations and highlights the underlying trend in wind power generation. The green curve corresponds to the rolling standard deviation over the same window, indicating local volatility; higher values coincide with periods of rapid changes in WP, while lower values reflect more stable behavior. Together, the rolling mean and standard deviation provide a clear visual representation of both the trend and variability in the predicted WP, enabling the identification of stable periods versus highly volatile intervals. This approach is useful for assessing forecast reliability and for guiding further model evaluation or smoothing strategies..

Table 1 summarizes the performance of the developed models for short-term wind power forecasting using RMSE, MAPE, and R^2 as evaluation metrics. Basic models, including GRU, LSTM, and Bi-GRU, achieved similar RMSE values (0.091-0.093 kW), although BiGRU exhibited a high MAPE (56.74%), indicating larger relative errors at lower power levels. Deeper and hybrid architectures improved accuracy; for instance, the Deeper LSTM reduced RMSE to 0.080 kW and MAPE to 22.54%.

Table 1. Comparison of the developed models for the PW forecasting.

Model	**RMSE (kW)**	**MAPE (%)**	**R^2 (%)**
GRU	0.093	28.77	96.14
LSTM	0.092	25.03	96.17
BiGRU	0.091	56.74	96.38
BiLSTM	0.093	34.78	96.25
LSTM-GRU	0.092	30.28	96.10
Deeper LSTM	0.080	22.54	96.51
SA-LSTM	0.074	15.90	97.20
SA-GRU	0.075	15.87	97.13

Models incorporating attention mechanisms, such as SA-LSTM and SA-GRU, achieved the best performance, with RMSE between 0.074 and 0.075 kW, MAPE around 15.9%, and R^2 approximately 0.97, demonstrating their effectiveness in capturing the most relevant temporal patterns. Figure 7a compares RMSE, MAPE, and R^2 across all models, while Figure 7b presents a composite ranking highlighting the superior performance of attention-based frameworks.

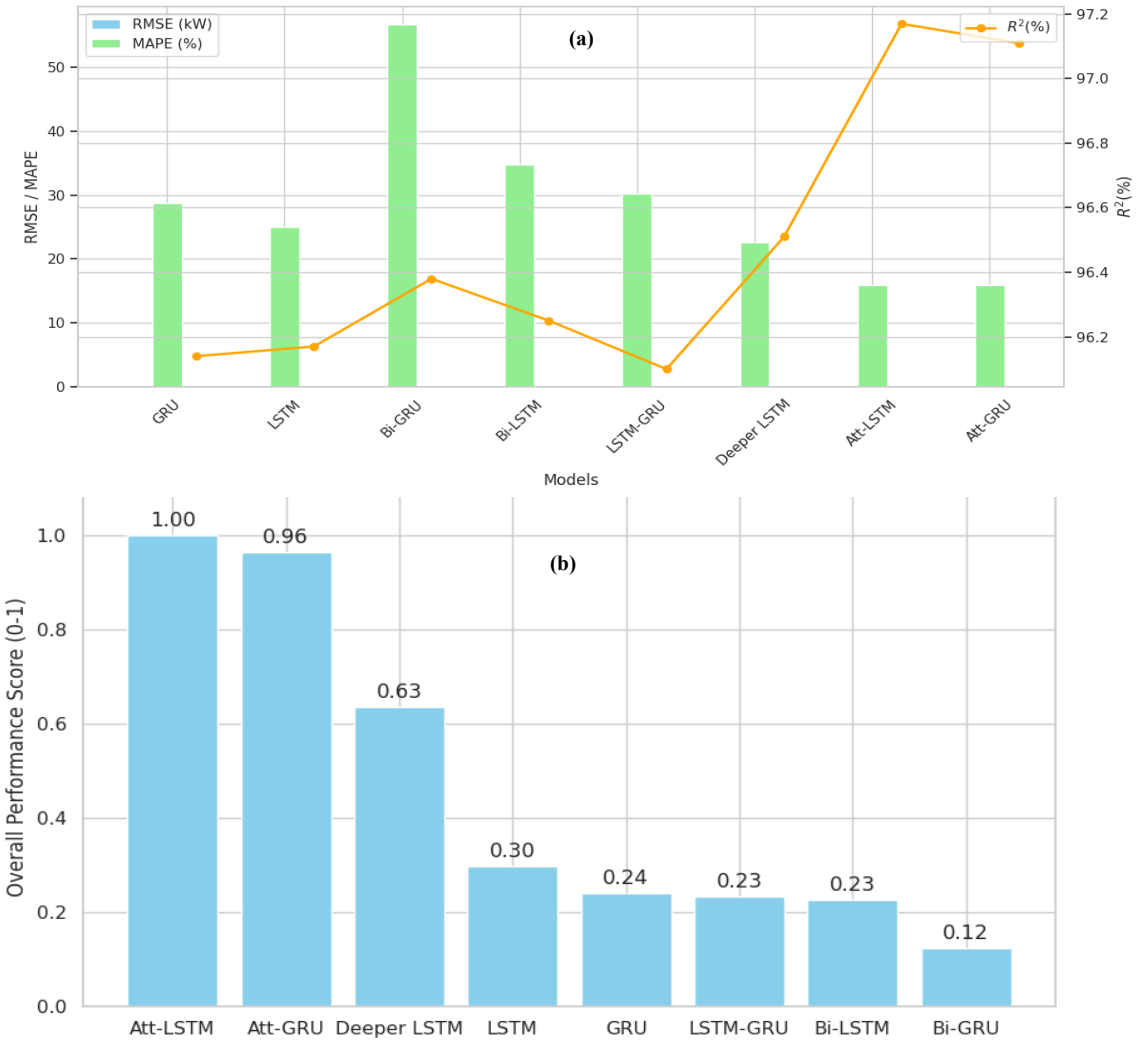


Fig. 7. a) Comparison of WP forecasting models and b) Ranking of WP forecasting models based on composite Score

5. Conclusion and perspectives

In this work, we assessed several DRNN models for ultra short-term wind power forecasting. The findings highlight the capability of these models to provide highly accurate wind power predictions. The forecasted wind power can be efficiently used for microgrid control and energy management. Among the evaluated models, the developed GRU and LSTM architectures, based on an attention mechanism (SA-GRU and SA-LSTM), showed clear advantages over standard deep RNN

architectures. They were able to better detect time dependencies, focus on the most relevant historical patterns, and reduce both RMSE and the MAPE, thus improving overall predictive performance.

Future research directions include scaling these models to multi-site and geographically distributed wind farms, incorporating probabilistic forecasting to quantify uncertainty, and exploring lightweight DL architectures suitable for immediate deployment in embedded systems. Furthermore, integrating weather forecasts and other environmental variables may further enhance the accuracy and robustness of the forecasts.

Appendix

Error metrics

RMSE measures the square root of the average squared difference between the predicted and measured values, providing an indication of the overall prediction accuracy while penalizing large errors more heavily

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{M}\left(WP_i - \widehat{WP}_i\right)^2}$$

where WP_i and $\widehat{WP}_i$ represent the measured and predicted values, respectively, and N is the number of samples.

MAPE quantifies the average relative error between predicted and actual values in percentage form, making it particularly useful for assessing forecasting accuracy in a normalized manner. It is expressed as:

$$MAPE = \sqrt{\frac{100}{M}\sum_{i=1}^{M}\frac{(WP_i - \widehat{WP}_i)}{WP_i}}$$

R^2 evaluates the strength of the linear relationship between predicted and measured values, indicating how well the model captures the temporal variability of the data. It is given by:

$$R^2 = \frac{\sum_{i=1}^{M}(WP_i - \overline{WP}_i)\left(\widehat{WP}_i - \overline{\widehat{WP}_i}\right)}{\sum_{i=1}^{M}(WP_i - \overline{WP}_i)^2 \sum_{i=1}^{N}\left(\widehat{WP}_i - \overline{\widehat{WP}_i}\right)^2}$$

where $\overline{WPi}$ and $\overline{\widehat{WPi}}$ denote the mean values of the measured and predicted data, respectively. Together, these metrics provide a comprehensive assessment of model accuracy, relative error magnitude, and correlation with observed data.

Acknowledgements

The author would like to acknowledge support from the International Center of Theoretical Physics (ICTP), Trieste (Italy) through the Associates Program (2023–2028).

Funding

The authors declare that this research received no external funding.

Data Availability Statement

Not applicable.

Conflicts of Interest

There is no conflict of interest, the writers assert.

References

[1] Wind power capacity in the Net Zero Scenario, 2015–2030, Int. Energy Agency (IEA). Available: https://www.iea.org/data-and-statistics/charts/wind-power-capacity-in-the-net-zero-scenario-2015-2030

[2] F. Rodríguez, A.M. Florez-Tapia, L. Fontán, A. Galarza, Very short-term wind power density forecasting through artificial neural networks for microgrid control, Renew. Energy 145 (2020) 1517–1527.

[3] G. Yu, L. Shen, Q. Dong, G. Cui, S. Wang, D. Xin, X. Chen, W. Lu, Ultra-short-term wind power forecasting techniques: Comparative analysis and future trends, Front. Energy Res. 11 (2024) 1345004.

[4] W.C. Tsai, C.M. Hong, C.S. Tu, W.M. Lin, C.H. Chen, A review of modern wind power generation forecasting technologies, Sustainability 15 (2023) 10757.

[5] M.H. Lipu, M.S. Miah, M.A. Hannan, A. Hussain, M.R. Sarker, A. Ayob, M.H.M. Saad, M.S. Mahmud, Artificial intelligence based hybrid forecasting approaches for wind power generation: Progress, challenges and prospects, IEEE Access 9 (2021) 102460–102489.

[6] H.H. Çevik, M. Çunkaş, K. Polat, A new multistage short-term wind power forecast model using decomposition and artificial intelligence methods, Physica A 534 (2019) 122177.

[7] B. Kosovic, S.E. Haupt, D. Adriaansen, S. Alessandrini, G. Wiener, L. Delle Monache, Y. Liu, S. Linden, T. Jensen, W. Cheng, M. Politovich, A comprehensive wind power forecasting system integrating artificial intelligence and numerical weather prediction, Energies 13 (2020) 1372.

[8] M. Godinho, R. Castro, Comparative performance of AI methods for wind power forecast in Portugal, Wind Energy 24 (2021) 39–53.

[9] L. Xiang, J. Liu, X. Yang, A. Hu, H. Su, Ultra-short term wind power prediction applying a novel model named SATCN-LSTM, Energy Convers. Manag. 252 (2022) 115036.

[10] S. Qureshi, F. Shaikh, L. Kumar, F. Ali, M. Awais, A.E. Gürel, Short-term forecasting of wind power generation using artificial intelligence, Environ. Chall. 11 (2023) 100722.

[11] L. Liu, J. Liu, Y. Ye, H. Liu, K. Chen, D. Li, X. Dong, M. Sun, Ultra-short-term wind power forecasting based on deep Bayesian model with uncertainty, Renew. Energy 205 (2023) 598–607.

[12] Q. Li, X. Ren, F. Zhang, L. Gao, B. Hao, A novel ultra-short-term wind power forecasting method based on TCN and Informer models, Comput. Electr. Eng. 120 (2024) 109632.

[13] S. Yang, J. Zhou, X. Gu, Y. Mei, J. Duan, A comprehensive framework of the decomposition-based hybrid method for ultra-short-term wind power forecasting with on-site application, Energy 313 (2024) 133911.

[14] B. Kishore, K. Kabilan, L.S. Rahul, G.P. Priyadarshan, E.P. Vishnutheerth, R. Satheesh, M.L. Kolhe, Advancing short-term wind power forecasting by AI-driven models for improved accuracy, Electr. Eng. 107 (2025) 14795–14810.

[15] C. Zhang, X. Xiao, Y. Wang, M.Z. Hou, S. Huang, W. Hu, M. Hu, R. Huang, Fine-grained ultra-short-term wind power forecasting based on Temporal Fusion Transformers integrated with turbine power time series clustering, Energy (2025) 137995.

[16] X. Shu, L. Gao, F. Zhang, X. Ren, L. Qin, Y. Wang, X. Wu, Ultra-short-term wind power forecasting method based on optimized decomposition and deep learning, Energy Convers. Manag. X (2025) 101315.

[17] W. Li, T. Yang, J. Yang, L. Peng, Predicting ultra-short-term wind power combinations under extreme weather conditions, IEEE Access (2025).

[18] P. Xu, S. Zhang, P. Luo, X. Wang, Y. Yan, D. Qi, Data-feature hybrid optimization for ultra-short-term wind power forecasting, Int. J. Electr. Power Energy Syst. 176 (2026) 111742.

[19] Mellit, A.M. Pavan, V. Lughi, Deep learning neural networks for short-term photovoltaic power forecasting, Renew. Energy 172 (2021) 276–288.

[20] Hochreiter, S. and Schmidhuber, J.,. Long short-term memory. Neural computation, 9(8), (1997) pp.1735-1780.

[21] Han, Z., Zhao, J., Leung, H., Ma, K.F. and Wang, W., 2019. A review of deep learning models for time series prediction. IEEE Sensors Journal, 21(6), pp.7833-7848.

[22] Boubaker, S., Benghanem, M., Mellit, A., Lefza, A., Kahouli, O. and Kolsi, L., Deep neural networks for predicting solar radiation at Hail Region, Saudi Arabia. IEEE Access, 2021, 36719-36729.

[23] Gamboa, J.C.B. Deep learning for time-series analysis. arXiv, 2017, preprint arXiv:1701.01887.

[24] Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. On the properties of neural machine translation: Encoder–decoder approaches. 2014, ArXiv:1409.1259.

[25] Xu, L., Xu, W., Cui, Q., Li, M., Luo, B. and Tang, Y., 2023. Deep heuristic evolutionary regression model based on the fusion of BiGRU and BiLSTM. Cognitive Computation, 15(5), pp.1672-1686.

[26] Sagheer, A. and Kotb, M.,. Unsupervised pre-training of a deep LSTM-based stacked autoencoder for multivariate time series forecasting problems. Scientific reports, 9(1), (2019), p.19038.

[27] W. Li, F. Qi, M. Tang, Z. Yu, Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification, Neurocomputing 387 (2020) 63–77.

[28] Kilinc, H.C., Apak, S., Ozkan, F., Ergin, M.E. and Yurtsever, A., 2024. Multimodal fusion of optimized GRU–LSTM with self-attention layer for hydrological time series forecasting. Water Resources Management, 38(15), pp.6045-6062.

[29] Z. Fu, H. Yang, A.M.C. So, W. Lam, L. Bing, N. Collier, On the effectiveness of parameter-efficient fine-tuning, Proc. AAAI Conf. Artif. Intell. 37 (2023) 12799–12807.

[30] P. Dhal, C. Azad, A fine-tuning deep learning with multi-objective-based feature selection approach for the classification of text, Neural Comput. Appl. 36 (2024) 3525–3553.

Islamic University Journal of Applied Sciences (IUJAS)
https://journals.iu.edu.sa/jesc

Volume VIII, Issue I, July 2026, Pages 22-45

Magnetic Ferrite/TiO_2 Nanocomposites for Enhanced Photocatalytic Water and Air Remediation

Hossameldin G. Mohamedbakr[1],*, K F. Hassan[1], Medhat M. El-Moselhy[1], Mohamed S. Thabet[1] Zeinhom H. Mohamed[1] and Usama. A. Soliman[1]

[1] Chemistry Division, Department of Physical Sciences, College of Science, Jazan University, , P.O. Box. 114, Jazan, 45142, Kingdom of Saudi Arabia

*Corresponding author: (Hossameldin G. Mohamedbakr), hmohamedbakr@jazanu.edu.sa

Abstract

Melding ferrite nanoparticles into the TiO_2 semiconductor lattice endows ferrite/TiO_2 nanocomposites, with a two-fold capability: they can be magnetically retrieved and upon light exposure they break down contaminants. This dual action directly tackles pressing woes stubborn organic pollutants lingering in water and greenhouse gases hovering in the atmosphere by sharpening charge separation and stretching light absorption deeper into the visible spectrum. Moreover, the spinel ferrite's tunable cation arrangement reshapes both its electronic band structure and overall behavior thereby dictating efficiency and the material's recyclability. A thin silica layer improves charge transfer and stability, while anatase–rutile–brookite TiO_2 mixtures enhance charge separation. These engineered photocatalysts degrade dyes and pharmaceutical pollutants in water and also drive hydrogen production through photocatalytic water splitting. Thanks, to their nature retrieving the catalyst is straightforward which curtails waste and trims operational expenditures. Taken together the nanocomposites chart a path toward environmental remediation and renewable power generation marrying photocatalytic prowess, with magnetic pull for straightforward recovery.

Keywords: Magnetic ferrite/TiO_2 nanocomposites, photocatalysis, water remediation, air purification, charge separation and visible light absorption.

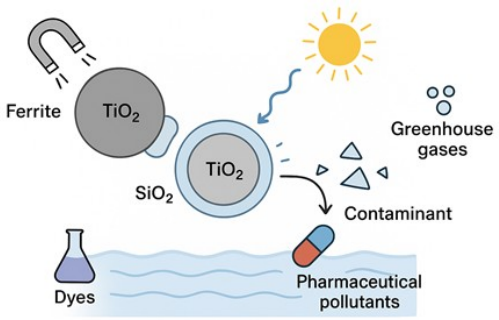
https://doi.org/10.63070/jesc.2026.015
Received 16 November 2025; Revised 23 January 2026; Accepted 14 February 2026.
Available online 04 March 2026.
Published by Islamic University of Madinah on behalf of *Islamic University Journal of Applied Sciences*.
This is an open access article published under the Creative Commons Attribution 4.0 International license (CC BY 4.0):https://creativecommons.org/licenses/by/4.0/

1. Introduction

Ferrite/TiO_2 nanocomposites belong to a material family that fuses magnetic ferrite nanoparticles with the semiconductor titanium dioxide endowing the composite with both magnetic and photocatalytic traits. The motivation for this marriage is straightforward: ferrites deliver a high saturation magnetization, which makes separation after a reaction virtually effortless while TiO_2 contributes potent photocatalytic activity under UV or solar light enabling the degradation of pollutants, in aqueous environments[1, 2]. This synergy reshapes the composite into a platform that can both catalyst chemical reactions and be easily reclaimed, confronting efficiency and sustainability challenges in environmental remediation systems. The routes to prepare composites fall into two categories “top-down" and "bottom-up" approaches. In the approach bulk material is broken down into nanoscale particles using physical techniques such, as ball-milling or laser ablation which typically yields a polydisperse particle size distribution[3]. Conversely "bottom-up" synthesis relies on controlled chemical or solvothermal routes that enable precise tweaking of cation distribution and particle shape. This method is often preferred when the aim is to maximize surface area and improve interface quality because catalytic activity is highly sensitive, to how active sitesre dispersed. Nonetheless both approaches have their constraints; maintaining uniform particle size and a well-ordered cation arrangement is essential to preserve the desired magnetic properties while extracting the best photocatalytic performance. Viewing it from a material-science angle spinel ferrites are especially appealing because their crystal lattice readily welcomes dopants of either divalent or trivalent metal ions. Such doping simultaneously tweaks the electronic band structure in ways that boost activity and alters the magnetic anisotropy that facilitates easy separation[4]. Experimental work on mixed ferrite systems e.g. Zn–Ni ferrite embedded in freudenbergite has revealed an enhancement in breaking down dyes, like methyl orange and methylene blue when the material is exposed to light. The co-habitation of metal cations within a spinel matrix tends to broaden the catalytic response window and at the same time reinforce magnetic-recovery capabilities. In the arena TiO_2 enjoys a solid reputation for a relatively high quantum yield under UV irradiation a benefit that springs from the electron-hole pairs it creates and that fuels redox reactions, on its surface. Still the naked form of TiO_2 is plagued by charge recombination. Embedding ferrites gives rise to heterojunctions (for example n-SnO_2/p-ferrite, architectures) where interfacial charge transfer lingers delaying recombination and boosting the photodegradation rate of organic pollutants[5]. Depending on how the ferrite’s band-gap aligns these heterojunctions can also pull the absorption edge into the spectrum making solar-driven water-treatment on a large-scale more practicable. The environmental target compounds that ferrite/TiO_2 photocatalysts set their sights on span dyes such, as the R250 hue, a gamut of phenolic constituents’ pharmaceutical agents exemplified by carbamazepine and the tangled stew of pollutants

found in complex industrial effluents[3, 6]. In runs degradation efficiencies tend to climb as time ticks by eventually levelling off; at a dye concentration of 100 ppm notable removal percentages have been reported when illumination is finely tuned to optimal conditions [3]. Notably when the system is kept in the dark the catalytic reactions scarcely move; the nanocomposites we crafted themselves show no activity under those conditions when tested against phenol-laden water[6]. highlighting that photoactivation is the driver. Magnetically recoverable nanocomposites also alleviate a flaw of many powder catalysts their clumsy removal from the treated water, which usually forces filtration steps that are prone, to choking. High-coercivity ferrites pose a risk of complicating demagnetization when they're recycled; for that reason, researchers have proposed low-coercivity alternatives such, as $SrFe_{12}O_{19}$ combined with TiO_2/SiO_2/graphene oxide where reversible magnetization enables reuse cycles without a substantial loss of catalytic capacity [1]. Sustainability arguments favour these designs because they cut down the secondary waste streams that typically arise from disposable single-use catalysts. Expanding the discussion past wastewater treatment researchers are now probing ferrites for biomedical uses. In hyperthermia nanoparticles of Fe_3O_4, $CoFe_2O_4$ and $MnFe_2O_4$ are harnessed to produce localized heating under alternating magnetic fields a consequence of their magnetic losses [4] . Though this medical avenue pursues objectives that differ starkly from those of photocatalysis it underscores the remarkable flexibility of ferrite materials. Looking at the problem, from a perspective the lessons learned about steering crystallinity and particle interfaces in one arena could very well inform the other. Surface chemistry matters as much; zeta-potential measurements reveal how electrostatic forces steer the deposition sequence first laying down a silica interlayer on a ferrite core and then adding a TiO_2 coating. These electrokinetic characteristics shape colloidal stability throughout the synthesis and dictate the uniformity of the shell's thickness[6]. Designed core-shell structures not only guard the magnetic phase against corrosion but also maximize the contact between pollutant molecules and the reactive TiO_2 facets, during illumination. Taken as a whole the design ethos behind ferrite/TiO_2 nanocomposites is a hybrid marrying the pull of magnetism which makes recovery a breeze with the push of semiconductor chemistry, which drives transformation. Fine-tuning these hybrids demands a trade-off among a suite of structural variables: how broadly the particle sizes are distributed, which cations populate the ferrite lattice, how uniformly the TiO_2 coating blankets the particles, the degree of coercivity tweaking required for easy recyclability and the quality of the heterojunction, at the interface that governs efficient charge-separation kinetics. Synthesizing insights from several investigations [1, 3, 5]with a deliberate engineering strategy these hybrid platforms appear as versatile candidates, for confronting the worldwide water-pollution challenge marrying robust contaminant degradation with energy-frugal recovery techniques.

Schematic 1: magnetic Ferrite/TiO_2 nanocomposites for enhanced photocatalytic water and air remediation



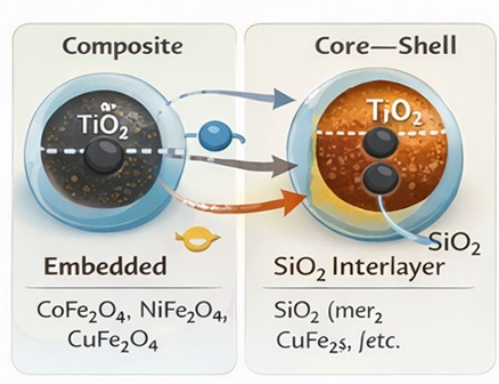
2 – Digging into the Background Uncovering Its Significance.

2.1 Environmental predicaments that photocatalysis can assist in resolving.

2.1.1 Water contamination sparked by a cocktail of compounds.

The worsening environmental pollution crisis driven by industrial waste streams, pharmaceutical residues and pesticide runoff calls for cutting-edge, sustainable cleanup methods. One promising avenue is photocatalysis, where light energy powers the oxidative breakdown of contaminants. In this arena titanium dioxide (TiO_2) stands out; it has been examined extensively because of its oxidizing ability, solid chemical stability, low cost and inherent non-toxicity[7, 8]. Nonetheless the mass deployment of TiO_2 collides with a pair of entrenched limitations: its expansive bandgap fences its photo-responsiveness to the ultraviolet (UV) fringe of sunlight just, about 4-5 % of the solar spectrum and the fleeting existence of photogenerated electron-hole pairs, which recombine so swiftly they whittle away its quantum efficiency[9, 10].

Figure 1 : Ferrite/TiO_2 nanocomposites



The allure of Ferrite/TiO_2 nanocomposites stems from an engineering trick: they not only patch the usual drawbacks of TiO_2 photocatalysts but also slip in a prized new trait magnetism. When a magnetic

ferrite think $CoFe_2O_4$, $CuFe_2O_4$ or $NiFe_2O_4$ is paired with TiO_2, the resulting behaves, like a synergistic team. Consequently, the photocatalytic window widens into the visible-light region. The catalyst can be coaxed out of the treated water with a simple external magnet making recovery a breeze. By confronting an operational bottleneck in slurry-based photocatalytic systems this work clears a path toward economical and reusable technologies, for environmental remediation [4, 6, 8]. The latest surveys have catalogued about 1×10^5 toxic chemical species detectable, across a diverse array of water sources worldwide. Among the contaminants synthetic dyes, pesticides and pharmaceutical residues stand out for their persistence and ongoing biological activity. Their ubiquity isn't accidental; the textile sector for example pumps out roughly 10^4 dyes annually a total that tops 7×10^5 tones entering the environment [7]. Such staggering quantities inevitably load the ecosystem because even the faintest traces of dyes can tint water, curtail penetration interfere with photosynthesis and upset the balance of aquatic habitats. The dogged persistence of pollutants traces straight to the intricacy of their molecular scaffolds. Synthetic dyes for example are packed with interlocking ring systems and a tangled lattice of double bonds endowing them with a stubborn resistance, to the degradation pathways nature normally exploits. Likewise, many pharmaceuticals bear halogen-laden substituents or fused heterocyclic frameworks that further thwart breakdown. As a result, conventional wastewater-treatment schemes that depend on separation or biological digestion often stumble when they confront these recalcitrant molecular backbones. Then being fully broken down into harmless CO_2 and H_2O many of these compounds stall halfway leaving behind partial degradation products that can prove even more toxic than the original molecule. This lingering persistence drives bioaccumulation in organisms, which then pass up the food chain. To make matters worse effluents seldom contain one type of contaminant. Wastewater from textile factories, for example is a cocktail of azo dyes, dispersants, surfactants and heavy metals. When runoff from farms spills, into lakes, rivers or streams it delivers a cocktail of pesticides, nitrogen-rich fertilizers and herbicides. The way these chemicals mingle can dramatically reshape their fate sometimes they accelerate degradation through synergistic interactions other times they retard it via antagonistic effects. For instance, mixtures of contaminants can alter adsorption onto particles or nanomaterials because they vie for the reactive sites thereby shifting binding affinities. Photocatalysis has risen as an oxidation method that attacks these molecules producing potent reactive oxygen species whenever light energizes semiconductor catalysts like TiO_2. Yet in its bare form TiO_2 suffers from poor charge separation because photogenerated electrons and holes tend to recombine, which dramatically limits its efficiency. Tailoring the catalyst surface with compounds can slow this recombination and as a bonus can endow the material, with magnetic traits that simplify its recovery after use. At glance it appears mainly as a catalyst-recycling upgrade yet it also accelerates overall treatment throughput because a quicker regeneration step trims

the downtime between remediation cycles. The urgency of the problem becomes starkly evident when one examines case studies that use model pollutants. Methyl orange (MO) methylene blue (MB) and the R250 dye are frequently employed as contaminants, in evaluating remediation approaches since they mirror the characteristics typical of real effluent streams. When the initial concentrations are relatively high 100 ppm for instance the photocatalytic rate can collapse if the catalyst dose is too low. This deceleration stems from dye molecules hoarding photons, which compresses the light-penetration depth in the reaction medium and simultaneously occludes sites on the catalyst [11]. Reducing the catalyst dosage while ramping up the load aggravates the problem as it throttles the generation of electron-hole pairs, per irradiated volume. Even though many different contaminants often travel down oxidative degradation routes such as the breaking of azo bonds in dye molecules their reaction rates can vary dramatically governed by pH the presence of co-contaminants and the particular surface chemistry of the catalyst employed[12]. A number of studies that couple UV-C radiation with anatase TiO_2 have demonstrated that dye-removal efficiency is tightly tied to alkaline conditions ($pH \approx 10$) because such an environment triggers a pronounced spike, in formation. Hand-visible-light-active composites such as ZrO_2/Ag-TiO_2 have proven capable of slashing up to 90 % of pollutants AZT being a notable example within eight hours under, neutral-pH conditions. This performance hints that by tweaking the illumination wavelength and reshaping the architecture the approach could be broadened to tackle a much wider spectrum of contaminants. Beyond the chemical hazards untreated organic waste also leaves a visual scar cloudy water, odd hues that can tarnish a water body's reputation and erode its economic value. In tourism-dependent regions a river or coastal stretch that suddenly takes on a hue from effluent can devastate local livelihoods even when measured contaminant levels technically sit within nominal safety limits. Moreover, prolonged low-dose exposure, to these compounds has been linked to endocrine-disrupting effects and a rising cancer risk endangering both wildlife and the people who draw drinking water from the source. Confronted with this challenge remediation strategies must prioritize not just raw degradation efficiency but also the capacity to adapt to shifting influent characteristics over time. Real wastewater plants are anything but steady pollutant loads can swing from day to day echoing production timetables or seasonal agricultural rhythms [4]. Hence a photocatalyst's value hinges on its knack, for preserving performance without the baggage of pre-conditioning, even as those loads dance unpredictably. Seen through the prism of probing water contamination by compounds the quest is inseparably bound to forging catalytic materials that can deftly tackle a kaleidoscope of molecular shapes whether the targets are textile-derived chromophores flaunting extensive π-systems or hydrophobic pharmaceutical scaffolds that cling readily to microplastics as they drift through waterways[13]. The multifaceted nature of ferrite/TiO_2 composites hands engineer a route for contaminant removal merging physicochemical pathways: a heightened

light-driven oxidation coupled with a simple magnetic pull-out. This tandem not accelerates the primary degradation but also damps the formation of secondary pollutants that frequently hitch a ride, with clean-up schemes that lean on bulky flocculant dosages or the painstaking back-wash of membrane filters. Even as researchers keep honing these catalysts navigating obstacles such, as a light-absorption band or fouling from natural organic matter the core impetus remains unchanged: to grapple with the tenacious complex organic pollutants that still inundate freshwater supplies worldwide.

2.2.2 Complications, in Charge-Carrier Recombination

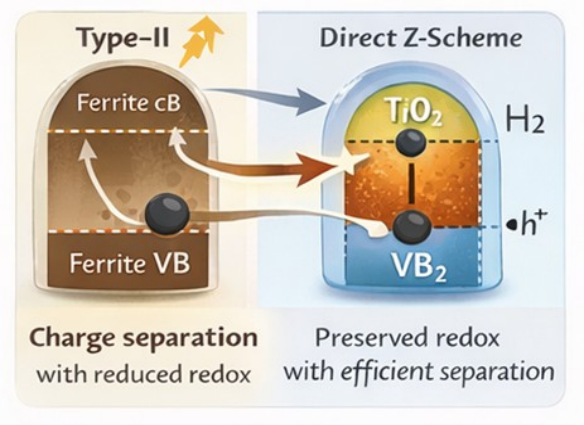
Photocatalysis relies on the generation of electron–hole (e^-/h^+) pairs upon light absorption and their subsequent migration to surface reaction sites. In TiO_2-based photocatalysts, however, rapid recombination of these carriersoften occurring on nanosecond timescales represents a dominant loss pathway, dissipating absorbed photon energy as heat or radiative emission rather than driving redox reactions. This intrinsic limitation substantially constrains the quantum efficiency of pristine TiO_2 systems. [8,14]. This intrinsic recombination remains one of the principal bottlenecks in TiO_2 photocatalysis. Photocatalysis relies on the generation of electron–hole (e^-/h^+) pairs upon light absorption and their subsequent migration to surface reaction sites. In TiO_2-based photocatalysts, however, rapid recombination of these carriers often occurring on nanosecond timescales represents a dominant loss pathway, dissipating absorbed photon energy as heat or radiative emission rather than driving redox reactions. This intrinsic limitation substantially constrains the quantum efficiency of pristine TiO_2 systems.

The formation of ferrite/TiO_2 heterojunctions has emerged as an effective strategy to suppress recombination by promoting spatial charge separation. As illustrated schematically in **Figure 2**, two distinct charge-transfer configurations are commonly discussed: Type-II and Z-scheme heterojunctions. In Type-II systems (Figure 2a), electrons and holes migrate to different semiconductor phases, reducing recombination probability but simultaneously lowering the redox potentials of the surviving carriers. While this configuration improves charge separation, it often compromises oxidative and reductive driving forces.

In contrast, Z-scheme ferrite/TiO_2 heterojunctions (Figure 2b) selectively recombine low-energy electrons and holes at the interface, leaving high-energy electrons in the TiO_2 conduction band and strong oxidizing holes in the ferrite valence band. This mechanism preserves redox strength while prolonging carrier lifetimes, thereby enhancing photocatalytic efficiency under both UV and visible-

light irradiation. The superiority of Z-scheme architectures in balancing recombination suppression and redox activity explains their frequent adoption in ferrite/TiO_2 composite designs.

Figure 2: a) Type-II and b) Z-scheme heterojunctions



a b

Despite these advantages, recombination cannot be entirely eliminated. several operational factors contribute to performance degradation over time. Lattice mismatch at the ferrite–TiO_2 interface can generate strain fields that evolve into dislocations, which act as defect-mediated recombination centres during prolonged irradiation. Surface fouling by reaction intermediates or strongly adsorbed pollutants further exacerbates recombination by blocking active sites required for effective charge utilization. Coupling TiO_2 with ferrite semiconductors to form heterojunctions has proven effective in suppressing recombination. Favourable band alignment at the interface generates an internal electric field that promotes spatial charge separation. In Z-scheme ferrite/TiO_2 systems, such as $CoFe_2O_4$/TiO_2, electrons in the ferrite conduction band preferentially recombine with holes in the TiO_2 valence band, leaving highly energetic electrons in the TiO_2 conduction band and holes in the ferrite valence band. This configuration preserves strong redox potentials while extending carrier lifetimes, thereby enhancing photocatalytic activity under both UV and visible light [4,15]. Similar improvements have been reported for TiO_2/$CuFe_2O_4$ architectures, where efficient charge separation significantly enhances pollutant degradation efficiency [10].

Despite these advantages, heterojunction systems are susceptible to performance degradation during prolonged operation. Lattice mismatch at the ferrite–TiO_2 interface can generate strain that evolves into dislocations, which act as recombination centres under repeated irradiation cycles. Additionally, surface fouling by reaction intermediates or strongly adsorbed pollutants can block active sites, redirecting charge carriers toward recombination pathways [6]. Particle aggregation represents another challenge, particularly for magnetic ferrite-based composites. Dipole–dipole interactions between

magnetic cores promote agglomeration, reducing effective surface area and increasing local carrier density, which in turn accelerates recombination [5]. Thin silica interlayers have been introduced to mitigate aggregation by physically separating particles while still permitting tunnelling-assisted charge transfer. However, excessive interlayer thickness impedes carrier mobility and suppresses interfacial redox reactions [6].

The stoichiometry of the ferrite phase further influences recombination dynamics. Adjusting cation composition alters band structure and interfacial charge transfer; for example, increasing Co^{2+} content relative to Zn^{2+} has been shown to prolong charge-separation lifetimes and improve photocatalytic performance [4]. Over-substitution, however, may introduce defect states or induce phase segregation, counteracting recombination suppression by creating additional non-radiative pathways. From a kinetic perspective, recombination reduces the effective rate constants of oxidation and reduction half-reactions by depleting the pool of reactive carriers below the threshold required for sustained radical-mediated degradation. This behaviour explains the commonly observed performance plateau in dye and pollutant degradation studies, where initial rapid removal is followed by a marked decline despite continuous illumination [3]. In contrast, optimized systems such as CuCe-ferrite/TiO_2 demonstrate near-complete degradation within short reaction times by maintaining prolonged charge separation at heterojunction interfaces. Recombination suppression can be conceptually described by maximizing the carrier lifetime (τ_s) and mean free path (l_s), according to:

$$R_{rec} \propto 1 / (\tau_s \cdot l_s)$$

Accordingly, rational catalyst design through controlled band alignment, interfacial engineering, dopant optimization, and aggregation management remains essential for reducing recombination losses and translating laboratory scale efficiencies into practical photocatalytic applications. While ferrite/TiO_2 composites offer multiple pathways to mitigate recombination, careful optimization is required to avoid trade-offs that hinder charge mobility or suppress reactive species generation.

3. Ferrite Materials, in Photocatalysis

3.1 Delving, into the Spinel Ferrite Crystal Structure

Ferrites metal oxides invariably adopt the MFe_2O_4 framework, where M denotes a divalent metal cation such as Fe^{2+}, Co^{2+}, Ni^{2+}, Cu^{2+}, Zn^{2+} or any mixture of these. Their appeal, in applications stems from a handful of standout characteristics:

Narrow Bandgap: Ferrites feature a bandgap in the 1.5–2.5 window a sweet spot that lets them harvest visible photons with impressive efficiency [8, 14]. Leveraging these formulations that embed ferrite particles can markedly boost TiO_2's photo-activity across the visible spectrum. ferrites the composite

gains magnetic separability, turning the retrieval of the photocatalyst for reuse, into a straightforward task a conclusion supported by numerous reports [3, 16, 17].In many frameworks metal cations tend to serve as co-catalysts upping surface reactivity and coaxing oxidants peroxymonosulfate (PMS) being a key example into forming the energetic sulfate radical ($SO_4^{\bullet-}$) [7, 10, 18].Choosing an M^{2+} cation opens the door to tweaking the composite's magnetic traits. Cobalt ferrite ($CoFe_2O_4$) finds application, its pronounced coercivity coupled with resilient chemical stability positioning it as a pivotal constituent, within Z-scheme photocatalytic constructs that facilitate CO_2 reduction while simultaneously decomposing dye molecules [4, 6, 19].Copper ferrite ($CuFe_2O_4$) and its kin, such as CuCe-ferrite turn out to be surprisingly effective at stripping color from dyes and breaking down pesticides a performance that originates from copper's redox hopping, between Cu^+ and Cu^{2+} ions [3, 10].Nickel ferrite ($NiFe_2O_4$) has found a niche in composites. Is also quite effective, at photodegrading dyes [16, 20]Both barium ferrite ($BaFe_{12}O_{19}$) and its strontium counterpart ($SrFe_{12}O_{19}$) sit snugly within the material family. In practice they're frequently woven into matrices, where they serve as microwave-absorbing agents and simultaneously kick-start advanced oxidation processes [1, 21, 22]. Inverse spinels $NiFe_2O_4$ being a textbook example feature an occupancy: half of the Fe^{3+} cations nestle in the tetrahedral A sites whereas the balance of the Fe^{3+} ions together with all Ni^{2+} cations are relegated to the octahedral B sites[16]. Vibrational spectroscopy offers a glimpse into this structural blueprint. In the FT-IR spectrum the absorption shoulders that cluster between 580 and 573 cm^{-1} can be traced back to the stretching of the M_Td–O bonds i.e. the dance of oxygen against metal ions, in a tetrahedral coordination sphere. The band cluster, in the 396–390 cm^{-1} window is a fingerprint of stretching motions of bound cations i.e. the M_OH–O bonds [9, 23]. The precise displacement of these absorptions correlates with which metal occupies the site and how large its ionic radius is. In magnesium ferrites the high-frequency tetrahedral stretch is chiefly driven by Fe–O, Mg-O and when zinc is present Zn-O bonds, whereas Co-O and Ni-O whose strong propensity is, for sites predominantly shape the lower-frequency band. Strong site preference steers the morphology of nanoscale particles: Co/Ni-rich systems tend to generate the particles their occupancy buttressing condensed lattice domains; Mn-containing ferrites settle at an intermediate size (15 nm) reflecting a willingness to occupy both site types[4]. The cation distribution between A and B sites isn't a fixed attribute it can shift under treatment. Annealing at temperatures promotes greater ordering, within the cation sublattice. A striking illustration comes from ferrites. When these are annealed modestly around 950 °C their saturation magnetization tends to fall. Pushing the temperature up to about 1250 °C flips the trend and the magnetization rises again. The higher heat drives the cations to reorganize into an inverse spinel pattern that's more favorable for magnetic alignment[24].

3.2 Magnetic affect photocatalysis.

Magnetic properties in ferrite-based photocatalysts do more than simply enable post-reaction recovery; they can quietly modulate the process itself by affecting charge-carrier dynamics stabilising particle dispersion and even tweaking light-absorption behavior. Ferrites such as Fe_3O_4, $CoFe_2O_4$ or $BaFe_{12}O_{19}$ when incorporated into TiO_2 matrices impart a saturation magnetisation (M_s) that hinges not on the ferrite phase but also on structural tweaks such as coatings, with non-magnetic layers. When silica or titania shells are deposited onto cores, the measured saturation magnetization (M_s) drops relative, to the uncoated material simply because the added non-magnetic coating inflates the total mass. Intriguingly X-ray fluorescence analysis often reveals that this decrease is modest roughly a 14 % reduction compared with bare magnetite confirming that the ferrite core stays essentially intact despite the coating. Preserving that magnetization is vital since letting M_s fall below a threshold would compromise the efficiency of magnetic separation after the photocatalytic step. Viewed through a lens, coercivity (Hc) and remanence (Mr) emerge as pivotal factors for recyclability and stability, in photocatalysis. Magnetic hysteresis measurements reveal that when particle dimensions lie within the single-domain window certain ferrite composites adopt a super-paramagnetic character exhibiting very low Hc and Mr values [6]. Adjusting the process subtly reshapes the interaction strength among individual particles suspended in the liquid, which in turn nudges the zeta potential and dictates how close the particles can come before magnetic attraction overwhelms electrostatic repulsion. The sweet spot an optimum arises when electrostatic forces keep the particles from clumping while the magnetic pull remains strong enough to enable efficient recovery. Recyclability data highlight where magnetism and catalytic efficiency intersect in practice. For example, recycling experiments with $CoFe_2O_4/TiO_2$ catalysts up to five cycles show a barely perceptible drop, in pollutant-removal performance. With carefully managed washing and drying steps the material sheds only about five percent of its mass each separation cycle [25]. Introducing zinc typically dampens anisotropy, which in turn pulls down coercivity. The net effect is that the particles become easier to remix for a light exposure though the trade-off may be a weaker ability to lock in long-lived carriers that benefit from tighter spin coupling. The surrounding environment adds another layer of complexity: the pH shifts and ionic-strength variations encountered in wastewater-treatment operations tweak both the surface-charge magnitude and the magnetic coupling, between catalyst grains. Under conditions the electrical double layers that coat the particles get squeezed curtailing the electrostatic repulsion that normally keeps them apart. Neglecting to adjust the saturation magnetization (Ms) enough leaves the particles prone to clump together which in turn blocks light and drags the photocatalytic efficiency down despite a theoretically favorable band structure. Ultimately reclaiming spent catalysts can shave off energy that would

otherwise be spent on filtration an approach that often succumbs to fouling yet this upside hinges on the particles maintaining a viable M_s throughout their service life without leaching or the ferrite cores crumbling under the relentless oxidative assault of repeated illumination cycles[26]. Encapsulating the cores, in a silica veil has shown promise in slowing photo dissolution while still preserving an acceptable M s. Steering clear of thick barrier layers is essential since they throw off the tunnelling distances electrons need to reach TiO_2 and wipe out the catalytic gains that careful interface design promises. In turn magnetic properties are woven into photocatalyst behaviour not just a handy handle for recovery. They influence recombination rates by shaping electronic structures dictate particle dispersion through a tug-of-war, between dipolar attractions and electrostatic repulsion determine thermal and mechanical durability across reuse cycles and might even lend a slight edge to light absorption via local electromagnetic field effects[6]. Constructing the ferrite/TiO_2 composite demands a panoramic appraisal of the myriad influences, at play so that the magnetic gains pursued never inadvertently shut off active site accessibility or throttle carrier mobility both of which underpin sustained photocatalytic efficiency.

4 Titanium Dioxide, in Photocatalysis

4.1 Anatase, Rutile and Brookite (The trio of TiO_2 polymorphs).

Titanium dioxide appears in three crystal modifications anatase, rutile and brookite each bearing distinct physicochemical traits that shape its photocatalytic character. In practice anatase and rutile have attracted the lion's share of attention mainly because their synthesis routes are comparatively straightforward and their performance benchmarks are well-charted[27]. Anatase adopts a structure and remains metastable at room temperature whereas rutile, also tetragonal but denser occupies the thermodynamically stable niche. Brookite carries an orthorhombic lattice that seldom appears in contexts owing to the need for finely tuned control over nucleation pathways during fabrication. From a standpoint anatase often takes precedence because its electron–hole recombination proceeds more slowly than that of rutile [25]. Titanium dioxide (TiO_2) remains the benchmark photocatalyst due to its unparalleled advantages. It primarily exists in three crystalline phases: anatase, rutile, and brookite. Anatase is generally considered the most photocatalytically active phase due to its higher Fermi level and slower charge carrier recombination rate [8, 14]. Defects perched at the boundary become trapping sites, slowing electron–hole recombination and thereby supplying more carriers, for photocatalytic reactions. Seeing mixed-phase systems, in action makes it obvious that a material often deemed inferior can actually play a role when it's embedded in a well-designed structure. Brookite however stays a bit of a puzzle high-purity samples are scarce and the phase tends to shift into anatase or rutile during heat treatment. Still brookite has proven to be an UV absorber and its orthorhombic

coordination may endow it with distinctive surface chemistry. Its conduction- and valence-band edges sit a shade apart from those of anatase or rutile hinting that unconventional heterojunction configurations could be crafted to expand the material's spectral responsiveness. Phase changes are tightly coupled to the synthesis route and calcination conditions. For example TiO_2 obtained via hydrolysis of titanium alkoxides typically yields anatase when the calcination temperature hovers, around 400–500 ∘C [25]. Pushing the heat beyond 600 ∘C sparks the nucleation of rutile domains and by 700 ^∘C or higher rutile usually dominates unless dopants are introduced to kinetically lock in the anatase phase[16]. The escalating pollution crisis driven by a surge of industrial effluents lingering pharmaceutical residues and pesticide runoff demands remediation approaches that are both cutting-edge and sustainably practicable. Heterogeneous photocatalysis, which captures energy to power the dismantling of contaminants has risen to prominence as a leading contender in this field. Within the family of photocatalysts, titanium dioxide (TiO_2) commands the share of attention thanks, to its formidable oxidative strength, chemical durability, low cost and non-toxic character [8, 14]. The ambition to embed TiO_2 on a large scale runs into two persistent hurdles: an excessively wide bandgap that confines its activity, to the UV slice of sunlight barely 4–5% and the swift recombination of photogenerated electron-hole pairs, which saps its overall quantum efficiency [9, 10, 28].

Ferrite/TiO_2 nanocomposites stand out because clever engineering mitigates their drawbacks while adding an unexpectedly useful trait: magnetism. By coupling a ferrite such, as $CoFe_2O_4$, $CuFe_2O_4$ or $NiFe_2O_4$ with TiO_2 a synergistic composite is formed. This union not pushes the photocatalytic activity further into the visible-light range but also makes it possible to separate and recover the catalyst from treated water with a simple external magnet. By tackling a snag in slurry-based photocatalytic systems this work clears a pathway toward cheap reusable technologies, for environmental cleanup [6, 8].

4.2 Activation Using UV Light.

Historically TiO_2 has enjoyed a reputation as an efficient photocatalyst whenever ultraviolet light shines on it thanks to its broad bandgap and formidable oxidative power[29]. Yet relying solely on UV to fire up the reaction introduces a bottleneck, for harnessing solar energy since ultraviolet radiation makes up only about 3–5 % of the sunlight that reaches Earth's surface. In practice this imposes a bottleneck on any natural-sunlight-driven process forcing UV-based catalytic systems to either resort to high-intensity artificial illumination or to endure markedly longer reaction times to reach comparable pollutant-degradation levels. The anatase phase of TiO_2 with its bandgap hovering around 3.2 eV primarily captures UV photons; its rutile counterpart squeezes the gap down to 3.0 eV yet this modest narrowing still leaves it far, from the bandgap required for efficient visible-light activation[30]. When the system is driven by UV light, the photogenerated electrons that climb into the conduction

band together with the holes left in the valence band spark redox reactions capable of fully mineralizing complex organic molecules into harmless CO_2 and H_2O. Yet this tactic squanders the visible portion of sunlight essentially throwing away the lion's share of ambient solar energy. Moreover, restricting activation, to UV photons also imposes a selectivity shaping how the reactions proceed. Robust oxidative agents such as hydroxyl radicals (·OH) and superoxide anions ($\cdot O_2^-$) generated by TiO_2 under UV illumination act as oxidants, capable of dismantling a vast array of chemical structures [30, 31].

5 Ferrite/TiO_2 Nanocomposites.

5.1 Interface Engineering.

The photocatalytic performance of ferrite/TiO_2 nanocomposites isn't the sum of its parts; it's largely dictated by the atomic-level interactions at the interface, between ferrite and TiO_2. Interface engineering means deliberately designing and tweaking that junction to steer charge carriers improve stability and ultimately push photocatalytic efficiency to its peak. The goal is to craft a seamless intimate electronically coupled heterojunction that strongly suppresses charge recombination and eases reactions.

Where ferrite and TiO_2 encounter each other lies the interface a corridor linking two distinct semiconductors. The way their energy bands the apex of the valence band and the floor of the conduction band align in this space dictates what becomes of the electrons (e^-) and holes (h^+) generated by light. When the junction is deliberately engineered the resulting internal electric field can be coaxed to shepherd these charge carriers apart enhancing their separation.

In an arrangement with staggered band structures illumination causes electrons to cascade from the higher-lying conduction band into the lower one while holes drift from the upper valence band down, to its lower counterpart. This spatial partitioning of the photogenerated carriers effectively suppresses recombination. In the hybrid used for 2,4-D degradation the prevailing view is that photogenerated electrons residing in the $CuFe_2O_4$ conduction band migrate into TiO_2 while holes created in TiO_2's valence band move in the opposite direction, toward $CuFe_2O_4$. This bidirectional charge transfer, which effectively separates electrons and holes is identified as the primary driver of the markedly enhanced photocatalytic activity [10]. Likewise, the superior dye-decolorization observed with MnCo-Ferrite/TiO_2 and CuCe-ferrite/TiO_2 composites is ascribed to the interfacial charge-separation mechanism[3, 17, 32].

What emerges is an advanced generally more efficient approach that emulates natural photosynthesis. In a Z-scheme the low-energy electrons lingering in the TiO_2 conduction band bearing weaker reducing power pair up with the low-energy holes in the ferrite valence band possessing weaker oxidizing power

right, at their interface. Consequently the high-energy electrons stay in the ferrite conduction band while the strongest holes remain in the TiO_2 valence band poised to drive the reactions. This scheme has been explicitly examined for Ferrite/TiO_2-nanocomposites employed in CO_2 conversion. The investigation revealed a link, between the Co-Zn metal composition of the ferrite and its catalytic performance a relationship that directly reshapes the band structure and boosts the efficiency of the Z-scheme charge-transfer pathway at the interface [4].Consequently the system retains redox potentials while delivering outstanding charge separation.

These employed techniques let TiO_2 nucleate and grow straight onto pre-formed ferrite nanoparticles resulting in a core-shell or tightly bonded architecture. For example, TiO_2-$CoFe_2O_4$ and TiO_2-$CuFe_2O_4$ composite films have been prepared through a route that relies on the thermal decomposition of metal-organic precursors, a process that yields a uniform distribution and strong interfacial adhesion[9]. Whether the ferrite particles are cultivated within a TiO_2 matrix or the oxide is deposited onto the ferrite the method guarantees contact area, between the two phases. The Ni ferrite/TiO_2 nanofibrous composite was produced by electrospinning, entrapping the ferrite precursors within a TiO_2-polymer fiber matrix; calcination then fuses everything into an interwoven architecture [16]. With this progress several hurdles remain. Achieving pinpoint control over the atomic arrangement suppressing defect sites that serve as recombination traps and definitively establishing the charge-transfer pathway whether it follows a Z-scheme or a Type-II mechanism are far, from easy tasks. Future research directions include:

Doping at the Interface: Introducing dopants (Ce [13]) directly at the interface to establish "bridge" energy levels that further promote charge transfer. Construction of Multi-Component Interfaces: Adding a material such, as graphene or other semiconductors (e.g., GO-doped systems [1, 31]) to build multi-interface structures that can precisely steer charge separation and provide conductive pathways. Advanced Characterization: By harnessing in-situ techniques we can watch the charge dynamics and chemical environment at the ferrite/TiO_2 interface unfold in time as photocatalytic reactions proceed.

All things considered it’s the interface that truly holds the key to crafting high-performance ferrite/TiO_2 nanocomposites. When researchers move past the habit of merely mixing the powders together and instead meticulously design a nanoscale heterojunction, they unlock the full punch of these materials paving the way, for efficient down-to-earth environmental remediation.

5.2 Charge-Transfer Mechanisms in Ferrite-TiO_2 Nanocomposites.

The remarkable photocatalytic efficiency of Ferrite/TiO_2 nanocomposites is attributed to a series of rapid charge-transfer events occurring across the heterojunction interface. Understanding these events

how light generates electron-hole pairs, how charges separate, move, and are eventually utilized is crucial for any material design initiative. The referenced studies delineate the fundamental mechanisms and the variables influencing them.

The dynamic synergy between ferrite and TiO_2 establishes a pathway for charge migration, with academic literature identifying two primary mechanisms. The standard model of charge transfer is essentially based on band alignment. In a ferrite/TiO_2 composite, the conduction band (CB) and valence band (VB) of narrow-bandgap ferrites, such as $CuFe_2O_4$ or $CoFe_2O_4$, are situated at higher energy levels than the corresponding bands in TiO_2.Both semiconductors act as photon absorbers; ferrite is excited by light, while TiO_2 remains inactive unless exposed to ultraviolet photons. Electrons in the ferrite's high-energy conduction band find a lower-energy state in TiO_2's conduction band and migrate across the interface. Concurrently, photogenerated holes in TiO_2's valence band move into the higher-energy valence band of the ferrite. This leads to a spatial separation of charge carriers electrons accumulate in TiO_2 while holes remain in the ferrite significantly reducing the likelihood of recombination. This mechanism frequently explains the activity observed in systems such as MnCo-Ferrite/TiO_2 and CuCe-ferrite/TiO_2 composites, which can decolorize dyes under light, a performance attributed to a Type-II charge flow that effectively separates charges [3, 17].

Direct Z-Scheme Mechanism, This mechanism carefully maintains the high redox potential of each component. In a Z-scheme, the interfacial region acts as a conduit, facilitating the recombination of charge carriers that have minimal impact on performance. Photons excite both semiconductor layers. Electrons in TiO_2's conduction band, with reducing power, move to the neighboring semiconductor's band and pair with holes in the ferrite's valence band, which exhibit weak oxidizing ability at the interface. Post-recombination, the ferrite's conduction band retains high-energy electrons, while TiO_2's valence band accumulates potent holes. The system surprisingly maintains both reducing and oxidizing abilities simultaneously while achieving charge separation, useful for high-energy reactions such as CO_2 reduction. This mechanism is supported by research on Co-Zn ferrite/TiO_2 nanocomposites used for CO_2 conversion, where metal composition is linked to performance, indicating that tuning the ferrite's band structure directly influences Z-scheme charge-transfer efficiency[4]. The redox potential serves as a distinctive indicator of this pathway.

Ferrite as an Electron Pump and Sensitizer, In visible-light-driven catalysis, ferrite functions as a photosensitizer. Visible light excites the material, leading to electron-hole pair formation. Electrons are transferred from the ferrite's conduction band to TiO_2's conduction band, loading TiO_2 with electrons even though it does not need to absorb light. This process transforms wide-bandgap TiO_2 from a semiconductor into a light-driven catalyst, enabling electrons in TiO_2's conduction band to

reduce O_2 and generate superoxide radicals ($\bullet O_2^-$).This sensitizing effect is key to the activity observed in composites like Ni-ferrite/TiO_2 nanofibrous composite and $TiO_2/CuFe_2O_4$[10].

5.3-Enhancing Dynamics through Defect Engineering and Dopants.

Slight doping-induced modifications of an interface's structure can tailor charge flow, doping with cerium (Ce) introduces energy states within the bandgap. These energy levels act as a conduit for charge carriers, improving multi-step transfer efficiency over direct jumps, and trapping electrons or holes to further reduce recombination. Magnetic TiO_2 doped with cerium shows enhanced performance, and Ce^{3+}-doped Ni-Zn ferrite/TiO_2 composites exhibit significant behavioral improvements [13, 20, 23]. In both systems, cerium species act as charge-transfer mediators, accelerating charge separation at the interface.

Synergistic Charge-Transfer in Oxidation Processes (AOPs Peroxymonosulfate (PMS) activation weaves charge-transfer dynamics into a photochemical cycle.Photochemical Activation: Light-excited TiO_2 or ferrite generates electrons that directly activate persulfate via the pathway

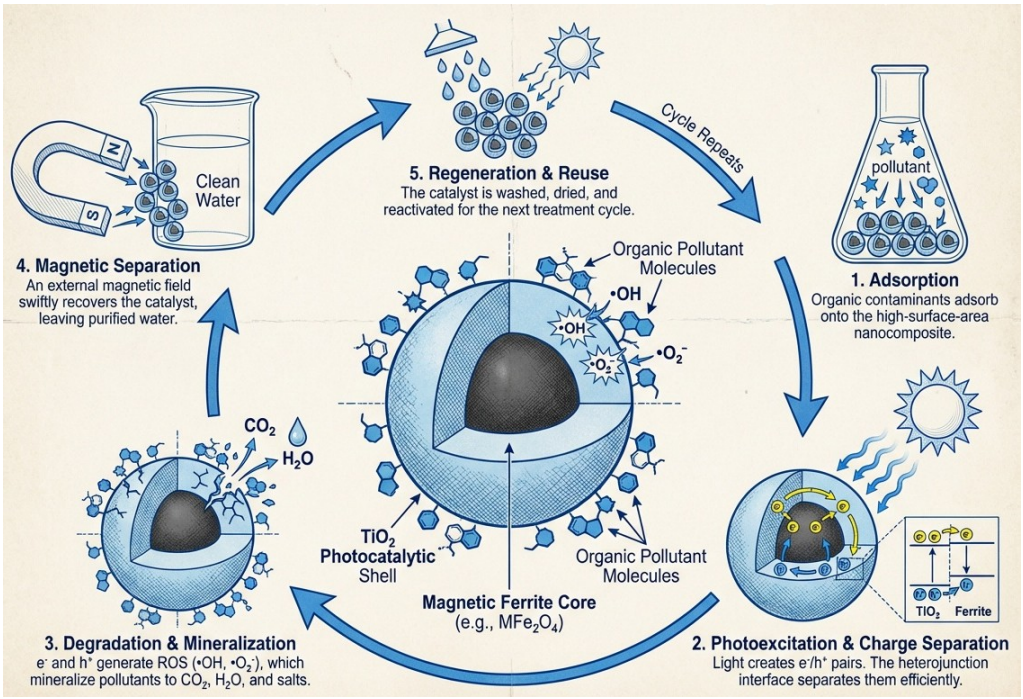
$$HSO_5^- + e^- \rightarrow SO_4\bullet^- + OH^-.$$

Catalytic Activation, Metal ions within the ferrite, such as Co^{2+}/Co^{3+} and Cu^{2+}/Cu^+, efficiently decompose PMS.

Dynamic Synergy: The heterojunction acts as a source of electrons (or holes), sustaining the metal-ion redox cycle. For instance, electrons transferred to TiO_2 can reduce Co^{3+} to Co^{2+}, facilitating efficient catalytic PMS activation. Studies show synergy in TiO_2/Fe_2O_3/Chitosan [18] for degradation and $CoFe_2O_4$/Gr [7] for amoxicillin breakdown. The heterojunction's ability to separate charge carriers is closely linked to sulfate radical ($SO_4\bullet^-$) generation.

The charge-transfer dynamics in Ferrite/TiO_2 nanocomposites cannot be simplified into a single pathway; rather, they unfold as a multifaceted synergistic interaction with competing routes. The prevalence of either a Type-II cascade or a Z-scheme depends on the ferrite formulation, the synthesis method shaping interfacial quality, and the specific photocatalytic reaction under investigation. Thus, the ultimate goal of interface engineering is to direct these dynamics so that reactive charge carriers not only persist longer but also concentrate at the surface, enabling highly efficient pollutant degradation.

Schematic 2: The Self-Sustaining Cycle of a Magnetically Recoverable Ferrite/TiO_2 Photocatalyst



6. Real-world uses of Ferrite/TiO_2 nanocomposites.

6.1 Organic Pollutants Break Down in Water.

The synthesis of these composites involves various methods, including sol-gel, co-precipitation, hydrothermal processes, and electrospinning, to ensure an intimate contact between the ferrite and TiO_2 phases, which is crucial for efficient charge transfer [9, 27, 33]. Synthetic dyes are often chosen as stand-in pollutants in laboratory studies because they are vividly colored and chemically varied. In practice methylene blue (MB) and crystal violet (CV) serve as model substrates, for assessing photocatalytic performance. In visible-light-driven degradation trials TiO_2 catalysts loaded with nickel ferrite showed a striking uptick in pseudo-first-order rate constants compared with bare TiO_2. The boost is mainly credited to the inverse-spinel $NiFe_2O_4$ phase, which damps electron-hole recombination [10]. This isn't a lab-side note; real textile effluents often contain dye levels above 100 ppm and fast breakdown is crucial, for reinstating water quality before discharge. When the apparent reaction rate climbs under light the system taps solar irradiation more efficiently thereby mitigating a key drawback of UV-only activation setups [23]. Pharmaceuticals constitute another class of pollutants that ferrite/TiO_2 can tackle. In experiments with sulfamethoxazole (SMX) antipyrine (AP) and ibuprofen (IBP) $CoFe_2O_4/TiO_2$ catalysts achieved near-complete removal and their catalytic activity remained essentially unchanged, over five reuse cycles [34]. Recyclability here hinges on retrieval by snatching the catalyst out magnetically loss between runs is virtually eliminated and the surface stays free of fouling. The sturdiness of these hybrids traces back to an interfacial strategy: a feather-thin SiO_2 sheath draped over ferrite cores acts as a shield against photodissolution while still

permitting tunneling-assisted electron flow, from TiO_2 into the ferrite's conduction bands [4]. Stability becomes a linchpin when tackling waste streams that host a mélange of molecular species and often contain chloride or sulfate ions competing for reactive radicals; if the interfacial layer collapses midway catalytic activity nosedives and residual pollutants linger. Visible-light-driven designs push the envelope further enabling degradation without the crutch of UV illumination. For example, the ternary nanocomposite g-C_3N_4/CdS/$CuFe_2O_4$ has demonstrated up to 85 % removal of agents such as AZT, within a 90-minute window [23]. The intricate heterojunction scaffold channels charge along a web of routes that reconfigure with the wavelength of incoming light simultaneously igniting oxidative and reductive pathways at separate locales while exploiting the distinct spectral windows of each constituent. Although CdS-based platforms are plagued by photocorrosion and the specter of heavy-metal leaching under operation the underlying concept deliberately pairing narrow-bandgap and wide-bandgap semiconductors with magnetic ferrites offers a roadmap toward safer composites built entirely from environmentally benign building blocks. In practice the speed at which pollutants are decomposed inevitably hinges on the pollutant's concentration, to the amount of catalyst applied. High dye loading with a low catalyst mass makes the dye molecules themselves gobble up photons cutting the depth that light can travel a phenomenon called the filter effect, which in turn damps photogeneration rates inside the catalyst [28]. Such uniform dispersion alleviates mass-transfer constraints ensuring pollutants encounter sites, throughout the reactor rather than being trapped in aggregated clusters that are shielded from irradiation. Moreover, tailoring the architecture such as coating ferrite cores with mesoporous TiO_2 shells simultaneously enlarges the accessible surface and boosts permeability, for larger pollutant molecules that would otherwise diffuse sluggishly toward the reactive zones [7]. When photocatalytic conditions are optimal reaction intermediates tend to accumulate fleetingly before they undergo complete mineralization. Certain intermediates may adsorb tenaciously onto the catalyst surface creating a blockage risk if they are not promptly removed by oxidation steps; designing the system for a high turnover frequency diminishes this fouling propensity. Tuning the illumination wavelengths and intensity profiles throughout treatment cycles can speed up the breakdown of intermediates; for example, alternating UV-rich phases with phases leverages the differing dominance of reactive oxygen species across the spectrum and capitalizes on the responsiveness of multicomponent composites [34]. Longer operational lifespans directly enhance feasibility for large-scale deployments such as municipal wastewater plants or decentralized rural installations. Magnetic recovery also reduces reliance on filtration which is prone, to clogging when handling high-turbidity influents loaded with organic matter and suspended solids[4]. So by using ferrite/TiO_2 catalysts that can be magnetically pulled out after each cycle fresh particulate pollutants stay out of the treated water a sustainability win, over powders that must be replaced after a short

lifespan. Embedding SiO_2 layers helps balance accessibility with chemical robustness keeping performance over many cycles close, to the initial output[34]. Still design trade-offs linger: raising coercivity by tweaking the ferrite composition improves retrievability. It can provoke lingering aggregation because residual magnetization remains once the particles leave the magnetic field[9]. Lowering coercivity eases redispersion before the cycle but it may hinder recovery efficiency in large-scale setups unless strong external magnetic gradients are applied to pull all active material out of solution quickly. Striking a balance between constraints and photocatalytic metrics still unfolds as an iterative back-and-forth, guided by both field trials and tightly controlled laboratory experiments. Consequently, whether the target molecule is a chromophoric azo dye of textile effluents or a robust pharmaceutical contaminant from hospital wastewater embedding magnetic ferrites, within TiO2 matrices yields a durable effective remediation sustaining photocatalytic activity while enabling a facile post-treatment separation. By weaving the adsorption perks inherent to ferrite surface chemistry with the electronic boons of engineered heterojunctions and grounding the whole architecture in morphologies that stay well-dispersed in water ferrite/TiO_2 nanocomposites emerge as a multipurpose toolkit, for breaking down organic pollutants across a panoply of environmental scenarios while sidestepping steep operating costs and secondary-contamination pitfalls [10, 20, 34].

6.2 Photocatalytic Generation of Hydrogen

Photocatalytic hydrogen production, with ferrite/TiO_2 nanocomposites blends the light-driven activity of a semiconductor with the benefit of magnetic recovery so water can be split and the catalyst reclaimed in the same step. The process hinges on exciting TiO_2: electrons are promoted into the conduction band (E_CB) where they reduce protons to hydrogen while the complementary holes left in the valence band (E_VB) oxidize water or any sacrificial agents present. In a TiO_2 system this two-half-reaction cycle is often thwarted by fast electron-hole recombination and a narrow spectral window. Linking the system with ferrites such as $NiFe._2O_4$, $CoFe_2O_4$ or $ZnFe_2O_4$ offers a route around those limitations. It does so by forming heterojunctions that boost charge separation and depending on the ferrite's composition can push absorption further into the visible spectrum[35, 36]. Band-edge positions are crucial: the absolute electronegativity χ of each component, with its band gap E_g sets the conduction- and valence-band energies through established relations.

$$E_{\mathrm{CB}} = \chi - E_C - \frac{E_g}{2}$$
$$E_{\mathrm{VB}} = E_{\mathrm{CB}} + E_g$$

Here E_C denotes the free-electron energy on the hydrogen scale 4.5 eV [16]. In a $NiFe_2O_4/TiO_2$ heterojunction $NiFe_2O_4$ with $\chi = 4.65$ eV interfaced to TiO_2 with $\chi = 5.8$ eV the resulting offset nudges

electrons, from TiO_2 into the ferrite under illumination, which in turn slows recombination and keeps the reduction potentials aligned for H_2 evolution. The specific material synthesis pathway chosen can markedly shape these benefits. Sol-gel derived $CoFe_2O_4/TiO_2$ has shown a stable hydrogen-production output over many visible-light cycles a robustness that stems from its uniformly dispersed particles and a meticulously tuned interfacial contact[37, 38]. Hydrothermal routes on the hand can produce inverse-spinel ferrites such, as $NiFe_2O_4$; their nanometer-scale particles enlarge the surface area and pile up reaction sites thereby boosting both photonic absorption and proton-reduction activity. Researchers have begun inserting an interlayer most often SiO_2 between magnetic cores and TiO_2 shells to guard against photo dissolution while still preserving the short-range tunneling needed for charge carriers to reach the reactive zones where water oxidation occurs[39, 40]. By tuning the design levers along these axes ferrite/TiO_2 nanocomposites shift from merely viable to potentially superior candidates for large-scale photocatalytic hydrogen generation marrying renewable-energy conversion, with simple catalyst handling and recyclability [16].

In conclusion, Ferrite/TiO_2 nanocomposites present a highly promising and rationally designed platform for photocatalytic hydrogen generation. By leveraging visible-light absorption through the ferrite and enabling superior charge separation via engineered heterojunctions, these materials translate the proven principles of enhanced photocatalytic remediation directly to the critical challenge of sustainable fuel production.

7. Conclusion

Ferrite–TiO_2 nanocomposites are carving out a niche as a promising suite of multifunctional materials that fuse magnetic behavior with photocatalytic action to meet pressing environmental demands. By stitching magnetic ferrite nanoparticles and the semiconductor titanium dioxide these hybrids can swiftly dismantle recalcitrant organic pollutants in water and air and they can be recovered with a simple magnetic pull. The cooperative interaction, between the TiO_2 phases boosts charge-carrier separation slashing the recombination losses that typically throttle photocatalytic performance in ordinary semiconductor systems. In practice the gain translates into a bump in quantum yields and a sped-up demolition of pollutants particularly when heterojunctions are painstakingly engineered to align the bands just right and to ease the flow of charge, across the interface.

In water-treatment settings ferrite/TiO_2 composites have proven effective at breaking down synthetic dyes, pharmaceuticals and other tenacious organics thanks, to their boosted adsorption capacity and lengthened charge-carrier lifetimes. Their magnetic retrievability curtails secondary-pollution concerns stemming from powder loss or filter fouling allowing the catalysts to be recovered and reused

across cycles. Likewise in hydrogen generation these nanocomposites exploit heterojunction-driven charge separation and visible-light absorption to lift hydrogen evolution rates while preserving catalyst stability and recyclability. When magnetism and photocatalysis dance together they dictate how particles spread out and how they grab light and that choreography bumps, up the system's efficiency.

At the end of the day ferrite/TiO_2 nanocomposites serve as a toolkit that can outmaneuver the built-in constraints of classic photocatalysts pairing amplified light-induced reactivity with an effortlessly recoverable magnetic handle. Their knack for being tuned to a spectrum of pollutants operating conditions and illumination regimes makes them attractive prospects for large-scale, low-energy water cleaning, air-quality management and the production of hydrogen. As the field pushes these composites forward, they are set to play a substantive part in the worldwide drive, to tame environmental pollution and foster clean-energy solutions.

Conflicts of Interest: There is no conflict of interest.

References

[1] M. Suneetha *et al.*, "Green Visible light catalysis for dye effluent Degradation with special reference Alizarin red and Antibacterial activity using $SrFe_{12}O_{19}/TiO_2/SiO_2$/GO Nanocomposite," in *International Conference on Bio-Based Environment for Sustainable Territory (ICBEST 2024)*, ed: Atlantis Press, 2025, pp. 195-220 , 94-6463-648-3.

[2] A. Khan *et al.*, "$FeSe_2/TiO_2$ heterostructure as an efficient photocatalyst and their electrochemical energy storage applications," *Materials Chemistry and Physics,* vol. 303, pp. 127793 . 0254-0584, 2023.

[3] E. T. Helmy, U. A. Soliman, A. M. Elbasiony, and B.-S. Nguyen, "CuCe-ferrite/TiO_2 nanocomposite as an efficient magnetically separable photocatalyst for dye pollutants decolorization," *Topics in Catalysis,* vol. 66, no. 1, pp. 53-63, 2023.

[4] R.-G. Ciocarlan *et al.*, "Ferrite/ TiO_2-nanocomposites as Z-scheme photocatalysts for CO_2 conversion: Insight into the correlation of the Co-Zn metal composition and the catalytic activity," *Journal of CO_2 Utilization,* vol. 36, pp. 177-186, 2020.

[5] K. Duan *et al.*, "A facile route to synthesize n-SnO_2/p-$CuFe_2O_4$ to rapidly degrade toxic methylene blue dye under natural sunlight," *RSC advances,* vol. 12, no. 26, pp. 16544-16553, 2022.

[6] A. Zielinska-Jurek *et al.*, "Design and application of magnetic photocatalysts for water treatment. The effect of particle charge on surface functionality," *Catalysts,* vol. 7, no. 12, p. 360, 2017.

[7] E. B. Lashkaryani, B. Kakavandi, R. R. Kalantary, A. J. Jafari, and M. Gholami, "Activation of peroxymonosulfate into amoxicillin degradation using cobalt ferrite nanoparticles anchored on graphene ($CoFe_2O_4$/ Gr)," *Toxin Reviews,* 2021.

[8] P. Shandilya, S. Sambyal, R. Sharma, A. Kumar, and D.-V. N. Vo, "Recent advancement on ferrite based heterojunction for photocatalytic degradation of organic pollutants: a review," *Ferrite: Nanostructures with tunable properties and diverse applications,* pp. 121-161, 2021.

[9] D. Balatskiy, Y. Budnikova, S. Bratskaya, and M. Vasilyeva, "TiO_2-$CoFe_2O_4$ and TiO_2-$CuFe_2O_4$ composite films: A new approach to synthesis, characterization, and optical and photocatalytic properties," *Journal of Composites Science,* vol. 7, no. 7, p. 295, 2023.

[10] M. Golshan, B. Kakavandi, M. Ahmadi, and M. Azizi, "Photocatalytic activation of peroxymonosulfate by TiO_2 anchored on cupper ferrite (TiO_2/ $CuFe_2O_4$) into 2, 4-D degradation: Process feasibility, mechanism and pathway," *Journal of hazardous materials,* vol. 359, pp. 325-337, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0304389418305090?via%3Dihub.

[11] E. T. Helmy, M. A. Ayyad, M. A. Ali, H. G. Mohamedbakr, and J. H. Pan, "Biochemical, Histological Changes, Protein Electrophoretic Pattern, and Field Application of CuPb–Ferrite/TiO_2 Nanocomposites for Controlling Terrestrial Gastropod Eobania vermiculata (Müller)," *Journal of Agricultural and Food Chemistry,* vol. 71, no. 17, pp. 6626-6634, 2023.

[12] N. I. Zouli, "Photodegradation of a Broad-Spectrum Antibiotic Azithromycin Using H_2O_2 under Ultraviolet Irradiation," *International Journal of Molecular Sciences,* vol. 25, no. 12, pp. 6702 , 1422-0067, 2024.

[13] T. Rahmiati, C.-C. Chen, S. A. Saad, N. Eko, and A. Apriana, "Surface Character of Magnetic Ce-doped TiO_2 for Photocatalytic Performance Enhancement," *Recent in Engineering Science and Technology,* vol. 1, no. 01, pp. 1-6, 2023.

[14] H. Ma and C. Liu, "A mini-review of ferrites-based photocatalyst on application of hydrogen production," *Frontiers in Energy,* vol. 15, no. 3, pp. 621-630, 2021.

[15] R. Suresh, S. Rajendran, P. S. Kumar, D.-V. N. Vo, and L. Cornejo-Ponce, "RETRACTED: Recent advancements of spinel ferrite based binary nanocomposite photocatalysts in wastewater treatment," *Chemosphere,* vol. 274, p. 129734, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045653521002034?via%3Dihub.

[16] H. C. T. Firmino *et al.*, "Nickel ferrite/TiO_2 nanofibrous composite: enhanced photocatalytic dye degradation under visible light," *Materials Research,* vol. 27, p. e20230391, 2024.

[17] S. Yousefi-Mohammadi, M. Movahedi, and H. Salavati, "MnCo–Ferrite/TiO_2 composite as an efficient magnetically separable photocatalyst for decolorization of dye pollutants in aqueous solution," *Surfaces and Interfaces,* vol. 11, pp. 91-97, 2018.

[18] M. Ahmadmoazzam, H. Akbari, A. Adibzadeh, S. Pourfadakari, and H. Akbari, "Visible-light-driven TiO_2/ Fe_2O_3/Chitosan nanocomposite with promoted photodegradation of meropenem and imipenem antibiotics by peroxymonosulfate," *Environmental Technology,* vol. 45, no. 17, pp. 3456-3467, 2024.

[19] S. Rohilla, "Rietveld refinement and structural characterization of TiO_2/$CoFe_2O_4$nanocomposites," in *IOP Conference Series: Materials Science and Engineering* vol. 872, ed: IOP Publishing, 2020, pp. 012171 , 1757-899X.

[20] S. Das *et al.*, "Structural, magnetic, and dielectric behaviour of Ce^{3+} doped Ni-Zn ferrite/TiO_2 magneto-dielectric nanocomposites," *Applied Physics A,* vol. 131, no. 10, pp. 1-21, 2025.

[21] R. Peymanfar and M. Rahmanisaghieh, "Preparation of neat and capped BaFe2O4 nanoparticles and investigation of morphology, magnetic, and polarization effects on its microwave and optical performance," *Materials Research Express,* vol. 5, no. 10, p. 105012, 2018.

[22] J. Zhao *et al.*, "Lanthanum and neodymium doped barium ferrite-TiO_2/MCNTs/poly (3-methyl thiophene) composites with nest structures: preparation, characterization and electromagnetic microwave absorption properties," *Scientific reports,* vol. 6, no. 1, p. 20496, 2016.

[23] S. Das, G. C. Nayak, S. K. Sahu, P. C. Routray, A. K. Roy, and H. Baskey, "Titania-coated magnetite and Ni-ferrite nanocomposite-based RADAR absorbing materials for camouflaging application," *Polymer-Plastics Technology and Engineering,* vol. 54, no. 14, pp. 1483-1493, 2015.

[24] L. Saini, M. K. Patra, R. K. Jani, G. K. Gupta, A. Dixit, and S. R. Vadera, "Tunable Twin Matching Frequency (f m1/f m2) Behavior of Ni1− xZnxFe2O4/NBR Composites over 2–

12.4 GHz: A Strategic Material System for Stealth Applications," *Scientific Reports,* vol. 7, no. 1, pp. 44457 , 2045-2322, 2017.

[25] H. Gong, "Photodegradation of pharmaceuticals with a recyclable catalyst $CoFe_2O_4/TiO_2$ in aqueous phase," 2018.

[26] H. O. U. Cuiling, L. I. Tiehu, Z. Xueyun, and C. Xiaowen, "Double-layer Electromagnetic Wave Absorber Based on Carbon Nanotubes Doped with La $(NO_3)_3$ and Fe_3O_4 Nanoparticles," *Materials Science,* vol. 23, no. 3, pp. 200-204, 2017.

[27] S. Rohilla, "Rietveld refinement and structural characterization of TiO2/CoFe2O4nanocomposites," 2020, vol. 872: IOP Publishing, pp. 012171 , 1757-899X.

[28] A. A. Alamri *et al.*, "Green magnetically separable molluscicide Ba–Ce–Cu ferrite/TiO2 nanocomposite for controlling terrestrial gastropods Monacha Cartusiana," *Scientific Reports,* vol. 15, no. 1, p. 2888, 2025. [Online]. Available: https://www.nature.com/articles/s41598-025-85730-8.pdf.

[29] R. Peymanfar, F. Norouzi, and S. Javanshir, "A novel approach to prepare one-pot Fe/PPy nanocomposite and evaluation of its microwave, magnetic, and optical performance," *Materials Research Express,* vol. 6, no. 3, p. 035024, 2018.

[30] S. S. Rasooly, M. Anwer, and G. Tsnim, "A review on treatment methods for pesticide contaminated water," *Journal of Environmental Science e, Toxicology and Food Technology,* vol. 16, p. 24, 2022.

[31] A. Rahimi, M. H. Zonoozi, R. Rahimi, and B. Zahabiyoun, "Photocatalytic activity of GO-doped bismuth-based photocatalyst for Methyl Orange decolorization under visible light irradiation," *Desalination and Water Treatment,* vol. 180, pp. 360-369, 2020.

[32] A. M. Elbasiony, A. A. Alamri, U. A. Soliman, H. G. Mohamedbakr, A. M. Wahba, and E. T. Helmy, "Green synthesis of N, B co-doped TiO2 nanoparticles with enhanced photocatalytic and antioxidant activities," *Journal of the Taiwan Institute of Chemical Engineers,* pp. 105626 , 1876-1070, 2024.

[33] E. E. Zaghloul, A. S. Amin, M. A. Diab, A. Elsonbati, and M. Ibrahim, "Removing 2, 4-dichlorophenoxyacetic acid (2, 4-D) from polluted water using zinc ferrite nanoparticles," *Egyptian Journal of Chemistry,* vol. 63, no. 4, pp. 1411-1428, 2020.

[34] P. Jiang, Y. Yu, and K. Li, "Hydrophilic TiO2/ MWCNT/PVDF membrane for enhanced photodegradation of methyl orange in water," *Fullerenes, Nanotubes and Carbon Nanostructures,* vol. 31, no. 12, pp. 1185-1191, 2023.

[35] C. Mahala, M. D. Sharma, and M. Basu, "2D nanostructures of CoFe2O4 and NiFe2O4: efficient oxygen evolution catalyst," *Electrochimica Acta,* vol. 273, pp. 462-473 , 0013-4686, 2018.

[36] O. y. m. N. Avcı, L. Sementa, and A. Fortunelli, "Mechanisms of the oxygen evolution reaction on NiFe2O4 and CoFe2O4 inverse-spinel oxides," *ACS catalysis,* vol. 12, no. 15, pp. 9058-9073 , 2155-5435, 2022.

[37] K. M. Gopalakrishnan, R. Ragul, K. Vallarasu, R. Anitha, and V. Vijayalakshmi, "Exploring CoFe 2 O 4/TiO 2 NCs for high-efficiency UV-driven organic dye degradation," *Journal of Materials Chemistry C,* vol. 13, no. 42, pp. 21561-21573, 2025.

[38] M. Sahni *et al.*, "Investigating the performance of CoFe2O4/ TiO2 nanomaterial as multifunctional cathode catalyst for simultaneous dye degradation and oxygen reduction in microbial desalination cell," *Desalination,* vol. 587, pp. 117958 , 0011-9164, 2024.

[39] H. K. Channi and P. Kumar, "Spinel and Inverse Spinel Ferrites for Environmental Remediation," in *Spinel and Inverse Spinel Ferrites*: Chapman and Hall/CRC, 2025, pp. 130-149 , 1032665653.

[40] J. Tan, W. Zhang, and A.-L. Xia, "Facile synthesis of inverse spinel NiFe2O4 nanocrystals and their superparamagnetic properties," *Materials Research,* vol. 16, pp. 237-241 ,1516-1439, 2013.

Islamic University Journal of Applied Sciences (IUJAS)
https://journals.iu.edu.sa/jesc

Volume VIII, Issue I, July 2026, Pages 46-61

Mapping Research Trends in Generative Artificial Intelligence: A Bibliometric and Network Analysis

Abdulbasid S. Banga [1, *]

[1] College of Computing and Informatics, Saudi Electronic University, Riyadh 11673, Saudi Arabia, a.banga@seu.edu.sa

*Corresponding author: (Abdulbasid S. Banga), *Email Address*: a.banga@seu.edu.sa

Abstract

This study investigates the integration of user-generated social media data into Public Health Surveillance (PHS) through a dynamic, multidimensional analytical framework. Existing approaches largely rely on offline processing of static social network data, leaving a gap in methods capable of handling high-velocity, continuously evolving social streams. We propose a dynamic model that updates multidimensional representations of social data in real time using unsupervised text-mining techniques. By jointly analyzing semantic content and temporal posting patterns, the framework identifies emerging events, topics, and influential or relevant users. To enhance the utility of social data for PHS, we introduce quantitative quality measures that filter low-value or out-of-domain user profiles. The approach is evaluated on a multi-year Twitter data stream, demonstrating its effectiveness in isolating meaningful signals and excluding noisy contributors. We further outline procedures for deriving user profiles from self-descriptions to support targeted filtering. The results show that the proposed model enables robust topic and event detection, audience characterization, and impact assessment. Overall, the dynamic multidimensional framework provides a scalable and adaptable foundation for incorporating social media intelligence into PHS systems.

Keywords: Health surveillance; Social network analysis; Multidimensional analysis; Text mining.

https://doi.org/10.63070/jesc.2026.016
Received 20 November 2025; Revised 23 January 2026; Accepted 21 February 2026;
Available online 04 March 2026.
Published by Islamic University of Madinah on behalf of *Islamic University Journal of Applied Sciences*.
This is an open access article published under the Creative Commons Attribution 4.0 International license (CC BY 4.0):https://creativecommons.org/licenses/by/4.0/

1. Introduction

PHS is the process of continually gathering, analyzing, and interpreting health-related data, then promptly sharing these findings with public health practitioners, clinicians, and policymakers for use in controlling and preventing disease [1]. In the last ten years, there have been a number of studies showing how social media data can be used along with traditional surveillance methods to provide information on the prevalence of disease, emergencies, indicators of outbreaks, and how people react to official public health messages [2,3].

Historically, public health surveillance has been based primarily on clinic-based data and periodic survey data collected from providers and pharmacies. While social media is not a substitute for official health data, it can add value to the PHS process by providing timely, easily accessible, and inexpensive information that reflects the public's concerns about and behaviors towards health. The benefit of using social media in PHS, however, is not just through collecting such data it also includes the ability to filter, integrate, and analyze a large amount of different and rapidly changing data. Current methods of utilizing these types of data typically use static datasets and ad hoc analyses on each dataset to answer only specific questions. As a result, these types of analyses have limited generalizability and scalability to other public health settings [3,4]. That being the case, their methods and output aren't transferable to other scenarios. Nevertheless, PHS agencies need the creation of versatile technologies equipped with sophisticated techniques to capture, integrate, and analyze social media streams for various objectives [2].

This research is primarily aimed at providing an adaptable analytical tool to assist public health (PH) agencies with analysis of the evolution of social media postings using multiple viewpoints. The method proposed by this program will allow PH agencies to analyze who is posting, what topics are being discussed, and how information flows over time within social media streams. In addition to providing real-time monitoring of emerging patterns, warnings, and events through continuous access to live social media feeds, the system supports both online analysis and historical analysis in order to allow PH practitioners to convert massive amounts of unstructured social media data into actionable public health information.

2. Procedures

Despite only a segment of people who use social media sharing health information explicitly, the total number of social media users can provide helpful data that is in addition to traditional public health data systems, which support PHS efforts [21,54]. Social media (and other digital) posts frequently

reflect public opinion, experience, and reaction towards a given health event and are therefore useful for surveillance purposes. However, while social media data can be valuable to generate predictive and descriptive attributes of health, extracting reliable and valid health data requires systematic methods to filter out and analyze the raw data, which can be heterogeneous, noisy, and ever-evolving over time. Consequently, social media data should be considered as an additional, rather than a replacement, source of information for the established PHS services.

2.1. Receiving parties

Social Media Users There are two different types of social media users, Organizations and Individuals. Organizations use social media to represent themselves including associations, laboratories, government agencies, journals, companies, etc.. Individual Social Media Users include everyone else from health care professionals to the General Public. As a result, lay social media users typically share information related to symptoms, diagnoses, opinions, emotions, and personal experiences. On the other hand, Organizations use social media primarily to make announcements, promote campaigns/activities, disseminate updates regarding studies, perform marketing activities, etc., and provide scientific findings.

Organizational and Individual Social Media Users exhibit distinctly different engagement patterns. Whereas many health organizations are primarily concerned with heightened public awareness of disease prevention practices, research-oriented accounts are focused on encouraging the dissemination of scientific knowledge. In general, Organizational JMRI accounts tend to have much higher levels of engagement and overall credibility than Individual User accounts. From a PHS perspective, the level of health knowledge of all Users could be classified as Expert or Professional and the level of personal engagement could be classified as Patient, Relative, and/or Observer.

2.2. Publishes articles

It is possible to use the words in the postings not just to find and sort them, but also to label them with subjects that provide varying degrees of information about what they include. Medical condition, health campaign, and financial service are the wide categories, skin disease, WHO alert, and AIDS treatment are the narrower ones, but varied topics fulfill one. One can also easily detect annotation elements that refer to real-world named entities, which include: people, places, organizations, diseases, and medications. Such information items as the location of posting users, hashtags (# sign), and screen names (@ sign) should be taken into consideration when annotating postings.

2.3. Polarity of posts

Social media has become very common as people get to express themselves, whether positively or negatively. For social awareness and acceptability analysis, alarm and conflict detection, and processing post contents into polarity values, [19] is a valuable tool.

Post's timeliness. Because users update their accounts as events unfold, PHS can rely on social media for up-to-the-minute news. Health problems that manifest consistently, seasonally, or as a result of an epidemic are the most often mentioned in the postings. Researchers may be able to sort postings by time and filter them based on their temporal dimension (e.g., daily, weekly, monthly, etc.). Tracking the development of facts, identifying responses to announcements, finding fresh information, and even filtering out noisy data would all become much simpler with this. For instance, if the distribution curve of melanoma postings has a peak, it might mean that there are a lot of useless posts in response to a prominent person's diagnosis.

We propose a method wherein various features are represented by semantic annotations automatically generated by intelligent analysis of incoming postings. Additionally, massive volumes of social media data may be filtered and classified by this technique. This allows analysts to categorize the postings and provide various summaries by combining some or all of these elements.

Seasonal ads informing about melanoma prevention strategies, for instance, may pique the attention of PHS officers who want to gauge public response. The resultant collection may be huge and challenging to handle after collecting the relevant articles using keywords. In our method, these postings are tagged by organizations, general lay users, or patients based on the advanced processing. They are also tagged with views that are good or negative in tone, among other things. Next, we may find useful indicators that evaluate social responses from multiple perspectives by imposing conditions on these features and the postings' temporality.

Factors including age, sex, locality, and ethnicity are additional considerations in demographic and epidemiological research pertaining to social media users. There have been several attempts to derive this data from post contents, as lay users do not mention them in their profiles [10,17,55]. At this time, our research tools do not take these factors into consideration.

2.4. A Multi-Factor Analysis of Streams from Twitter

Our method is based on multidimensional data models that are created dynamically over streams of Twitter data [31]. Figure 1 shows the star-based schema that we propose for PHS. To facilitate analysis in accordance with PHS requirements, this schema primarily aims to provide stream-level summaries of themes and events.

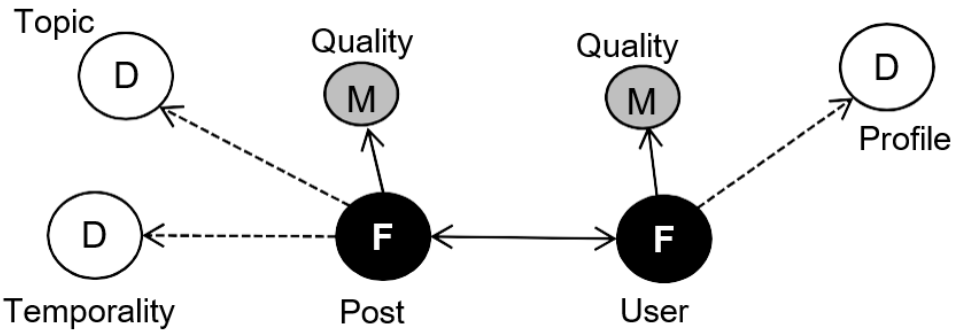


Figure 1. Analyzing Twitter streams using a multi-dimensional model: a brief overview. Dimensions are denoted by (D), facts by (F), and metrics by (M), using the standard nomenclature for multidimensional models. Dimensions are produced dynamically from the streaming data, as shown by the dotted lines.

Two distinct fact types are used to represent the social media data items. The Post Facts include data derived from the post's text, as well as its temporal aspects, subjects, and other quality indicators for tweets, such as the number of responses or retweets. From a different perspective, the User Facts are drawn from the profile data comprising the descriptions along with the other quality factors located at the user tier, e.g., the official verification status of the sending account. Keep in mind that it is possible to get certain user metrics from the sum of a user's post metrics across all of their tweets. The two-headed arrow that goes from "Post" to "User" shows the binary relationship that exists between a user's postings and the tweet's poster. Metrics like a user's total number of responses to relevant posts need this communication.

In the suggested paradigm, the values of the dimensions are determined by the incoming data, making them dynamic. One example is the Topic Dimension, which takes into consideration the primary topics discussed in tweets based on statistical inferences made after processing a certain number of tweets. In particular, in this article, we give a bigram-based temporal analysis study in order to determine these themes within the Twitter stream. Posts may also be characterized by the way the published material changes over time. A tweet may be an opinion piece, a component of a larger news story, or even part of a campaign to raise awareness about health issues. So, to correctly categorize and analyze the stream's data, we need to determine the Temporality Dimension of postings.

The quality of the posts, as well as that of the users, is a significant part of our business. The measurements of the content of the posts (achieved by processing the text part of posts) and the measurements provided by Twitter in the posts metadata (e.g., the number of responses to the tweet or the number of followers of a poster) can be combined to form parameters of the Quality Dimension.

The information's reliability in relation to the current analysis is shown by the quality metrics that are produced. As an example, a large number of people who post often could be spammers whose postings have a strong bias towards certain subjects. Because of complicated elements like metaphors, irony, and sarcasm, or because keywords are ambiguous, some communications might be little more than background noise. This latter point is well recognized in the literature due to the sarcastic usage of some medical terms in other fields, such as politics, such as cancer and metastasis. The purpose of quality metrics is to evaluate the data processing's conformity to the analysis domain.

Finally, there will be a Profile Dimension that the users can use to identify their main objectives upon engaging the social media platform. When doing social network research, user profiling is useful for gaining insight into the demographics of content creators and consumers. A dynamic differentiation between journalists, healthcare providers and services, and concerned citizens is something we'd want to see implemented for PHS. Those who have a genuine interest in the health conditions under observation are considered concerned parties (e.g., cancer survivors, patients' families). Users are profiled based on the little biographical information they supply on their Twitter profiles. In particular, we extract relevant frequent bigrams from user descriptions to identify and manually categorize a collection of individuals. We then use the identified customers to train a neural-network-based classifier to be able to predict the profiles of new users.

We set up a Twitter multi-keyword feed with some of the PHS domain themes (see Table 1) to demonstrate the effectiveness of the proposed streamed multidimensional research methodology. Reproducing a big and diverse stream of tweets for monitoring is the goal of the query.

Table 1. Terms related to the public health surveillance's Twitter feed.

Category	Details
Conditions	cancer, tumor, carcinoma, leukemia, lymphoma, metastasis, sarcoma, skin cancer, melanoma
Treatments	chemotherapy, homeopathy
Observations	UV, Cholesterol, LDL, MRI

Domain coherence and vocabulary variety are two quality measures that we create for user analysis based on the information they post. The word distribution in the tweets serves as the basis for both measures' language models. We find the domain coherence by adding the language model of the user to a distribution of UMLS Meta Thesaurus ® medical words. In our tests, we discovered that UMLS encompassed a lot of medical terms, such as the most commonly used forms in which non-medical

individuals refer to illnesses and ailments. To quantify the range of words used, we propose using the I index by Yule [56] on non-UMLS terms, too.

3. Results

We collected an extensive number of tweets in six months (June to December 2019) with the help of two methods: The Twitter API search method and keywords mentioned in Table 1. In this stream, 395,804 individuals provided a total of 777,778 tweets; 77 percent of them had made only one tweet. In regard to the overall quality of the stream, it is worth mentioning that the obtained data are very redundant; only 281,260 tweets contained original data, whereas the rest were copies of already existing tweets. Because of this, it's easy to see how some streams may favor the most popular tweets. Last but not least, remember that we analyzed user interaction by treating the retweets in the stream independently.

3.1 Users Analysis

While the majority of spam and automated accounts will have an extremely high posting frequency along with low Yule's I's (generally <30), they also tend to post content that does not relate to the health domain. By contrast, individuals who have high Yule's I scores but low domain coherence numbers also tend to be prolific posters; for instance, they will likely have screen names containing streams keyword phrases, yet their posts will have nothing to do with health. Figure 2 illustrates how the filters created using these metrics help to successfully identify and eliminate irrelevant users from consideration. The result of such filtering would eliminate around 15% of total tweets originally collected. Table 2 displays further information on the metrics of the top users.

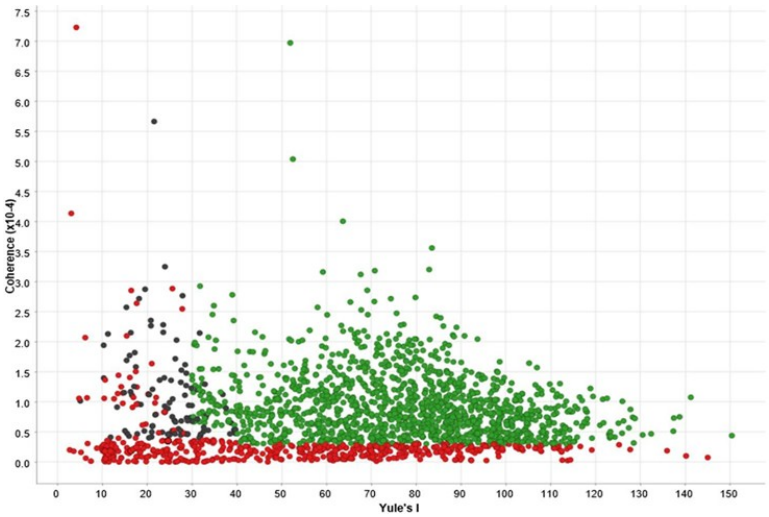


Figure 2. User quality measures according to the number of tweets published in the stream of the two thousand highest users. Positive users are represented by green points, users who need human examination are represented by black points, and users that should be filtered out are represented by red points.

Table 2. Productivity indicators for the top ten users. Stream keywords are bolded in the screen names column. Rows that are shaded represent users who are not in the domain.

Screen Name (User)	#Tweets	Yule's I	Coherence (10^{-4})	Description
Qunol Official	3306	42.3	0.88	Forum
ldl_bot	2556	94.3	0.21	Comics bot
Lymphoma Papers	2110	75.4	1.39	Academic bot
treda10	1740	70.6	1.15	Company CEO
Avos From Mexico	1623	49.9	0.33	Food sales
Eureka Mag	1194	71.2	1.87	Magazine
med vizor	1053	56.6	0.65	Forum
beechnutwx	1010	11.2	0.31	Weather bot
SnoMiddleForkWX	963	29.8	0.32	Weather bot
Sara_sarcoma	947	113.7	0.02	Influencer

3.2 Events and Topics

The identification of health-related events and subjects within the Twitter stream are performed through the application of a temporal frequency analysis on bigrams. In order to create a list of candidates from the data, three conditions must be satisfied for any given bigram:

1. kurtosis ≥ 10 thus, it will exclude all bigrams that do not show sufficient temporal variation

2. peak occurrence ≥ 20 tweets therefore, this allows for inclusion of only those bigrams that were associated with notable spikes of tweet activity

3. temporal window ≤ 10 days as such, this is used to define how long a specific bigram may remain in the Twitter stream.

Applying these thresholds led to the creation of more than 3,000 potential candidates for bigrams. These candidates have been grouped into 260 distinct health-related topics/events based on a graph modularity clustering algorithm [57].

Events account for just around 7% of all tweets, whereas subjects account for 51% of all incoming tweets. Furthermore, there is a correlation between at least one subject and 60% of event-related tweets. Therefore, we need to be mindful of event-related tweets when we analyze certain issues because of the influence they might have.

The most common occurrences identified in English postings are shown in Table 3, while those in Spanish are shown in Table 4. Most of the time, events revolve around news stories involving well-known politicians or athletes. A large number of incidents in Spanish concern youngsters whose illnesses spread to other parts of the world. The quality measures used in these two tables are Y, the average Yule, I, %Ver., the percentage of verified accounts, and OnDom., the average tweets per user

in the stream. In addition, the number of people who tweeted and retweeted about these occurrences is shown in the last column.

Table 3. Top Frequent Events (English)

Event (bigram)	#Tweets	Main Topic	User's Quality (Y, %Ver., On Dom.)	#Users Tweet./Retw.
Alex Trebek	2659	Pancreas cancer	(70, 31%, 6)	2200 / 4535
Carlos Carrasco	659	Leukemia	(72, 28%, 4)	589 / 2473
Tapeworm woman	554	False tumor	(65, 39%, 6)	437 / 332
Ross Perot	357	Leukemia	(85, 23%, 4)	335 / 1093
Bank holidays (chemo canceled)	86	Chemotherapy	(68, 0%, 5)	84 / 27

Table 4. Top Frequent Events (Spanish)

Event (bigram)	#Tweets	Main Topic	User's Quality (Y, %Ver., OnDom.)	#Users Tweet./Retw.
Tabare Vázquez	921	Lung Tumor	(57, 27%, 6)	691 / 1521
Carlos Carrasco	854	Leukemia	(55, 33%, 5)	262 / 622
Francia homeopatía	338	Homeopathy	(59, 11%, 6)	293 / 638
Luis Enrique	282	Cancer	(71, 13%, 3)	102 / 1331
fármaco evita	108	Cancer	(61, 7%, 7)	103 / 84

The number of chosen bigrams for subjects is close to four thousand, which is the same as for events. There are 139 distinct ways to classify these bigrams. Table 5 shows the most common bigrams related to various cancer kinds in English, whereas Table 6 shows the same information for Spanish. The topics containing two words, e.g., "leukemia" and "melanoma were represented by unigrams as opposed to bigrams because their semantics are associated with one word. It should be mentioned that topics involve the incidents relevant to subjects; hence, we need to be cognizant of how these events affect the topics (for instance, the "leukemia" topic). Take note that, owing to awareness initiatives linked with themes, topic-related users tend to be more active than event-related users.

Table 5. Top Frequent Topics (English Subset)

Topic	#Tweets	User's Quality (Y, %Ver., OnDom.)	#Users Tweet./Retw.	%Tweets (Replies, 1st Person)
leukemia	27,000	(76, 6%, 12)	3,200 / 17,300	(16%, 12%)
melanoma	26,791	(78, 8%, 15)	2,700 / 3,364	(14%, 7%)
skin cancer	19,000	(72, 7%, 5)	8,219 / 23,022	(23%, 12%)
brain tumor	18,900	(70, 5%, 5)	6,091 / 11,268	(37%, 28%)
lung cancer	1,500	(75, 4%, 45)	401 / 396	(18%, 8%)

According to Table 5, discussions about leukemia and melanoma made up most of the English-language content considered in this study. There was a considerable amount of public interest in these diseases, as evidenced by the many replies to and sharing of information about the topics. Skin cancer and brain tumors were also widely discussed, and many people responded to posts about those types of cancers by sharing their own experiences. Overall, it appears that users of English-language Twitter were talking about a combination of public awareness efforts and sharing their own experiences with significant cancer types.

Table 6. Top Frequent Topics (Spanish Subset)

Topic	#Tweets	User's Quality (Y, %Ver., OnDom.)	#Users Tweet./Retw.	%Tweets (Replies, 1st Person)
leucemia	12,390	(57, 8%, 4)	1,388 / 11,108	(23%, 7%)
tumor cerebral	4,451	(67, 5%, 3)	1,264 / 9,951	(23%, 16%)
melanoma	3,726	(60, 15%, 9)	208 / 421	(8%, 3%)
cáncer mama	2,800	(58, 4%, 6)	221 / 188	(17%, 12%)
cáncer páncreas	1,100	(57, 3%, 4)	259 / 242	(2%, 3%)

Table 6 shows that the top topic of discussion among Spanish speaking users are conditions related to Leukaemia, while Brain Cancer and Melanoma followed. There are fewer Verified Users discussed in Spanish when compared to English users; however, because there are more retweets in relation to Spanish Tweets, this indicates a more dynamic flow of information in Spanish based on first-hand experiences by those impacted by these conditions and their families and friends.

The first-person perspective and the response percentages for each subject (the final columns of Tables 5 and 6) are two noteworthy aspects that we examine. Politics is a valid subject for around 4% of melanoma-related tweets. By examining the most common subtopics (bigrams) linked to melanoma, this fact has been readily discovered. Compared to the overall melanoma topic's response rate of 14%, we see that 70% of the relevant tweets are responses in this instance. The ratio of persons directly impacted by the illnesses is shown by using the first person; on the other hand.

Lastly, Figure 3 provides an analysis of the subjects and events organized by user profiles, showing the writers (blue) and the audience (red). Section 3 explains that the people who submit relevant tweets are called authors, and that the people who are interested in retweeting such tweets are called the audience. Then, users may be categorized into distinct profiles based on the language models used in their descriptions. The language models in these profiles are presented in Table 7.

Table 7. Language Models for Different Profiles According to the Users' Descriptions

Profile	Definition	Top Words in Users' Descriptions
professional	Specialists	radiosurgery, oncoplastic, oncologist, haemato-oncology, consultants
ex-professional	Retired people	retired, former, senate, viet, colonel
student	Students	student, thesis, undergraduate, engineering, studying
sports	Sportsman	runner, marathon, athlete, skater, rider
religion	Religious terms	amin, allah, savior, hindu, jesus
p. services	Public services	traffic, 24h, breaking, weather, protection
h. services	Health-care services	urgencias, screenings, specialties, uninsured, visitors
concerned	Disease-concerned people	survivor, reminder_ribbon, survivorship, warrior
pseudo	Alternative therapies	reiki, meditation, yoga, ayurveda, remedial
journalist	Press and publications	medline-indexed, issn, journal, indexed, open-access
others	Other people	zombies, yin-yang, weekends, voracious, virgo (all with low probs)

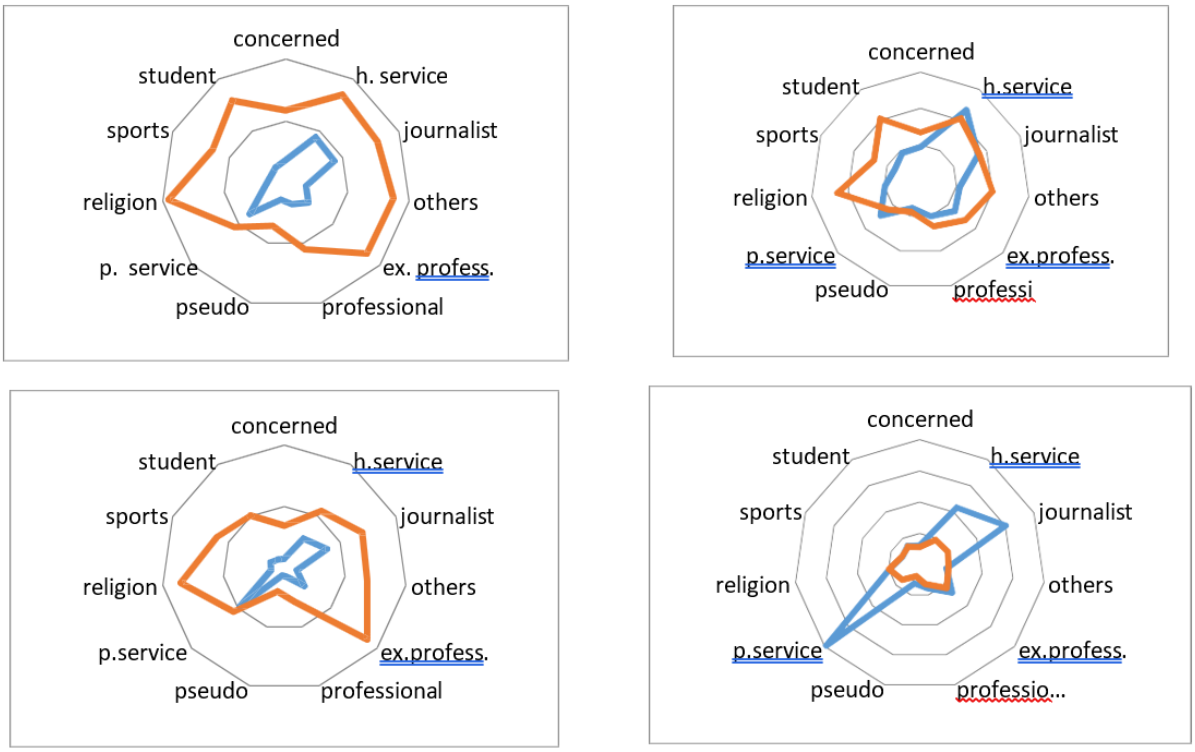


Figure 3. Using blue for writers and red for attendees, we can compare profiles from various events and subjects: (a) First, the subject of leukemia; second, the occurrence involving Carlos Carrasco. part (b) Cancer of the skin (topic) and tapeworm (event) (bottom to top).

Take note of how the profiles linked with various event types and topics change in terms of nature and influence. The postings on the "Carlos Carrasco" case and the "leukemia" issue in Figure 3 (on the left) clearly originate from journalists and the public, or health services. Except for those who employ alternative therapies (pseudo in Figure 3), it similarly depicts a huge audience that is evenly dispersed. On the other hand, the statistic shows that there are more diverse sorts of writers contributing to the "skin cancer" issue. This includes a plethora of tweets generated by public and health services, likely in the context of preventative initiatives. Finally, many writers have reported the unusual incidence of the "tapeworm" (perhaps in an effort to define it or avoid its effects), but only a tiny number of others have retweeted it.

3.3 Contentious and Open to Debate Subjects

Finally, to illustrate the application of the proposed analytical model, we have selected a group of events and subjects the peculiarities of which underlie that they may be observed in the course of PHS works. Table 8 displays a few events and themes that do not have a high number of users or an appropriate audience. You may have noticed that, except for the first subject, none of their audience members have verified accounts. Issues 1 and 2 are campaign-related topics. There are a lot of

messages in Topic 4 and Topic 3; however, the Yule's I scores are poor. For all intents and purposes, they are spam subjects; they do not provide any useful information. Lastly, an example of an unconfirmed occurrence with a miracle treatment is Topic 5.

Table 8. Controversial Events/Topics

Topic/Event	User's Quality (Y, %Ver., OnDom.)	#Users Twitting/Retw.	%Tweets (Replies, 1st Person)
1. black skin	(85, 5%, 2)	180 / 5915	(43%, 22%)
2. force chemo	(97, 0%, 5)	66 / 15	(3%, 0%)
3. *evita metástasis*	(24, 0%, 282)	109 / 185	(0%, 0%)
4. *fármaco frenar*	(25, 0%, 255)	115 / 366	(0%, 0%)
5. *desaparece oración*	(56, 0%, 4)	83 / 3	(0%, 0%)

Regarding the content of the tweets concerning these events/topics, among the hot issues are the color black, that is, the skin, which corresponds to the comment, black skin, as far as skin cancer is concerned, the responses comprise 43%. Only 3% of respondents have voiced a view on the subject of "force chemo," suggesting a lack of strong opposition to a government mandate. Even though numerous people have written on the remaining topics/events in Table 8, none of them have received any replies.

4. Discussion

To enhance the use of social media information in their multidimensional models in the decision support systems, several firms are now using cutting-edge Social Business Intelligence technology to create solutions. In this context, developing systems that generate information in real-time is a significant challenge, as is modifying the analytical models to accommodate dynamic data streams. Using this technology to build health monitoring systems is the primary contribution of this paper's work. Thus, the study presents a multi-dimensional model for analyzing Twitter streams with the purpose of intelligently using public health information and shows how it may be used for PHS tasks.

Posts are characterized by metrics and features that indicate their quality, duration, users, and content in the suggested model. After that, the incoming data stream may be filtered and summarized using these semantic annotations. With the final system in place, we can monitor PHS tasks for mentions of certain themes and events on social media. Incorporating quality measures into the multidimensional model is a strength of the suggested strategy that allows for accurate PHS analysis. Another good feature to be mentioned is the utilization of unguided statistical methods to determine themes, events, and user profiles. This allows for continuous tuning of the multidimensional model to the observed data stream. This research examines application scenarios that show how the suggested tool might be beneficial for PHS-related intelligent analysis of social data and value generation. By using this

conceptual framework, several dimensions may be derived to elucidate occurrences and themes that can be subject to on-demand process monitoring. This is because profiles and user-generated data are summarized and linked to the users' activities. The suggested technique can allow health officers to address the major applications of social media in the PHS: social monitoring, situational awareness, and the surveillance of communications.

Insights into the most noteworthy issues and occurrences connected to the monitored subjects on Twitter, with an emphasis on the kind and caliber of the individuals participating, may be gained from the examined use cases. We can do a better job rating the author’s credibility, for example, is the content from authorities, real users or bots, by looking at the author dimension and quality metrics. A suspected illness epidemic and associated occurrences may be quickly examined by grouping post metrics by often relevant subjects. Our tests, for instance, used data on first-person tweets and responses tagged with "leukemia" to find self-reported treatment experiences. However, event detection provides an opportunity to potentially estimate the present illness incidence rate. During times of crisis, the same approach might be used to identify self-reported need for humanitarian aid. By identifying the user type that generates the data and the people who consume it, as well as by tracking their responses to specific messages, we may learn about their interrelationships and track how people are aware of and perceive PHS events.

5. Conclusions

To aid PHS systems, we have introduced a new approach to multidimensional analysis of Twitter streams in this article. The proposed solution does not need a manual, and it can automatically detect themes and events within the stream according to its semantics and timeline attributes, indicating the capabilities of the solution to provide an actual summary of the contents of these objects and the people who published or interacted with the content. Our experiments illustrate the application of such multidimensional models to PHS tasks based on a real-life long-term cancer stream. Prioritizing data quality and the target audience for observed events and subjects is important to the suggested approach. This study addressed the objectives through a direct examination of the impacts of the multidimensional model. The evidence produced shows how well the model works to develop real-time identification and analysis systems for Health-related Topics, Events and User Response based on Social Media data streams.

In the future, our focus will be on establishing clear key performance indicators (KPIs) for PHS applications. For example: the time from initial detection of an outbreak until public notification is sent , how quickly the topic of public health alert spreads, and measuring the degree of the user's

influence based on their reliability, credibility, and trustworthiness. Additionally, we intend to further develop an automated method for classifying events and topics by linking them to existing medical taxonomies, such as the Unified Medical Language System (UMLS) Meta-thesaurus, as well as utilizing advanced techniques from text mining. Ultimately, the goal of these improvements is to improve the framework's interpretability, robustness, and operational applicability for real-world PHS applications and enhance the overall reliability and value of the proposed system.

Conflicts of Interest: There is no conflict of interest, the writers assert.

Ethical and Privacy Considerations

The present study has been done with only publicly available Twitter data according to the platform's terms of service and data use policies without attempting to identify or deanonymize individual users. Analyses were done using aggregate-level data in order to reduce privacy risks. The present research adheres to the data protection principles that are defined in laws, such as the General Data Protection Regulation (GDPR), which encompass responsible handling, minimizing, and ethical use of user-generated content for public health research.

Funding

The authors declare that this research received no external funding.

Data Availability Statement

Not applicable.

References

[1] B. M. Gomes, C. B. Rebelo, and L. A. De Sousa, “Public health, surveillance systems and preventive medicine in an interconnected world,” in *One Health*, 2022, pp. 33–71. doi: 10.1016/b978-0-12-822794-7.00006-x.

[2] A. Gupta and R. Katarya, “Social media based surveillance systems for healthcare using machine learning: A systematic review,” *Journal of Biomedical Informatics*, vol. 108, p. 103500, Jul. 2020, doi: 10.1016/j.jbi.2020.103500.

[3] O. Edo-Osagie, B. De La Iglesia, I. Lake, and O. Edeghere, “A scoping review of the use of Twitter for public health research,” *Computers in Biology and Medicine*, vol. 122, p. 103770, May 2020, doi: 10.1016/j.compbiomed.2020.103770.

[4] L. Zhou, D. Zhang, C. C. Yang, and Y. Wang, “Harnessing social media for health information management,” *Electronic Commerce Research and Applications*, vol. 27, pp. 139–151, Dec. 2017, doi: 10.1016/j.elerap.2017.12.003.

[5] H. Ögel-Balaban, “The use of online social network sites during the COVID-19 pandemic as a protective or risk factor for well-being of university students,” *Cyberpsychology Journal of Psychosocial Research on Cyberspace*, vol. 16, no. 3, Jun. 2022, doi: 10.5817/cp2022-3-4.

[6] Y. Wang *et al.*, "From virus to viral: content analysis of HIV-related Twitter messages among young men in the U.S.," *BMC Digital Health*, vol. 2, no. 1, Jul. 2024, doi: 10.1186/s44247-024-00102-y.
[7] O. Edo-Osagie, B. De La Iglesia, I. Lake, and O. Edeghere, "A scoping review of the use of Twitter for public health research," *Computers in Biology and Medicine*, vol. 122, p. 103770, May 2020, doi: 10.1016/j.compbiomed.2020.103770.
[8] J. Dietrich *et al.*, "Adverse Events in Twitter-Development of a Benchmark Reference Dataset: Results from IMI WEB-RADR," *Drug Safety*, vol. 43, no. 5, pp. 467–478, Jan. 2020, doi: 10.1007/s40264-020-00912-9.
[9] A. Rawat, M. A. Wani, M. ElAffendi, A. S. Imran, Z. Kastrati, and S. M. Daudpota, "Drug adverse event detection using Text-Based Convolutional Neural Networks (TextCNN) technique," *Electronics*, vol. 11, no. 20, p. 3336, Oct. 2022, doi: 10.3390/electronics11203336.
[10] "Public health surveillance for COVID-19 using Twitter sentiment analysis," *Texila International Journal of Public Health*, vol. 12, no. 3, Sep. 2024, doi: 10.21522/tijph.2013.12.03.art034.
[11] M. Myslín, S.-H. Zhu, W. Chapman, and M. Conway, "Using Twitter to examine smoking behavior and perceptions of emerging tobacco products," *Journal of Medical Internet Research*, vol. 15, no. 8, p. e174, Aug. 2013, doi: 10.2196/jmir.2534.
[12] X. Ji, S. A. Chun, Z. Wei, and J. Geller, "Twitter sentiment classification for measuring public health concerns," *Social Network Analysis and Mining*, vol. 5, no. 1, p. 13, May 2015, doi: 10.1007/s13278-015-0253-5.
[13] A. Alessa and M. Faezipour, "Flu outbreak prediction using Twitter posts Classification and linear regression with Historical Centers for Disease Control and Prevention Reports: Prediction Framework Study," *JMIR Public Health and Surveillance*, vol. 5, no. 2, p. e12383, May 2019, doi: 10.2196/12383.
[14] P. P. Morita, I. Z. Hussain, J. Kaur, M. Lotto, and Z. A. Butt, "Tweeting for health Using Real-time mining and Artificial Intelligence–Based Analytics: Design and development of a big data ecosystem for detecting and analyzing misinformation on Twitter," *Journal of Medical Internet Research*, vol. 25, p. e44356, Mar. 2023, doi: 10.2196/44356.
[15] P. Missier *et al.*, "Tracking dengue epidemics using Twitter content classification and topic modelling," in *Lecture notes in computer science*, 2016, pp. 80–92. doi: 10.1007/978-3-319-46963-8_7.
[16] K. W. Prier, M. S. Smith, C. Giraud-Carrier, and C. L. Hanson, "Identifying Health-Related topics on Twitter," in *Lecture notes in computer science*, 2011, pp. 18–25. doi: 10.1007/978-3-642-19656-0_4.
[17] W. H. Bangyal, M. Iqbal, A. Bashir, and G. Ubakanma, *Polarity Classification of Twitter Data Using Machine Learning Approach*. 2023, pp. 1–6. doi: 10.1109/hccs59561.2023.10452557.
[18] Y. Liscano, L. a. A. Arrieta, J. F. Montenegro, D. Prieto-Alvarado, and J. Ordoñez, "Early warning of infectious disease outbreaks using social media and digital data: A scoping review," *International Journal of Environmental Research and Public Health*, vol. 22, no. 7, p. 1104, Jul. 2025, doi: 10.3390/ijerph22071104.
[19] M. D. Soufi, P. R. Hachesu, and R. Ferdousi, "Oncology Informatics for Lynch Syndrome Research and Care: A Literature review," *JCO Clinical Cancer Informatics*, vol. 6, no. 6, p. e2200087, Nov. 2022, doi: 10.1200/cci.22.00087.
[20] S. Amin, M. I. Uddin, M. A. Zeb, A. A. Alarood, M. Mahmoud, and M. H. Alkinani, "Detecting information on the spread of dengue on Twitter using artificial neural networks," *Computers, Materials & Continua/Computers, Materials & Continua (Print)*, vol. 67, no. 1, pp. 1317–1332, Jan. 2021, doi: 10.32604/cmc.2021.014733.
[21] J. Sleigh, J. Amann, M. Schneider, and E. Vayena, "Qualitative analysis of visual risk communication on twitter during the Covid-19 pandemic," *BMC Public Health*, vol. 21, no. 1, p. 810, Apr. 2021, doi: 10.1186/s12889-021-10851-4.

[22] A. H. Zadeh, H. M. Zolbanin, R. Sharda, and D. Delen, "Social media for nowcasting flu activity: Spatio-Temporal Big data analysis," *Information Systems Frontiers*, vol. 21, no. 4, pp. 743–760, Jan. 2019, doi: 10.1007/s10796-018-9893-0.
[23] A. Z. Klein, A. Magge, K. O'Connor, J. I. F. Amaro, D. Weissenbacher, and G. G. Hernandez, "Toward using Twitter for tracking COVID-19: a natural language processing pipeline and exploratory data set," *Journal of Medical Internet Research*, vol. 23, no. 1, p. e25314, Jan. 2021, doi: 10.2196/25314.
[24] G. Livraga, A. Motta, and M. Viviani, *Assessing User Privacy on Social Media: The Twitter Case Study*. 2022, pp. 1–9. doi: 10.1145/3524010.3539502.
[25] Leon, "Data Warehouse Building: A comprehensive Overview," *BI4Dynamics - Business Intelligence for Microsoft Dynamics*, Jan. 06, 2025. https://www.bi4dynamics.com/data-warehouse-building-a-comprehensive-overview/
[26] E. Blomqvist, "The use of Semantic Web technologies for decision support – a survey," *Semantic Web*, vol. 5, no. 3, pp. 177–201, Jan. 2014, doi: 10.3233/sw-2012-0084.
[27] S. Moon and D. Iacobucci, "Social Media Analytics and its applications in marketing," *Foundations and Trends® in Marketing*, vol. 15, no. 4, pp. 213–292, Mar. 2022, doi: 10.1561/1700000073.
[28] E. French and B. T. McInnes, "An overview of biomedical entity linking throughout the years," *Journal of Biomedical Informatics*, vol. 137, p. 104252, Dec. 2022, doi: 10.1016/j.jbi.2022.104252.
[29] K. Mouyassir, M. Hanine, and H. Ouahmane, "Business Intelligence Model to analyze Social Media through Big Data analytics," *SHS Web of Conferences*, vol. 119, p. 07006, Jan. 2021, doi: 10.1051/shsconf/202111907006.
[30] B. Alabduallah, A. Banga, N. Iqbal, A. Ikram, and H. Diab, "Advancing cryptographic security with a new Delannoy-Derived Chaotic S-Box," *IEEE Access*, vol. 12, pp. 82926–82937, Jan. 2024, doi: 10.1109/access.2024.3410668.
[31] A. Mahajan, B. Kaushik, M. K. I. Rahmani, and A. S. Banga, "A hybrid feature selection and ensemble stacked learning models on Multi-Variant CVD datasets for effective classification," *IEEE Access*, vol. 12, pp. 87023–87038, Jan. 2024, doi: 10.1109/access.2024.3412077.
[32] E. F. Siddiqui, M. Haleem, S. F. Ahmad, A. Salhi, A. T. Zamani, and N. Varish, "A Multi-Layered AI-Driven cybersecurity architecture: integrating entropy analytics, fuzzy reasoning, game theory, and Multi-Agent reinforcement learning for adaptive threat defense," *IEEE Access*, vol. 13, pp. 170235–170257, Jan. 2025, doi: 10.1109/access.2025.3610526.

Islamic University Journal of Applied Sciences (IUJAS)
https://journals.iu.edu.sa/jesc

Volume VIII, Issue I, July 2026, Pages 62-77

Hydrodynamic Behavior of Offset Jets Under Various Situations

Mohammed Awdah B Alshehre[1], Amani Amamou[2, 3], Ali Assoudi[2], H. I. Elsaeedy[1] and Nejla Mahjoub Said[1, *]

[1]Department of Physics, College of Science, King Khalid University, Abha, 61413, Saudi Arabia
445818821@kku.edu.sa, halsayede@kku.edu.sa, nalmahjoub@kku.edu.sa

[2]LGM, National Engineering School of Monastir, University of Monastir, 5000 Monastir, Tunisia

[3]National Engineering School of Gasfa, University of Gafsa, 2112 Gafsa, Tunisia
amani.amamou@yahoo.fr

*Corresponding author: (Nejla Mahjoub Said), *Email Address*: nalmahjoub@kku.edu.sa

Abstract

In this paper, experimental and numerical results of three-dimensional turbulent offset jet flows are investigated through different situations. Velocity measurements are carried out using Laser Doppler Velocimetry (LDV) technique. Numerical simulations are performed by solving the Navier-Stokes equations with the finite volume method using the Reynolds Stress Model (RSM). The different situations of the offset jet are examined; the single offset jet (SOJ), the offset in coflow stream and the dual jet (DJ), for an offset ratio OR = 4 for the SOJ and OR = 0.5 for the DJ configuration. Flow field evolution and hydrodynamic features are presented. It is found that the behavior of the offset jet varies from one situation to another, leading to that the interaction between the offset jet and its surroundings influences the mean velocity decay and the dispersion of the jet flow.

Keywords: Offset jet; Dual jet; Turbulence; Numerical simulation.

https://doi.org/10.63070/jesc.2026.017
Received 18 November 2025; Revised 24 December 2025; Accepted 10 February 2026.
Available online 04 March 2026.
Published by Islamic University of Madinah on behalf of *Islamic University Journal of Applied Sciences*.
This is an open access article published under the Creative Commons Attribution 4.0 International license (CC BY 4.0):https://creativecommons.org/licenses/by/4.0/

1. Introduction

A jet is generated when a fluid is discharged into an ambient medium through a nozzle or orifice with an ejection velocity U_j. The majority of the flows found in nature and engineering applications are turbulent jets. Therefore, considerable researches on jet flows has been conducted in the last few decades to better understand their turbulent behavior and its impact on heat and mass transfer. An offset jet flow is formed by the ejection of a fluid from a nozzle exit into an ambient medium at a certain distance from the boundary surface. The proximity of the jet to the surface influences the dynamic behavior of the flow [1]. The structure of the offset jet flow is described by three different regions. In the first region, the entrainment of the ambient fluid creates a low-pressure zone under the jet due to the presence of the wall. This causes the jet to deviate towards the wall surface and possibly attach to it to form a recirculation zone upstream of the impingement point. The velocity of the jet decreases as it approaches the impingement surface, which allows the pressure inside the jet to increase until it reaches its maximum value at the attachment point. Beyond this point, the pressure gradually decreases along the wall, and the jet continues to develop to form a wall jet.

Many parameters may influence the behavior of an offset jet flow. In early studies [2-5], the reattachment length has usually been the primary parameter of interest to study the characteristics of an offset jet flow. Experimental measurements were made with different techniques such as hotwires, Pitot tubes, Laser Doppler Anemometry (LDA) and Particle Image Velocimetry (PIV). Agelin-chaabet al. [6] examined the effects of the offset ratio (OR) that designed the ratio between the offset height and the nozzle width (h/w), and the Reynolds number on the structure of the turbulent round offset jet using the Particle Image Velocimetry (PIV) at three different Reynolds numbers of 5000, 10000 and 20000. They found that the decay and spread rates are nearly independent of the Reynolds number and the offset height.

Recent numerical investigations [7-9] explored heat transfer enhancement in offset jet configurations using different control strategies. Ajmi et al. [7-8] and Rathore and Verma [9] conducted numerical investigations of turbulent offset jet flows to assess how the jet inclination angle, Reynolds number, and offset ratio influence fluid dynamics and heat transfer. Simulations were performed using CFD with the SST k–ω turbulence model. Results show that increasing the jet obliquity and Reynolds number enhances both local and average Nusselt numbers, while a lower offset ratio improves thermal performance, providing insights for optimizing heating and cooling jet designs. They indicated that offset jet heat transfer can be improved through either jet orientation control or surface geometry modification. Behera and Rathore [10] used CFD with a low-Reynolds-number turbulence model to examine conjugate heat transfer of a turbulent offset jet impinging on a moving flat plate, accounting

for both fluid convection and solid conduction. Results show that smaller offset ratios and higher plate-to-jet velocity ratios enhance the Nusselt number, highlighting the influence of plate motion and geometry on convective cooling performance. Mele et al. [11] analyzed flow and heat transfer in a plate with embedded discrete heaters cooled by a turbulent offset jet through steady-state convection–conduction simulations. The study highlights that material properties and heater configurations strongly influence conjugate heat transfer in plates cooled by turbulent offset jets. They observed that polymer plates experience higher heat transfer but greater thermal gradients, while ceramic plates exhibit lower and more uniform heating. In addition to RANS-based studies, more recent LES simulations of a planar offset wall-jet [12, 13] (Choudhury A. R., Anupindi; Hossen et al.) were conducted. Choudhury and Anupindi [12] provided higher-fidelity insight into turbulence dynamics, showing three peaks in the streamwise Nusselt number and highlighting enhanced heat transfer in the recirculation region, for Re = 7500 and an offset ratio OR = 5. In the case of sinusoidal pulsating offset jet, Aghaei et al. [14] numerically investigated turbulent sinusoidally pulsating offset jets, showing that low-frequency, small-amplitude oscillations enhance heat transfer and significantly influence thermal performance.

A recent experimental study by Nyantekyi-Kwakye [15] examined the flow development and turbulence characteristics of turbulent offset jets issuing from sharp-edged rectangular nozzles. Velocity and turbulence statistics were obtained using Acoustic Doppler Velocimetry (ADV), revealing the role of nozzle geometry and expansion conditions on jet decay, recirculation, large-scale vortex structures, and mixing behavior. Complementing these findings, Hakimi and Khorsandi [16] investigated turbulent round offset jets in quiescent and coflowing environments, showing that coflow modifies jet spreading, decay rates, and turbulence intensity, providing further insights into how external conditions influence offset jet flow development and entrainment. The Coandă effect is investigated in shallow-water free surface offset jets by Sollevanti et al. [17], extending classic air-jet studies to hydraulically relevant conditions. The paper experimentally shows that free-surface dynamics enhance jet attachment, yielding shorter reattachment lengths, and proposes simple scaling laws that unify jet behavior across flow regimes.

The combination between an offset jet and a wall jet is refers to as dual jet (DJ). Murphy et al. [18-20] conducted experimental studies of wall-bounded dual jets. First, they investigated [8] the effects of velocity ratio ($R = U_j/U_\infty = 0.5–2$), Reynolds number (Re = 5500–12,000) and offset ratio using infrared thermography and Particle Image Velocimetry (PIV). They found that for $R < 1$, the offset jet dominated while for $R > 1$, the wall jet constrained the offset jet and stabilized the flow. Likewise, the heat transfer is shown to be strongly influenced by R. Then, Murphy et al. [20] validated their

experimental setup by comparing a single wall jet and a single offset jet results with results from literature. By using PIV measurements, a dual jet flow field is investigated for Reynolds numbers ranging from 5500 to 12000, with an offset ratio and a velocity ratio equal to the unity. They concluded that the rising of Re leads to a slight reduction in the size of the recirculation region together with the increase of the defection of both jets. The study of Altaharwah and Hsu [21] investigated dual parallel plane jets with a fixed offset ratio. They used smoke visualization and hot-wire measurements to examine how Reynolds number ratios between the two jets and pulsation control flow interaction and mixing influence the dynamics of dual plane jets. Results show that varying the Reynolds ratio changes the dominant jet, significantly affecting lateral spreading, turbulence, and overall mixing efficiency. Kumar and Murugan [22] extended previous analyses by using PIV to capture the full velocity field and coherent flow structures of the interacting jets. Mondal et al. [23] investigated the behavior of a dual jet with and without a parallel co-flow stream, using the standard k–ω turbulence model. They focused on how the co-flow velocity (CFV) affects heat transfer when the bottom wall is kept at a constant temperature. The authors found that increasing the co-flow velocity weakens the heat transfer in dual jet systems, mainly by lowering the Nusselt number in the merging region. However, Xu et al. [24] used realizable k–ε model and transient PISO solver to investigate the energy separation phenomenon in the shear layer of two parallel free air jets under a range of Reynolds numbers. Results show that increasing Reynolds number intensifies thermal separation, while jet interaction geometry partially suppresses the effect.

Based on the over-mentioned rsearches, single offset and dual jet configurations have been extensively studied, whether experimentally or numerically. However, this paper aims to proof the use of a single offset jet with a jet density $\rho_j = 1.3$ and an offset ratio $OR = h/d_h = 4$ by ARECO group (ARECO®: Air Refreshing Control) as a specific application for nebulisation to preserve freshness and nutritional qualities of vegetables. In fact, d_h is hydraulic diameter and h is the jet offset height for SOJ and refers to the separation distance between the two parallel jet for the dual jet arrangement. The present paper investigates the behavior of a three-dimensional turbulent plane offset jet injected under different situations and combines experimental measurements and numerical simulations. Experiments are conducted using a Laser Doppler Velocimetry (LDV) technique for the SOJ case. Computations are performed with the Ansys-Fluent software using the RSM model and applied on a non-uniform grid mesh selected after a carful mesh sensibility test for both SOJ and DJ cases. The maximum mean velocity decay along the axial direction and the vertical distributions of the axial mean velocity in the plane of symmetry ($z = 0$) for different streamwise locations and for different velocity ratios are investigated.

2. Experimental and numerical set-up

2.1 Experimental set-up

Experimental measurements are conducted in a test section of 1500mm × 400mm × 400mm in the longitudinal, vertical and lateral directions, respectively (Figure 1).

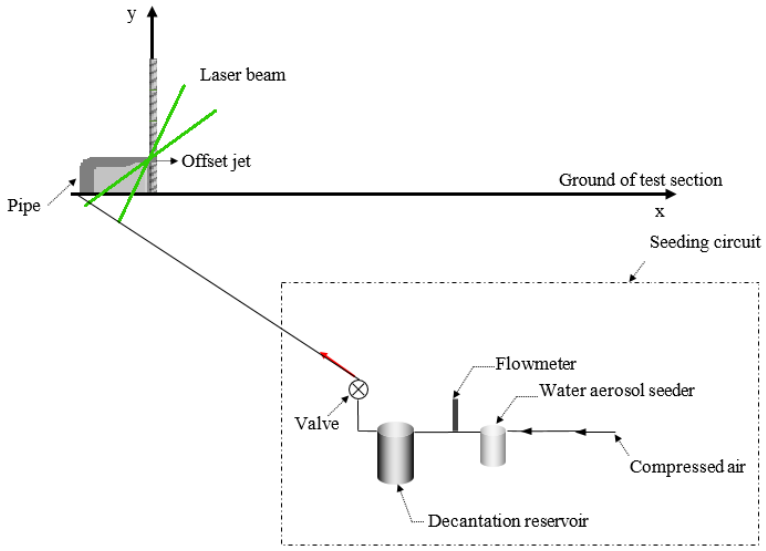


Figure 1. Skeleton view of the experimental setup

The jet flow is discharged from a rectangular nozzle (L×w) with an offset height h above the horizontal wall. Test section specifications are recapitulated in Table 1.

Table 1. Test section specifications

Parameter	Value
Nozzle dimension (mm²)	L×w = 30×6
Offset height (mm)	h = 40
Jet hydraulic diameter (mm)	d_h = 10

The Laser Tomography System (LTS), for the flow visualization, and the Laser Doppler Velocimetry (LDV) technique, for the velocity measurements, are used to investigate the flow structure. The jet is formed by an airflow seeded with small particles of water cloud of approximately 3µm diameter

(Nebulization, ARECO®: Air Refreshing Control). LDV system specifications are resumed in Table 2.

Table 2. LDV system specifications

Parameter	**Specification**
LDV type	Three-component LDV
Laser operation mode	Monochromatic laser
Scattering configuration	Optical probe volume at beam intersection
Seeding method	Nebulization
Seeding material	Water droplets (aerosol)
Seeding particle diameter	3 µm
Data processing software	FLOWSIZER

2.2 Numerical set-up

The flow is assumed steady, incompressible, turbulent and three-dimensional. The averaged equations describing the behavior of the offset jet flow is written as:

$$\frac{\partial U_i}{\partial x_i} = 0 \tag{1}$$

$$\frac{\partial}{\partial x_k}(U_k U_i) = -\frac{1}{\rho}\frac{\partial P}{\partial x_i} + \frac{\partial}{\partial x_k}\left[\nu\left(\frac{\partial U_i}{\partial x_k} + \frac{\partial U_k}{\partial x_i}\right) - \overline{u_i' u_k'}\right] \tag{2}$$

Where $i, k = 1, 2, 3$ referring to x, y and z respectively. $\overline{u_i' u_k'}$ are the Reynolds stresses terms. The closure turbulence model RSM (Reynolds Stress model) is used to solve this system of equations.

$$\underbrace{\frac{\partial}{\partial x_k}(U_k \overline{U_i' U_j'})}_{C_{ij}} = \underbrace{\frac{\partial}{\partial x_k}\nu\frac{\partial}{\partial x_k}(\overline{U_i' U_j'})}_{D_{ij}^L} - \underbrace{\left[\overline{U_i' U_k'}\frac{\partial U_j}{\partial x_k} + \overline{U_j' U_k'}\frac{\partial U_i}{\partial x_k}\right]}_{P_{ij}} + D_{ij}^T + G_{ij} + \phi_{ij} + \varepsilon_{ij} \tag{3}$$

C_{ij} denotes the convective term and D_{ij}^L, P_{ij}, D_{ij}^T, G_{ij}, φ_{ij}, ε_{ij}, represent respectively, the molecular diffusion, the stress production, the turbulent diffusion, the buoyancy production, the pressure strain and the dissipation rate [10]. The equation of the turbulent kinetic energy (k) and that of the dissipation rate of the kinetic energy (ε) which are relative to the second-order model are defined as follows [25]:

$$U_i \frac{\partial K}{\partial x_i} = \frac{\partial}{\partial x_k}\left[\left(\nu + \frac{\nu_t}{\sigma_k}\right)\frac{\partial k}{\partial x_k}\right] + \frac{1}{2}P_{ii} - \varepsilon \tag{4}$$

$$U_i \frac{\partial \varepsilon}{\partial x_i} = \frac{\partial}{\partial x_k}\left[\left(\nu + \frac{\nu_t}{\sigma_\varepsilon}\right)\frac{\partial \varepsilon}{\partial x_k}\right] + C_{\varepsilon 1}\frac{1}{2}P_{ii}\frac{\varepsilon}{k} - C_{\varepsilon 2}\frac{\varepsilon^2}{k} \quad (5)$$

where $C_{\varepsilon 1} = 1.44, C_{\varepsilon 2} = 1.92$ and $\sigma_\varepsilon = 1$.

The governing equations, along with the corresponding boundary conditions (Table 3) and imposed experimental values (Table 4), are solved using the CFD software Ansys Fluent [26] based on the Finite Volume Method (FVM). The discretization of the equations is carried out using a first-order upwind scheme. The SIMPLE algorithm is employed to ensure proper coupling between pressure and velocity fields. The resulting algebraic equations are solved iteratively through the Tri-Diagonal Matrix Algorithm (TDMA). The absolute convergence criterion was established by requiring the residuals of the continuity and momentum equations to fall below. After satisfying this requirement, the same threshold was subsequently applied to the energy equation, ensuring complete convergence of the coupled flow and thermal fields.

Table 3. Boundary conditions details

Boundaries	**Velocity**
Offset jet inlet	$u = U_j, v = 0, w = 0$
External flow	$u = U_\infty, v = 0, w = 0$
Vertical wall	$u = 0, v = 0, w = 0$
Horizontal wall	$u = 0, v = 0, w = 0$
Rest of the domain	$\frac{\partial u}{\partial n} = 0, \frac{\partial v}{\partial n} = 0, \frac{\partial w}{\partial n} = 0$

Table 4. Experimental conditions details

Parameter	**Value**
Jet exit velocity (m s^{-1})	Uj = 3.4
Jet density (kg/m^3)	ρ_j = 1.3
Reynolds number (-)	Re_j = 2455

The experimental set-up, described above, is reproduced fairly using the CFD procedure. Figure 2 illustrates a mesh sensitivity study based on the maximum velocity decay (U_m/U_j) for a SOJ. Several computational grids are evaluated to examine the influence of mesh resolution on the numerical results confronted to present experimental data for OR = 4. Four representative grids are reported in this study

varying from coarse to very fine meshes. As shown in Figure 2, the normalized maximum velocity decay obtained with Mesh 3 and Mesh 4 closely matches the experimental data. Therefore, Mesh 3 is considered sufficiently refined to yield grid-independent predicted results. After this careful grid independence investigation, the picked out grid is 1.3×10^6 cells with details reported in Table 5.

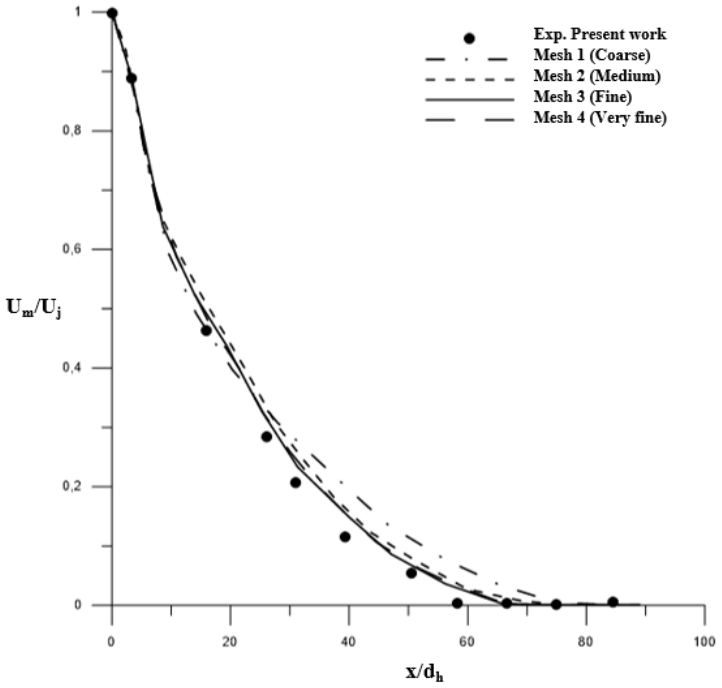


Figure 2. Grid independence test

Table 5. Grid details

n_x	n_y	n_z	y^+	Total cells
180	90	80	$30 < y^+ < 100$	$1.3\ 10^{-6}$

3. Results and discussion

3.1 Single offset jet

The evolution of a single offset jet (SOJ) with a density of $\rho_j = 1.3$ is presented in Figure 3 for OR = 4. In the symmetry plane (z = 0), the jet emerges and diffuses through the ambient surrounding immediately downstream of the nozzle exit. The jet is deflected and rolled up, causing the formation of shear layer vortices (SLV). In the near field, the offset height seems to influence the development

of the jet indicated by the breakdown of the flow after a short axial distance downstream of the nozzle exit. This breakdown leads to the attachment of the jet to the bottom wall, located in the impingement point.

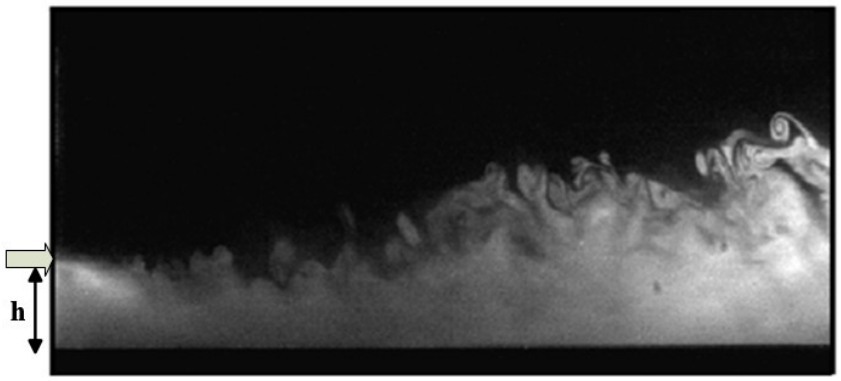


Figure 3. Flow visualization in the symmetry plane for OR = 4

The variation of the maximum velocity normalized by the jet velocity (U_m/U_j) is plotted along the normalized axial direction (x/d_h) in Figure 4. Experimental data from Agelin-Chaab and Tachie [6] are considered for comparison together with present LDV experimental measurements. It is obvious that the two curves follows the same decay beginning with a higher initial decrease in the near jet exit region. A small discrepancy is observed between the present experimental measurements and those reported by Agelin-Chaab and Tachie [6]. This difference may be attributed to the use of different jet fluids: a water jet in [6] and an air jet in the present experiments. Predicted results fairly follow both experimental data. Further downstream, the effect of the offset height is notable, when the offset height is smaller, the maximum velocity decays rapidly.

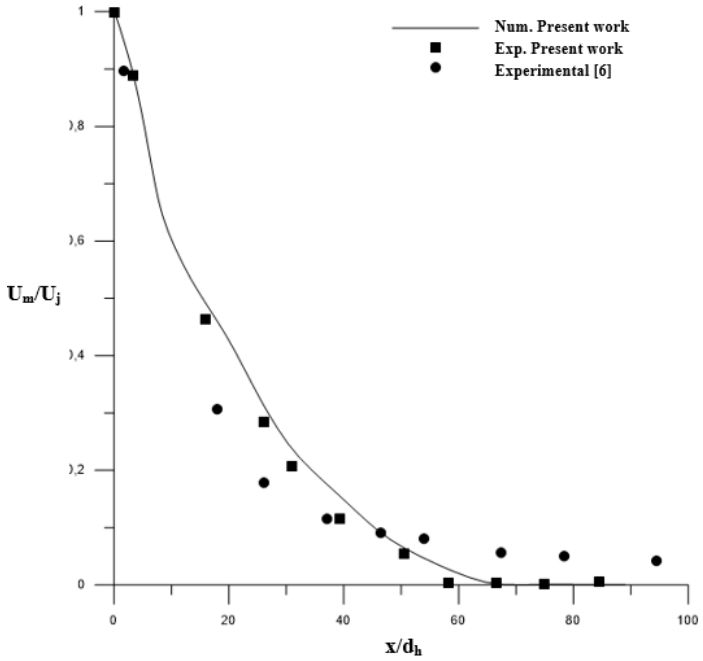


Figure 4. Maximum mean velocity decay

3.2 Offset jet into coflow stream

The air jet flow is discharged from a vertical wall offset by a height h above the horizontal wall with an ejection velocity U_j into an external flow, in coflow direction, stagnant or with a uniform velocity U_∞. The vertical distributions of the axial mean velocity component in the plane of symmetry (z = 0) for different streamwise locations ($x/d_h = 4$, $x/d_h = 7$ and $x/d_h = 10$) and for different velocity ratios (R = 1, R< 1, R > 1) are shown in Figure 5.

It is clear that, downstream of the jet nozzle exit, the profiles of the normalized axial velocity (U/U_m) exhibit negative values near the horizontal wall for the three velocity ratios pointing out the presence of a recirculation zone. This considered region is marked by the presence of vortices in contrast with the case where the offset jet is emerged in quiescent ambient; $R = \infty$ [27]. Moving vertically away from the nozzle exit, the mean axial velocity increases with the normalized vertical coordinate (y/d_h) to reach the uniform velocity of the external fluid leading to the maximum value $U/U_m = 1$, far from the interaction zone. Besides the presence of an external flow stream in coflow direction influences the parameters of the offset jet (the reattachment length and the attachment point), it is found that increasing the coflow velocity (i.e. the velocity ratio decreases) leads to the decrease of the mean axial velocity.

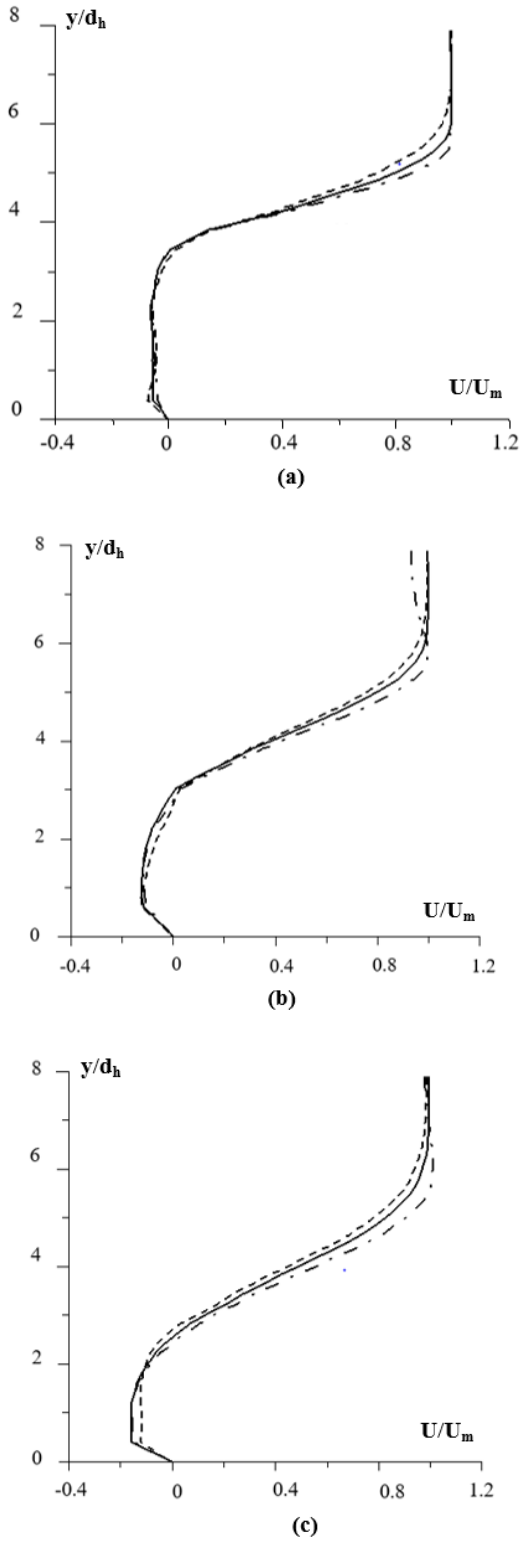


Figure 5. Axial mean velocity profiles for different velocity ratios at various streamwise positions (a) $x/d_h = 4$, (b) $x/d_h = 7$, (c) $x/d_h = 10$

3.3 Dual jet

The evolution of a single offset jet (SOJ) and a dual jet (DJ) at an OR = 0.5 is illustrated through the streamlines of axial velocity in the Figure 6 (a) and Figure 6 (b), respectively. The most remarkable observation is the pronounced recirculation region below the jet in the case of a single offset jet and between the two nozzles of two parallel plane jets. This leads to determine the reattachment point (rp) for a SOJ, defined as the first point on the impingement wall where the fluid attaches, and the merging point (mp) for a DJ at which the two jets deflect to each other and merge together. There two characteristic points are presented in Figure 6 (a) and Figure 6 (b), respectively. It is found that the merging point is located at $x_{mp}/d_h = 9.3$ and the reattachment point is situated at $x_{rp}/d_h = 11.5$.

The DJ flow is characterized by two vortices rotating in opposite directions inside the recirculation zone, opposite the SOJ flow, where only one vortex is formed in the recirculation zone between the nozzle and the impingement surface. Furthermore, it is observed that the SOJ's recirculation region is larger than that of DJ, resulting that the locus of the reattachment point is far from the merging point.

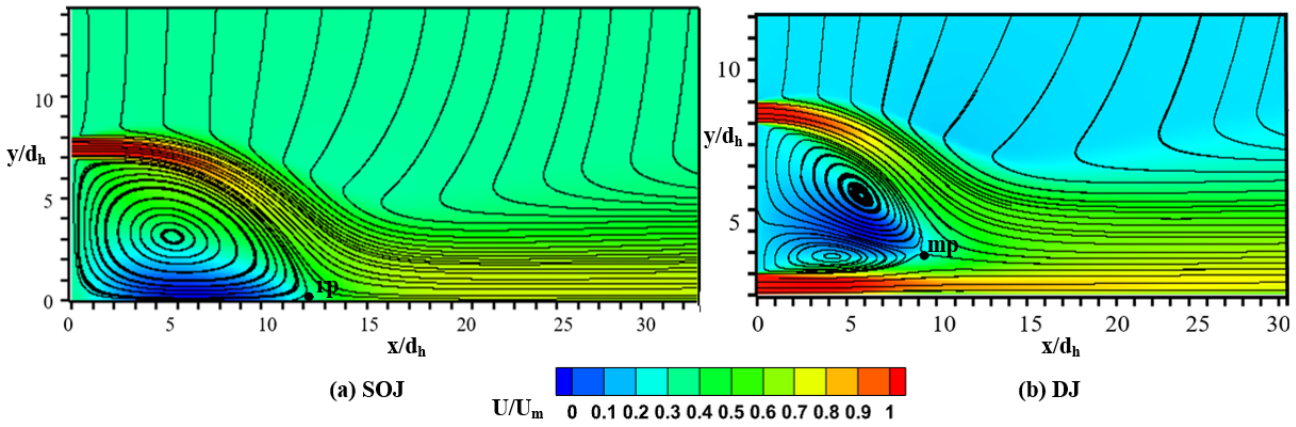


Figure 6. Streamlines of axial velocity of (a) a single offset jet and (b) a dual jet

Figure 7 exhibits the predicted maximum velocity decays of both single offset jet (SOJ) and dual jet (DJ) confronted to experimental data of Wang and Tan [28]. The normalized maximum velocity remains constant for few hydraulic diameters downstream of the nozzle exit and then decreases gradually whilst moving away from the initial jet region. It is worth noting that the maximum velocity in the case of DJ decays more rapid than that of SOJ in the converging zone (the near flow field) due to the strong interaction between the two parallel jets. These indicates that the SOJ potential core is larger than that of DJ. Downstream the initial region, the both jet configurations show similar decay rate except that the maximum velocity in the case of SOJ decreases faster than that in the case of DJ.

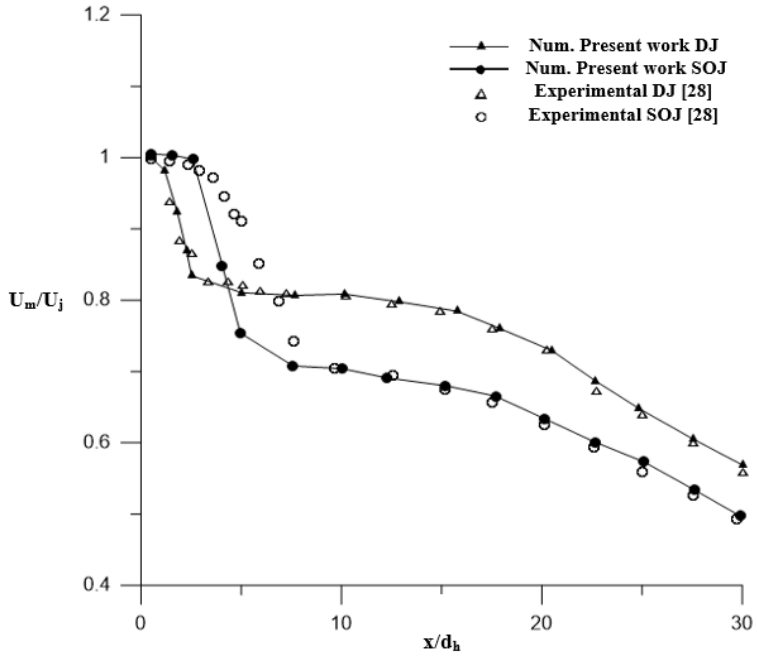


Figure 7. Normalized velocity decay of a dual jet compared to that of a single offset jet for OR = 0.5

4. Conclusion

An experimental and computational study has been carried out to investigate the hydrodynamic behavior of the offset jet configuration under various situations; a single offset jet, an offset jet into coflow stream and a dual jet. Experimental data are measured with Laser Doppler Velocimetry (LDV) technique for the single offset jet and numerical results are predicted using the CFD software Ansys-Fluent for both single offset jet and dual jet configurations. The study addresses an industrially relevant topic by investigating the application of an offset jet within the ARECO nebulization system as a strategy to reduce food waste.

The comparison between experimental and predicted results as well as data from literature exhibits good agreement. It is clear that the offset height has a notable influence on the maximum velocity decay showing that smaller offset height leads to rapid maximum velocity decay. In the presence of a coflow stream, increasing the coflow velocity (i.e., decreasing the velocity ratio) leads to a reduction in the mean axial velocity. Numerical results illustrate the formation of a recirculation zone beneath the jet for the single offset jet and between the two nozzles for the dual parallel plane jets, with the recirculation region being larger in the single offset jet configuration. Furthermore, the potential core of the single offset jet is longer than that of the dual jet, resulting in a more rapid decay of the maximum velocity for the dual jet in the near-field region.

Funding

The authors declare that this research received no external funding.

Data Availability Statement

Not applicable.

Conflicts of Interest

There is no conflict of interest, the writers assert.

References

[1] A. Assoudi, A. Amamou, N.M. Said, H. Bournot, Characteristics and analysis of a turbulent offset jet including the effect of density and offset height, Int. J. Mech. Sci. 174 (2020) 105477.

[2] D.I. McRee, H.L. Moses, The effect of aspect ratio and offset ratio on nozzle flow and jet reattachment, in: Advances in fluids, F.T. Brown (Eds.), ASME Press, New York, 1967, pp. 142-161.

[3] N. Rajaratnam, K. Subramanya, Plane turbulent reattached wall jets, J. Hydraul. Div. 94 (1968) 95-112.

[4] J. Hoch, L.M. Jiji, Theoretical and experimental temperature distribution in two-dimensional turbulent jet boundary interaction, J. Heat Transf. 103 (1981) 331-335.

[5] A. Nasr, J. Lai, A turbulent plane offset jet with small offset ratio. Exp. Fluid. 24 (1998) 47–57.

[6] M. Agelin-Chaab, M.F. Tachie Characteristics and structure of turbulent 3D offset jets, Int. J. Heat Fluid Flow. 32:3https://www.sciencedirect.com/journal/international-journal-of-heat-and-fluid-flow/vol/32/issue/3 (2011) 608-620.

[7] M. Ajmi, N. Hnaien, S. Marzouk, L. Kolsi, K. Ghachem, H. Ben Aissia, M.A. Almeshaal, Numerical investigation of heat transfer enhancement of an inclined heated offset jet. Int. Commun. Heat Mass Transf. 116 (2020) 104682.

[8] M. Ajmi, N. Hnaien, S. Marzouk, L. Kolsi, K. Ghachem, H. Ben Aissia, Numerical investigation and triple-parameters correlations development on the dynamic characteristics of a turbulent offset jet, J. Turbul. 22:6 (2021) 325-352.

[9] S.S. Rathore, S.K. Verma, Numerical investigation on the efficacy of jet obliquity for fluid flow and thermal characteristics of turbulent offset jet, Heat Mass Transf. 58 (2022) 1223–1246.

[10] V.M. Behera, S.K. Rathore, The effect of plate motion on heat transfer enhancement using turbulent offset jet flow: A conjugate approach, Int. Commun. Heat Mass Transf. 136 (2022) 106173.

[11] B. Mele, G. Ruocco, Conjugate heat transfer from a discrete-source heated plate by offset jet impingement, Int. J. Heat Mass Transf. 217 (2023) 124651.

[12] A.R. Choudhury, K. Anupindi, LES of a planar offset wall jet with heat transfer-characterization, turbulent kinetic energy and Reynolds shear stress budgets, Int. J. Heat Mass Transf. 191 (2022) 122847.

[13] M.K. Hossen, E. Asgari, M. Saeedi, B. Nyantekyi-Kwakye, Large-eddy simulation of an offset jet flow over surface-mounted ribs. Proc. CSME Congress, Canada, 2025.

[14] A. Aghaei, S. Jahangiri, A.I. Khdair, R.S. Faisal, M. Afrand, Investigation of turbulent heat transfer in sinusoidal pulsating offset jets: a numerical study, Int. J. Therm. Sci. 220 (2026) 110427.

[15] B. Nyantekyi-Kwakye, Large-scale motion of submerged offset jets issuing from sharp-edged rectangular nozzles, Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci. 235:1 (2020) 78-93.

[16] F. Hakimi, B. Khorsandi, Dynamics of round turbulent offset jets in quiescent and coflowing environments, Phys. Fluid. 36 (2024) 105156.

[17] M. Sollevanti, M. Miozzi, G.P. Romano, Coandă effect in free-surface shallow-water offset jets. Exp. Fluids. 66 (2025) 57.

[18] P.J. Murphy, S. Alimohammadi, S.M. O'Shaughnessy An experimental investigation of the effect of the velocity ratio on the flow and wall heat transfer characteristics of a wall-bounded dual jet. Therm. Sci. Eng. Progress. 66 (2025) 104007.

[19] P.J. Murphy, S. Alimohammadi, S.M. O'Shaughnessy, Experimental investigation of the effect of the offset ratio on the flow characteristics and heat transfer behaviour of a wall-bounded dual jet flow, Int. J. Thermofluid. 26 (2025) 101125.

[20] P.J. Murphy, S. Alimohammadi, S.M. O'Shaughnessy, Experimental investigation of a wall-bounded dual jet flow for varying Reynolds number: Flow visualisation, hydrodynamic characteristics, and associated heat transfer, Int. J. Therm. Sci. 211 (2025) 109699.

[21] Y.A. Altaharwah, C.M. Hsu, Flow and mixing characteristics of single-pulsed dual parallel plane jets. Results in Engineering, 25 (2025) 103741.

[22] S. Kumar, S. Murugan, Investigation of flow characteristics and velocity fields of excited two parallel plane jets, Phys. Fluids. 36 (2024) 015142.

[23] T. Mondal, N. Hnaien, M. Ajmi, K. Ghachem, L. Kolsi, CFD Investigation of Thermal Characteristics for a Dual Jet with a Parallel Co-flow, ACS Omega 7 (2022) 27864−27875.

[24] W. Xu, R. Dong, S. Liang, Numerical simulation study of energy separation in jet shear layer, Process. 12:2 (2024) 331.

[25] A. Amamou, A. Elkaroui, R. Ben Khalifa, N.M. Said, H.I. El-Saeedy, Computational investigation of highly heated homogeneous subsonic turbulent jets issuing from plane and circular nozzles, Int. J. Therm. Sci. 187 (2023) 108163.

[26] ANSYS Fluent Theory Guide, ANSYS, Inc., 275 Technology Drive Canonsburg, PA 15317, July 2021.

[27] A. Assoudi, S. Habli, S.N. Mahjoub, H. Bournot, G. Le Palec, Experimental and numerical study of an offset jet with different velocity and offset ratios, Eng. Appl. Comput. Fluid. Mech. 9:1 (2015) 490-512.

[28] X.K. Wang, S.K. Tan, Experimental investigation of the interaction between a plane wall jet and a parallel offset jet, Exp. Fluids. 42:4 (2007) 551–562.

Islamic University Journal of Applied Sciences (IUJAS)
https://journals.iu.edu.sa/jesc

Volume VIII, Issue I, July 2026, Pages 78-109

Enhancing Cybersecurity in IoT-Based Maternal Health Monitoring Systems Using Machine Learning Algorithms

Abdulbasid S. Banga [1] *

[1] College of Computing and Informatics, Saudi Electronic University, Riyadh 11673, Saudi Arabia, a.banga@seu.edu.sa

*Corresponding author: (Abdulbasid S. Banga), Email Address: a.banga@seu.edu.sa

Abstract

The rapid expansion of IoT devices in maternal health monitoring enables continuous data collection and improved clinical assessment; however, it also introduces significant security and privacy concerns due to the sensitivity of maternal health information. This study investigates how artificial intelligence (AI) and machine learning (ML) can enhance both analytical performance and data protection in IoT-based maternal monitoring systems. The proposed framework employs Random Forest, Decision Tree, Support Vector Machine, and a stacking–bagging ensemble to improve maternal risk prediction and anomaly detection. Privacy-preserving techniques are integrated to secure physiological parameters: homomorphic encryption ensures data confidentiality during processing, while differential privacy limits information leakage from model outputs. Experimental results show that the stacking classifier combined with Random Forest achieved the highest accuracy of 82.3%, demonstrating greater robustness than traditional algorithms. Although differential privacy strengthened data protection, it reduced precision and F1-score, highlighting a trade-off between privacy and accuracy. Overall, integrating ensemble learning with privacy-preserving methods improves the security, accuracy, and reliability of IoT-driven maternal health monitoring systems.

Keywords: Maternal Health; Anomaly Detection; Homomorphic Encryption; Ensemble Learning.

https://doi.org/10.63070/jesc.2026.018
Received 21 November 2025; Revised 13 February 2026; Accepted 16 February 2026.
Available online 04 March 2026.
Published by Islamic University of Madinah on behalf of *Islamic University Journal of Applied Sciences*.
This is an open access article published under the Creative Commons Attribution 4.0 International license (CC BY 4.0):https://creativecommons.org/licenses/by/4.0/

1. Introduction

The adoption of the Internet of Things (IoT) in maternal health surveillance has transformed personal healthcare by enabling real-time evaluation and continuous monitoring of physiological conditions. These technologies support improved maternal care through predictive analytics and individualized treatment planning. However, the rapid growth of data collection introduces substantial cybersecurity risks, particularly due to the sensitivity of maternal medical records, making strong protective measures essential for maintaining patient trust and safeguarding clinical outcomes. Recent research increasingly focuses on applying modern machine learning techniques to enhance the safety and performance of IoT-driven maternal healthcare solutions. This study aims to develop an intelligent data framework that integrates advanced encryption techniques to secure maternal medical records while ensuring accurate fetal and maternal risk predictions. Secure IoT systems enhanced with AI analytics have the potential to transform maternity healthcare operations. These measures improve clinical operations while simultaneously strengthening data protection.

The health status of a pregnant woman defends both the fetus and the mother and triggers health impacts that last throughout pregnancy to the postnatal years. It has been scientifically proven that hypertension, along with gestational diabetes in pregnancy, results in severe health issues that are acute maternal complications that manifest in adulthood [1]. Comprehensive maternal care provides essential benefits by preventing emergency complications and supporting long-term health outcomes for individuals and communities. In 2015, newborns accounted for half of all under-five child deaths, with preterm birth complications and intrapartum infections such as sepsis and meningitis each contributing 15% of these deaths [2]. According to UNICEF (2018), Pakistan has one of the highest neonatal mortality rates globally, with 46 deaths per 1,000 live births [3]. Prenatal monitoring is therefore fundamental to preventing avoidable risks and adverse outcomes, including long-term harm and mortality [4,5]. The advances in the technologies of artificial intelligence enable medical workers to prepare individual treatment plans and conduct efficient diagnostic forecasts in maternal health care. The current medical models and convolutional algorithms used in modern artificial intelligence systems are used to develop accurate medical tests that can give a prediction of illnesses in patients by analyzing different medical inputs [6]. When incorporating IoT maternal health monitoring into artificial intelligence technology, data privacy protection frameworks should be enforced to address the possible cybersecurity threats that are introduced in the process of exchanging them in favor of the use of higher diagnostic accuracy.

Research on maternal health has accelerated in recent years with the integration of IoT, Machine Learning, and Cyber-Physical Systems into healthcare environments. A wide range of studies has examined maternal healthcare risks and the technological methods used to address them. The study of [7] involved the use of IoT systems to identify the risks of maternal healthcare in far geographical regions. As the research indicates, the Internet of Things systems can offer much-needed healthcare services to the developing population centers and improve the medical tracking capabilities and performance monitoring [7]. In their research, [8] used ML technology to achieve correct maternal health risk identification. Research studies established that ML methods were able to identify significant risk factors that yielded further opportunities for early health care treatment [8]. Medical CPS implementations benefit the healthcare industry, as [9] explains the structure of healthcare systems of the future in the context of secure healthcare systems [9]. As demonstrated in [10], smart technologies play a significant role in strengthening public health systems by supporting the assessment of maternal risk and improving maternity-care outcomes. Building on this, [11] focused on evaluating maternal pregnancy health using ensemble learning–based feature engineering, highlighting the potential of advanced machine-learning techniques to enhance predictive accuracy in maternal health monitoring.

Wearable technology is an indispensable part of systems that are deployed to ensure the health of mothers is tracked. [12] examined wireless sensor networks and wearable devices in relation to their use in the e-health system in monitoring the health of the mother [12]. [13] used data mining techniques to identify preterm birth risk factors and demonstrated how analytic techniques create healthcare solutions to preterm birth prevention [13]. The use of blockchain systems along with UAVs contributes to the rise in the provision of healthcare services. [14] introduced VAHAK as a new system that integrates UAVs with blockchain to create services of Healthcare 4.0 provision in the field of maternal healthcare logistics optimization [14]. [15] assessed the technological attributes and the use of the technologies of Medical 4.0 in the healthcare system. [16] state that IoT applications introduce privacy risks to the maternal services by investigating the data protection requirements of maternity healthcare IoT systems [10]. In their effort to strengthen the security of healthcare applications, [17] examined the use of ensemble machine-learning techniques for malware classification [16]. Effective communication across healthcare systems depends on the seamless exchange of medical data, which is essential for delivering efficient and coordinated care [18]. Related work has demonstrated how medical devices interact to achieve patient-data interoperability, enabling continuous maternal health monitoring [17]. Furthermore, [19] emphasized that big data analytics holds significant potential for advancing medical learning and supporting functional decision-making in healthcare environments

[18]. The study reviewed the scientific contributions on the topic of automated solutions between healthcare 4.0 and biomedical engineering and health informatics to transform maternal healthcare [19]. The combined results of these studies show that the use of technological synergy can change the practices of medical care in prenatal care to develop research in the future. The developments in maternal healthcare via IoT and ML applications have been achieved, yet further efforts are required to integrate security practices with high-performance models. Maternal health data must be secured at all costs, as information leakages would cause massive ethical and legal consequences. Machine Learning for Security Architectures proposes a zero-trust framework that leverages machine learning to address cybersecurity threats in healthcare IoT systems [20]. The approach employs convolution-based machine learning techniques to detect, prevent, and mitigate potential threats, thereby enhancing data privacy and integrity. The zero-trust model operates based on the principle of enhancing security by authenticating every access request at all times. IoT Healthcare Intrusion Detection Systems (IDS) have now become a significant component of IoT cybersecurity. A hybrid IDS with the Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) is very accurate in detecting the presence of malicious entities in IoT networks [21]. The system was evaluated using the CICIoT2023 dataset and achieved an accuracy of 98.42%, demonstrating strong potential for real-world applications. IoT Healthcare Authentication and Risk Mitigation. The optimization of ML to the security aspects of healthcare IoT is a well-surveyed area, and its focus has shifted to the authentication procedure and methods of risk elimination [22]. The current study highlights the importance of robust authentication protocols that rely on machine learning and deep learning algorithms to prevent unauthorized access and mitigate potential cyber threats. Ensuring effective anomaly detection through ensemble learning has also been emphasized. To strengthen IoT cybersecurity, an ensemble-based anomaly detection approach has been applied. Additionally, recent work has utilized Bayesian hyperparameter optimization to develop an adaptive framework capable of identifying irregularities across diverse IoT networks [23]. The model demonstrated a great predictive capacity with a heterogeneous dataset, and the tool will be effective when IoT-based maternal health monitoring systems are involved. Real-Time Monitoring of the Patient and Data Security. The offered intelligent IoT system, with the implementation of ML algorithms to conduct a real-time monitoring of the patient, can enhance data security [24]. This system not only ensures that sensitive information is accessible only to authorized personnel but also prevents any tampering with patient data—an essential requirement when handling highly confidential maternal health records. Within this context, Healthcare Artificial Intelligence (AI) and AI-driven cybersecurity for IoT have been widely discussed as sustainable solutions for securing interconnected medical environments. Prior research has suggested that AI-powered models can effectively detect cyber threats and strengthen the security of

IoT-based maternal healthcare monitoring systems [25]. However, the proposed architectures still faced challenges related to long-term sustainability and minimizing exposure to emerging cyber risks. To detect anomalies within the hospital IoT Ecosystems, it uses Neural Networks. The hybrid model LSTM-CNN has been used in one of the recent studies that sought to find anomalies in the hospital IoT networks and revealed high accuracy in detecting malicious activities [26]. This type of model implementation enhances the cybersecurity of hospitals, which in turn facilitates the maintenance of maternal health monitoring systems because of a secure and stable environment to transmit patient data. The reviewed literature demonstrates that the role of ML-based cybersecurity systems in the IoT healthcare is growing. Beginning with the zero-trust architecture to advanced IDS and anomaly detection models, the ML-based solutions have significantly improved security structures. Though the issue of IoT-enabled maternal healthcare, machine-learning prediction models, and cybersecurity frameworks are well-elaborated, the problem statement provided cannot define what should be achieved in the current study. The literature highlights several concerns, including the rising risks associated with pregnancy, the vulnerability of pregnant women's personal information, and the lack of secure AI-based forecasting systems. However, the specific aims of the referenced research are not clearly defined. Establishing well-articulated objectives is essential for outlining the study's scope, guiding the researcher's direction, and demonstrating how the proposed secure, AI-driven maternal mental-health prediction framework addresses the identified gaps. The lack of these objectives forms a lack of clarity and disrupts the relevance of the research problem, methodology, and suggested architecture. These methodologies applied directly to maternal health monitoring should become the topic of further studies, as the safety of the data and patients will be secured. Despite the present literature pointing to the potential existence of maternal mental-health risks in the course of pregnancy, including postpartum depression, anxiety, and behavioral alterations, the existing literature seldom discusses the security and privacy risks of gathering and processing this sensitive data. The majority of AI-based maternal health models have been built on the accuracy of prediction, but without secure data-handling systems. This gap highlights a critical challenge: how to develop an AI-based maternal mental-health prediction system that is both clinically reliable and cyber-secure. The present study addresses this issue by integrating state-of-the-art machine-learning models with robust privacy-preservation mechanisms, including Homomorphic Encryption and Differential Privacy.

The study will fulfill its objectives by having three broad objectives.

I. The implementation of homomorphic encryption will ensure the protection of sensitive maternal health records and provide access features that can be used. This is how secure information processing operations may be implemented in the maternal healthcare privacy systems.

II. The study examines the performance of various machine learning algorithms in environments where data remains encrypted, providing a comprehensive analysis of their behavior under such conditions. Among the evaluated models, a Random Forest combined with Gradient Boosting in a series configuration demonstrates the most optimal performance. Furthermore, the use of advanced data-protection configuration techniques enhances overall system performance while simultaneously strengthening security.

III. The evaluation procedure has to conduct studies on the performance of traditional ML algorithms against ensemble learning methods in terms of joint capacity to be data secure and predictive.

The study aims to develop a comprehensive approach for enhancing the security and predictive capabilities of IoT-based maternal health monitoring systems, thereby improving the overall safety and efficiency of maternal healthcare solutions.

2. Methodology

2.1 Procedures

When pregnant mothers experience physical or emotional pregnancy complications, the health of both mother and the unborn is placed in a very high-risk state. Up-to-date data science solutions allow machine learning models to incorporate effective predictive control and remedies for the prevention of health threats. The study uses data obtained from Kaggle, where an Internet of Things (IoT)–based monitoring system recorded maternal health information across various medical settings, including hospitals, community clinics, and dedicated maternal healthcare facilities. The system uses vital health functions that are measured by workers in order to identify the possible risks of pregnancy. The predictive features present in the dataset are applied in maternal health risk assessment, and they consist of Age and SystolicBP, DiastolicBP, BP measurements, and the variables BS, Heart Rate, and Body Temperature. Risk analysis of patient groups as being low, mid, or high-risk happens by means of collecting these physiological signs and utilizing them to determine the Risk Level variable.

Table 1. A representative sample of the dataset is as follows

Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	Risk Level
25	130	80	15	98	86	High Risk
35	140	90	13	98	70	High Risk
29	90	70	8	100	80	High Risk
30	140	85	7	98	70	High Risk
35	120	60	6.1	98	76	Low Risk

23	140	80	7.01	98	70	High Risk
23	130	70	7.01	98	78	Mid Risk
35	85	60	11	102	86	High Risk
32	120	90	6.9	98	70	Mid Risk
42	130	80	18	98	70	High Risk
23	90	60	7.01	98	76	Low Risk

A sample of 1,500 records from the larger database was used as the dataset shown in Table 1. The study includes participants of varying ages and healthcare backgrounds, making the dataset a valuable research resource for maternal health risk analysis. The data follows the key physiological parameters, which are blood pressure and the level of blood sugar, and incorporates body temperature and heart rate parameters, which play crucial roles in the health assessment of the patient during pregnancy risk identification. The dataset includes both periodic check-up records and continuous monitoring data, thus making a full picture of the maternal health check-up. This database has exemplary value since the variety of health measurements is detailed to allow predictive models to be developed to anticipate situations of high risk with a patient.

The researchers employed several approaches to ensure the validity, reliability, and security of their findings. The key components of the methodology are outlined below.

2.1.1 Dataset Overview

- The dataset consists of 1,500 records containing measurements for Age, Systolic Blood Pressure, Diastolic Blood Pressure, Blood Sugar (BS), Heart Rate, and Body Temperature. These assessment features were selected because of their proven effectiveness in evaluating maternal health risks.
- The Risk Level variable is the target of the machine learning models and is split into three risk levels, among them being Low Risk, High Risk, and Mid Risk, using health parameter analysis.

2.1.2 Data Preprocessing

- The appropriate data imputation techniques, such as mean or median techniques to fill the gaps in the data before the execution of the analysis, were appropriate since not having full records would lead to bias.
- Evaluation of outliers used z-score normalization as the detection method. Standardization through this technique protects data results from extreme values while you conduct health research, which must prevent data skewing.

- The predictive models were fed with the standardized feature values either using min-max scaling or z-score normalization to prevent the occurrence of an individual feature controlling the model outputs due to its magnitude. The dataset showed a clear class imbalance, with High-Risk cases occurring more frequently than the Mid- and Low-Risk categories. To prevent biased learning, the models were trained using class-weighted loss functions and stratified 5-fold cross-validation, ensuring each fold preserved the original class proportions. SMOTE oversampling was also tested, but class weights provided more stable results. Macro-F1 and weighted metrics were used because they better represent performance across imbalanced classes. Privacy-preserving models, especially those using differential privacy, showed increased sensitivity to imbalance, as added noise disproportionately affected minority-class predictions

2.1.3 Feature Selection

The primary feature-selection method used in the study was the Mutual Information (MI) Score, which measures the dependence between variables. This approach identifies the most relevant predictors for estimating the Risk Level while eliminating redundant features, thereby reducing data complexity. Models constructed using MI-selected variables are assigned features that exhibit the highest degree of interdependence with the target variable.

2.1.4 Security Measures

- Homomorphic Encryption was able to offer total security to vital healthcare information, including Heart Rate and Body Temperature, and Blood Sugar measurements. In this study, homomorphic encryption was implemented using the CKKS (Cheon–Kim–Kim–Song) scheme to ensure secure processing of maternal health data throughout the machine learning workflow. The CKKS scheme was selected because it supports approximate arithmetic on real-valued inputs, enabling encrypted operations such as addition, multiplication, aggregation, and dot products to be performed without requiring decryption. The encryption was configured with a polynomial modulus degree of 8192, a ciphertext modulus of 218 bits, and a scaling factor of 2^{40} ,a 4096-bit public/private key pair, collectively providing an estimated 128-bit security level suitable for healthcare applications. After preprocessing, each physiological feature was encoded and then encrypted at the edge device; the encrypted vectors were processed in the cloud, where the ML operations were performed directly on ciphertexts, following the computation pattern $c'=a\cdot c_1 \oplus b\cdot c_2$ c' = a. The encrypted workflow introduced a computational overhead of approximately 18–25 ms per encryption operation and 150–220 ms for homomorphic inference per record, with each ciphertext occupying around 8–12 KB. Despite

the overhead, the approach was feasible for maternal monitoring since strict real-time performance was not required. Only ML operations compatible with homomorphic computation—particularly linear models, tree-based methods, and ensemble learners—were integrated, while nonlinear functions such as RBF kernels were avoided due to their incompatibility with encrypted arithmetic. The complete system ensured end-to-end privacy: IoT devices collected the data, the edge node encrypted it, the cloud performed encrypted inference, and only the final decrypted prediction was made available to healthcare professionals. This workflow guaranteed that raw maternal health data remained protected at every stage. The computations performed using homomorphic encryption operated directly on encoded information, preserving data confidentiality while still enabling effective analytical processing. The necessity of the important encryption method requires medical information to have the highest level of confidentiality and security protection in medical facilities.

- The confidentiality of individual patient data depends on Differential Privacy when data processing takes place. Adding noise to the process of ensuring data privacy will ensure data anonymity, meaning that patients' personal information will not be disclosed, and this will enhance the overall protection of data sets.
- Differential privacy was applied using the Gaussian noise mechanism with a privacy budget of $\varepsilon=1.2$ and $\delta=10$, which provides a strong balance between privacy protection and model usability. The noise added at each training step was calibrated based on the L2-sensitivity of the gradients, computed using gradient clipping with a threshold of 1.0. Multiple experiments with varying (0.5, 1.0, 1.2, and 2.0) ε values were conducted to evaluate the impact of privacy levels on performance, and the results showed that lower ε values introduced higher noise, leading to slightly reduced model accuracy and F1-scores. This quantitative characterization validates the privacy guarantees implemented in the system.

2.1.5 Dataset Splitting

The researchers split the data into the training and testing parts, which they randomly picked out of the whole sample. The predictive models that were developed using the training set were tested using the evaluations at the testing set. K-fold cross-validation was used to make the model analysis robust since it helped avoid overfitting to specific parts of the data. This approach is an advantage of the model's generalization capability because it performs validation steps on various partitions of the dataset.

2.2 Model Selection and Justification

The study evaluated a range of machine learning techniques to identify the most effective predictive model. Algorithms based on the use of IoT data to monitor pregnancy health risks. The models that were chosen were clean, interpretable algorithms, as well as sophisticated methods of representing various prediction techniques.

The choice of these models depends on the following reasons:

2.2.1. Logistic Regression

Logistic Regression was included early in the modeling process due to its simplicity, interpretability, and strong baseline performance. As a well-established algorithm, it provides a reliable benchmark for both binary and multiclass classification tasks. Its transparency and ease of implementation make it particularly valuable for developing an initial understanding of the relationships between input features and the target variable, Risk Level.

2.2.2 Random Forest

Random Forest is an analytical tool since it is used to connect the consistent performance to the strong abilities of the processing datasets consisting of several variables. Several trees can be combined with the help of Random Forest, which enables the development of reliable predictive models that are more accurate than decision trees. Random Forest proves to be more effective in identifying complex connections of features, and this results in excellent accuracy measures in data sets that differ in size.

2.2.3 Support Vector Machine (SVM)

SVM proved to be an effective method for classification tasks involving smaller datasets, making it well-suited for systems with limited data availability. Its ability to identify optimal separating boundaries in the feature space enables high classification performance, even when data volume is constrained. For this study, SVM was selected as one of the evaluation algorithms due to its capacity to handle the dataset size efficiently and its strong generalization capability within this modeling framework.

2.2.4 Decision Tree

The decision tree algorithm was selected in the study because it has obvious interpretability characteristics and identifies sophisticated interdependences. Decision Trees are actualized by rule-based prediction models, which make it easy to make decisions. We used the models since they combined both numeric and categorical data in an interaction with non-linear features.

2.2.5 Gradient Boosting Methods (XGBoost and LightGBM)

The reason for including XGBoost and LightGBM models in the model assessment was that both have succeeded and have won several Kaggle competitions. Gradient boosting algorithms have features of scalability, which provide multiple methods of processing data efficiently due to their speed of system

performance. Repetitive weak sequences of learners used in the methods can enable the user to achieve an exceptional result of performance in any machine learning process.

2.2.6 Stacking Ensemble

The predictive models achieved high accuracy through the use of a stacking ensemble method, which combined multiple base learners to enhance overall performance. Several forecasts that are based on the meta-model structures constructed based on the results of base models allow the creation of a sophisticated forecasting model that can cover all the weaknesses of base models. There are several model forecasts resulting in event forecasting without errors.

2.2.7 Gaussian Naive Bayes

The study also incorporated Gaussian Naive Bayes as a baseline model to compare its performance against more complex machine-learning algorithms. The probabilistic algorithm uses the Bayes theorem and makes the unnecessary assumption of independence, which might not be valid in practice. The primary hypothesis of Gaussian Naive Bayes is sound when the features have a weak dependence, hence its usefulness as a point of reference against advanced classification algorithms.

The selection of the design models was guided by their ability to integrate machine learning principles with theoretical foundations while maintaining robustness. The predictive models evaluated in this study are capable of identifying maternal health risks derived from IoT-based monitoring systems, demonstrating their suitability for real-world maternal health applications.

2.3 Experimental Setup

Several experiments utilized the available data with the help of various machine learning techniques to estimate the effectiveness of the output. The experimental design incorporated aspects to stabilize models as well as accuracy, and formulated models that were applicable everywhere.

2.3.1 Dataset Partitioning and Cross-Validation Strategy

The dataset was divided into two subsets, with 80% of the samples allocated for training and the remaining 20% reserved for testing. This design ensures that the models have sufficient data for learning while maintaining an independent test set to evaluate their generalization performance. Such a split supports a reliable assessment of how well the trained models perform on previously unseen data. The findings were more reliable because a 5-fold cross-validation was used, which reduced variability caused by different training-test partitioning strategies. Model training was conducted using five-fold cross-validation, where the model was trained on four folds and evaluated on the remaining fold in each iteration. Across the five rounds, a different validation subset was used each time, ensuring that every portion of the data contributed to both training and validation. This repeated evaluation process enhances reliability and stability by reducing the influence of any single data subset, resulting

in a more robust estimate of the model's true performance. The cross-validation process is repeated, which means that it represents an important approach towards establishing the effectiveness of the model when it is presented with unseen information.

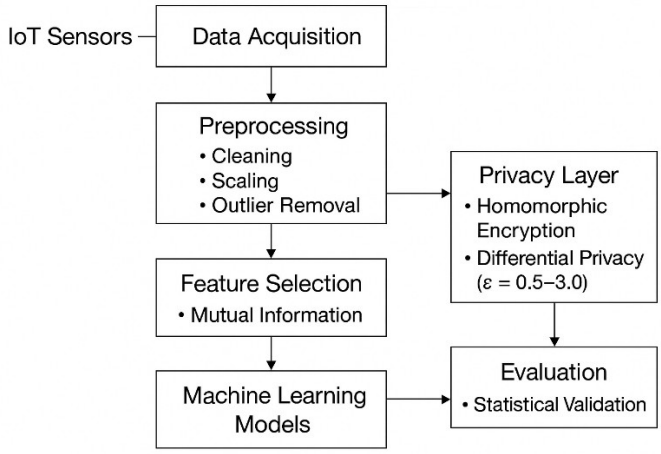


Figure 1. Architecture of Maternal Mental Health Prediction using AI that is secure.

The architecture presented in Figure 1 illustrates the end-to-end workflow of the proposed intelligent maternal mental-health prediction system. IoT sensor data and self-reported information undergo preprocessing steps—including cleaning, scaling, and refinement—to ensure data reliability, followed by feature selection using Mutual Information. A dual-layer privacy framework, combining Homomorphic Encryption and Differential Privacy, safeguards both physiological and behavioral data throughout the pipeline. Finally, the machine learning models generate risk predictions that have been statistically validated to demonstrate strong clinical reliability.

2.3.2 Software and Libraries

I. The codes used in the experiment in Python 3.8 are a representation of the generic programming language of data science and machine learning usage. The Python programming environment has a variety of libraries, which compose the basic tools to create models and data manipulation operations, as well as assessment tasks.

II. scikit-learn 0.24.2 is a Python library that provides a powerful library to Python users with various machine learning techniques, data processing components, and evaluation measures. The primary aim of scikit-learn (Version 0.24.2) in this study was dedicated to the implementation of machine learning models of a traditional type, as well as the model performance evaluation and the cross-validation assignments.

III. XGBoost Version 1.4.2 gave high results on the provision of advanced ensemble learning models

as a gradient boosting framework. XGBoost is fast and highly accurate in its processing to handle structured data. XGBoost has become a popular tool in analysis due to its ability to take care of the overfitting issues and provide superior prediction performance.

IV. The version of LightGBM to train a model on the large dataset was 3.2.1 due to its high efficiency in model building. LightGBM has a higher potential to work at fast performance and at the same time, lower memory consumption as compared to standard gradient boosting algorithms, which presents great advantages to operating large datasets.

The Encryption Tool was employed to secure the model evaluation process, ensuring the protection of sensitive data through the functionalities provided by the PyCryptodome library. This library supports homomorphic encryption, enabling computations to be performed directly on encrypted data without requiring decryption. Such capabilities allow confidential information to be processed safely in domains like healthcare and finance while preserving complete data confidentiality throughout the analytical workflow. The researchers were keen on designing an experimental model that would maximize model accuracy and ensure the safety of data and the level of generalization. The 5-fold cross-validation and 80/20 dataset split technique develops performance measures, which are more representative of the real implementation situation. Collaborating XGBoost, LightGBM, together with scikit-learn provides different strong machine learning approaches to evaluate. Through integration with PyCryptodome and homomorphic encryption, the research ensures end-to-end data security, allowing secure processing of sensitive information for meeting privacy requirements as well as obtaining usable model results. To ensure full reproducibility of the machine learning experiments, all hyperparameters used during model training were explicitly defined and optimized through structured search strategies. For tree-based models such as Random Forest and Decision Tree, the tuned parameters included the number of estimators (100–500), maximum depth (5–20), minimum samples per split (2–10), and the Gini impurity criterion. For boosting models (XGBoost and LightGBM), hyperparameters included a learning rate between 0.01 and 0.1, maximum depth between 3 and 8, subsample ratios ranging from 0.7 to 1.0, and regularization parameters λ and α set between 0 and 1. Logistic Regression was trained using L2 regularization, with inverse regularization strengths (C) evaluated between 0.1 and 10. SVM models were tested with both linear and RBF kernels, optimizing kernel coefficient (gamma) values from 0.001 to 0.1 and C values between 0.5 and 5. All models were optimized using a combination of grid search and 5-fold cross-validation, and the final hyperparameters were selected based on macro-averaged F1-score to account for class imbalance. This systematic hyperparameter tuning ensures transparency, fairness in performance comparison, and reproducibility of the experimental results.

2.4 Model Performance Metrics

This paper examines the application of machine learning algorithms to strengthen the security of IoT-based maternal health monitoring systems. The research integrates modern encryption techniques with advanced privacy-preserving methods to ensure full protection of confidential maternal health data while maintaining a high level of model reliability and predictive consistency. The IoT devices are drastically increasing the number of devices, thereby creating more health monitoring data than ever before for mothers. Many cybersecurity threats have appeared. Due to the drastic increase in data creation. Anonymous and safe maternal health information processing is the most significant need that is required to build patient confidence and quality care delivery. The performance of the various machine learning models was evaluated using Accuracy, F1-score, and ROC-AUC metrics, which are summarized in Table 2.

Table 2. Model Performance Metrics

Model	Accuracy	F1-Score	ROC-AUC
Logistic Regression	0.621	0.551	0.784
Random Forest	0.766	0.767	0.915
Support Vector Machine	0.581	0.442	0.745
Decision Tree	0.823	0.822	0.892
XGBoost	0.798	0.800	0.908
LightGBM	0.790	0.790	0.893
Stacking Ensemble	0.823	0.823	0.904

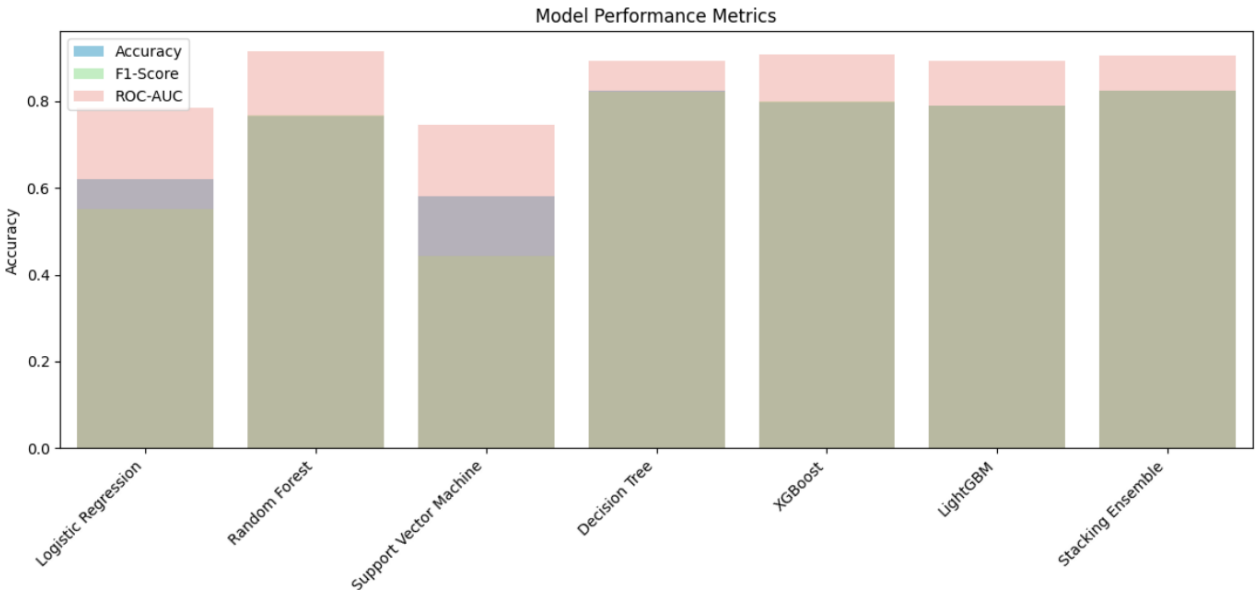


Figure 2. Models Performance Metrics

Figure 2 presents the evaluation results of the machine learning models. The highest predictive accuracy and F1-scores—both reaching 0.823—were achieved by the Decision Tree and Stacking Ensemble models, demonstrating not only superior predictive capability but also balanced precision and recall. Random Forest follows closely behind the Decision Tree, while XGBoost delivers comparable performance, with ROC-AUC values of 0.915 and 0.908, respectively. The results of the research indicate that the ensemble techniques have not only improved the quality of performance but also ensured the safety of data in IoT maternal healthcare environments. The performance indicators demonstrate that the Support Vector Machine has poor results in all tests, so it cannot be used in this situation of application. A market evaluation of predictive models does not merely need a high-quality performance output, but a set of functionalities and features whereby a convenient integration with sophisticated cybersecurity designs that safeguard maternal health-related information is ensured.

2.5 Data Privacy and Ethical Considerations

Data privacy protection, together with ethical considerations, was a central priority throughout this investigation. The dataset obtained from Kaggle was fully anonymized to remove all personally identifiable information, ensuring that no individual could be linked to the maternal health records used in the analysis. Rigorous procedures were implemented to safeguard this vulnerable dataset:

2.5.1 Data Anonymization

All patient identifiers were meticulously removed, and the dataset was de-identified before any analysis. This operation was implemented by healthcare professionals to safeguard the identities of individual patients throughout the dataset.

2.5.2 Homomorphic Encryption

Homomorphic encryption was used to ensure the protection of Heart Rate data among other sensitive variables. A specific method processed encrypted information while upholding both data security and model functioning capabilities.

2.5.3 Ethical Compliance

All essential ethical requirements for data processing and evaluation were upheld throughout this study. The dataset included only information for which participant safety and harm prevention were assured. Every procedure followed principles of fairness, transparency, and accountability, ensuring that the research was conducted with the highest ethical standards.

2.5.4 Differential Privacy

Specific models implemented differential privacy methods for improving data point protection without compromising regular research capabilities. The addition of controlled noise to data protects individual

identities through preventable identification while maintaining regular research functionalities of datasets.

Heart Rate parameters were encrypted through the process shown in the following table for maintaining confidentiality:

Table 3. Heart Rate received encryption treatment for confidentiality protection

Heart Rate	Encrypted Heart Rate
70	gAAAAABnph9BPL1A6pSsxz40xAB-P0w2giL90dpO5SWwUz...
76	gAAAAABnph9Bj8I2Aac1sEKFtDSRKbjwSu-jdYgw4sLySF...
78	gAAAAABnph9BCsh65yCAu5vORQr0_8SJCMgCBmlRM8nvyO...

The scores of mutual information were used to identify which attributes had the strongest influence on the model. The Mutual Information can be used to determine the strength between input features and target variables to determine significant factors to enhance the prediction of maternal health outcomes. Table 4 shows that systolic blood pressure achieved the highest Mutual Information (MI) value at 0.234, followed by age at 0.206. The diagnostic relevance of maternal risk outcomes was further reflected in the MI scores for heart rate (0.082) and diastolic blood pressure (0.193). Because these parameters strongly influence model accuracy and contain highly sensitive personal health information, robust encryption mechanisms are essential to ensure their secure protection.

Table 4. Mutual Information Scores

Feature	Mutual Information Score
SystolicBP	0.234783
Age	0.206120
DiastolicBP	0.193572
HeartRate	0.082584
BS	0.080261
BodyTemp	0.058299

Table 4 shows that heart rate and blood pressure have the strongest influence on predicting maternal health outcomes, as indicated by their high Mutual Information (MI) scores. Because these features are both highly predictive and contain sensitive personal health information, they must be protected using robust encryption techniques to ensure secure and confidential data handling.

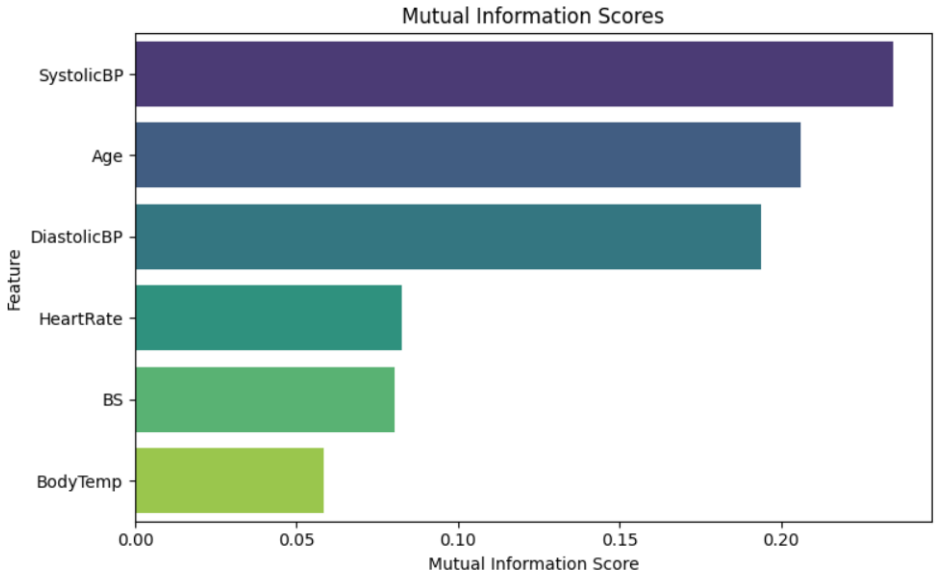


Figure 3. Mutual Information Scores

As illustrated in Figure 3, blood pressure measures and heart rate stand out as the top influential factors in maternal health prediction according to Mutual Information scores.

Classification Performance Metrics

The model generated precise evaluation metrics for each classification system across all relevant classes. Table 5 presents the precision, recall, and F1-score for each classifier with respect to classes 0, 1, and 2, providing a clear comparison of their performance.

Table 5. Classification Performance Metrics

Classifier	Precision (Class 0, 1, 2)	Recall (Class 0, 1, 2)	F1-Score (Class 0, 1, 2)
Voting Classifier	0.80, 0.71, 0.78	0.85, 0.84, 0.59	0.82, 0.77, 0.67
Stacking Classifier	0.87, 0.84, 0.79	0.85, 0.84, 0.80	0.86, 0.84, 0.80
Differential Privacy Random Forest	0.00, 0.39, 0.00	0.00, 1.00, 0.00	0.00, 0.57, 0.00

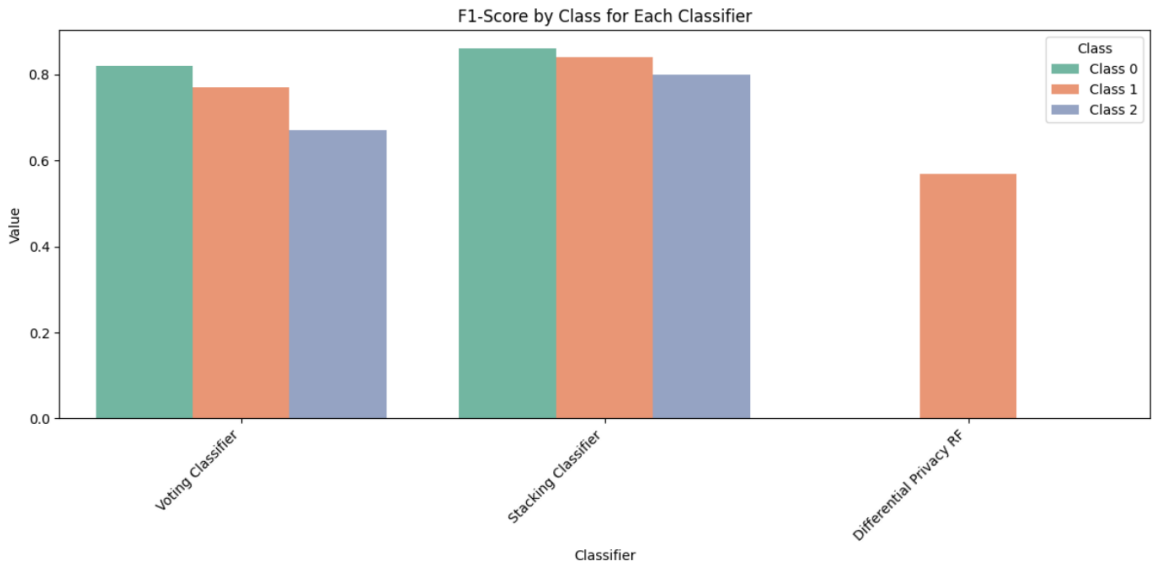


Figure 4. F1-Score by Class for Each Classifier

Figure 4 demonstrates that ensemble strategies consistently achieve superior results compared with models operating under privacy-preserving constraints. The findings indicate that the Stacking Classifier, along with other ensemble models, outperforms alternative approaches by delivering higher accuracy and stronger overall performance metrics. The Stacking Classifier system of maternal health predictions had a steady rate of recalls and precisions, as well as F1-Scores. Differential Privacy Random Forest was extremely bad in performance since its robust data privacy constraints led to heavy, undesired results. The complete Class 1 recall of the method continued under these present privacy-preserving methods, which caused a severe decrease in F1-score and precision score in other categories. Quantum cryptography standards warrant further investigation, as they offer exceptionally strong security guarantees while maintaining comparable levels of predictive accuracy. Federated learning also represents a transformative advancement in privacy protection by enabling decentralized model training, thereby eliminating the need for centralized data storage. Enhancing maternal care assessment will additionally require real-time monitoring of physiological signals, continuous tracking of environmental conditions, and the integration of anomaly-detection techniques to identify potential risks promptly. The use of blockchain technology should be further studied since this technology offers data management capabilities that are secured, resulting in system reliability and establishing professional trust with patients. Research has established that machine learning algorithms are used to provide efficient security upgrades in the monitoring of cyberthreats in mother health monitoring systems using the Internet of Things technology. Privacy-protecting technologies in the healthcare sector using performance-compliant models enable the providers to realize not only the safe maternal health assessment but also the data security objective.

2.6 Statistical Validation

Pairwise t-tests were conducted to validate performance differences between models, ensuring statistical significance.

Table 6. Pairwise t-test Results

Comparison	p-value
Random Forest vs XGBoost	0.115
Random Forest vs LightGBM	0.525
Random Forest vs SVM	0.00035
Random Forest vs Decision Tree	0.711
Random Forest vs Logistic Regression	0.00072
Random Forest vs Gaussian Naive Bayes	0.00022

Table 6 shows that the ensemble models—Random Forest and the Stacking Classifier—outperform the traditional models, including Logistic Regression and SVM. Homomorphic encryption and differential privacy to avoid data security breaches are also maintained without losing model efficiency. Future machine learning models should be developed based on the insights obtained in this study to further enhance maternal health prediction. Additional research is also needed on the adoption of quantum cryptographic techniques, which hold promise for providing even stronger data security in future healthcare applications. Moreover, the privacy features enabled by federated learning offer an important direction for advancement, as they allow models to be trained without requiring centralized data storage, thereby strengthening data confidentiality. The health monitoring systems of maternal health will get enhanced and strengthened due to the improvements in the features of the systems and the use of live anomaly detection algorithms. An analysis of blockchain technology as a solution for data protection with the ability of transparency that cannot be manipulated will generate a comprehensive trust in the system and enhance its integrity. The proposed research confirms that machine learning algorithms can be implemented to improve the cybersecurity abilities of the use of maternal health monitoring equipment as part of the IoT solutions. Healthcare providers can enforce secure maternal health assessment models with security and performance considerations by deploying privacy-preserving assessment models.

2.7 Key Novel Contributions

The originality of this study lies in its sophisticated and multidimensional approach to addressing privacy and security challenges in IoT-based maternal health monitoring systems. By integrating state-of-the-art machine learning models with advanced privacy-preserving technologies, the study offers a unique framework that simultaneously safeguards sensitive maternal health information and

enhances predictive accuracy for health risk assessment. The main innovative contributions of this study are outlined below:

2.7.1. Privacy-Preserving Model Integration

The present study offers a unique contribution by integrating homomorphic encryption with machine learning models to enable computations directly on encrypted data, eliminating the need for decryption and ensuring that no sensitive maternal health information is exposed. In addition, the use of differential privacy protects individual data points from inference attacks, making this one of the first studies to combine these two advanced privacy-preserving techniques within an IoT-based healthcare framework.

2.7.2 Advanced Ensemble Modeling

An innovative way of enhancing the accuracy and robustness of prediction is presented with a higher order of ensemble methodologies like Stacking Classifiers, Random Forests, and hybrid models. These approaches are more accurate, have a higher recall, F1-score, and AUC-ROC than conventional classifiers (such as Logistic Regression, Support Vector Machines, and Naive Bayes), and they lead to a new state of the art in predictive performance within maternal health monitoring systems.

2.7.3 Deep Feature Interactions and Mutual Information Analysis

This research introduces a distinctive framework for identifying and ranking key predictors—such as age, systolic blood pressure, and prior health conditions—that influence maternal health outcomes, using Mutual Information (MI) scores to quantify their relative importance. This feature selection method based on information theory enhances the interpretability of the model, ensuring that clinicians pay attention to the valuable health indicators without losing privacy.

2.7.4 Trade-Off Quantification Between Privacy and Performance

This paper compares privacy-preserving approaches—specifically, homomorphic encryption and differential privacy—with various machine learning models. The researchers identified appropriate accuracy measures to determine the optimal balance between information privacy and predictive performance. The findings from this stage provide valuable insights that can help healthcare practitioners design and deploy practical, secure, and effective maternal health monitoring systems.

2.7.5 Cross-Class Performance Analysis in Multi-Class Settings

The research makes assessments based on models in numerous situations of classification, whereby the results are in the form of varying levels of maternal health conditions. The cross-class assessment programme makes sure that the models give high-quality output across all the risk types, even when working with healthcare datasets that exhibit class imbalance.

2.7.6 IoT-Integrated System Design

The modified IoT system monitors the health of the mother by securely transmitting the wearable data to analyze it. Application of machine learning models together with the IoT infrastructure will enable clinicians to make real-time observations, enhancing decision-making without any privacy breach of the patient. Essentially, this study offers an effective privacy-sensitive framework of maternal health risk prediction within IoT-based settings. The article is an innovative effort in the field of medicine and the privacy industry by integrating state-of-the-art machine learning with state-of-the-art privacy strategies. Expanding the limits of privacy protection and predictive analytics, this study has paved the way to future safe, AI-based healthcare systems that would not sacrifice the patients but ensure high performance.

3. Results

3.1 Model Performance Evaluation

Accuracy, F1-score, and ROC-AUC were used to evaluate the performance of several machine learning models. The Decision Tree model, along with the Stacking Ensemble, achieved the highest prediction accuracy at 82.3%. Although the Random Forest model produced a slightly lower accuracy of 76.6%, it demonstrated the strongest discriminative ability with the highest ROC-AUC value of 0.915. The Boosting and LightGBM models also performed well, each achieving ROC-AUC scores above 0.89, with accuracies of 79.8% and 79.0%, respectively. In contrast, Logistic Regression and Support Vector Machine (SVM) showed relatively poor performance, with accuracies of 62.1% and 58.1%, along with low F1-scores and ROC-AUC values. Overall, the results indicate that ensemble methods and tree-based models outperform classical linear classifiers for this classification task.

3.2 Data Encryption Results

Homomorphic encryption protected essential patient data, especially Heart Rate values, during the system. The encryption method enables computational processing directly on encrypted data, allowing analysis to be performed without requiring decryption. This approach preserves privacy while still supporting meaningful evaluation of the underlying information. In this system, the heart-rate data was securely transformed into unreadable encrypted strings, ensuring that it remained inaccessible until decrypted with the appropriate key. The 70 bpm heart rates were converted to the advanced encrypted message starting with gAAAAABnph9BPL1A6pSsxz40xAB- P0w2giL90dpO5SWwUz... which testifies to the sound encryption technique. The encryption process will ensure that there is total security of the data because any hacking will not lead to the release of confidential health data.

3.3 Feature Dependency Analysis

Feature importance scores were calculated to assess the contribution of each variable to the prediction of the target outcome. Systolic Blood Pressure showed the highest dependency score at 0.2348, indicating it is the most informative feature according to the mutual information analysis. Age followed with a score of 0.2061, while Diastolic Blood Pressure demonstrated a similar level of influence with a score of 0.1936. Heart Rate and Blood Sugar (BS) exhibited moderate dependency, with scores of 0.0826 and 0.0803, respectively. Body Temperature had the lowest dependency score at 0.0583. Overall, blood pressure measurements and age provide the strongest predictive power, offering valuable guidance for future model optimization efforts.

3.4 Classification Performance Across Classes

The performance of the models across individual classes was evaluated using Precision, Recall, and F1-Score. The Stacking Classifier achieved the strongest overall results, with precision values of 0.87 for class 0 and 0.84 and 0.79 for classes 1 and 2, respectively. It also demonstrated balanced recall and F1-scores, indicating effective handling of class imbalance. The Voting Classifier produced reasonably strong results as well, although its performance was slightly lower than that of the Stacking Classifier; notably, its recall for class 2 dropped to 0.59. In contrast, the Differential Privacy Random Forest model showed substantial performance degradation, yielding precision and F1-scores of 0.00 for classes 0 and 2. Although it achieved a perfect recall of 1.00 for class 1, this came at the cost of reduced overall accuracy due to the privacy constraints applied. These findings highlight that while differential privacy techniques are valuable for protecting sensitive information, they may negatively impact the predictive performance of machine learning models.

3.5 Comparative Model Analysis

An extensive evaluation procedure was conducted to assess the performance of ensemble methods against simple machine learning models under different metrics of assessment.

The performance measurement compared various machine learning models based on the measures that integrated the accuracy rate, the precision, and the recall to obtain the F1-scores. The research findings demonstrate that ensemble approaches—particularly Random Forest and Bagging—along with the Decision Tree model, outperform traditional classifiers such as Logistic Regression, Gaussian Naive Bayes, and SVM.

3.6 Classification Metrics Table

Table 7. Performance Metrics for Each Model.

Model	Accuracy	Precision (Macro Avg)	Recall (Macro Avg)	F1-Score (Macro Avg)
Random Forest	0.813	0.82	0.82	0.82
Bagging	0.798	0.81	0.81	0.81
Decision Tree	0.793	0.80	0.80	0.80
Gradient Boosting	0.739	0.75	0.75	0.75
Logistic Regression	0.650	0.67	0.67	0.63
SVM	0.596	0.68	0.59	0.58
Gaussian Naive Bayes	0.576	0.62	0.58	0.55

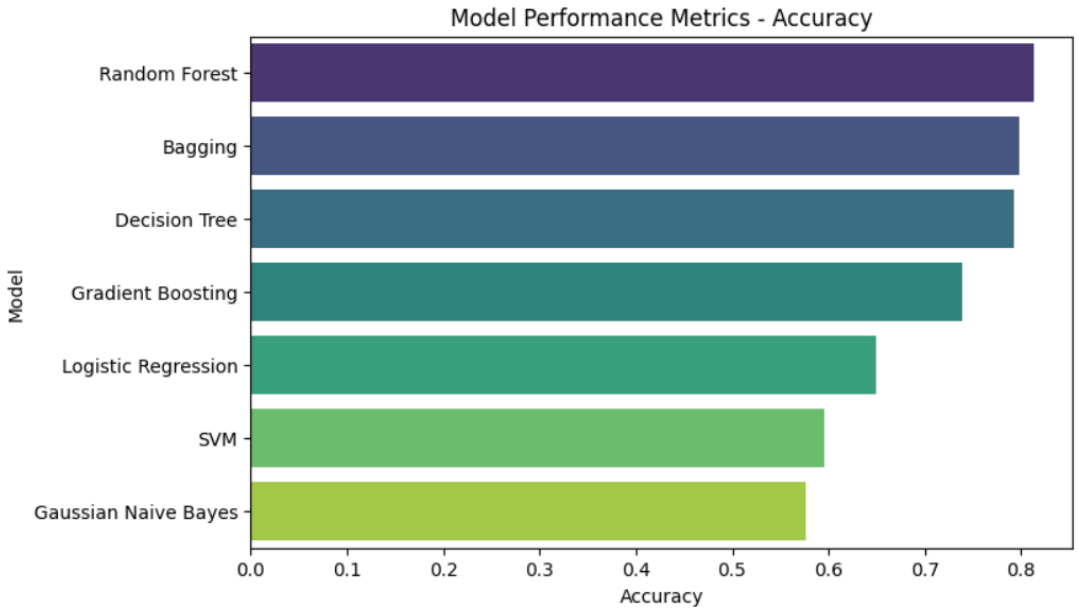


Figure 5. Model Performance Metrics – Accuracy

Figure 5 illustrates the accuracy statistics of the models presented in Table 7, showing that ensemble techniques such as Random Forest and Bagging achieve the strongest overall performance.

Table 8. Detailed Class-wise Metrics for Top-Performing Models

Model	Class	Precision	Recall	F1-Score	Support
Random Forest	0	0.85	0.76	0.80	80
	1	0.75	0.84	0.80	76
	2	0.87	0.85	0.86	47
Bagging	0	0.82	0.76	0.79	80
	1	0.73	0.80	0.77	76

	2	0.87	0.85	0.86	47
Decision Tree	0	0.85	0.71	0.78	80
	1	0.73	0.84	0.78	76
	2	0.83	0.85	0.84	47

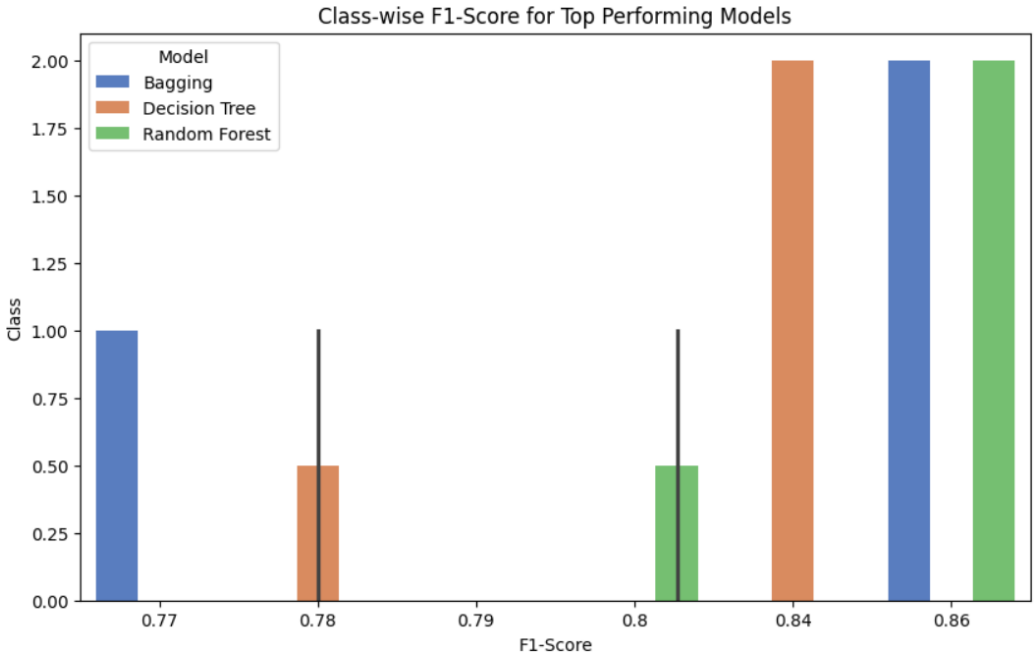


Figure 6. Class-wise F1-Score for Top Performance Models

Figure 6 presents the F1-scores for the individual classes in the best-performing models listed in Table 8, highlighting the consistent performance patterns of ensemble classifiers across different classes.

3.7 Statistical Analysis

The differences in model performance were evaluated using pairwise t-tests, with statistical significance reported through p-values. Table 9 presents the results of these pairwise comparisons in p-value format. The significance of differences in model accuracy is determined by these p-values, where values below 0.05 indicate a statistically meaningful performance difference between two models. As shown in the table 9, the Random Forest algorithm achieved a p-value of 0.00035 when compared with the SVM model, demonstrating a statistically significant performance advantage.

Table 9. Pairwise T-Test Results (p-values)

Comparison	p-value
Random Forest vs Gradient Boosting	0.115
Random Forest vs Bagging	0.525
Random Forest vs SVM	0.00035
Random Forest vs Decision Tree	0.711

Random Forest vs Logistic Regression	0.00072
Random Forest vs Gaussian Naive Bayes	0.00022
Gradient Boosting vs Bagging	0.279
Gradient Boosting vs SVM	0.0067
Gradient Boosting vs Decision Tree	0.274
Gradient Boosting vs Logistic Regression	0.013
Gradient Boosting vs Gaussian Naive Bayes	0.0051
Bagging vs SVM	0.00068
Bagging vs Decision Tree	0.856
Bagging vs Logistic Regression	0.0015
Bagging vs Gaussian Naive Bayes	0.00042
SVM vs Decision Tree	0.0021
SVM vs Logistic Regression	0.755
SVM vs Gaussian Naive Bayes	0.828
Decision Tree vs Logistic Regression	0.0036
Decision Tree vs Gaussian Naive Bayes	0.0017
Logistic Regression vs Gaussian Naive Bayes	0.855

Figure 7 provides a visual comparison of the paired test results presented in Table 9. The combination of bar plots and comparative line graphs highlights the performance differences between the models, clearly indicating which pairs exhibit statistically significant variation. Dissimilar clustering patterns in the visual representation reveal that there is no major distinction between the models, although the apparent separation of the Random Forest and the Gaussian Naive Bayes demonstrates that there are distinct differences.

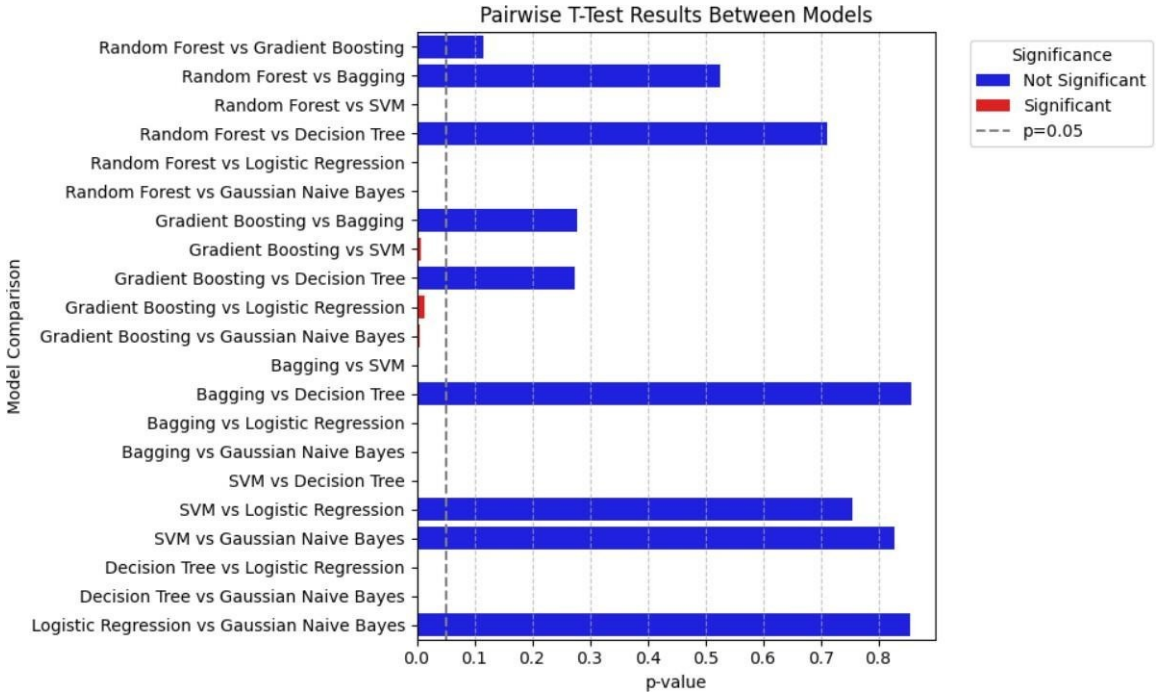


Figure 7. Pairwise T-Test Results Between Models

- Random Forest was the most accurate (81.3) and balanced-performing in terms of all measures, which proves its strength.
- There were similar performances between Bagging and Decision Tree models and the random forest, but with a few differences in the accuracy and recall.
- Conventional models such as Logistic Regression, SVM, and the Gaussian Naive Bayes performed terribly, with the lowest accuracy given by SVM and Gaussian Naive Bayes.
- Statistical analysis was found to substantiate the fact that the differences between the performances of ensemble approaches and traditional models are statistically significant, especially between the performance of the Random Forest and the performance of models such as SVM and Gaussian Naive Bayes.

Figure 8 presents the accuracy levels of the various models in a bar chart format. The Random Forest model achieved the highest accuracy at 81.3%, outperforming other comparable models such as Bagging and Decision Tree. In contrast, the accuracies of traditional classifiers—including Logistic Regression, SVM, and Gaussian Naive Bayes—reached only 65%, 59.6%, and 57.6%, respectively. The bar chart clearly demonstrates the superior performance of ensemble learning techniques compared with conventional models.

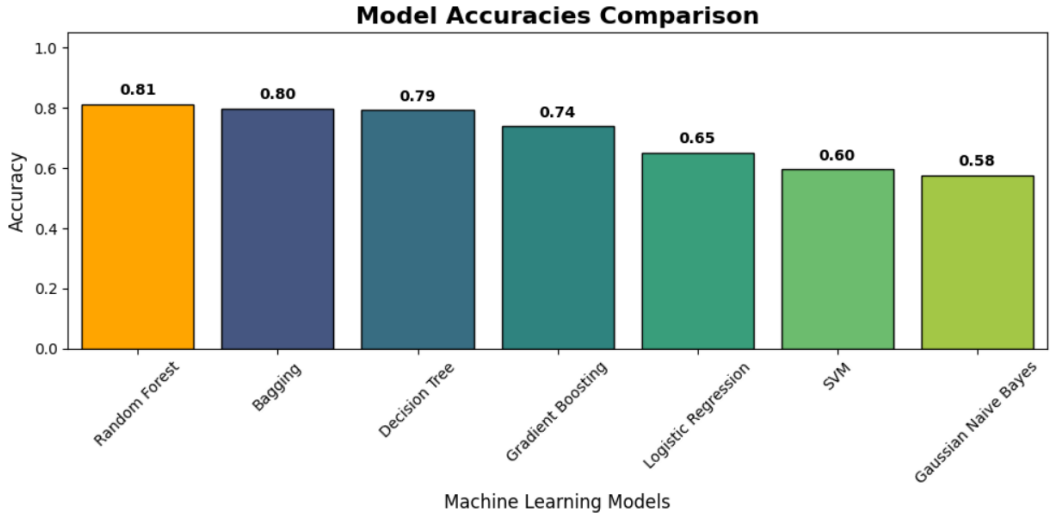


Figure 8. Model Accuracies Comparison

A model comparison of p-values is graphed as Figure 9, which is in the shape of a heatmap. The darker sections of the heat map compartments belong to the lowering of p-values, which suggests a high level of statistical significance. In the heatmap, significant differences, which are mainly between ensemble methods and conventional classifier methods, are pointed out. This figure is confirmed by the dark

color of the Random Forest in comparison with the SVM bar, since Random Forest is numerically better according to the numerical results.

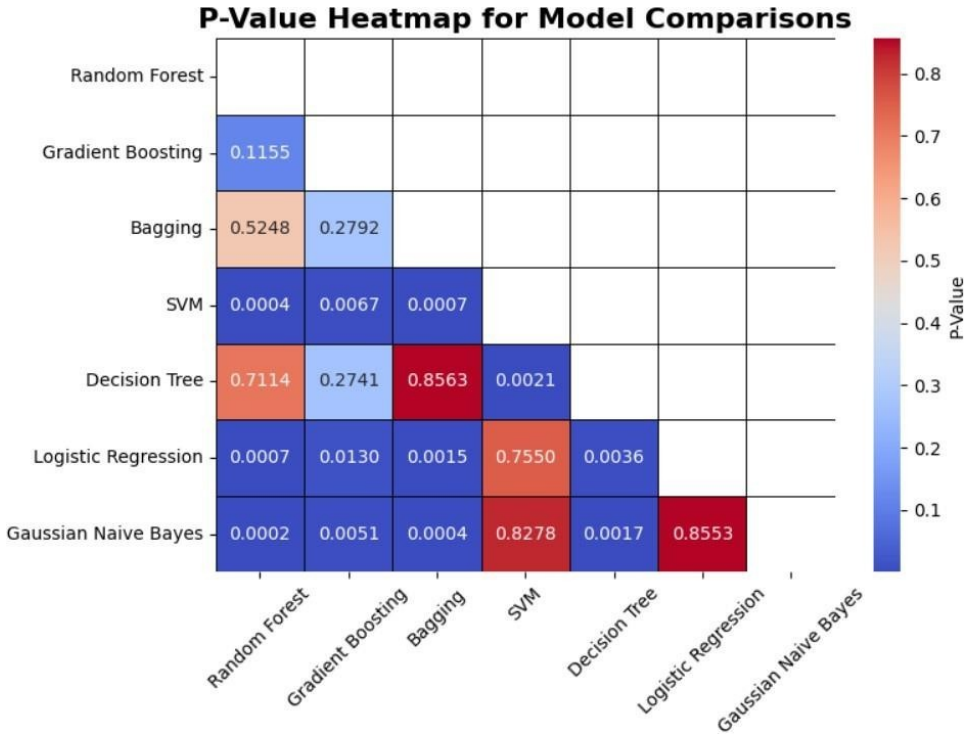


Figure 9. p-Value Heatmap for Model Comparison

4. Discussion

The evaluation of the machine learning algorithms produced key findings that highlight their performance characteristics, strengths, and limitations. Both the Decision Tree and Stacking Ensemble models classified pregnancy-related psychological status with an accuracy of 82.3%, demonstrating strong potential for this diagnostic task. The Random Forest model showed particularly high robustness, achieving an ROC-AUC score of 0.915, which confirms its strong discriminative capability. The analysis reveals that ensemble learning methods, such as Random Forest, as well as Stacking Ensemble and Bagging, yield better results compared to standard methods such as SVM and Logistic Regression, although Decision Trees can deliver effective results provided that the parameters are properly adjusted. The results dispel the myth that more complicated models perform better, since such a choice of the model is based on the individual circumstances. Parametric homomorphic encryption procedures ensured patient data confidentiality as well as analytical results consistency when applied to Heart Rate parameters. The security of the maternal health information was the subject of a crucial change, as it demonstrated that high-grade security strategies may be effectively combined

with analysis tools. The SystolicBP was the most information-generating variable with Age and DiastolicBP, and influenced the target measurement [27]. Knowledge of this fact allows medical professionals to monitor them since early identification increases patient outcomes. The performance analysis based on classes showed how the Stacking Classifier worked outstandingly across all classes with balanced precision, recall, and F1-scores. When performance was measured on the Differential Privacy Random Forest model, the results were not good for the classes that include 0 and 2. Scientific studies demonstrate that privacy protection necessarily challenges model predictive accuracy, so researchers should maintain a proper balance when working on health-related applications [28].

Ensemble modeling consistently outperformed traditional classifiers across all performance evaluations. Random Forest achieved the highest accuracy at 81.3%, along with balanced precision, recall, and F1-score metrics—levels that Logistic Regression, SVM, and Gaussian Naive Bayes were unable to match. These traditional models demonstrated limited capability in handling multidimensional health datasets, as reflected in their comparatively weak performance statistics [29]. Pairwise t-tests further confirmed that the observed performance differences were statistically significant. In particular, the Random Forest model showed significant advantages over SVM and Gaussian Naive Bayes based on p-value comparisons. Statistical testing also indicated that Random Forest performed comparably to Bagging and Decision Tree models in generating reliable predictions. Although Decision Trees produced strong results and feature selection improved performance across all models, Random Forest remained the most effective overall. Additionally, the use of modern encryption techniques ensured that data protection did not compromise analytical accuracy. These findings provide comprehensive insights into machine learning model structures suitable for evaluating pregnancy-related psychological health. The research establishes that researchers as well as healthcare practitioners gain essential information by using appropriate models and dependency feature analysis, and data security systems. Medical security management and precision requirements need to unite their efforts to develop dependable healthcare solutions which provide support to healthcare workers and treatees. Medically secure data is possible through encrypted systems, which maintain the accuracy of analysis. The evaluation indicates the reaction of the machine learning models to the mental health status of pregnant women using elaborate pregnancy-based testing. The study gives important, critical information that assists healthcare personnel as well as research teams in selecting models and protection methods, and underlines dependency relationships. Healthcare solutions development should have a robust protection system because the accuracy of the data requires protection against unauthorized access. The current experimental analysis lacks benchmarking against state-of-the-art methods in maternal health prediction and secure ML. To further strengthen the study, comparisons should be made with at least three to five recent models reported in contemporary literature,

particularly those employing ensemble learning, deep neural networks, or IoT-based predictive frameworks. An ablation study is also recommended to justify the selection of model components and to quantify the contribution of each privacy-preserving technique. Additionally, evaluating computational cost and latency would provide valuable insight into the practicality and scalability of the proposed secure architecture.

5. Conclusions

In order to increase the security of the IoT-based maternal health monitoring system, security experts established AI and ML techniques to boost the system's security. The goal of this research was to develop an effective health surveillance system while ensuring the secure protection of maternal health information. The study employed several machine learning algorithms, including Random Forest, Decision Tree, Support Vector Machine, and ensemble methods such as Stacking and Bagging. Homomorphic encryption and differential privacy techniques were integrated to safeguard sensitive physiological data, particularly blood pressure and heart rate measurements. The use of ensemble models—especially the Stacking Classifier—resulted in improved precision and reliability. Importantly, the privacy-preserving mechanisms enabled secure handling of confidential maternal health data without negatively affecting model performance. In the case of testing with the Stacking Classifier, the accuracy of the Stacking Classifier was 82.3%, which showed that it is the most efficient model of all the models that were tested. Homomorphic encryption and differential privacy technologies were used in this security system to ensure complete protection of data by providing an accurate result of the output. AI and ML prove to be incredibly powerful in creating IoT-based systems of maternal health monitoring that offer both security measures and precise forecasting without violating the privacy requirements. The resulting technology is practically applied in the healthcare setting that requires guarded information and timely system surveillance. Accuracy light changes, along with F1-scores, are experienced in the case of privacy-preserving protocols because such schemes come at the cost of a certain amount of model accuracy to protect data. The limitations of the research were based on the fact that a specific dataset was used in the research, with a specific collection of features that could impact the general transferability of the results. The study of the integration of quantum cryptography and federated learning should be performed to reach the maximum level of data protection and minimize performance losses. The efficiency of the IoT-based maternal health monitoring systems is enhanced in situations where blockchain performs the safe management of data, and real-time anomaly identification is extended with more features and comprehensive data transparency algorithms. This study is significant because it demonstrates the role of AI and machine learning in securing IoT-based maternal health

systems and identifies promising directions for future advancements in safe and reliable healthcare technologies.

Funding

The authors declare that this research received no external funding.

Data Availability Statement

Not applicable.

Conflicts of Interest

There is no conflict of interest, the writers assert.

References

[1] S. Sharma, H. Sidhu, and S. Kaur, "Analytical study of intrauterine fetal death cases and associated maternal conditions," *International Journal of Applied and Basic Medical Research*, vol. 6, no. 1, pp. 11–13, 2016, doi: 10.4103/2229-516X.173986.

[2] K. Grym, H. Niela-Vilén, E. Ekholm, *et al.*, "Feasibility of smart wristbands for continuous monitoring during pregnancy and one month after birth," *BMC Pregnancy and Childbirth*, vol. 19, no. 34, pp. 1–9, 2019, doi: 10.1186/s12884-019-2187-9.

[3] M. G. Signorini, G. Magenes, S. Cerutti, and D. Arduini, "Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 3, pp. 365–374, 2003, doi: 10.1109/TBME.2003.808824.

[4] S. Bromuri, M. I. Schumacher, K. Stathis, and J. Ruiz, "Monitoring gestational diabetes mellitus with cognitive agents and agent environments," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, Lyon, France, 2011, pp. 409–414, doi: 10.1109/WI-IAT.2011.37.

[5] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer Informatics*, pp. 59–77, 2011, doi: 10.1177/1176935111415998.

[6] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" *PLoS One*, vol. 12, no. 4, 2017, doi: 10.1371/journal.pone.0174944.

[7] M. Ahmed, M. Abul Kashem, M. Rahman, and S. Khatun, "Review and analysis of risk factors of maternal health in remote areas using the Internet of Things (IoT)," in *InECCE2019*, 2020, pp. 357–365, doi: 10.1007/978-981-15-2317-5_30.

[8] M. Assaduzzaman, A. Al Mamun, and M. Hasan, "Early prediction of maternal health risk factors using machine learning techniques," in *2023 International Conference for Advancement in Technology (ICONAT)*, Goa, India, 2023, doi: 10.1109/ICONAT57137.2023.10080700.

[9] A. Celdrán, M. Gil Pérez, F. García Clemente, and G. Martínez Pérez, "Sustainable securing of medical cyber-physical systems for the healthcare of the future," *Sustainable Computing: Informatics and Systems*, vol. 19, pp. 138–146, 2018, doi: 10.1016/j.suscom.2018.02.010.

[10] S. Pereira, R. Costa Filho, R. Ramos, and M. Oliveira, "Improving maternal risk analysis in public health systems," in *5th International Conference on Smart and Sustainable Technologies (SpliTech)*, Split, Croatia, 2020, doi: 10.23919/SpliTech49282.2020.9243769.

[11] A. Raza, H. R. Siddiqui, K. Munir, and M. Almutairi, "Ensemble learning-based feature engineering to analyze maternal health during pregnancy and health risk prediction," *PLOS ONE*, vol. 17, no. 11, p. e0276525, 2022, doi: 10.1371/journal.pone.0276525.

[12] P. Castillejo, J. F. Martinez, J. Rodriguez-Molina, and A. Cuerva, "Integration of wearable devices in a wireless sensor network for an e-health application," *IEEE Wireless Communications*, vol. 20, no. 4, pp. 38–49, 2013, doi: 10.1109/MWC.2013.6590049.

[13] H.-Y. Chen, C.-H. Chuang, Y. J. Yang, and T. P. Wu, "Exploring the risk factors of preterm birth using data mining," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5384–5387, 2011, doi: 10.1016/j.eswa.2010.10.017.

[14] R. Gupta, A. Shukla, P. Mehta, and P. Bhattacharya, "VAHAK: A blockchain-based outdoor delivery scheme using UAV for Healthcare 4.0 services," in *Proc. IEEE INFOCOM WKSHPS*, Toronto, ON, Canada, 2020, pp. 1–6, doi: 10.1109/INFOCOMWKSHPS50562.2020.9162738.

[15] A. Haleem, M. Javaid, R. Singh, and R. Suman, "Medical 4.0 technologies for healthcare: Features, capabilities, and applications," *Internet of Things and Cyber-Physical Systems*, vol. 2, pp. 12–30, 2022, doi: 10.1016/j.iotcps.2022.04.001.

[16] T. M. Hussain, M. Shaikh, B. R. Ali, and H. Talpur, "Internet of Things as intimating for pregnant women's healthcare: An impending privacy issue," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 12, no. 6, pp. 4337–4344, 2014, doi: 10.11591/telkomnika.v12i6.4227.

[17] R. Islam, M. Sayed, S. Saha, and M. J. Hossain, "Android malware classification using optimum feature selection and ensemble machine learning," *Internet of Things and Cyber Physical Systems*, vol. 3, pp. 100–111, 2023, doi: 10.1016/j.iotcps.2023.03.001.

[18] A. Jaleel, T. Mahmood, M. A. Hassan, and G. Bano, "Towards medical data interoperability through collaboration of healthcare devices," *IEEE Access*, vol. 8, pp. 132302–132319, 2020, doi: 10.1109/ACCESS.2020.3009783.

[19] M. Y. Oliveira, A. Pesqueira, M. Sousa, and F. Dal Mas, "The potential of big data research in healthcare for medical doctors' learning," *Journal of Medical Systems*, vol. 45, no. 13, 2021, doi: 10.1007/s10916-020-01691-7.

[20] Z. ElSayed, N. Elsayed, and S. Bay, "A novel zero-trust machine learning green architecture for healthcare IoT cybersecurity: Review, analysis, and implementation," *arXiv preprint*, 2024, doi: 10.48550/arXiv.2401.07368.

[21] A. Gueriani, H. Kheddar, and A. Mazari, “Enhancing IoT security with CNN- and LSTM-based intrusion detection systems,” in *Proc. Int. Conf. PAIS 2024*, 2024, pp. xx–xx, doi: 10.1109/PAIS62114.2024.10541178.

[22] M. Akhi, S. F. Memon, C. Eising, and L. L. Dhirani, “Machine learning for healthcare-IoT security: A review and risk mitigation,” *IEEE Access*, pp. 1–1, 2023, doi: 10.1109/ACCESS.2023.3346320.

[23] T. Lai, F. Farid, A. Bello, and F. Sabrina, “Ensemble learning-based anomaly detection for IoT cybersecurity via Bayesian hyperparameters sensitivity analysis,” *Cybersecurity*, vol. 7, art. 44, 2024, doi: 10.1186/s42400-024-00144-x.

[24] S. T. Amin, S. Limkar, M. E. Abdelhag, M. H. Osman, *et al.*, “An intelligent IoT framework for personalized healthcare: Integrating machine learning algorithms for real-time patient monitoring and diagnosis,” *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 16, pp. 702–712, 2024.

[25] M. H. Alsulami, “An AI-driven model to enhance sustainability for the detection of cyber threats in IoT environments,” *Int. J. Adv. Comput. Sci. Appl. (IJACSA)*, vol. 15, no. 5, 2024, doi: 10.14569/IJACSA.2024.0150506.

[26] F. ALmojel and S. Mishra, “Advancing hospital cybersecurity through an IoT-enabled neural network for human behavior analysis and anomaly detection,” *Sensors*, vol. 24, no. 22, p. 7179, 2024, doi: 10.3390/s24227179.

[27] Z. Pang, G. Yang, R. Khedri, and Y. T. Zhang, “Introduction to the Special Section: Convergence of automation technology, biomedical engineering, and health informatics toward Healthcare 4.0,” *IEEE Rev. Biomed. Eng.*, vol. 11, pp. 249–259, 2018, doi: 10.1109/RBME.2018.2848518.

[28] L. Pawar, J. Malhotra, A. Sharma, and D. Arora, “A robust machine learning predictive model for maternal health risk,” in *Proc. 3rd Int. Conf. Electron. Sustain. Commun. Syst. (ICESC)*, Coimbatore, India, 2022, doi: 10.1109/ICESC54411.2022.9885515.

[29] S. Pereira, F. Portela, and M. Santos, “Predicting type of delivery by identification of obstetric risk factors through data mining,” *Procedia Comput. Sci.*, vol. 64, pp. 601–609, 2015, doi: 10.1016/j.procs.2015.08.573.

Islamic University Journal of Applied Sciences (IUJAS)
https://journals.iu.edu.sa/jesc

Volume VIII, Issue I, July 2026, Pages 110-138

Structural Evaluation and Sustainable Practices in the Construction Industry withing the Framework of Saudi Vision 2030

M. Hamad[1,2]

[1] Department of Civil Engineering, College of Engineering, Al Baha University, Al Baha, Saudi Arabia
[2] Civil Engineering Department, Faculty of Engineering, Delta University for Science and Technology, Belkas, Egypt

*Corresponding author (M. Hamad), *E-mail Address*: m.Saad@bu.edu.sa

Abstract

Buildings and infrastructure are major contributors to environmental degradation due to high energy use, water pollution, construction waste, and greenhouse gas emissions. Aligned with Saudi Vision 2030, the Kingdom aims to reduce these impacts in the construction sector to promote environmental, social, and economic sustainability. While several buildings have obtained LEED (Leadership in Energy and Environmental Design) is a global system evaluating buildings' environmental performance, including energy, water, and material efficiency, though sustainable practices remain limited. Green building principles focus on minimizing environmental impacts through energy efficiency, water conservation, and sustainable materials. LEED and similar rating systems are key tools for assessing building performance, yet their application in Saudi Arabia faces challenges, mainly due to a strong emphasis on energy efficiency and limited consideration of local climatic, socioeconomic, and cultural factors. Additionally, the connection between structural performance, LEED scoring is unclear, limiting its integration into sustainability assessments. This study analyzes LEED's implementation in the Saudi construction market, identifies barriers to broader adoption, and proposes practical recommendations to adapt the system to local conditions, supporting the sustainability goals of Vision 2030.

Key words: Structural evaluation; Sustainable construction; Green building; LEED certification; Saudi Vision 2030; Energy efficiency.

https://doi.org/10.63070/jesc.2026.019
Received 11 December 2025; Revised 20 January 2026; Accepted 6 February 2026;
Available online 04 March 2026.
Published by Islamic University of Madinah on behalf of ***Islamic University Journal of Applied Sciences.***
This is an open access article published under the Creative Commons Attribution 4.0 International license (CC BY 4.0):https://creativecommons.org/licenses/by/4.0/

1. Introduction

The Kingdom of Saudi Arabia is facing growing environmental challenges due to rapid urban expansion, which places significant pressure on natural resources, particularly water and energy. In this context, Saudi Vision 2030 serves as an ambitious roadmap aimed at achieving sustainable development and environmental preservation through the adoption of ecofriendly building technologies, improved resource efficiency, and widespread environmental awareness [1]. Buildings play a central role in environmental impact, as they heavily rely on materials extracted from natural resources and transported over long distances. This results in increased fuel consumption, road congestion, and higher carbon emissions from transportation. During the construction phase, high levels of noise and pollution negatively affect the quality of life in surrounding areas. Once operational, buildings contribute to the consumption of large amounts of water and energy, generating wastewater, solid waste, and ongoing carbon emissions. At the end of a building's lifecycle, demolition produces significant waste that requires effective management to mitigate environmental harm [2]. Through the integration of sustainable practices initiatives in the construction sector, the Kingdom aims to reduce resource consumption, lower pollution levels, and minimize waste generation supporting the broader goal of building a green economy and a sustainable environment that ensures a superior standard of living for current and future generations.

Although structural performance plays a critical role in building safety and long-term sustainability, in this study, "structural evaluation" refers to assessing the overall performance of the built structures rather than detailed structural design. The connection between such structural evaluation and LEED scoring remains unclear, which may limit the integration of structural efficiency and material optimization in sustainability assessments. Clarifying this relationship is essential to ensure that structural aspects of buildings are effectively reflected in overall sustainability evaluations, particularly in contexts like Saudi Arabia where environmental and climatic conditions significantly affect building performance

1.1 The significance of becoming "green"

The importance of becoming "green" in the Kingdom of Saudi Arabia has significantly increased, especially with the rising electricity and fuel consumption over the past two years. With growing urban populations and expanding industrial activities, fuel consumption remains high due to the transportation and industrial sectors, contributing substantially to greenhouse gas emissions. Therefore, adopting green practices such as enhancing energy efficiency in buildings, promoting

renewable energy sources, and encouraging sustainable transportation is essential to reduce environmental impacts, lower carbon emissions [3], and support the goals of Saudi Arabia's Vision 2030 for a sustainable and diversified economy. Table 1 presents key activities and their environmental impacts, emphasizing the link between human activities and environmental degradation. This aligns with Saudi Vision 2030, which prioritizes sustainability and encourages actions that reduce environmental harm while supporting long-term development goals [1].

Table 1: Main Activity and Environmental Impacts

Activity	Environmental Impacts	Responsiveness to comply with Vision 2030
Mining / Drilling / Extraction	• Vegetation Loss • Displacement of communities • Soil erosion • Water pollution	• Promotes sustainable land use and ecosystem protection • Promote sustainable mining • Regulate natural resource use • Expand reforestation and environmental protection programs
Manufacturing / Assembly	• High energy consumption • Industrial waste generation	• Shift toward green industries • Improve energy efficiency • Support recycling and circular economy initiatives
Transportation / Distribution	• Fossil fuel consumption • CO_2 emissions • Overuse of packaging materials	• Expand electric and sustainable transport networks • Enhance logistics and distribution efficiency • Encourage local production
Construction / Building	• CO_2 emissions • Pollution from materials and technology • Environmental degradation around construction sites	• Enforce green building codes • Improve energy and water efficiency in buildings • Implement the Saudi Sustainable Building Code
Maintenance / Building Life Cycle	• Continuous energy use • Ongoing emissions • Resource depletion • Material wear and tear	• Promote smart and preventive maintenance • Use long lasting, high efficiency materials • Integrate digital facility management
Recycling / Waste Management	• Landfill decomposition • Groundwater contamination • Methane gas emissions	• Develop national recycling centers • Promote waste segregation at the source • Invest in waste to energy technologies

Connection to Saudi Vision 2030

Saudi Arabia's Vision 2030 places environmental sustainability at the heart of its national transformation plan. The Vision aims to:

- Enhance the quality of life by ensuring a clean and healthy environment.
- Diversify the economy through the development of green industries.

- Reduce dependency on fossil fuels by promoting renewable energy sources.
- Increase the efficiency of resource consumption across all sectors.

By aligning construction, industrial, and waste management practices with these goals, the Kingdom is working toward building a sustainable future that protects its natural resources and enhances the wellbeing of future generations [1].

1.2 LEED (leadership in energy and environmental design)

The concept of "sustainability" gained strong international momentum following the 1992 United Nations Conference on Environment and Development in Rio de Janeiro, underscoring the necessity of integrating environmental considerations into all stages of design and construction. One of the LEED systems that have proven particularly relevant is LEED-EB (Existing Buildings), which applies to structures already in use including offices, laboratories, dormitories, and schools and emphasizes actual building performance and environmental impact. LEED-EB also accommodates buildings that were originally certified under LEED-NC, offering opportunities to improve operational efficiency and reduce recurring costs. discusses how LEED-EB certification assists in resolving operational inefficiencies and lowering building operating expenses [4].

To implement LEED-EB effectively, designers must perform comprehensive sustainability assessments. Recent studies highlight BIM as a tool that can enhance LEED evaluations, streamline documentation, and reduce time and cost. Techniques include third-party software, cloud-based BIM, and API plug-ins, or combining BIM with optimization and machine learning to improve LEED credits and energy efficiency. These examples are provided for context; BIM integration is not part of this study. [5]

Additionally, it is required that any project seeking LEED certification include a minimum level of commissioning. A LEED-accredited professional often guides the project team through a re-commissioning management guide to ensure sustained performance over time emphasized this requirement to safeguard against performance decay after initial certification. Moreover, LEED-EB version 4.1 has introduced performance-based re-assessment and offers bonus points for buildings previously certified under other LEED categories, encouraging continuous performance compliance [6].

1.2.1 challenges of implementing leed in saudi arabia

1. **Lack of Expertise and Multidisciplinary Skills**

A major barrier to the implementation of green building projects is the limited availability of professionals with expertise in sustainable construction. Forming project teams with multidisciplinary knowledge of sustainability remains difficult.

2. **High Initial Costs**

Developers often identify the higher upfront costs of sustainable materials and green technologies, compared to conventional alternatives, as a significant barrier to implementing green building projects.

3. **Harsh Desert Climate and Extreme Weather Conditions**

High heat, low humidity, and frequent sandstorms affect both building performance and maintenance, thus requiring materials and designs that tolerate severe environmental conditions.

4. **Water Scarcity and Water Sustainability**

Given the Kingdom's limited water resources, green building projects face added pressure to apply water saving systems, efficient irrigation, and technologies that reduce water usage.

5. **Weak or Insufficient Regulatory Frameworks**

Although interest in sustainability is growing, there are still gaps in national legislation or building codes that mandate sustainable building practices, resulting in inconsistent adoption.

6. **Limited Awareness and Professional Education**

Many architects, engineers, and contractors are not fully familiar with LEED requirements or the benefits of green building, which hinders wider uptake across the building sector.

7. **Complex Documentation and Administrative Burden**

The process of gathering required documentation for LEED certification is often complex and time consuming, discouraging some teams from fully committing to the certification process [7].

1.2.2 advantages & opportunities of adopting leed in Saudi Arabia

1. **Pioneering Achievements Such as KAPSARC**

The King Abdullah Petroleum Studies and Research Center (KAPSARC) in Riyadh has achieved five LEED-EBOM Platinum certifications for its existing buildings, totaling approximately 52,284 square meters, establishing it as a local benchmark in sustainable building performance [8].

2. **Enhanced Environmental Reputation and Investment Attraction**
   LEED certified buildings provide a marketing edge and attract international investors who prioritize environmentally responsible projects.
3. **Growing Government Support and Alignment with Vision 2030**
   The objectives of Vision 2030 align well with promoting sustainability in the built environment, opening up opportunities for incentives, facilitation, and supportive policies for green projects [1].
4. **Long Term Operational Cost Savings**
   Buildings that meet high sustainability standards tend to use less energy and water, resulting in significant reductions in operating costs over their lifecycle.
5. **Boosting Innovation in Green Technologies and Materials**
   LEEDEB adoption drives demand for advanced materials and technologies, which encourages local industries to invest in research and development in sustainability.

1.2.3 practical suggestions to strengthen leed implementation in Saudi Arabia

1. **Professional Training and Workshops**

Organize specialized training sessions and workshops for architects, engineers, and contractors to deepen their understanding of LEED standards and sustainable design practices. World Business Council for Sustainable Development (WBCSD)

2. **Financial Incentives**

Provide government support, tax reductions, or subsidies to projects aiming for LEED certification to lower the financial burden of initial investments.

3. **Development of Local Guidelines for LEED in Saudi Conditions**

Create guiding documents and best practices that account for desert climate, water scarcity, and regional material availability to ensure consistent application across all regions.

4. **Integration of Modern Technologies and Smart Monitoring Systems**

Employ sensors and real time performance monitoring tools to ensure continual compliance with LEED requirements and assess building behavior after occupancy.

5. **Strengthening Laws and Regulations**

Embed sustainability criteria into the Saudi Building Code and consider mandating certain green certifications or standards, especially in government-initiated projects.

6. **Showcase or Pilot Projects**

Establish pilot or showcase green building projects in major cities to demonstrate the economic, environmental, and technical feasibility of LEED, serving as examples for wider adoption.

1.3 Terminologies “green versus sustainable”

While the terms “green” and “sustainable” are often used interchangeably, they carry distinct meanings, especially in the context of environmental design and construction. A “green building” typically refers to structures designed to reduce environmental impact by improving energy efficiency, conserving water, and utilizing ecofriendly materials. The focus is primarily on minimizing harm to the environment throughout the building's lifecycle from design to operation [9].

On the other hand, “sustainability” is a broader and more holistic concept that extends beyond environmental concerns to include economic viability and social responsibility**.** Sustainable buildings these are initiatives that not only minimize environmental impact but also generate long-term economic value and improve the quality of life for both occupants and surrounding communities. In essence, all sustainable buildings are green, but not all green buildings can be considered truly sustainable [10].

As Saudi Arabia moves forward with its Vision 2030**,** understanding the difference between “green” and “sustainable” becomes increasingly important. Integrating both approaches into the construction and development sectors is vital Aiming to create buildings that are environmentally responsible, economically viable, and socially beneficial, thereby contributing to a more resilient and future-ready Kingdom [1].

Theory #1:

Figure 1 illustrates the distinctions between commonly used terms, focusing on energy consumption and environmental impact regenerative design represents the highest sustainability level, followed by reconciliation and restorative approaches, while sustainability reflects a neutral environmental impact. Green practices achieve relative improvement through rating systems such as LEED, whereas conventional practice is limited to legal compliance. The progression of environmentally conscious design, as shown in the figure, follows an encouraging upward trajectory moving from traditional design to green design, and ultimately toward fully restorative and regenerative approaches. In this context, adopting green practices can be seen as an important step on the path to achieving true sustainability [11].

The distinctions between commonly used terms in environmentally conscious design, such as “green” and “sustainable,” can be understood through their relative impacts on energy consumption and

environmental footprint. The evolution from traditional building practices to green design, and ultimately to sustainable and regenerative design, represents a progressive improvement in minimizing ecological harm and enhancing resource efficiency. Green building is often considered an important step toward sustainability, focusing on reducing energy and water use, minimizing site disturbances, and improving indoor environmental quality compared to conventional buildings. This includes careful consideration of lifecycle impacts, from material selection and construction to operation and end of life reuse or disposal. Sustainability, however, encompasses a broader philosophy, emphasizing the balance between environmental stewardship, economic viability, and social wellbeing to ensure that future generations can enjoy a comparable quality of life It involves not only using resources efficiently but also creating resilient systems and communities that can thrive over the long term. Therefore, while all green buildings aim to reduce environmental impacts, truly sustainable projects integrate economic and social dimensions to support lasting benefits for people and the planet [12].

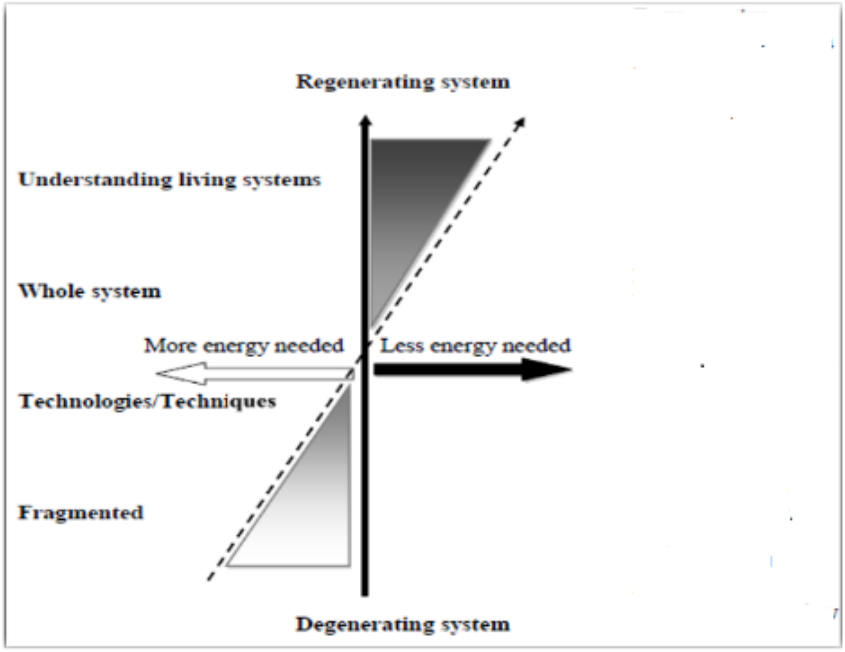


Fig. 1. Illustrates The Differences Between Commonly Used Terminologies Based On Energy Consumption and Environmental Impact. [11]

Theory #2:

From an alternative perspective, the relationship between sustainability and green design is illustrated in Figure 2. In this depiction, sustainability encompasses environmentally stable, technologically advanced, materially efficient, and ecologically responsible building design. Conversely, the term "green" is viewed more abstractly, integrating concepts of ecology, performance, and sustainability [13]. Within this framework, ecology refers to how a building interacts with and maintains harmony with the natural environment. It is possible, therefore, for a building to be sustainable and stable yet not necessarily "green" if it exhibits poor environmental performance or negative ecological impact. Similarly, a building may demonstrate high performance but lack ecological balance or long term stability, underscoring the nuanced distinction between these concepts. Devoid of stability, ecology, etc. [14].

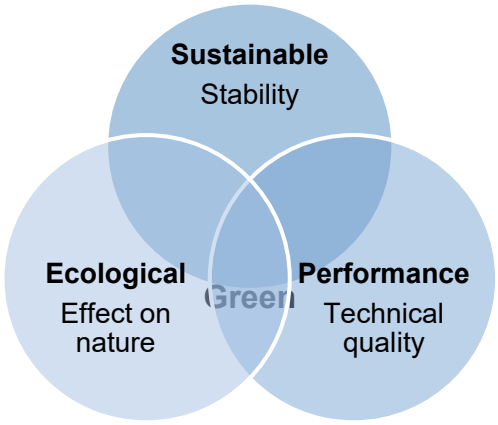


Fig. 2. Relationship Between Green, Sustainability, Ecology, And Performance [13]

2. Research Objective

The primary objective of this paper is to evaluate selected construction projects in the Kingdom of Saudi Arabia that have obtained LEED certification, with a particular focus on identifying the weaknesses that affect their scoring. To date, numerous construction projects in major cities such as Riyadh, Jeddah, and Dammam have been awarded LEED certifications, reflecting the Kingdom's growing commitment to sustainable building practices. Figure 3, illustrates the geographic distribution of these certified projects across various regions of the Kingdom, highlighting the progress and challenges in implementing green building standards nationwide.

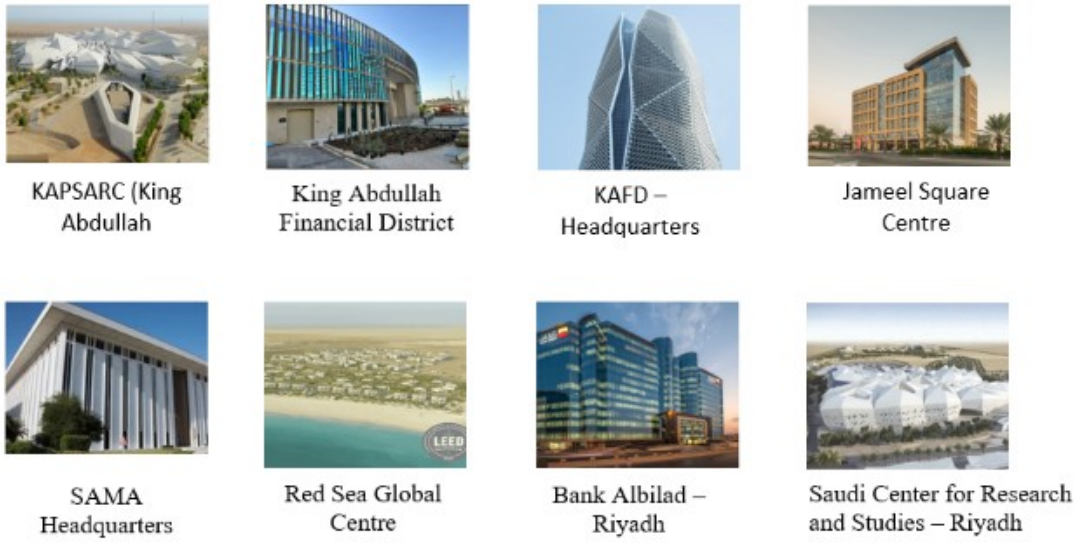


Fig. 3. Several Construction Projects in The Kingdom of Saudi Arabia Have Achieved LEED Certification (All images sourced from the Internet and official building websites.)

3. Methodology

Figure 4 illustrates the detailed research methodology applied in this study, which consists of five main steps. First, a selection of construction projects in Saudi Arabia was identified based on their LEED certification status, and relevant data, including building type, size, and LEED scorecards, were collected to form the study dataset. Next, the collected data were examined for completeness and consistency, and preliminary assessments were conducted to evaluate project compliance with LEED criteria across all relevant categories. Each project was then scored according to the LEED rating system, and scores were categorized by performance areas (e.g., Energy & Atmosphere, Sustainable Sites, Water Efficiency) to facilitate comparative analysis. Statistical analysis was applied to the evaluation results to determine overall project performance, identify trends, and quantify strengths and weaknesses in each LEED category. Finally, the findings were summarized, gaps and recurring weaknesses in current project practices were highlighted, and practical recommendations were proposed to improve the accuracy of the scoring system and enhance the implementation of sustainable strategies in future projects.

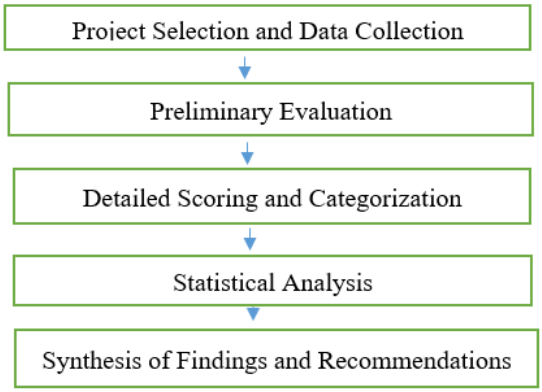


Fig. 4. Research Methodology

Statistics show a significant variation among Middle Eastern countries in the number of projects certified under the **LEED** (Leadership in Energy and Environmental Design) system as show in figure 5. The Kingdom of Saudi Arabia leads with 1,029 certified projects, reflecting its growing commitment to sustainability and alignment with the goals of Saudi Vision 2030. It is followed by the United Arab Emirates, which has 749 certified projects, benefiting from an early adoption of green building standards, particularly in Dubai and Abu Dhabi. Qatar ranks third with 162 projects, many of which are part of modern infrastructure and sustainable stadium developments. Egypt has achieved 46 certified projects, amid growing efforts to implement sustainability standards in national development projects. Meanwhile, Kuwait has 23 certified projects, indicating room for further expansion in the prospective implementation of green building practices [15].

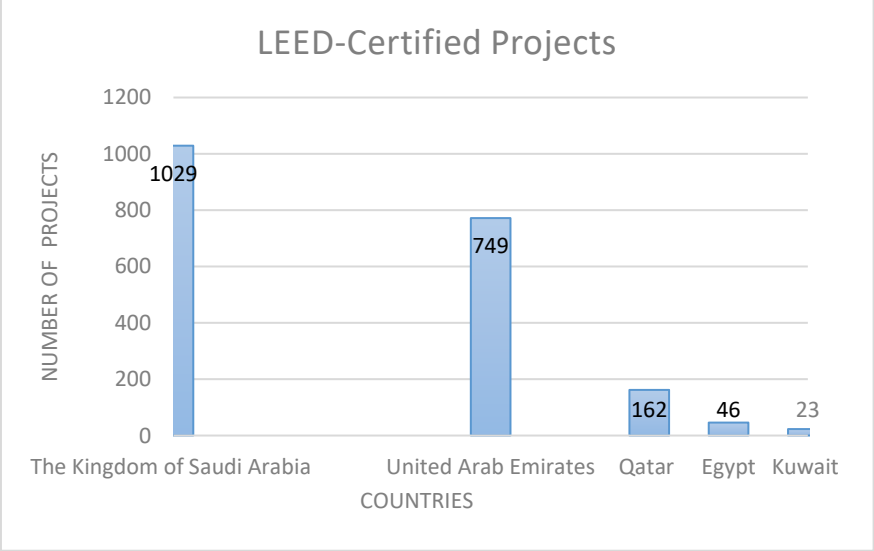


Fig. 5. Comparison Between Some Middle Eastern Countries to Obtain LEED

As show in figure 6, the number of projects certified under the LEED (Leadership in Energy and Environmental Design) system in the Kingdom of Saudi Arabia has reached 1,029 projects, reflecting the country's growing commitment to sustainability in the construction and urban development sectors.

These certifications are distributed across various levels: 240 projects achieved Platinum certification, the highest level, 174 received Gold, 569 were awarded Silver, and 46 projects obtained basic Certification [15]. This research aims to increase awareness and understanding of the LEED system and its importance in promoting energy efficiency and environmental sustainability, as well as to encourage the development of more certified projects in alignment with Saudi Vision 2030, which seeks to build a sustainable future and achieve comprehensive development across all sectors.

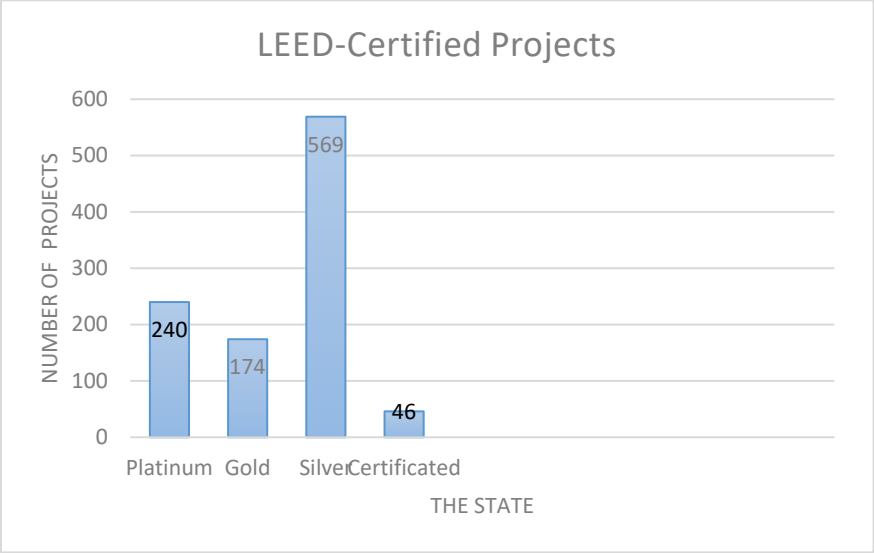


Fig. 6. Classification of LEED Certified Projects in The Kingdom of Saudi Arabia.

This study will analyze the projects certified under LEED in the Kingdom of Saudi Arabia, aiming to gain a deeper understanding of these projects and to develop recommendations that contribute to improving environmental performance and future sustainability of buildings. The projects have been classified into three main categories: Gold, Silver, and Certified, with the objective of identifying shortcomings and challenges within each category, and striving to achieve the higher. The selected projects were chosen randomly, based on the availability of necessary data for analysis. Detailed information about these projects is presented in the following tables, which serve as the foundation for the analysis and conclusions of this study.

As shown in the following tables, Table 2 includes projects with LEED Certified status, presenting individual scores for key sustainability categories such as Energy and Atmosphere, Materials and Resources, Indoor Environmental Quality, Sustainable Sites, Water Efficiency, and Innovation in Design, highlighting strengths and areas for improvement. Table 3 presents projects that attained LEED Silver certification, with a breakdown of scores in the same categories to assess environmental performance. Table 4 showcases projects that achieved LEED Gold certification, detailing scores across all major categories to emphasize achievements and potential improvements. Additionally,

Table 5 presents fifteen LEED Gold projects, Table 7 shows eight LEED Silver projects, and Table 9 includes seven LEED Certified projects, all with detailed category scores. This presentation provides a clear evaluation of environmental performance and supports strategies to enhance sustainability outcomes and achieve higher LEED ratings [15].

Table 2: Projects Certified LEED Gold in The Kingdom of Saudi Arabia and Total Score Achieved

#	**Project & Address**	**Project Certification Level**	**Scoring**
1	**KAAUH Riyadh # Airport Road, Riyadh, Saudi Arabia, 11552**	Gold	60
2	**4HP9R9R Al Khams's Rd Road Fanateer, Al Jubal, Saudi Arabia, 35811**	Gold	70
3	**VOX Cinema sin the Esplanade Mall The Esplanade Mall, Riyadh, Saudi Arabia, 12371**	Gold	69
4	**VOX Cinema sin The Roof in Riyadh King Abdul-Aziz Rd Al Burj, Riyadh, Saudi Arabia, 18278**	Gold	69
5	**DAR Engineering Al Olay a, Riyadh, Saudi Arabia, 11642**	Gold	62
6	**Dr Suleiman Al Habib Al Qassim Al Safra Unit1 Buraydah, Al Qassim, Saudi Arabia, 52381**	Gold	67
7	**Dr Suleiman Al Habib Al Khobar King Fahad Bin Abdul-Aziz Road Banda, Al Khobar, Saudi Arabia, 34423**	Gold	65
8	**DrSulaimanAlHabibAlOlayaWadiHanifaOlayaUnit42Riy, Riyadh, Saudi Arabia, 12214**	Gold	62
9	**VOX Cinema sin Roshn Front Riyadh, Riyadh, Saudi Arabia,**	Gold	60
10	**Olaya Towers by Osool –Al Olaya Street, Riyadh, Saudi Arabia, 12213**	Gold	71
11	**Jamel Square Prince Mohammed Bin Abdul-Aziz St Al and, Jeddah, Saudi Arabia, 23326**	Gold	63
12	**Bank Al Bilad Tower Al Mutamarat 8229 Building No23952, Riyadh, Saudi Arabia, 12711**	Gold	60
13	**Mon shaat Head Office 7739OlayaSt Al Ghadir247013311https:mon shaatgovsa, Riyadh, Saudi**	Gold	67
14	**Saudi Electricity Company HQ Abha 68303129Almahalah Abha62561, Abha, Saudi Arabia, 62561**	Gold	72
15	**King Abdul-Aziz Center for World Culture Dhahran Jubail Highway Saudi Arabia, Dhahran, Saudi Arabia, 31311**	Gold	72

Table 3: Projects Certified LEED Silver in The Kingdom of Saudi Arabia and Total Score Achieved

#	Project & Address	Project Certification Level	Scoring
1	**NEOM SATCO Layd own office Wade Ainounah Neom, Neom, Saudi Arabia, 49637**	silver	51
2	**Panda202 ArRawdah District, Jeddah, Saudi Arabia, 23434**	silver	59
3	**KAUST RP Mosque 4700 King Abdullah University, Jeddah Thuwal, Saudi Arabia, 23955**	silver	55
4	**IPSO911Riyadh Storage Building Salboukh Road, Riyadh, Saudi Arabia**	silver	50
5	**KFMC Rehabilitation Hospital Al Sulay maneyya Riyadh, Riyadh, Saudi Arabia, 11525**	silver	55
6	**Asharqia Chamber Tower Plotnumber6DBlocknumber1Masterp, Dammam, Saudi Arabia,**	silver	50
7	**U.S. Consulate General Dhahran 889 Salah Elden Al Ayoubi St, Al Andalus, Al Khobar, Saudi Arabia, 31261**	silver	55
8	**KAUST RP Administration Building 4700 King Abdullah University, Jeddah Thuwal, Saudi Arabia, 23955**	silver	52

Table 4: Projects Certified LEED The Kingdom of Saudi Arabia and Total Score Achieved

#	Project & Address	Project Certification Level	Scoring
1	**Al Woroud City Admin Building Al Sail Road Taif, Taif, Saudi Arabia, 26311**	certified	49
2	**The VUE Obhor, Jeddah, Saudi Arabia, 23734**	certified	41

3	**Ministry of Environment Water and Agric. Ministry of Environment Water and Agric., Riyadh, Saudi Arabia,**	certified	47
4	**Diyar Al Salam Jeddah, Jeddah, Saudi Arabia, 21411**	certified	44
5	**NGHA Jeddah Neuroscience Hospital NGHA Jeddah Campus, Jeddah, Saudi Arabia,**	certified	48
6	**SABIC Academy Building Saudi Arabia, Riyadh, Saudi Arabia,**	certified	46
7	**Tamer Group Main Distribution Center Kumar Area, King Faisal Road, Near Old Check Point, Jeddah, Saudi**	certified	47

The LEED rating system is one of the most prominent global standards for assessing sustainable buildings. It is used to measure how well projects adhere to environmental sustainability criteria. Certifications are awarded based on the total number of points a project earns through the implementation of sustainable design and construction strategies. LEED certification is divided into four main levels:

- Certified: Awarded to projects that earn between 40 and 49 points.
- Silver: Awarded to projects that earn between 50 and 59 points.
- Gold: Awarded to projects that earn between 60 and 79 points.
- Platinum: The highest level, awarded to projects that earn 80 points or more [15].

These certification levels reflect the building's commitment to environmentally responsible practices, encouraging reduced environmental impact, improved resource efficiency, and the creation of healthy and sustainable indoor environments.

The LEED rating system evaluates projects based on a set of fundamental environmental criteria, which are divided into several main categories. Projects earn points in each category, and the total number of points determines the final level of certification. These categories include:

- Sustainable Sites: Assesses site selection and development strategies that minimize environmental impact, such as preserving natural habitats and reducing pollution.
- Water Efficiency: Focuses on methods and technologies that reduce water consumption both indoors and outdoors.
- Energy and Atmosphere: Addresses energy efficiency, the use of renewable energy sources, and the optimization of heating, cooling, and ventilation systems.

- Materials and Resources: Evaluates the use of sustainable, recycled, and regionally sourced materials, as well as waste reduction practices during construction.
- Indoor Environmental Quality: Assesses indoor air quality, lighting, thermal comfort, and acoustic performance to promote healthier indoor environments.
- Innovation in Design: Awards additional points to projects that implement innovative strategies exceeding standard sustainability practices.
- Regional Priority: Provides extra points for addressing specific environmental concerns relevant to the project's geographic location.

Points earned across all categories are totaled to determine the final certification level, encouraging designers and developers to adopt integrated solutions that enhance long term environmental sustainability.

As shown in the following tables, Table 6 focuses on projects that achieved LEED Gold certification, showing each item's score as a percentage of the total possible points, which highlights the categories that contribute most to Gold-level certification and indicates areas with the greatest impact on sustainability performance. Table 8 presents the relative scores for projects with LEED Silver certification, allowing for comparison across categories to identify both strengths and areas that require improvement to meet Silver-level requirements. Table 10 illustrates the relative scores for projects with LEED Certified status, providing insight into the most influential categories at the Certified level and supporting targeted strategies to enhance environmental performance.

Table 5: Presents The Fifteen Projects That Have Achieved LEED Gold Certification, along with The Separated Scores for Each Individual Scoring Category

#	Project Adress	Project Certification Level	scoring	SUSTAINABLE SITES (4)	WATER EFFICIENCY (15)	ENERGY & ATMOSPHERE (35)	MATERIAL & RESOURCES (9)	INDOOR ENVIRONMENTAL QUALITY (22)	INNOVATION (1)	LOCATION & TRANSPORTATION (14)
1	KAAUH-Riyadh # Airport Road, Riyadh, Saudi Arabia, 11552	Gold	60	0	6	20	7	15	0	12
2	4HP9R9RAlKhameesRdRoadFanateer, AlJubayl, Saudi Arabia, 35811	Gold	70	0	9	21	7	20	1	12
3	VOXCinemasintheEsplanadeMall-TheEsplanadeMall, Riyadh, Saudi Arabia, 12371	Gold	69	0	8	19	7	22	1	12
4	VOXCinemasinTheRoofinRiyadh- KingAbdulazizRdAlBurj, Riyadh, Saudi Arabia, 18278	Gold	69	0	9	17	7	22	1	13
5	DAREngineering-AlOlaya, Riyadh, Saudi Arabia, 11642	Gold	62	0	9	24	7	19	0	10
6	nanAlHabibAlQassim-AlSafraUnit1Buraydah, AlQassim, Saudi Arabi	Gold	67	0	9	25	4	17	1	11
7	DrSulaimanAlHabibAlKhobar-KingFahadBinAbdulazizRoadBanda, AlKhobar, Saudi Arabia, 34423	Gold	65	0	11	19	5	17	1	12
8	anAlHabibAlOlaya-WadiHanifaOlayaUnit42Riy, Riyadh, Saudi Arabi	Gold	62	0	7	21	8	16	1	9
9	VOXCinemasinRoshnFront - Riyadh, Riyadh, Saudi Arabia, 00000	Gold	60	0	11	14	5	20	1	9
10	OlayaTowersByOsool -AlOlayaStreet, Riyadh, Saudi Arabia, 12213	Gold	71	0	10	28	5	15	1	12
11	JameelSquare Address: PrinceMohammedBinAbdulazizStAlAnd, Jeddah, Saudi Arabia, 23326	Gold	63	0	9	19	6	18	1	10
12	BankAlBiladTower Address: AlMutamarat8229BuildingNo23952, Riyadh, Saudi Arabia, 12711	Gold	60	1	9	21	6	12	0	11
13	MonshaatHeadOffice Address: 7739OlayaStAlGhadir2470133111https:monshaatgovsa, Riyadh, Saudi	Gold	67	1	12	22	6	15	0	11
14	SaudiElectricityCompanyHQAbha Address: 68303129AlmahalahAbha62561, Abha, Saudi Arabia, 62561	Gold	72	0	11	24	7	17	1	12
15	KingAbdulazizCenterforWorldCulture Address: DhahranJubailHighwaySaudiArabia, Dhahran, Saudi Arabia, 31311	Gold	72	4	11	16	8	20	1	12

Table 6: The Relative Score of Each Scoring Item of LEED Gold

	Sustainable Sites	**Water Efficiency**	**Energy & Atmosphere**	**Material & Resources**	**Indoor Environmental Quality**	**Innovation**	**Location & Transportation**
sum	6	141	310	95	265	11	168
total sum	60	225	525	135	330	15	210
relative score	0.1	0.6266	0.5905	0.7037	0.8030	0.7333	0.8

Table 7: Presents The Eight Projects That Have Achieved LEED Silver Certification, along with The Separated Scores for Each Individual Scoring Category

#	Project Adress	Project Certification Level	scoring	LOCATION & TRANSPORTATION (14)	SUSTAINABLE SITES (4)	WATER EFFICIENCY (15)	ENERGY & ATMOSPHERE (35)	MATERIAL & RESOURCES (9)	INDOOR ENVIRONMENTAL QUALITY (22)	INNOVATION (1)
1	NEOMSATCOLayd own office WadiAinounahNeom, Neom, Saudi Arabia, 49637	silver	51	8	0	12	16	5	9	1
2	Panda202 ArRawdahDistrict, Jeddah, Saudi Arabia, 23434	silver	59	10	0	10	17	3	19	0
3	KAUST RP Mosque 4700 King Abdullah University, Jeddah-Thuwal, Saudi Arabia, 23955	silver	55	11	1	9	16	6	11	1
4	IPSO911Riyadh Storage Building SalboukhRoad, Riyadh, Saudi Arabia	silver	50	9	0	11	14	5	10	1
5	KFMC Rehabilitation Hospital AlSulaymaneyyaRiyadh, Riyadh, Saudi Arabia, 11525	silver	55	14	1	8	15	6	11	0
6	Asharqia Chamber Tower Plotnumber6DBlocknumber1Masterp, Dammam, Saudi Arabia,	silver	50	9	0	12	14	6	8	1
7	U.S. Consulate General Dhahran 889 Salaheddin Al Ayoubi St, Al Andalus, Al Khobar, Saudi Arabia, 31261	silver	55	10	1	10	13	7	13	1
8	KAUST RP Administration Building 4700 King Abdullah University, Jeddah-Thuwal, Saudi Arabia, 23955	silver	52	11	0	9	12	6	13	1

Table 8: The Relative Score of Each Scoring Item of LEED Silver

	Sustainable Sites	**Water Efficiency**	**Energy & Atmosphere**	**Material & Resources**	**Indoor Environmental Quality**	**Innovation**	**Location & Transportation**
Sum	3	81	117	44	94	6	82
Total sum	32	120	280	72	176	8	112
Relative score	0.09375	0.675	0.4179	0.6111	0.5341	0.75	0.7321

Table 9: Presents The Seven Projects That Have Achieved LEED Certificated Only, along with The Separated Scores for Each Individual Scoring Category

#	Project Adress	Project Certification Level	scoring	location & transportation (20)	sustainable sites (10)	water efficiency (11)	Eenergy & atmosphere (20)	material & resources (13)	indoor environmental quality (16)	Regional priority credits (4)	innovation (6)
1	Al Woroud City Admin Building AlSailRoadTaif, Taif, Saudi Arabia, 26311	certified	49	3	1	9	18	0	8	4	6
2	The VUE Obhor, Jeddah, Saudi Arabia, 23734	certified	41	5	2	5	7	5	9	3	5
3	Ministry of Environment Water and Agric Ministry of Environment Water andAgric, Riyadh, Saudi Arabia,	certified	47	12	3	5	13	4	6	2	2
4	Diyar Al Salam Jeddah, Jeddah, Saudi Arabia, 21411	certified	44	13	7	2	3	5	8	3	3
5	NGHA - Jeddah Neuroscience Hospital NGHA Jeddah Campus, Jeddah, Saudi Arabia,	certified	48	5	5	8	9	5	11	4	1
6	SABIC Academy Building Saudi Arabia, Riyadh, Saudi Arabia,	certified	46	5	5	8	11	1	7	5	4
7	Tamer Group - Main Distribution Center Kohmra Area, King Faisal Road, Near Old Check Point, Jeddah, Saudi	certified	47	7	7	7	11	1	5	5	4

Table 10: The Relative Score of Each Scoring Item of LEED Certificated

	Sustainable Sites	**Water Efficiency**	**Energy & Atmosphere**	**Material & Resources**	**Indoor Environmental Quality**	**Innovation**	**Location & Transportation**
Ssum	30	44	72	21	54	25	50
Total sum	70	77	140	91	112	42	140
Relative score	0.4286	0.5714	0.5143	0.2308	0.4822	0.5953	0.3572

4. Analysis of result

4.1. Analysis of performance results for leed gold certified projects

The analysis of the relative scores across the fifteen LEED Gold certified construction projects reveals notable variations in performance across the different sustainability categories. The relative score for each category is as follows: Sustainable Sites: 6/60 = 0.10, Water Efficiency: 141/225 = 0.63, Energy & Atmosphere: 310/525 = 0.59, Materials & Resources: 95/135 = 0.70, Indoor Environmental Quality: 265/330 = 0.80, Innovation: 11/15 = 0.73, Location & Transportation: 168/210 = 0.80. These results indicate strong performance in categories such as Indoor Environmental Quality, Location & Transportation, and Innovation, each achieving over 70% of the possible points. The Materials & Resources and Water Efficiency categories also demonstrate solid performance, with scores above 60%.

However, the Sustainable Sites category shows a significant underperformance, achieving only 10% of the total possible points (6 out of 60), the low score (10% for Gold projects) in the Sustainable Sites category is mainly due to Saudi Arabia's arid climate, scarce vegetation, and water scarcity, which limit landscaping and irrigation options. Urban density and site constraints further restrict sustainable site practices. LEED criteria, designed for temperate regions, are not fully adapted to desert conditions. Locally tailored strategies, such as drought-resistant plants, water reuse, and reduced impervious surfaces, are needed to improve performance in this category, supporting Vision 2030 sustainability goals, which heavily affects the overall certification score.

Additionally, the Energy & Atmosphere category, despite being a critical aspect of environmental performance, scored only 59% for gold projects, indicating room for improvement. This relatively low score is primarily attributed to Saudi Arabia's hot and arid climate, which significantly increases cooling demands, conventional building designs that limit the use of passive energy strategies, and the limited adoption of renewable energy systems. Overall, the cumulative performance across all categories did not exceed 50% of the total possible points, highlighting the need for strategic interventions, particularly in site selection and energy management. Addressing these weaknesses can enhance overall environmental efficiency and increase the potential for achieving higher LEED certification levels in future projects

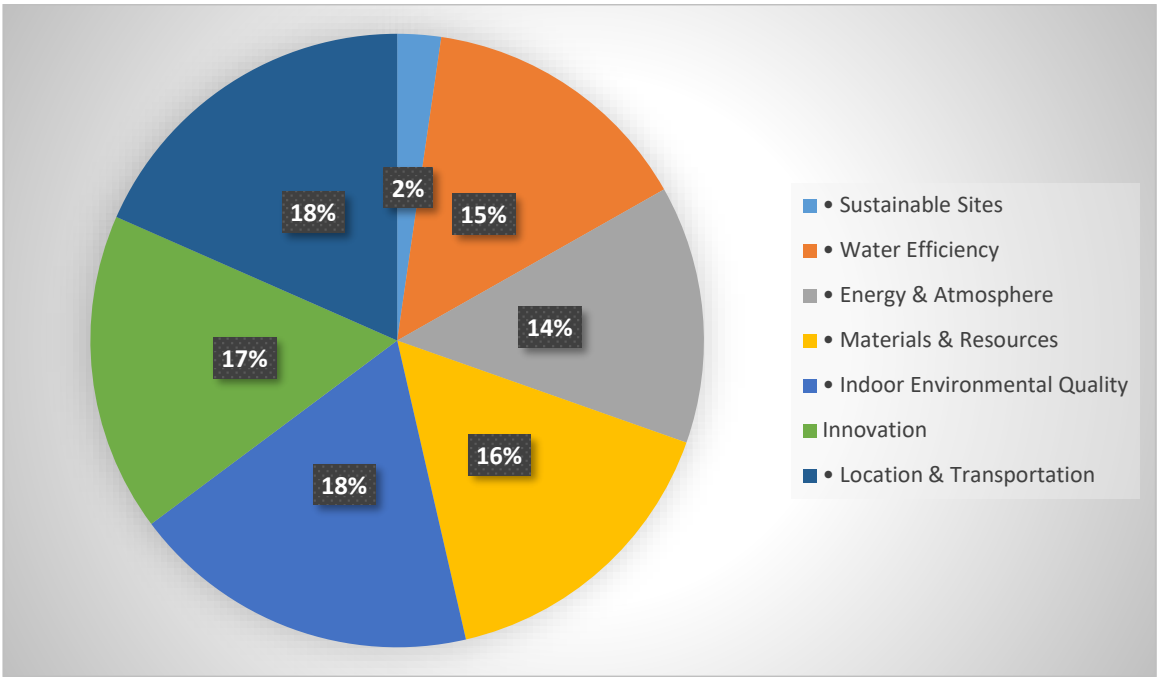


Fig. 7. Scoring Leeds Items Percentage for LEED Gold Certified Projects.

Figure 7 illustrates the percentage distribution of LEED scoring across various sustainability categories for projects that achieved LEED Gold certification. The breakdown is as follows: Sustainable Sites: 2.3% Water Efficiency: 14.5% Energy & Atmosphere: 13.6% Materials & Resources: 16% Indoor Environmental Quality: 18.4% Innovation: 16.4% Location & Transportation: 18.4% This distribution reveals a clear imbalance in score contribution among the evaluated categories. Notably, Sustainable Sites received the lowest percentage (2.3%), which indicates either underperformance in site selection and ecological integration or a strategic reprioritization of this category in design and planning phases. In contrast, Indoor Environmental Quality and Location & Transportation both scored the highest at 18.4%, reflecting strong attention to occupant health, comfort, connectivity, and access to sustainable mobility. The relatively high scores in Innovation (16.4%) and Materials & Resources (16%) suggest that these projects incorporate advanced construction strategies and place value on responsible material usage, potentially including recycled and locally sourced components. However, the moderate score in Energy & Atmosphere (13.6%), which is typically one of the most critical categories in determining a building's sustainability performance, points to room for improvement in energy efficiency measures, renewable integration, and overall carbon footprint reduction. This aligns with sustainable design principles discussed in literature. For instance, [12] highlights the importance of practical, performance driven housing design that incorporates tight construction, high insulation, controlled ventilation, and energy recovery systems all of which contribute to lower carbon emissions and enhanced energy efficiency. Such strategies are not only effective but also accessible and replicable, allowing widespread adoption of sustainable building practices.

4.2. Analysis of performance results for leed silver certified projects

An evaluation of the relative scores for the eight construction projects certified under the LEED Silver level highlights noticeable disparities in performance across the various sustainability categories. The relative score for each category is as follows: Sustainable Sites: 3/32 = 0.09, Water Efficiency: 81/120 = 0.675, Energy & Atmosphere: 117/280 = 0.42, Materials & Resources: 44/72 = 0.61, Indoor Environmental Quality: 94/176 = 0.54, Innovation: 6/8 = 0.75, Location & Transportation: 82/112 = 0.73

Key Observations

Sustainable Sites (0.09)

This category demonstrates the lowest relative score (9%), indicating a major weakness in site planning and ecological considerations. This underperformance may result from poor integration of sustainable land use practices, minimal efforts to mitigate site disruption, or lack of prioritization during project design.

Water Efficiency (0.675) and Materials & Resources (0.61)

These two categories show strong performance, reflecting effective implementation of water saving fixtures, landscape strategies, and responsible material selection potentially including recycled and locally sourced materials.

Location & Transportation (0.73) and Innovation (0.75)

These are the topper forming categories, suggesting that the projects focused on connectivity, walkability, public transportation access, and also incorporated creative or forward thinking sustainability strategies that exceed standard LEED requirements.

Energy & Atmosphere (0.42)

Despite being one of the most critical categories in environmental performance, this score is relatively low, revealing opportunities for improvement in energy optimization, renewable energy integration, and HVAC system efficiency.

Indoor Environmental Quality (0.54)

This is a moderate score, indicating some effort to ensure thermal comfort, air quality, and lighting control. However, there is still room for enhancement to meet higher occupant wellbeing standards.

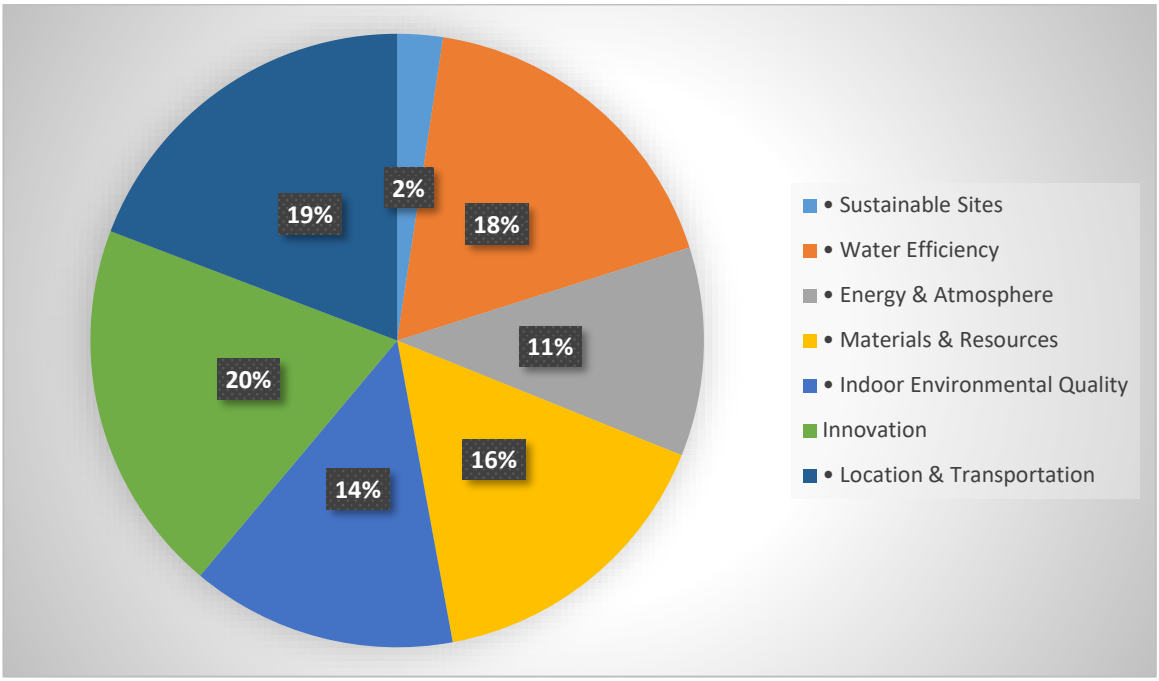


Fig. 8. Scoring Leeds items percentage for LEED Silver Certified Projects

Figure (8) illustrates the relative distribution of scores across the various sustainability categories for projects certified under LEED Silver. The highest percentages were recorded in the categories of Innovation (19.7%) and Location & Transportation (19.2%), indicating a strong focus on integrating creative solutions and effective urban connectivity that supports sustainable mobility and access to public transportation. This success reflects the adoption of design strategies that exceed standard LEED requirements and can be leveraged to enhance performance in other categories. In contrast, moderate performance was observed in Indoor Environmental Quality (14%) and Water Efficiency (17.7%), suggesting a reasonable level of attention to indoor comfort, air quality, and water saving technologies. However, there remains room for improvement in natural ventilation, the use of high efficiency plumbing fixtures, and water efficient irrigation systems. The Energy & Atmosphere category scored just 11%, which is relatively low given its critical role in reducing carbon emissions and improving energy efficiency. This weakness highlights the need to strengthen the integration of renewable energy technologies, as well as the performance of HVAC systems and thermal insulation. On the other hand, the Sustainable Sites category recorded the lowest score at 2.4%, indicating significant shortcomings in site planning and selection regarding environmental impact mitigation and resource preservation. Therefore, it is essential to prioritize site choices that protect ecosystems, reduce urban pollution, and support sustainable land development practices in future projects. Overall, the score distribution demonstrates clear variation in performance across the categories, emphasizing the importance of

adopting a balanced approach that addresses areas of weakness particularly in energy and site sustainability while maintaining excellence in innovation and urban connectivity.

4.3. Analysis of performance results for leed certified projects

The relative score evaluation of seven construction projects certified at the LEED Certified level reveals clear disparities in performance across various sustainability categories. The relative scores for each category are as follows: Sustainable Sites: 30/70 = 0.43, Water Efficiency: 44/77 = 0.57, Energy & Atmosphere: 72/140 = 0.51, Materials & Resources: 21/91 = 0.23, Indoor Environmental Quality: 54/112 = 0.48, Innovation: 25/42 = 0.6, Location & Transportation: 50/140 = 0.36

Category Performance Overview

- **Materials & Resources (0.23):** This category has the lowest score, indicating weak implementation of sustainable material selection and recycling practices.
- **Location & Transportation (0.36):** A relatively low score, reflecting limited integration with public transport infrastructure or selection of sites that do not support urban sustainability.
- **Sustainable Sites (0.43):** A moderate performance level, indicating some environmental considerations in site planning, but without strong results in reducing environmental impact.
- **Indoor Environmental Quality (0.48):** Also a moderate score, suggesting partial implementation of air quality, thermal comfort, and indoor environmental standards.
- **Energy & Atmosphere (0.51):** This score reflects average performance, with some application of energy efficiency strategies, though lacking robust renewable energy use or advanced HVAC systems.
- **Water Efficiency (0.57):** A relatively strong category, showing clear attention to water conservation both indoors and outdoors.
- **Innovation (0.60):** The highest scoring category, reflecting the integration of creative, beyond standard LEED strategies within project designs

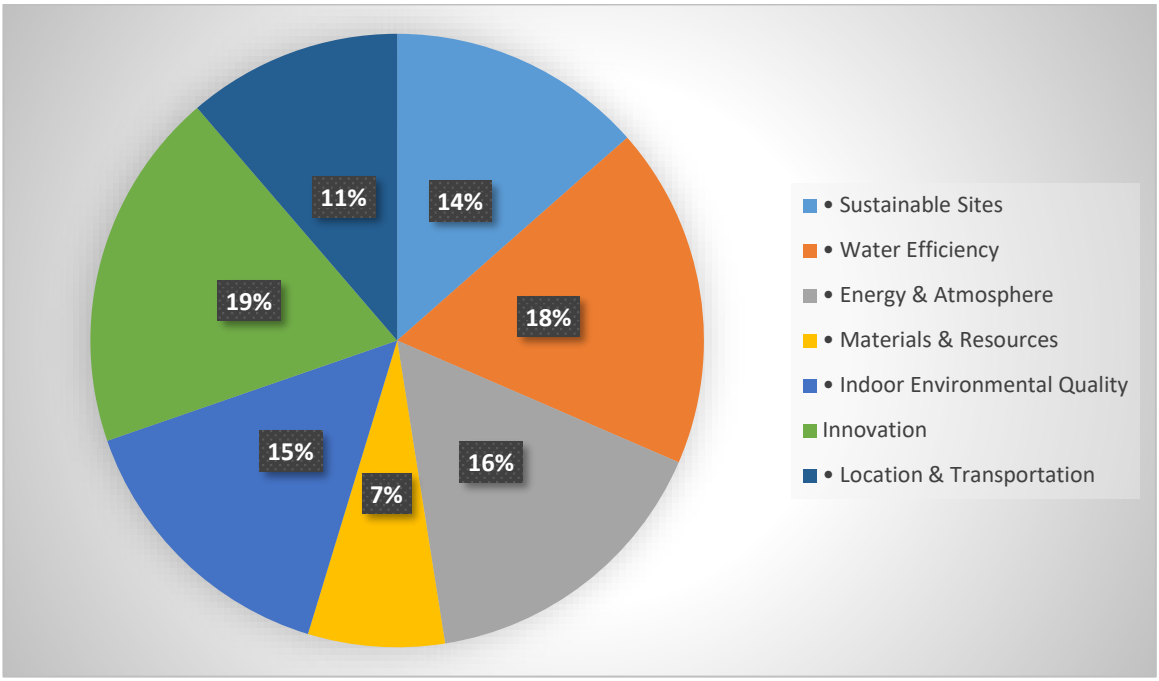


Fig. 9. Scoring Leeds Items Percentage for LEED Certified Projects

Figure 9 illustrates the relative percentage distribution of scores achieved across the sustainability categories for projects certified under LEED. The results are as follows: The Sustainable Sites category scored 13.5%, indicating a moderate level of attention to site planning, selection, and environmental impact reduction. Water Efficiency achieved a good percentage of 18%, reflecting the implementation of effective water saving strategies. The Energy & Atmosphere category scored 16%, a moderate level that shows efforts to improve energy efficiency and reduce emissions, though there is room for greater improvement in the use of renewable energy sources and HVAC systems. Materials & Resources recorded the lowest score at 7.2%, reflecting a clear weakness in the use of sustainable materials and waste management. Indoor Environmental Quality scored 15%, indicating noticeable efforts to enhance indoor air quality and thermal comfort. Innovation ranked highest at 19%, demonstrating that the projects adopt innovative solutions that go beyond traditional standards. Finally, Location & Transportation scored 11.3%, indicating a need to strengthen connectivity with sustainable transportation and improve accessibility. Overall, the distribution reveals disparities in performance among the categories, with clear strengths in innovation and water efficiency, and weaknesses in materials management and location & transportation, highlighting the need to adopt balanced strategies to enhance overall sustainability.

5. Conclusion

The analysis of LEED Gold certified projects showed varied performance across sustainability categories. Strong results appeared in Indoor Environmental Quality and Location and Transportation, while weaknesses were noted in Sustainable Sites and Energy and Atmosphere. This highlights the need for balanced improvement across all categories to achieve holistic sustainability and higher certification levels in the future.

Silver certified projects demonstrate clear strengths in Innovation and Location & Transportation. However, there is a noticeable weakness in the Energy & Atmosphere and Sustainable Sites categories. Strengthening performance in these areas while leveraging modern building science and design principles will be essential to advancing toward higher LEED certification levels such as Gold or Platinum

The results indicate that LEED Certified projects exhibit varied performance across sustainability categories, with notable strengths in Innovation and Water Efficiency, and significant weaknesses in Materials & Resources and Location & Transportation. This imbalance highlights the need for a more holistic and evenly distributed application of sustainable strategies to reach higher certification levels in future developments.

To improve the Sustainable Sites category, future projects should implement advanced storm water management, increase green and permeable areas, optimize site orientation for natural ventilation and daylight, preserve native vegetation, and enhance connectivity to public transportation.

In addition, improving Energy & Atmosphere, Materials & Resources, and other weaker categories alongside Sustainable Sites will help projects achieve higher LEED scores and contribute to more sustainable and resilient built environments

6. Recommendations for improvement

- Improving Site Planning: Focus on careful site selection and enhancing integration with the natural environment to reduce environmental impact.
- Enhancing Energy Efficiency: Strengthen thermal insulation and adopt renewable energy sources to improve performance in the Energy and Atmosphere category.
- Developing Water Conservation Systems: Continue adopting smart irrigation and water saving technologies to achieve higher performance in the Water Efficiency category.
- Promoting the Use of Sustainable Materials: Increase reliance on recycled and low carbon materials in the Materials and Resources category.
- Maintaining Indoor Environmental Quality: Continue prioritizing user comfort and indoor air quality in building design.
- Maintaining Integration with Sustainable Transportation: Continue selecting sites with easy access to public transportation.
- Leveraging Innovation: Utilize innovative solutions to enhance the performance of weaker categories and achieve overall balance in environmental performance

Funding

The author declares that no funds, grants, or other support were received during the preparation of this manuscript

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Environmental Sustainability and Vision 2030 Ministry of Economy and Planning. (2021). Vision 2030 programs and environmental sustainability. Government of Saudi Arabia. https://www.vision2030.gov.sa

[2] Environmental Impact of the Building Sector (Global Reference) United Nations Environment Programme. (2022). 2022 Global status report for buildings and construction: Towards a zero-emissions, efficient and resilient buildings and construction sector. UNEP. https://www.unep.org

[3] Transport Sector and Greenhouse Gas Emissions International Energy Agency. (2023). Transport. https://www.iea.org/topics/transport

[4] Zuo, J., Zhao, Z.-Y., & Zhao, X. (2023). LEED Certification in Building Energy Efficiency: A Review of Its Performance Efficacy and Global Applicability. Sustainability, 17(5), 1876. https://doi.org/10.3390/su17051876

[5] Rahman, M., Mim, S. A., & Oshin, S. A. (2021). Integration of Building Information Modeling (BIM) and LEED for a green building rating. Journal of Engineering Science, 12(2), 47–57. https://doi.org/10.3329/jes.v12i2.54630

[6] U.S. Green Building Council. (n.d.). LEED v4.1 existing building: Performance path and recertification. U.S. Green Building Council. Retrieved from https://www.usgbc.org/leed/v41

[7] Taleb, H. M., & Sharples, S. (2011). Developing sustainable residential buildings in Saudi Arabia: A case study. Applied Energy, 88(1), 383–391. https://doi.org/10.1016/j.apenergy.2010.07.029

[8] King Abdullah Petroleum Studies and Research Center (2025). LEED certifications and sustainability achievements. KAPSARC. Retrieved from. https://www.kapsarc.org/about-kapsarc/our-facilities

[9] Bersouli, F., & Samira, A. S. (2023). Green buildings and reducing the environmental impact. Journal of Economic Sciences Institute, 26(1), 131–150

[10] Yuan, H., & Shen, L. (2020). Revisiting triple bottom line within the context of sustainable construction: A systematic review. Journal of Cleaner Production, 252, 119884. https://doi.org/10.1016/j.jclepro.2019.119884

[11] Jerry Yudelson (McGraw-Hill, 2009). Green Building Through Integrated Design. https://openlibrary.org/books/OL16975803M/Green_building_through_integrated_design

[12] Lu, Y. et. al., (2025). A literature review of sustainable building research. Buildings, 15(19), Article 3609. https://doi.org/10.3390/BUILDINGS15193609

[13] Osman Attmann (2010). Green Architecture: Advanced Technologies and Materials Open Library. https://openlibrary.org/works/OL15612106W/Green_architecture

[14] Almusaed, A., Yitmen, I., Myhren, J. A., & Almssad, A. (2024). Assessing the impact of recycled building materials on environmental sustainability and energy efficiency: A comprehensive framework for reducing greenhouse gas emissions. Buildings, 14(6), 1566. https://doi.org/10.3390/buildings14061566

[15] U.S. Green Building Council. (n.d.). LEED rating system. USGBC. Retrieved from https://www.usgbc.org/leed

Islamic University Journal of Applied Sciences (IUJAS)
https://journals.iu.edu.sa/jesc

Volume VIII, Issue I, July 2026, Pages 139-152

Simulation and Experimental Assessment of Various MPPT Methods for Photovoltaic Systems

Nassir Rouibah [1] *, Abdelouahab Bouttout [2], Sofiane Haddad [3]

[1]National Center of Studies and Integrated Research of Building, CNERIB, 16097, Algeria, rouibahnassir@gmail.com

[2] National Center of Studies and Integrated Research of Building, CNERIB, 16097, Algeria, bouttout@gmail.com

[3] Renewable Energy Laboratory, Electronics Department, MSB, University of Jijel,18000, Algeria, sof_had@yahoo.fr

*Corresponding author: Nassir Rouibah , *Email Address*: rouibahnassir@gmail.com

Abstract

Maximum Power Point Tracking (MPPT) is a power electronic system used to control the electrical circuit that interfaces with the photovoltaic array and adjusts the electrical operating point of the module to operate at its MPP. Several methods have been developed over the years to implement MPPT effectively. This paper covers the simulation and experimental implementation of several MPPT methods. Among these methods are conventional, hybrid, and soft competing methods. The simulations were carried out using MATLAB/SIMULINK. The Arduino Mega 2560 controller implemented the hardware configuration and real-time data acquisition. All experimental work is done at the Renewable Energies Laboratory of the University of Jijel. A thorough explanation of the platforms for implementing software and hardware is given. Test the flexibility and comparison of the MPPT technique has been performed. The results of MPPT methods obtained from simulation and experimental setups were presented and discussed.

Keywords: MPPT techniques; Implementation; Photovoltaic system; MATLAB/Simulink.

https://doi.org/10.63070/jesc.2025.020
Received 25 December 2025; Revised 18 February 2026; Accepted 01 March 2026.
Available online 10 March 2026.
Published by Islamic University of Madinah on behalf of *Islamic University Journal of Applied Sciences*.
This is an open access article published under the Creative Commons Attribution 4.0 International license (CC BY 4.0):https://creativecommons.org/licenses/by/4.0/

1. Introduction

The International Energy Agency (IEA) has released its annual report, projecting that the world's renewable electricity capacity will increase by over 60% from 2020 levels to more than 4,800 GW by 2026 [1]. This represents the equivalent of the current total global energy capacity from nuclear power and fossil fuels combined. The report forecasts that renewable energy will account for approximately 95% of the global energy capacity growth through 2026, with solar photovoltaics alone contributing to more than half of this increase [2-3].

In order to increase PV system efficiency using MPP tracing, MPPT controllers have been used. 1968 saw the introduction of the first PV system with MPPT for a space system [4-5]. Following that, there was a sharp increase in the dependability, precision, speed, efficiency, and simplicity of MPPT controllers. The best MPPT algorithm often follows the quick changes in output power and operates at a high speed with minimal oscillations around the maximum power point (MPP). Today, there are a lot of applicable MPPT control methods accessible [6]. These controllers were used to monitor a photovoltaic system with several maximum power points or the MPP under uniform illumination for a single MPP [7].

A power electronic system is called Maximum Power Point Tracking (MPPT)that is used for controlling the electrical circuit that interfaces with the photovoltaic array to adjust the electrical operating point of the module such that it operates at its MPP [8]. This is done because the MPP of a solar power device depends on the temperature and the irradiation level. Because of this, if the electrical resistance of the load is fixed, the operating voltage of the PV module will have to be changed to obtain the maximum power [9]. In general, the system controller makes a decision based on the estimated MPP to adjust the required voltage. This is done by continuously sweeping the operating voltage of the PV array and measuring the power output. Once the power output has been calculated, it is compared to the power output at the previous voltage increment. The controller then increases the voltage and continues the process [10-11-12].

In the literature, different categories of MPPT have been presented. These methods can be broadly categorized into two groups: advanced methods and traditional methods [13]. Traditional MPPT techniques are commonly used because they were introduced a long time ago. They are valued for their simplicity and ease of implementation. However, these techniques are limited to monitoring the power point of solar systems under certain conditions [14]. The well-known traditional methods are the perturb and observe (P&O), hill climbing (HC), fuzzy logic control (FLC), artificial neural network (ANN) and, incremental inductance (In Cond) etc. [15-16]. On the other hand, advanced MPPT

techniques, such as artificial intelligence (AI) or bioinspired (BI) methods, have been developed [17-18]. These advanced techniques outperform traditional approaches in tracking despite being relatively complex. As a result, they are capable of extracting the global power point maximum (GMPP) in situations of partial shading. The most used GMPPT methods in the science are partial swarm optimization (PSO), gray wolf optimization (GWO) and Ant colony optimization (ACO) [19]. The latest developed modern bioinspired (BI) methods for tracking the GMPPT are: Horse Herd Optimization (HAO), Flying Squirrel Search Optimization (FSSO) and Salp Swarm Algorithm (SSA) [20]. The goals of this paper are to simulate and test the performance of various MPPT methods in two types of weather: uniform irradiation conditions (UICs) and non-uniform irradiation conditions (N-UICs). The simulated and verified experimental MPPT methods are P&O, modified version of P&O, In Cond, modified version of InCond, HC, FLC, adaptive P&O_FL and, PSO method [21-22].

Enhancing the efficiency of photovoltaic systems requires reliable extraction of maximum power under constantly changing conditions such as irradiance, temperature, and partial shading. To address this challenge, numerous MPPT techniques—ranging from classical P&O and INC methods to optimization-based and intelligent algorithms—have been proposed. However, their reported performance varies significantly depending on the models, panel characteristics, environmental assumptions, and evaluation metrics used in different studies. This variability makes it difficult to determine which approach is truly the most suitable in real-world applications. Moreover, simulation outcomes often fail to match experimental results, adding further uncertainty. Consequently, a major issue arises: identifying the MPPT method that best balances rapid tracking, accuracy, stability, and implementation simplicity[23]. This highlights the need for rigorous comparative assessments, combining both simulation and experimental validation, to better understand the actual behavior of different MPPT strategies under varying operational conditions [24].

The main objective of this work is to provide a comprehensive experimental benchmarking of eight widely used MPPT algorithms implemented on a unified low-cost hardware platform. Unlike many previous comparative studies that are limited to simulation-based evaluations or a reduced number of algorithms, this study combines MATLAB/Simulink simulations with real-time implementation using an Arduino Mega 2560 controller. All algorithms are evaluated under identical operating conditions, including uniform and dynamically varying irradiation levels, ensuring a fair and consistent performance assessment. This approach offers practical insights into implementation complexity, dynamic response, and steady-state behavior, thereby bridging the gap between theoretical analysis and real-world applications.

The structure of the paper is as follows: Section II presents a description of the test bench, and provide a thorough explanation of the hardware and software. Section III presents an Experimental setup. Section VI present results and discussions, Finally, A conclusion is given in Section V.

2. Description of the test bench

In this section, we presented the test bench carried out at the RELab. As shown in Figure 1, The test bench consists of the following elements: a GPV, a chopper type DC/DC converter, a resistive load, current and voltage sensors, a digital oscilloscope, DC power supply, Arduino Mega 2560 bord and a computer.

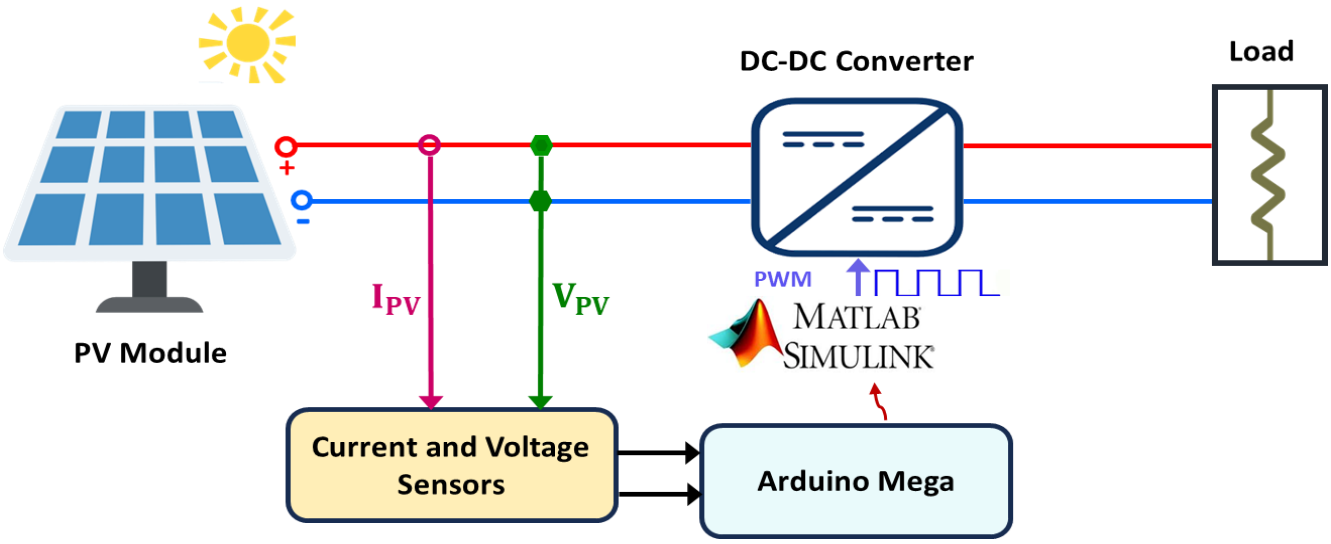


Figure 1. Photovoltaic system architect.

2.1.Hardware Design

This subsection describes the hardware architecture of the test bench of photovoltaic system and explains the main parts of the system. The PV module type that is being used is the MSX- BP Solar 120 watts, which is produced using multi-crystalline technology. Table 1 displays its electrical parameters under standard test conditions (STC: 1000 W/m2 and 25 °C).

A DC-DC chopper converter is a device that provides highly efficient, dependable, and flexible electrical power regulation in DC circuits. Figure 2 shows the electrical circuit of the chopper converter. Its primary components are two switches (MOSFET, Diode), two capacitors (Cint, Cout), and an inductance (L). The values of the utilized components are compiled in Table 2.

Table 1. Specifications of the MSX-120 solar module

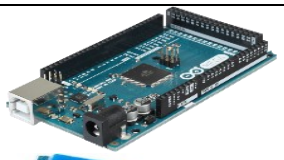
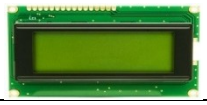
Description	**Value**
Open circuit voltage (Voc)	42.1 V
Short circuit current (Isc)	3.87 A
Rated voltage (Vmpp)	33.70 V
Rated current (Impp)	3.56 A
Rated power (Pmax)	120 W

Table 2. DC-DC Chopper specifications

Description	Value
Input capacitor (Cint)	50 µF
Output capacitor (Cout)	220 µF
Bobin (L)	1 mH
Load (lamp)	5 W - 12 V
Frequency	25 KHz

A data acquisition system is created to gather, process, save, and transmit information from the data acquisition board Arduino Mega and voltage and current sensors to control the designed DC-DC chopper. Below are the main components of the data acquisition system designed for PV systems. In Table 3, it is important to note that the parts used to create the data acquisition system are easily obtainable in the market and are reasonably priced. The total cost of the designed data acquisition system is estimated to be 21.41 $.

Table 3. The reset component of prototype.

Components	Role /cost	Figure
Arduino Mega 2560 Bord	Acquisition Measurement data the and control DC-DC chopper 10.67$	
Sensor voltage	Measurement of the output voltage of the PV Module (Vpv) 2.3$	
Sensor current AllegroACS712	Measurement of the output current of the PV Module (Ipv) 2.1$	
LCD screen 16x4	Show the values of measurement parameters 6.34 $	

2.2. Software Design

In this work, we are not intended to give theoretical backgrounds about the used algorithms: Perturb and Observe (P&O), modified version of (P&O), Incremental Conductance (In Cond), modified version of (In Cond), Hill Climbing (HC), modified version of In Cond, FLC, adaptive P&O_FL and, PSO method, as these algorithms are already developed and simulated. So, in this study, we will focus more on the simulation and experimental implementation in an Arduino board microcontroller of these methods. The MPPT method using the Simulink library is used to determine the duty cycle of the

DC/DC step-up chopper in use, and the Arduino Mega 2560 board is directly used to provide the PWM signal control. As previously indicated, for acquisition and control. The MOSFET transistor of the chopper converter is controlled by the first board, which also implements MPPT algorithms. One DC-DC converter with a switching frequency of 25 kHz and a PWM generator make up the power stage interface between the PV module and the load. The MPPT controller calculates the necessary duty cycle and transmits it to the PWM generator based on the measured Ipv and Vpv. The latter changes the duty cycle of the MPPT into a PWM signal that activates and deactivates the power switch of the DC-DC chopper .

3. Experimental Setup

The facility test has been carried out at the Renewable Energies Laboratory of the University of Jijel. Figure 5 shows the hardware implementation platform that was utilized for the MPPT technique's experimental evaluation and implementation. The test bench includes a DC-DC chopper converter, digital oscilloscope, resistive load, a computer for display results, and a data acquisition board that consists of an Arduino Mega Board, current and voltage sensors, and an LCD screen. Figures 3-4 show the designed DC-DC chopper converter and the data acquisition board, respectively.

During experimental testing, irradiance levels were monitored using a calibrated reference solar cell. The reference cell is installed in the same plane and orientation as the photovoltaic panels to ensure accurate irradiance measurement. Solar irradiance is determined by measuring the voltage across two resistors integrated within the reference cell. The system is calibrated using a pyranometer and a luxmeter, with a calibration coefficient of K=1000.The voltage and current sensors were calibrated using a precision digital multimeter prior to data acquisition to correct offset and gain errors [5].

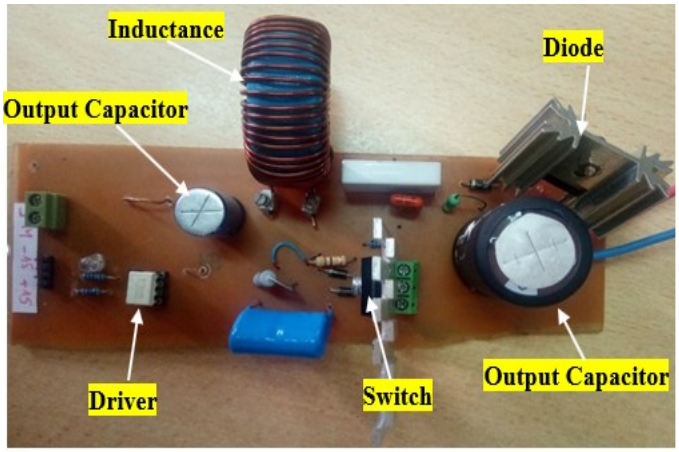


Figure 3. DC-DC chopper converter.

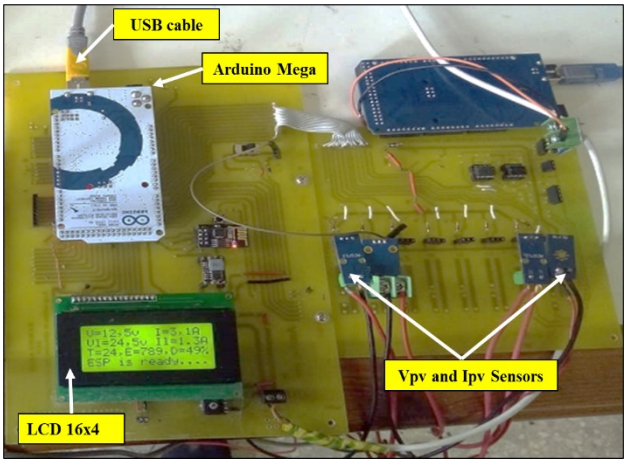


Figure 4. Data acquisition board.



Figure 5. Data acquisition board.

4. RESULTS AND DISCUSSION

4.1. SIMULATION RESULTS

In order to evaluate the effectiveness of the MPPT methods, we conducted various simulations using the MATLAB/Simulink software under different conditions. These conditions included uniform irradiation and rapidly changing atmospheric conditions. Figure 6 illustrates the implementation of the entire standalone photovoltaic system using MATLAB/Simulink. The key components involved are the PV generator, two blocks for current and voltage sensors to measure Ipv and Vpv, a resistive load, a DC/DC chopper converter, and a PWM generator.

- **Simulation under uniform irradiation conditions**

Temperature T = 25 °C is maintained constant during this test and each method is tested with a uniform irradiation level. The results obtained using the test under UICs are presented in Figure 7(a-b-c-d). When the MPPT algorithms search on the MPP, the optimum voltage and current cause oscillations in PV voltage and current. These current and voltage oscillations result in power oscillations. Figure 7 (a-b-c-d), shows that the convergence time of all the methods will not drop back to 0.03 seconds except the P&O method, which takes almost 0.05 seconds. The output power is stable as shown in Figure 7(c-d), except with the two methods P&O and InCond (see figure 7-a-b), which contain small oscillations around the MPP.

- **Simulation under variations of irradiation levels**

This test maintains a constant temperature of T = 25°C while introducing various irradiation step adjustments. The initial irradiation is between G = 400W/m and 800 W/m, then increased and

decreased. The results obtained are presented in figure 8. Figure (8-c) shows that there are oscillations around MPP and a slow response when there is a change of irradiation levels for the HC method, especially in the low irradiation levels of 200 W/m2. On the other hand, the LF method presents a faster response during irradiation changes, and oscillations are almost neglected. Figure (8-e) shows that the PSO and P&O_FL methods show oscillations at the beginning of the MPP search but have a faster response during irradiation change.

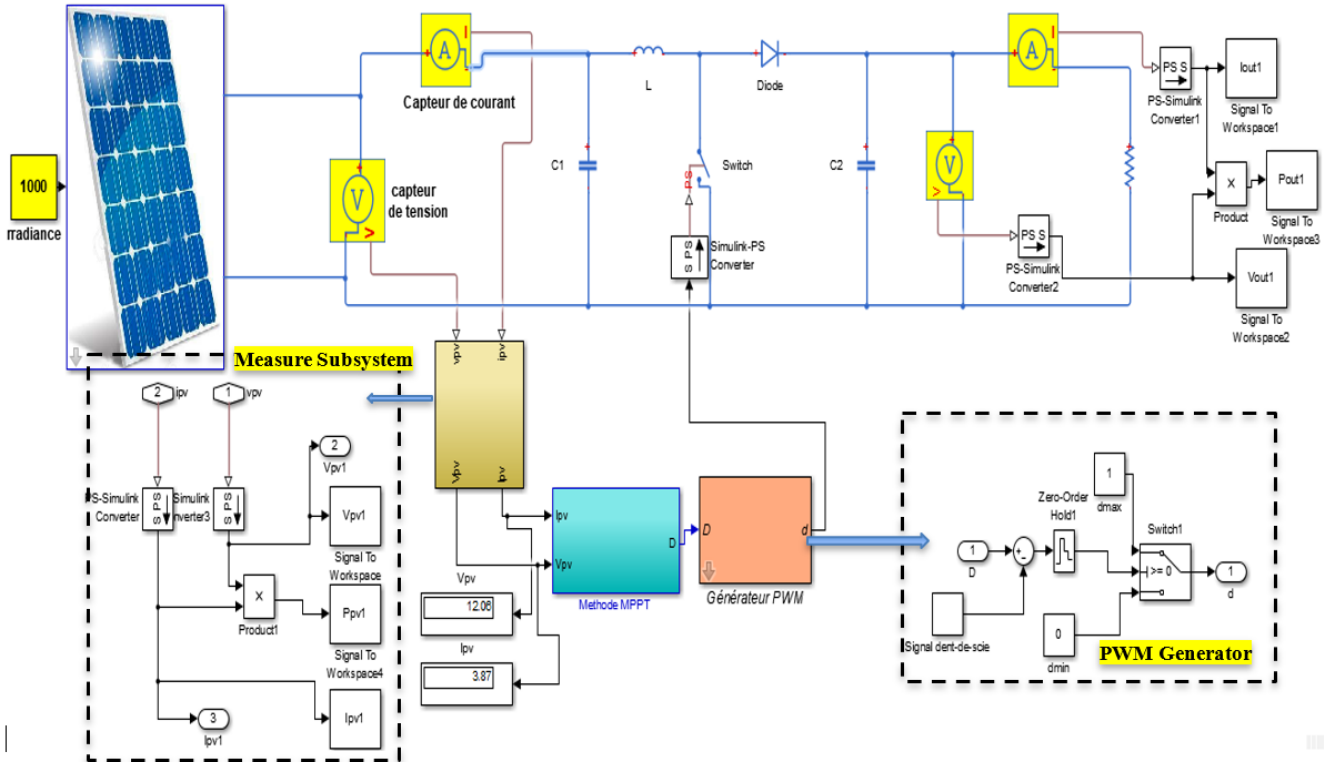
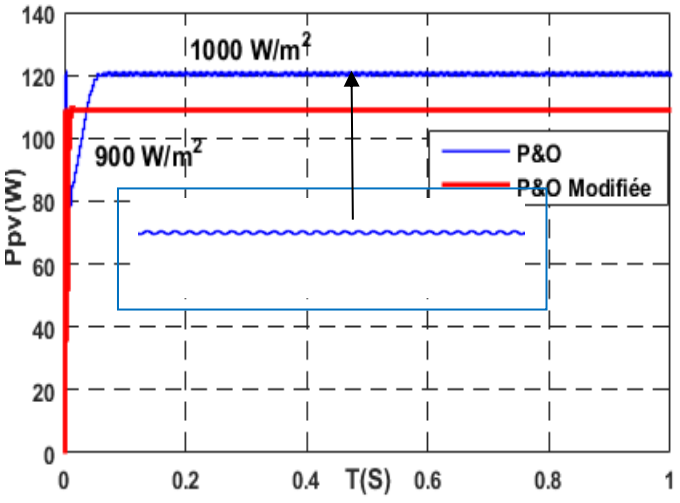
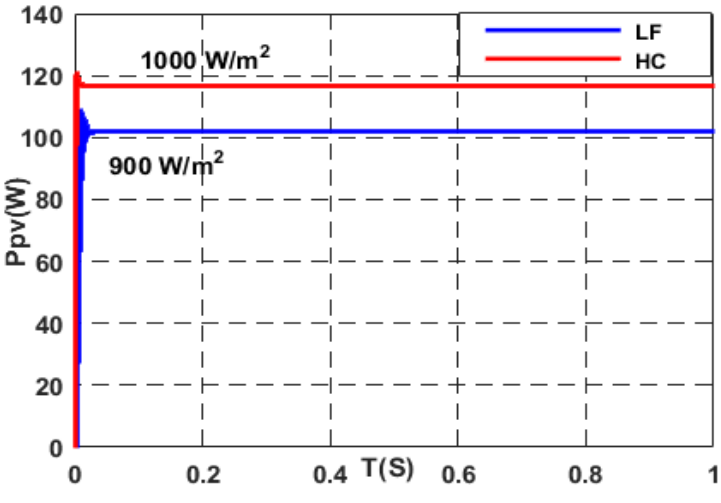


Figure 6. Model of a stand-alone PV system using MATLAB/Simulink.



(a)



(b)

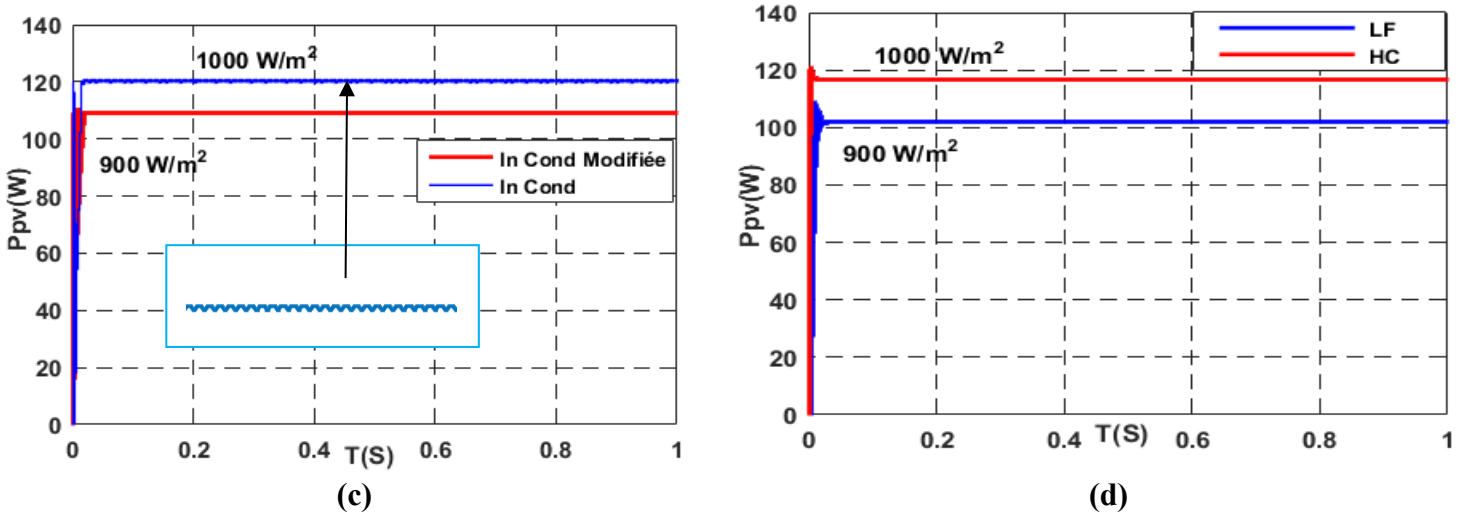


Figure 7. PV power of different MPPTs methods: a) P&O and InCond methods; b) P&O and modified In Cond; c) HC and LF; d) Adaptive P&O_LF and PSO.

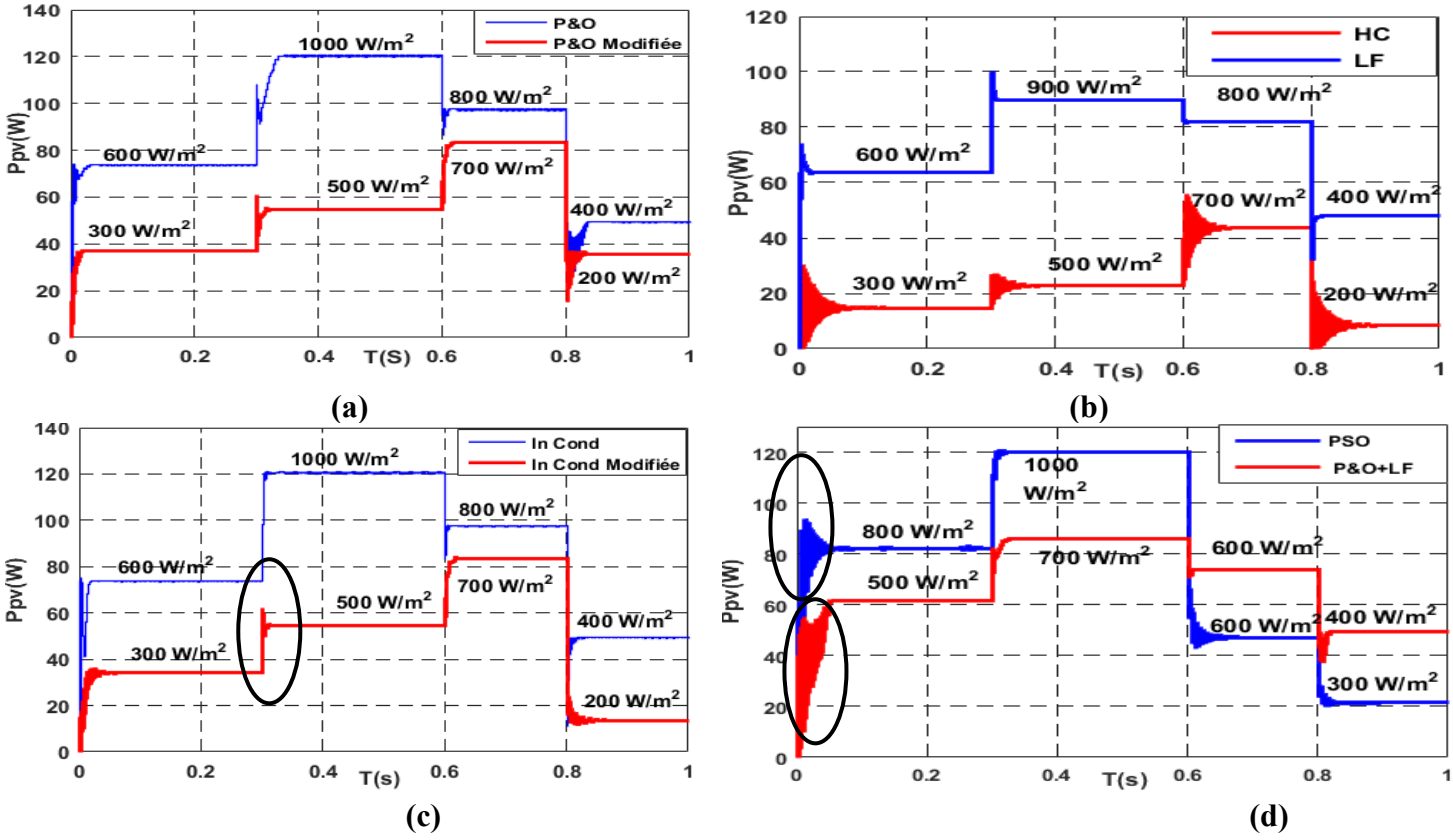


Figure 8. PV power of different MPPTs methods: a) P&O and In Cond methods; b) P&O and modified In Cond; c) HC and LF; d) PSO and P&O_LF.

4.2 EXPERIMENTAL RESULTS

To test the MPPT methods under CIUs, experiments were conducted on three BP Solar 60 W PV modules connected in series. The hardware implementation platform shown in Figure .5 was designed at the Relab. The measured characteristics of PV voltage, current, and power are given in Figure 10-

a-b-c-d-e-f. The experimental conditions were conducted under ambient temperatures ranging between 25 °C and 28 °C, with irradiance levels varying between approximately 200 W/m² and 800 W/m² depending on weather conditions and shading adjustments

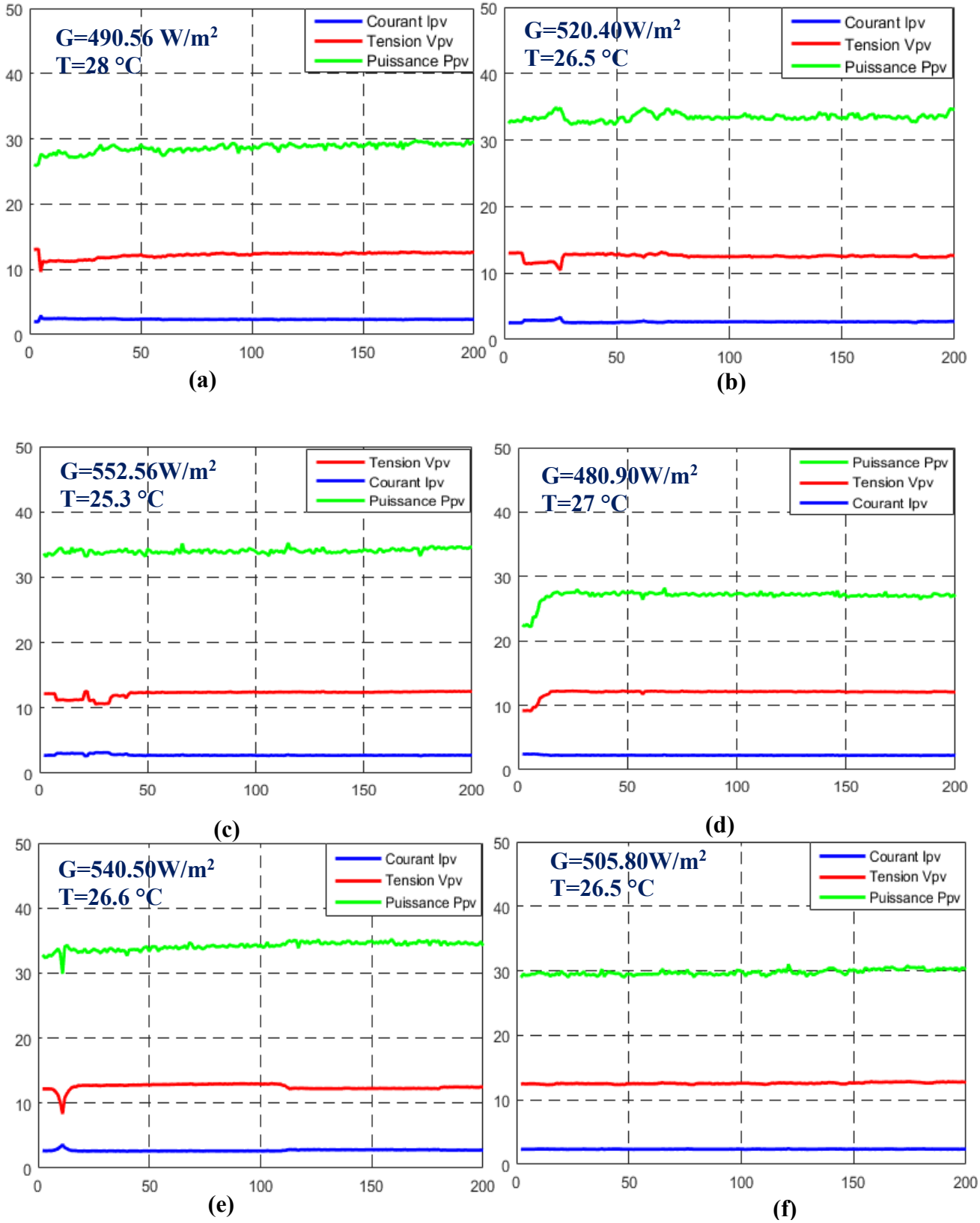


Figure 9. Experimental power, current and voltage curves of different MPPT methods:

(a)P&O, (b)InCond, (c) HC, (d) FL(e) FL_P&o (f) PSO.

Table 4. Comparative evaluation of different MPPT methods.

MPPTs Methods	Irradiation W/m² Temperature °C	Convergence time(ms)	Efficiency %	Average Power(W)	Used sensors	Ability to track under COP	Oscillation at PPM
P&O	490.56 28.00	05.00	94.93	28.5250	V, I	No	Exists
InCond	520.40 26.50	25.00	95,99	33.4852	V, I	No	Exists
HC	552.56 25.30	24.00	99.40	34.0394	V, I	No	Exists
LF	480.90 27.00	10.00	97.13	34.1784	V, I	Yes	Exists
LF_ P&O	540.5 26.6	5.500	96.15	29.8037	V, I	Yes	Exists
PSO	505.80 26.50	03.30	95.73	27.0304	V, I	Yes	Exists

Under constant temperature (25 °C) and uniform irradiation, all MPPT methods were able to track the maximum power point. However, clear differences in dynamic behavior and stability were observed. The P&O method exhibited the slowest convergence time, reaching steady state after approximately 0.05 s, whereas the other techniques converged faster, within about 0.03 s. This slower response is mainly attributed to the fixed perturbation step size of the conventional P&O algorithm, which limits its ability to rapidly reach the MPP. The modified version of the P&O algorithm introduces an adaptive perturbation step size that decreases as the operating point approaches the MPP, thereby reducing steady-state oscillations while maintaining fast convergence. In contrast, the conventional P&O uses a fixed perturbation step. Similarly, the modified Incremental Conductance algorithm incorporates an adaptive decision threshold and variable duty cycle adjustment to improve stability and tracking accuracy under rapidly varying irradiation conditions.

Moreover, noticeable oscillations in PV voltage, current, and output power were observed for the P&O and InCond methods. These oscillations around the peak power point reduce the effective harvested energy and indicate lower steady-state efficiency. In contrast, methods such as HC, LF, PSO, and the adaptive P&O_LF demonstrated reduced oscillations and improved stability, confirming their superior tracking precision under steady environmental conditions.

When irradiation levels were varied while maintaining constant temperature, the performance differences among the MPPT methods became more pronounced. The HC method showed a slower

dynamic response and significant oscillations around the MPP, particularly at low irradiation levels (200 W/m²). This behavior highlights the sensitivity of hill-climbing approaches to rapid environmental changes and reduced irradiance. The LF-based method, however, demonstrated a much faster response and minimal oscillations during irradiation transitions. This confirms its ability to adapt efficiently to dynamic operating conditions. Similarly, the PSO and P&O_LF methods showed fast convergence following irradiation changes, although initial oscillations were observed during the MPP search phase. Despite this, their overall tracking speed and accuracy were superior to conventional algorithms.

5. Conclusion

In this paper, we have presented the results of our simulation and experimental implementation of Maximum Power Point Tracking (MPPT) methods under Uniform Irradiance Conditions (UICs) and variations in irradiation levels. The results are divided into two parts. The first part includes various simulations using the MATLAB/Simulink software under different conditions, specifically: uniform irradiance conditions and rapidly varying atmospheric conditions. The second part comprises the validation of different MPPT methods implemented experimentally and the simulation of MPPT methods, including the Perturb and Observe method (P&O), modified P&O method, Incremental Conductance method (InCond), modified (improved) InCond method, Hill Climbing method (HC), Fuzzy Logic-based (FL) method, Adaptive P&O_FL method, and Particle Swarm Optimization method (PSO). The flexibility and comparison of the MPPT methods have been tested.

Funding

The authors declare that this research received no external funding.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] International Energy Agency [Online]. Available.at http://www.iea.org/. [Accessed: Jan. 03, 2026].

[2] Abdulaziz F. Shendi, Ahmad M. Nahhas. "Power generation by using Photovoltaic systems for Yanbu and Rabigh regions in Saudi Arabia: a costeffective study". Islamic University Journal of Applied Sciences, Volume V, Issue I, July 2023, Pages 10-44. https://doi.org/10.63070/jesc.2023.001.

[3] Habib Muhammad Usman1, Nirma Kumari Sharma, Sani Saminu, Abdulbasit Bashir Yero, Abdurrahman Salisu Yahya, Farouk Isah Muhammad. "Enhancing Photovoltaic Efficiency through Engine Oil Coatings: A Comparative Analysis of New, Partially Used, and Degraded Oils". The Islamic University Journal of Applied Sciences, Issue II, Volume VI, December 2024, Pages 134-154. . https://doi.org/10.63070/jesc.2024.018.

[4] Motahhir S, El Ghzizal A, Sebti S, Derouich A. MIL and SIL and PIL tests for MPPT algorithm. Cogent Engineering. 2017 Jan 1 ;4(1) :1378475.

[5] Rouibah, N., Elhammoumi, A., Bouttout, A., Haddad, S., Oukaci, S., Limam, A. and Benghanem, M., 2025. Smart monitoring of photovoltaic energy systems: An IoT-based prototype approach. Scientific African, p.e02973.

[6] Messai A, Mellit A. Design and implementation of maximum power point tracking algorithm using fuzzy logic and genetic algorithm. In Assessment and Simulation Tools for Sustainable Energy Systems 2013 (pp. 285-307). Springer, London.

[7] Zainuri MA, Radzi MA, Che Soh A, Rahim NA. Development of adaptive perturb and observe-fuzzy control maximum power point tracking for photovoltaic boost dc–dc converter. IET Renewable Power Generation. 2014 Mar;8(2):183-94.

[8] Rouibah, N., Barazane, L., Rabhi, A., Hajji, B., Bouhedir, R., Hamied, A. and Mellit, A., 2021. Experimental Assessment of Perturb & Observe, Incremental Conductance and Hill Climbing MPPTs for Photovoltaic Systems. In Proceedings of the 2nd International Conference on Electronic Engineering and Renewable Energy Systems: ICEERE 2020, 13-15 April 2020, Saidia, Morocco (pp. 461-467). Springer Singapore.

[9] Li, P., Zhang, J., Xu, R., Zhou, J. and Gao, Z., 2024. Integration of MPPT algorithms with spacecraft applications: Review, classification and future development outlook. *Energy*, *308*, p.132927.

[10] F. Belhachat, C.Larbes,A review of global maximum power point tracking techniques of pv system under partial shading conditions, Renewable and Sustainable Energy Reviews (2018) 513–553.

[11] Villegas-Mier, C.G., Rodriguez-Resendiz, J., Álvarez-Alvarado, J.M., Rodriguez-Resendiz, H., Herrera-Navarro, A.M. and Rodríguez-Abreo, O., 2021. Artificial neural networks in MPPT algorithms for optimization of photovoltaic power systems: A review. *Micromachines*, *12*(10), p.1260.

[12] D Rekioua, E. Matagne, Optimization of photovoltaic power systems: modelization, simulation and control, Springer Science & Business Media (2012) 113–148.

[13] Rouibah, N., Bouttout, A., Oukaci, S., Limam, A. and Derradji, L., 2024, October. The Recent Global MPPT Methods: State-of-the-Art. In International Conference on Artificial Intelligence in Renewable Energetic Systems (pp. 504-515). Cham: Springer Nature Switzerland.

[14] B Bendib, H Belmili, F Krim, A survey of the most used MPPT methods: Conventional and advanced algorithms applied for photovoltaic systems, Renewable and Sustainable Energy Reviews 45 (2015) 637–648.

[15] Z Salam, J Ahmed, BS Merugu, the application of soft computing methods for MPPT of PV system: A technological and status review, Applied energy 107 (2013)135–148.

[16] Ali, A.I.M., Mousa, H.H., Mohamed, H.R.A., Kamel, S., Hassan, A.S., Alaas, Z.M., Mohamed, E.E. and Abdallah, A.R.Y., 2023. An enhanced P&O MPPT algorithm with concise search area for grid-tied PV systems. *IEEE Access*, *11*, pp.79408-79421.

[17] Jately, V., Azzopardi, B., Joshi, J., Sharma, A. and Arora, S., 2021. Experimental Analysis of hill-climbing MPPT algorithms under low irradiance levels. *Renewable and Sustainable Energy Reviews*, *150*, p.111467.

[18] HA Sher, AF Murtaza, A Noman, KE Addoweesh, K Al-Haddad, M Chiaberge, A new sensorless hybrid MPPT algorithm based on fractional short-circuit current measurement and P&O MPPT, IEEE Transactions on Sustainable Energy 6 (2015).

[19] De Brito, M. A. G., Galotto, L., Sampaio, L. P., e Melo, G. d. A., et Canesin, C. A., "Evaluation of the main MPPT techniques for photovoltaic applications," IEEE transactions on industrial electronics, V. 60, no 3, (2013), 1156-1167.

[20] Zemmit, A., Loukriz, A., Belhouchet, K., Alharthi, Y.Z., Alshareef, M., Paramasivam, P. and Ghoneim, S.S., 2025. GWO and WOA variable step MPPT algorithms-based PV system output power optimization. *Scientific Reports*, *15*(1), p.7810.

[21] Bouchakour, A., Zarour, L., Bessous, N., Bechouat, M., Borni, A., Zaghba, L., Rabehi, A., Alwabli, A., El-Abd, M. and Ghoneim, S.S., 2024. MPPT algorithm based on metaheuristic techniques (PSO & GA) dedicated to improve wind energy water pumping system performance. *scientific reports*, *14*(1), p.17891.

[22] Celikel, R., Aydogmus, O. and Yilmaz, M., 2025. A modified perturb and observe MPPT algorithm for PEMFC with rapid convergence and low power oscillation. *Scientific Reports*, *15*(1), p.27848.

[23] Xia, K., Li, Y. and Zhu, B., 2024. Improved photovoltaic MPPT algorithm based on ant colony optimization and fuzzy logic under conditions of partial shading. *IEEE Access*, *12*, pp.44817-44825.

[24] Bouksaim, M., Mekhfioui, M. and Srifi, M.N., 2025, September. A Comprehensive Decade-Long Review of Advanced MPPT Algorithms for Enhanced Photovoltaic Efficiency. In Solar (Vol. 5, No. 3, p. 44). MDPI.

Islamic University Journal of Applied Sciences (IUJAS)
https://journals.iu.edu.sa/jesc

Volume VIII, Issue I, July 2026, Pages 153-167

Numerical Vibration Analysis and Nonlinear Stability Prediction of Hydrodynamic Journal Bearings: Short and Long Bearing Configurations

Amira Amamou1[1,2], Amani Amamou [1,3], Nejla Mahjoub Said [4,*]

[1] Laboratory of Mechanical Engineering, National Engineering School of Monastir, University of Monastir, 5000 Monastir, Tunisia
[2]Faculty of Sciences of Monastir, University of Monastir, 5000 Monastir, Tunisia, amira.amamou@fsm.rnu.tn
[3] National Engineering School of Gasfa, University of Gafsa, 2112 Gafsa, Tunisia
amani.amamou@yahoo.fr
[4] Department of Physics, College of Science, King Khalid University, Abha, 61413, Saudi Arabia,

* Corresponding author: Nejla Mahjoub Said, *Email Address*: nalmahjoub@kku.edu.sa

Abstract

This study investigates a two-degree-of-freedom model of a rigid, symmetrical, and balanced rotor supported by two identical journal bearings. Hydrodynamic fluid forces are described using half Sommerfeld solutions combined with both short and long bearing approximations. Accurate prediction of stability boundaries and nonlinear behaviors near these limits, such as stable and unstable limit cycles, hysteresis, and jump phenomena, is essential for reliable rotor-bearing system design and operation. To achieve this, a continuation approach based on a predictor–corrector methodology is employed to trace equilibrium paths and bifurcation points. The results show that for short bearings (aspect ratio $D/L < 0.5$), a bearing parameter of 0.8 produces subcritical bifurcation at rotational speeds below the dimensionless threshold of 2.64. In contrast, for long bearings ($D/L > 1$), a parameter of 0.42 leads to supercritical bifurcation for speeds exceeding 1.67. These findings highlight the significant influence of bearing parameters and rotor dynamics on system stability and nonlinear response. Finally, numerical integration of the journal center motion validates the accuracy and robustness of the continuation-based predictions.

Keywords: Nonlinear stability; Rotor bearing system; Hopf bifurcation; Numerical integration; Numerical continuation.

https://doi.org/10.63070/jesc.2026.021
Received 24 December 2025; Revised 03 April 2026; Accepted 10 April 2026.
Available online 15 April 2026.
Published by Islamic University of Madinah on behalf of *Islamic University Journal of Applied Sciences*.
This is an open access article published under the Creative Commons Attribution 4.0 International license (CC BY 4.0):https://creativecommons.org/licenses/by/4.0/

1. Introduction

The stability of journal bearings is of utmost importance to ensure the safe operation of rotating machines. While modelling the rotor bearing system's dynamic characteristics, nonlinearities are induced due to self-excited vibrations caused by hydrodynamic bearing instability. Linear analysis is usually used to predict stability threshold speeds; the prediction is only valid for small motion about the equilibrium position. Nonlinear analysis, however, is required to predict rotor whirl motions which may cause rotors failure during operation.

Several authors have studied the dynamic behavior of journal bearings and how it affects rotor system vibration. Newkirk [1] reported on the self-excited instability motions of rotor bearings systems based on an experimental examination. He noticed that orbital motions appear at a frequency roughly equal to half the rotor speed. These findings were published in Hollis and Taylor [2]. In 1960, Holmes [3] estimated the oil whirl threshold speeds and then developed a stability chart for the rotor-bearing system considering oil film as elastic in a viscous system. Linear analysis of a rotor supported by hydrodynamic bearings has been also performed to forecast the instability threshold speed for the long bearing approximation model [4].

The investigation of the nonlinear effect on bearing stability indicated that either a supercritical or subcritical Hopf bifurcation indicates the beginning of instability. For supercritical behavior, a stable equilibrium solution of the equations of motion loses stability as the rotor speed increases, and a stable limit cycle emerges smoothly from the bifurcation point. However, unstable limit cycles appear at speeds below the stability threshold speed in a subcritical bifurcation point, creating the equilibrium point basin of attraction. Myers [5] used the Hopf bifurcation theory to forecast periodic orbits at rotor speeds near the critical stability speed of a long journal bearing. He discovered the existence of three distinct bifurcation areas in the parameter space of the steady-state eccentricity ratio. Related research for the short journal bearing was reported by Hollis and Taylor [2].

To analyze the stability of short journal bearings while taking the turbulence impact into consideration, Khonsari and Wang [6] applied the Hopf bifurcation theory. The analytical investigation of the stability of hydrodynamic journal bearings reported in several references [5-7] is restricted to predicting the Hopf bifurcation point, the kind of bifurcation, and the size and form of the limit cycles in the vicinity of the stability threshold speed. Several other foundational studies back up the results. In 1993, Childs [8] focused on rotordynamic stability analyses for fluid-film bearings. Ehrich [9], in 1992, reported subsynchronous whirl and hysteresis in experiments, characteristic of with subcritical Hopf bifurcation. San Andrés and Taylor [10] conducted experiments that validated nonlinear force interactions in rotor–bearing test rigs. Brockett and El-Sayed [11] analysed nonlinear transient

responses in journal-bearing systems focusing on the complex dynamic transitions caused by fluid-film forces. More recently, Tondl [12] along with Zhang and Meng [13] developed nonlinear models that describe nonlinear phenomena governing the transition from a stable equilibrium point to unstable limit cycles. Amamou [14] used numerical continuation and bifurcation analysis to predict the rotor safe operating speed range by determining an extended stability chart including the domain of the journal threshold stability speed and the boundaries of limit cycle bifurcations.

Tools for the continuation and bifurcation detection of equilibrium points and limit cycles of nonlinear ordinary differential equations have been extensively developed within the field of numerical continuation [15,16,17]. Classical continuation algorithms allow systematic tracing of solution branches and detecting qualitative changes such as saddle-node, pitchfork, and Hopf bifurcation, as well as the birth of periodic solutions. Among the early and influential developments are the software packages AUTO [18], CONTENT [19], and MATCONT [20], which provide robust algorithm for branch switching, Floquet multiplier computation, and stability analysis. Taken together, these tools form a powerful computational framework for understanding both local and global behaviours in nonlinear dynamical systems.

Building on our previous works [17], [21], in which the nonlinear analysis and stability of short and long hydrodynamic bearings were studied separately, the present work goes a step further by combining both short and long bearing configurations within a unified framework. This allows a direct comparison of their nonlinear dynamic behavior and bifurcation characteristics, revealing how bearing geometry, through the aspect ratio D/L, governs the transition between subcritical and supercritical regimes. In the present work, the stability analysis of a rigid and balanced rotor supported by hydrodynamic journal bearings is carried out using numerical continuation and verified by numerical integration. Both short and long models of a journal bearing under the half Sommerfeld boundary conditions are considered. The findings of the path following approach are supported by numerical simulation for a chosen set of bearing design parameters under specific working circumstances. This paper is organized as follows. First, a nonlinear dynamic model of a rotor bearing system with short and long journal bearing is introduced, expressed in terms of the nondimensional journal speed and bearing modulus. Next, numerical continuation techniques and bifurcation analyses of equilibrium points and limit cycles are presented to investigate the stability and bifurcation behavior of the rotor–bearing system. Finally, the main numerical results and findings of the present study are verified using numerical integration method focusing on the nonlinear response and stability characteristics of the investigated journal bearings.

2. Rotor bearing system modelling

The rotor, with a mass $2M$ is mounted in two symmetrical journal bearings. In Figure 1, O_j and O_b are respectively the journal center and the bearing center. The hydrodynamic forces $\boldsymbol{F}_\varepsilon$, $\boldsymbol{F}_\theta$ and constant load $\boldsymbol{W} = M\boldsymbol{g}$ are the external forces acting on the journal. The motion of the journal center is modeled through equations of motion written in terms of derivatives with respect to nondimensional time $\tau = \omega t$ are as follows:

$$\begin{aligned} \ddot{\varepsilon} - \varepsilon\dot{\theta}^2 &= \frac{-F_\varepsilon}{Mc\omega^2} + \frac{g}{c\omega^2}\cos\theta \\ \varepsilon\ddot{\theta} + 2\dot{\varepsilon}\dot{\theta} &= \frac{F_\theta}{Mc\omega^2} - \frac{g}{c\omega^2}\sin\theta \end{aligned} \tag{1}$$

ε and θ are respectively the eccentricity ratio and the attitude angle of the journal center. c is the radial bearing clearance and ω is the rotor rotating speed.

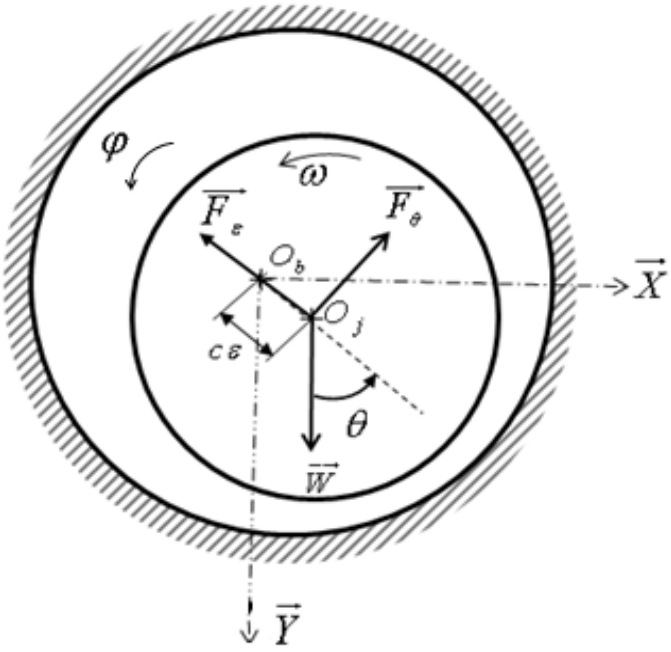


Figure 1. A hydrodynamic bearing cross section [17]

2.1 The short journal bearing model

Short bearings are commonly employed to minimize axial pressure variation and reduce frictional losses. Hydrodynamic force components in the rotating coordinates can be written as [2]:

$$\begin{cases} F_{\varepsilon} = \dfrac{\mu R L^3 \omega}{2c^2}\left[\dfrac{2\left(1-2\dot{\theta}\right)\varepsilon^2}{\left(1-\varepsilon^2\right)^2}+\dfrac{\pi\dot{\varepsilon}\left(1+2\varepsilon^2\right)}{\left(1-\varepsilon^2\right)^{2.5}}\right] \\ F_{\theta} = \dfrac{\mu R L^3 \omega}{2c^2\varepsilon}\left[\dfrac{\pi\left(1-2\dot{\theta}\right)\varepsilon^2}{2\left(1-\varepsilon^2\right)^{1.5}}+\dfrac{4\dot{\varepsilon}\varepsilon^2}{\left(1-\varepsilon^2\right)^2}\right] \end{cases} \tag{2}$$

where μ is the viscosity of the oil film, R is the journal radius and L is the bearing length.

Substituting the expressions of the hydrodynamic reactions (2) in the system of equations (1) and using the dimensionless bearing parameter $\Gamma = \dfrac{\mu R L^3}{2Mc^{2.5}g^{0.5}}$, the dimensionless journal speed $\varpi = \omega\sqrt{\dfrac{c}{g}}$ and the vector of state variables $\boldsymbol{X} = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 \end{pmatrix}^T = \begin{pmatrix} \varepsilon & \theta & \dot{\varepsilon} & \dot{\theta} \end{pmatrix}^T$, the journal center equations of motion can be written in the following state space form:

$$\begin{cases} \dot{x}_1 = x_3 \\ \dot{x}_2 = x_4 \\ \dot{x}_3 = x_1 x_4^2 - \dfrac{\Gamma}{\varpi}\left[\dfrac{2\left(1-2x_4\right)x_1^2}{\left(1-x_1^2\right)^2}+\dfrac{\pi x_3\left(1+2x_1^2\right)}{\left(1-x_1^2\right)^{2.5}}\right] \\ +\dfrac{1}{\varpi^2}\cos\left(x_2\right) \\ \dot{x}_4 = \dfrac{-2x_3x_4}{x_1}+\dfrac{\Gamma}{\varpi}\left[\dfrac{\pi\left(1-2x_4\right)}{2\left(1-x_1^2\right)^{1.5}}+\dfrac{4x_3}{\left(1-x_1^2\right)^2}\right] \\ -\dfrac{1}{x_1\varpi^2}\sin\left(x_2\right) \end{cases} \tag{3}$$

For a given bearing parameter Γ, the system of equations (3) depends on ϖ and has the following form $\boldsymbol{X} = \boldsymbol{f}(\boldsymbol{X}, \varpi)$

2.2. The long journal bearing model

In the case of a long bearing, the axial flow can be neglected. The flow is then purely circumferential. The expressions of hydrodynamic force components are obtained as follows [21]:

$$\begin{cases} F_{\varepsilon} = \dfrac{6\mu LR^3\omega}{c^2}\left\{\dfrac{2\varepsilon^2\left(1-2\dot{\theta}\right)}{\left(2+\varepsilon^2\right)\left(1-\varepsilon^2\right)}+\dfrac{\left[\pi^2\left(2+\varepsilon^2\right)-16\right]\dot{\varepsilon}}{\pi\left(2+\varepsilon^2\right)\left(1-\varepsilon^2\right)^{1.5}}\right\} \\ F_{\theta} = \dfrac{6\mu LR^3\omega}{c^2}\left\{\dfrac{\pi\varepsilon\left(1-2\dot{\theta}\right)}{\left(2+\varepsilon^2\right)\left(1-\varepsilon^2\right)^{0.5}}+\dfrac{4\varepsilon\dot{\varepsilon}}{\left(2+\varepsilon^2\right)\left(1-\varepsilon^2\right)}\right\} \end{cases} \tag{4}$$

Using the nondimensional bearing parameter $\lambda = \dfrac{6\mu LR^3}{M\,c^{2.5}g^{0.5}}$ and $\varpi = \omega\sqrt{\dfrac{c}{g}}$, the dimensionless state space form $\boldsymbol{X} = \boldsymbol{f}(\boldsymbol{X},\varpi)$ that characterizes the development of the system state is $\boldsymbol{X} = \left(x_1 \quad x_2 \quad x_3 \quad x_4\right)^T = \left(\varepsilon \quad \theta \quad \dot{\varepsilon} \quad \dot{\theta}\right)^T$ produced and the following system of equations is obtained:

$$\begin{cases} \dot{x}_1 = x_3 \\ \dot{x}_2 = x_4 \\ \dot{x}_3 = x_1x_4^2 - \dfrac{\lambda}{\varpi}\left\{\dfrac{2x_1^2\left(1-2x_4\right)}{\left(1-x_1^2\right)\left(2+x_1^2\right)}+\dfrac{\left[\pi^2\left(2+x_1^2\right)-16\right]x_3}{\pi\left(2+x_1^2\right)\left(1-x_1^2\right)^{1.5}}\right\} \\ \quad +\dfrac{1}{\varpi^2}\cos\left(x_2\right) \\ \dot{x}_4 = \dfrac{-2x_3x_4}{x_1}+\dfrac{\lambda}{\varpi}\left\{\dfrac{\pi\left(1-2x_4\right)}{\left(2+x_1^2\right)\left(1-x_1^2\right)^{0.5}}+\dfrac{4x_3}{\left(2+x_1^2\right)\left(1-x_1^2\right)}\right\} \\ \quad -\dfrac{1}{x_1\varpi^2}\sin\left(x_2\right) \end{cases} \tag{5}$$

For a given bearing parameter λ and rotor speed ϖ, the nonlinear system of second order differential equations mentioned above may be solved.

3. Numerical continuation

It is necessary to have an initial steady state solution to use a continuation algorithm. For a given bearing parameter Γ and a nondimensional rotor speed ϖ, an equilibrium point for an autonomous nonlinear system of ordinary differential equations may be derived. This equilibrium point is unstable for a rotor speed greater than the stability threshold speed, and stable otherwise.

The numerical continuation is performed using the MatCont toolbox in MATLAB [20], which implements a predictor-corrector scheme. A tangent predictor generates an initial guess along the solution branch, with an initial step size of 0.1, adaptively adjusted within the range [10^{-5}, 0.5] based on the local behavior of the branch. The corrector phase relies on a pseudo-arclength Newton-Raphson

method, ensuring robust tracking of the solution branch including around turning points and bifurcations. Convergence is achieved when the residual norm satisfies $\|F(x)\| < 10^{-6}$, within a maximum of 10 Newton iterations per continuation step.

Subsequently, the branch of equilibrium points is found by numerical continuation using the rotor speed as a control parameter. The branch of limit cycles is then determined by continuation of limit cycles, starting from the Hopf bifurcation point. We are interested in finding possible bifurcations of limit cycles of the rigid rotor-bearing system for a selected value of the bearing parameter.

3.1 The short journal bearing: equilibrium point continuation and bifurcation detection

For a given bearing parameter Γ, numerical continuation extends the initial equilibrium point and tracks its path as the nondimensional rotating speed ϖ is varied where ϖ serves as the control parameter. When a Hopf bifurcation point 'H' is detected at $\varpi = \varpi_H = \varpi_c$, the first Lyapunov coefficient is automatically computed by MatCont. A negative Lyapunov coefficient indicates a supercritical Hopf bifurcation, in which stable limit cycles emerge from the bifurcation point for $\varpi \geq \varpi_H$. Conversely, a positive Lyapunov coefficient characterizes a subcritical Hopf bifurcation, where unstable limit cycles bifurcate from the Hopf point for $\varpi \leq \varpi_H$, potentially coexisting with the stable equilibrium and leading to more complex dynamic behavior.

As reported in [17], for short bearings, the value $\Gamma = 0.6$ acts as a transition threshold separating the two regimes: bearing parameters below this value give rise to supercritical behavior, while values above it lead to a subcritical behavior. To illustrate the subcritical regime, $\Gamma = 0.8$ is selected as a representative value, for which a subcritical Hopf bifurcation is observed at the threshold speed ϖ_H. Figure 2 illustrates the path of the initial equilibrium point obtained by numerical continuation, for the selected bearing parameter $\Gamma = 0.8$, as the nondimensional rotating speed ϖ is varied. The equilibrium curve reveals two distinct regions: the equilibrium point is stable for rotor speeds below the Hopf bifurcation point H, shown as a solid line, and becomes unstable beyond it, shown as a dashed line.

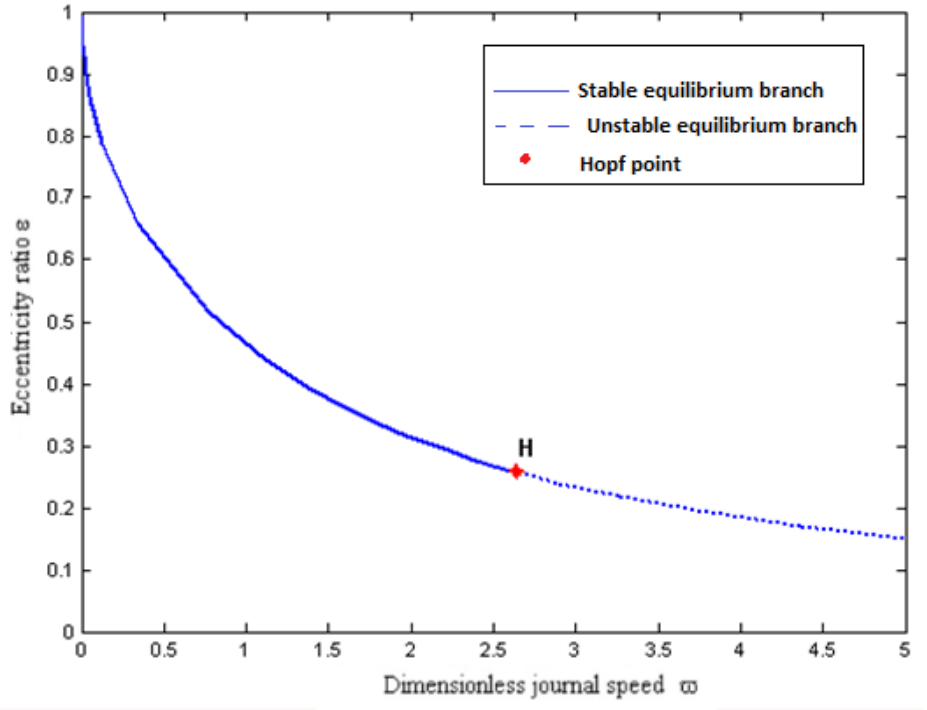


Figure 2: Equilibrium manifold in the (ϖ, ε) plane for a short journal bearing with bearing parameter $\Gamma = 0.8$

The Lyapunov coefficient for the Hopf point is equal to 1.86×10^{-2}. As a result, the Hopf point should give rise to an unstable limit cycle branch. The journal bearing behavior is then subcritical for rotational speeds below the threshold speed $\varpi < \varpi_c = 2.64$. The branch of limit cycles is determined by numerical continuation, which starts with the Hopf point and uses ϖ as a control parameter. Other bifurcations may also occur in limit cycles. For a rigid and balanced rotor, as long as the eccentricity limitations $\varepsilon = 1$ are not reached [15,22], limit cycles are only scheduled to cross a Limit Point of Cycles bifurcation (LPC).

At an LPC bifurcation point, stable and unstable periodic solution collide and disappear. An unstable limit cycle branch bifurcates from a stable limit cycle branch. The same thing happens when a branch of unstable limit cycles bifurcates into a branch of stable limit cycles. For the chosen bearing parameter $\Gamma = 0.8$, Only one bifurcation point has been found as seen in Figure 3.

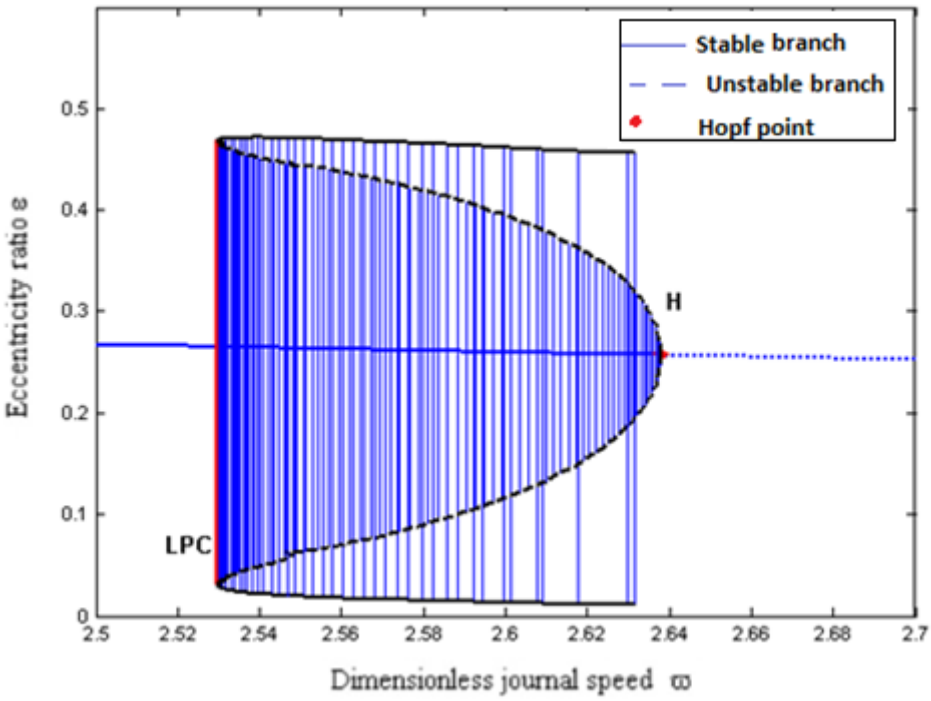


Figure 3: Subcritical limit cycle bifurcations for a bearing parameter $\Gamma = 0.8$

The unstable limit cycle in the range $\varpi_{LPC} < \varpi < \varpi_H$ defines the stable equilibrium solution's basin of attraction. The journal center returns to the stable equilibrium position if it is perturbed inside the basin of attraction. It diverges away from the basin of attraction to bigger stable limit cycles for greater disturbances, as seen in Figure 3.

For subcritical bifurcations, large unstable limit cycle exists until the saddle node (LPC) speed is reached. The ϖ_H and ϖ_{LPC} specify the rotor bearing system's hysteresis loop's boundaries. An unstable solution exists within the hysteresis loop. At a specific running speed, large disturbance exceeding the basin of stability causes the journal center to shift to large stable limit cycles, whereas disturbance within unstable orbit will decay to the equilibrium position.

The system of equations (3) for the short bearing model is numerically integrated to determine the journal center's trajectories to validate the predictions made by the numerical continuation approach. The pair (Γ, ϖ) must be chosen for the numerical simulation. Equations (3) are then integrated from the initial time up to a chosen end time, starting from a selected initial point $\left(\varepsilon_0 \quad \theta_0 \quad \dot{\varepsilon}_0 \quad \dot{\theta}_0\right)$ in the vicinity of the stability point.

The system tends to asymptotically approach the steady state equilibrium point for $\varpi = 2.54 < \varpi_c = 2.64$, if the journal center is released from a location inside the unstable periodic solution, in accordance with the properties of unstable periodic solutions. The initial positions are

shown with asterisk mark. Figure 4 depicts the locus of journal center that has been released from a position close to equilibrium point and has now settled to the steady state equilibrium point.

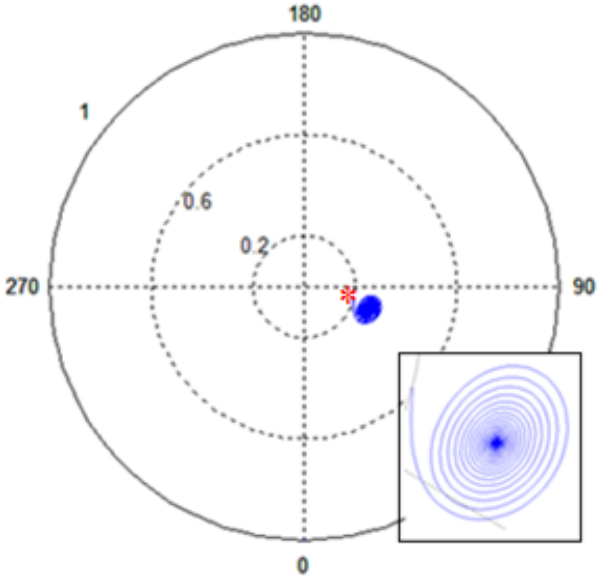


Figure 4: The journal center's trajectory deviated from its initial location inside the unstable limit cycle for $\Gamma = 0.8$ and $\varpi = 2.54$

The shaft is moving at the same speed $\varpi = 2.54$ during the second simulation, but it is let go from a starting position that is outside the unstable limit cycle. As shown in Figure 5, the rotor bearing system, in this instance, converges to a stable limit cycle.

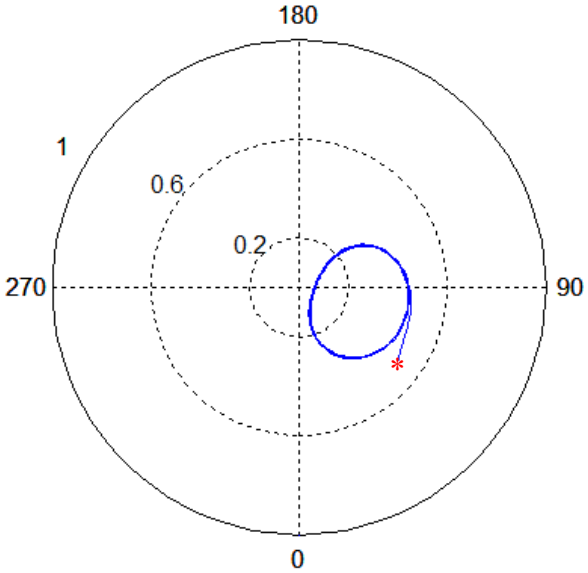


Figure 5: The journal center's trajectory deviated from a point outside the unstable limit cycle for $\Gamma = 0.8$ and $\varpi = 2.54$

3.2 The long journal bearing: equilibrium point continuation and bifurcation detection

The one parameter bifurcation analysis of equilibrium points and the detection of their bifurcations for the rotor bearing system with a journal bearing parameter $\lambda = 0.42$ is carried out. As reported in [5, 22], the bifurcation is subcritical for $\lambda \leq 0.27$ and $\lambda \geq 4.43$, and supercritical for $0.27 < \lambda < 4.43$. Accordingly, a bearing parameter $\lambda = 0.42$ is selected as a representative value within the supercritical range, for which a stable supercritical behavior is predicted.

The one parameter bifurcation analysis of equilibrium points and the detection of their bifurcations for the rotor bearing system with a journal parameter $\lambda = 0.42$ is carried out. A stable supercritical behavior is predicted, as reported in work [5, 22], for running speeds $\varpi > \varpi_c = 1.67$. A trajectory that converges with the steady state equilibrium position represents a behavior at a stable point. Beyond a supercritical bifurcation, the trajectories of the journal center must converge to a periodic limit cycle at beginning positions within or outside the limit cycle envelope for speeds higher than the critical stability speed ϖ_c. By numerical continuation, an equilibrium point is located and then continued with respect to the journal speed ϖ which represents the continuation parameter to define the path of the equilibrium point as shown in Figure 6.

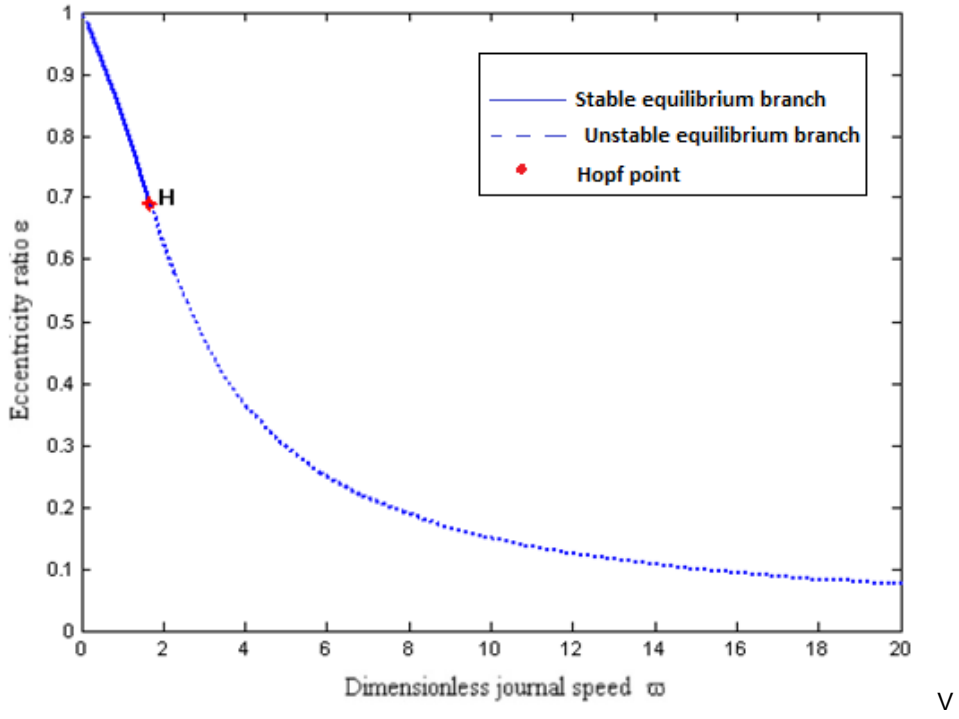


Figure 6: Equilibrium manifold in the (ϖ, ε) plane for a long journal bearing with bearing parameter $\lambda = 0.42$

Lyapunov coefficient is equal to -3.24×10^{-1} at $\varpi = \varpi_H = 1.67$, indicating a supercritical Hopf bifurcation since it is negative. Once a Hopf bifurcation is detected, numerical continuation of limit cycles is applied to determine the branch of stable limit cycles as shown in Figure 7.

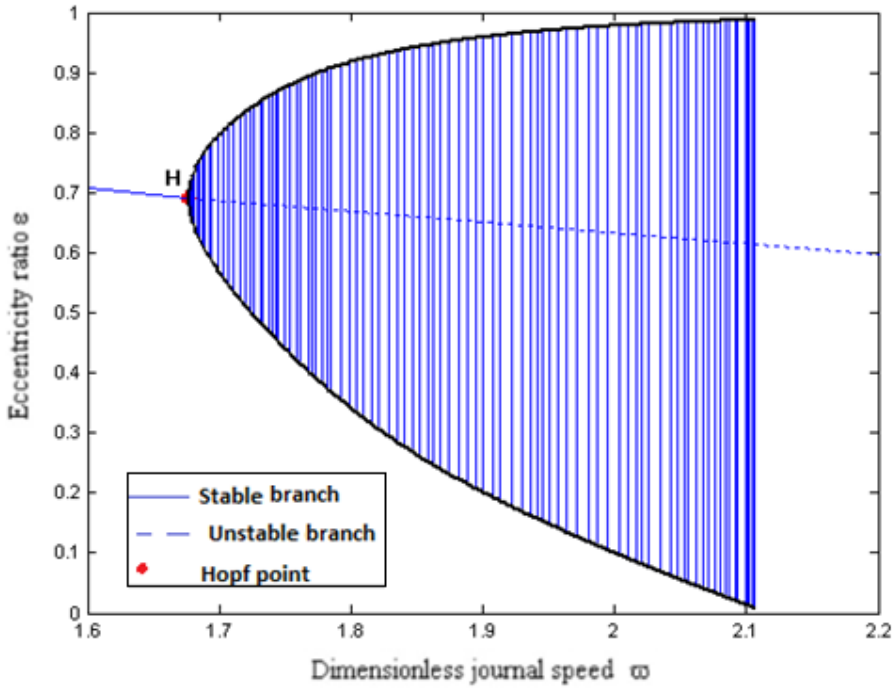


Figure 7: The branch of stable limit cycles bifurcating from the Hopf point H for a bearing parameter $\lambda = 4.2$

The stable limit cycle branch in Figure 7 has been extended until the clearance limit is reached. No LPC bifurcation is detected for the supercritical limit cycles before reaching the clearance limit. Increasing the running speed ϖ, the size of the limit cycles grows gradually with the running speed ϖ. Figure 7 reveals that the periodic solutions may be of high amplitude and involve contact with the bearing's housing.

Notably, a low amplitude limit cycle appears directly as the rotor speed exceeds the threshold speed. A numerical simulation is carried out to verify the above results and to investigate how the whirl orbits develop at rotor speeds away from the threshold speed.

Figures 8 (a) and 8 (b) show the existence of stable limit cycles for dimensionless speed $\varpi = 2 > \varpi_c = 1.67$.

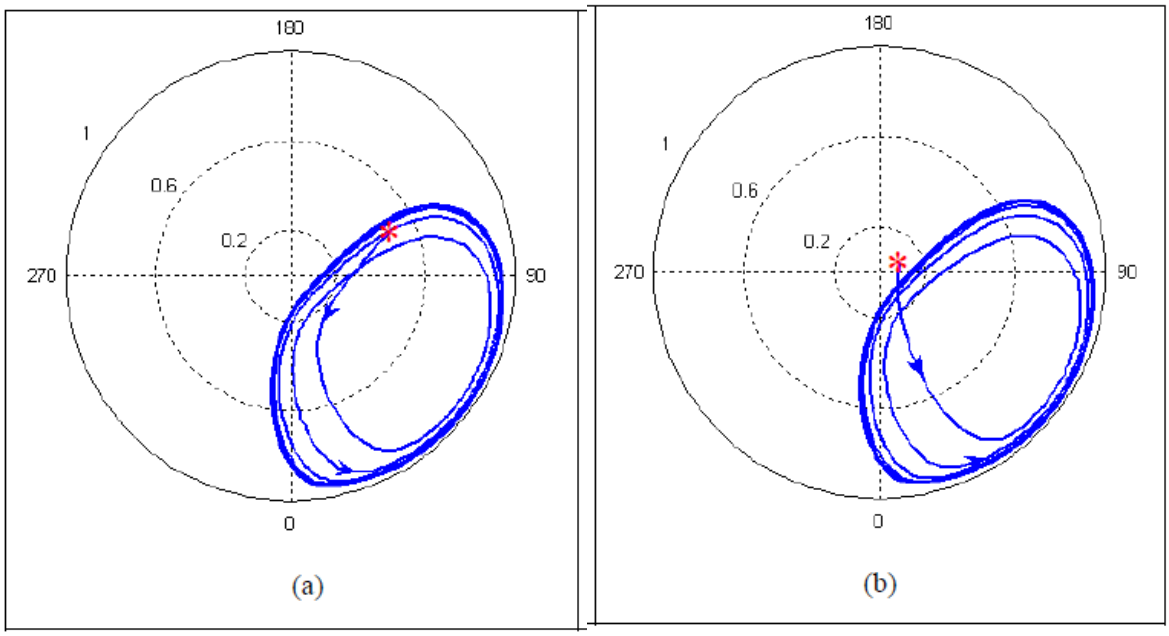


Figure 8: Phase plane trajectories of the journal center for $\lambda = 0.42, \varpi = 2$: (a) initial point in the stable region, (b) initial point outside the stable region

It can be noted that the trajectories of journal center converge to the same limit cycle whether the initial point is chosen inside the limit cycle envelope, Figure 8 (a), or outside the limit cycle envelope, Figure 8 (b).

4. Conclusion

A comprehensive nonlinear analysis of stability and bifurcation behaviors is presented for a perfectly balanced rigid rotor supported by identical hydrodynamic journal bearings using numerical continuation via the MatCont toolbox. Both short and long bearing models, formulated under the half-Sommerfeld boundary conditions, are considered, allowing a broad and realistic representation of bearing configurations.

The study reveals fundamentally different dynamic behaviors depending on key design parameters such as rotational speed, applied load, the bearings lubrication characteristics, and the bearing geometry, as characterized by the aspect ratio D/L. Indeed, Hopf bifurcations may be either subcritical or supercritical. As reported in [17], for the short bearing (D/L < 0.5), the bifurcation is subcritical for $\Gamma \geq 0.6$ and supercritical for $\Gamma < 0.6$, while for the long bearing (D/L > 1), the bifurcation is subcritical for $\lambda \leq 0.27$ and $\lambda \geq 4.43$, and supercritical for $0.27 < \lambda < 4.43$ as demonstrated in [5,22].

To illustrate these contrasting behaviors, two representative cases are examined. For the short bearing with $\Gamma = 0.8$, a subcritical Hopf bifurcation is observed at the threshold speed $\varpi_H = 2.64$, giving rise

to unstable limit cycles below this speed and indicating an abrupt and potentially dangerous loss of stability. In contrast, for the long bearing with $\lambda = 0.42$, a supercritical Hopf bifurcation occurs at the threshold speed $\varpi_H = 1.67$, producing stable limit cycles beyond this speed and reflecting a safer and more predictable dynamic transition. The short bearing reaches instability at a higher threshold speed but in a more abrupt and dangerous manner, while the long bearing loses stability earlier but more gradually. These two cases clearly highlight how the aspect ratio D/L, together with the bearing parameter, governs the nature of the bifurcation and consequently the safety of the rotor bearing system.

Numerical continuation of the limit cycles emanating from the Hopf bifurcation points, together with the detection of limit point of cycles (LPC) bifurcations, enables an accurate delineation of the system's stability boundaries and the identification of operating regions prone to hysteresis phenomena. A detailed numerical investigation corroborates the results obtained through continuation methods, confirming the validity and robustness of the proposed approach.

The findings of this work provide valuable guidelines for selecting appropriate system parameters to prevent undesirable journal center trajectories and self-excited vibrations, thereby enhancing the operational reliability, durability, and overall service life of rotor bearing systems. Future work will focus on extending this analysis to more complex configurations, including floating ring and tilting-pad bearings, and on investigating the influence of thermal effects, bearing misalignment, and shaft flexibility on the predicted bifurcation thresholds.

Funding

The authors declare that this research received no external funding.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] B. L. Newkirk, Shaft Whipping due to Oil Action in Journal Bearings, G.E. Rev. 27(1924) 169-178.

[2] P. Hollis, D. Taylor, Hopf Bifurcation to Limit Cycles in Fluid Film Bearings. J. Tribol. Trans. ASME. 108(1998) 184-189.

[3] R. Holmes, The Vibration of a Rigid Shaft on Short Sleeve Bearings, J. Mech. Eng. Sci. 2 (1960) 337-341.

[4] M.M. Reddi, P.R. Trumpler, Stability of the high-speed journal bearing under steady load, ASME J. Eng. Ind. 84 (1962) 351–358.

[5] C. J. Myers, Bifurcation theory applied to oil whirl in plain cylindrical journal bearings, J. Appl. Mech. Trans. ASME. 51(1984) 244-250.
[6] J. K. Wang, M. M. Khonsari, Application of Hopf bifurcation theory to rotor-bearing systems with consideration of turbulent effects, J. Tribol.Int.39(7) (2006) 701- 714.
[7] J. W. Lund, Oil Whip Whirl Orbit of a Rotor in Sleeve Bearings, ASME J. Eng. Ind. 89 (1967) 813-823.
[8] D. W. Childs, Turbomachinery Rotordynamics: Phenomena, Modeling, and Analysis.
Wiley, New York,1993.
[9] F.F. Ehrich, Handbook of Rotordynamics, McGraw-Hill, New York,1992.
[10] L. San Andrés, C. F. Taylor, Experimental identification of nonlinear rotordynamic coefficients in fluid-film bearings, J. Tribol.Trans. ASME. 118 (1996) 39–47.
[11] R. W. Brockett, A. El-Sayed, Nonlinear transient response of journal-bearing supported rotors. J. Sound Vib. 201(1) (1997) 1–15.
[12] A. Tondl, To the problem of self-excited vibrations in rotating machinery.Eng.Mech**.**7(1) (2000) 1–15.
[13] H. Zhang, G. Meng, Nonlinear dynamic modeling of rotor–bearing systems with fluid-film forces. J. Sound Vib. 273 (2004) 175–193.
[14] A. Amamou, Nonlinear stability analysis and numerical continuation of bifurcations of a rotor supported by floating ring bearings. Proc. IMechE Part C: J. Mech. Eng. Sci. 236 (2021) 1-13.
[15] Yu. A. Kuznetsov, Elements of Applied Bifurcation Theory, Springer-Verlag, 1998.
[16] B. Krauskopf, Numerical continuation methods for dynamical systems: path following and boundary value problem, Springer,2007.
[17] M. Chouchane, A. Amamou, Bifurcation of Limit Cycles in Fluid Film Bearings, Int. J. Non-Linear Mech.46 (2011) 1258-1264.
[18] E. J. Doedel et al., AUTO-07P: Continuation and Bifurcation Software for Ordinary Differential Equations (manual), January 2012
[19] Yu. A. Kuznetsov, H. G. E. Meijer, Numerical Bifurcation Analysis of Maps, Cambridge University Press, New York,2019.
[20] A. Dhooge, W. Govaerts, and Yu. A. Kuznetsov, MATCONT: A MATLAB package for continuation and bifurcation analysis, ACM Trans. Math. Softw. 29 (2003) 141 - 164.
[21] J. Frene, et al., Hydrodynamic Lubrication Bearings and Thrust Bearings, Elsevier, Tribology Series, Amsterdam,1997.
[22] A. Amamou, M. Chouchane, Nonlinear stability analysis of long hydrodynamic journal bearings using numerical continuation, Mech. Mach. Theory. 72 (2014) 17-24.

Islamic University Journal of Applied Sciences (IUJAS)

https://journals.iu.edu.sa/jesc

Volume VIII, Issue I, July 2026, Pages 168-188

Theoretical investigation using DFT Calculations, Vibrational Assignments, Infrared Spectrum and Normal Coordinate Analysis of (2-Chloro-4-fluorophenyl) boronic Acid

Usama A. Soliman[1]*

[1] Chemistry Division, Department of Physical Sciences, College of Science, Jazan University, P.O. Box. 114, Jazan, 45142, Kingdom of Saudi Arabia.

* Corresponding author: (Usama A. Soliman) , *Email Address*: usoliman@jazanu.edu.sa

Abstract

The geometry, vibrational frequencies, and thermodynamic parameters of 2-chloro-4-fluorophenylboronic acid (CFPB; $C_6H_5BClFO_2$) have been computed using the Gaussian 09 software package. Initially, four distinct rotamers were proposed for this compound, arising from the internal rotation of the hydroxyl groups around the C–B bond. A conformational stability analysis for an isolated molecule was performed using the B3LYP functional with the standard 6-311++G(d,p) basis set. Our calculations identified the CT structure where one OH group is *cis* to both the phenyl Cl and F atoms, and the other OH is *trans* to the phenyl Cl atom—as the most stable conformation. The energy differences between the CT conformer and the others (2-4) range from 1700 to 2000 cm^{-1} (4.88 – 5.72 kcal/mol). This stability is attributed to stabilizing O_5...H14 and Cl_6....H15 intramolecular hydrogen bonds. A complete vibrational assignment of the observed IR bands is provided, supported by calculated wavenumbers, IR intensities, and a normal coordinate analysis. The computed wavenumbers and IR intensities for the CT form agree well with the experimental data.

Keywords: 2-Chloro-4-fluorophenylboronic acid; B3LYP calculations; Infrared spectroscopy; Normal coordinate analysis.

https://doi.org/10.63070/jesc.2026.022

Received 17 January 2026; Revised 27 March 2026; Accepted 10 April 2026.

Available online 15 April 2026.

Published by Islamic University of Madinah on behalf of *Islamic University Journal of Applied Sciences*.

This is an open access article published under the Creative Commons Attribution 4.0 International license (CC BY 4.0):https://creativecommons.org/licenses/by/4.0/

1. Introduction

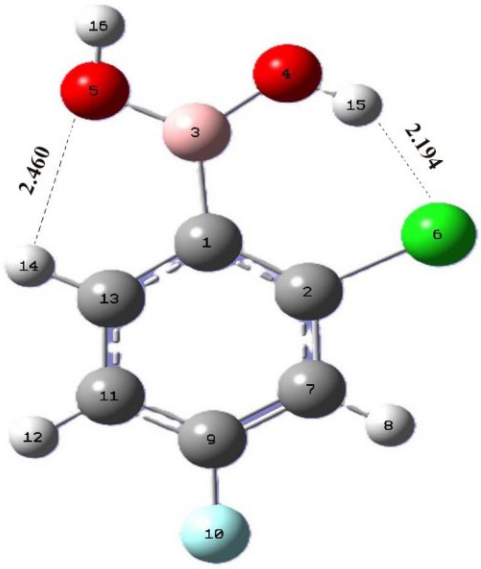
Boronic acid derivatives represent a class of compounds of significant interest due to their broad applications in pharmaceuticals [1,2] and diverse physiological activities [3,4]. They are also vital in synthesizing chemical building blocks, intermediates [5], biaryl compounds [6,7], and antihypertensive agents [8]. Their biological utility extends to roles as enzyme inhibitors, catalysts, protecting groups, saccharide-sensing receptors, neutron capture therapy agents, and anti-metabolites for cancer treatment [9-12]. Phenylboronic acid, featuring a phenyl ring substituted with a boron atom bonded to two hydroxyl groups, is a fundamental structure among these derivatives. The single-crystal X-ray structure of phenylboronic acid was determined by Rettig and Trotter [16], while its dimer was reported by Cyrański et al. [17]. The structure of 2,4-difluorophenylboronic acid was detailed by Rodriguez et al. [18]. Additionally, X-ray diffraction studies have been conducted on 4-chloro- and 4-bromophenylboronic acids, as well as the hydrates of 2- and 4-iodophenylboronic acid [19]. The infrared spectra of phenylboronic acid [20], diphenyl phenylboronate [20], and several derivatives—including 4-chloro- [21], 3-bromo- [22], 3,5-dichloro- [23,24], and 2,6-dimethoxyphenylboronic acid [25] have been examined both experimentally and theoretically. Conformational analysis, vibrational spectra, and electronic transitions have also been explored for 2,3- and 3,5-difluorophenylboronic acid molecules [26,27] and methylboronic acid [28]. To the best of our knowledge, a comprehensive theoretical or experimental study on 2-chloro-4-fluorophenylboronic acid (CFPB) has not been reported, although its IR spectrum is not available. This study aims to address this gap by investigating the conformational stability and vibrational assignments of CFPB using DFT quantum mechanical calculations. We also performed a normal coordinate analysis to elucidate the normal modes of vibration, supported by potential energy distributions (PEDs) and force constant determinations.

2. Computational Details

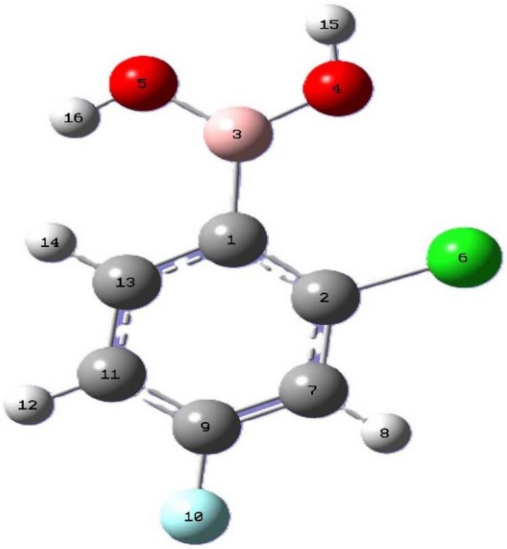
2.1. Conformational Isomerism and DFT Calculations

The rotational isomerism in CFPB (with an assumed C_s symmetry) arises from the two hydroxyl groups attached to the boron atom, which is connected to the phenyl ring at the C_1 position. These structures are designated as **CT, TC, TT, and CC**, where 'C' denotes *cis* and 'T' denotes *trans* orientations of the OH groups relative to the phenyl C_1 atom. Calculating vibrational spectra is computationally more intensive than geometry optimization [29], as it requires a normal mode analysis involving the second derivatives of energy with respect to nuclear displacements.

For molecules of moderate size, Density Functional Theory (DFT) calculations have proven highly effective for both structural and frequency analyses [30, 31]. The availability of analytical second derivative techniques has significantly advanced quantum mechanical calculations, making vibrational frequency computations feasible for larger molecules [32]. Optimizing the CFPB conformers can be sensitive to initial approximations, and shifts in nuclear positions during optimization may lead to saddle points or local energy minima. Saddle points are problematic as they yield imaginary vibrational frequencies, while local minima can misrepresent conformational interconversions and vibrational spectral features. The inherent flexibility and mobility of the hydroxyl group around the single bond, coupled with potential intramolecular hydrogen bonding, contribute to these optimization challenges. To mitigate these issues, the molecular geometry of 2-chloro-4-fluorophenylboronic acid was constrained to C_s symmetry during the optimization and frequency calculations for all conformers. Geometry optimizations for the CFPB conformers **(Fig. 1)** were performed using the hybrid B3LYP functional [33-35] and the 6-311++G(d,p) basis set, as implemented in the Gaussian 09 package [36]. This basis set includes polarization and diffusion functions. Energy minima with respect to nuclear coordinates were located by relaxing all geometric parameters simultaneously using the Pulay gradient method [37], with specific dihedral angles constrained to maintain C_s symmetry. Vibrational frequencies were subsequently calculated at the same level of theory. Full convergence was achieved, with the maximum force below 0.00045 mdyne and the maximum displacement below 0.0018 Å. The B3LYP/6-311++G(d,p) results conclusively identified the CT structure **(Conformer 1 in Fig. 1)** as the global minimum, characterized by the lowest energy. In contrast, the other isomers (TC, TT, CC) were identified as transition states, as evidenced by their higher relative energies and the presence of one imaginary frequency in their vibrational analysis; they were therefore excluded from further analysis. The energy differences between the CT conformer and the others (2-4) range from 1700 to 2000 cm^{-1} (4.88 – 5.72 kcal/mol), as illustrated in **Fig. 1**. Selected structural parameters (SPs) calculated at the B3LYP/6-31G(d) and B3LYP/6-311++G(d,p) levels for the most stable CT conformer are listed in **Table 1**, alongside X-ray crystallographic data for relevant phenylboronic acid derivatives [16,38,18].

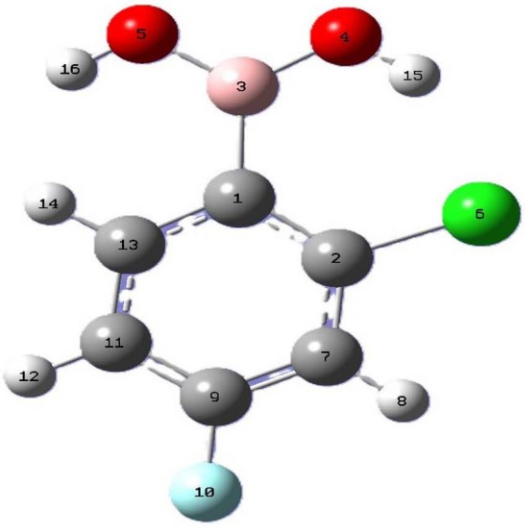


Structure CT
Minimum



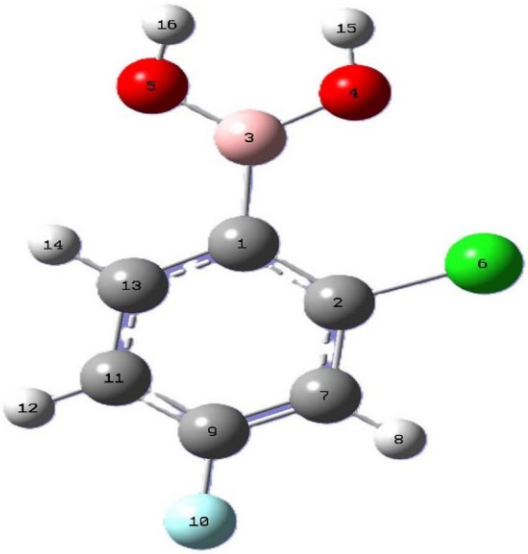
Structure TC

ΔE= 1720cm^{-1}(4.92 kcal/mol)
1707cm^{-1}(4.88 kcal/mol)



Structure CC

ΔE= 1728cm^{-1}(4.94 kcal/mol)
1722cm^{-1}(4.92 kcal/mol)



Structure TT

ΔE= 2000cm^{-1}(5.72 kcal/mol)
1930cm^{-1}(5.52 kcal/mol)

Fig 1. Total energies differences (***bold italic*** utilizing 6-311++G** and normal line utilizing 6-31G* basis sets) between **CT** conformer and other **CC, TT** and **TC** conformers at b3lyp method in cm^{-1}and kcal/mol

TABLE.1. B3LYP structural parameters for 2-Chloro-4-florophenylboronic acid (conformer CT) utilizing 6-31G(d) and 6-311++G(d,p) basis sets.

Parameters	X-ray Ref. [16]	X-ray Ref. [38]	X-ray Ref. [18]	*B3LYP for Structure(CT)* 6-31G(d)	6-311++G(d,p)
r($C_1C_{2\ ring}$)	1.404	1.406	1.382	1.405	1.402
r($C_2C_{7\ ring}$)	1.389	1.377	1.370	1.394	1.391
r($C_7C_{9\ ring}$)	1.378	1.365	1.363	1.389	1.385
r ($C_9C_{11\ ring}$)	1.384	1.375	1.366	1.389	1.385
r($C_{11}C_{13\ ring}$)	1.390	1.387	1.374	1.391	1.389
r($C_{13}C_{1ring}$)	1.402	1.398	1.394	1.414	1.412
r($C_7H_{8\ ring}$)	1.00	0.93	0.93	1.083	1.081
r($C_{11}H_{12\ ring}$)	1.00	0.93	0.93	1.085	1.082
r($C_{13}H_{14\ ring}$)	1.00	0.93	0.93	1.086	1.083
r(C_2— Cl_6)	—	—	—	1.780	1.777
r(C_9—F_{10})	—	—	1.364	1.345	1.350
r(C_1—B_3)	1.568	1.562	1.566	1.583	1.579
r(B_3—O_4)	1.378	1.343	1.388	1.367	1.366
r(B_3—O_5)	1.362	1.366	1.361	1.367	1.366
r(O_4H_{15})	0.75	0.855	0.841	0.968	0.963
r(O_5H_{16})	0.75	0.855	0.841	0.968	0.963
r(Cl.....H_{15})	—	—	2.16	2.200	2.194
r(O_5.....H_{14})	—	—	—	2.453	2.460
∠($C_1C_2C_7$)	121.8	123.9	125.1	123.4	123.2
∠($C_2C_7C_9$)	119.5	118.6	116.4	118.2	118.1
∠($C_7C_9C_{11}$)	120.3	118.4	123.0	121.9	122.2
∠($C_9C_{11}C_{13}$)	120.1	119.7	117.9	117.8	117.7
∠($C_{11}C_{13}C_1$)	121.1	121.8	122.9	123.6	123.5
∠($C_{13}C_1C_2$)	117.2	117.6	114.6	115.0	115.3
∠($C_2C_1B_3$)	122.0	122.1	125.3	129.3	128.9
∠ ($C_{13}C_1B_3$)	120.8	120.1	120.1	115.6	115.8
∠ ($C_1B_3O_4$)	118.7	119.0	123.8	121.3	121.3
∠ ($C_1B_3O_5$)	125.0	122.7	117.4	121.3	121.3
∠ ($B_3O_4H_{15}$)	111.0	116.0	116.0	111.8	113.2
∠ ($B_3O_5H_{16}$)	111.0	124.0	115.0	111.8	113.2
∠ ($O_4B_3O_5$)	116.3	118.2	118.7	117.5	117.4
∠ ($C_1C_2Cl_6$)	120.0	—	—	122.2	122.2
∠ ($C_7C_2Cl_6$)	120.0	—	—	114.5	114.6
∠ ($C_2C_7H_8$)	120.0	—	121.8	121.6	121.6
∠ ($C_9C_7H_8$)	120.0	—	121.8	120.2	120.3
∠ ($C_7C_9F_{10}$)	—	118.1	118.2	118.5	118.4
∠ ($C_{11}C_9F_{10}$)	—	118.0	116.7	119.6	119.4
∠ ($C_9C_{11}H_{12}$)	120.0	119.1	121.0	120.0	120.2
∠ ($C_{13}C_{11}H_{12}$)	120.0	119.1	121.0	122.1	122.1
∠ ($C_{11}C_{13}H_{14}$)	120.0	120.7	118.5	118.9	119.0
∠ ($C_1C_{13}H_{14}$)	120.0	120.7	118.5	117.4	117.6

2.2. Calculated IR Spectra

The IR spectrum of the CT conformer was simulated using the computed wavenumbers and IR intensities and compared with the experimentally observed spectra recorded in KBr disk and Nujol mull **(Figs. 2A, 2B)** from Ref. [39]. The frequencies and infrared intensities were derived from the *ab initio* calculations. IR intensities were determined based on the derivatives of the dipole moment with respect to Cartesian coordinates. These derivatives, obtained from the *ab initio* output, were transformed into normal coordinates using the equation[39]: using $\frac{\partial \mu_u}{\partial Q_i} = \left(\sum_j \frac{\partial \mu_u}{\partial X_j}\right) L_{ij}$,where$Q_i$ is the ith Cartesian displacement coordinate, and $L_{ij}$is the transformation matrix between the Cartesian displacement coordinates and normal coordinates. The infrared intensities were then calculated with the following formula[39]: $I_i = \frac{N\pi}{3c^2}\left[\left(\frac{\partial \mu_x}{\partial Q_i}\right)^2 + \left(\frac{\partial \mu_y}{\partial Q_i}\right)^2 + \left(\frac{\partial \mu_z}{\partial Q_i}\right)^2\right]$. The simulated IR spectrum is in reasonably good agreements with the observed spectrum of the sample with the exception of some bands with high intensities are reversed in the calculated spectrum.

2.3. Normal Coordinate Analysis

The molecule possesses 42 fundamental vibrational modes, distributed as 29 in-plane (A') and 13 out-of-plane (A") modes, all of which are both IR and Raman active. A set of 55 internal coordinates **(Fig. 3)** was used to construct 42 local symmetry coordinates **(Table 2)** following Wilson's GF matrix method [40]. While 13 of these internal coordinates are redundant, they are necessary for determining the force constants. The Cartesian coordinates from the optimized structure were input into a G-matrix program to generate the B-matrix, which was then used in an F-matrix program similar to the one developed by Schachtschneider [41]. This program converts the force constants (FCs) from Cartesian coordinates into FCs in the chosen internal coordinates. The resulting diagonal FCs in internal coordinates, along with the implemented scaling factors (SFs), are listed in **Table 3**. These FCs were employed in a mass-weighted Cartesian coordinate calculation to reproduce the vibrational frequencies and determine the potential energy distributions (PEDs). It is well-known that theoretical FCs and vibrational frequencies tend to be

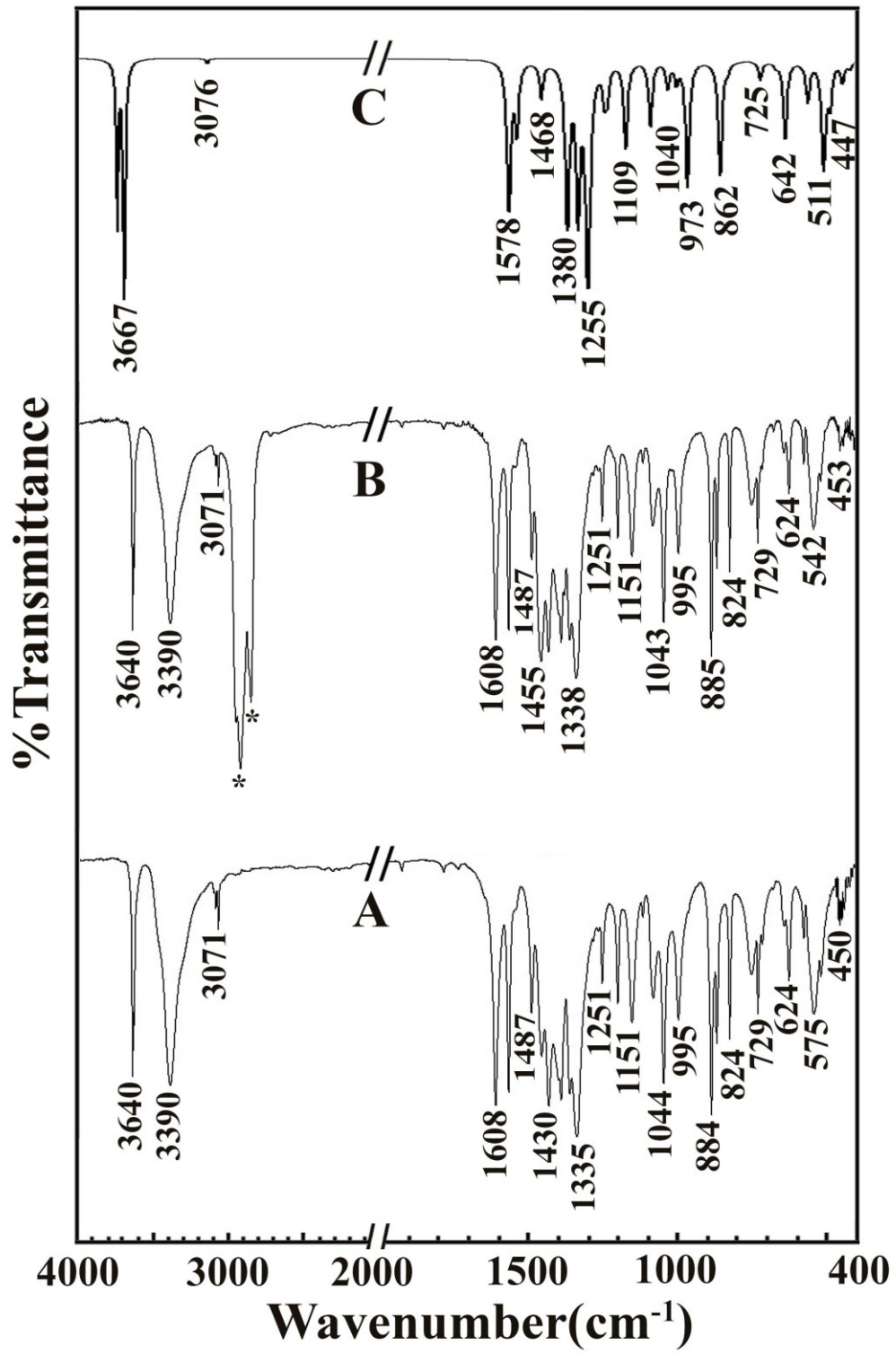


Fig. 2. IR spectrum in KBr disk of 2-Floro-4-chlorophenylboronic acid (A) from 400 to 4000 cm^{-1}; (B) Experimental IR spectrum in Nujol (C) scaled B3LYP/6-311++G(d,P) for 2-chloro-4-Florophenyl boronic acid (**CT**) conformer.

TABLE.2. Symmetry coordinates[a] for 2-Chloro-4-florophenylboronic acid, $C_6H_5FClBO_2$ (-d_0), $C_6H_3FClD_2BO_2$ (2OD)[b] and $C_6H_2FClD_3BO_2$ (3CD)[c] molecules.

ν_i[d]	Definition	Symmetry Coordinate
ν_1	OH/OD[b] Stretch	$S_1 = m_1 + m_2$
ν_2	OH/OD[b] Stretch	$S_2 = m_1 - m_2$
ν_3	CH/CD[c] Stretch	$S_3 = r_1 + r_2 + r_3$
ν_4	CH/CD[c] Stretch	$S_4 = 2r_1 - r_2 - r_3$
ν_5	CH/CD[c] Stretch	$S_5 = r_2 - r_3$
ν_6	CC Ring Stretch	$S_6 = q_1 + q_3 + q_5 - q_2 - q_4 - q_6$
ν_7	CC Ring Stretch	$S_7 = 2q_1 - q_3 - q_5 + 2q_2 - q_4 - q_6$
ν_8	CH/CD[c] In plan deformation	$S_8 = \alpha_1 - \alpha_2$
ν_9	CC Ring Stretch	$S_9 = q_3 - q_5 + q_4 - q_6$
ν_{10}	B−O Stretch	$S_{10} = a_1 + a_2$
ν_{11}	B−O Stretch	$S_{11} = a_1 - a_2$
ν_{12}	CC Ring Stretch	$S_{12} = 2q_1 - q_3 - q_5 - 2q_2 + q_4 + q_6$
ν_{13}	CH/CD[c] In plan deformation	$S_{13} = \alpha_3 - \alpha_4$
ν_{14}	C−F Stretch	$S_{14} = R$
ν_{15}	CH/CD[c] In plan deformation	$S_{15} = \alpha_5 - \alpha_6$
ν_{16}	C−B Stretch	$S_{16} = T$
ν_{17}	CC Ring Stretch	$S_{17} = q_3 - q_5 - q_4 + q_6$
ν_{18}	OH/OD[b] In plan deformation	$S_{18} = \lambda_1 - \lambda_2$
ν_{19}	OH/OD[b] In plan deformation	$S_{19} = \lambda_1 + \lambda_2$
ν_{20}	CH/CD[c] wag	$S_{20} = \gamma_1$
ν_{21}	Ring breath	$S_{21} = q_1 + q_3 + q_5 + q_2 + q_4 + q_6$
ν_{22}	CH/CD[c] wag	$S_{22} = \gamma_2$
ν_{23}	CH/CD[c] wag	$S_{23} = \gamma_3$
ν_{24}	CB wag	$S_{24} = \gamma_6$
ν_{25}	CCC Ring bending	$S_{25} = \theta_1 + \theta_3 + \theta_5 - \theta_2 - \theta_4 - \theta_6$
ν_{26}	Ring puckering	$S_{26} = \tau_1 - \tau_2 + \tau_3 - \tau_4 + \tau_5 - \tau_6$
ν_{27}	C−Cl Stretch	$S_{27} = E$
ν_{28}	OH/OD[b] torsion	$S_{28} = \eta_1$
ν_{29}	CF wag	$S_{29} = \gamma_5$
ν_{30}	CCC Ring bending	$S_{30} = \theta_2 - \theta_3 + \theta_5 - \theta_6$
ν_{31}	OH/OD[b] torsion	$S_{31} = \eta_2$
ν_{32}	C−F In plan deformation	$S_{32} = \beta_1 - \beta_2$
ν_{33}	Ring puckering	$S_{33} = \tau_1 - \tau_2 + \tau_4 - \tau_5$
ν_{34}	$B(OH)_2$ Scissor	$S_{34} = 2\Phi - \rho_1 - \rho_2$
ν_{35}	BO_2 Rock	$S_{35} = \rho_1 - \rho_2$
ν_{36}	CCC Ring bending	$S_{36} = 2\theta_1 - \theta_2 - \theta_3 + 2\theta_4 - \theta_5 - \theta_6$
ν_{37}	C−Cl In plan deformation	$S_{37} = \delta_1 - \delta_2$
ν_{38}	Ring twist	$S_{38} = \eta_3$
ν_{39}	CCl wag	$S_{39} = \gamma_4$

TABLE.2. Continued.

ν_i[d]	Definition	Symmetry Coordinate
ν_{40}	$CCB(OH)_2$ bending	$S_{40} = \varepsilon_1 - \varepsilon_2$
ν_{41}	Ring puckering	$S_{41} = 2\tau_3 - \tau_1 - \tau_2 + 2\tau_6 - \tau_4 - \tau_5$
ν_{42}	CCBO torsion	$S_{42} = \eta_4$

[a]Symmetry coordinates are not normalized.

[b]2-4-CFPB -d_2, $C_6H_3FClD_2BO_2$ deutrated two OH group.

[c]2-4-CFPB -d_3, $C_6H_2FClD_3BO_2$ deutrated three CH group.

[d]ν_{20},ν_{22}-ν_{24},ν_{26},ν_{28},ν_{29},ν_{31},ν_{33},ν_{38},ν_{39},ν_{41},ν_{42} belongs to A″ species and other symmetries belongs to A′species.

Overestimated due to systematic errors, including the neglect of electron correlation and basis set limitations [42]. Therefore, the wavenumbers calculated at the B3LYP/6-311++G(d,p) level were scaled by factors of 0.93 for stretches, 0.945 for in-plane deformations, and 0.978 for out-of-plane bending modes (wagging and torsions) to improve agreement with experimental values [43]. Furthermore, frequency calculations were performed for two isotopes of CFPB: C_2FPB-d_2 (O-D deuterated) and CFPB-d_3 (C-D deuterated). This helped to unequivocally distinguish between OH and CH bending modes and other fundamentals appearing in similar spectral regions. The final scaled harmonic vibrational frequencies, infrared intensities, Raman activities, and PEDs for all normal modes are presented in **Table 4**.

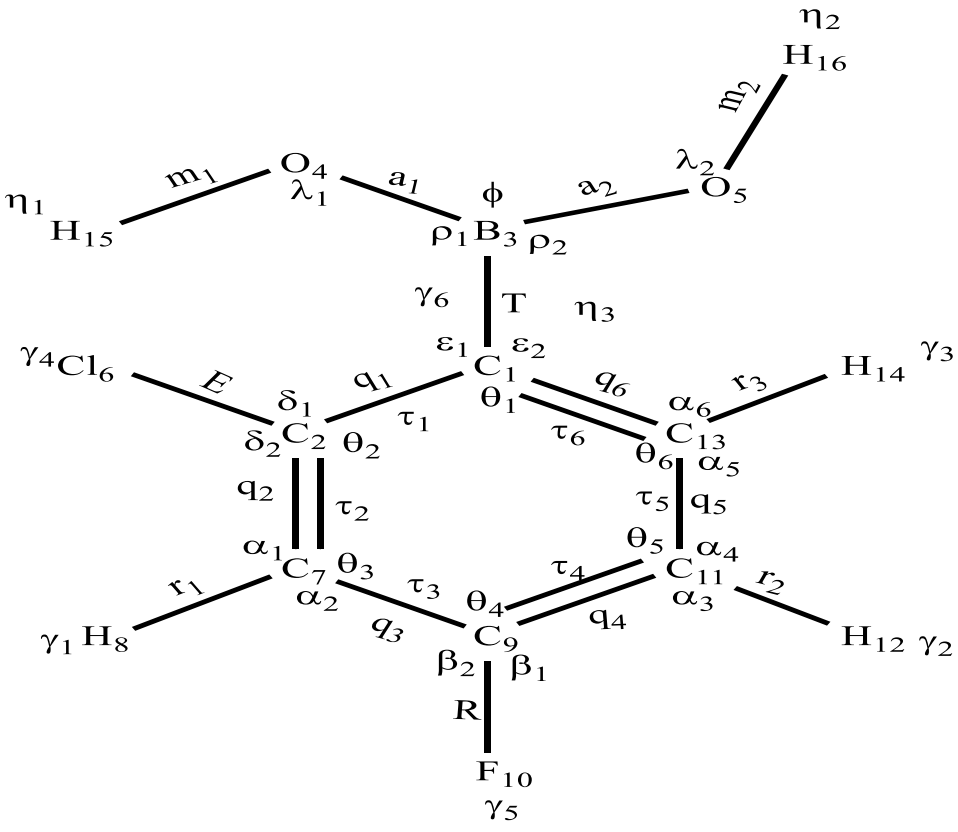


Fig. 3 Internal coordinates definition of 2-Chloro-4-fluorophenyl boronic acid(CFPB) **CT** coformer.

TABLE. 3. B3LYP/6-311++G(d,p) unscaled force constants (mdyne/Å) for 2-Chloro-4-florophenylboronic acid (conformer CT) molecule

Atom #	Vib. modes	Symbol[a]	Scaling factor[b]	Diagonal force constant[c]
C_1—C_2	Stretch	q_1	0.930	5.096
C_2—C_7	Stretch	q_2	0.930	5.192
C_7—C_9	Stretch	q_3	0.930	5.325
C_9—C_{11}	Stretch	q_4	0.930	5.320
C_{11}—C_{13}	Stretch	q_5	0.930	5.292
C_{13}—C_1	Stretch	q_6	0.930	4.801
C_1—B_3	Stretch	T	0.930	3.988
C_2—F_6	Stretch	R	0.930	6.389
C_9—Cl_{10}	Stretch	E	0.930	3.154
B_3—O_4	Stretch	a_1	0.930	6.034
B_3—O_5	Stretch	a_2	0.930	6.513
O_4H_{15}	Stretch	m_1	0.930	8.092
O_5H_{16}	Stretch	m_2	0.930	8.309
C_7H_8	Stretch	r_1	0.930	5.670
C_9H_{10}	Stretch	r_2	0.930	5.628
$C_{13}H_{14}$	Stretch	r_3	0.930	5.590
$B_3C_1C_2$	Bending	ε_1	0.945	1.030
$B_3C_1C_{13}$	Bending	ε_2	0.945	0.807
$C_1C_2F_6$	Bending	β_1	0.945	0.942
$C_7C_2F_6$	Bending	β_2	0.945	0.856
$C_7C_9Cl_{10}$	Bending	δ_1	0.945	0.663
$C_{11}C_9Cl_{10}$	Bending	δ_2	0.945	0.659
$C_2C_7H_8$	Bending	α_1	0.945	0.433
$C_9C_7H_8$	Bending	α_2	0.945	0.420
$C_9C_{11}H_{12}$	Bending	α_3	0.945	0.420
$C_{13}C_{11}H_{12}$	Bending	α_4	0.945	0.427
$C_{11}C_{13}H_{14}$	Bending	α_5	0.945	0.455
$C_1C_{13}H_{14}$	Bending	α_6	0.945	0.471
$C_1B_3O_4$	Bending	ρ_1	0.945	0.796
$C_1B_3O_5$	Bending	ρ_2	0.945	0.622
$O_4B_3O_5$	Bending	Φ	0.945	0.679
$H_{15}O_4B_3$	Bending	λ_1	0.945	0.626
$H_{16}O_5B_3$	Bending	λ_2	0.945	0.508
$C_{13}C_1C_2$	Bending	θ_1	0.945	0.787
$C_1C_2C_7$	Bending	θ_2	0.945	0.742
$C_2C_7C_9$	Bending	θ_3	0.945	0.708
$C_7C_9C_{11}$	Bending	θ_4	0.945	0.716
$C_9C_{11}C_{13}$	Bending	θ_5	0.945	0.706
$C_{11}C_{13}C_1$	Bending	θ_6	0.945	0.717
$C_2C_7H_8C_9$	Wag	γ_1	0.978	0.233
$C_9C_{11}H_{12}C_{13}$	Wag	γ_2	0.978	0.190
$C_{11}C_{13}H_{14}C_1$	Wag	γ_3	0.978	0.302
$C_1C_2F_6C_7$	Wag	γ_4	0.978	0.288
$C_7C_9Cl_{10}C_{11}$	Wag	γ_5	0.978	0.355
$C_2C_1B_3C_{13}$	Wag	γ_6	0.978	0.265
$C_2C_1B_3O_4$	Wag	η_3	0.978	0.705
$C_1B_3O_4H_{15}$	Torsion	η_1	0.978	0.139
$C_1B_3O_5H_{16}$	Torsion	η_2	0.978	0.120
$C_{13}C_1B_3O_5$	Torsion	η_4	0.978	0.424
$B_3C_1C_2C_7$	Puckering	τ_1	0.978	0.070
$C_1C_2C_7C_9$	Puckering	τ_2	0.978	0.165
$C_2C_7C_9C_{11}$	Puckering	τ_3	0.978	0.099
$C_7C_9C_{11}C_{13}$	Puckering	τ_4	0.978	0.077
$C_9C_{11}C_{13}C_1$	Puckering	τ_5	0.978	0.080
$C_{11}C_{13}C_1C_2$	Puckering	τ_6	0.978	0.104

[a]See **Fig. 2** for atom numbering and internal coordinates definitions.
[b]Implemented scaling factors to reproduce scaled frequencies in **Table 4**.
[c]Diagonal force constant at B3LYP/6-311++G(d,P) level.

3. Results and Discussion

3.1. Structural Parameters and Force Constants

As the crystal structure of CFPB is not available, the optimized structural parameters (SPs) were compared with experimental data for phenylboronic acid [16], 3-fluorophenylboronic acid [38], and 2,4-difluorophenylboronic acid [18]. The carbon-carbon bond distances show good agreement, typically within ±0.026 Å, and most bond angles align within ±0.1-7.2°. A notable exception is the ∠($B_3O_5H_{16}$) angle in 3-fluorophenylboronic acid [38], which deviates by about 13°. The calculated B–O bond distance of 1.366 Å is very close to the observed value of 1.359 Å for pentafluorophenylboronic acid [44], differing by only 0.007 Å. The calculated B–C bond distance of 1.579 Å also agrees well with X-ray data [16,38,18]. Substituting a C–H bond in benzene derivatives with a C–X (X = F, Cl, Br...) bond leads to a predictable increase in bond length [45]. For similar boronic acids [46,23,47], the observed C–Cl bond length ranges from 1.73 to 1.75 Å, which is consistent with our calculated value of 1.777 Å. Excellent correlations (R^2 values of 0.992, 0.985, and 0.984) were found between the experimental bond lengths [18,38,16] and those predicted by B3LYP/6-311++G(d,p). The CCC bond angles in the benzene ring range from 115.1° to 124.8° in our study, compared to 117.2°–121.8° in phenylboronic acid [16], 117.6°–123.9° in 3-fluorophenylboronic acid [38], and 114.6°–125.1° in 2,4-difluorophenylboronic acid [18]. The R^2 values for bond angles were 0.812 [18], 0.702 [38], and 0.683 [16]. Discrepancies in bond angles between experimental and theoretical values [16,18] are expected, as our calculations model an isolated molecule more representative of the gas phase, whereas X-ray data reflect the crystalline solid state. Overall, the predicted SPs show good agreement with X-ray data, particularly for bond distances. The calculated intramolecular hydrogen bond (HB) distances, O...H, range from 2.453 to 2.460 Å. These are shorter than the sum of the van der Waals radii for oxygen and hydrogen (2.72 Å) by 0.260–0.267 Å [48,49], indicating significant interaction. Similarly, the calculated Cl….H distances range from 2.194 to 2.200 Å, which are considerably shorter than the sum of the van der Waals radii for fluorine and hydrogen (3.0 Å) by 0.800–0.806 Å [50]. This suggests strong intramolecular HB interactions for O_5...H_{14} and Cl_6....H_{15}, in addition to the expected intermolecular HBs in the solid phase **(see Table 1)**.

3.2. Vibrational Assignments

The vibrational frequencies calculated using DFT methods show improved agreement with experimental values **(Table 4)**, suggesting that these methods partially account for anharmonic effects. Therefore, the vibrational assignments proposed here are primarily based on the B3LYP/6-311++G(d,p) frequencies compared to the observed wavenumbers [39]. The differences between the IR spectra obtained in KBr disk and Nujol mull were negligible (±2 cm^{-1}).

3.2.1. OH and CH Fundamental Vibrations

The assignments for the OH/OD and CH/CD stretching modes in C_2FPB **(Table 4)** are consistent with observed bands in the 3600–3000 cm^{-1} region. The O–H stretches were observed at 3640(s) and 3390(s, br), matching the calculated values at 3716 and 3667 cm^{-1}, respectively. These values align well with those reported for other boronic acid derivatives [1,16,21,23,24,26,51]. The O–H in-plane bending vibrations (δ_{ip}; ν_{17} and ν_{19}) were computed at 1040 and 973 cm^{-1}, corresponding well to the observed bands at 1044(s) and 884(s) cm^{-1}. Two OH out-of-plane bending modes are expected. The first, observed at 590 cm^{-1}, matches the calculated frequency at 585 cm^{-1} (ν_{28}). The second (δ_{oop}O–H), calculated at 511 cm^{-1} (ν_{31}), appears to be obscured in the low-wavenumber region below 541 cm^{-1}. These findings are consistent with both experimental data and literature reports [16,21,23,24,26,51].

The aromatic C–H stretching vibrations, typically observed between 3100 and 3000 cm^{-1}, are largely independent of the nature of the substituents [52-55]. Accordingly, the observed IR bands at 3095, 3085, and 3071 cm^{-1} are assigned to the C–H stretches (ν_3–ν_5), matching the calculated values at 3102, 3093, and 3076 cm^{-1}. Aromatic C–H in-plane deformation modes (δ_{ip}) for benzene derivatives generally fall in the 1300–1000 cm^{-1} range [56,57], while the out-of-plane deformations (δ_{oop}) occur between 1000 and 600 cm^{-1} [58]. The frequencies of these deformations are influenced more by the substituent's position than its identity [57,59]. In this study, the δ_{ip}CH modes (ν_8, ν_{13}, ν_{15}) were calculated at 1468, 1245, and 1109 cm^{-1}, aligning with the observed bands at 1487, 1198, and 1100 cm^{-1}. In agreement with Ref. [60], the observed IR bands at 866 and 824 cm^{-1} were assigned to δ_{oop} CH modes, corresponding to the predicted frequencies at 862 (ν_{22}) and 836 (ν_{23}) cm^{-1}. Notably, the δ_{ip} OH (ν_{19}) and a δ_{oop}CH (ν_{20}) mode, both calculated at 973 cm^{-1}, are likely overlapped and both contribute to the very strong IR band observed at 884 cm^{-1}.

TABLE.4. Observed and calculated frequencies (cm^{-1}) for 2-Chloro-4-florophenylboronic acid (CFPB) Structure CT.

Species	ν_i*	Assignment[a]	B3LYP/6-311++G(d,p)						
			Unscaled	Fixed[b] scaled	(X–D)[c]	IR Int[d].	Raman Act[e].	IR KBr[g]	PED[f]
	ν_1	υ_s OH / OD	3853	3716	2808	107.4	153.2	3640s	$51S_1\ 49S_2$
	ν_2	υ_s OH / OD	3803	3667	2790	206.7	50.3	3390s,br	$51S_2\ 49S_1$
	ν_3	υ_{as} CH / CD	3217	3102	2368	1.2	88.8	3095sh	$65S_3\ 35S_4$
	ν_4	υ_s CH / CD	3207	3093	2375	1.8	160.9	3085sh	$57S_4\ 29S_3\ 14S_5$
	ν_5	υ_{as} CH / CD	3189	3076	2340	0.22	47.1	3071wm	$86S_5\ 10S_4$
	ν_6	υ_s CC Ring	1633	1578		191.1	34.6	1608s	$34S_6\ 16S_7\ 16S_{16}\ 10S_{27}$
	ν_7	υ_s CC Ring	1606	1552		71.5	5.4	1564s	$50S_7\ 18S_6\ 10S_{29}$
	ν_8	δ_{ip} CH/ CD	1515	1468	914	35.1	2.8	1487m	$12S_8\ 22S_{13}\ 14S_6\ 13S_9\ 12S_{16}$
	ν_9	υ_s CC Ring	1426	1380		220.7	2.5	1430s	$14S_9\ 35S_{10}\ 10S_6 10S_{15}\ 10S_{16}$
A′	ν_{10}	υ_s B−O Stretch	1389	1344		209.2	5.9	1388s	$33S_{10}\ 12S_{12}\ 10S_9\ 10S_{14}\ 10S_{15}$
	ν_{11}	υ_s B−O Stretch	1355	1310		400.6	21.1	1335vs	$24S_{11}\ 20S_{16}\ 17S_{19}\ 10S_{10}\ 10S_{15}$
	ν_{12}	υ_s CC Ring	1296	1255		35.1	6.7	1251m	$50S_{12}\ 17S_{13}\ 10S_8$
	ν_{13}	δ_{ip} CH/ CD	1288	1245	850	31.0	3.8	1198m	$23S_{13}\ 30S_{12}\ 10S_7\ 10S_9\ 10S_{14}$
	ν_{14}	υ_s C−F Stretch	1222	1182		92.3	11.7	1151s	$35S_{14}\ 22S_8\ 20S_{20}\ 10S_{22}$
	ν_{15}	δ_{ip} CH/ CD	1143	1109	820	2.6	3.0	1100w	$41S_{15}\ 15S_{18}\ 13S_9\ 12S_{13}$
	ν_{16}	υ_s C−B Stretch	1133	1098		65.9	13.8	1079m	$38S_{16}\ 20S_{19}\ 18S_{20}\ 10S_9 10S_{22}$
	ν_{17}	δ_{ip} OH/ OD	1072	1040		24.9	2.0	1044s	$47S_{17}\ 10S_{18}\ 10S_{19}$
	ν_{18}	υ_s CC Ring	1041	1010	841	18.4	5.6	995s	$45S_{18}\ 22S_{17}\ 21S_{22}$
	ν_{19}	δ_{ip} OH/ OD	1004	973	800	150.8	1.7	(884vs)	$36S_{19}\ 38S_{11} 18S_{17}$
A″	ν_{20}	γ_{wag} CH / CD	983	973	810	0.22	0.04	(884vs)	$55S_{20}\ 18S_{23}\ 11S_{27}\ 10S_{40}$
A′	ν_{21}	Ring breath	892	863		93.8	13.8	(866s)	$22S_{21}\ 22S_{14}\ 20S_{25} 14S_{30}$
	ν_{22}	γ_{wag} CH / CD	872	862	757	37.7	0.13	(866s)	$65S_{22}\ 10S_{23}\ 10S_{27}$
A″	ν_{23}	γ_{wag} CH / CD	843	836	648	7.6	0.05	824m	$48S_{23}\ 26S_{20}\ 12S_{22}$
	ν_{24}	γ_{wag} CB	735	725		18.2	0.45	(729m)	$66S_{24}\ 11S_{29} 10S_{27}$
	ν_{25}	υ_s C−Cl Stretch	694	672		4.1	6.8	(729m)	$33S_{25}\ 55S_{34}$
A′	ν_{26}	δ_{ip} CCC Ring	663	642		17.5	5.8	638sh	$14S_{26}$ $27S_{16} 13S_{30} 11S_{11} 10S_{14} 10S_{32}$
A″	ν_{27}	δ Ring puckering	650	641		60.8	0.20	624m	$25S_{27}\ 45S_{24}\ 21S_{33}$
A″	ν_{28}	τ OH/ OD	592	585	442	1.4	0.24	590w	$63S_{28} 16S_{33} 15S_{37}$
	ν_{29}	γ_{wag} C−F	569	564		41.2	0.21	575m	$21S_{29}\ 29S_{40}\ 24S_{28}\ 19S_{31}$
A′	ν_{30}	δ_{ip} CCC Ring	517	502		11.6	1.5	541br,m	$23S_{30}\ 17S_{32}\ 16S_{34}\ 13S_{38} 10S_{39}$
A″	ν_{31}	τ OH/ OD	518	511	380	118.2	1.9	—	$72S_{31}\ 10S_{28}\ 10S_{29}$
A′	ν_{32}	δ_{ip} C−F	504	489		38.6	3.0	(450w)	$18S_{32}\ 34S_{36}\ 15S_{30}\ 10S_{38}\ 10S_{39}$
A″	ν_{33}	δ Ring puckering	452	447		18.3	0.54	(450w)	$26S_{33}\ 16S_{41}\ 13S_{40} 10S_{23} 10S_{29}$
	ν_{34}	δ_{ip} CCC Ring	435	422		6.0	8.3	(450w)	$28S_{34}\ 28S_{25}\ 18S_{32}\ 10S_{21}$
A′	ν_{35}	δ_{Rock} BO_2	344	334		5.1	2.2	—	$35S_{35}\ 22S_{32} 18S_{25}$
	ν_{36}	$\delta_{Scissor}$ $B(OH)_2$	322	312		3.9	3.3	—	$23S_{36}\ 29S_{30}\ 23S_{16}\ 10S_{35}$
A″	ν_{37}	Ring Twist	284	281		1.8	2.0	—	$59S_{37}\ 22S_{27}$

	ν_{38}	δ_{ip} C–Cl	247	240	0.78	1.3	—	$70S_{38}\ 10S_{32}$
A′	ν_{39}	$CCB(OH)_2$ bending	187	182	3.7	0.33	—	$69S_{39}\ 22S_{35}$
	ν_{40}	γ_{wag} CCl	180	176	0.01	0.77	—	$25S_{40}\ 20S_{41}\ 16S_{22}11S_{27}10S_{33}$
	ν_{41}	δ Ring puckering	99	99	0.93	1.0	—	$35S_{41}\ 33S_{33}\ 15S_{42}$
	ν_{42}	τ CCBO	29	28	1.0	0.05	—	$85S_{42}\ 10S_{33}$

a Notations for fundamentals: m, stretch; d, bending; c, wagging; R, ring and s, torsion. These fundamentals could be pure in minor cases; however they are coupled with other normal modes as given in PEDs.
b Scaling factors 0.93 for all stretches. Moreover 0.945 used for bends and 0.978 for wags and torsions.
c The calculated unscaled frequencies for d_1-4-2-CFPB (X-D = C-D or O-D) isotopmers.
d Calculated infrared intensities in km/mol (IUPAC, Glossary of terms used in Photochemistry, 3rd Edition, July 2004).
e Calculated Raman activities in Å4/amu.
f Contributions less than 10% are omitted.
g Bands between brackets are used for two or three normal modes of vibration.

3.2.2. $B(OH)_2$ Fundamental Vibrations

The B–O stretching frequency typically gives rise to a strong band in the 1300–1400 cm^{-1} region in FT-IR spectra [1,21,23,24,26,51]. Vargas et al. assigned the ν(B–O) stretch in trinuclear boron complexes around 1370 cm^{-1} [61]. In our study, the B–O stretching modes are calculated at 1344 (ν_{10}) and 1310 (ν_{11}) cm^{-1}, fitting the observed IR bands at 1388 and 1335 cm^{-1}, respectively. Fundamentals below 400 cm^{-1} namely the $B(OH)_2$ scissoring mode (ν_{35}), the BO_2 rocking (ν_{35}), CCB bending (ν_{40}), and torsion (ν_{42}) modes fall outside our experimental range and could not be observed.

3.2.3. C–X (X = B, F, Cl) Fundamental Vibrations

Three C–X (X = B, F, Cl) stretches are expected in the solid-state IR spectrum of C_2FPB. The C–F and C–Cl stretching frequencies produce strong and medium IR bands, respectively. Their predicted infrared intensities (92.3 and 4.1 km/mol) support their assignment to the observed bands at 1151 cm^{-1} and 729 cm^{-1}, consistent with literature [23,24,26]. The C–B stretch (ν_{16}) is observed at 1098 cm^{-1} and calculated at 1079 cm^{-1}, agreeing well with a reported value of 1085 cm^{-1} [62], though this mode is sometimes overlooked [1]. Furthermore, the observed IR bands at 729 and 575 cm^{-1} are assigned to the C–B (ν_{24}) and C–F (ν_{29}) wagging modes, respectively. The C–Cl in-plane (ν_{38}) and out-of-plane (ν_{40}) bends were predicted at 240 and 176 cm^{-1} but could not be observed experimentally, as mentioned.

3.2.4. Ring Fundamental Vibrations

Ring stretching vibrations were predicted in the 1600–865 cm^{-1} range, consistent with similar molecular systems [1,16,21,23,24,26,58,63]. The computed wavenumbers show slight deviations from the observed values but are supported by their predicted IR intensities. The observed bands at 1609 (ν_6), 1564 (ν_7), and 1430 (ν_9) cm^{-1} correspond to calculated intensities of 191.1, 71.5, and 220.7 km/mol, respectively. The medium and strong IR bands at 1251 (ν_{12}) and 995 (ν_{18}) cm^{-1} match the calculated values at 1255 and 1010 cm^{-1}. The ring breathing mode (ν_{21}) is assigned to the strong band at 866 cm^{-1}, consistent with its high predicted IR intensity of 93.8 km/mol. The ring deformation modes (ν_{26} and ν_{30}), calculated at 642 and 502 cm^{-1}, fit the observed bands at 638 and 541 cm^{-1}, in agreement with experimental values [64,65]. Other fundamentals (ν_{37}, ν_{41}) lie below 400 cm^{-1} and were not observed.

3.3. OH Barriers to Internal Rotation

The 3D potential energy surface for the rotation of both the hydroxyl groups in the CT conformer was explored. The O_5H_{16} (*trans*) and O_4H_{15} (*cis*) group was rotated in 10° increments around the C–B bond together at same time to determine any other minimum structure **(Fig. 4)**. Starting from the TC (E = -967.278089 Hartree), the energy increased to a maximum at a dihedral angle $\tau(H_{15}O_4B_3C_1)$ of -100° and $\tau(H_{16}O_4B_3C_1)$ of 80° (E = -967.249231 Hartree), corresponding to a nearly perpendicular OH orientation relative to the benzene ring. Further rotation led to a decrease in energy toward the CT conformer minimum. The barrier between this maximum and the CT isomer is 22.2 kcal/mol (7749 cm^{-1}). The barrier from TT to CT was found to be 6.7 kcal/mol (2356 cm^{-1}).

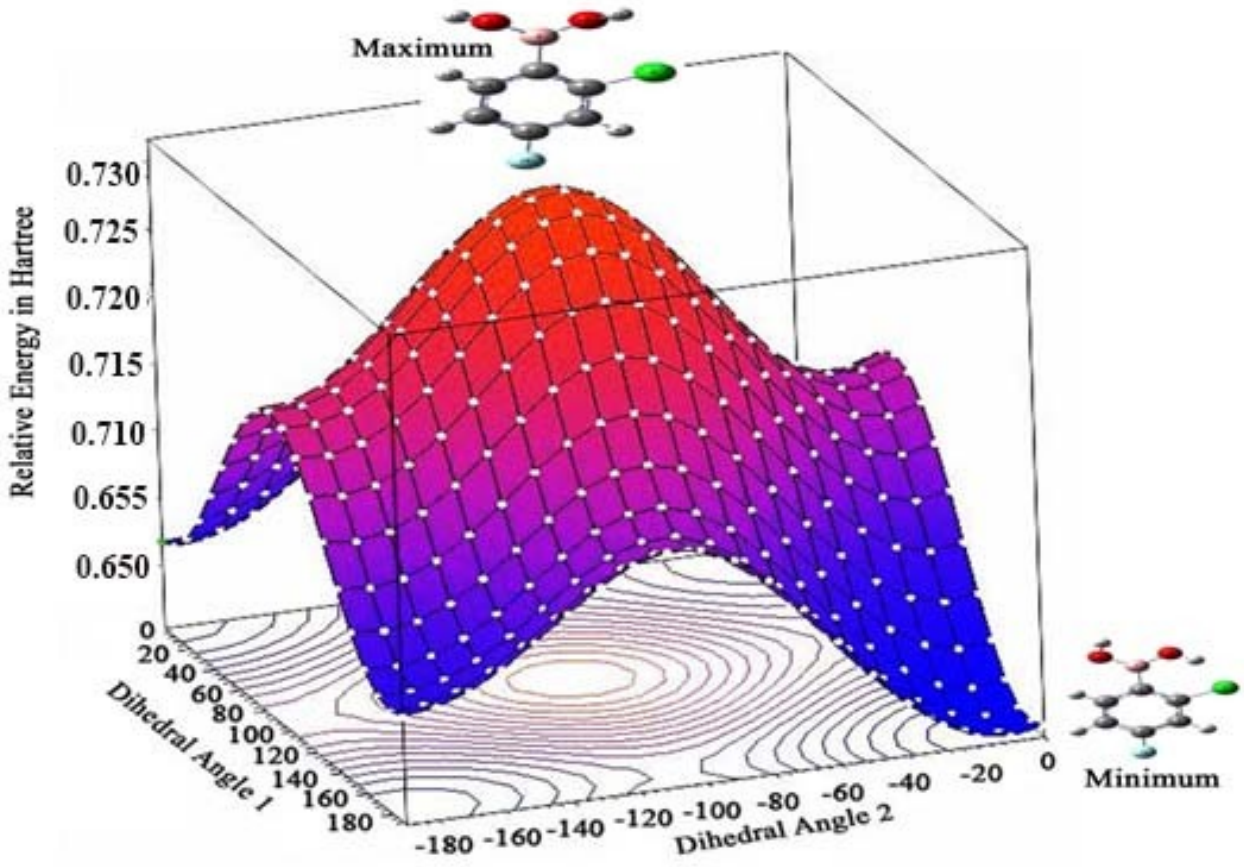


Fig. 4. 3D Barriers to internal rotations of two OH groups of 2-Chloro-4-fluorophenyl boronic acid as obtained from B3LYP /6-311++G(d,p) potential surface scan.

The barrier between CC to CT barrier is 4.9 kcal/mol (1722 cm^{-1}). These values are consistent with V_2 barriers reported for 2-aminophenol and p-fluorophenol (1213 and 1007 cm^{-1}, respectively) [63,66,67]. These significant rotational barriers, along with the calculated imaginary frequencies for other conformers, firmly justify the exclusion of high-energy isomers and confirm CT as the most stable structure.

4. Conclusion

B3LYP/6-311++G(d,p) computations establish that the **CT** conformer of **(2-chloro-4-fluorophenyl)boronic acid** is the most stable structure in the solid phase. In this conformation, one OH group is *cis* to the phenyl C_1 and Cl atoms, while the other is *trans* to the phenyl C_1 atom. Supported by frequency calculations, normal coordinate analysis, and experimental IR spectroscopy, comprehensive and accurate vibrational assignments have been proposed for all observed fundamental modes.

Funding

The authors declare that this research received no external funding.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Mariana Pereira Silva, Lucília Saraiva, Madalena Pinto and Maria Emília Sousa, Boronic Acids and Their Derivatives in Medicinal Chemistry: Synthesis and Biological Applications. *Molecules* **2020**, *25*(18), 4323; https://doi.org/10.3390/molecules25184323

[2] Deniz A. Anakök, Andrea Angeli, Ilaria D'Agostino, Gioele Renzi, Maria Luisa Massardi, Camilla Tavani, Servet Çete, Simone Carradori, Roberto Ronca, Clemente Capasso, Fabrizio Carta, Claudiu T. Supuran, A Journey Around Boronic Acids: Sulfonyl Hydrazone-Containing Derivatives as Carbonic Anhydrase Inhibitors. *Chemical Biology & Drug Design* **09 April 2025,** 105(4), e70108; *https://doi.org/10.1111/cbdd.70108*

[3] Hall, D. G. *Boronic Acids: Preparation and Applications in Organic Synthesis and Medicine*; Wiley-VCH: Weinheim, 2005.

[4] Suzuki, A. Cross-Coupling Reactions Of Organoboranes: An Easy Way To Create C–C Bonds. *Angew. Chem. Int. Ed.* **2011**, *50* (30), 6722–6737.

[5] Yang, W.; Gao, X.; Wang, B. Boronic Acid Compounds as Potential Pharmaceutical Agents. *Med. Res. Rev.* **2003**, *23* (3), 346–368.

[6] Bull, S. D.; Davidson, M. G.; Van den Elsen, J. M. H.; Fossey, J. S.; Jenkins, A. T. A.; Jiang, Y.-B.; Kubo, Y.; Marken, F.; Sakurai, K.; Zhao, J.; James, T. D. Boronic Acid-Based Sensor. *Acc. Chem. Res.* **2019**, *52* (10), 2879–2890.

[7] Wu, X.; Li, Z.; Chen, X.-X.; Fossey, J. S.; James, T. D.; Jiang, Y.-B. Selective Sensing of Saccharides Using Simple Boronic Acids and Their Aggregates. *Chem. Soc. Rev.* **2013**, *42* (20), 8032–8048.

[8] Rettig, S. J.; Trotter, J. Refinement of the Crystal Structure of Phenylboronic Acid. *Can. J. Chem.* **1977**, *55* (17), 3071–3075.

[9] Kurt, M. Theoretical and Experimental Study on the Molecular Structure and Vibration Spectra of 4-Chlorophenylboronic Acid. *J. Mol. Struct.* **2009**, *919* (1–3), 215–223.

[10] Karabacak, M.; Kose, E.; Atac, A.; Sas, E. B.; Asiri, A. M.; Kurt, M. Experimental (FT-IR, FT-Raman, UV and NMR) and Theoretical (EHMO and DFT) Analysis of 3-Bromophenylboronic Acid. *Spectrochim. Acta A* **2012**, *97*, 892–904.

[11] Rodríguez-Cuamatzi, P.; Vargas-Díaz, G.; Hopfl, H. 2,4-Difluorophenylboronic Acid. *Acta Crystallogr. E* **2005**, *61* (6), o44–o46.

[12] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; et al. *Gaussian 09, Revision D.01*; Gaussian, Inc.: Wallingford, CT, 2009.

[13] Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38* (6), 3098–3100.

[14] Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37* (2), 785–789.

[15] Vosko, S. H.; Wilk, L.; Nusair, M. Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: A Critical Analysis. *Can. J. Phys.* **1980**, *58* (8), 1200–1211.

[16] Wilson, E. B.; Decius, J. C.; Cross, P. C. *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*; McGraw-Hill: New York, 1955.

[17] NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101, Release 22, May 2024.

[18] Sass, E.; Kose, E.; Kurt, M.; Karabacak, M. FT-IR, FT-Raman, Molecular Structure and Natural Bond Orbital Analysis of 2-Fluorophenylboronic Acid. *Spectrochim. Acta A* **2015**, *137*, 1315–1325.

[19] Yang, W.; He, H.; Drueckhammer, D. G. Computer-Guided Design in Molecular Recognition: Design and Synthesis of a Glucopyranose Receptor. *Angew. Chem. Int. Ed.* **2001**, *40* (9), 1714–1718.

[20] Larsen, R. D.; King, A. O.; Chen, C. Y.; Corley, E. G.; Foster, B. S.; Roberts, F. E.; Yang, C.; Lieberman, D. R.; Reamer, R. A. Efficient Synthesis of Losartan, A Nonpeptide Angiotensin II Receptor Antagonist. *J. Org. Chem.* **1994**, *59* (21), 6391–6394.

[21] Tjarks, W.; Anisuzzaman, A. K. M.; Liu, L.; Soloway, A. H.; Barth, R. F.; Perkins, D. J.; Adams, D. M. Synthesis and in Vivo Evaluation of Bis-Mercaptoundecahydro-Closo-Dodecaborate Derivatives. *J. Med. Chem.* **1992**, *35* (9), 1628–1633.

[22] Yamamoto, Y.; Asao, N. Selective Reactions Using Allylic Metals. *Chem. Rev.* **1993**, *93* (6), 2207–2293.

[23] Alam, F.; Soloway, A. H.; Barth, R. F.; Mafune, N.; Adams, D. M.; Knoth, W. H. Boron Neutron Capture Therapy: Linkage of a Boronated Macrocyclic Molecule to Monoclonal Antibodies. *J. Med. Chem.* **1989**, *32* (10), 2326–2330.

[24] O'Donovan, M. R.; Mee, C. D.; Fenner, S.; Teasdale, A.; Phillips, D. H. Boronic Acids-A Novel Class of Bacterial Mutagen. *Mutat. Res.* **2011**, *724* (1–2), 1–6.

[25] Cambre, J. N.; Sumerlin, B. S. Biomedical Applications of Boronic Acid Polymers. *Polymer* **2011**, *52* (21), 4631–4643.

[26] Westmark, P. R.; Smith, B. D. Boronic Acid Sensors for Saccharides. *J. Pharm. Sci.* **1996**, *85* (3), 266–269.

[27] Cyranski, M. K.; Jezierska, A.; Klimentowska, P.; Panek, J. J.; Sporzynski, A. On the Aromatic Character of the Benzene Ring in Substituted Benzenes. *J. Phys. Org. Chem.* **2008**, *21* (6), 472–482.

[28] Faniran, J. A.; Shurvell, H. F. Infrared Spectra of Phenylboronic Acid and Diphenyl Phenylboronate. *Can. J. Chem.* **1968**, *46* (12), 2089–2095.

[29] Kurt, M. FT-Raman and FT-IR Spectra and DFT Calculations of 4-Chlorophenylboronic Acid. *J. Raman Spectrosc.* **2009**, *40* (1), 67–75.

[30] Karabacak, M.; Cinar, M.; Kurt, M. Molecular Structure and Vibrational Spectra of 3,5-Dichlorophenylboronic Acid. *J. Mol. Struct.* **2008**, *885* (1–3), 28–35.

[31] Rai, A. K.; Kumar, S.; Rai, A. K. Vibrational Spectra of 2,6-Dimethoxyphenylboronic Acid. *Vib. Spectrosc.* **2007**, *42* (2), 397–403.

[32] Rettig, S. J.; Trotter, J. Crystal Structure of Phenylboronic Acid. *Can. J. Chem.* **1977**, *55* (17), 3071–3075.

[33] Cyranski, M. K.; Jezierska, A.; Klimentowska, P.; Panek, J. J.; Sporzynski, A. Structural and Energetic Analysis of Molecular Crystals of Phenylboronic Acid. *J. Phys. Org. Chem.* **2008**, *21* (6), 472–482.

[34] Rodríguez-Cuamatzi, P.; Vargas-Díaz, G.; Hopfl, H. Crystal Structure of 2,4-Difluorophenylboronic Acid. *Acta Crystallogr. E* **2005**, *61* (6), o44–o46.

[35] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; et al. *Gaussian 09, Revision D.01*; Gaussian, Inc.: Wallingford, CT, 2009.

[36] Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648–5652.

[37] Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37* (2), 785–789.

[38] Vosko, S. H.; Wilk, L.; Nusair, M. Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: A Critical Analysis. *Can. J. Phys.* **1980**, *58* (8), 1200–1211.

[39] Pulay, P. Ab Initio Calculation of Force Constants and Equilibrium Geometries in Polyatomic Molecules. *Mol. Phys.* **1969**, *17* (2), 197–204.

[40] Schlegel, H. B. Optimization of Equilibrium Geometries and Transition Structures. *J. Comput. Chem.* **1982**, *3* (2), 214–218.

[41] Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986.

[42] Peng, C.; Ayala, P. Y.; Schlegel, H. B.; Frisch, M. J. Using Redundant Internal Coordinates to Optimize Equilibrium Geometries and Transition States. *J. Comput. Chem.* **1996**, *17* (1), 49–56.

[43] Wilson, E. B.; Decius, J. C.; Cross, P. C. *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*; McGraw-Hill: New York, 1955.

[44] Schachtschneider, J. H. *Vibrational Analysis of Polyatomic Molecules*; Shell Development Co.: Houston, TX, 1964–1965.

[45] NIST Computational Chemistry Comparison and Benchmark Database. http://cccbdb.nist.gov/vibscale.asp (accessed May 2024).

[46] Wu, Y.; Dong, C.; Liu, S.; Zhu, H.; Wu, M. Y. Crystal Structure of 3-Fluorophenylboronic Acid. *Acta Cryst. E* **2006**, *62* (6), o4236–o4237.

[47] Horton, P. N.; Hursthouse, M. B.; Becket, M. A.; Hankey, M. A. Crystal Structure of Pentafluorophenylboronic Acid. *Acta Cryst. E* **2004**, *60* (12), o2204–o2205.

[48] Shimpl, M.; Seetha, N. L.; Pedireddi, V. R. Crystal Structures of 4-Chloro- and 4-Bromophenylboronic Acids. *Cryst. Growth Des.* **2007**, *7* (10), 1958–1963.

[49] David, J.; Hallam, H. E. Infrared Spectra of Boronic Acids and Their Anhydrides. *Spectrochim. Acta* **1965**, *21* (5), 841–848.

[50] Krueger, P. J. Vibrational Spectra of Boron Compounds. *Tetrahedron* **1970**, *26* (22), 4753–4766.

[51] Karabacak, M.; Kurt, M. FT-IR, FT-Raman and DFT Studies of 2,3-Difluorophenylboronic Acid. *J. Mol. Struct.* **2009**, *919* (1–3), 215–223.

[52] Socrates, G. *Infrared and Raman Characteristic Group Frequencies: Tables and Charts*, 3rd ed.; John Wiley & Sons: Chichester, 2001.

[53] Dwivedi, C. P.; Sharma, S. N. Vibrational Spectra of Substituted Benzenes. *Indian J. Pure Appl. Phys.* **1973**, *11* (6), 447–452.

[54] Varsanyi, G. *Assignments for Vibrational Spectra of Seven Hundred Benzene Derivatives*; Adam Hilger: London, 1974.

[55] Singh, N. P.; Yadav, R. A. Vibrational Spectra of Chloro- and Fluorobenzenes. *Indian J. Phys. B* **2001**, *75* (4), 347–352.

[56] Doddamani, S.; Ramoji, A.; Yenagi, J.; Tonannavar, J. Vibrational Spectra of Disubstituted Benzenes. *Spectrochim. Acta A* **2007**, *67* (1), 150–156.

[57] Jag, M. *Organic Spectroscopy - Principles and Applications*; Narosa Publishing House: New Delhi, 2001.

[58] Balachandran, V.; Parimala, K. Vibrational Spectroscopic Studies of 2,4-Dichlorophenylboronic Acid. *J. Mol. Struct.* **2012**, *1007*, 136–146.

[59] Kalsi, P. S. *Spectroscopy of Organic Compounds*, 6th ed.; New Age International: New Delhi, 2004.

[60] Jag, M. *Organic Spectroscopy: Principles and Applications*, 2nd ed.; Narosa Publishing House: New Delhi, 2010.

[61] Vargas, G.; Hernandez, I.; Hopfl, H.; Ochoa, M. E.; Castillo, D.; Farfan, N.; Santillan, R.; Gomez, E. Trinuclear Boron Complexes. *Inorg. Chem.* **2004**, *43* (3), 8490–8497.

[62] Romanos, J.; Beckner, M.; Stalla, D.; Tekcei, A.; Suppes, G.; Jalisatgi, S.; Lee, M. W.; Hawthorne, M. F.; Robertson, J. D.; Firlej, L.; et al. Infrared Spectroscopy of Boron Compounds. *Carbon* **2013**, *54*, 208–212.

[63] Soliman, U. A.; Hassan, A. M.; Mohamed, T. A. Vibrational Analysis of 2-Aminophenol. *Spectrochim. Acta A* **2007**, *68* (3), 688–696.

[64] Yamaguchi, Y.; Frisch, M.; Gaw, J.; Schaefer, H. F.; Binkley, J. S. Analytical Second Derivatives for Molecular Orbital Theory. *J. Chem. Phys.* **1986**, *84* (4), 2262–2268.

[65] Sathyanarayanan, D. *Vibrational Spectroscopy: Theory and Applications*; New Age International: New Delhi, 2004.

[66] Larsen, N. W.; Nicolaisen, F. M. Vibrational Spectra and Conformational Analysis of p-Fluorophenol. *J. Mol. Struct.* **1979**, *22* (1), 29–36.

[67] Martinez de la Hoz, J. M.; Balbuena, P. B. The DFT Route to Geometry Optimization and Vibrational Spectrum Calculation. In *Density Functional Theory*; IntechOpen: London, 2013; pp 277–301.

Islamic University Journal of Applied Sciences (IUJAS)
https://journals.iu.edu.sa/jesc

Volume VIII, Issue I, July 2026, Pages 189-217

Explainable Deep Learning Models for Citrus Leaf Disease Identification and Agricultural Decision Support

Sami Saad Albouq [1] *

[1] Department of Computer and Information Systems, Islamic University of Madinah, 42351, Medina, Saudi Arabia, salbouq1@iu.edu.sa

*Corresponding author: (Sami Saad Albouq), *Email Address*: salbouq1@iu.edu.sa

Abstract

Citrus is a commercially important crop affected by diseases such as anthracnose, bacterial blight, and citrus canker, which damage trees and reduce fruit yield. Classical diagnosis is time-consuming and expert-dependent, limiting scalability. In this paper, we propose explainable deep learning models for multiclass citrus leaf disease detection, incorporating custom pre-processing and extensive data augmentation for robustness. A custom CNN and transfer learning-based VGG16 were fine-tuned on an augmented dataset of nine citrus leaf categories, including both healthy and unhealthy specimens. Performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrices. The CNN achieved 97% accuracy, outperforming VGG16's 95%, demonstrating the effectiveness of our preprocessing approach. For interpretability, Grad-CAM, SHAP, and LIME were uniformly applied to highlight symptom-relevant regions aligned with agronomic knowledge, enhancing transparency and user trust. The framework effectively combines predictability and interpretability, supporting practical, real-time, and sustainable agricultural decision systems.

Keywords: Citrus Leaf Diseases; Deep Learning; CNN, VGG16; Explainable AI; Grad-CAM; Agricultural Decision Support.

https://doi.org/10.63070/jesc.2026.023
Received 25 December 2025; 15 February 2026; Accepted 10 April 2026.
Available online 15 April 2026.
Published by Islamic University of Madinah on behalf of *Islamic University Journal of Applied Sciences*.
This is an open access article published under the Creative Commons Attribution 4.0 International license (CC BY 4.0):https://creativecommons.org/licenses/by/4.0/

1. Introduction

Citrus belongs to one of the most significant fruit crops, cultivated worldwide with high economic, nutritional, and cultural value. However, its productivity is continually compromised by numerous leaf pathologies, such as anthracnose, bacterial blight, citrus canker, and curvularia virus, among others, which cause drastic losses in productivity and quality [1]. These diseases, in addition to having a significant impact on local food security, also have a substantial impact on international commerce. The existing methods of diagnostic procedures in most instances rely on trained eyesight and lab tests, which are time-consuming, labor-intensive and prone to errors in the agricultural context. Hence, the intelligent prediction model that is scalable, reliable, and achieves early screening and intervention is greatly in need [2].

With the development of deep learning methods, plant disease recognition has also been revolutionised to achieve excellent accuracy in image-based classification. CNNs and transfer learning architectures, such as VGG16, have yielded excellent results on agricultural datasets. Upcoming networks, such as Vision Transformers, continue to challenge the limits of accuracy and the feasibility of heavy computation [3]. Meanwhile, as powerful predictive models, they function as a “black box”, making their decisions completely obscure. This black box challenge results in a trust deficit, as stakeholders in agriculture do not only need accurate predictions but also human-understandable explanations to make informed disease management decisions [4].

Explainable artificial intelligence (XAI) is being employed to address this issue and enhance transparency and trust in deep learning. Gradient-weighted Class Activation Mapping (Grad-CAM), SHapley Additive exPlanations (SHAP), and Local Interpretable Model-agnostic Explanations (LIME) are some of the techniques aimed at providing visual clues to the areas within an image that influence model decisions [5]. Such approaches, when incorporated into citrus leaf disease detection systems, not only reveal the principles underlying the system's decision but also identify patterns of symptoms that are relevant and in agreement with agronomic intuition [6]. This is the connection that is lacking between predictive performance and practical usability, which favours AI-driven systems that can be interpreted, are reliable, and suitable for in-field agricultural decision support. In this research, the following objectives are planned:

- Develop robust CNN and VGG-16-Based models to detect multi-class citrus leaf diseases effectively.
- Implement customized preprocessing and augmentation techniques to enhance the models' generalization capabilities.
- Integrate explainability methods such as Grad-CAM, SHAP, and LIME into the detection pipeline for improved visual interpretation of model predictions.
- Evaluate models based on accuracy, precision, recall, F1-score, and confusion matrices.
- Provide a decision-support framework that translates AI outcomes into easily understandable insights for farmers and policymakers.

The contributions of this study are clarified as follows:

- An explainable citrus leaf disease classification framework is developed by jointly integrating Grad-CAM, SHAP, and LIME with CNN and VGG16 models.
- A lightweight custom CNN (0.33M parameters) is benchmarked against VGG16 under identical training conditions, achieving 97% accuracy with reduced computational cost.
- The study demonstrates that robust preprocessing and multi-XAI validation enable RGB-based models to perform competitively with complex YOLO and hyperspectral baselines.

The rest of this paper follows the following structure. Section 2 conducts a literature review of the gaps in citrus disease detection and interpretability. Section 3 describes the preparation of datasets, preprocessing, and augmentation. The suggested CNN and VGG16 models, which include Grad-CAM, SHAP, and LIME, are depicted. In Section 4, the experimental outcomes are discussed, including performance measures and visual interpretations. Section 5 discusses about the implications of explainable AI on agricultural decision support, and the conclusion is made in Section 6.

2. Related Work

2.1 Plant Disease Detection: Traditional ML vs Deep Learning

Wang et al. [7] introduced a citrus pest and disease detection system with deep learning models based on YOLOv3 and YOLOv4 and trained on field images with high detection rates of Huanglongbing, citrus canker, and healthy leaves. The framework had a great potential of edge based, real-time monitoring of the orchards. Nonetheless, the method is mainly concerned with the performance of detection of objects and lacks insight that can be explained in regard to disease-specific visual characteristics. Contrastingly, the suggested study highlights the image-level classification of disease along with multi-method explainable AI to aid in greater transparency and diagnostic explainability.

Sujatha et al. [8] proposed an automatic leaf plant disease detection system based on combination of deep learning and machine learning. The research utilised CNN models in the feature extraction process, specifically VGG19 and Inception v3, along with classifiers such as SVM and kNN, to enhance the results. Performance on the Banana, Custard Apple, Fig, and Potato leaf datasets varied: the highest accuracy was obtained for the Custard Apple dataset, at 99.1%. This study demonstrates the flexibility of introducing hybrid AI models to address various types of plant disease detection problems.

2.2 Deep Learning Architectures in Agriculture: CNNs and VGG16

CNNs have become the cornerstone of plant disease detection as they capture spatial hierarchies in images. In the meantime, among the CNN-based models, VGG16 is a relatively heavyweight architecture compared to the shallower solutions, which, however, is simple enough to provide strong generalisation in the field of agriculture. Various works support the idea that VGG 16 outperforms shallower CNNs for classification tasks,

albeit at a drastically higher computational load. Generally, multiple researchers use CNN to obtain a low FLOPs baseline, and VGG to implement transfer learning backend. As a result, they are the most popular models in agricultural image analysis.

Dhiman et al. [9] proposed an intelligent citrus detection system that integrates CNN and LSTM models, combined with edge computing. The approach utilises better feature extraction, downsampling, and feature fusion to fit the best models on limited-resource devices. The model was tested on the Kaggle and Plant Village datasets, which included 2,950 citrus fruit images, using pruning and post-quantisation methods. Experimental results demonstrated better accuracy, with 97.18% for pruning and 98.25% for pruning plus post-quantization compared with previous CNN-based techniques.

2.3 Explainable AI (XAI) in Limited Agricultural Use

Even though CNNs and VGG16 are accurate, they are frequently referred to as black boxes, which do not remain very interpretable. To address the above issue, various XAI technologies, including Grad-CAM, LIME, and SHAP, have emerged based on the interpretation of model predictions. These techniques are commonly used in the medical and industrial sectors, but are scarcely utilised in agriculture. Current agricultural research often falls short in terms of accuracy and lacks explainability, which undermines farmer trust. Therefore, the instantiation of XAI in CNN and VGG16 predictions presents a new trend in detecting plant diseases.

Kondaveeti and Simhadri [10] proposed a three-stage approach to assess DLMs in rice leaf disease detection through explainable AI. Conventional metrics were combined with a qualitative and quantitative XAI-based analysis, which included LIME visualisations, as well as IoU, DSC, and a new overfitting ratio. ResNet50 was the topmost performing model with a classification accuracy of (99.13%) and higher reliability for feature selection from eight evaluated models. The novel approach enhances the transparency and credibility of the model, both of which have broader applications beyond agriculture.

2.4 Beyond Conventional CNNs: Detection, Biochemical, and Graph-Based Methods

Xu et al. [11] focused on detecting Huanglongbing (HLB) infection in citrus plants through the application of volatile organic compound (VOC) profiling with GC 13 MS, GC 13 IMS, and OPLS-DA modelling. Their experiment showed extremely high predictability among various citrus species, which showed that VOC-biochemical signatures are highly discriminating in the early detection of HLB. Nonetheless, the method is based on specific laboratory equipment and chemical analysis, which restricts its applicability and scalability in the field conditions in real-time. In addition, the approach lacks image learning and explainability of visual disease appraisal.

Guo et al. [12] suggested YOLOv8n-DE, a lightweight object detection model with dilated reparameterized feature modules and shared parameter detection head to identify citrus disease. The model offered high detection accuracy and highly decreased the number of parameters and computational cost therefore it is applicable in orchard deployment. Regardless of these benefits, the approach puts emphasis on detection that uses bounding-

boxes and fails to offer model interpretability or pixel-level description of disease symptoms, which becomes more and more significant in precision agriculture to motivate trust and decision support.

Aswini and Vijayakumaran [13] proposed a citrus disease classification system based on a graph neural network that combines feature extraction through FA-ResNet, gated fusion adaptive graph learning with a bio-inspired optimization algorithm. Their hybrid design realised significant performance increase over conventional CNN-based classifiers especially in complicated texture and light conditions. Nevertheless, multiple tightly coupled stages, such as segmentation, graph construction, and evolutionary optimization, take part in the model, and it leads to the complexity of the system and low transparency. Also, clear explainable AI (XAI) evaluation to decode specific visual clues on diseases was not mentioned.

2.5 CNN-Based Plant Disease Detection with Explainable AI

Recent studies demonstrated efficiency of convolutional neural networks in classifying plant and leaf diseases with the help of RGB images. Some of these works use such transfer-learning backbones as VGG, ResNet, and MobileNet to enhance classification accuracy in limited data conditions. Nevertheless, a large number of these methods mainly involve predictive performance, and offer little insight on model decisions.

To overcome this shortcoming, a recent literature on agricultural algorithms has embraced the explainable AI methods, which are mostly Grad-CAM, to visualize disease-relevant regions in leaf images. Although these studies can help make qualitative conclusions on CNN attention patterns, the explainability of their results is commonly limited to analysis of one method and a small sample. Contrary to that, the suggested framework combines several complementary XAI methods (Grad-CAM, SHAP, and LIME) and implements them uniformly in both a custom CNN and a transfer-learning model, which allows obtaining a more holistic and trustworthy view of disease-specific characteristics.

Additionally , in contrast to detection- or hyperspectral-based systems, which rely on complicated architectures or specialized sensors, the suggested design can compete with the others with minimal technical overhead of the standard RGB images and transparent priorities on the reliability of the model and support of real-world agricultural decision making.

2.6 Identified Gaps

The literature reveals a few gaps important to this study. First, although CNNs and VGG16 have achieved impressive results, the black-box nature of these models hinders their trustworthiness and applicability in the agricultural sector. Second, the methods of explainability have not been significantly explored in the context of plant diseases, which reduces the relationship between prediction and practical farming activities. Third, there are limited works that combine outputs of AI with real-world applications to assist growers in making decisions. To resolve these problems, CNN and VGG16 need to be employed in conjunction with XAI tools, so that the high accuracy is associated with understandable and farmer-friendly recommendations. Table 1 consolidates recent advances in citrus disease detection, underscoring both high accuracy and the emerging need for explainable, decision-oriented models.

Table 1: Recent Advances in Explainable Citrus Leaf Disease Detection

Authors / Year	Methodology / Models	Explainability / Analysis	Performance (Accuracy / Metrics)	Key Contribution
Zhu et al., 2024 [14]	Multi-Model Fusion Network (MMFN) combining AlexNet, VGG, ResNet, Pre-ResNet34, Pre-ResNet50 with Transfer Learning	Comparative model evaluation, transfer learning analysis	99.72% (healthy vs diseased), 98.68% (multi-class diseases)	Developed MMFN with transfer learning achieving state-of-the-art citrus leaf disease classification
Kaur et al., 2025 [15]	Nonlinear Fuzzy Rank-Based Ensemble (NL-FuRBE) combining VGG19, AlexNet, Xception with image enhancement (VAD, morphological filtering)	Image quality metrics (PSNR, SSIM, NIQE); ensemble explainability via fuzzy ranking	Avg. accuracy 96.51% (5-fold CV)	Introduced NL-FuRBE with enhanced image preprocessing and fuzzy ensemble, improving robustness
Mo & Wei, 2024 [16]	YOLOv8n-CDDA with Cross-Domain Dynamic Attention (CDDA), ARMS augmentation, pruning + knowledge distillation	Grad-CAM++ for feature visualization	Accuracy 90.89%, Recall 81.12%, mAP50 88.36%	Proposed lightweight YOLOv8n-CDDA for real-time citrus disease detection on edge devices
Min et al., 2024 [17]	Hyperspectral Reflectance + Fluorescence Imaging with ML classifiers (SVM, Discriminant Analysis, PCA feature selection)	Hyperspectral band analysis for interpretability	Accuracy up to 94.5% (SVM & DA classifiers)	Combined hyperspectral reflectance & fluorescence imaging for high-precision citrus disease detection

3. Proposed Methodology

The methodology of this study is structured in a way that explainable deep learning models are created and evaluated to classify various lemon leaf diseases precisely. A comprehensive workflow was constructed, encompassing dataset pre-processing and augmentation for robustness, as well as the design of both a specific CNN model and a transfer learning strategy utilising VGG16. The models were trained with optimal hyperparameters, and the evaluation metrics were carefully chosen to measure performance at both global and class levels. In addition, Explainable AI tools such as Grad-CAM and SHAP were incorporated to improve interpretability by generating visual explanations consistent with agronomic knowledge.

Figure 1 provides a structured methodology for the present study. Using nine classes of lemon leaf diseases images, the pre-processing and augmentations techniques like (normalization, rotation, shift shear zoom flip etc) were applied to obtain a balanced dataset. Two model structures were used as follows: one is a customised CNN with convolution, pooling, dense, and dropout layers, the other is transfer learning based on the VGG16 network with a refining dense layer. Training and validation were performed using the Adam optimiser with Categorical cross-entropy loss, over 30 epochs, and a batch size of 30. Assessment was conducted using accuracy, precision, recall, F1-score, confusion matrices, ROC, and AUC. To enhance interpretability,

explainable AI methods such as Grad-CAM, LIME, and SHAP were incorporated to view results classified diseases with visual explanations and decision-support insights for agriculture.

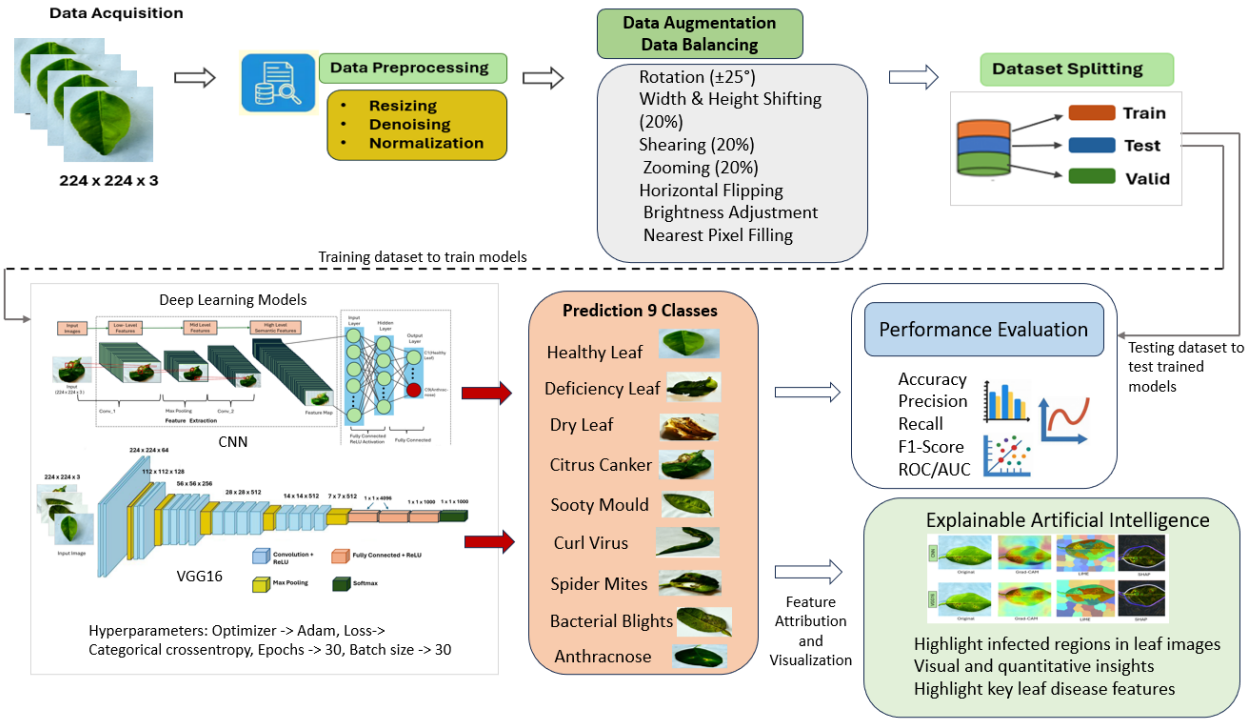


Figure 1: Proposed Methodology for Explainable Citrus Leaf Disease Detection

3.1 Dataset Overview

In this work, we utilise the Comprehensive Lemon Leaf Disease Dataset for advanced detection and sustainable agriculture as the reference citrus dataset benchmark for computer vision-based disease recognition. This dataset comprises 1,354 high-resolution raw images (800×800 pixels) of a lemon leaf, captured in Melandaha, Jamalpur, Bangladesh, from July to September 2024. It includes images of nine clearly defined categories — anthracnose, bacterial blight, citrus canker, curl virus deficiency leaf, dry leaf, healthy leaf, sooty mould, spider mites — annotated by agricultural experts to ensure accuracy and consistency in classification. These original images retain natural differences in lighting, orientation, and condition of the leaves, resulting in a dataset that reflects true field conditions.

Figure 2 presents the example images of nine categories from the lemon leaf dataset, which contain both healthy and diseased leaves. The diseases include common illnesses such as Anthracnose, Bacterial Blight, Citrus Canker, Curl Virus, Deficiency Leaf, Dry Leaf, and others, including Sooty Mould and Spider Mites. These visual examples illustrate the diversity in the dataset, which is crucial for training robust deep learning models.



Figure 2: Representative Samples of Lemon Leaf Disease Dataset

3.2 Data Preprocessing & Augmentation

All images were normalized and rescaled to a consistent range of pixel intensities, allowing for the convergence of models during training. We employed various augmentation strategies, including random rotation, zoom, width and height shifts, shear, and horizontal flip, to enhance the model's robustness. Although the dataset had a fair proportion of classes, augmentation was applied to increase the intra-class variation, enabling the models to generalise more effectively to unseen conditions (i.e., leaf orientation, lighting, and background variation). Not only does such a preprocessing pipeline alleviate the overfitting problem, but it also enhances the model's ability to cope with real-life agricultural situations where highly non-uniform image acquisition conditions are encountered. The dataset was further split into 80% for training and 20 % for testing for model training and evaluation.

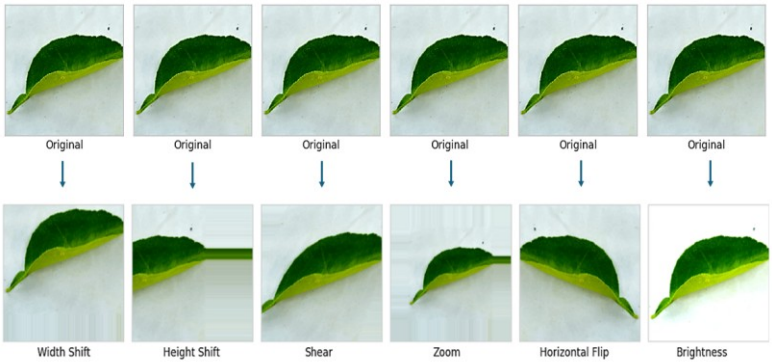


Figure 3: Data Augmentation Techniques for Lemon Leaf Images

Figure 3 illustrates the effectiveness of different types of augmentation employed to diversify dataset and improve model generalization. From the lemon leaf image, several transformations were applied, including width shift, height shift, shear, zoom range (zoom), horizontal flip, and brightness change. Such augmentations mimic possible natural changes in orientation, lighting and scale of the images to ensure that the deep learning models are robust when detecting citrus leaf diseases under different conditions.

Table 2: Image Augmentation Parameters for Lemon Leaf Dataset

Augmentation Technique	Applied Range / Value	Purpose
Rescaling	1./255	Normalize pixel values to [0,1]
Rotation	±25°	Simulate natural leaf orientation changes
Width Shift	±20% (0.2)	Handle horizontal displacement
Height Shift	±20% (0.2)	Handle vertical displacement
Shear Transformation	0.2	Distort images for geometric robustness
Zoom	0.2	Simulate different distances to camera
Horizontal Flip	True	Capture symmetry in leaf orientation
Validation Split	20%	Ensure reliable evaluation
Brightness Adjustment	Range = [0.8, 1.2]	Vary lighting to mimic real-world scenarios

Table 2 summarises the augmentation methods employed to enhance model generalisation and robustness. All transformations, including rotation, shift, shear, zoom, flipping, and brightness, were properly parameterised to model the actual variation observed in leaf images. Such methods helped create a dataset that represents various states, thereby enhancing the trustworthiness of deep learning models in disease recognition.

Augmentation ranges listed in Table 2 fall within commonly used ImageDataGenerator configurations, but were chosen deliberately because of the size, the class imbalance, and the visual diversity of the citrus leaf data instead of being chosen randomly. The dataset also has moderate variations on leaf orientation, scale, illumination and background conditions hence conservative augmentation ranges were emulated to avoid introducing unrealistic variations which may affect disease texture pattern. Early trials with augmentation levels that were stronger led to lower validation performance and the final parameters are a trade-off between the diversity of the data and retention of features. The augmentation strategy was therefore informed by the dataset characteristics and empirical validation and not arbitrary defaults.

3.3 Model Architectures

3.3.1 Convolutional Neural Network (CNN)

For our baseline we used a custom architecture CNN which designed for the task of multi-class lemon leaf disease classification. The architecture consisted of a series of convolutional layers with ReLU activation, followed by max-pooling layers to reduce spatial dimensionality while preserving discriminative information [18]. A flattening layer was inserted between the convolutional backbone and fully connected dense layers, although dropout was employed to prevent overfitting. The model was trained using the Adam optimiser with a learning rate of 1e-4, a batch size of 32, and trained for a maximum of 30 epochs with validation checks. This CNN served as a baseline for the comparison to more sophisticated transfer learning models [19].

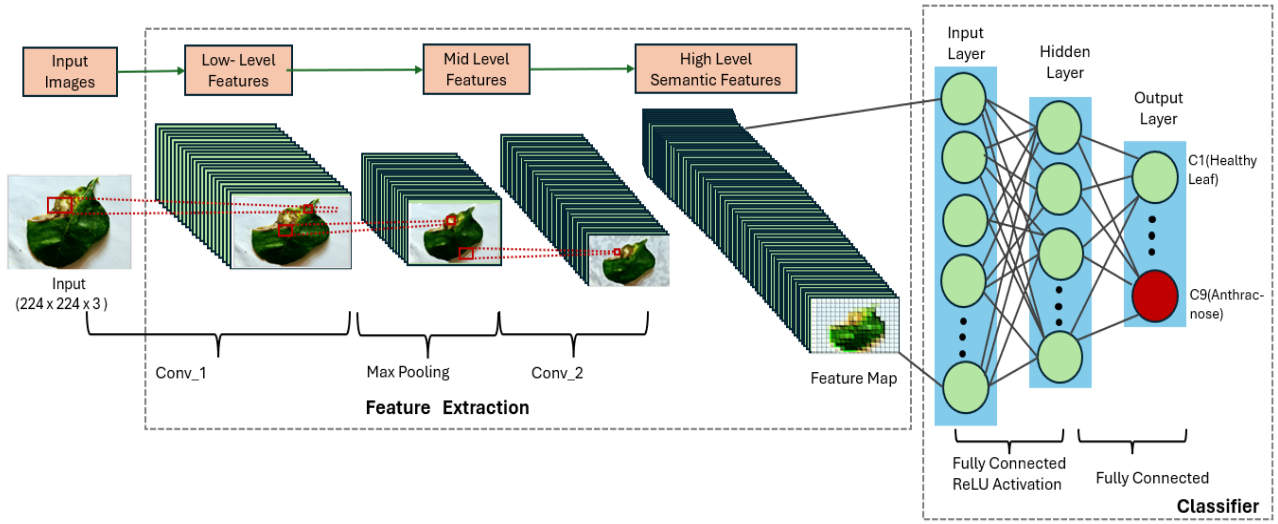


Figure 4: Custom CNN Architecture for Lemon Leaf Disease Detection

The architecture of the specific CNN used in this work is shown in Figure 4. The model leverages hierarchical features from input images through convolutional and pooling operations, capturing different patterns (low-level, mid-level, and high-level). These feature maps are then flattened and fed through two fully connected layers with ReLU activation, which classifies the presence of diseased leaf into nine categories (including both healthy and infected leaves).

3.3.2 Transfer Learning with VGG16

To utilise the potential of pre-trained models, we implemented VGG-16 as our backbone network, initialised with ImageNet weights. The convolutional base was partially frozen to retain low-level, generic features; however, the top of the CNN architecture was fine-tuned for domain-specific classification of citrus leaf diseases. On top of the backbone, a Global Average Pooling layer was added, followed by two fully connected dense layers with ReLU activation and dropout for regularisation, and finally a softmax output layer for the nine disease classes.

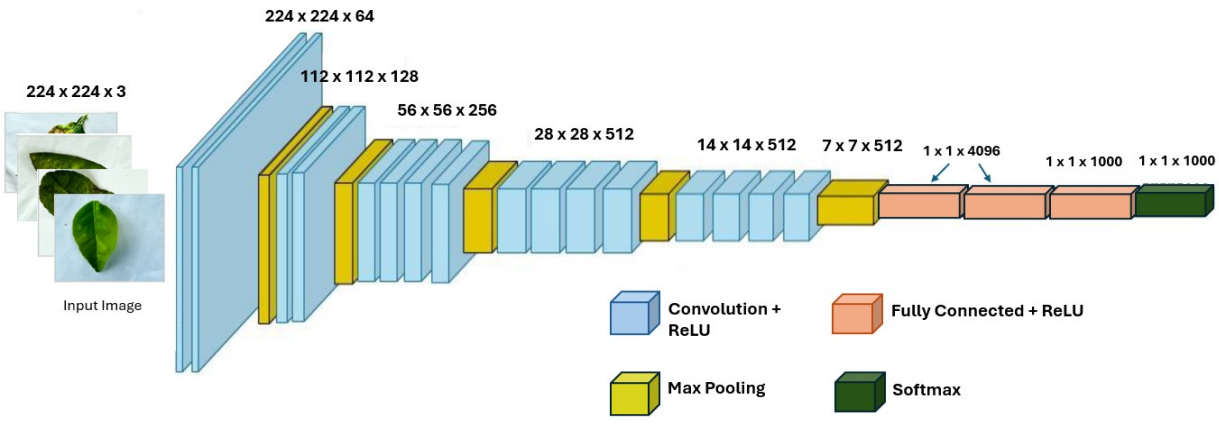


Figure 5: VGG16 Transfer Learning Architecture for Lemon Leaf Disease Detection

This hybrid transfer-learning and fine-tuning method combines the advantages of transfer learning and task-specific fine-tuning, resulting in improved classification performance in both seen and unseen datasets [20].

The VGG-16 transfer learning model used during the experiment is illustrated in Figure 5. Trained on ImageNet, the network utilises several stacked convolutional layers and max pooling to extract complex semantic features. Fully connected layers with ReLU and softmax activation are tuned for nine-class classification of lemon leaf diseases. The model can leverage transfer learning in the form of fine-tuning from pre-trained weights.

Table 3: Hyper Parameters settings for models.

Parameter	CNN (Proposed)	VGG16 (Transfer Learning)
Epochs	30	30
Batch Size	32	32
Total Parameters	≈ 0.33M	≈ 14.9M
Trainable Parameters	≈ 0.11M	≈ 0.067M
Non-Trainable Parameters	0	≈ 14.7M
Learning Rate	0.0001	0.0001
Loss Function	Categorical Crossentropy	Categorical Crossentropy
Optimizer	Adam	Adam

The training configurations of the CNN and VGG16 models are listed in Table 3, demonstrating their divergence in terms of complexity and scale. Although CNN is lightweight (≈0.33M parameters), VGG16 is much deeper (14.9M parameters), and both networks used the Adam optimiser with similar hyperparameters for a fair comparison. Layer-wise implementation for both CNN and VGG16 as follows. The models were trained with the categorical crossentropy loss function together with a softmax output layer since it is a multi-class classification task and the category is mutually exclusive.

1) Convolution Layer: Filters are used to capture the local spatial information of the input images. Learnable filter applied to input patches for capturing local features. As shown in (Eq. 1), each output pixel is obtained by a cross-correlation of the input window and the filter weight. This operation is capable of maintaining spatial position and location relationships, as well as capturing hierarchical structures (edges → textures → objects).

$$\text{Output}(i,j) = m\sum n\sum \text{Input}(i+m, j+n) \cdot \text{Filter}(m,n) + \text{Bias} \quad (1)$$

2) Batch Normalization: Normalizes the activations to accelerate training and improve stability. Speeds train up for normalization by zero mean and unit variance. The normalization step (Eq. 2a) eliminates internal covariate shift and the scale/shift transformation (Eq. 2b) maintains representational capacity.

Given input xx, BN computes:

$$x = \frac{x-\mu}{\sqrt{a^2}-e} \quad (2a)$$

where μ and σ^2 are the mean and variance of the mini-batch. The normalized output is then scaled/shifted:

$$y = \gamma x + \beta \quad (2b)$$

γ (scale) and β (shift) are learnable parameters; $\epsilon\epsilon$ avoids division by zero.

3) Clipped ReLU Layer: Adds non-linearity and avoids activation explosion. To introduce non-linearity but avoiding unbounded activations. Compared to the standard ReLU (Eq. 3), this restriction limits the maximum value (clip_value) to prevent saturation in deep networks.

$$ClippedReLU(x) = min(max(0, x), clip_value) \quad (3)$$

With clip_value = 6 , clip_value = 6, activations are constrained to [0,6].

4) Grouped Convolution Layer: Computational efficiency is enhanced by factorizing the filters into groups. As derived in (Eq. 4), which independently operate on a subset of input channels, and results in 1/k parameter reduction (k is the group number). This approach delivers broader networks at the same computation cost.

$$Groupk = Conv(Inputk, Filterk) \quad (4)$$

5) Fully Connected Layer: **Purpose**: Projects learned features to class scores via a linear transformation (Eq. 5). The weight matrix W describes combination rules for high-level features, but modern designs now use GAP instead to reduce the number of parameters.

$$y = Wx + b \quad (5)$$

W: Weight matrix (shape: input_dim×output_diminput_dim×output_dim).

x: Flattened input features.

6) Softmax Layer: The logit-to-probability conversion using the exponential normalization from (Eq. 6). The denominator normalizes so that output values sum to 1 and can be interpreted as a probability. Because of the exponential function.

$$P(yi) = \frac{e^{zi}}{\sum_{j=1}^{c} e^{zj}} \quad (6)$$

C is the number of classes; outputs a probability distribution.

7) Classification Layer: Predicts the winning class by the argmax (Eq. 7). This is a deterministic decision rule that transforms the softmax probability distribution into a scalar prediction and then selects the class with the highest softmax probability.

$$\text{Prediction} = \text{argmax}(P(yi)) \quad (7)$$

Table 4. Custom CNN Architecture

Layer Type	Output Shape	Parameters	Key Notes
Input	224 × 224 × 3	0	RGB image input
Conv2D (32) + MaxPooling	111 × 111 × 32	896	3×3 convolution, ReLU activation
Conv2D (64) + MaxPooling	54 × 54 × 64	18,496	Deeper feature extraction
Conv2D (128) + MaxPooling	26 × 26 × 128	73,856	High-level spatial features
Global Average Pooling	128	0	Replaces flattening to reduce parameters
Dense (128) + Dropout (0.3)	128	16,512	Compact fully connected layer
Dense (9) + Softmax	9	1,161	Multi-class classification
Total Parameters	—	332,765	Lightweight model (< 1M parameters)
Trainable Parameters	—	110,921	Optimized for efficiency

Architectural details of the two deep learning models used in this study are provided in Tables 4 and 5. The custom CNN (Table 4) is a lightweight baseline network consisting of three convolutional–pooling blocks, one fully connected layer, and one softmax classifier, total ≈0.33M parameters. On the other hand, the transfer learning model that is built on top of VGG16 (Table 5) uses a pre-trained convolutional backbone from ImageNet with global average pooling and dense layers connected to dropout regularization (total 14.9 million parameters). With these architectures, the tradeoff between efficiency and performance is effectively addressed for citrus leaf disease classification. The VGG16 convolutional base was frozen during training, and only the task-specific dense layers were fine-tuned, resulting in a small number of trainable parameters.

Table 5. VGG16 Transfer Learning Architecture

Layer Type	Output Shape	Parameters	Key Notes
Input (VGG16 Conv Base, Frozen)	224 × 224 × 3	14,714,688	Pre-trained ImageNet convolutional layers
Global Average Pooling	512	0	Feature aggregation
Dense (128) + Dropout (0.3)	128	65,664	Task-specific classification head
Dense (9) + Softmax	9	1,161	Multi-class output layer
Total Parameters	—	14,915,165	Includes optimizer parameters
Trainable Parameters	—	66,825	Only classification head trained
Non-Trainable Parameters	—	14,714,688	Frozen convolutional backbone

3.4 Explainable AI Integration

Although CNNs and VGG16 can achieve high prediction quality, deep learning models are black-box systems, which reduces the transparency and trust in their performance in agricultural settings. To address this need, we integrated XAI methods which enable visibility into how models make decisions visually. These approaches further promote interpretability of the areas contributing to the classification and contribute to a more practical useful method for farmers and agronomists by localising the most discriminative areas in citrus leaves.

3.4.1 Gradient-weighted Class Activation Mapping (Grad-CAM)

Class-discriminative heatmaps were created using Grad-CAM, which utilises the gradient information flowing into the final convolutional layers of the model. These heatmaps indicate the most important parts of input images for predictions. For citrus leaf disease diagnosis, Grad-CAM consistently drew attention to symptom-specific areas (lesions, discolouration, and fungal spots) as experts observed, yielding supporting evidence for the model's decisions.

3.4.2 SHapley Additive exPlanations (SHAP)

To measure the contribution of individual pixels or groups of pixels to a specific prediction, SHAP values were utilized. This framework enables the generation of a consistent and theoretically valid measure of feature importance by approximating Shapley values in cooperative game theory. All explanatory images in this study illustrated how colour intensities and textural methods influenced the predicted probability of a diseased or healthy leaf, providing an interpretable decomposition of model judgments.

3.4.3 Local Interpretable Model-Agnostic Explanations (LIME)

The LIME was used to interpret the predictions based on local superpixel segmentation. LIME identifies the super pixels responsible for classification by perturbing image regions and monitoring the impact on prediction. For citrus leaf-based diseases, the overlaid LIMEs effectively highlighted infected patches and suppressed irrelevant background, providing an intuitive and farmer-friendly interpretation of how the model distinguishes different diseases.

In the case of Grad-CAM analysis, feature attribution was calculated based on the last convolutional layer of each model because these layers are the most relevant spatial representations of features of diseases localization. The final Conv2D block was employed in the custom CNN, but, in VGG16, the block5_conv3 layer was applied in Grad-CAM. The representativeness was achieved by the conduct of explainability analysis on a representative half (90 test images) that was chosen randomly among the samples correctly classified in terms of disease to cover classes. Grad-CAM, SHAP, and LIME visualizations were created and compared with each other on a case-by-case basis and across images.

In the reviewed samples, there were recurrent areas of emphasis that were always correlated with the apparent symptoms of the disease like lesions, discoloration, and fungal spots in most instances. Qualitatively, more than

80 percent of the assessed explanations focused on disease-impacted areas as opposed to background or healthy leaf areas, which shows that the models used clinically and agronomically important characteristics to forecast. These results support the validity and clearness of the planned clarifiable model.

3.5 Evaluation Metrics

Standard metrics, including accuracy, precision, recall, and the F1-score, were used to evaluate the models' performance both at the overall and class levels. Correct and incorrect predictions across the nine leaf disease categories were presented through confusion matrices. In contrast, ROC curves and AUC values were calculated to assess the models discriminative capacity in a multi-class scenario. Table 6 presents the evaluation metrics adopted in our work, along with their verbal and mathematical definitions. It helps in having clear interpretation of model performance and enables cross comparison with other methodologies being more standardized.

Table 6: Evaluation Metrics of Model

Metric	Description	Formula
Accuracy	Proportion of total correct predictions among all predictions.	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	Measures how many predicted positives are actually correct.	$\frac{TP}{TP + FP}$
Recall (Sensitivity)	Measures how many actual positives are correctly predicted.	$\frac{TP}{TP + FN}$
F1-Score	Harmonic mean of precision and recall. Balances false positives/negatives.	$2 * \frac{Precision * Recall}{Precision + Recall}$
Specificity	Measures how many actual negatives are correctly predicted.	$\frac{TN}{TN + FP}$
ROC Curve	Graphical representation of TPR (Recall) vs. FPR across thresholds.	TPR = TP / (TP + FN), FPR = FP / (FP + TN)

3.6 Experimental setup environment

The data were randomly sampled and classified by a stratified 80% training and 20% validation division in order to maintain the class balance in all nine categories. The randomness was kept by means of a constant seed

(seed = 42). Both CNN and VGG16 models were also trained using the Adam optimizer with a learning rate of 1 x 10 -4, a batch size of 30, and 30 epochs in a workstation with an NVIDIA GPU and TensorFlow 2.x.

4. Result Evaluation

The experimental findings indicate a high level of detection and classification by the proposed DNN models in identifying citrus leaf diseases. Both the custom CNN and transfer learning based on VGG16 achieved high performance in prediction, and in all cases, the accuracy on the augmented dataset was above 90%. Large-scale quantitative studies that employ precision, recall, F1-score and class-wise confusion matrices have also indicated the influence of efficacy features and models developed. By incorporating explainable AI methods, such as Grad-CAM, SHAP, and LIME, the visual explanation of how the model arrives at its predictions was significantly enhanced, and it was found that the chosen areas correspond to the areas of disease-affected leaves. Besides the overall accuracy, there were fine differences in the pairwise comparison of the two models. The CNN was faster converging and less expensive to calculate and could therefore be deployed in the field on a lightweight device. Conversely, it was slightly lower than VGG16 in terms of precision and recall across various disease classes, with minimal differences even within disease lookalike similarity. These two competences are complementary, demonstrating the trade-off between efficiency and accuracy, and motivating the use of deep learning models to provide reliable AI in decision support systems that can lead to sustainable farming, provided that explainability is achieved.

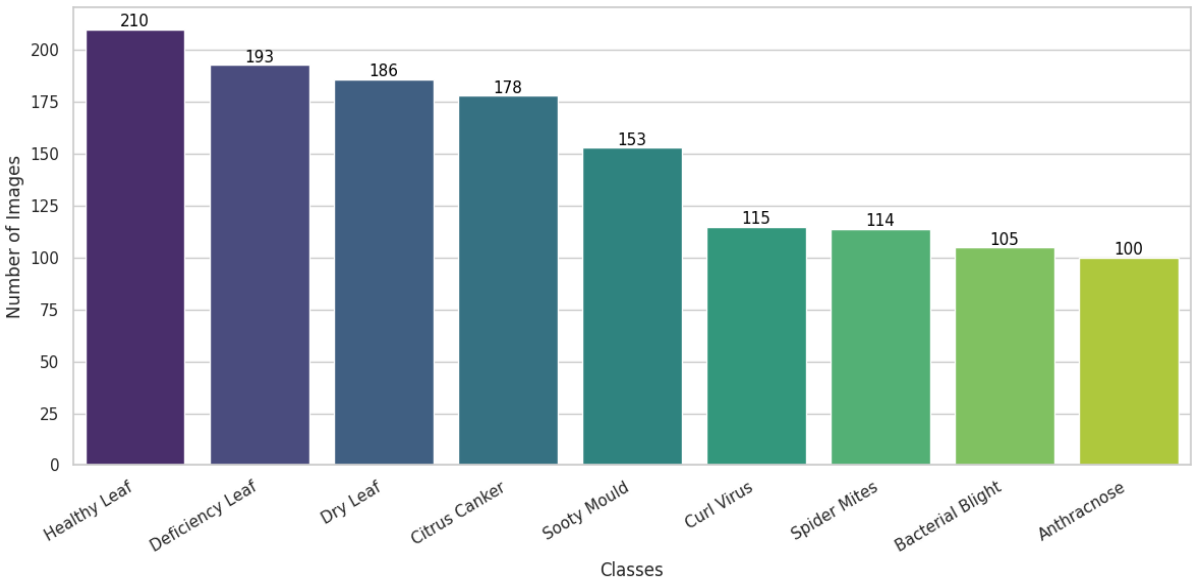


Figure 6: Lemon Leaf Disease Dataset – Class Distribution

Figure 6 shows the class distribution of the original dataset. It is clear from this figure that the categories' Healthy Leaf' and 'Deficiency Leaf' are overrepresented, whereas 'Anthracnose' and 'Bacterial Blight' are

underrepresented. It could harm model training if it is biased towards the majority class. To address the above-mentioned issue and provide even-handedness of all disease types, data augmentation methods were applied to create a well-balanced dataset, resulting in highly robust and generalizable deep learning models.

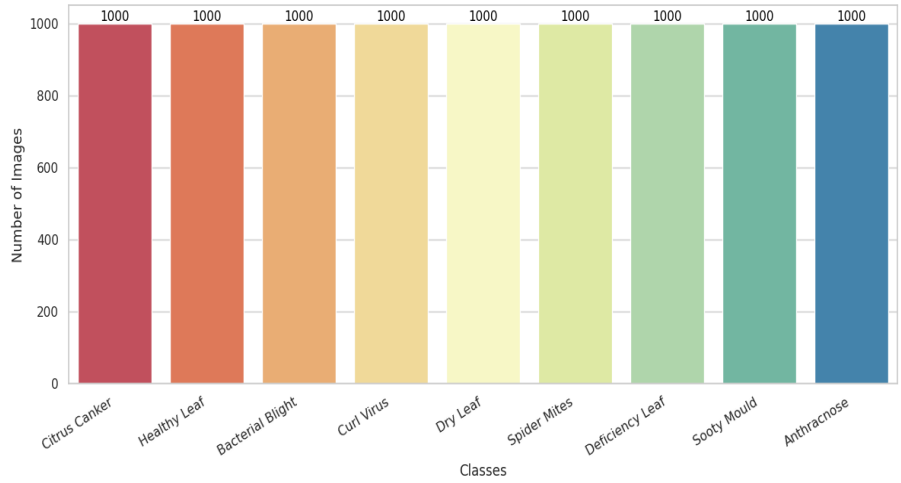


Figure 7: Balanced Lemon Leaf Disease Dataset – Post Augmentation

Figure 7 shows the balanced class distribution after employing systematic data augmentation, where there are 1,000 images per each of the total nine classes. This balanced data generation prevents bias in finding disease-specific features caused by an imbalanced original dataset, and thus permits the learning of diverse category-dependent characteristics equally. We enhanced the generalisation of models, lowered misclassification risks, and increased the reliability of citrus disease detection in practical applications with real agricultural scenes by balancing the dataset.

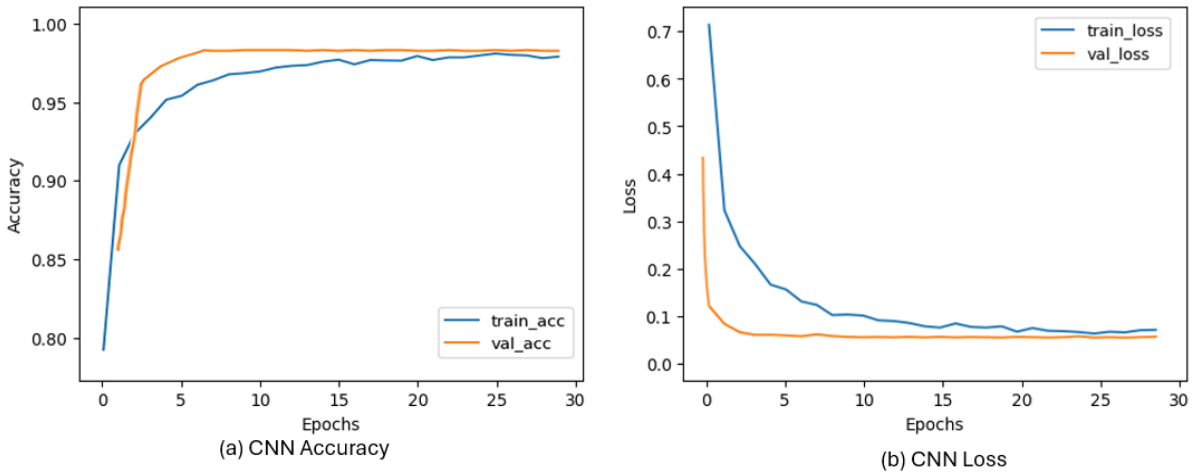


Figure 8: Training and Validation Performance of CNN

Figure 8 displays the accuracy and loss curves for the training and validation sets (30 epochs) of the CNN model. The accuracy curve increases rapidly in the early epochs, then levels off around 98%, with a slightly higher validation set accuracy than the training set accuracy, suggesting good generalisation. Similarly, the loss curves show a steep drop in training and validation losses, that smoothly converge at low values powering overfitting. These results indicate that the CNN model is capable of learning discriminative representations for the identification of citrus leaf diseases.

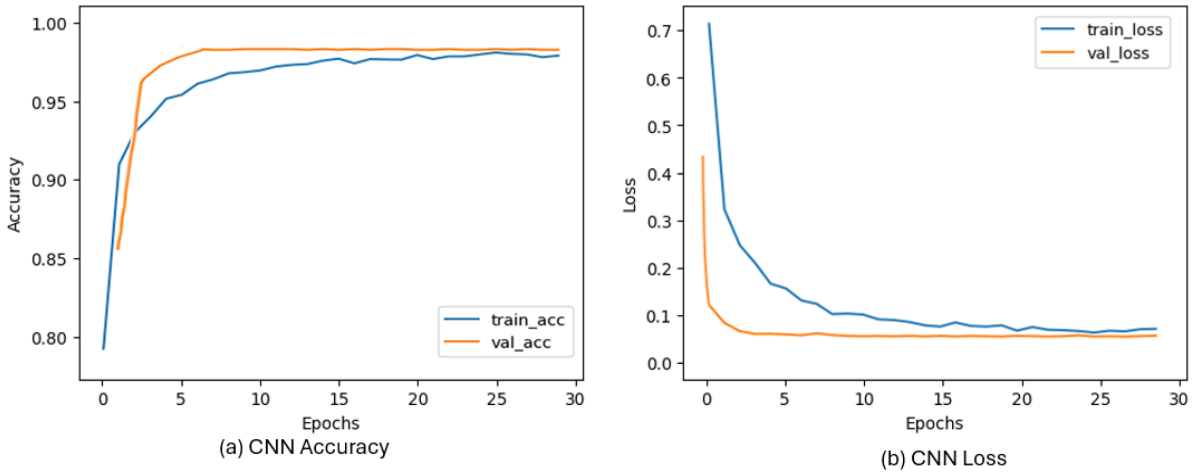


Figure 9: CNN Training and Validation Accuracy and Loss

Figure 9 show the training process of the CNN model. Subfigure (a) depicts the curves of accuracy, showing that both training accuracy and validation accuracy rise rapidly and converge above 97%, indicating that the model consistently learns. Subfigure (b) shows the loss curves, which both decrease gradually to low values before stabilising, indicating satisfactory convergence with minimal overfitting. Such phenomena also proved the generalization ability of the CNN in citrus leaf disease diagnosis.

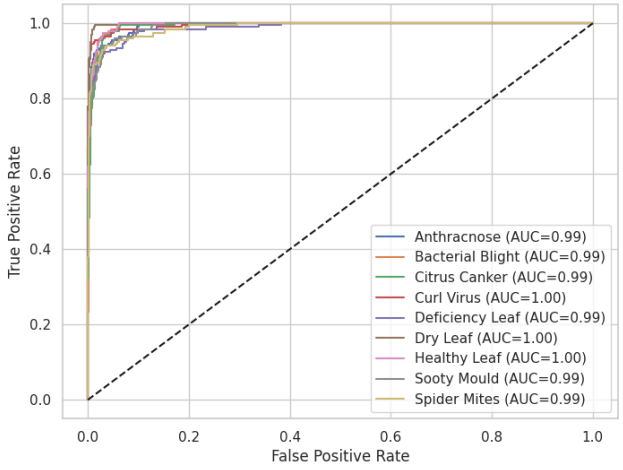


Figure 10: ROC Curve for CNN Model

As shown in Figure 10, the ROC curve highlights the class-wise performance of the CNN model for citrus leaf disease identification. The ROC analysis shows that the performance of the discriminative analysis is strong in all the nine classes with AUC values of 0.99 in Anthracnose, Bacterial Blight, Citrus Canker, Deficiency Leaf, Sooty Mould, and Spider Mites and 1.00 perfect AUC values in Curl Virus, Dry Leaf, and Healthy Leaf. Most classes have true positive rates of more than 0.95 at false positive rates less than 0.05 and thus have high sensitivity and low misclassification. These findings provide a quantitative support of the strength and stability of the proposed model in the process of multi-class citrus leaf disease detection.

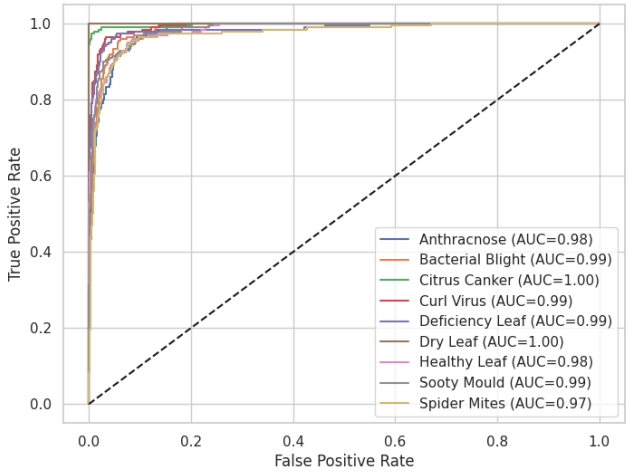


Figure 11 Class-Wise ROC Curve for VGG16

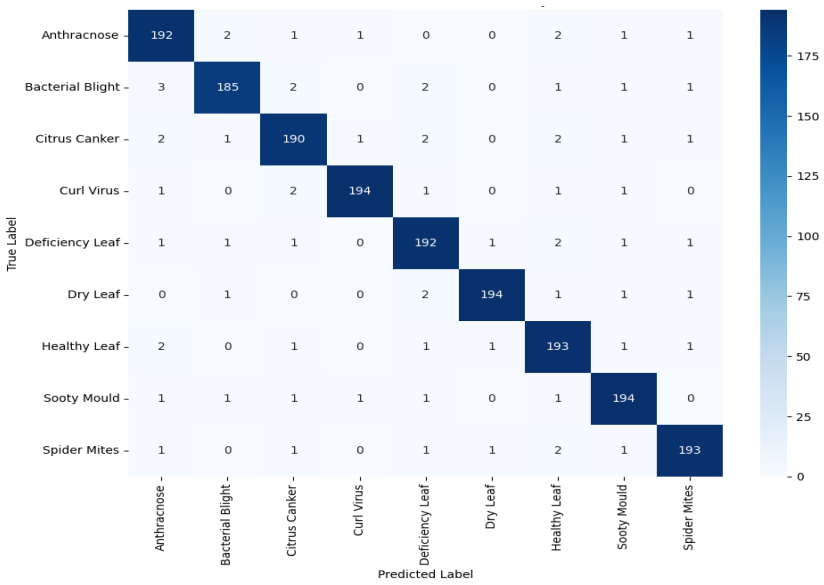


Figure 12: Confusion Matrix for CNN Model

The ROC curve shown in Figure 11 represents the classification performance of VGG16 model for all categories of citrus leaf diseases. The AUC values are between 0.97 and 1.00 with Citrus Canker and Dry Leaf having the highest AUCs of 1.00, Anthracnose having AUC of 0.98 and the lowest AUC of 0.97 of Spider Mites. Most classes achieve the true positive rates of more than 0.90 when the false positive rate is less than 0.10 meaning that it has high but slightly lower discriminative power as compared to the proposed model. Altogether, the findings validate the classification performance of all disease classes with slight and limited differences in the sensitivity of the visually similar conditions.

The confusion matrix represents the performance of CNN model on nine classes of citrus leaves as presented in Figure 12. Many predictions are in the top left quadrant, indicating high TP rates for all categories; Anthracnose, Curl Virus, Dry Leaf and Sooty Mould had particularly high precision. Some false assignments are distracting for classes like Bacterial Blight and Citrus Canker but those are still few. These results show that the CNN exhibits excellent classification performance and is well balanced between different disease pattern recognitions.

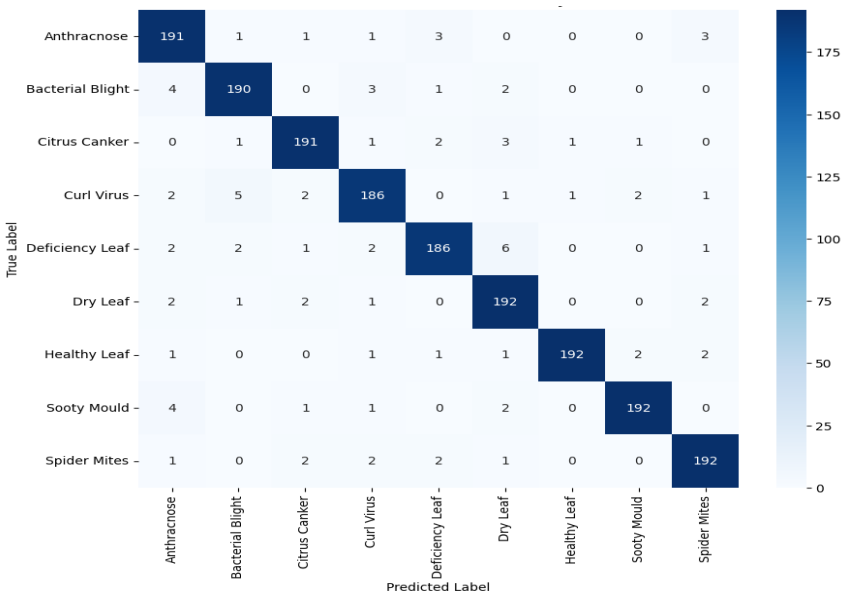


Figure 13: Confusion Matrix for VGG16 Model

The confusion matrix in Figure 13 shows the classification performance of the VGG16 model for nine classes of citrus leaf diseases. Although most predictions fall on the diagonal (showing good recognition for most of the categories), misclassifications tend to be more common than CNN. Curl Virus and Deficiency Leaf, etc., have higher overlaps with their neighbours, under sensitivity to visual closeness. Nevertheless, the model performs well in classification, achieving a general accuracy of 95%, which supports the effectiveness of our proposed transfer learning method.

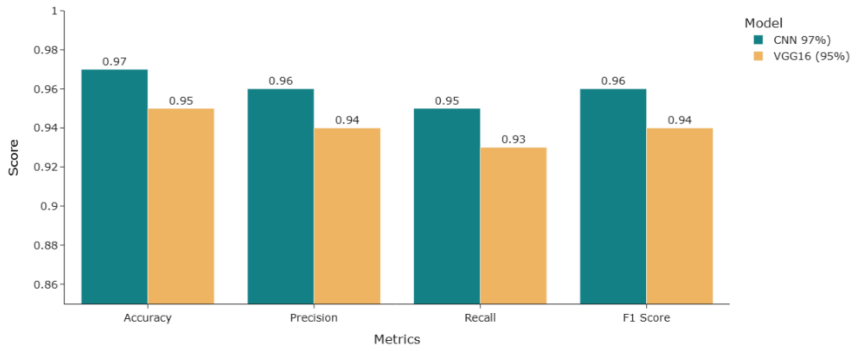


Figure 14: Model Comparison on Performance Metrics

Figure 14 illustrates the comparative performance of CNN (97%) and VGG16 (95%) on four key metrics: Accuracy, Precision, Recall, and F1-Score. The results in all categories are almost always higher for the CNN model, with an accuracy of 0.97, precision of 0.96, recall of 0.95, and an F1-score of 0.96, while VGG16 lags behind with values of 0.95, 0.94, 0.93, and 0.94, respectively. Findings demonstrate that the customised CNN architecture exhibits better generalisation and robustness than the transfer learning-based VGG16 for the automated classification of citrus leaf diseases.

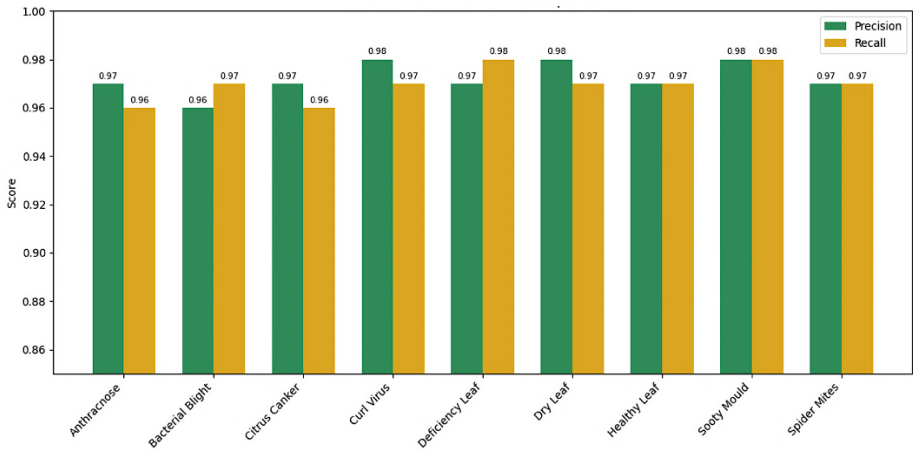


Figure 15: Precision and Recall per Class of CNN Model

Figure 15 shows the average class-wise precision and recall of the CNN model, with an overall accuracy rate of 97%. The obtained results, indicative of consistent behaviour in all nine classes with precision and recall values between 0.96 and 0.98, are tabulated. The categories of Dry Leaf, Curl Virus, and Sooty Mould receive the highest scores (0.98), which demonstrates that this model performs well for different types of diseases affecting citrus tree leaves. The near coincidence of precision and recall implies that there is a strong compromise between false positive and false negative minimisations, guaranteeing reliable disease detection.

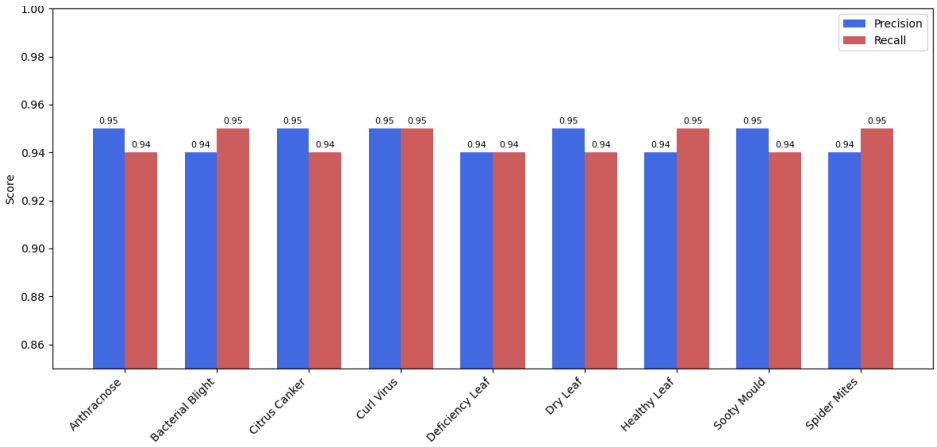


Figure 16: Precision and Recall per Class of VGG16 Model

Figure 16 illustrates the class-wise precisions and recalls of our VGG16 model, which achieved 95% accuracy. Both precision and recall fall within a tight range of 0.94–0.95, indicating that the method exhibits consistent performance across all nine citrus leaf disease classes. Although smaller than CNN, VGG16 possesses an equally good detection ability, thus guaranteeing that it avoids being biased in any one disease category. The results reveal that VGG16 is a very consistent transfer learning approach, but with slightly lower sensitivity than the CNN model.

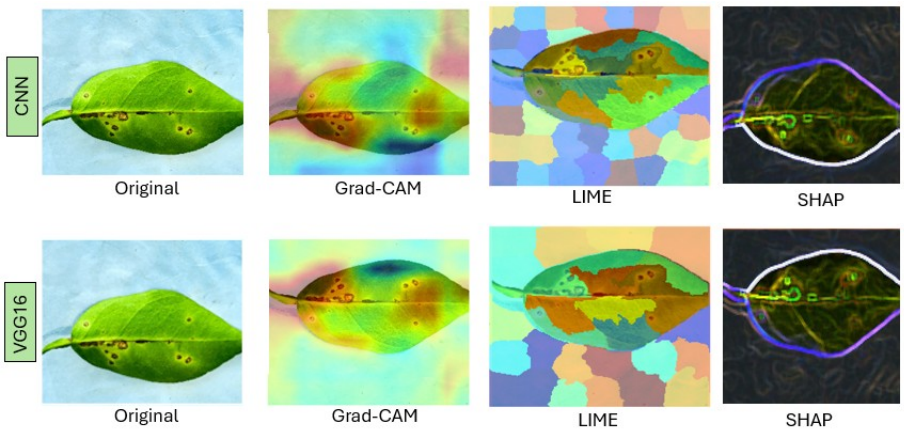


Figure 17: Comparative Explainable AI Visualizations for Citrus Leaf Disease Detection

Figure 17 presents a comparison of explanation outputs for CNN and VGG16, which use Grad-CAM, LIME, and SHAP, on a single sheet with an identical diseased leave image as input. Grad-CAM visually attains the broader lesion region of interest for classification, whereas LIME segments and accentuates superpixels contributing to predictions. The SHAP provides attribution of pixels that identify important symptomatic areas, consistent with expert diagnosis. In combination, these visualisations corroborate the models' logic, thereby increasing interpretability and confidence in real-world agricultural decision support.

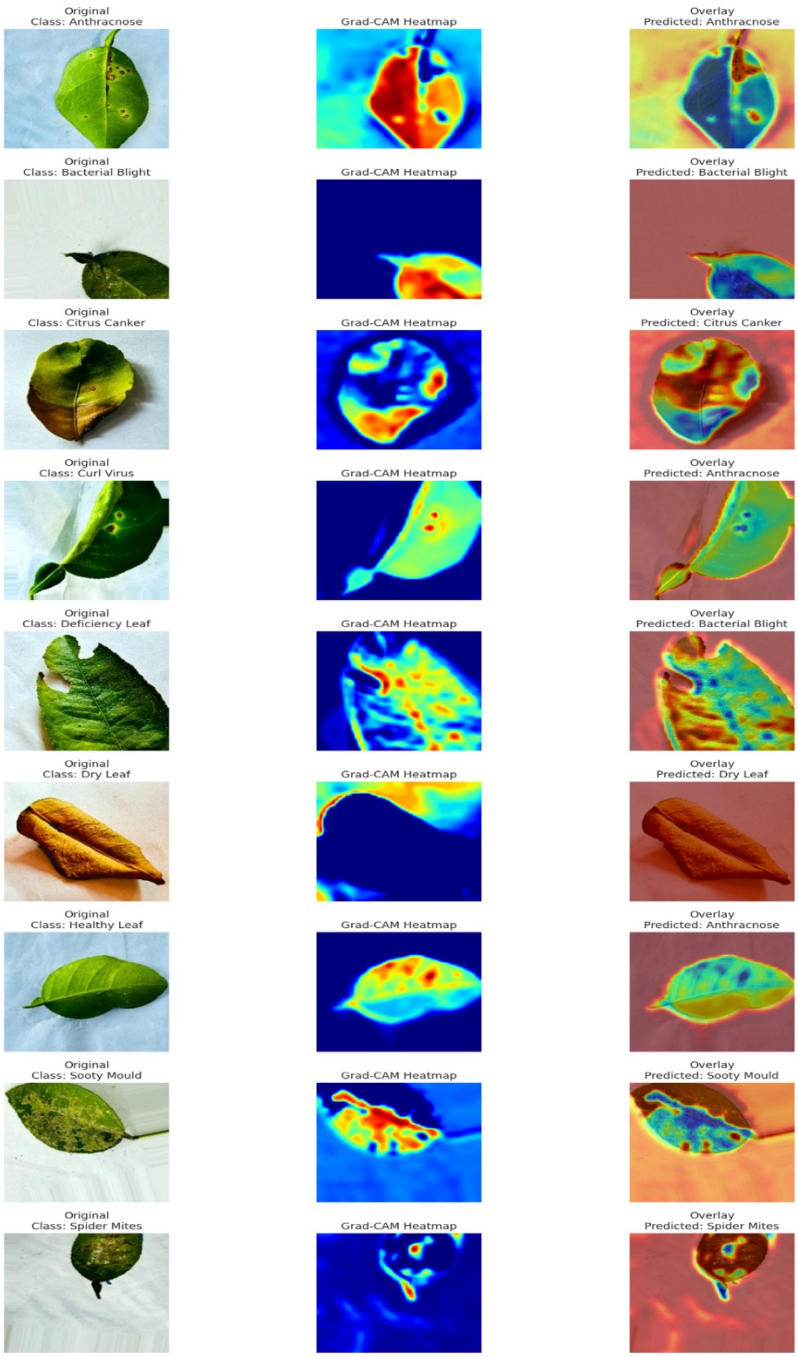


Figure 18: Grad-CAM–Based Visual Explanation of CNN

Figure 18 depicts the visual explanations (Grad-CAM) of the last convolutional layer of the CNN model to classify citrus leaf diseases. The disease-relevant areas including lesions, discoloration, and texture abnormalities are identified by the heatmaps and employed to make predictions with the models. Having proper classifications demonstrate the high coincidence between activated areas and observed pathological phenomena, whereas wrongly classified samples demonstrate the coincidence of visual cues belonging to different classes. These findings indicate that CNN makes decisions basing on significant leaf features and not background facts.

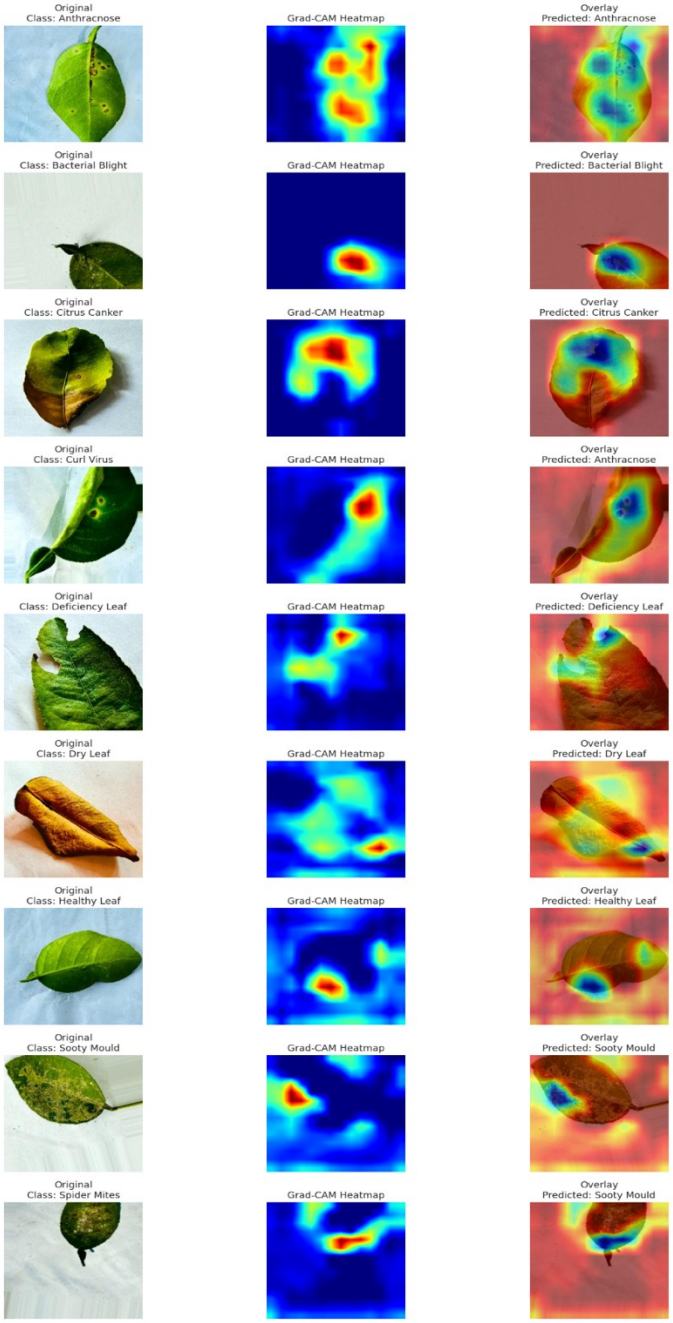


Figure 19: Grad-CAM–Based Visual Explanation of VGG16

Figure 19 demonstrates Grad-CAM visualizations generated on the basis of the last convolutional layer of VGG16 model to understand citrus leaf disease predictions. The salient areas of the disease identified with visual salience are the lesions, discoloration, and surface irregularities that were trained in ImageNet-pretrained features. Right predictions show high similarities between the activated and pathological areas whereas the misclassified ones show overlapping features activation across visually similar classes. Through these illustrations, it is possible to understand how transfer learning facilitates VGG16 to learn discriminative disease information and retain model interpretability.

The results have shown that, on average, CNN and VGG16 performed well in modelling for citrus leaf disease classification, though CNN slightly outperformed VGG16 with regard to overall accuracy, precision, recall, and F1-score. Explainable AI techniques, namely Grad-CAM, LIME, and SHAP, were incorporated to identify regions of interest as indicated by the models, which aligned well with observable disease symptoms. These observations not only confirm the robustness of the models but also enhance their interpretability, thereby improving their potential for agricultural adoption. In summary, the results suggest that integrating deep learning and XAI can provide well-performing, interpretable decision support systems for sustainable crop production.

5. Discussion

The results of this paper reinforce the transformational nature of deep learning in precision agriculture, particularly in the automated detection of citrus leaf disease. We were able to experimentally demonstrate that the custom CNN was more accurate than the VGG16 transfer learning model, highlighting the importance of domain-specific model design. Although both models worked efficiently in various disease classes, the CNN was consistently able to produce superior recall and F1-scores, which ensured a more reliable detection of underlying disease symptoms. Notably, a balanced and augmented dataset was used, which ensured resistance to class imbalance, a phenomenon frequently reported in studies of plant diseases. The findings suggest that the accuracy and reliability of agricultural artificial intelligence systems can be significantly enhanced when specifically designed architectures are implemented with the aid of various training datasets.

In addition to performance, this work focuses on the explainability aspect to ensure that AI in agriculture is reliable and can be effectively applied. Grad-CAM, LIME, and SHAP were not only effective in revealing the essential visual symptoms but also in good agreement with agronomic knowledge of experts, facilitating interpretability between the precision of the computer and the discernibility by the farmer. Our models can inform decisions by interpreting them, which can serve as a guide for practice during the early phases of disease monitoring and specific interventions. This combined high-performance classification with easy interpretability makes the proposed framework a viable and scalable solution for sustainable agriculture, benefiting direct farmers, researchers, and authorities.

Table 7 provides a comparison of the proposed explainable CNN-VGG16 framework with the newest state-of-the-art citrus leaf disease detection methods on the basis of model structure, data type, support of explainability and reported performance. In contrast to the YOLO-based and hyperspectral approaches that either focus on the accuracy of the detection, or use special imaging tools, the proposed study focuses on lightweight classification, multi-method explainability, and their practical deployability on normal RGB images. As noted by the comparison, the proposed approach is competitive in its accuracy, and it is the only one that offers transparent and understandable predictions that can be applied in the real-world support of agricultural decisions.

To contextualize the achieved accuracies of 97% for the custom CNN and 95% for VGG16, Table 7 provides the comparative analysis of the given results with the representative citrus and leaf disease detection methods in the literature. The comparison demonstrates that the suggested models perform competitively in comparison to detection-based and hyperspectral systems and is the only one that includes explainable AI practices, and their computational complexity is lower with the usage of standard RGB images.

Table 7. Comparative Positioning of the Proposed Method Against Recent Citrus Leaf Disease

Study	Core Methodology	Data Modality	Explainability	Performance	Key Limitation vs. This Work
Zhu et al., 2025 [21]	YOLOv8-CMS with CSPPC, attention modules + segmentation	RGB images	Not integrated	Accuracy: 97.9%, mAP50: 98.2%	Heavy detection model; no interpretability or decision transparency
Frederick et al., 2025 [22]	YOLOv8 with hyperspectral band selection	Hyperspectral (HSI)	Not integrated	Weighted F1: 0.896	Requires expensive sensors; limited field scalability
Zhu et al., 2025 [23]	CBACA-YOLOv5 with CBAM + coordinate attention	RGB images	Not integrated	Accuracy: 96.1%, mAP: 92.1%	Focused on detection; lacks interpretability and decision support
Frederick et al., 2024 [24]	YOLOv8 with supervised spectral band selection	Hyperspectral (adaxial/abaxial)	Not integrated	Accuracy: 87.1%	Complex imaging setup; lower accuracy than proposed CNN
Xiao et al., 2026 [25]	Spectral–physiological fusion with ML classifiers	Spectral + physiological	Not integrated	Accuracy: up to 98.9%	Single-disease focus (anthracnose); not image-based, no XAI
This Work	Custom CNN + VGG16 with unified XAI (Grad-CAM, SHAP, LIME)	RGB images	Multi-XAI integrated	Accuracy: 97%, F1: 0.96	Lightweight, interpretable, decision-oriented

Existing approaches often apply either no explainability strategy or use one visualization method, which is problematic when it comes to the assessment of trust. This research shows much stronger evidence that the learned features reflect the real disease symptoms as opposed to spurious correlations by comparatively analyzing Grad-CAM, SHAP, and LIME outputs of two different model families. The multi-XAI validation is a realistic development on the conventional CNN based classification pipelines.

6. Conclusion

This study proposed a deep learning model to explainable detect citrus leaf disease, which integrates the predictive power of the CNN and VGG16 models with interpretability methods, including Grad-CAM, LIME, and SHAP. The experimental findings revealed that the custom CNN achieved state-of-the-art accuracy (97%),

surpassing VGG16 (95%) in terms of accuracy, precision, recall, and F1-score, and was made robust by utilising advanced preprocessing and augmentation techniques. Explainable AI techniques enabled the clear visualisation of the disease-impacted areas of the land, which increased confidence in the results received by the models and matched the predictions with professional agronomic observations. These results demonstrate the need to strike a very delicate balance between predictive accuracy and transparency in the development of trustworthy AI-based systems in agriculture.

The broader impact of this study is that it can help develop sustainable and intelligent agricultural practices. The proposed framework can provide an early-warning system for crop health management, as well as a powerful decision-support tool, empowering farmers, agronomists, and policymakers by combining high-performance disease detection with explainability. This paper provides a basis for incorporating AI into real-time agricultural surveillance frameworks, which will enhance food security, minimise financial losses, and ensure environmentally friendly farming practices. Future directions include generalising the framework to other crops, utilising multimodal data (e.g., hyperspectral and environmental data), and running lightweight models on edge devices to implement the framework on a large scale in the field.

Funding

The authors declare that this research received no external funding.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] P. Sharma and P. Abrol, “Multi-component image analysis for citrus disease detection using convolutional neural networks,” Crop Protection, vol. 193, p. 107181, Jul. 2025, doi: 10.1016/j.cropro.2025.107181.
[2] A. Goyal and K. Lakhwani, “Integrating advanced deep learning techniques for enhanced detection and classification of citrus leaf and fruit diseases,” Scientific Reports, vol. 15, no. 1, p. 12659, Apr. 2025, doi: 10.1038/s41598-025-97159-0.
[3] K. Yan et al., “Three-dimensional fluorescence spectral characteristic of flavonoids for citrus Huanglongbing disease early detection,” Microchemical Journal, vol. 208, p. 112263, Jan. 2025, doi: 10.1016/j.microc.2024.112263.

[4] S. A. Higgins et al., “Plant-derived, nodule-specific cysteine-rich peptides as a novel source of biopesticides for controlling citrus greening disease,” Phytopathology, vol. 114, no. 5, pp. 971–981, May 2024, doi: 10.1094/PHYTO-09-23-0322-KC.
[5] P. K. Yadav et al., “Automated classification of citrus disease on fruits and leaves using convolutional neural network generated features from hyperspectral images and machine learning classifiers,” Journal of Applied Remote Sensing, vol. 18, no. 1, p. 014512, Jan. 2024, doi: 10.1117/1.JRS.18.014512.
[6] P. Dinesh and R. Lakshmanan, “Deep learning-driven citrus disease detection: A novel approach with DeepOverlay L-UNet and VGG-RefineNet,” International Journal of Advanced Computer Science and Applications, vol. 15, no. 7, pp. 1023–1041, Jul. 2024.
[7] C. Wang et al., “A simple and effective VIGS system facilitates the control of citrus canker by silencing CsLOB1,” Phytopathology Research, vol. 6, no. 1, p. 12, Mar. 2024, doi: 10.1186/s42483-024-00234-z.
[8] R. Sujatha, S. Krishnan, J. M. Chatterjee, and A. H. Gandomi, "Advancing plant leaf disease detection integrating machine learning and deep learning," Scientific Reports, vol. 15, Art. no. 11552, Apr. 2025, doi: 10.1038/s41598-024-72197-2.
[9] P. Dhiman, A. Kaur, Y. Hamid, E. Alabdulkreem, H. Elmannai, and N. Ababneh, "Smart Disease Detection System for Citrus Fruits Using Deep Learning with Edge Computing," Sustainability, vol. 15, no. 5, p. 4576, Mar. 2023, doi: 10.3390/su15054576.
[10] H. K. Kondaveeti and C. G. Simhadri, "Evaluation of deep learning models using explainable AI with qualitative and quantitative analysis for rice leaf disease detection," Scientific Reports, vol. 15, Art. no. 31850, Aug. 2025, doi: 10.1038/s41598-025-14306-3.
[11] Q. Xu, L. Ma, R. Zhou, C. Wang, J. Bai, L. Sun, and J. Cai, “Detection of huanglongbing infection in citrus using compositional analysis of volatile organic compounds,” Plant Pathology, vol. 73, no. 8, pp. 2084–2100, Oct. 2024, doi: 10.1111/ppa.13964.
[12] X. Guo, X. Wang, W. Zhu, S. X. Yang, L. Song, P. Li, and Q. Li, “Citrus disease detection based on dilated reparameterized feature enhancement and shared parameter head,” Sensors, vol. 25, no. 7, p. 1971, Mar. 2025, doi: 10.3390/s25071971.
[13] E. Aswini and C. Vijayakumaran, “Citrus plant classification using gated fusion adaptive graph neural network with harbor seal whiskers optimization algorithm,” Knowledge-Based Systems, vol. 329, p. 114093, Nov. 2025, doi: 10.1016/j.knosys.2025.114093.
[14] H. Zhu, D. Wang, Y. Wei, X. Zhang, and L. Li, "Combining Transfer Learning and Ensemble Algorithms for Improved Citrus Leaf Disease Classification," Agriculture, vol. 14, no. 9, p. 1549, Sep. 2024. doi: 10.3390/agriculture14091549. [Online]. Available: https://doi.org/10.3390/agriculture14091549
[15] B. Kaur, S. K. Gupta, M. Janarthan, D. M. Alsekait, and D. S. AbdElminaam, "Precision diagnosis of citrus leaf diseases using image enhancement and nonlinear fuzzy ranking ensemble approach NLFuRBe," Scientific Reports, vol. 15, Art. no. 32296, 2025. doi: 10.1038/s41598-025-16923-4. [Online]. Available: https://doi.org/10.1038/s41598-025-16923-4
[16] H. Mo and L. Wei, "Lightweight citrus leaf disease detection model based on ARMS and cross-domain dynamic attention," Journal of King Saud University – Computer and Information Sciences, vol. 36, no. 10, p. 102133, Oct. 2024. doi: 10.1016/j.jksuci.2024.102133. [Online]. Available: https://doi.org/10.1016/j.jksuci.2024.102133

[17] H. J. Min, J. Qin, P. K. Yadav, Q. Frederick, T. Burks, M. Dewdney, I. Baek, and M. Kim, "Classification of Citrus Leaf Diseases Using Hyperspectral Reflectance and Fluorescence Imaging and Machine Learning Techniques," Horticulturae, vol. 10, no. 11, p. 1124, Oct. 2024. doi: 10.3390/horticulturae10111124. [Online]. Available: https://doi.org/10.3390/horticulturae10111124
[18] A. Bora and P. Sarma, "A Visionary Analysis on Ways and Techniques Used for Identification of Citrus Plant Disease," 2025 3rd International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC), Silchar, India, Feb. 2025. doi: 10.1109/ISACC65211.2025.10969180.
[19] M. Franco-Hernández, V. Mejía-Domínguez, and Y. Rodriguez-Gallo, "Advanced Image Processing Methods for Early Detection and Monitoring of Aphid Infestations," 2024 IEEE Central America and Panama Student Conference (CONESCAPAN), Panama, Panama, Sep. 2024. doi: 10.1109/CONESCAPAN62181.2024.10891106.
[20] N. Valarmathi, M. Bharathi Kannan, M. Elankumaran, M. Gowri Shankar, and B. Gowtham, "Fruit Disease Prediction with Fertilizer Recommendation for Citrus Family using Deep Learning," 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, Mar. 2023. doi: 10.1109/ICSCDS56580.2023.10104601.
[21] H. Zhu, D. Wang, Y. Wei, P. Wang, and M. Su, "YOLOV8-CMS: A high-accuracy deep learning model for automated citrus leaf disease classification and grading," Plant Methods, vol. 21, no. 1, p. 88, Jun. 2025, doi: 10.1186/s13007-025-01396-3.
[22] Q. Frederick, T. Burks, J. A. Watson, P. K. Yadav, J. Qin, M. Kim, and M. M. Dewdney, "Supervised hyperspectral band selection using texture features for classification of citrus leaf diseases with YOLOv8," Sensors, vol. 25, no. 4, p. 1034, Feb. 2025, doi: 10.3390/s25041034.
[23] J. Zhu, J. Chen, H. He, W. Bai, and T. Zhou, "CBACA-YOLOv5: A symmetric and asymmetric attention-driven detection framework for citrus leaf disease identification," Symmetry, vol. 17, no. 4, p. 617, Apr. 2025, doi: 10.3390/sym17040617.
[24] Q. Frederick, T. Burks, P. K. Yadav, J. Qin, M. Kim, and M. Dewdney, "Classifying adaxial and abaxial sides of diseased citrus leaves with selected hyperspectral bands and YOLOv8," Smart Agricultural Technology, vol. 9, p. 100600, Dec. 2024, doi: 10.1016/j.atech.2024.100600.
[25] Z. Xiao, C. Yang, Z. Wei, X. Fu, C. Zhao, and G. Li, "A spectral–physiological feature fusion model for the early detection of anthracnose in citrus leaves," Computers and Electronics in Agriculture, vol. 243, p. 111410, Mar. 2026, doi: 10.1016/j.compag.2026.111410.

Islamic University Journal of Applied Sciences (IUJAS)
https://journals.iu.edu.sa/jesc

Volume VIII, Issue I, July 2026, Pages 218-229

Numerical study of air flow and temperature distribution in a room equipped with a cooling ceiling

Rim Ben Khalifa [1], Mohammed Awdah B Alshehre [2], Amamou Amani [1,3] and Nejla Mahjoub Said [2,*]

[1] LGM, National Engineering School of Monastir, 5000 Monastir, University of Monastir, Tunisia, rimbenkalifa@yahoo.fr

[2] Department of Physics, College of Science, King Khalid University, Abha 61413, Saudi Arabia, 445818821@kku.edu.sa, nalmahjoub@kku.edu.sa

[3] National Engineering School of Gasfa, University of Gafsa, 2112 Gafsa, Tunisia, amani.amamou@yahoo.fr

*Corresponding author: (Nejla Mahjoub Said), *Email Address*: nalmahjoub@kku.edu.sa

Abstract

The purpose of this paper is to simulate the most important problems in thermal comfort requirements, mainly a cooling ceiling system Using the RANS approach. A field flow model based on computational fluid dynamics was used to simulate the heat transfer of a room with the ceiling cooling system. The simulation is carried out using the Shear Stress Transport k-ω model. The main objective is to evaluate the heat transfer of a room equipped with a cooling ceiling. Firstly, the study describes the thermal performance of the cooling ceiling. Then, the effect of the water velocity on the performance of the floor is presented. Finally, the cooling ceiling system in a room is analyzed in terms of hydrodynamic and thermal performance for different Reynolds number. In addition, the results indicate that the cooling capacity increases nearly linearly with the temperature difference between the air-room and the cooling ceiling system.

Keywords: Dynamics (CFD); RANS turbulence modelling; Thermal comfort; Heat transfer and Reynolds number effects.

https://doi.org/10.63070/jesc.2026.024
Received 24 December 2025; Revised 12 April 2026; Accepted 16 April 2026.
Available online 20 April 2026.
Published by Islamic University of Madinah on behalf of *Islamic University Journal of Applied Sciences*.
This is an open access article published under the Creative Commons Attribution 4.0 International license (CC BY 4.0):https://creativecommons.org/licenses/by/4.0/

1. Introduction

Regarding changes in the climate and fossil fuel limitations, the growing demand of energy all over the world is becoming a serious problem. New and renewable energy sources seem to be the solution. The building sector is one of the largest consumers of energy with a large amount directly related to the heating and cooling of buildings.

As is well known in buildings, the presence of people, equipment, sunlight, or lighting generates a very large amount of heat that makes cooled-air temperature regulation systems insufficient. Many previous studies present the various modes and processes of heat transfer. Increasing the energy efficiency of the building is an alternative to reduce energy consumption while improving thermal comfort. The “Future of Cool”, where new technologies like cooling ceiling system should continue to grow in popularity for many buildings. This new technology to cool air uses much less energy and with a higher comfortable degree in the field of air-conditioning. Currently, ceiling cooling systems are increasingly in use. They are used for both cooling in the summer period and heating in cold periods.

Several researchers have studied the cooling performance under the influence of various parameters. The various modes and processes of heat transfer are discussed by Graziani et al. [1] and Xia et al. [2].

They made the transfer of heat to or from a room equipped with a cooling ceiling by two-mode of transfers: convection and radiation modes. In an experimental study Chicote et al. [3] studied the cooling capacity and heat transfer coefficients. In a test room equipped with a cooled radiant ceiling, they found that radiant and convective phenomena must be examined separately. They also determined that the radiant heat transfer coefficient attains lower values due to the influence of the ceiling surface emissivity. The total cooling capacity of the radiant ceiling can be determined by taking into consideration both radiant and convective heat fluxes.

The cooling capacity and heat transfer coefficients were the objectives of the numerical study of Refet [4]. With this study, the authors gave a new equation related to the convective heat transfer coefficient, heat transfer coefficient and the ratio of heat transfer coefficients. This equation depends only on the temperature difference between the indoor air and ceiling surface.

With the study of Jeong et al. [5], they made A comparison between the pilot DOAS/ radiant panel cooling system and a conventional all-air VAV system with air-side economizer. They found that the pilot DOAS/ radiant panel cooling system using more pumping energy. Furthermore, they showed that when the ceiling coverage ratio of suspended panels decreases, panel-cooling capacity increases. Chiang, et al [6] examined with a full-scale experiment a radiant cooling ceiling system installed in an

office. The results were compared with the values from the CFD simulation. Many parameters were analyzed such as air temperature from the diffuser, surface temperature, and the area of the cooling panels. This study improves indoor thermal comfort and energy efficiency using a radiant cooling ceiling system.

A series of numerical research on the ceiling cooling system were carried out. We can find the numerical simulation of Xie et al. [7]. In this computational fluid dynamics (CFD) simulation study, the author examined the heat transfer of water flow in the capillary of ceiling radiant cooling panel. Six influencing factors such as chilled water inlet parameters, conditions of gypsum plaster and capillary mats structural parameters were examined to obtain heat transfer performance. Yin et al [8] examined with a numerical simulation, a metal ceiling radiant cooling panels with a serpentine tube arrangement (CRCP-s). Further in this study, the research investigated the influences of the mean panel temperature, outlet water temperature and cooling capacity. The results show that all the parameters have a significant effect on the cooling capacity.

Ashrae [9] conducted experiments to examine the performance of using a cooling ceiling panel to transfer heat. This research refers to the effect of chilled water temperature, supplied chilled water flow rate, and the cooling performance of the panel. The experiment showed that supplied chilled water temperature has a direct influence on the surface temperature of the panel, the room temperature, and the heat transfer rate of the panel.

Causone et al. [10] evaluated experimentally the characters of heat transfer in the case of radiant cooling ceiling equipped in a room. It is remarked that several parameters could be affected the heat transfer coefficient: surface temperature distributions, internal gains, air movements. Experimental data emphasizes the fundamental role of the reference temperature to determine the heat transfer coefficients. The reference temperature is defined as the average unheated surface temperature (AUST) calculated considering the view factors between surfaces

In these terms, the study of Su et al. [11] analyzed the concrete ceiling radiant cooling panel. The temperature distribution of the concrete panel is represented under different conditions: water temperatures, distances of tubes, and water flow rate.

Also, the authors found that the large thermal inertia of the concrete panel caused the temperatures of the panel and indoor air to drop slowly. However, it reduced the adverse impact of supply water temperature fluctuation. This previous work is aimed at developing practical models to estimate the cooling output on indoor thermal comfort.

Karathanassis et al [12] investigated, with a numerical simulation, the flow and temperature fields in the case of ceiling cooling systems. In another study, Menter [13] introduced the Shear Stress Transport (SST) k-ω model. They, also, evaluated the prediction of the turbulent flow and heat transfer. The flow phenomena emerging inside the tubes were examined and thoroughly discussed. Sikula et al [14] analyzed the dynamic behavior of the temperature field of the cooling ceiling. The simulation was carried out using the Fluent 6.3 computer programs. The model of turbulence chosen was RNG k-ε to quantify in detail the cooling ceiling capacity and the ceiling surface temperatures for different Reynolds Number.

Many parameters are examined in different study like Reynolds Number, the dimension of the room

Also, we found a numerical and experimental study are focused to analyze the effect of Water velocity. Fredriksson and al. [15] found different phases of velocity: after an initial acceleration phase the velocity is almost constant. Then there is a slow decay in velocity with the distance from the chilled beam. This is attributed to the finite length of the chilled beam, which gives rise to entrainment of air at the ends of the plume. A CFD simulations is realized with the aim to analyse the airflow pattern and the velocity fields inside the test room in the study of Catalina [16]. It was shown in their work that a local discomfort at feet/ankle zone was observed but with good values of air velocity for the rest of the test room.

This paper reports a numerical study of a cooling ceiling installed in a test room. Ansys Fluent is a powerful CFD software that is powerful and up to date in all simulation fields related to fluid mechanics and heat transfer. The Shear Stress Transport (SST) k-ω model already implemented to predict the mean flow field and the heat transfer in the boundary layer. The mean temperatures, the behavior of the global flow generated by a cooling ceiling are presented. Additionally, the effects of the velocity of water on the thermal performance are examined. Finally, the flow and temperature fields in a room equipped with a cooling ceiling are analyzed.

2 Numerical simulations

2.1 Governing equations

Stokes equations (RANS equations), based on the Reynolds decomposition, are numerically solved. For such flows, the following assumptions are made in this numerical simulation: steady and incompressible fluid flow. Based on these assumptions, the governing equations including the continuity (1), the momentum (2) and the energy equation (3), are considered as follows:

$$\frac{\partial u_i}{\partial x_i} = 0 \qquad (1)$$

$$\frac{\partial}{\partial x_k}(\rho u_k u_i) = -\frac{\partial P}{\partial x_i} + \frac{\partial}{\partial x_k}\left[\mu\left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i}\right) - \rho\overline{u_i' u_k'}\right] \quad (2)$$

$$\frac{\partial}{\partial x_k}(\rho u_k T) = \frac{\partial}{\partial x_k}\left(\frac{\mu}{Pr}\frac{\partial T}{\partial x_k} - \rho\overline{u_k' T'}\right) \quad (3)$$

Where i, k =1, 2, 3 referring to x, y and z respectively. u_i and x_i are the mean velocity and the coordinate in the i direction, respectively. P represents the pressure and ρ is the density. μ and Pr are the dynamic viscosity and the Prandtl number, respectively. T designates the temperature. The Reynolds Averaged Navier-Stokes (RANS) equations formulated above contain the Reynolds stresses terms $\overline{u_i' u_k'}$ and the turbulent heat flux $\overline{u_k' T'}$. The Shear Stress Transport (SST) k-ω model introduced by Graziani et al. [1] is used for the calculation of these terms.

2.2. Flow configuration and boundary conditions

The main dimensions of the test room with a cooling ceiling are given in Figure 1. In order to examine the cooling performance, the study was conducted in an office room (1.2m × 0.55 m × 0.6 m). It is equipped with a hydronic ceiling composed of six serpentine tubes. Airflow will enter the room through a door and check the measurements through a window. The rest of the domain was considered an adiabatic wall.

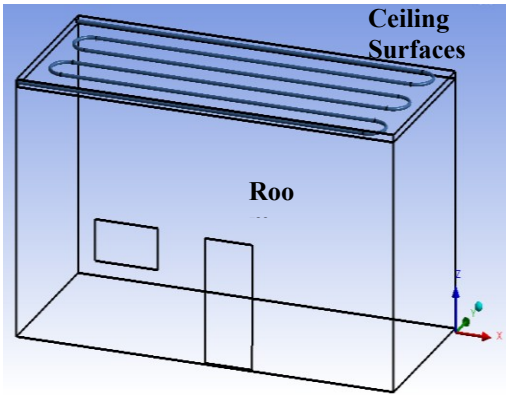


Figure 1. The main dimensions of Room with Cooling Ceiling System

The ceiling cooling panel is installed in the room. The cooling device comprises gypsum serpentine tubes thermally bonded to rectangular aluminum. The simulation was performed under these conditions: the surface temperature of all walls was set ambient temperature (T_a=24°C) and the cooling water temperature "T_w" in the tubes of the ceiling was set to T_w=10°C. The air temperature of the chamber was kept constant at 24 °C. The dimension of the main numerical domain is listed in Table 1. The cooling ceiling configuration is inspired from the flow configuration of Yin et al. [8]. The dimensions of the radiant panel comprise three distinct layers: an insulation layer, a water pipe layer, and a surface layer. (Figure 1).

Table 1: Main dimensions and materials

Items	Values
Tube outside diameter (D)	0.009 m
Number of tubes (N)	6
Panel thickness (spanwise direction (z))	0.015m
Panel material	Gypsum
Length (axial direction (x))	1.2 m
Width of ceiling surface (vertical direction (y))	0.55m
Velocity air (V_a)	0.3 m^2/s
width of Room ((axial direction (z))	0.6m

3. Numerical method

The resolution of the governing equations with appropriate boundary conditions is carried out, by using Ansys Fluent [17]. The numerical code uses the finite volume method; the solution of equations being based on the algorithm SIMPLER proposed by Patankar [18]. For the algebraic equations obtained, we use an implicit scheme.

The convergence absolute criterion is set to a residual value less than 10^{-5} for the continuity and velocity component equations and afterward, the energy equation.

The computational domains for both two parts "room and cooling ceiling" was discretized using unstructured grids. A sufficiently fine grid was closely positioned adjacent to the solid ceiling panel where prevails a very high gradient of different variables.

4. Results and discussion

4.1 Analysis the total thermal performance of the cooling ceiling

4.1.1 Flow Field

Figure 2 gives an example of the evolution of the surface temperature on the lower face of the ceiling.

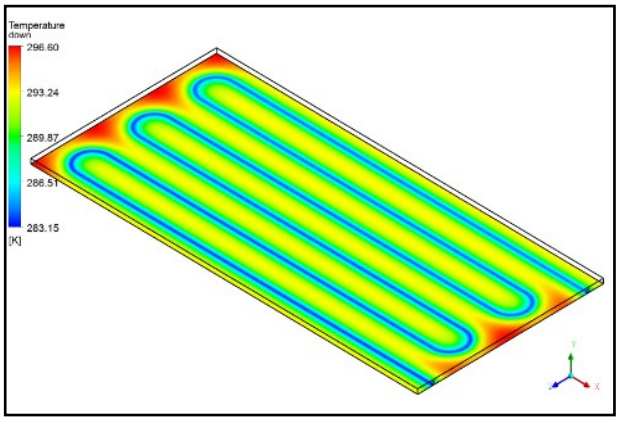


Figure 2. Scheme of temperature fields of ceiling surface

Tests were conducted for a Reynolds number of 2077, based on the diameter of pipe "D" and the water velocity "V_w".). The Reynolds number for pipe flow is defined as $Re = V_w \times D/\nu$.

As expected, the lowest temperatures are below the cooling pipes. It is obvious that the surface temperature distributions could be found a similar trend depicted in literature [7].

A large temperature gradient occurs between water fluid and gypsum ceiling, in all cases, to be driven by convection from the adjacent walls. It can be seen from this figure the values were uniformly located at the panel surface. One can also notice ceiling surface temperatures near the outdoor walls that are significantly higher - around 24 C.

The temperature fields in a longitudinal plane x–y for three planes (x = 0.4 m, x = 0.6 m and x = 1.15m) is shown in figure 3. This figure shows streamlines of temperature, observed for different positions to determine the spreading behavior of the temperature. The water temperature affects the surface temperature of the panel. Cold fluid is deflected toward the outer tube wall, causing the diminution of the ceiling temperature. Downstream of the water pipe, the total temperature distribution is uniform. With a cooling ceiling the temperature of water or its flow rate can be modified so that the surface temperature adapt to the desired conditions.

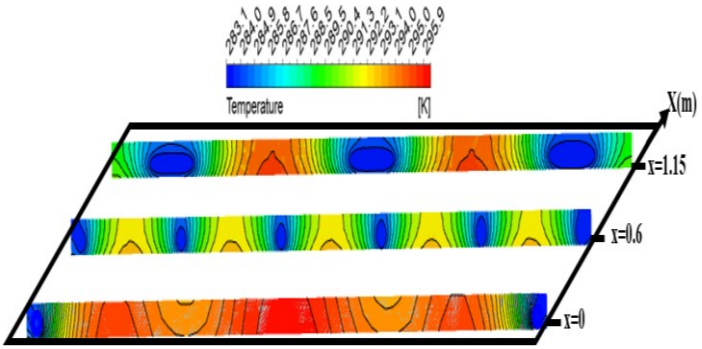


Figure 3. Two-dimensional temperature distribution of the cooling ceiling panel

Figure 4 (a, b) presents the distribution of water velocity in the symmetry plane (y=0 m) and the streamwise velocity profiles of the downstream pipe flow. The cool water flow enters the pipe at a uniform velocity (0.3 m/s). The region beyond the entrance region in which the velocity is changing significantly. As is clear from Figure 4 (a), the velocity of the fluid at the midsection of the pipe must increase to keep the mass flow rate through the pipe constant. As a result, a velocity gradient develops along with the pip. It is clear that the velocity profile appears distorted downstream of the inlet water, at the curved region (Figure 4 b). When the fluid in the pipe is not cooled, in this case, the velocity profile is remains unchanged.

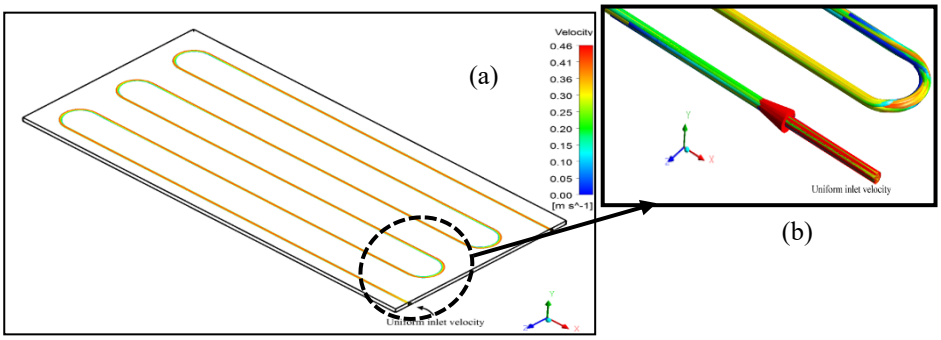


Figure 4. (a) Scheme of velocity fields the symmetry plane, (b) Streamwise velocity profiles of the downstream pipe flow

4.1.2 Effect of the Reynolds number

To further elucidate the global flow of the cooling system, Karathanassis et al [12] examined the flow in the cooling configurations for lower values of the Reynolds number. In our study, the temperature distribution covering a wide range of Reynolds numbers was investigated (Re=692-2770). The Reynolds number for pipe flow is defined as $Re = V_w \times D/\nu$ with ν is the kinematic viscosity of the water.

The distribution of the vertical temperature along the x-direction for different Reynolds numbers is illustrated in figure 5. As shown in this figure, the heat transfer rate depends on the cooling capacity rate of water temperatures. Therefore, the water flow temperature decreases with increasing Reynolds number.

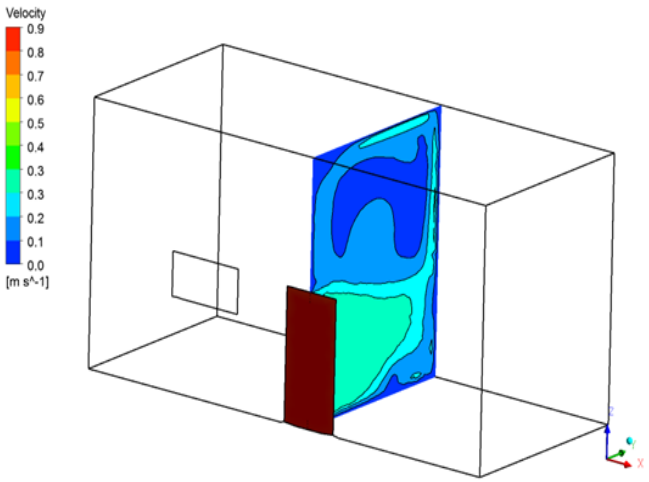


Figure 5. Distributions of the vertical temperature along the x-direction for different Reynolds number

For Reynolds number values greater than 1384, the cooling effect was enhanced, however, with the velocity increasing. The total heat transfer and the cooling capacity reduces with the reduction of

velocity water. Therefore, it can conclude that the velocity water should be large enough for greater cooling capacity.

4.2 Analysis of the behaviour of the cooling ceiling in an office room

In our study, the investigation of the heat exchange between the cooling ceiling with the room thermal is presented. Temperature of the surface is an important factor exercising significant influence on the process of heat transfer to air-water mist. The temperature of the surface is shown in Fig. 6 for different Reynolds numbers. The ambient air velocity enters to the room through the door is equal to 0.3 m/s. The vertical temperature fields in the test room in the plane x =0.6 m is examined in this Figure.6.

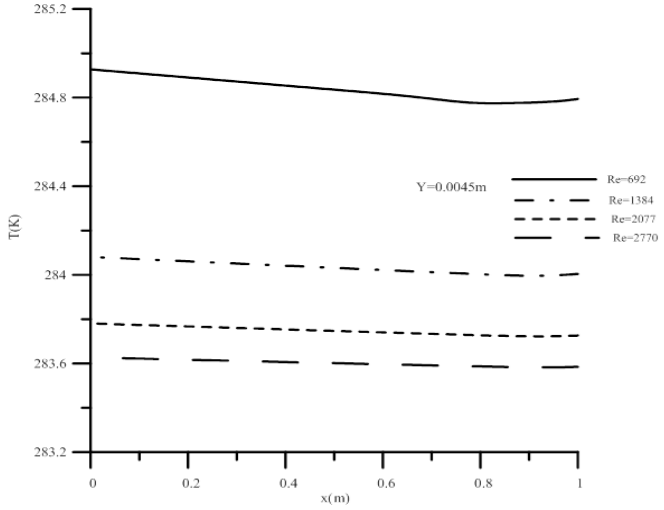


Figure 6. The total temperature fields inside the test room in the plane x=0.6m.

The air movements are occurring in the test room. The warm air enters the room from the door with ambient temperature and is rising to the ceiling. The presence of a cooling ceiling allows the generation of a cold downdraught. The rising air creates an interior low pressure that should pull in large volumes. The heat exchange increases nearly linearly with the temperature difference between the room air and the cooling ceiling system. In this condition, a stable thermal stratification developed inside the room, with warm still air close to the heated ceiling and low convective heat transfer. According with the Figure ,6 Surface temperature decreases with increasing Reynolds number. Because of evaporation of water, surface temperature decreases in the air-water mist with increasing water flow density at constant heat flux. It can be seen so that the higher the heat capacity ratio, the higher the convective heat transfer coefficient. This is because a higher heat capacity ratio leads to a more balanced flow and thus an increase in heat transfer rate.

Figure 7 presents the behavior of the velocity inside the room with the presence of a cooling ceiling system and in the plane x=0.6m. This figure indicates that the room will be divided into a lower region

where the air is like the supply condition and an upper region above the stationary front that is well mixed cold air. It is remarked also that when an airstream strikes the cooling ceiling surface, it tends to spread and follow the surface until the velocity dissipates. Therefore, it is interesting to examine the influence of air velocity because it is seemed to contribute more significantly to the enhancement of cooling panel capacity.

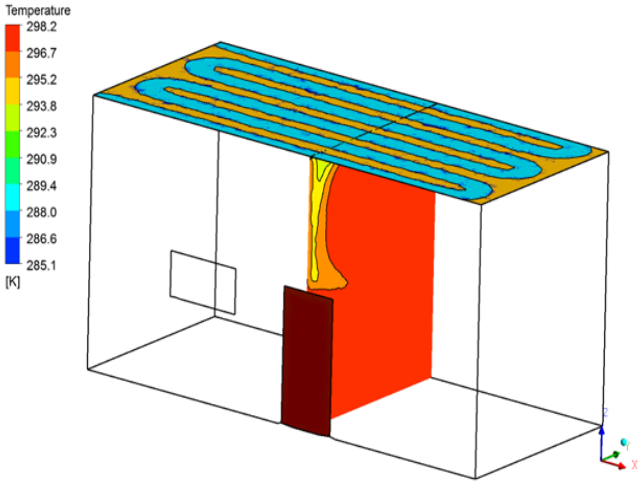


Figure 7. Velocity fields inside the test room in the plane x=0.6m.

5. Conclusion

The performance of the cooling ceiling installed in a room has been investigated in this paper using CFD simulations. This examination was conducted numerically employing of the finite volume method together with the Shear Stress Transport (SST) k-ω and a non -uniform mesh system.

Coldwater is spread out along the pipes causing the diminution of the ceiling surface temperature. A higher Reynolds number leads to a lower outlet temperature of water. It was established that an increase in the Reynolds number affects heat transfer. For practical applications, the Reynolds number should be large enough to keep the cooling capacity and the thermal comfort of the room. Secondly, the dynamic characteristics and the thermal behavior in the room with the presence of a cooling system are examined.

The room receives a very even temperature spread between the ceiling and the floor. Significant mixing, indicated by reduced temperature gradients, was evident in the upper part of the room with ceiling cooling. Velocity and Temperature measurements indicated that with greater heat gains the plumes have enough momentum to drive flow across the ceiling surface and down the walls.

Funding

The authors declare that this research received no external funding.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] S. Graziano, S., Jens, P. Marta, M.S. Assessing energy and thermal comfort of different low-energy cooling concepts for non-residential buildings, Energy Conversion and Management, Volume 76, December 2013, Pages 332-341.

[2] Yizai Xia, Mumma, A. Ceiling Radiant Cooling Panels Employing Heat-Conducting Rails: Deriving the Governing Heat Transfer Equations, ASHRAE Transactions. 2006, Vol. 112 Issue 1, p34-41. 8p.

[3] Andrés-Chicote, M. Tejero-González, A. Velasco-Gómez,E. F. J. Rey-Martínez, E. Experimental study on the cooling capacity of a radiant cooled ceiling system, Energy and Buildings, Volume 54, November 2012, Pages 207-214.

[4] K. Refet, K. New approach relevant to total heat transfer coefficient including the effect of radiation and convection at the ceiling in a cooled ceiling room. Applied Thermal Engineering, 29 (8-9), 1561-1565.

[5] Jeong, J.W. Mumma, S.A. Bahnfleth, W.P. Energy Conservation Benefits of a Dedicated Outdoor Air System with Parallel Sensible Cooling by Ceiling Radiant Panels, ASHRAE Transactions; Atlanta Vol. 109, (2003): 627-636.

[6] Chiang, W.H. Wang, C.Y. Huang, J.S. Evaluation of cooling ceiling and mechanical ventilation systems on thermal comfort using CFD study in an office for subtropical region, Building and Environment, Volume 48, February 2012, Pages 113-127.

[7] Xie, D. Wang, Y. Hanqing, W. Shunquan, M. Maili, L. Numerical analysis of temperature non-uniformity and cooling capacity for capillary ceiling radiant cooling panel, Renewable Energy, Volume 87, Part 3, March 2016, Pages 1154-1161.

[8] Yin, Y. L. Wang, R. Z. Zhai, X. Q. Ishugah, T. F. Experimental investigation on the heat transfer performance and water condensation phenomenon of radiant cooling panels, Building and Environment, Volume 71, January 2014, Pages 15-23.

[9] ASHRAE, HVAC Systems and Equipment (SI), (2008).

[10] Causone, F. Corgnati, S.P. Filippi, M. Olesen, B.W. Experimental evaluation of heat transfer coefficients between radiant ceiling and room, Energy and Buildings, 41 (2009)

[11] Su, L. Li, N. Zhang, X. Sun, Y. Qian, J. Heat transfer and cooling characteristics of concrete ceiling radiant cooling panel, Applied Thermal Engineering, ATE, 84 (2015)

[12] Karathanassis, I.K. Papanicolaou, E. Belessiotis, V. Bergeles, G.C. Flow and Temperature Fields in Cooling Devices with Embedded Serpentine Tubes, Numerical Heat Transfer, Part A: Applications 66 (4), 349-378, 2014.

[13] Menter, F.R. Two-equation eddy-viscosity turbulence models for engineering applications, AIAA J, 32, 1598–1605 (1994).

[14] Sikula, O. Hirs, J. Analysis of the dynamic behavior of the cooling ceiling in an office room with a view of thermal comfort, ventilation and comfort strategies, 3294- 3302, Proceedings The 23rd IIR International Congress of Refrigeration, 1-8, (2011) .

[15] Fredriksson , J.Sandberg , M. Moshfegh.B. Experimental investigation of the velocity field and airflow pattern generated by cooling ceiling beams. Building and Environment Volume 36, Issue 7, Pages 891-899 (2001).

[16] Catalina ,T. Virgone,J. Kuznik,J. Evaluation of thermal comfort using combined CFD and experimentation study in a test room equipped with a cooling ceiling, Building and Environment

Volume 44, Issue 8, Pages 1740-1750 (2009).

[17] Ansys Fluent 13.0.0, Ansys Fluent Inc., (2010)

[18] Patankar, S. V. Numerical Heat Transfer and Fluid Flow, McGraw-Hill, New York, 1980.

Islamic University Journal of Applied Sciences (IUJAS)
https://journals.iu.edu.sa/jesc

Volume VIII, Issue I, July 2026, Pages 230-248

OCSA: An Optimized Cuckoo Search Algorithm with Guided Exploration and Refined Local Search

Riad Bouakacha[1], Sofiane Haddad[2,*], Ammar Boulaiche[3] and Abdelhamid Rabhi[4]

[1] Electrical Department, MSB Jijel University, BP 98, Ouled Aissa, Jijel-1800 Algeria
[2] REL Laboratory, Electronics Department, MSB Jijel University, Algeria
[3] LaRIA Laboratory, Faculty of Exact Sciences, MSB Jijel University, Jijel 18000, Algeria.
[4] MIS Laboratory, Université de Picardie Jules Verne, Amiens, Picardie, France

*Corresponding author: (Sofiane Haddad), *Email Address*: s_haddad@univ-jijel.dz

Abstract

This paper introduces an optimized version of the Cuckoo Search Algorithm (CSA), termed the Optimized CSA (OCSA), designed to address key limitations of traditional population-based optimization methods, including slow convergence, limited precision, and high memory usage. OCSA improves both exploration and exploitation through two main contributions. First, it employs a multi-directional guided displacement strategy around the current best solution, enhancing the algorithm's ability to explore promising regions more effectively. Second, it strengthens local search by having each cuckoo generate multiple candidate solutions and selecting the best among them, which improves solution accuracy and accelerates convergence. The performance of OCSA is evaluated using benchmark functions, engineering design problems, and a real-world application in economic dispatch. Comparative evaluations against established algorithms, including Cooperative Coevolutionary CS (CCCS), Particle Swarm Optimization (PSO), and Genetic Algorithm (GA), as well as mathematical solvers such as Interior Point OPTimizer (IPOPT) and Linear and Objective Quadratic (based on an infeasible primal-dual interior-point) Optimizer (LOQO), demonstrate that OCSA achieves superior convergence speed, solution quality, and scalability. These results establish OCSA as an efficient optimization approach for complex problems.

Keywords: Optimized Cuckoo Search Algorithm; Metaheuristic Optimization; Exploration; Exploitation; Enhanced Local Search.

https://doi.org/10.63070/jesc.2026.025
Received 09 April 2026; Revised 18 April 2026; Accepted 25 April 2026.
Available online 30 April 2026.
Published by Islamic University of Madinah on behalf of *Islamic University Journal of Applied Sciences.*
This is an open access article published under the Creative Commons Attribution 4.0 International license (CC BY 4.0):https://creativecommons.org/licenses/by/4.0/

List of abbreviations

Abbreviation	**Full Form**
ABC	Artificial Bee Colony
AIS	Artificial Immune System
AIS-GA	Artificial Immune System combined with Genetic Algorithm
ALPSO	Augmented Lagrangian Particle Swarm Optimization
BBA	Binary Bat Algorithm
BFO	Bacterial Foraging Optimization
CCCS	Cooperative Coevolutionary Cuckoo Search
CCGA	Cooperative Coevolutionary Genetic Algorithm
Charibde	Cooperative Hybrid Algorithm using Reliable Interval-Based Methods and Differential Evolution
CHc	Constraint Handling (coevolutionary)
CHp	Constraint Handling (penalty)
COPs	Constrained Optimization Problems
CPSO	Cooperative Particle Swarm Optimization
CPSO-SK	Cooperative Particle Swarm Optimization – Single K (collaborative model)
CS / CSA	Cuckoo Search (Algorithm)
DE	Differential Evolution
EA	Evolutionary Algorithm
EMS	Evolutionary Multi-agent System
FA	Firefly Algorithm
GA	Genetic Algorithm
GDACS	Grouped Dynamic Adaptive Cuckoo Search
HAGCS	Hierarchical Adaptive Cuckoo Search
HS	Harmony Search
HS-SQP	Harmony Search combined with Sequential Quadratic Programming
IPOPT	Interior Point Optimizer (primal-dual interior-point algorithm)
LOQO	Optimization solver based on an infeasible primal-dual interior-point method
MECA	M-elite Coevolutionary Algorithm
MINOS	Modular In-core Nonlinear Optimization System
MOCSA	Multi-Objective Cuckoo Search Algorithm
OCSA	Optimized Cuckoo Search Algorithm
OEA	Organizational Evolutionary Algorithm
PSO	Particle Swarm Optimization
PSOA	Particle Swarm Optimization Algorithm
PSOAN	Particle Swarm Optimization with Adaptive Neighborhood
PSOSTR	Particle Swarm Optimization with Self-adaptive Technique for Mixed-variable Response
RA	Random/Restart Approach *(context-dependent)*
SA	Simulated Annealing
SA-DS	Simulated Annealing with Downhill Simplex
SBM	Socio-Behavioral Model
SCA	Sine Cosine Algorithm
SMES	Simple Multi-member Evolution Strategy
SQP	Sequential Quadratic Programming
UCOPs	Unconstrained Optimization Problems
YCSA	Yang's Cuckoo Search Algorithm *(CSA attributed to Xin-She Yang)*

1. Introduction

The Cuckoo Search Algorithm (CSA), proposed by Yang and Deb, is a nature-inspired metaheuristic that models the brood parasitism behavior of certain cuckoo species. These species select high-quality host nests to maximize the survival probability of their eggs, a strategy that CSA emulates to efficiently explore the search space and identify optimal solutions.

CSA has emerged as an effective global optimization technique, integrating fundamental concepts from algorithms such as Particle Swarm Optimization (PSO), Differential Evolution (DE), and Simulated Annealing (SA). Its simplicity and ease of implementation, combined with minimal parameter tuning requirements—typically limited to population size and discovery probability—contribute to its robustness across a wide range of applications [1–2].

The algorithm has been successfully applied in numerous domains, including engineering design optimization [3], economic load dispatch [4], unit commitment [5], multi-objective power scheduling [6], optimal power flow [7], manufacturing optimization [8], financial modeling [9], image processing [10], and aerodynamic design [11]. Comparative studies [12] have demonstrated its superior performance in terms of convergence speed, robustness, and solution accuracy relative to several other metaheuristic approaches.

Nevertheless, similar to many population-based optimization methods, CSA encounters limitations when addressing large-scale and high-dimensional problems, primarily due to increased computational cost and slower convergence rates [2], [9]. To mitigate these issues, this study proposes an enhanced variant of CSA that reduces the population size while improving convergence efficiency without compromising solution quality.

Numerous modifications of CSA have been reported in the literature to improve its performance. Lévy flight-based mechanisms were introduced in [4] to strengthen exploration capabilities, while improvements in exploitation have been achieved through guided selection and evaluation strategies [2], [9]. The authors in [13] proposed a hierarchical adaptive cuckoo search algorithm (HAGCS) aimed at enhancing global search capability and adaptability across diverse optimization scenarios. Their approach incorporates distributional entropy-based analysis, weighted fitness–distance balanced ranking, an adaptive global search strategy, and a hierarchical nest replacement scheme to improve convergence and avoid local optima. Adaptive and hybrid variants, including self-adaptive techniques [14], co-evolutionary models [15], and dynamic parameter tuning strategies [16], have also demonstrated improved performance. Additionally, parameter sensitivity analyses have been investigated in [17].

Further advancements include the Grouped Dynamic Adaptive Cuckoo Search (GDACS) introduced in [18], which enhances population diversity during initialization using chaotic maps and replaces the standard update mechanism with Cauchy and Lévy-based operators. The method further employs dynamic population grouping based on fitness levels and applies adaptive step-size mechanisms to each group, thereby achieving a better balance between exploration and exploitation.

In addition, [19] presented a comprehensive survey of multi-objective CSA (MOCSA) variants and their applications. The study reviewed 123 distinct variants reported in the literature, analyzing their theoretical developments, methodological adaptations, and applications across various domains. It further categorizes these variants into original, modified, and hybrid approaches, evaluates their performance characteristics, and outlines potential directions for future research.

Despite these advancements, the standard CSA remains heavily reliant on stochastic processes, which tend to favor exploration at the expense of exploitation. To address this imbalance, the proposed method introduces enhancements to both phases. Specifically, a guided exploration mechanism is employed to direct the search toward promising regions of the solution space, while an improved exploitation strategy generates multiple candidate solutions around elite individuals and selects the most promising ones.

These improvements result in the Optimized Cuckoo Search Algorithm (OCSA), which achieves a more effective balance between exploration and exploitation while preserving the inherent strengths of the original CSA.

The remainder of this paper is organized as follows. Section II reviews the standard CSA. Section III presents the proposed enhancements. Section IV discusses the simulation results obtained from benchmark and real-world problems. Finally, Section V concludes the paper and outlines directions for future work.

2 Cuckoo search algorithm

The CSA, proposed by Yang and Deb in 2009, is a metaheuristic optimization method inspired by the brood parasitism of certain cuckoo species, which lay their eggs in the nests of other birds instead of raising their own offspring.

2.1 Behaviour of Cuckoo Bird

Certain cuckoo species, such as Ani and Guira, lay their eggs in the nests of other birds. If detected, the host may remove the eggs or abandon the nest. To increase survival, some cuckoos mimic the appearance of host eggs and time their laying so their chicks hatch first and dominate the nest [20].

Based on this behavior, the CSA follows three main rules [3]: each cuckoo lays one egg in a randomly chosen nest, the best nests are kept for the next generation, and a fraction of nests is replaced when parasitic eggs are discovered with probability $\boldsymbol{P_a}$. In practice, this is modeled by replacing a portion $\boldsymbol{P_a}$ of the population with new random solutions.

CSA combines local and global random search, controlled by the parameter $\boldsymbol{P_a}$, to balance exploration and exploitation [1].

$$x_i^{t+1} = \boldsymbol{x_i^t} + \alpha\, \boldsymbol{\rho} \otimes H(P_a, \varepsilon) \otimes \left(\boldsymbol{x_j^t} - \boldsymbol{x_k^t}\right) . \quad (1)$$

where x^t_j and x^t_k are two randomly selected solutions, $H(u)$ is the Heaviside function, ε is a uniformly distributed random number, and ρ is the step size. The global search is performed using Lévy flights [1].

$$\boldsymbol{x_i^{t+1}} = \boldsymbol{x_i^t} + \alpha L(s) . \quad (2)$$

Here, α > 0 is the step size scaling factor, usually set relative to the problem's characteristic scale LLL. Typically, α = O(L/10), though smaller values like α = O(L/100) can be more effective to avoid overly large steps [1].

To generate the Lévy flight random walk from a nest, we:

- ✓ Select a random direction sampled from a uniform distribution
- ✓ Generate step sizes using the Mantegna algorithm, which yields:

$$\boldsymbol{s} = \boldsymbol{u} / \left(\boldsymbol{v}^{\frac{1}{\beta}}\right) . \quad (3)$$

While $1 \leq \beta \leq 2$ and $|s| \geq |s_0| >> 0$, and u and v are drawn from normal distribution:

$$u \sim N(0, \sigma_u^2); v \sim N(0, \sigma_v^2)$$

Where

$$\sigma_u = \left[\frac{\Gamma(1+\beta) sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right)\beta 2^{\frac{\beta-1}{2}}}\right]^{1/\beta} ;\ \sigma_v = 1 . \quad (4)$$

2.2 Pseudo Code of the Cuckoo Search Algorithm

Based on the aforementioned rules, the CSA pseudo-code is thoroughly detailed in [4] and can be summarized as follows (Algorithm 1):

Algorithm 1- Pseudo Code of the Cuckoo Search Algorithm.

```
BEGIN

// 1. Objective Definition
DEFINE objective function:
   Minimize f(x), where x ∈ ℝ^d

// 2. Initialization
GENERATE n host nests x_i (i = 1, 2, ..., n)
SET discovery probability P_a ∈ [0, 1]
SET maximum generations MaxGen
SET t = 0

// 3. Initial Fitness Evaluation
EVALUATE fitness F_i = f(x_i) for all nests

// 4. Main Loop
WHILE (t < MaxGen) DO

   // Cuckoo search phase
   FOR each cuckoo i DO
      // a) Lévy flight
      x_i_new = x_i + α ⊕ Lévy(λ)
      // b) Evaluate new solution
      F_i_new = f(x_i_new)

      // c) Random nest selection
      Randomly choose nest j
      // d) Greedy replacement (minimization)
      IF (F_i_new < F_j) THEN
         REPLACE x_j with x_i_new
      END IF
   END FOR

   // Abandonment phase
   ABANDON a fraction P_a of worst nests
   GENERATE new random solutions for abandoned nests

   // Re-evaluation
   EVALUATE fitness of new solutions

   // Update best solution
   RANK all nests by fitness
   IDENTIFY current best solution

   // Increment generation counter
   t = t + 1

END WHILE

// 5. Output
OUTPUT best solution found

END
```

3 Optimized CSA

3.1. Implementation of Enhanced Exploration in CSA

The original CSA, as implemented by Yang, relies on stochastic processes during both the exploration and exploitation phases. This stochastic nature helps prevent the algorithm from becoming trapped in local optima, particularly through the use of Equation (2), which generates new solutions based solely on the current population. However, while effective at maintaining diversity, this strategy can sometimes slow down convergence.

To address this limitation, our modified algorithm retains the core stochastic mechanisms but introduces additional features designed to accelerate convergence while maintaining global search

capability. Drawing inspiration from natural cuckoo behavior, where cuckoos engage in extensive exploratory movements before selecting a secure, resource-rich nesting site, our approach incorporates multiple exploratory phases.

The process begins with an initial exploration using Lévy flights (Figures 1 and 2) to identify promising regions in the search space. This is followed by random exploratory displacements originating from the initial points to further sample nearby regions. Finally, the algorithm performs directed movements toward the current best nest, which include both lateral displacements (left or right of the starting point) and gradient-inspired movements (upstream or downstream relative to the best nest location).

By combining stochastic and quasi-random exploration with guided exploitation, this enhanced strategy improves both the efficiency and reliability of the search process. The subsequent selection process for determining the optimal nesting sites proceeds as follows (Algorithm 2):

Algorithm 2- Optimized Cuckoo Nesting Algorithm with Lévy Flights and Asymmetric Local Moves.

BEGIN

FOR i = 1 TO number_of_cuckoo_nests DO
$x_i^{(1)}$ = Lévy_flight_move(x_i^0)
$x_i^{(2)}$ = downstream_move ($x_i^{(1)}$, best_nest)
$x_i^{(3)}$ = upstream_move($x_i^{(1)}$, best_nest)
$x_i^{(4)}$ = right_move($x_i^{(1)}$)
$x_i^{(5)}$ = left_move($x_i^{(1)}$)

x_i = BEST($x_i^{(1)}$, $x_i^{(2)}$, $x_i^{(3)}$, $x_i^{(4)}$, $x_i^{(5)}$)
END FOR
END

3.2. Implementation of Enhanced Exploitation Phase in CSA

In contrast to the standard CSA assumption described in [5], where each cuckoo lays only one egg, we introduce an intensified exploitation mechanism inspired by biological observations. In nature, certain parasitic cuckoo species lay multiple eggs in host nests, with only the best-adapted offspring surviving. We incorporate this behavior into our algorithm to enhance its exploitation capability.

This modification introduces several key features: a multi-egg deposition strategy, a competitive selection process, and an adaptive survival mechanism. Together, these elements allow the algorithm to generate multiple candidate solutions within each iteration, from which only the most promising are retained. The resulting enhanced exploitation phase operates as follows (Algorithm 3).

Algorithm 3. Competitive Multi-Egg Cuckoo Search Algorithm with Adaptive Survival Strategy.

BEGIN
 	
FOR k = 1 TO number_of_eggs_per_cuckoo DO
 	
// Abandon egg with probability Pa
IF random() ≥ Pa THEN
 	
// Randomly select another nest j
j = random_integer(1, Nnest)
 	
// Generate new solution via differential mutation
	xi'(k) = xi(k) + random() * (xj - xi)
	
	ELSE
	// Keep original egg (implicitly xi'(k) = xi(k))
	END IF
	
	END FOR
	
	// Keep best egg (for minimization)
	xi = ARGMIN{ f(xi'(k)) } // or BEST OF {xi'}
	
	END

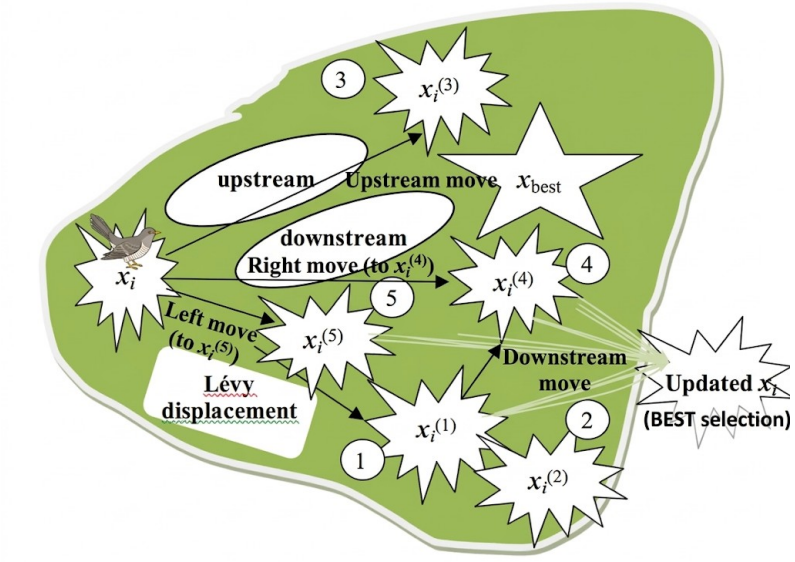


Fig. 1 Exploration the area by the cuckoo bird

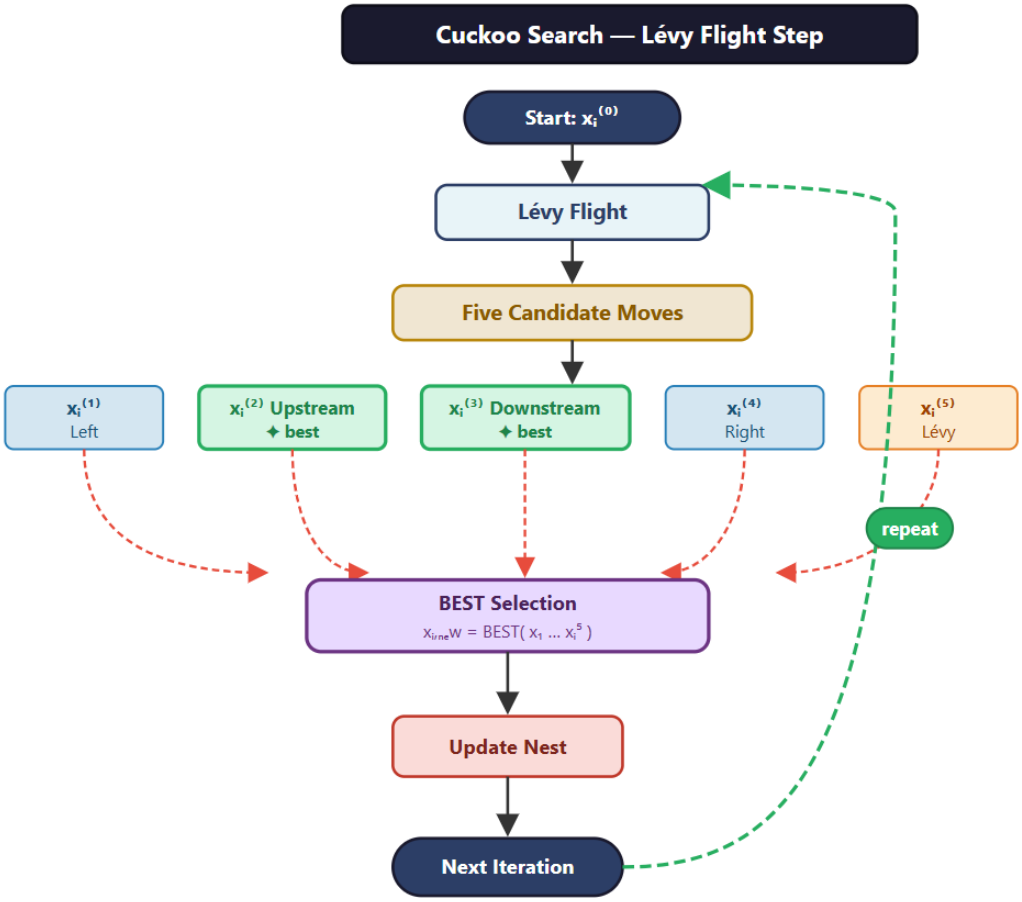


Fig. 2 Flowchart- OCSA Lévy Flight Step

4. Simulation Methodology

In our simulation study, we maintained consistent problem formulations while replicating most of the experimental applications reported by Zheng *et al.* [15]. Our validation approach consisted of two main stages to thoroughly assess the performance of the proposed algorithm.

First, we conducted comprehensive benchmark testing that included both constrained and unconstrained optimization problems drawn from references [22–24]. In addition, we evaluated the algorithm on a set of challenging global optimization problems reported in [25], which further tested its robustness and versatility.

Second, we carried out specific performance evaluations focusing on test function F9, as illustrated in Figure 4. These tests included augmenting the dimensions to large scales to examine the algorithm's scalability and computational efficiency under extreme problem sizes.

4.1 Results on Unconstrained Optimization Problems

In this study, we evaluated the performance of our proposed algorithm using 15 benchmark functions (F1–F15) under consistent test conditions [23]. For functions F1 to F13, the dimensionality was set to 30, while for F14 and F15, it was increased to 100. Each test involved 2,500 function evaluations and was repeated 50 times to ensure statistical significance.

The comparative results between the CCCS and OCSA algorithms, as summarized in Table 1, demonstrate the superior convergence behavior of OCSA toward the exact global optima. Notably, OCSA consistently attained optimal values for both the best and average solutions across most functions. In contrast, CCCS exhibited substantially inferior results in terms of solution quality metrics, indicating a significant performance gap.

Figures 3 and 4 further illustrate this advantage. For example, on function F8, OCSA reached convergence within 500 evaluations, whereas CCCS required approximately 1,000 evaluations. Similarly, for the high-dimensional function F15, OCSA converged before 500 evaluations, while CCCS needed nearly 1,500 evaluations.

Additionally, a comparative analysis involving other algorithms such as MECA and OEA (Table 2) confirms that OCSA maintains superior average solution quality. Across the majority of the test functions, OCSA consistently outperforms these methods, reinforcing its robustness and efficiency in global optimization.

Table 1: Comparison between CCCS and OCSA over 50 trials [15]

Function	fmin	CCCS Best Value	CCCS Mean Value	OCSA Best Value	OCSA Mean Value
F1	0	8.804e-242	2.011e-217	0	0
F2	0	4.271e-143	1.628e-130	0	0
F3	0	6.280e-192	5.464e-154	0	0
F4	0	1.085e-121	1.333e-098	0	0
F5	0	25.0366	25.4182	2.521e-29	1.525e-25
F6	0	0	0	0	0
F7	0	5.309e-05	3.080e-04	0	0
F8	-12569.5	-1.25694870e+04	-1.25694867e+04	-1.25694866e+04	-1.25694866e+04
F9	0	0	0	0	0
F10	0	1.655e-07	3.663e-05	-8.882e-16	-8.882e-16
F11	0	0	0	0	0
F12	0	6.845e-11	9.518e-07	1.571e-32	1.571e-32
F13	0	2.073e-13	1.723e-07	1.350e-32	1.350e-32
F14	-99.60	-96.6008	-83.7328	-96.4322	-94.3259
F15	-78.33236	-78.33233	-78.33232	-78.3323	-78.3323

Table 2: Comparison between MECA, OEA, CCCS, YCSA and OCSA over 50 trials for mean function values [15]

Function	**fmin**	**CCCS**	**MECA**	**OEA**	**OCSA**
F1	0	2.011e-217	4.228e-183	2.481e-30	0
F2	0	1.628e-130	1.845e-110	2.068e-13	0
F3	0	5.464e-154	3.274e-95	1.883e-09	0
F4	0	1.333e-098	5.124e-02	8.821e-02	0
F5	0	25.4182	7.973e-02	0.227	1.525e-25
F6	0	0	0	0	0
F7	0	3.080e-04	4.084e-04	3.297e-03	0
F8	-12569.5	-1.25694867e+04	-1.25694866e+04	-1.25694866e+04	-1.25694866e+04
F9	0	0	0	5.430e-17	0
F10	0	3.663e-05	0	5.336e-14	-8.882e-16
F11	0	0	3.844e-03	1.317e-02	0
F12	0	9.518e-07	1.571e-32	9.207e-30	1.571e-32
F13	0	1.723e-07	1.350e-32	4.323e-19	1.350e-32
F14	-99.60	-83.7328	-98.7095	-99.5024	-94.32596
F15	-78.33236	-78.33232	-78.33233	-78.33233	-78.33232

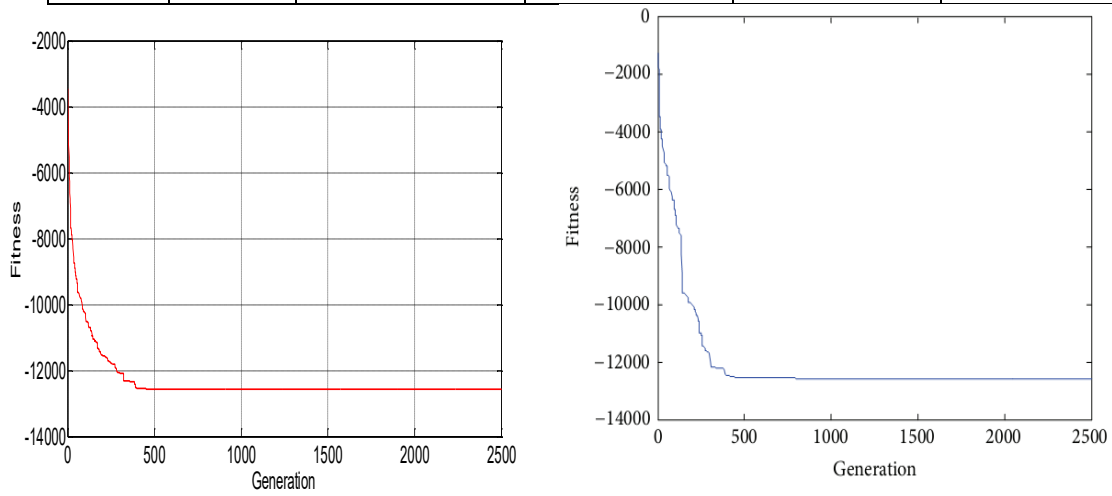


Figure 3: Convergence curves of our application (figure on the left) and of CCCS of F8

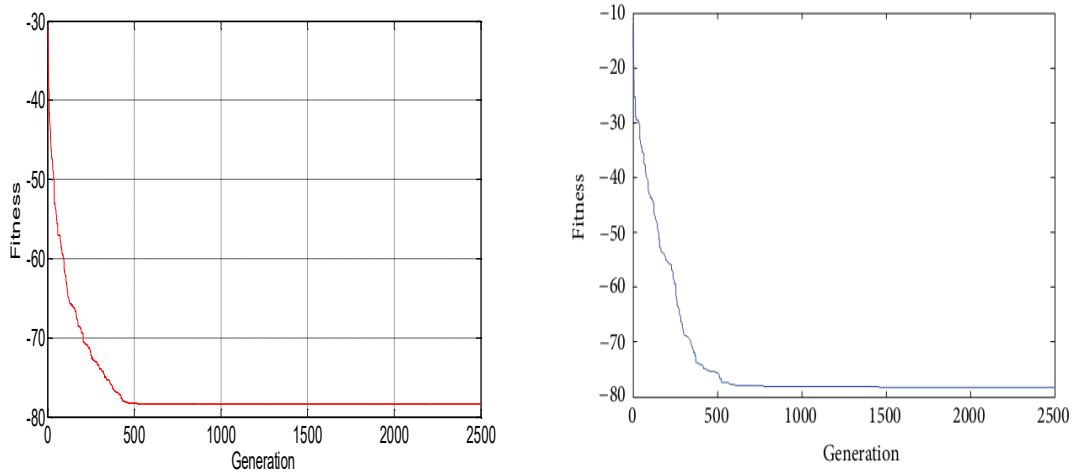


Figure 4: Convergence curves of our application (figure on the left) and of CCCS of F15

4.2 Results on Constrained Optimization Problem

This section evaluates the performance of the proposed algorithm on a diverse set of constrained optimization challenges. The test suite includes nine benchmark problems (G01 to G11), excluding G02 and G03 due to ambiguous or inconsistent results reported in the literature. Additionally, two engineering design problems from [15] are considered to assess practical applicability. Finally, an economic load dispatch problem that incorporates valve-point effects is included to demonstrate the algorithm's capability in handling complex real-world scenarios.

4.2.1 Benchmark Function Analysis

In accordance with the CCCS methodology, we conducted 2,500 function evaluations per test. It is worth noting that the comparative algorithm OEA utilized 24,000 evaluations per test. Table 3 summarizes the statistical results obtained from 50 trials for each function, although our implementation employed fewer trials in certain cases.

The key findings indicate that our algorithm consistently reached the global optimum for the majority of the tested functions. Furthermore, it demonstrated superior performance when compared to several benchmark methods, including CCCS, OEA CHp, OEA CHc, and EMS. Overall, these results clearly highlight the competitive advantage of our method in solving constrained optimization problems efficiently and reliably.

Table 3: Comparison between OEA, SMES, CCCS best function values and OCSA function values [15]

Function	fmin	CCCS	OEA CHp	OEA CHc	SMES	OCSA
G01	15	15	15	15	15	15
G04	-30665.539	-3.066554e+04	-30665.539	-30665.539	-30665.539	-30665.52411
G05	5126.498	5.126498e+03	5126.497	5126.532	5126.599	5126.498077
G06	-6961.814	-6.961814e+03	-6961.814	-6961.814	-6961.814	-6961.813875
G07	24.306	24.30621	24.308	24.307	24.327	24.306206
G08	-0.095825	-9.5825e-02	-0.095825	-0.095825	-0.095825	-0.105460
G09	680.630	680.630	680.630	680.630	680.632	680.630057
G10	7049.331	7.0492e+03	7052.236	7100.030	7051.903	7049.239
G11	0.750	0.7500	0.750	0.750	0.750	0.749950

4.2.2 Welded Beam Design Case Study

This section provides a comparative analysis of different optimization methods reported in the literature for the welded beam design problem. The comparison focuses on two main aspects: performance metrics across various evaluation budgets and the identification of solutions that violate problem constraints, which are marked as unreliable.

As shown in Table 4, our proposed method achieves the best-known optimal solution for the problem. Additionally, it demonstrates superior convergence efficiency by requiring the fewest iterations among all feasible solutions. These results highlight the effectiveness and reliability of our approach compared to existing methods.

Table 4: Welded beam problem: comparison of OCSA results with the literature [14].

Researcher / Study	Method	m	H	l	t	b	Cost / Objective	No. of Eval.	Constraint Violation
Coello	GA	0.2088	3.4205	8.9975	0.2100	1.7483	N.A.	N.A.	Yes
Leite & Topping	GA	0.2489	6.1097	8.2484	0.2485	2.4000	6273	No	No
Deb	GA	0.2489	6.1730	8.1789	0.2533	2.4331	320,080	No	No
Lemonge & Barbosa	GA	0.2443	6.2117	8.3015	0.2443	2.3816	320,000	No	No
Bernardino et al.	AISa-GA	0.2444	6.2183	8.2912	0.2444	2.3812	320,000	No	No
Atiqullah & Rao	SAb	0.2471	6.1451	8.2721	0.2495	2.4148	N.A.	No	No
Hedar & Fukushima	SA	0.2444	6.2175	8.2915	0.2444	2.3810	N.A.	No	No
Liu	SA-GA	0.2231	1.5815	12.8468	0.2245	2.2500	26,466	Yes	Yes
Hwang & He	SA-DSc	0.2444	6.2158	8.2939	0.2444	2.3811	56,243	No	No
Parsopoulos & Vrahatis	PSO	N.A.	N.A.	N.A.	N.A.	1.9220	100,000	-	-
He et al.	PSO	0.2444	6.2175	8.2915	0.2444	2.3810	30,000	No	No
Zhang et al.	EAd	0.2443	6.2201	8.2940	0.2444	2.3816	28,897	No	No
Coello	EA	N.A.	N.A.	N.A.	N.A.	1.8245	N.A.	??	??
Lee & Geem	HSe	0.2442	6.2231	8.2915	0.2443	2.3810	110,000	Yes	Yes
Mahdavi et al.	HS	0.2057	3.4705	9.0366	0.2057	1.7248	200,000	Yes	Yes
Fesanghary et al.	HS-SQP	0.2057	3.4706	9.0368	0.2057	1.7248	90,000	Yes	Yes
Siddall	RAg	0.2444	6.2819	8.2915	0.2444	2.3815	N.A.	No	No
Akhtar et al.	SBMh	0.2407	6.4851	8.2399	0.2497	2.4426	19,259	No	No
Ray & Liew	SCAi	0.2444	6.2380	8.2886	0.2446	2.3854	33,095	No	No
Montes & Oca	BFOj	0.2536	7.1410	7.1044	0.2536	2.3398	N.A.	Yes	Yes
Zhang et al.	DEk	0.2444	6.2175	8.2915	0.2444	2.3810	24,000	No	No
Gandomi et al.	FA	0.2015	3.5620	9.0414	0.2057	1.7312	50,000	Yes	Yes
Hongqing Zheng	CCCS	0.3312	10.000	2.4095	0.3456	2.1730	25,000	Yes	Yes
Present study	OCSA	0.20573	7.0924	9.03663	0.20573	2.21815	35 (0.9 sec)	No	No

4.2.3 Case of pressure design

Table 5, which presents the results obtained for the pressure design problem, also demonstrates the superiority of our algorithm. It achieves the best optimal solution with fewer function evaluations compared to various solutions reported in the literature.

Table 5: Pressure vessel design: comparison of OCSA results with the literature [15].

Method	R	L	Ts	Th	Cost	Iter. Max / Time
CCCS	40.3196	200.0000	0.7782	0.3846	5885.3	1000
ALPSO	41.3500	200.0000	0.7980	0.3950	6234.0	7590
PSOA	N.A.	N.A.	N.A.	N.A.	6292.0	6506
PSOSTR	N.A.	N.A.	N.A.	N.A.	6272.0	3723
He et al.	N.A.	N.A.	N.A.	N.A.	6290.0	30000
Akhtar et al.	N.A.	N.A.	N.A.	N.A.	6335.0	12630
OCSA	40.31962	200.0000	0.77817	0.38304	5880.67	55 (0.7 sec)

4.2.4 Case of economic dispatch problem with valve point effect

In this subsection, the economic dispatch problem with three generating units was re-examined, incorporating the valve-point loading effect of steam admission [22]. Consequently, the cost function exhibits a non-smooth, rippling shape, as illustrated in Figure 6 for the case of two generating units.

Table 6 summarizes the results, including the errors associated with the nonlinear equality constraint for each optimization method. The proposed algorithm demonstrated superior performance by achieving the lowest total cost, followed by the Artificial Bee Colony (ABC) algorithm, the Particle Swarm Optimization (PSO) method, the Lambda Iteration method, and the Genetic Algorithm (GA). In contrast, the Binary Bat Algorithm (BBT) produced an inaccurate solution, with a deviation of 20 MW in the power balance equation.

Table 6: Economic dispatch with valve point effect results [25].

Load demand	Parameter	Lambda	GA	PSO	ABC	BBT	**OCSA**
850 MW	P1 MW	382.258	382.2552	394.5243	300.266	296.3495	**300.265979611306**
	P2 MW	127.419	127.4184	200.000	149.733	199.6002	**149.733100113962**
	P3 MW	340.323	340.3202	255.4756	400.000	334.0503	**399.999999999999**
	Error	0	6.2e-003	1e-005	1e-003	-20	**9.2e-004**
Total cost Rs/h		8575.68	8575.64	8280.81	8253.10	8121.8568	**8253.08824756006**

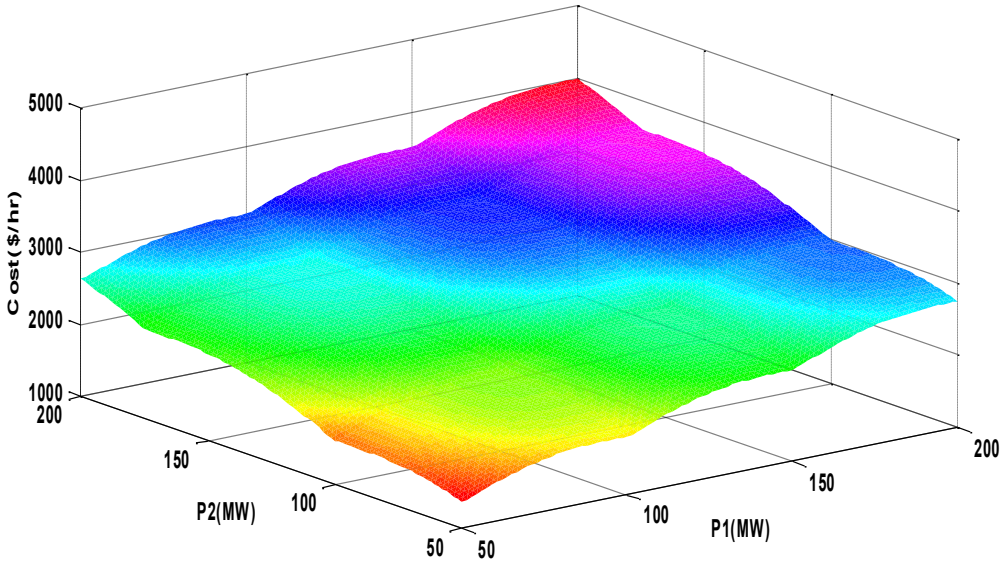


Figure 5: Illustration cost function for two production units

4.3 Simulation of difficult global optimization problems

In this section, for comparison with the optimization solvers Ipopt, LOQO, MINOS, PGAPack, PSwarm, and Couenne, we conducted the same set of simulations as Vanaret *et al.* [26]. In their study, a hybrid algorithm combining an Evolutionary Algorithm with interval-based methods (numerically certified using Charibde) was applied to a set of challenging optimization problems, including Michalewicz, Sine Envelope, Shekel, Egg Holder, Rana, and Keane functions. However, the Shekel function was excluded from our tests due to ambiguity in its definition, as multiple variants of the Shekel function exist and the specific version used was not clearly identified by the authors.

The simulation results are summarized in Table 7, presenting the obtained minimum values, along with the number of function evaluations and/or computation time when available. Based on these results, it is evident that our proposed OCSA algorithm outperforms several of the compared solvers, both in terms of global optimality and computational efficiency.

Table 7: State-of-the-art solvers against OCSA on test functions [26]

Method	Metric	Michalewicz (d=50)	Sine Envelope (d=5)	Egg Holder (d=5)	Rana (d=5)	Keane (d=4)
Ipopt	Value	-19.773742	-5.8351843	-3586.3131827	-75.512076	-0.2010427
	Eval	167	24	8	16	8
LOQO	Value	-0.0048572	-5.8351843	25.5609238	-69.5206	-0.0983083
	Eval	88	17	5924	138	50
MINOS	Value	0	-5.87878	-3586.313183	-233.592	-0.2347459
	Eval	3	38	3	1	3

PGAPack	Value	-37.60465	-5.569544	-3010.073	-2091.068	—
	Eval	9582	9615	9626	9622	—
PSwarm	Value	-24.38158	-5.835182	-2840.799	-1595.056	—
	Eval	2035	2049	2040	2046	—
Couenne	Value	-49.619042	-5.9660007	-3719.7287498	-2046.8320657	-0.6222999
	Time	265s	0.4s	20.6s	20.3s	2s
Charibde	Value	-49.6248323	-5.9659811	-3719.7248363	-2046.8320657	-0.6222810
	Time	8.4s	219s	0.8s	17.8s	14860s
	Eval	717100	227911400	742600	19186150	6351693297
OCSA	Value	-49.6248323040744	-5.966048150900620	-3719.724835658383	-2046.832065725112	-0.6222810255678127
	Time	223.3s	0.8s	13.1s	16.4s	12.9s
	Eval	8872	108	1702	2000	1000

4.4 Convergence Analysis

To further assess the convergence capabilities of the proposed Optimized Cuckoo Search Algorithm (OCSA), we conducted a series of rigorous experiments using the benchmark function F9, as illustrated in Figure 5. These tests were intentionally designed under extreme conditions to challenge the algorithm's robustness and efficiency. Specifically, we imposed a highly stringent termination criterion, requiring the relative error to fall below $\varepsilon < 10^{-400}$ with respect to the known global optimum. Additionally, we progressively increased the problem dimensionality, ranging from 10^1 to 10^6, to evaluate scalability. For comparison, we used the YCSA algorithm as a baseline, which has been reported to outperform established methods such as Particle Swarm Optimization (PSO) and Simulated Annealing (SA) [1, 3].

The results, summarized in Table 8, reveal several key insights. First, OCSA consistently outperforms YCSA across all tested dimensions, demonstrating superior convergence behavior and reliability in reaching the global optimum. Notably, while YCSA is recognized for its high performance, OCSA achieves even better results, indicating its robustness and enhanced optimization capacity. Furthermore, the algorithm shows remarkable scalability, maintaining its performance advantage even as the dimensionality increases to extremely high levels.

Convergence metrics, presented in Figure 7, further reinforce these findings. OCSA successfully reaches the global minimum with a final function value of fmin=0. Even at the highest tested dimensionality ($d=10^6$), the algorithm maintains efficient performance, with a total computation time of approximately 130 seconds. This result underscores the algorithm's practicality and effectiveness for solving large-scale, complex optimization problems.

Table 8: Comparison between YCSA [3] and OCSA of variation dimension of F9

Fun.	Meth.	d	10	10^2	10^3	10^4	10^5	10^6
F9	OCSA	*t*(sec)	0.031809	0.040294	0.103599	0.656107	13.066879	134.508003
		f_{min}	0	0	0	0	0	0
F9	YCSA	*t*(sec)	1.640971	*failed*	*failed*	*failed*	*failed*	*failed*
		f_{min}	0	*failed*	*failed*	*failed*	*failed*	*failed*
F9	SA	*t*(sec)	*failed*	*failed*	*failed*	*failed*	*failed*	*failed*
		f_{min}	*failed*	*failed*	*failed*	*failed*	*failed*	*failed*

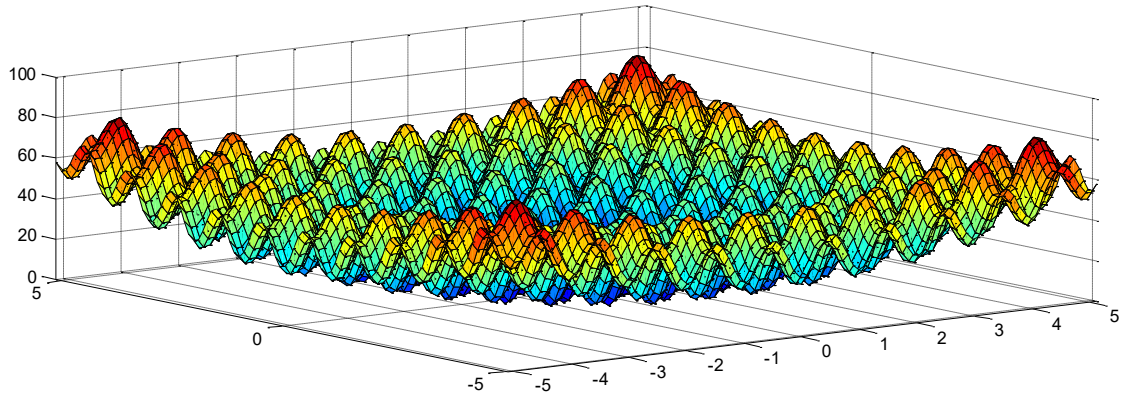


Figure 6: The landscape of F9's 2D function

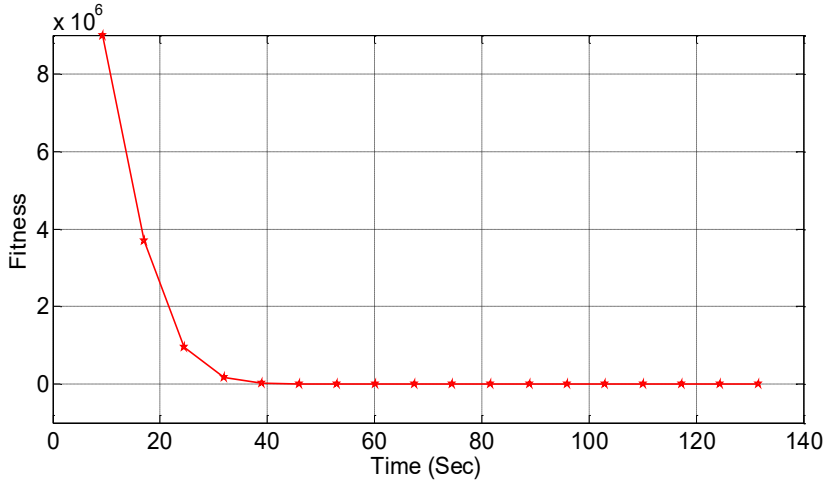


Figure 7: Convergence curve for the dimension d=10^6 of F9

5. Conclusion

This study presents an enhanced version of the Cuckoo Search Algorithm (CSA) with several improvements to address common limitations in metaheuristic optimization. First, a hybrid exploration strategy combines random and quasi-random search patterns to improve coverage of the solution space and maintain diversity among candidate solutions. Second, the algorithm strengthens exploitation while preserving global search capability. Carefully designed refinement phases enhance promising solutions without causing premature convergence. Additionally, the proposed method improves

computational efficiency through faster convergence, lower memory usage, and better performance on large-scale problems, making it suitable for high-dimensional applications.

The main contribution is the bidirectional integration of exploration and exploitation. Exploitation mechanisms operate during exploration, while exploration continues during exploitation, ensuring a balanced and adaptive search process. As a result, the algorithm effectively targets high-quality solutions, avoids suboptimal regions, prevents local optima stagnation, and maintains consistent performance across challenging scenarios. Comparative evaluations demonstrate its superiority over conventional algorithms and specialized solvers, with improved speed, accuracy, constraint handling, and strong potential for integration into other population-based metaheuristic frameworks.

Competing interests

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

Funding

Not applicable.

References

[1] Yang XS, *Cuckoo Search and Firefly Algorithm: Theory and Applications*, Studies in Computational Intelligence, vol. 516, Springer, Warsaw, Poland, 2014.

[2] Tuba M, Subotic M, Stanarevic N, Modified cuckoo search algorithm for unconstrained optimization problems, in: Proc. 5th European Conf. on European Computing Conference (ECC'11), 2011, pp. 263–268.

[3] Yang XS, Deb S, Engineering optimization by cuckoo search, *Int. J. Math. Model. Numer. Optim.*, 1 (2010) 330–343.

[4] Bindu AH, Reddy MD, Economic load dispatch using cuckoo search algorithm, *Int. J. Eng. Res. Appl.*, 3 (2013) 498–502.

[5] Alireza G, Rozbeh J, A new approach for solving the unit commitment problem by cuckoo search algorithm, *Indian J. Sci. Technol.*, 6 (2013) 5235–5242.

[6] Chandrasekaran K, Simon SP, Multi-objective scheduling problem: Hybrid approach using fuzzy assisted cuckoo search algorithm, *Swarm Evol. Comput.*, 5 (2012) 1–16.

[7] Rao MR Mohana, Naresh Babu AV, Optimal power flow using cuckoo optimization algorithm, *Int. J. Adv. Res. Electr. Electron. Instrum. Eng.*, 2 (2013) 4213–4218.

[8] Yildiz AR, Cuckoo search algorithm for the selection of optimal machining parameters in milling operations, *Int. J. Adv. Manuf. Technol.*, 64 (2013) 55–61.

[9] Tawfik AS, Badr AA, Abdel-Rahman IF, One rank cuckoo search algorithm with application to algorithmic trading systems optimization, *Int. J. Comput. Appl.*, 64 (2013) 30–37.

[10] Agrawal S, Panda R, Bhuyan S, Panigrahi BK, Tsallis entropy based optimal multilevel thresholding using cuckoo search algorithm, Swarm Evol. Comput., 11 (2013) 16–30.

[11] Naumann DS, Evans B, Walton S, Hassan O, A novel implementation of computational aerodynamic shape optimisation using modified cuckoo search, *Appl. Math. Model.*, 40 (2016) 4543–4559.

[12] Farahani SM, Abshouri AA, Nasiri B, Meybodi MR, A Gaussian firefly algorithm, *Int. J. Mach. Learn. Comput.*, 1 (2011) 448–454.

[13] Ouyang, C., Liu, X., Zhu, D. et al. Hierarchical adaptive cuckoo search algorithm for global optimization. Cluster Comput 28, 321 (2025). https://doi.org/10.1007/s10586-024-04924-3.

[14] Mlakar U, Fister I Jr, Fister I, Hybrid self-adaptive cuckoo search for global optimization, *Swarm Evol. Comput.*, 11 (2016) 1–26.

[15] Zheng H, Zhou Y, A cooperative coevolutionary cuckoo search algorithm for optimization problem, *J. Appl. Math.*, 2013 (2013), Article ID 912056, 9 pages.

[16] Valian E, Mohanna S, Tavakoli S, Improved cuckoo search algorithm for global optimization, *Int. J. Commun. Inf. Technol.*, 1 (2011) 31–44.

[17] Tuba M, Cuckoo search optimization metaheuristic adjustment, in: *Recent Advances in Knowledge Engineering and Systems Science*, 2013, pp. 144–149.

[18] Shao-Qiang Ye, Azlan Mohd Zain, Yusliza Yusoff, Improved Cuckoo Search Algorithm for Engineering Optimization Problems, Computers, Materials and Continua, 87, 1, 2026, https://doi.org/10.32604/cmc.2025.073411.

[19] Makhadmeh, S.N., Awadallah, M.A., Kassaymeh, S. et al. Recent advances in Multi-objective Cuckoo Search Algorithm, its variants and applications. Arch Computat Methods Eng 32, 3213–3240 (2025). https://doi.org/10.1007/s11831-025-10240-9.

[20] Payne RB, Sorenson MD, Klitz K, *The Cuckoos*, Oxford University Press, Oxford, 2005.

[21] Yang XS, *Nature-Inspired Metaheuristic Algorithms*, 2nd ed., Luniver Press, UK, 2010.

[22] Yao X, Liu Y, Lin GM, Evolutionary programming made faster, *IEEE Trans. Evol. Comput.*, 3 (1999) 82–102.

[23] Leung YW, Wang YP, An orthogonal genetic algorithm with quantization for global numerical optimization, *IEEE Trans. Evol. Comput.*, 5 (2001) 41–53.

[24] Runarsson TP, Yao X, Stochastic ranking for constrained evolutionary optimization, *IEEE Trans. Evol. Comput.*, 4 (2000) 284–294.

[25] Mallikarjuna B, Reddy KH, Hemakesavulu O, Economic load dispatch problem with valve-point effect using a binary bat algorithm, *ACEEE Int. J. Electr. Power Eng.*, 4 (2013) 33–38.

[26] Vanaret C, Gottel JB, Durand N, Alliot JM, Certified global minima for a benchmark of difficult optimization problems, *HAL Id: hal-00996713*, 2014. Available: https://hal-enac.archives-ouvertes.fr/hal-00996713

Islamic University Journal of Applied Sciences (IUJAS)
https://journals.iu.edu.sa/jesc

Volume VIII, Issue I, July 2026, Pages 249-268

Silicon and Silicon Carbide Technologies for Nuclear Medicine: Performance, AI-Augmented Imaging, and Future Prospects

Olfat. S. Ahmed[1] and Belqees Hassan[2,*]

[1]Department of Physics, College of Science, Qassim University, Buraydah 51452, Saudi Arabia

[2]Department of Physics, College of Science, Ibb University, Ibb, Yemen

*Corresponding author: (Belqees Hassan), *Email Address*: Belqees.ahmad@yahoo.com

Abstract

Silicon-based technologies have become central to modern nuclear medicine, owing to their superior spatial resolution, mature microfabrication infrastructure, compact form factor, and seamless compatibility with advanced readout electronics. In parallel, silicon carbide (SiC) has emerged as a promising wide-bandgap semiconductor for radiation detection and dosimetry in demanding environments, particularly where radiation hardness, thermal stability, and high-temperature operability are required. This study examines the physical basis of silicon detectors, their operation principles, and their applications in nuclear medicine imaging, dosimetry, and medical electronics, with emphasis on positron emission tomography (PET), single-photon emission computed tomography (SPECT), and hybrid imaging systems. The article also compares silicon and SiC in terms of bandgap, charge transport, radiation tolerance, breakdown strength, energy resolution, and fabrication maturity. In addition, the role of silicon technologies in the development of digital medicine and artificial intelligence-assisted imaging is discussed, along with the current progress of nuclear medicine infrastructure and initiatives in Saudi Arabia. The study concludes that silicon remains a mature and highly effective detector material for high-resolution nuclear medicine applications, while SiC offers a strong future pathway for next-generation dosimetry and radiation-hard detector systems.

Keywords: Silicon detectors; Nuclear medicine; Radiation dosimetry; Silicon carbide detectors.

https://doi.org/10.63070/jesc.2026.026
Received 03 March 2026; Revised 14 April 2026; Accepted 22 April 2026.
Available online 30 April 2026.
Published by Islamic University of Madinah on behalf of *Islamic University Journal of Applied Sciences.*
This is an open access article published under the Creative Commons Attribution 4.0 International license (CC BY 4.0):https://creativecommons.org/licenses/by/4.0/

1. Introduction

Silicon (Si, atomic number 14) is the second most abundant element in the Earth's crust and the seventh most plentiful in the universe, occurring naturally as silicon dioxide (SiO_2) in sand, silicate rocks, and clay minerals rather than in elemental form [1,2]. It was first isolated in 1824 by the Swedish chemist Jöns Jacob Berzelius, who produced it by reducing potassium fluorosilicate with potassium metal and carefully washing the resulting product [3]. Industrial production today involves heating sand (SiO_2) with carbon at temperatures approaching 2200°C, followed by further purification to yield electronic-grade material [1,5]. At room temperature, silicon exists in two allotropic forms: crystalline silicon, which exhibits a grayish hue with a metallic sheen and a diamond cubic crystal structure, and amorphous silicon, which appears as a brown powder. High-purity single crystals of the crystalline form are grown by the Czochralski technique, producing the wafers that underpin virtually all modern solid-state devices. When doped with trivalent elements such as boron or gallium, or with pentavalent elements such as phosphorus or arsenic, silicon acquires the controlled electrical conductivity required for transistors, solar cells, rectifiers, and microchips [1,5]. Beyond microelectronics, silicone-based materials - organosilicon polymers derived from silicon- have been used in biomedical applications for over sixty years, having entered commercial markets in 1946 [3,4]. Their combination of hydrophobicity, chemical and thermal stability, low surface tension, inherent biocompatibility, and bio-durability makes them suitable across a wide range of medical device, pharmaceutical, and personal-care applications, in forms ranging from gels, fluids, and adhesives to elastomers for lubricating, bonding, and damping [4]. In the context of nuclear medicine and radiation physics, silicon's most consequential property is its capacity to generate electron–hole pairs when traversed by ionising radiation, requiring only approximately 3.6 eV per pair - far less than the ~15 eV needed in gas-filled detectors and orders of magnitude less than scintillator systems -which translates directly into superior energy resolution and signal yield [1,6]. Combined with fast carrier mobility, compatibility with CMOS microfabrication, and the availability of ultra-pure crystals, silicon has become the material of choice for radiation detectors, medical imaging electronics, and dosimetry devices in modern nuclear medicine [7,8]. This study examines these roles systematically, from fundamental material properties through clinical applications to the evolving landscape of silicon-based nuclear medicine in the Kingdom of Saudi Arabia. The present work offers three original contributions: a systematic side-by-side comparison of Si and SiC detector characteristics grounded in published experimental data; a hypothetical quantitative simulation of detector signal yield and energy resolution under controlled conditions; and an integrated assessment of silicon-based nuclear medicine technologies within the context of Saudi Arabia's Vision 2030 national health strategy.

2. Physical and Radiological Properties of Silicon

Silicon (atomic number Z = 14, atomic mass ≈ 28.085 u) is a group-IV metalloid with a face-centered cubic (diamond cubic) crystal structure. The four valence electrons form covalent bonds with four neighboring atoms, giving silicon its characteristic semiconductor behavior [5]. As illustrated in Figure 1, the three categories of solids- metals, semiconductors, and insulators are distinguished by their electronic band structures: metals exhibit overlapping valence and conduction bands allowing free electron flow, whereas insulators possess a wide bandgap (E_g) that prevents thermal excitation of electrons under normal conditions. Silicon occupies an intermediate position as a semiconductor, with a moderate bandgap of 1.12 eV at 300 K. At absolute zero, all electrons are tightly bound; as temperature rises, thermal excitation promotes electrons across this bandgap, creating mobile electron–hole pairs (shown in the lower panels of Figure 1 as the symmetric n and p carrier distributions about

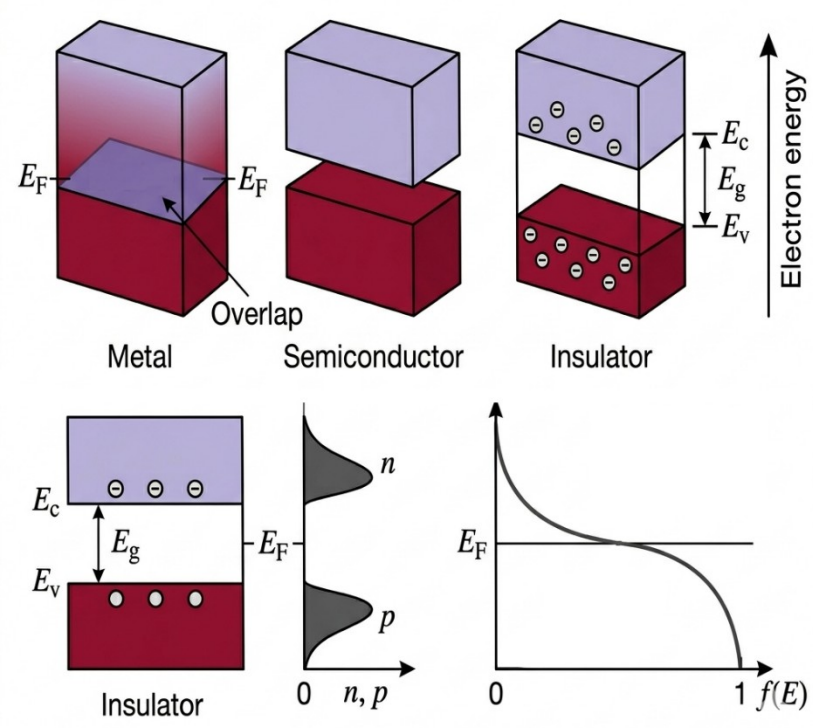


Figure 1: Band structure classification of solids (metal, semiconductor, and insulator), illustrating silicon's intermediate position as a semiconductor with a bandgap of 1.12 eV at 300 K

E_F) that underlie The Fermi–Dirac distribution f(E) confirms that at finite temperatures, a non-zero probability exists for electrons to occupy states above E_c, enabling conduction. These intrinsic material properties acquire particular clinical significance in nuclear medicine, where the sensitivity of PET and SPECT systems depends critically on the charge yield and timing characteristics of their constituent detectors. Silicon detectors have become a cornerstone of nuclear medicine -particularly in high-resolution Positron Emission Tomography (PET) and Single Photon Emission Computed

Tomography (SPECT), due to their exceptional signal characteristics and advanced fabrication potential [9]. These intrinsic material properties translate directly into measurable detector advantages; Table 1 summarises the three core physical properties of silicon, their technical specifications, and their clinical relevance in PET imaging.

Table 1: Core Physical Advantages of Silicon Detectors and Their Clinical Relevance in PET Imaging

Property	Key details	Clinical impact	Ref.
Detection efficiency Low ionisation energy & high resistivity	~3.6 eV per e-h pair (vs ~34 eV for air ionisation chambers; 400–1000 eV for scintillators); wide depletion zone minimises dark current	Superior energy resolution; high signal-to-noise ratio essential for low-activity medical imaging	[6,28]
Speed & material quality High carrier mobility & crystal purity	Electron mobility ~1450 $cm^2/V \cdot s$; Czochralski and float-zone growth yield ultra-low impurity concentrations and improved radiation hardness	Fast coincidence timing for PET; stable long-term performance in clinical radiation environments	[6,10]
Spatial resolution & integration CMOS compatibility and strip/pixel segmentation	Standard wafer processing enables on-chip ASIC readout; strip pitch 50–200 μm; position resolution below 10 μm achievable in laboratory conditions	Sub-millimetre resolution in clinical PET; far exceeds the 1–2 mm limit of conventional scintillator-based systems	[7,11]

Achieving the material purity required to exploit these properties demands extensive processing of naturally occurring silicon sources. Silicon does not occur in elemental form in nature; it is found principally as silicon dioxide (SiO_2) and in complex silicate minerals. Representative concentrations in geological materials are: volcanic rocks (281,500 ppm), sandstone (368,000 ppm), shale (73,000 ppm), and limestone (24,000 ppm). In aquatic environments, silicon concentrations are 6.5 ppm in fresh water and 3 ppm in seawater. Biological systems also incorporate silicon: land plants accumulate 120–6,000 ppm, and certain marine organisms up to 70,000 ppm [2]. Industrial production of metallurgical-grade silicon involves carbothermic reduction of quartz (SiO_2) with carbon at approximately 2200 °C. Electronic-grade silicon requires further purification via the Siemens process or fluidized-bed reactor deposition of trichlorosilane, achieving purity levels exceeding 99.9999999% (nine nines). Single-crystal boules are then grown by the Czochralski technique and sliced into wafers that form the substrates for detector and integrated circuit fabrication [1,5].

3. Applications of Silicon in Nuclear Medicine

3.1 Silicon Radiation Detectors: Operating Principles

Owing to its high purity, fast radiation response, and well-established fabrication technology, silicon is one of the most widely used semiconductor materials in nuclear radiation detectors [29,31]. Silicon radiation detectors operate on the principle of reverse-biased p-n junctions. Under an applied reverse bias voltage, a depletion region devoid of mobile charge carriers forms across the junction, creating a sensitive detection volume. When ionizing radiation (gamma rays, X-rays, or charged particles) traverses this depletion region, it generates electron–hole pairs proportional to the deposited energy. The built-in electric field separates these carriers, driving electrons toward the n-type contact and holes toward the p-type contact, producing a measurable current pulse at the output [6,7]. As illustrated in Figure 2, a silicon strip detector comprises p^+-type implant strips separated by SiO_2 passivation on a bulk n-type substrate (~300 µm thick), with a backplane n^+-type contact connected to a positive bias voltage; the generated electrons and holes drift in opposite directions under the electric field E and are collected at their respective electrodes, with the resulting signal amplified by an external preamplifier

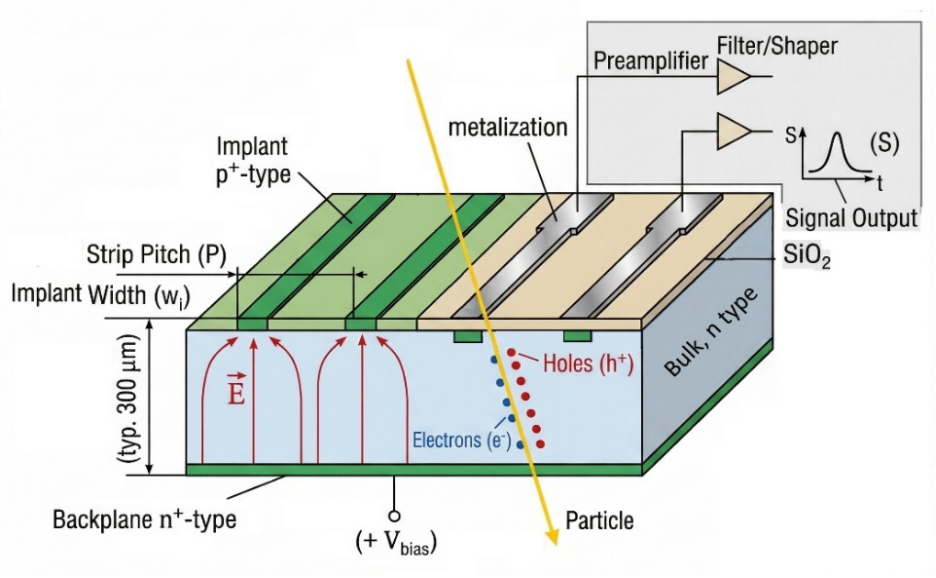


Figure 2: Principles of operation of a silicon strip detector

and filter/shaper chain. The device architecture, shown in Figure 3, consists of a p^+ top layer, a lightly doped n^- substrate of thickness D, and an n^+ bottom contact, all connected to an external amplifier under reverse bias, confirming the same charge-collection mechanism across detector geometries [31,32]. Common detector configurations include p-n, n-p, and n-in-p geometries. The n-in-p configuration is particularly favored for high-radiation applications, as electrons dominate charge collection and are less susceptible to trapping caused by radiation-induced bulk defects [14,15]. Silicon strip detectors segment the junction into parallel strips with a typical pitch of 50–200 µm, providing one-dimensional position information, while double-sided silicon strip detectors (DSSDs) add

orthogonal strips on both faces for 2D position resolution below 10 μm. Silicon photomultipliers (SiPMs) represent a further advancement: arrays of avalanche photodiode microcells operating in Geiger mode, offering single-photon sensitivity, compactness, and immunity to magnetic fields — properties that conventional gas-based photomultiplier tubes cannot match [16,17].

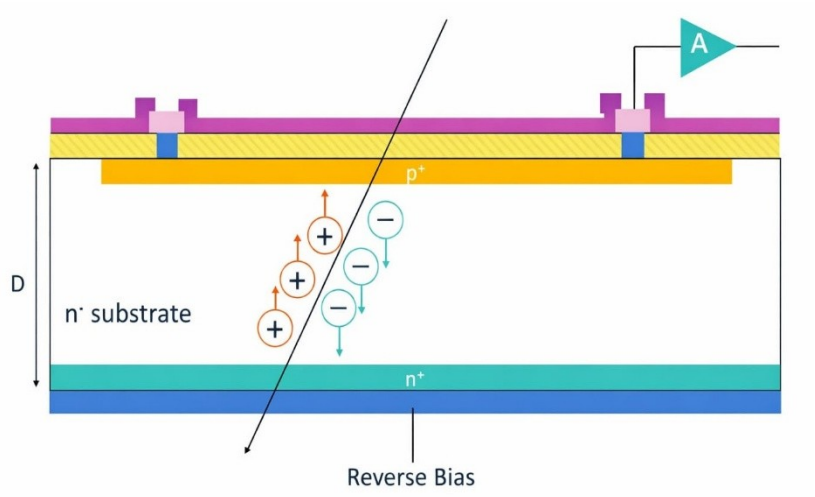


Figure 3: Silicon p^+-n^--n^+ radiation detector under reverse bias showing electron-hole pair generation and charge drift to the external amplifier.

3.2 Medical Imaging: PET and SPECT

3.2.1 Positron Emission Tomography (PET)

PET is a non-invasive functional imaging modality based on the coincidence detection of 511-keV annihilation photons produced when positrons emitted by a radiotracer annihilate with electrons in tissue [18]. The most widely used radiotracer is ^{18}F-fluorodeoxyglucose (FDG), which accumulates preferentially in tissues with elevated glucose metabolism, enabling detection of malignant tumors, neurological disorders, and cardiac dysfunction [18,19]. As illustrated in Figure 4, a PET ring is lined with silicon-based SiPM detectors that record the two 511-keV annihilation photons emitted simultaneously in opposite directions. In modern PET scanners, silicon-based SiPMs have largely replaced traditional photomultiplier tubes as the photodetection elements coupled to scintillating crystals such as LYSO or BGO, offering superior timing resolution critical for time-of-flight PET (ToF-PET), a compact form factor enabling MRI-compatible hybrid systems, and stability across clinical temperature ranges [25]. Silicon microelectronics additionally perform coincidence logic, time stamping, list-mode data acquisition, and iterative image reconstruction central to PET operation [9,26,27].

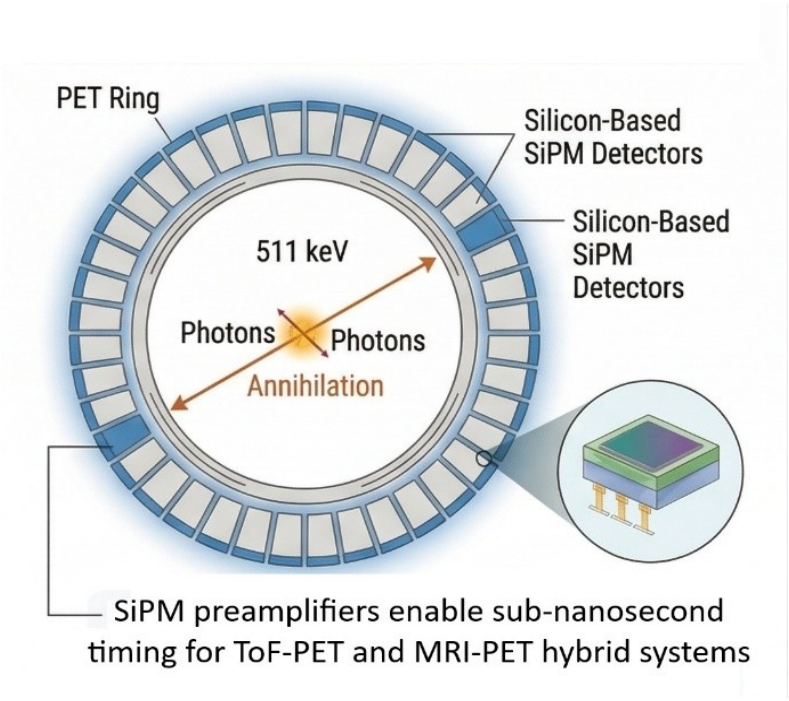


Figure 4: Silicon-based SiPM PET ring enabling 511-keV coincidence detection for ToF-PET and MRI-PET hybrid imaging.

3.2.2 Single-Photon Emission Computed Tomography (SPECT)

As illustrated in Figure 5, a clinical SPECT system consists of a sliding patient table, a pair of rotating dual-head gamma cameras that encircle the patient, and a dedicated workstation for image

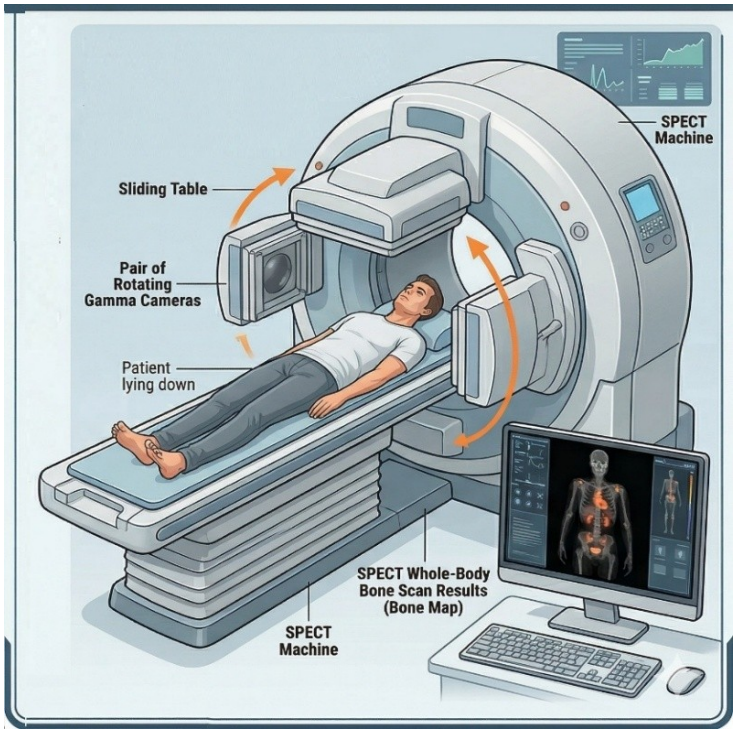


Figure 5: Clinical SPECT scanner with rotating dual-head gamma cameras and whole-body bone scan reconstruction.

reconstruction. The system measures the spatial distribution of gamma-emitting radiotracers, generating 3D tomographic maps of tracer concentration that are routinely used to assess myocardial perfusion, hepatobiliary function, renal perfusion, bone metabolism, thyroid pathology, and tumor staging [18,28]. Silicon detectors, particularly Compton cameras based on position-sensitive silicon strip detectors, offer an alternative imaging approach that eliminates the mechanical collimator, thereby recovering detection efficiency. As shown in Figure 6, the Compton camera design employs two silicon strip detector planes: a position-sensitive scattering detector that generates a recoiling electron, followed by an absorption detector that captures the scattered photon, enabling source localization without mechanical collimators. In combined SPECT/CT configurations, silicon detectors are integral to the CT subsystem for anatomical co-registration, as well as to data acquisition and image correction electronics [9].

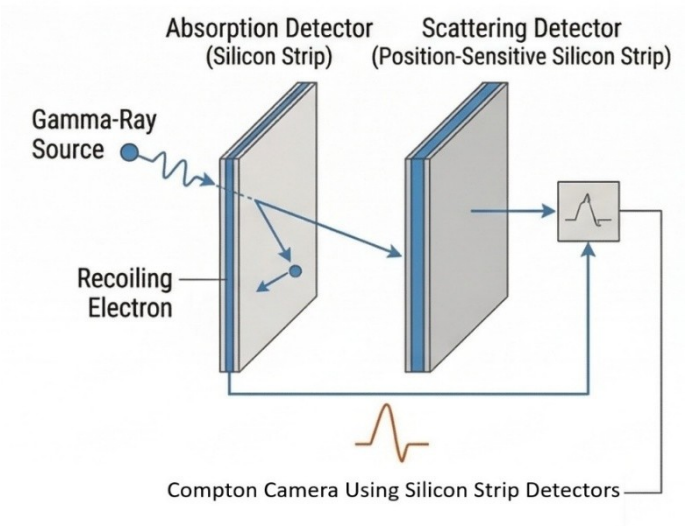


Figure 6: Compton camera using silicon strip detectors for collimator-free gamma-ray source localization

3.3 Silicon in Radiation Dosimetry

Accurate radiation dosimetry is paramount in both nuclear medicine imaging and radiotherapy to ensure therapeutic efficacy while protecting surrounding healthy tissue. Silicon diode dosimeters are widely deployed for this purpose, exploiting two key material advantages: near-tissue equivalence (Z_{eff} ≈ 14 vs. Z_{eff} ≈ 7.4 for soft tissue), which minimises energy-dependent correction factors, and a significantly greater charge yield per unit volume compared with air ionisation chambers, owing to silicon's substantially higher density (2.33 g/cm³) and lower electron–hole pair creation energy (3.6 eV vs. ~34 eV in air) [6,32]. For advanced radiotherapy modalities involving small radiation fields, steep dose gradients, and ultra-high dose rates such as FLASH radiotherapy, silicon dosimeters offer

spatial resolution below 1 mm², sub-millisecond response time, and the ability to be fabricated in body-implantable form factors for real-time in vivo dosimetry [6,33]. Beyond conventional silicon, silicon carbide (SiC) has emerged as a promising wide-bandgap material for next-generation radiation dosimetry, particularly in carbon ion therapy environments; a detailed comparative analysis is presented in Section 4 [33].

4. Comparative Analysis: Silicon vs. Silicon Carbide Detectors

Silicon carbide (SiC) has emerged as a wide-bandgap alternative to silicon for radiation detector applications requiring operation in high-temperature, high-radiation, or high-electric-field environments [21,33]. Among its polytypes, the 4H-SiC configuration has attracted the greatest research interest owing to its combination of a wide bandgap, high carrier saturation velocity, and superior thermal conductivity. Table 2 presents a systematic comparison of the principal physical properties of 4H-SiC and silicon at room temperature, and Figure 7 shows the cross-sectional architecture of a representative SiC p-n diode detector fabricated for dosimetry research.

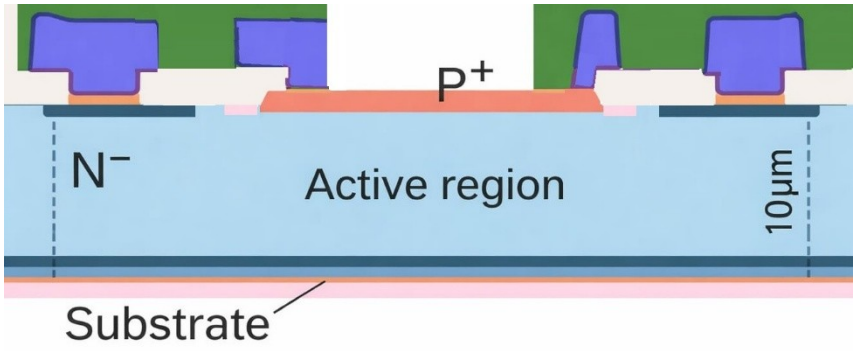


Figure 7: Cross-section of a SiC p-n diode with n⁻ epitaxial layer on substrate

As Table 2 demonstrates, SiC exhibits a bandgap approximately three times larger than silicon (3.23 eV vs. 1.12 eV), thermal conductivity roughly three times greater (4.9 vs. 1.5 W/cm·K), and a breakdown electric field ten times higher (3–4 × 10^6 vs. 3 × 10^5 V/cm). These properties make SiC inherently resistant to thermal noise and radiation-induced bulk damage, conferring an operational advantage near the Bragg peak in carbon ion therapy, where both high local ionisation density and elevated linear energy transfer (LET) present significant challenges for conventional silicon detectors. Matsumoto et al. investigated LET distribution measurements using an energy-dispersive SiC dosimeter exposed to pristine and spread-out Bragg peak (SOBP) carbon beams at 290 MeV/u, conducted across the plateau, SOBP, and post-Bragg peak regions. Relative biological effectiveness (RBE) was estimated using a linear-quadratic model based on dose-averaged LET values,

demonstrating the potential of SiC detectors for therapeutic dose calculation in carbon ion therapy, though further validation remains necessary [33].

Table 2. Comparative physical properties of 4H-SiC and Si at room temperature [21,33].

Property	4H-SiC	Si
Bandgap E_{gap} (eV)	3.23	1.12
Breakdown field E_breakdown (V/cm)	$3–4 \times 10^6$	3×10^5
Electron mobility μ_e ($cm^2/V \cdot s$)	800	1450
Hole mobility μ_h ($cm^2/V \cdot s$)	115	450
Saturation velocity V_sat (cm/s)	2×10^7	0.8×10^6
Atomic number Z	14/6 (average)	14
Relative permittivity ε_r	9.7	11.9
e-h pair creation energy (eV)	7.6–8.4	3.6
Density (g/cm^3)	3.22	2.33
Displacement energy E (eV)	30–40	13–15
Thermal conductivity (W/cm·K)	4.9	1.5

Silicon, however, retains decisive advantages in standard clinical environments: its lower e-h pair creation energy (3.6 eV vs. 7.6–8.4 eV in SiC) produces proportionally larger signal charges per unit deposited energy, its higher carrier mobilities enable faster timing, and its vastly more mature fabrication ecosystem delivers lower cost, better wafer uniformity, and seamless integration with CMOS readout electronics [6,7]. Table 3 quantifies these differences in detector energy resolution under alpha-particle irradiation, confirming that silicon achieves superior resolution under standard operating conditions.

Table 3. Measured ($FWHM^{det}$) and statistical-limit ($FWHM^{stat}$) energy resolutions for SiC and Si alpha-particle detectors [35].

Detector	$FWHM^{det}$ (keV)	$FWHM^{stat}$ (keV)
SiC	10.3	5.0
Si	7.3	3.5

The choice between Si and SiC is therefore application-driven: silicon is preferred for conventional PET/SPECT imaging and standard dosimetry owing to its superior signal yield and proven integration pathways, while SiC becomes the material of choice in high-radiation-dose or elevated-temperature environments, such as neutron dosimetry in reactor fields or heavy-ion therapy monitoring [33,35].

4.1 Hypothetical Simulation: Comparative Signal Analysis

To further illustrate the practical significance of the material differences summarised in Tables 2 and 3, a hypothetical simulation was conducted comparing Si and SiC detector performance under identical irradiation conditions. Assuming an absorbed energy of 5 MeV, pair creation energies of W_{Si} = 3.6 eV and W_{SiC} = 8.0 eV, and a Fano factor F = 0.1, consistent with established semiconductor detector theory [6], the number of generated electron–hole pairs N and the energy resolution R were computed according to:

$$N = \frac{E}{W} \qquad (1)$$

$$R = 2.35 \times \sqrt{\frac{F \times N}{N}} \qquad (2)$$

The results, presented in Table 4, confirm that silicon generates 2.22 times more electron–hole pairs than SiC under identical conditions, yielding proportionally superior energy resolution. This quantitative advantage directly underpins silicon's dominance in standard clinical PET/SPECT applications, while SiC retains its operational edge in high-radiation environments as discussed above.

Table 4. Hypothetical simulation results for Si and SiC detectors at E = 5 MeV, F = 0.1.

Parameter	Si	SiC
Pair creation energy W (eV)	3.6	8.0
e-h pairs N	1,388,889	625,000
Energy resolution R	6.31×10^{-4}	9.40×10^{-4}
FWHM (keV)	3.16	4.70
Relative signal strength	Strong	Moderate

Note: Values computed from standard detector physics formulae; W values from Table 2; F = 0.1 as typical for semiconductors [6]

As shown in Figure 8, silicon generates 2.22 times more electron–hole pairs than SiC under identical conditions (1,388,889 vs. 625,000 pairs), yielding a proportionally superior energy resolution (FWHM = 3.16 keV vs. 4.70 keV). This quantitative advantage directly underpins silicon's dominance in standard clinical PET/SPECT applications, while SiC retains its operational edge in high-radiation environments.



Figure 8. Hypothetical simulation comparing electron–hole pair yield, energy resolution, and signal strength for Si and SiC detectors under identical irradiation conditions.

5. Silicon detector advantages and the role of artificial intelligence in nuclear medicine

Semiconductor radiation detectors offer a set of advantages that collectively explain their dominance in nuclear medicine instrumentation. Their high density - 2.33 g/cm³ for silicon and 3.22 g/cm³ for SiC - maximises stopping power per unit thickness, permitting compact thin-detector designs with reduced parallax error in tomographic imaging [6,7]. The small diffusion length of charge carriers in solid-state materials, combined with fine strip or pixel segmentation, yields position resolution below 10 μm, far exceeding the capability of gas-filled detectors [7,14]. At 3.6 eV per electron–hole pair, silicon generates approximately 4–10 times more signal carriers than gas ionisation chambers (20–40 eV per ion pair) and 100–270 times more than scintillator systems (400–1000 eV per photoelectron), producing intrinsically superior energy resolution [6]. Silicon's full compatibility with CMOS fabrication further allows the detector and its readout electronics to be co-integrated on the same substrate, reducing parasitic capacitance, electronic noise, and system complexity, while carrier transit times below 30 ns support the tight coincidence timing windows required in PET systems and permit high-count-rate operation without significant pile-up distortion [7,8]. Engineered structures such as p-spray insulation and oxygen-enriched bulk silicon extend these advantages to radiation-tolerant designs adequate for cumulative doses below 1 kGy per year, which encompasses the majority of clinical nuclear medicine environments [7]. These hardware strengths are substantially amplified by artificial intelligence algorithms running on silicon-based processors — CPUs, GPUs, and custom neural processing units — which perform real-time image reconstruction, attenuation correction, tumour segmentation, lesion classification, and standardised uptake value (SUV) quantification in PET and SPECT workflows, as illustrated in Figure 9. Deep convolutional networks applied to low-count PET acquisitions have demonstrated reductions in administered radiotracer activity of up to 80% while preserving diagnostic image quality, automated tumour staging performance comparable to expert nuclear medicine physicians, and simultaneous multi-pathology classification on chest scintigraphy. The integration of AI inference engines directly into silicon ASIC readout chains, enabling event-level classification and adaptive thresholding at the point of acquisition, represents the next frontier in this convergence [19].

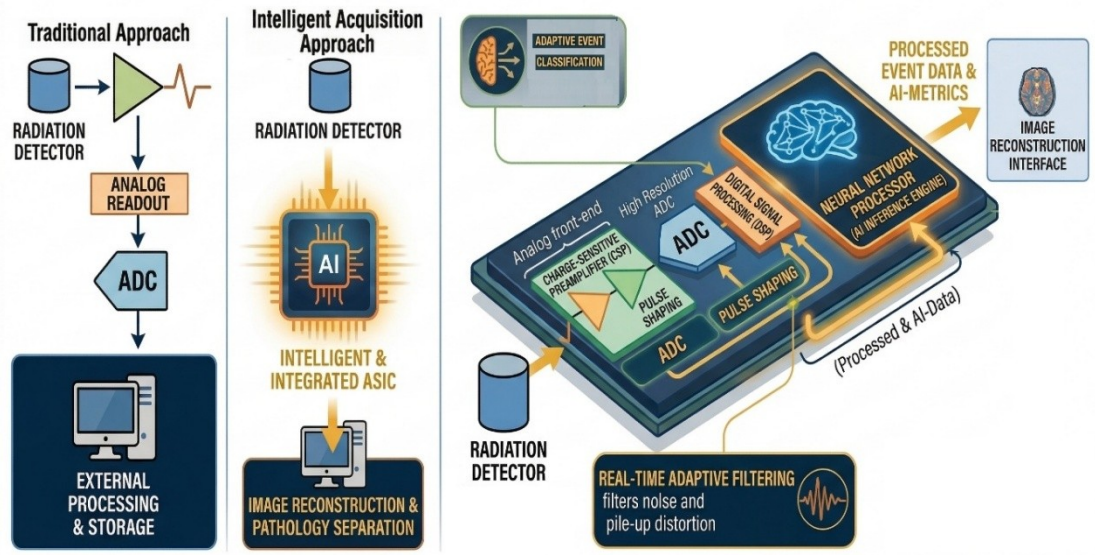


Figure 9: AI-integrated ASIC readout architecture for radiation detectors, enabling real-time event classification, adaptive filtering, and on-chip image reconstruction at the point of acquisition

6. Nuclear Medicine in Saudi Arabia: Current Status and Vision 2030 Initiatives

Saudi Arabia has made substantial investments in nuclear medicine infrastructure over the past two decades. The Kingdom operates 51 nuclear medicine facilities under government, private, and Ministry of Health management, collectively performing approximately 37,655 general nuclear medicine investigations and 12,387 cardiac studies annually, with a 2018 national survey documenting 21 PET/CT systems, 55 SPECT/CT systems, and 35 standalone SPECT and gamma camera units distributed across the Kingdom [37]. Major nuclear medicine centres are co-located with leading tertiary hospitals including King Abdulaziz Medical City, King Faisal Specialist Hospital and Research Centre, King Saud Medical City, and Almana Hospitals, offering comprehensive services encompassing oncology, cardiology, neurology, endocrinology, and nephrology [37,40]. The Nuclear and Radiological Regulatory Commission (NRRC) has established radiation safety regulations aligned with International Atomic Energy Agency (IAEA) standards and WHO guidelines. National Diagnostic Reference Levels (NDRLs) for radiographic, CT, and hybrid SPECT/CT examinations have been developed and validated through multicentre studies, providing an evidence-based framework for optimising patient doses while maintaining diagnostic quality [38,39,41]. Building on this foundation, the Kingdom's Vision 2030 health transformation agenda targets the systematic deployment of PET/CT for oncological, neurological, and cardiac imaging; the integration of nuclear medicine and medical physics curricula at Princess Nourah bint Abdulrahman University and King Saud University with American Board certification pathways for Saudi nationals; the development of

comprehensive regulatory frameworks for radiation source management and emergency preparedness under the NRRC; active participation in IAEA technical cooperation programmes for technology transfer and quality-assurance capacity building; and support for precision medicine and biomedical imaging research at King Abdullah University of Science and Technology (KAUST) [37]. These coordinated initiatives collectively position Saudi Arabia to transition from a technology-importing nation in nuclear medicine to one capable of regional leadership in advanced molecular imaging and silicon-based detector research.

7. Challenges and Future Research Directions

Despite the considerable maturity of silicon detector technology, several technical and translational challenges must be addressed before its full potential in nuclear medicine can be realised. The first and most established challenge is radiation damage and long-term stability. Prolonged exposure to high-energy particles induces bulk crystal defects (primarily vacancy clusters and interstitials), as well as surface oxide damage that progressively degrades detector performance by increasing leakage current and introducing charge trapping centres. Oxygen-enriched silicon and cryogenic operation at −20°C or below mitigate these effects but add system complexity and cost; future research should therefore prioritise defect-engineering strategies and the development of novel silicon allotropes with intrinsically enhanced radiation hardness [7]. A second frontier concerns ultra-high dose-rate dosimetry for FLASH radiotherapy. This emerging modality delivers therapeutic doses in microsecond pulses at rates exceeding 40 Gy/s, a regime in which conventional silicon diodes exhibit pronounced ion recombination and polarisation effects that compromise dosimetric accuracy. The development of diamond-like carbon thin-film sensors and SiC-based diodes specifically optimised for FLASH conditions represents a critically needed and actively pursued research direction [22]. Miniaturisation for in vivo dosimetry presents a third challenge. Implantable silicon dosimeters with wireless readout capability would enable real-time dose monitoring during brachytherapy and external-beam radiotherapy fractions; silicon's biocompatibility for implantable applications is well established, with silicone materials in clinical use since 1946 [3,4, 40], but achieving reliable power delivery and bidirectional data transmission within the body without introducing artefacts in concurrent MRI or CT imaging remains a significant engineering obstacle. The commercial scalability of SiC detectors constitutes a fourth constraint. Unlike silicon, SiC cannot be grown from a melt phase and requires vapour-phase epitaxy processes that are substantially more expensive and lower in throughput than the Czochralski method used for silicon. Achieving wafer-scale SiC fabrication with the uniformity and yield required for clinical nuclear medicine applications remains an open materials science challenge; hybrid Si/SiC detector arrays combining silicon's signal yield with SiC's radiation hardness in a single

module represent a pragmatic intermediate solution meriting systematic investigation [22,35]. Finally, the co-design of AI-capable silicon ASICs defines a fifth emerging frontier. Future nuclear medicine systems will increasingly require on-detector AI inference, with neural network engines embedded directly in the ASIC readout chain, enabling event-level classification, pile-up rejection, and adaptive thresholding in real time. Engineering silicon ASICs that incorporate neural processing capability while maintaining the noise floor, power consumption, and timing resolution demanded by medical detector operation represents an open and consequential problem at the intersection of semiconductor physics and machine learning [19].

8. Conclusions

This study has demonstrated that silicon occupies a central and indispensable position in the technological ecosystem of nuclear medicine. Its favourable physical properties (particularly low electron–hole pair creation energy, fast carrier mobility, compatibility with CMOS fabrication, and broad commercial availability) make it uniquely suited for radiation detection, PET and SPECT imaging systems, nuclear medical electronics, and precision dosimetry. The systematic comparison with silicon carbide (SiC) reveals complementary strengths: silicon excels in energy resolution and signal magnitude under standard conditions, while SiC's wider bandgap and higher radiation hardness offer advantages in extreme environments such as heavy-ion radiotherapy and high-neutron-flux facilities. This conclusion is further supported quantitatively by the hypothetical simulation, which demonstrates that silicon generates 2.22 times more electron–hole pairs than SiC under identical irradiation conditions, directly confirming its superior signal yield in standard clinical environments. Judicious selection of detector material based on application-specific requirements is therefore essential. The convergence of silicon hardware capabilities with artificial intelligence algorithms is driving a qualitative transformation in nuclear medicine diagnostics, enabling low-dose imaging, automated image interpretation, and adaptive scanning protocols with quantitative accuracy exceeding prior approaches. At the national level, Saudi Arabia's expanding nuclear medicine infrastructure and Vision 2030-aligned initiatives in workforce development, regulatory standardisation, and international scientific collaboration position the Kingdom to play an increasingly significant role in the adoption and development of silicon-based nuclear medicine technologies. Future research priorities should address radiation damage mitigation in silicon detectors, FLASH radiotherapy dosimetry, implantable in vivo dosimetry systems, scalable SiC fabrication, and the co-design of AI-capable silicon ASICs for next-generation nuclear medicine instrumentation.

Funding

The authors declare that this research received no external funding.

Data Availability Statement

Not applicable.

Conflicts of Interest

There is no conflict of interest, the writers assert.

References

[1] Wedepohl KH. The composition of the continental crust. *Geochimica et Cosmochimica Acta*. 1995;59(7):1217–1232. https://doi.org/10.1016/0016-7037(95)00038-2

[2] Datnoff LE, Snyder GH, Korndörfer GH, eds. *Silicon in Agriculture*. Studies in Plant Science, Vol. 8. Amsterdam: Elsevier; 2001. https://doi.org/10.1016/S0928-3420(01)80001-X

[3] Colas A, Curtis J. Silicone biomaterials: history and chemistry. In: Ratner BD, Hoffman AS, Schoen FJ, Lemons JE, eds. *Biomaterials Science: An Introduction to Materials in Medicine*. 2nd ed. Elsevier Academic Press; 2004. pp. 80–86. https://doi.org/10.1016/B978-012582463-7/50009-2

[4] Curtis J, Colas A. Medical applications of silicones. In: Ratner BD, Hoffman AS, Schoen FJ, Lemons JE, eds. *Biomaterials Science: An Introduction to Materials in Medicine*. 3rd ed. Academic Press; 2013. pp. 1106–1116. https://doi.org/10.1016/B978-0-08-087780-8.00096-3

[5] Lin W, Hoff H. Silicon materials. In: Doering R, Nishi Y, eds. *Handbook of Semiconductor Manufacturing Technology*. 2nd ed. Boca Raton: CRC Press; 2008. https://doi.org/10.1201/9781420017663

[6] Knoll GF. *Radiation Detection and Measurement*. 4th ed. New York: John Wiley & Sons; 2010. ISBN: 978-0-470-13148-0

[7] Dalla Betta GF, Boscardin M, Bomben M, Calderini G, Chauveau J, Chauveau J, Darbo G, Giacomini G, La Rosa A, Mendicino R, Muñoz FJ, Quirion D, Ronchin S, Sultan DMS. Silicon radiation detector technologies: from planar to 3D. *Chips*. 2023;2(2):83–101. https://doi.org/10.3390/chips2020006

[8] Michail C, Fountos G, Kalyvas N, Valais I. Radiation detectors and sensors in medical imaging. *Sensors (Basel)*. 2024;24(19):6251. https://doi.org/10.3390/s24196251

[9] Studen A, Brzezinski K, Chesi E, Cindro V, Clinthorne NH, Cochran E, Grošičar B, Grkovski M, Honscheid K, Kagan H, Lacasta C, Llosa G, Mikuž M, Stankova V, Weilhammer P, Žontar D. Silicon detectors for combined MR-PET and MR-SPECT imaging. *Nucl Instrum Methods Phys Res A*. 2012;702:88–90. https://doi.org/10.1016/j.nima.2012.08.040

[10] Härkönen J. Development of radiation hard particle detectors made of Czochralski grown silicon. *Acta Physica Polonica A*. 2013;124(2):372–376. https://doi.org/10.12693/aphyspola.124.372

[11] Park SJ, Rogers WL, Huh S, Kagan H, Honscheid K, Burdette D, Chesi E, Lacasta C, Llosa G, Mikuz M, Studen A, Weilhammer P, Clinthorne NH. A prototype of very high-resolution small animal PET scanner using silicon pad detectors. *Nucl Instrum Methods Phys Res A*. 2007;570(3):543–555. https://doi.org/10.1016/j.nima.2006.10.277

[12] Boguszewska-Czubara A, Pasternak K. Silicon in medicine and therapy. *J Elem*. 2011;16(3):489–497. https://doi.org/10.5601/jelem.2011.16.3.13

[13] Kramberger G. Radiation effects in silicon detectors. *Nuclear Instruments and Methods in Physics Research Section A*. 2019;924:192–197. https://doi.org/10.1016/j.nima.2018.09.060

[14] Bolla G, Andreev V, Baumbaugh A, Chramowicz J, Dagenhart D, Goldstein J, Gutierrez G, Hidas D, Hill C, Howell J, Jensen H, Johnson M, Kwan S, Lipton R, Noyes V, Pixley J, Sipos R, Skup KR, Whitmore J. Sensor development for the CMS pixel detector. *Nucl Instrum Methods Phys Res A*. 2002;485(1):89–97. https://doi.org/10.1016/S0168-9002(01)01904-1

[15] Grodzicka M, Moszyński M, Szczęśniak T, Kaczmarek M, Grodzicki M, Wolski D, Baszak J. Silicon photomultipliers in detectors for nuclear medicine. In: *Nuclear Medicine Imaging Detectors*. CRC Press; 2015. https://doi.org/10.1201/b18957-8

[16] Mazzillo M, Condorelli G, Sanfilippo D, Valvo G, Carbone B, Piana A, Fallica G, Billotta S, Belluso M, Bonanno G, Cosentino L, Pappalardo A, Finocchiaro P. Silicon photomultipliers for nuclear medical imaging applications. *Proc SPIE*. 2008;70030I. https://doi.org/10.1117/12.780204

[17] Cherry SR, Sorenson JA, Phelps ME. *Physics in Nuclear Medicine*. 4th ed. Philadelphia: Elsevier Saunders; 2012. https://doi.org/10.1016/B978-1-4160-5198-5.00001-0

[18] Tettey-Engmann F. Advances in artificial intelligence-based medical devices for healthcare applications. *Biomed Mater Devices*. 2025. https://doi.org/10.1007/s44174-025-00243-2

[19] International Atomic Energy Agency. *Annual Report 2020*. Vienna: IAEA; 2020. https://www.iaea.org/publications/reports/annual-report-2020

[20] Li H, Shi L, Ye Z, Jian Y, Chen B, Gong Z. Application of silicon nanowire field effect transistor (SiNW-FET) biosensor with high sensitivity. *Sensors*. 2023;23(15):6808. https://doi.org/10.3390/s23156808

[21] Petringa G, Cirrone GAP, Cuttone G, Fazzi A, Giordanengo S, Manti L, Marchetto F, Raffaele L, Rotondi M, Sabini MG, Valastro LM. Silicon carbide sensors in radiotherapy dosimetry: progress, challenges, and perspectives. *Front Sens*. 2025. https://doi.org/10.3389/fsens.2025.1608902

[22] Tooley MA. Electronics and biological signal processing. BJA Educ. 2023 Apr;23(4):122-127. doi: 10.1016/j.bjae.2022.12.004. Epub 2023 Feb 8. PMID: 36960437; PMCID: PMC10028393.

[23] Chircov C, Grumezescu AM. Microelectromechanical systems (MEMS) for biomedical applications. *Micromachines*. 2022;13(2):164. https://doi.org/10.3390/mi13020164

[24] Lecoq P. Development of new scintillators for medical applications. *Nucl Instrum Methods Phys Res A*. 2017;809:130–139. https://doi.org/10.1016/j.nima.2015.08.041

[25] Conti M, Bendriem B. The new opportunities for high time resolution clinical TOF PET. *Clin Transl Imaging*. 2019;7:139–147. https://doi.org/10.1007/s40336-019-00316-5

[26] Lozovoy KA, Kokhanenko AP, Dirko VV, Kukenov OI, Akimenko NS, Voitsekhovskii AV. Silicon-based avalanche photodiodes: advancements and applications in medical imaging. *Nanomaterials*. 2023;13(23):3078. https://doi.org/10.3390/nano13233078

[27] Owens A. *Compound Semiconductor Radiation Detectors*. CRC Press; 2022. https://doi.org/10.1201/9781315225425

[28] Lutz G. *Semiconductor Radiation Detectors: Device Physics*. Springer; 1999. https://doi.org/10.1007/978-3-540-71679-2

[29] Spieler H. *Semiconductor Detector Systems*. Oxford University Press; 2005. https://doi.org/10.1093/acprof:oso/9780198527848.001.0001

[30] Hartmann F. *Evolution of Silicon Sensor Technology in Particle Physics*. 2nd ed. Springer; 2017. https://doi.org/10.1007/978-3-319-64436-3

[31] Matsumoto T, Yamaguchi K, Yanagisawa R, Kubodera K, Arai Y, Makino T, Hirano Y, Ohno T. Development of a SiC semiconductor-based dosimeter for evaluating clinical dose distribution in carbon ion cancer therapy fields. *Nucl Instrum Methods Phys Res B*. 2023;542:151–157. https://doi.org/10.1016/j.nimb.2023.06.013

[32] Saini AS, Zhu TC. Energy dependence of commercially available diode detectors for in-vivo dosimetry. *Med Phys*. 2007;34(5):1704–1711. https://doi.org/10.1118/1.2719365

[33] Tudisco S, Altana C, Boscarino S, Cardella G, Cavallaro S, De Luca A, Dell'Aquila D, Di Toro M, Ferrarì D, Gao Z, Greco V, Lombardo I, Longo S, Maiolino C, Martorana N, Pagano A, Papa M, Pirrone S, Politi G, Porto F, Rizzo F, Russotto P, Scuderi V, Souliotis G, Stefanini AA. SiCILIA — silicon carbide detectors for intense luminosity investigations and applications. *Sensors*. 2018;18(7):2289. https://doi.org/10.3390/s18072289

[34] AlDawood Z, Alotaibi S, Alghamdi A, Alhazmi F, Alqahtani M, Alshahrani M. Emergency preparedness plan in radiology and nuclear medicine departments in the Kingdom of Saudi Arabia. *Radiat Phys Chem*. 2025;237. https://doi.org/10.1016/j.radphyschem.2025.112651

[35] Al Asmri M, Alotaibi J, Alwahbi A, Almadani N, Alabdullah AA, Alsharqi O, Hamidi S. The public health care system and primary care services in Saudi Arabia: a system in transition. *East Mediterr Health J*. 2020;26(4):468–476. https://doi.org/10.26719/emhj.20.050

[36] Alrehily F. Diagnostic reference levels of radiographic and CT examinations in Saudi Arabia: a systematic study. *Radiat Prot Dosimetry*. 2022;198(19):1451–1461. https://doi.org/10.1093/rpd/ncac183

[37] Al-Qahtani SM, Abuhaimed A, Alarifi A, Alghamdi S, Alrushoud A, Alsufiani H, Alsuhibani M, Alswieleh A, Alzimami K, Babikir E, Bashier A, Kharita MH. The establishment of national diagnostic reference levels for adult SPECT-CT in Saudi Arabia. *J Radiol Prot*. 2023;43(3). https://doi.org/10.1088/1361-6498/ace452

[38] Albahiti SK, Khafaji M, Batawil N, Catipay N, Alsafi K. Enhancing nuclear medicine practice in Saudi Arabia: advocating for comprehensive guidelines and local diagnostic reference levels. *Cureus*. 2024;16(2):e54230. https://doi.org/10.7759/cureus.54230

[39] Alsulimane ME. Evaluation of computed tomography radiation dose metrics in Saudi Arabia: comparison to national and international diagnostic reference levels. *Saudi Med J*. 2025;46(12):1409–1418. https://doi.org/10.15537/smj.2025.46.12.20250527

[40] Almuhammad A, Alrasheed M. "Design and Development of a Feeding Unit for Medical Devices Implanted Inside Human Body Using Artificial Intelligence Theories". Islamic University Journal of Applied Sciences. 2024;VI(I):151–167. https://doi.org/10.63070/jesc.2024.007

Islamic University Journal of Applied Sciences (IUJAS)
https://journals.iu.edu.sa/jesc

Volume VIII, Issue I, July 2026, Pages 269-279

Machine Learning Approaches for Indoor Temperature Prediction in Northern Saudi Arabia

Yazeed Basheer O. Alshammari [1, *]

[1] Department of AI & Data Science, College of Computer Science and Engineering, University of Hail, Hail, Saudi Arabia

*Corresponding author: (Yazeed Basheer O. Alshammari), Email: s202309381@uoh.edu.sa

Abstract

This paper compares ten different machine learning algorithms for estimating indoor temperatures using empirical sensor readings from a real-world environment in Hail, Saudi Arabia. The data was collected during July 2023, and includes a total sample size of 8,966 records with five minutes between each reading. Five distinct environmental factors were recorded: light level, humidity, pressure, altitude and temporal features. The machine learning algorithms evaluated were Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, K-Nearest Neighbors, SVR, Decision Tree, Random Forest, Gradient Boosting, and XGBoost. Each compared using R-squared, RMSE, MSE, and correlation metrics. Random Forest algorithm has achieved the highest level of performance with an R-squared value of 0.9307 and RMSE value of 0.4370 after hyperparameter tuning compared to traditional linear algorithms which have an R-squared value of below 0.33. As well, the Feature Importance analysis showed that the atmospheric pressure and humidity were the primary predictors. These findings validate tree based ensemble methods for modeling nonlinear relationships in indoor environmental systems and demonstrate practical applicability for smart building implementations in harsh climatic conditions.

Keywords: Indoor temperature prediction; Machine learning; Ensemble methods; Environmental sensors; Smart buildings

https://doi.org/10.63070/jesc.2026.027
Received 06 February 2026; Revised 20 April 2026; Accepted 25 April 2026.
Available online 30 April 2026.
Published by Islamic University of Madinah on behalf of *Islamic University Journal of Applied Sciences.*
This is an open access article published under the Creative Commons Attribution 4.0 International license (CC BY 4.0):https://creativecommons.org/licenses/by/4.0/

1. Introduction

Building upon the technologies of smart buildings, it is seen that unprecedented opportunities exist for a data-centric approach to environmental management. Indoor temperature is a key aspect that affects energy usage, as well as the productivity of occupants [1]. Indoor temperature can therefore be predicted using machine learning models to take proactive measures in terms of maintaining a proper atmosphere. This is more significant in regions of extreme climates, like Northern Saudi Arabia, where buildings are subjected to severe thermal conditions and consume up to 70% of residential electric energy for cooling in peak summer seasons [2]. Conventional reactive control systems adapt HVAC system control based purely on current temperature deviations [3]. Such an approach leads to an uncontrolled usage of energy and discomfort during transitions. Machine learning techniques can help achieve predictive control by learning complex nonlinear relationships between environmental factors without mathematical formulations [4]. Advances in low-cost sensors and computer capabilities have made it economically and technically possible to utilize sophisticated prediction models in real-time environments [5]. Despite considerable research into machine learning in building energy prediction, knowledge gaps still persist in this field of research. Most of the existing research has been based on temperate climate zones, with insufficient consideration of extreme conditions, which prevail in the Middle Eastern countries. Similarly, comparative studies tend to be based on a small subset of algorithms rather than on broader modeling approaches. Research results are mostly based on artificial or unavailable databases that cannot be verified elsewhere. This research fills the existing gaps by comparatively and comprehensively evaluating ten machine learning algorithms using real-world data collected by the Hail city sensors during peak summer. Some of the primary objectives of this research are identifying the optimal approaches for handling temperature prediction problems under extreme climate conditions, quantifying the variations between the standard statistical approaches and machine learning techniques, determining the significance of environment sensor variables, and formulating guidelines for practitioners of building management systems.

2. Literature Review

The traditional methodologies undertaken for indoor temperature prediction involved classical statistical techniques such as ARIMA models, exponential smoothing, and multiple linear regression techniques [6]. These methods have limited forecasting accuracy, especially when environmental interactions are involved. The drawback of these methods for indoor climate predictions lies with their inability to determine interaction effects among multiple environmental parameters. Machine learning applications in building energy systems have seen considerable progress in recent times [7]. Research

has shown that artificial neural network models have shown superior efficiency for energy consumption prediction tasks and temperature prediction tasks compared with other models. Support Vector Machines have been adequately used for HVAC energy prediction tasks, although sensitivity to hyperparameters remains a concern [8]. Tree-based ensemble techniques have been identified as excellent candidates for various tasks, considering the ability of these techniques for handling nonlinear relationships, the management of heterogeneous data, lack of overfitting with trees, and interpretability of these techniques [9]. Random Forest algorithms have been recorded as the best-performing algorithms for building energy system task benchmarks [10]. Gradient Boosting algorithms, including XGBoost, have shown superior accuracy with more computing power required for these techniques compared with the other algorithms. Even with progress, gaps persist across current research. While some work emphasizes lab settings or data gathered in mild weather, real-world use in scorching, dry zones like northern Saudi Arabia still lacks support. Additionally, numerous analyses test just a few machine learning methods - typically sticking to basic linear strategies or only a couple of combined models - which weakens how thoroughly different systems can be judged. Though tree-based models like Random Forest and Gradient Boosting excel at forecasting building energy use and indoor conditions - thanks to their skill with complex patterns and mixed sensor data - their tested success often ties closely to particular locations or datasets. Because of this, moving predictions across different environments stays a challenge, despite shared methodological roots. What often gets overlooked is how uneven the application of hyperparameter adjustments remains across studies, which may introduce bias. Because of this inconsistency, one algorithm may appear stronger just by setup - not skill. A quiet problem then emerges: comparisons lose balance when tuning effort varies without acknowledgment. Without equal refinement, claims about model superiority rest on shaky ground. Fair tests demand equal care in preparation - something still missing in much of the existing research. This work tackles existing gaps by comparing ten machine learning methods within one standardized setup, relying on actual sensor readings from a hot, dry region. Such an approach supports fairer judgments about how models function, especially when weather grows harsh - conditions rarely covered well in past studies. Recent studies have demonstrated strong performance of deep learning models for indoor temperature prediction in HVAC systems [11]. Additional research has validated machine learning approaches for indoor temperature prediction across different building environments [12].

3. Methodology

3.1 Dataset Description

The dataset was collected by University of Hail faculty and made publicly available through Ploms Science Press [13]. Data collection occurred during July 2023 within a residential building in Hail, Northern Saudi Arabia. The study period corresponds to peak summer when daily maximum temperatures frequently exceed 45 degrees Celsius. Environmental measurements were captured using IoT sensors at five-minute intervals, recording timestamp, light intensity measured in lux, temperature in degrees Celsius, humidity as a percentage, atmospheric pressure in Pascals, and altitude in meters. The raw dataset comprises 9,811 records; after removing 56 observations with missing values, 8,966 complete observations remained for analysis.

3.2 Data Preprocessing

With the help of extensive preprocessing, data quality was assured. The data present in the columns was represented uniformly by removing spaces from the column identifiers, also known as variables. Similarly, the variable ‘timestamp’ was converted from its original representation of time, where each time entry is specified with hours, minutes, and seconds, to a numerical form with all hours being normalized to an integer value from 0 to 86,399, representing the total time elapsed since the start of the day, which is captured by algorithms with diurnal patterns. The target variable used is the data on temperature, while the other five variables are considered as feature variables. Scale functions are not applied on the variables as they show invariance properties on addressed data, and experiments suggested that standardization did not impact the results.

3.3 Machine Learning Models

Ten regression algorithms were compared: Linear Regression, Ridge Regression with L2 regularization, Lasso with L1 regularization, Elastic Net combining L1 and L2 regularizations, K-Nearest Neighbors, Support Vector Regression with RBF kernel, Decision Tree, Random Forest, Gradient Boosting, and XGBoost. All the algorithms in this were implemented using scikit-learn version 1.1 and higher and XGBoost version 1.7 and higher [14][15]. Initial training was done on default hyperparameters to define baseline performance. This allows a fair assessment of intrinsic model behavior prior to optimization. After baseline evaluation, hyperparameter optimization was performed for the best algorithm based on the baseline: the randomized search with fivefold cross-validation evaluating 50 configurations.

3.4 Evaluation Protocol

The dataset was split into a training set comprising 80% of observations (7,173 samples) and a testing set comprising 20% of observations (1,793 samples). To preserve the temporal structure of the sensor data and avoid potential data leakage, the dataset was sorted chronologically based on the time feature prior to splitting. The first 80% of observations were used for training, while the remaining 20% were reserved for testing. Quantification of outcomes was performed using R-squared, which estimates the proportion of variance explained, Mean Squared Error (MSE), which measures average squared error, Root Mean Squared Error (RMSE), which is interpretable in degrees Celsius, Pearson correlation, and Spearman correlation. R-squared and RMSE were considered the primary performance indicators due to their interpretability and relevance to regression accuracy.

4. Results and Analysis

4.1 Model Performance Comparison

Performance metrics show significant variations of the modelling techniques. It can be observed that the highest performance of the baseline model was achieved by the Random Forest model with an R-squared value of 0.9233 and an RMSE of 0.4599 degrees. XGBoost model ranked second, achieving an R-squared value of 0.9071 and an RMSE of 0.5061 degrees. Decision Tree model ranked third, with an R-squared of 0.8660 and an RMSE of 0.6080 degrees, while the Gradient Boosting model obtained an R-squared of 0.7941 and an RMSE of 0.7534 degrees. High Pearson correlation of 0.9609 and Spearman correlation of 0.9381 for the Random Forest model indicate very close agreement between the predicted values and the actual temperature values. The linear models showed a performance that was substantially worse than others. Ridge and Linear Regression had almost identical performance, with an R-squared of around 0.32 and an RMSE of around 1.37 degrees. Lasso and Elastic Net models performed slightly worse than this. The poor performance of linear models with an R-squared of less than 0.33 reinforces the supposition of the non-linear relationships between sensor readings and temperature. The distance-based models had the poorest performance overall, with an R-squared of 0.0561 and an RMSE of 1.6133 degrees for SVR, and an R-squared of 0.0409 and an RMSE of 1.6262 degrees for KNN, which likely resulted from the curse of dimensionality and the use of default hyperparameters. Overall, the results demonstrate clear and consistent performance differences across model families, confirming that nonlinear ensemble methods are better suited for indoor temperature prediction under real-world sensor conditions. Similar performance advantages of ensemble learning methods have been reported in recent building energy prediction studies [16]. Deep learning hybrid models have also shown strong capability in capturing temporal dependencies in indoor temperature data [17]. Figure 1 provides a visualization of the performance of the models.

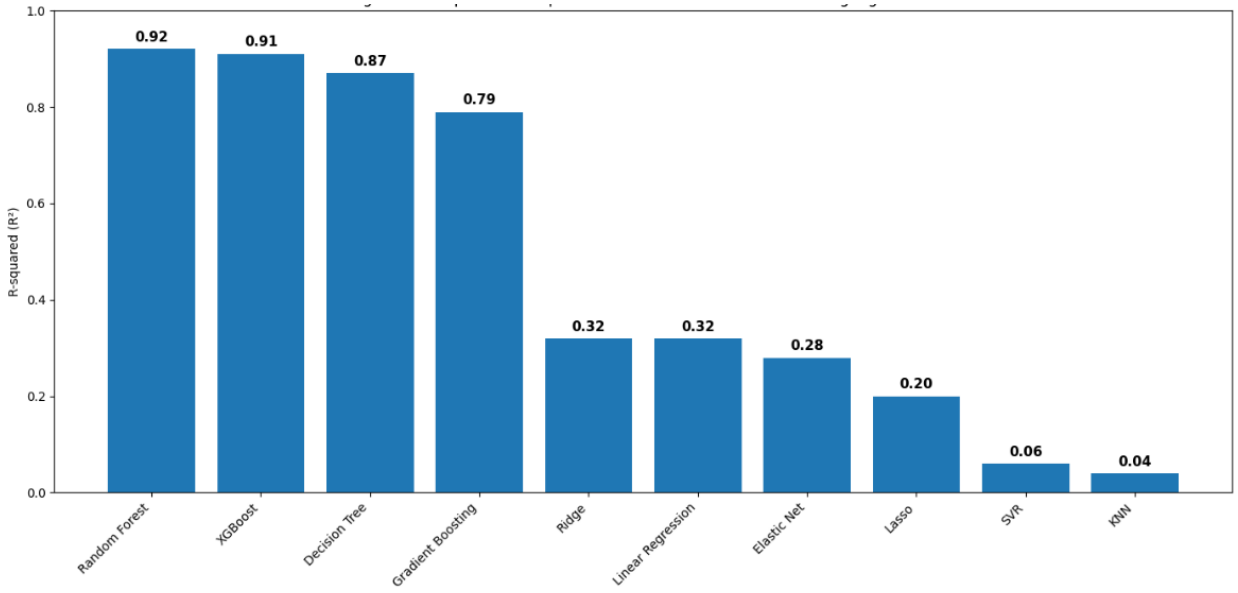


Figure 1. Comparative R-squared Performance of Machine Learning Algorithms

4.2 Hyperparameter Optimization

Comprehensive hyperparameter tuning of Random Forest using randomized search evaluated 50 configurations. Optimal configuration consisted of 300 estimators, unlimited depth, minimum 5 samples for splitting, minimum 1 sample per leaf, square root feature selection, and disabled bootstrap. This achieved cross-validation R-squared of 0.9013 and test performance R-squared of 0.9307 with RMSE of 0.4370 degrees. Optimization yielded meaningful improvements with R-squared increasing by 0.74 percentage points from 0.9233 to 0.9307, and RMSE decreasing 0.023 degrees from 0.4599 to 0.4370, representing 9.6% reduction in unexplained variance and 5% reduction in error magnitude with tangible operational benefits for building management.

4.3 Environmental Parameter Analysis

Analysis from the optimized Random Forest reveals relative contribution of environmental measurements to prediction accuracy. Atmospheric pressure emerged most influential at 38%, followed by humidity at 27%, time-of-day at 22%, light intensity at 11%, and altitude at 2%. Pressure dominance aligns with physical expectations as variations are associated with weather systems influencing outdoor conditions and building thermal loads. Humidity importance reflects the relationship with sensible heat transfer and HVAC latent cooling demands. Temporal feature captures systematic diurnal patterns from solar radiation, occupancy schedules, and HVAC modes. Light serves as a proxy for solar heat gain and occupancy-related internal heat. Minimal altitude importance is

expected for fixed sensor location. Figure 2 illustrates the relative contribution of environmental parameters.

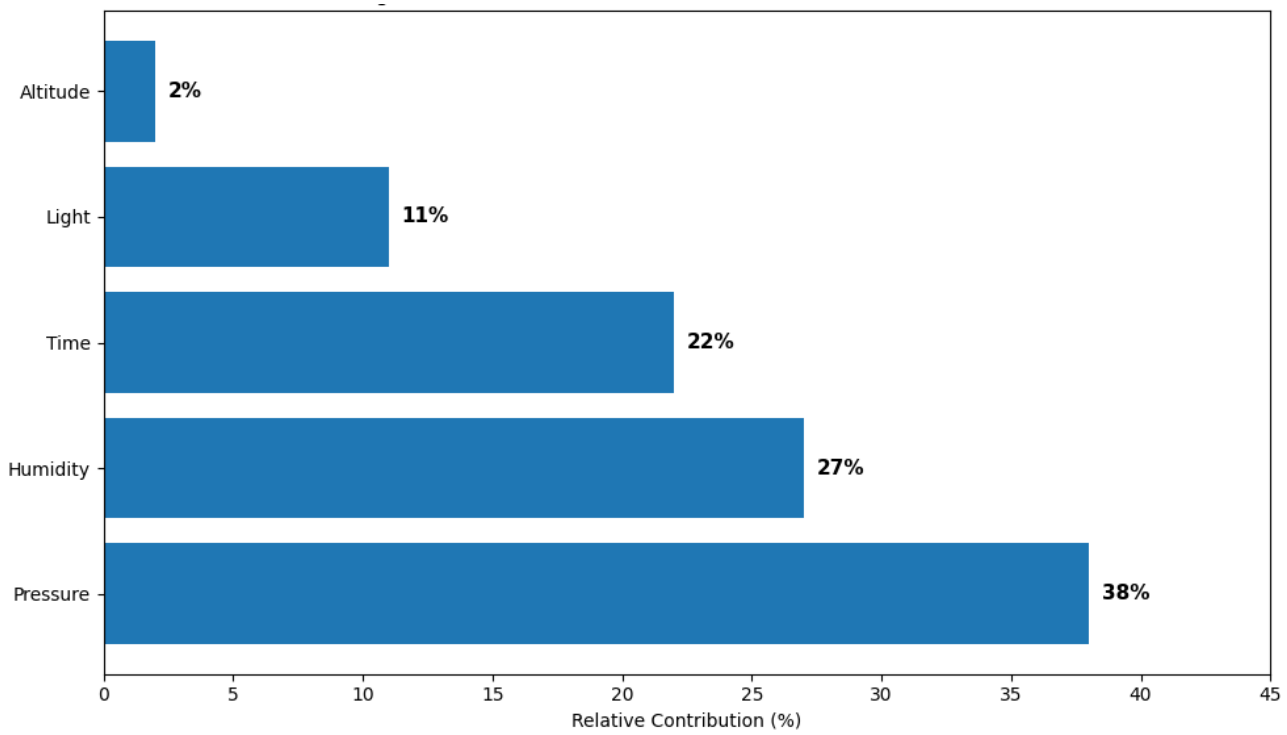


Figure 2. Relative Contribution of Environmental Parameters to Temperature Prediction

5. Discussion

5.1 Performance Interpretation

Despite their simplicity, linear approaches fall short when forecasting indoor temperatures, mainly because building thermal patterns involve complex shifts across multiple factors. Results show that ensemble techniques handle such complexity far better than traditional distance-driven or straight-line models. Nonlinear relationships - like those linking humidity, atmospheric pressure, and hour-by-hour changes - are too complex for basic methods to grasp accurately. As a result, more adaptive strategies emerge stronger in these conditions. Despite their differences, tree ensembles like Random Forest, XGBoost, and Gradient Boosting often outperform simpler models by splitting data dynamically while capturing complex patterns naturally. What distinguishes Random Forest apart is steady behavior across datasets - its strength comes from combining multiple trees built on resampled training cases, a method that smooths out errors and supports reliable predictions. Unlike more flexible approaches, linear methods - such as Linear, Ridge, Lasso, and Elastic Net - perform poorly, which suggests the process at play isn't governed by straight-line patterns. Because these models miss how environmental factors influence one another, they consistently fail to fit the data well. Performance issues with KNN and SVR reveal how strongly they react to unscaled features, tuning choices, and noisy, high-dimensional inputs. Without intense scaling or fine-tuned kernels, their results lag - ensemble

approaches handle such conditions better when left near defaults. Across different models, outcomes remain stable - a sign the method holds up under variation - while also showing ensemble techniques with nonlinear traits handle indoor environment forecasts more effectively. Results do not waver much, suggesting the setup works well, given how these predictive systems perform similarly when conditions change slightly. Predictions stay accurate because the approach adapts naturally to complex patterns found indoors, unlike simpler linear alternatives. Stability in performance implies trust in design, especially since varied setups lead to similar conclusions about forecasting ability. These findings align with broader research indicating that nonlinear models outperform linear approaches in complex environmental systems [18].

5.2 Practical Implications

The optimized Random Forest model, with an RMSE of 0.4370°C, demonstrates strong potential for supporting predictive control in HVAC systems by enabling short-term forecasting of indoor temperature trends. This allows building management systems to shift from purely reactive control to anticipatory adjustments that can improve thermal stability and operational efficiency. The model relies only on standard environmental sensors commonly available in modern building monitoring systems, eliminating the need for additional specialized hardware. This improves its feasibility for deployment in existing infrastructure. In terms of implementation, the model is computationally lightweight at inference time, making it suitable for real-time use in embedded or edge-based controllers. This supports integration into smart building systems where low-latency decision-making is required. Additionally, the use of feature importance analysis improves interpretability by identifying the most influential environmental variables, which can assist in system diagnostics and provide insight into sensor or environmental behaviour under different operating conditions. Overall, the results indicate that tree-based ensemble models can be effectively applied in smart building environments to support data-driven temperature prediction and enhance HVAC decision-making under real-world constraints. Data-driven HVAC control strategies have been widely studied for improving energy efficiency and system responsiveness [19].

5.3 Limitation and Future Work

Despite strong results, several limitations remain. The dataset is restricted to a single building over one summer month, limiting exposure to diverse weather conditions and usage patterns. This reduces the generalizability of the findings to other buildings, seasons, and climatic regions. The evaluation relies on a single chronological train–test split, which preserves temporal order but does not capture broader temporal variability. More robust approaches such as rolling or sliding window validation would better

reflect model performance over time. Model comparison is also slightly affected by uneven hyperparameter tuning, as only the Random Forest model was optimized while others used default configurations. Although baseline comparisons remain informative, this introduces potential bias in relative performance assessment. Finally, the study focuses on short-term prediction and does not evaluate integration with real HVAC control systems or energy consumption impact, leaving practical deployment outcomes untested. Future work should expand the dataset across multiple buildings and seasons, apply consistent hyperparameter optimization across all models, and adopt more rigorous time-series validation techniques. Integration with real HVAC systems would also allow evaluation of energy efficiency and operational performance in real-world conditions. Advanced forecasting techniques such as sequence-based deep learning models have shown promise for improving long term prediction accuracy [20]. Comparative studies of machine learning models further highlight the importance of robust validation in time-series prediction tasks [21].

6. Conclusion

From a residential structure in northern Saudi Arabia, sensors gathered real-world climate readings during intense summer periods, forming the foundation for assessing ten predictive algorithms. Tree ensembles like Random Forest and XGBoost lead in performance, outperforming simpler linear approaches along with proximity-driven ones. Every scoring measure aligns with this trend, hinting at complex, nonlinear interactions within indoor thermal patterns. Such outcomes suggest conventional models may lack sensitivity when handling erratic household environment shifts. Although data stems from one location, it still captures realistic indoor thermal dynamics under extreme climatic conditions. This work primarily offers a side-by-side analysis of various regression methods, tested uniformly through actual sensor readings taken in severe weather. While earlier research often examines only a few models or idealized settings, here a wider selection of algorithms undergoes assessment under identical conditions - making contrasts in results more reliable. What stands out is the coherence of testing applied across diverse approaches, something frequently absent in past efforts. Moreover, atmospheric pressure, along with humidity and time-based trends, emerges clearly as central to forecasting temperature, reinforcing their role in simulating indoor conditions. Evidence shows methods rooted in data are capable of reflecting intricate heat dynamics within built spaces. Despite these findings, conclusions face constraints due to reliance on one structure and a narrow seasonal window, limiting broader applicability. Moving ahead, research might cover diverse structures and annual cycles instead; stronger temporal analysis techniques could be applied alongside testing within actual heating and cooling operations to examine effects on power use and daily function.

Funding

The authors declare that this research received no external funding.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] H. X. Zhao and F. Magoulès, "A review on the prediction of building energy consumption," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 6, pp. 3586–3592, 2012. https://doi.org/10.1016/j.rser.2012.02.049

[2] A. Gassar and S. Cha, "Assessment of current energy consumption in residential buildings in Jeddah, Saudi Arabia," *Buildings*, vol. 9, no. 7, p. 163, 2019. https://doi.org/10.3390/buildings9070163

[3] A. Afram and F. Janabi-Sharifi, "Theory and applications of HVAC control systems: A review of model predictive control (MPC)," *Building and Environment*, vol. 72, pp. 343–355, 2014. https://doi.org/10.1016/j.buildenv.2013.11.016

[4] K. Amasyali and N. M. El-Gohary, "A review of data-driven building energy consumption prediction studies," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1192–1205, 2018. https://doi.org/10.1016/j.rser.2017.04.043

[5] C. Deb, F. Zhang, J. Yang, S. E. Lee, and K. W. Shah, "A review on time series forecasting techniques for building energy consumption," *Renewable and Sustainable Energy Reviews*, vol. 74, pp. 902–924, 2017. https://doi.org/10.1016/j.rser.2017.02.028

[6] C. Voyant et al., "Machine learning methods for solar radiation forecasting: A review," *Renewable Energy*, vol. 105, pp. 569–582, 2017. https://doi.org/10.1016/j.renene.2017.05.038

[7] T. Ahmad et al., "Machine learning in building energy prediction: A review," *Renewable and Sustainable Energy Reviews*, vol. 144, p. 110611, 2021. https://doi.org/10.1016/j.rser.2020.110611

[8] Y. Zhang, Z. O'Neill, B. Dong, and G. Augenbroe, "Comparisons of inverse modeling approaches for predicting building energy performance," *Building and Environment*, vol. 86, pp. 177–190, 2015. https://doi.org/10.1016/j.buildenv.2015.01.021

[9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. https://link.springer.com/article/10.1023/A:1010933404324

[10] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
https://projecteuclid.org/euclid.aos/1013203451

[11] R. Zhuang et al., "Deep learning-based room temperature prediction for HVAC systems," *Sensors*, vol. 23, no. 5, p. 2732, 2023.
https://doi.org/10.3390/s23052732

[12] H. Golshanfard et al., "Machine learning-based indoor temperature prediction in buildings," *Journal of Building Engineering*, vol. 52, p. 104432, 2022.
https://doi.org/10.1016/j.jobe.2022.104432

[13] University of Hail Faculty, "Temporal Variations Dataset for Indoor Environmental Parameters in Northern Saudi Arabia," 2023.
https://www.researchgate.net/publication/371764282

[14] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
https://jmlr.org/papers/v12/pedregosa11a.html

[15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
https://doi.org/10.1145/2939672.2939785

[16] F. Elmaz et al., "CNN-LSTM architecture for predictive indoor temperature modeling," *Building and Environment*, vol. 206, p. 108327, 2021.
https://doi.org/10.1016/j.buildenv.2021.108327

[17] N. Norouzi et al., "Applicability of deep learning algorithms for predicting indoor temperatures: Towards the development of digital twin HVAC systems," *Buildings*, vol. 13, no. 6, p. 1542, 2023.
https://doi.org/10.3390/buildings13061542

[18] J. Wei et al., "Deep learning for building energy prediction: A review," *Applied Energy*, vol. 272, p. 114683, 2020.
https://doi.org/10.1016/j.apenergy.2019.114683

[19] C. Fan et al., "Short-term building load prediction using machine learning methods," *Energy and Buildings*, vol. 236, p. 110373, 2021.
https://doi.org/10.1016/j.enbuild.2020.110373

[20] Y. Chen et al., "Machine learning for building energy prediction: Methods and applications," *Energy*, vol. 196, p. 117560, 2020.
https://doi.org/10.1016/j.energy.2020.117560

[21] G. Li et al., "Performance evaluation of sequence-to-sequence-attention model for short-term building energy prediction," *Energy*, vol. 259, p. 124915, 2022.
https://doi.org/10.1016/j.energy.2022.124915

Islamic University Journal of Applied Sciences (IUJAS)
https://journals.iu.edu.sa/jesc

Volume VIII, Issue 1, July 2026, Pages 280-298

Enhancing Arabic Dialectical Speech Recognition with Large Language Models

Rehab Albeladi[1]

[1] Department of Computer Science, Taibah University, Saudi Arabia

rbeladi@taibahu.edu.sa

Abstract

Automatic Speech Recognition (ASR) for Arabic faces unique challenges due to the language's diglossic nature, where Modern Standard Arabic (MSA) coexists with diverse dialects that differ substantially at lexical, phonological, and morphological levels. In this study, we present a comparative evaluation of state-of-the-art ASR models, Whisper (medium and large) and Wav2Vec2, applied to five Arabic varieties: Egyptian, Hijazi, Khaliji, Najdi, and MSA. We design a two-phase experiment: in the first phase, 50 speech samples per dialect are evaluated under three settings, raw transcriptions, and outputs corrected by two large language models (GPT and ALLAM). In the second phase, the best-performing ASR–LLM combination is scaled to the full test set for each dialect. Performance is assessed using Word Error Rate (WER) and semantic similarity distance. Results show that Whisper-Large combined with GPT consistently outperforms other systems across dialects. Improvements are especially pronounced in dialectal speech, while MSA achieves the lowest baseline WER. These findings highlight both the persistent difficulty of dialectal Arabic ASR and the effectiveness of LLM-based post-processing for reducing recognition errors.

Keywords: Automatic Speech Recognition; Large Language Models; Arabic Dialects.

https://doi.org/10.63070/jesc.2026.028
Received 15 February 2026; Revised 20 April 2026; Accepted 24 April 2026.
Available online 30 April 2026.
Published by Islamic University of Madinah on behalf of *Islamic University Journal of Applied Sciences.*
This is an open access article published under the Creative Commons Attribution 4.0 International license (CC BY 4.0):https://creativecommons.org/licenses/by/4.0/

1. Introduction

Automatic speech recognition (ASR) systems have made substantial progress in recent years, driven by advances in self-supervised learning and large-scale pretraining. Models such as Whisper [1] and Wav2Vec2.0 [2] have achieved near-human performance on several benchmark datasets, particularly for high-resource languages like English. However, the application of ASR to Arabic speech—especially its dialectal varieties—remains a significant challenge. Arabic is characterized by a diglossic structure, with Modern Standard Arabic (MSA) used in formal settings and a diverse set of regional dialects used in everyday speech. These dialects differ considerably in phonology, morphology, syntax, and lexicon, and they are often underrepresented in ASR training corpora.

Despite increasing interest in Arabic speech technologies, few studies have systematically evaluated the performance of state-of-the-art ASR models across a broad set of Arabic dialects. Moreover, raw ASR outputs often lack fluency and readability, particularly when dealing with dialectal input. Recent developments in large language models (LLMs), such as GPT-4 and Arabic-specific models like ALLAM, offer promising tools for post-processing ASR outputs. These models can correct transcription errors, restore grammaticality, and preserve meaning through context-aware generation.

In this work, we investigate how well two leading ASR systems—Whisper and Wav2Vec2—perform on Arabic dialectal speech using the SADA dataset, a multi-dialectal corpus covering five major Arabic varieties. We further examine the potential of LLMs to improve ASR outputs through generative correction, comparing the performance of GPT and ALLAM in this task. Our evaluation employs both Word Error Rate (WER), a traditional metric for transcription accuracy, and Semantic Distance (SemDist), which quantifies the preservation of meaning between hypotheses and reference texts. By analyzing results across dialects and system configurations, we aim to offer insights into the strengths and limitations of current tools for Arabic ASR, and to highlight the role of LLMs in bridging the gap in low-resource or high-variance speech settings. The focus of this study is on quantitative evaluation through objective metrics such as WER and semantic distance, complemented by statistical analysis across dialects. A qualitative investigation of recognition or correction errors is beyond the scope of this work.

2. Literature Review

This section presents a review of the existing literature on Automatic Speech Recognition (ASR) for the Arabic language. It begins with an overview of the field, outlining its principal challenges and historical development. Subsequently, it examines the application of deep learning models to Arabic ASR. The section then discusses the emerging practice of integrating Large Language Models (LLMs)

into ASR pipelines for output refinement. Finally, it concludes with a summary of relevant comparative studies that have benchmarked the performance of various ASR systems, providing context for the present study.

2.1 Overview of Arabic ASR Research

The evolution of Automatic Speech Recognition for Arabic is characterized by efforts to address a set of persistent linguistic challenges. The most significant of these is the diglossic state of the language, which involves a distinction between Modern Standard Arabic (MSA)—the form used in formal, official, and written contexts—and numerous colloquial dialects used as a spoken language for daily communication [3]. These dialects are not merely variations in accent but constitute distinct linguistic systems with substantial lexical, phonological, morphological, and syntactic divergences from MSA. For instance, the interrogative phrase "how are you?" translates to kayfa ḥāluk in MSA, but is commonly rendered as izzayak in Egyptian Arabic, shlōnak in Gulf Arabic, and kīfāk in Levantine Arabic. This degree of lexical divergence makes ASR systems trained exclusively on MSA data fail to recognize basic conversational phrases from dialectal speech.

Initial research in Arabic ASR focused predominantly on MSA. This focus was a pragmatic consequence of the greater availability of structured, annotated datasets derived from formal sources, such as broadcast news corpora. The main technology of this period was the Hidden Markov Model (HMM) framework, typically utilizing Gaussian Mixture Models (GMMs) to model the acoustic states of phonetic units [4]. The application of these GMM-HMM systems to Arabic was complicated by the language's complex morphology. A single Arabic root can yield hundreds of word forms through the affixation of proclitics (e.g., و"and", ف"so") and enclitics (e.g., pronoun suffixes). This morphological complexity results in high out-of-vocabulary (OOV) rates, thereby limiting model performance [5].

The performance of these MSA-centric models on dialectal Arabic exhibited substantial degradation. This linguistic gap underscored the necessity for dialect-specific resources. Consequently, a significant focus of subsequent research has been the development of large-scale, annotated speech corpora for various dialects. The creation of resources such as the SADA (Saudi Audio Dataset for Arabic) for Saudi dialects [6] and the Arabic Multi-Genre Broadcast MGB-2 dataset [7] are important contributions. The development of such corpora is a resource-intensive process that involves recording, manual transcription, and careful annotation to capture the phenomena of spontaneous speech, including disfluencies and code-switching. These datasets now provide the necessary foundation for the development and robust evaluation of dialect-aware ASR systems.

2.2 Deep Learning-Based ASR for Arabic

The application of deep learning to ASR marked a significant methodological shift, leading to substantial performance improvements and addressing many of the limitations of GMM-HMM systems. The initial integration of deep learning in ASR involved replacing GMMs with Deep Neural Networks (DNNs) within the hybrid GMM-HMM framework, enabling more discriminative modeling of acoustic features and resulting in immediate performance gains [8].

Subsequently, Recurrent Neural Networks (RNNs) architectures, including Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, became prevalent. The capacity of these models to capture long-range temporal dependencies in the speech signal was well-suited for language processing. For Arabic, this translated to an improved ability to model coarticulation effects and contextual phonetic information across long utterances, mitigating a known weakness of the first-order Markov assumption in HMMs [9].

A more fundamental change was the adoption of end-to-end (E2E) models. In these architectures, a sequence of acoustic features is directly mapped to a sequence of characters or words, unifying the acoustic, pronunciation, and language model components of traditional ASR into a single, jointly trained neural network. Particularly beneficial for Arabic dialects, this approach eliminates the need for manually constructed phonetic dictionaries, which are often unavailable for unwritten colloquial variants. Two primary E2E frameworks became prominent: Connectionist Temporal Classification (CTC) [10] and attention-based sequence-to-sequence (seq2seq) [11] models. CTC-based models learn a monotonic alignment between acoustic frames and output labels, while attention mechanisms permit the model to learn dynamic, non-monotonic alignments.

Current state-of-the-art systems are largely based on large-scale, pre-trained models that utilize the transformer architecture. The methodology of pre-training on extensive datasets and subsequently fine-tuning on a specific task has proven to be highly effective. Wav2Vec 2.0 [2] is a prominent example of self-supervised learning in this domain. These models are fine-tuneable on smaller, labeled dialectal datasets, which for instance, demonstrated effectively for Egyptian Arabic [12].

In parallel, Whisper [1]—trained in a weakly-supervised fashion on around 680,000 hours of multilingual, multitask audio—illustrates the potential of large-scale models in delivering "zero-shot" ASR performance. It can transcribe languages and dialects without explicit supervised fine-tuning, offering a robust baseline for under-resourced Arabic dialects [13].

2.3 LLM Integration in ASR Pipelines

The integration of large language models (LLMs) into the post-processing pipeline of automatic speech recognition (ASR) systems has emerged as a promising approach for error correction [14]. Traditional ASR systems, while increasingly accurate, still produce errors due to factors such as acoustic ambiguity, speaker variability, and domain mismatch. LLMs, with their advanced contextual understanding and generative capabilities, offer a promising solution for refining ASR outputs by correcting transcription errors and enhancing overall readability and accuracy.

Recent research has demonstrated that LLMs can be employed in both zero-shot and few-shot settings to perform ASR error correction without the need for extensive retraining on ASR-specific data. These models can leverage N-best lists—multiple hypotheses generated by the ASR system—to provide richer context for correction, or operate directly on the 1-best output for more generalizable solutions [15]. Approaches such as generative error correction (GEC) utilize LLMs to generate improved transcriptions by either unconstrained generation or by selecting the most plausible candidate from the N-best list [16]. Another approach investigated in the literature is in-context learning and prompt engineering, in which examples or specific instructions are provided in the prompt, so that LLMs can be guided to perform more accurate corrections, even without explicit retraining [17].

While an experimental study showed that LLM was unable to correct transcription errors [18], recent research has demonstrated that LLM-based correction reduces word error rate (WER) and improves transcription quality across a variety of ASR architectures, including both transducer and attention-based encoder-decoder models [15][19][20]. Notably, LLMs have shown remarkable generalization capabilities, effectively correcting outputs from different ASR systems even in black-box scenarios where the underlying ASR model is unknown [15]. Multilingual LLMs have also proven capable of correcting ASR errors in a wide range of languages, leveraging their cross-lingual knowledge to provide robust performance even when only the 1-best hypothesis is available [19].

For Arabic ASR, the post-correction stage has considerable utility. LLMs can be employed to address a range of prevalent error types, including the confusion between phonetically similar but orthographically distinct letters, errors in grammatical agreement (e.g., subject-verb, noun-adjective), which are common in Arabic and are frequently missed by ASR systems that lack deep syntactic awareness. However, the application of LLMs for ASR post-processing and error correction in Arabic speech remains less explored compared to English. Emerging research has begun to leverage LLMs for code-switched Arabic-English ASR, utilizing models like GPT-4 to generate synthetic code-switched data and improve ASR performance through N-best list rescoring and language model interpolation [24].

2.4 Comparative ASR Studies

As the variety of ASR models and methodologies for Arabic has expanded, the need for rigorous, data-driven comparison has become critical for benchmarking performance and guiding future research [7]. These comparative studies are essential for quantifying the relative merits of different architectures and understanding the specific conditions under which certain models excel. Initial comparative works focused on benchmarking traditional GMM-HMM systems against the first wave of hybrid DNN-HMM and end-to-end RNN-based models. These studies consistently reported the superior performance of deep learning approaches, particularly when larger training datasets were utilized [22][23].

The current focus of comparative analysis is on the new generation of large pre-trained models. Researchers are actively benchmarking fine-tuned Wav2Vec models against Whisper's zero-shot and fine-tuned performance across various Arabic speech datasets, such as the Multi-Genre Broadcast (MGB) challenge datasets [13][24]. These studies often reveal a trade-off: Whisper's broad, weakly-supervised pre-training provides a strong generalist model, while a Wav2Vec 2.0 model fine-tuned on a high-quality, domain-specific dataset can often achieve lower error rates for that specific domain or dialect. Recent comprehensive benchmarking efforts like the Open ASR Leaderboard [25] have systematically evaluated models like Whisper against other state-of-the-art systems, providing valuable insights into their capabilities on both Modern Standard and Dialectal Arabic. These benchmarks highlight that while pre-trained models have significantly advanced the field, performance can still vary widely depending on the specific dialect and acoustic conditions [13].

A critical aspect of these comparative studies is the choice of evaluation metric. For decades, the predominant metric has been the Word Error Rate (WER), which algorithmically computes a distance based on word-level substitutions, insertions, and deletions. However, the linguistic community has long critiqued WER's shortcomings for morphologically rich languages like Arabic. The tendency of WER to over-penalize minor morphological errors that may not impede human comprehension is a known issue [26]. In Arabic, for example, consider two potential errors in transcribing a spoken utterance. An error that changes kitābuhum ("their book") to kitābuhunna ("their [feminine] book") is a minor disagreement in grammatical gender. An error that changes it to maktabuhum ("their office") is a major semantic failure. Both scenarios could register as a single word error under the WER metric, despite their vastly different impact on meaning. This discrepancy illustrates that WER can be a poor proxy for semantic accuracy in Arabic [27].

The inadequacy of WER has motivated the use of semantically-oriented evaluation metrics. This research will therefore also employ Semantic Distance (SemDist), a metric that uses sentence embeddings from models like Sentence-BERT [28] to quantify the semantic similarity between an ASR hypothesis and a ground truth reference. The use of sentence-level semantic similarity has been explored in various Arabic NLP tasks, demonstrating its utility in capturing meaning beyond lexical overlap [29]. By representing sentences as vectors in a semantic space, SemDist evaluates the preservation of meaning, providing a useful complement to WER.

3. Methodology

Let X denote a set of Arabic speech utterances, each associated with a dialect label $d \in D$, where D is the set of dialects represented in the dataset. The objective is to evaluate the performance of state-of-the-art ASR models on transcribing X into text, and to assess whether post-processing these transcriptions with large language models (LLMs) can further optimize transcription quality. The evaluation is conducted using both Word Error Rate (WER) and Semantic Distance (SemDist) as metrics.

Formally, for each utterance $x \in X$, the ASR model f_{ASR} produces a hypothesis $h = f_{ASR}(x)$. The LLM post-processing step is defined as $h' = f_{LLM}(h)$. Both h and the final output h' are compared to the ground truth transcription y using WER and SemDist.

The experimental pipeline (figure 1) consists of the following stages:

1. ASR Inference: Each audio sample is transcribed using the benchmarked ASR models.
2. LLM Post-Processing: The ASR-generated hypotheses are passed to LLMs for error correction and refinement.
3. Evaluation: The outputs are evaluated against reference transcriptions using WER and SemDist.

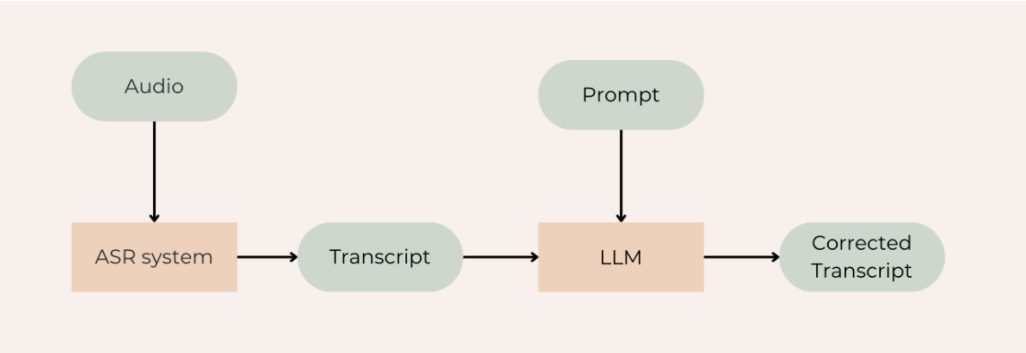


Fig. 1: Experimental Pipeline

This architecture enables a modular comparison of ASR models and the impact of LLM-based post-processing. Both the ASR models and LLMs are evaluated in a zero-shot setting, without any fine-tuning on the dataset or specific examples to guide the language model.

3.1 Dataset and pre-processing

The Saudi Audio Dataset for Arabic (SADA) [6] is a recently released benchmark that contains over 600 hours of audio covering multiple Arabic dialects. Each entry consists of an audio file, a ground-truth transcription, and metadata including dialect classification. All audio files were provided in WAV format with a sampling rate of 16 kHz and 16-bit PCM encoding. The dataset includes many Arabic dialects, mostly Saudi dialects, in addition to Egyptian, Yemeni, Levantine, besides Main Stream Arabic. The audio comes from television shows covering both spontaneous and speaking styles in diverse genres.

The dataset is split into training, validation, and test sets. For this study, we only use the test split of the SADA dataset. The test set contains 6193 audio segments. The average segment length in the set is 6.89 seconds while the median is 3.34. Figure 2 shows the speakers' genders and environment distribution, while Table 1 shows the dialects statistics of the test set in addition to calculated basic noise analytics. Out of 6193 entries in the test set, 134 contain "غير واضح" which means "unclear" in the transcription field. This means that these segments were not fully transcribed by the dataset providers. Such records are excluded from the evaluation set in the preprocessing step to guarantee that automatic speech recognition (ASR) performance metrics are computed only on utterances with fully verified reference transcriptions. To maintain consistency, transcripts are also normalized using Whisper's BasicTextNormalizer.

Table 1: SADA test set dialects

SpeakerDialect	*No. of audio segments*	*% of test set*	*Average SNR*	*Average clipping rate*
Najdi	1704	27.52%	13.32	0.0041
More than 1 speaker	1320	21.31%	13.58	0.0013
Khaliji	1150	18.57%	12.83	0.0034
Hijazi	809	13.06%	12.81	0.0035
Unknown	765	12.35%	11.71	0.0041
Notapplicable	167	2.70%	17.64	0.0042
ModernStandardArabic	157	2.54%	12.73	0.0017
Egyptian	96	1.55%	12.74	0.0034
Shamali	18	0.29%	15.61	0.0039
Yemini	7	0.11%	10.97	0.0036

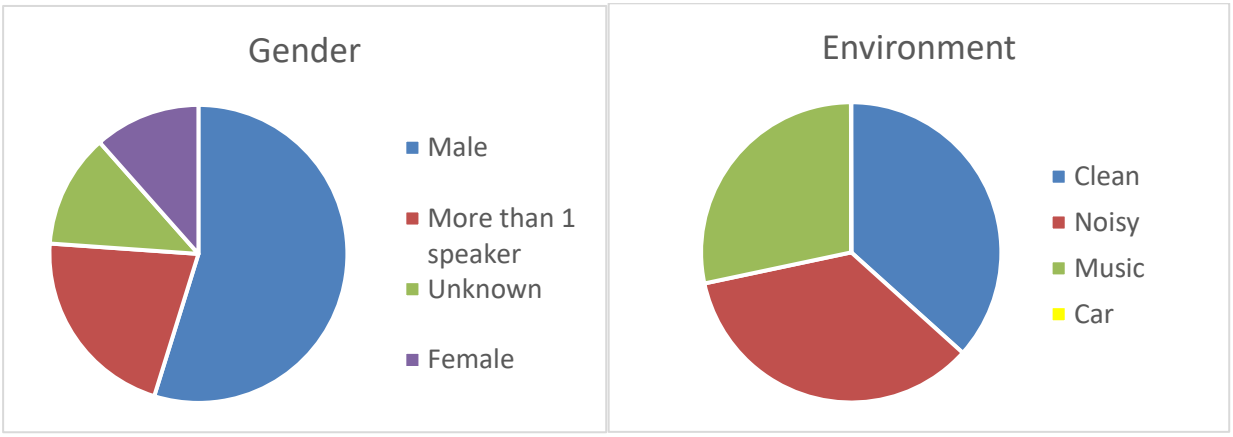


Fig. 2: Test set gender and environment distribution

3.2 ASR models

Two state-of-art ASR models are used in this study, namely, Whisper and Wav2vec. Whisper is an ASR system developed by OpenAI, designed to handle a wide range of languages and dialects. The architecture is based on an encoder-decoder Transformer, where input audio is first converted into a log-Mel spectrogram and then processed by the model to generate transcriptions. Whisper was trained on an extensive dataset exceeding 680,000 hours of multilingual and multitask supervised data, which includes a significant proportion of non-English audio. This large-scale, diverse training corpus enables Whisper to generalize robustly across languages, accents, and acoustic conditions, making it particularly effective for zero-shot transcription tasks [1]. Whisper is available in several pre-trained checkpoints of increasing size and capacity—tiny, base, small, medium, and large (including large-v2 and large-v3)—with each larger model offering improved transcription accuracy at the cost of greater computational requirements. Both the medium and large variants are applied in this study. Both models are multilingual and support direct transcription of Arabic speech, making them suitable for comparative studies involving dialectal Arabic.

Wav2Vec 2.0 is a self-supervised speech representation model developed by Facebook AI, which learns directly from raw audio waveforms. The jonatasgrosman/wav2vec2-large-xlsr-53-arabic model is a fine-tuned version of the large XLSR-53 checkpoint, specifically adapted for Arabic speech recognition. This model was trained using the Common Voice 6.1 and Arabic Speech Corpus datasets, allowing it to capture the phonetic and lexical characteristics of Arabic speech. The model operates at a 16kHz sampling rate. In its benchmark evaluations, the model achieved a WER of approximately 39.6% on the Common Voice Arabic test set.

3.3 LLMs

For the post processing of the transcribed audios, two large language models are employed. GPT (Generative Pre-trained Transformer) represents a family of large-scale language models developed by OpenAI, notable for their advanced natural language understanding and generation capabilities. GPT models are built on the Transformer architecture, leveraging self-attention mechanisms to capture complex contextual relationships within text. These models are pre-trained on vast and diverse text corpora, encompassing hundreds of billions of parameters, which enables them to generate highly coherent and contextually relevant responses to a wide range of prompts [30]. Recent versions, such as GPT-4, have further improved the model's ability to maintain context over longer interactions, handle multilingual content—including Arabic—and provide accurate outputs across a variety of applications. The model supports both zero-shot and in-context learning, making it appropriate for tasks such as ASR output correction.

The second model is ALLAM [31], which is a large language model specifically designed for Arabic language, addressing the unique linguistic characteristics of Arabic. Unlike general-purpose models such as GPT, ALLAM is trained predominantly on Arabic data, including Modern Standard Arabic and a wide spectrum of dialects, as well as English. Its architecture is based on the Transformer framework, similar to GPT, but with optimizations and pre-training strategies tailored for Arabic NLP tasks. The model demonstrates strong performance in a range of applications, including text generation, and summarization.

4. Experiments

4.1 Experiments setup

To evaluate the performance of automatic speech recognition (ASR) and language models on dialectal Arabic speech, we conducted a two-phase experimental setup. In the first phase, we compared the outputs of three ASR models—Whisper-medium, Whisper-large, and Wav2Vec 2.0—under three distinct conditions: (i) raw ASR output, (ii) ASR output corrected by GPT, and (iii) ASR output corrected by ALLAM. All models were evaluated in a zero-shot setting, without any fine-tuning or domain-specific adaptation. For each Arabic dialect under study (Khaliji, Najdi, Hijazi, Egyptian, and MSA), 50 audio segments were sampled from the test portion of the SADA dataset. These audio files were transcribed using each ASR model, and the resulting transcripts were either left unaltered (raw), or passed to an LLM for post-processing. Post-processing with GPT was carried out using GPT-4 via the OpenAI API (gpt-4 model), with carefully constructed prompts to encourage correction of standard

word forms while preserving dialectal meaning. For ALLAM, we used the ALLaM-7B model hosted on Hugging Face, accessed through the transformers library, and prompted similarly to correct the ASR output. This setup resulted in a total of nine system variants per dialect (three ASR models × three settings). To ensure fair comparison, evaluation metrics were computed for each system on normalized text.

In the second phase, the best-performing ASR and LLM pipeline from phase 1 was selected based on aggregate performance across dialects using two evaluation metrics: Word Error Rate (WER) and Semantic Distance (SemDist). The selected system was then applied to the entire test set of each dialect in the SADA corpus. This allowed us to conduct a more comprehensive evaluation of how well the optimal ASR+LLM configuration performs in transcribing dialectal Arabic speech and whether it generalizes effectively beyond the limited sample used in the first phase. WER was computed using standard Levenshtein distance (using python's jiwer library), while SemDist measured the semantic similarity between reference and hypothesis sentences using sentence embeddings and cosine distance.

4.2 Results and discussion

This section presents the evaluation results of ASR performance across five Arabic dialects using Whisper and Wav2Vec2 models, with and without post-processing by large language models (GPT and ALLAM). The results are analyzed in terms of Word Error Rate (WER) and Semantic Distance (SemDist) to assess both surface-level accuracy and semantic preservation.

Table 2: Phase 1 results

	WER					*SemDist*				
	Egy	Hjz	Klj	MSA	Njd	Egy	Hjz	Klj	MSA	Njd
Wav2Vec(raw)	0.96	0.93	0.92	0.88	0.94	0.39	0.42	0.40	0.46	0.35
Wav2Vec (GPT)	0.79	0.84	0.76	0.71	0.75	0.38	0.40	0.37	0.41	0.37
Wav2Vec (Allam)	0.88	0.91	0.98	0.84	0.89	0.43	0.43	0.46	0.48	0.41
Whisper-md (raw)	0.49	0.62	0.61	0.60	0.56	0.25	0.30	0.27	0.24	0.31
Whisper-md (GPT)	0.42	0.54	0.52	0.43	0.52	0.21	0.30	0.26	0.20	0.23
Whisper-md (Allam)	0.53	0.62	0.71	0.56	0.59	0.29	0.33	0.31	0.23	0.32
Whisper-lg (raw)	0.49	0.53	0.53	0.44	0.53	0.22	0.26	0.28	0.19	0.29
Whisper-lg (GPT)	**0.41**	**0.49**	**0.51**	**0.40**	**0.48**	**0.19**	**0.24**	**0.25**	**0.15**	**0.22**
Whisper-lg (Allam)	0.49	0.56	0.67	0.59	0.57	0.24	0.28	0.30	0.21	0.32

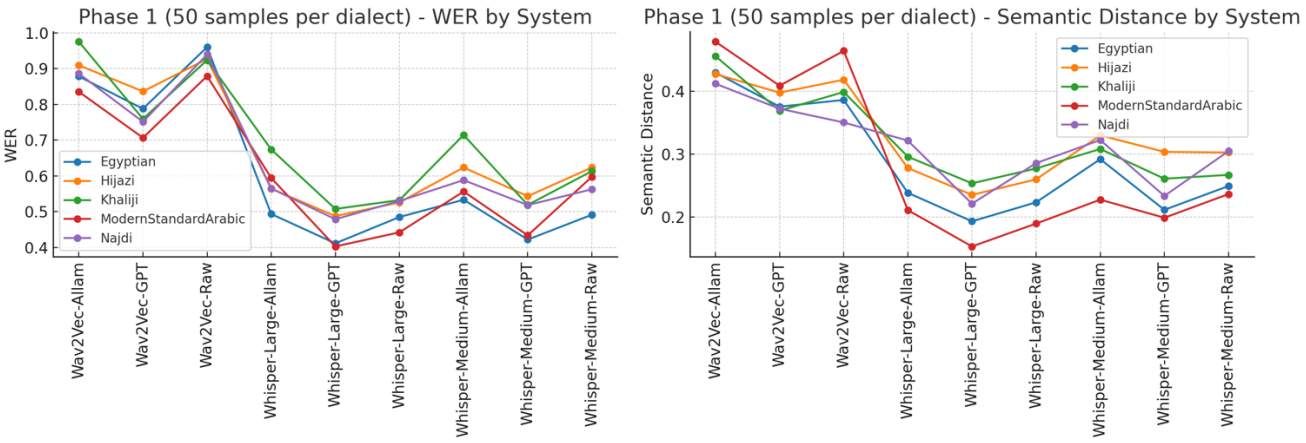


Fig. 3: WER and SemDist for phase 1

Results from the first phase are shown in table 2 and depicted in figure 3. They show that Whisper models consistently outperformed Wav2Vec2 across all dialects. In particular, Whisper-Large achieved the lowest Word Error Rate and Semantic Distance, outperforming Whisper-Medium and Wav2Vec2 in nearly all cases. This aligns with findings from Radford et al. [1], who demonstrated Whisper's robustness to noisy environments and multilingual input, partly due to its large-scale weakly supervised training on web-scale data. In contrast, Wav2Vec2, while strong in certain clean speech settings [2], appears less adapted to the pronunciation variability, loanwords, and informal constructions common in dialectal Arabic. The open universal Arabic ASR leaderboard [25] also shows that Whisper Large outperforms Whisper medium and Wav2Vec on SADA dataset.

Results also show that LLM post-processing appears to be consistently beneficial, with GPT showing more pronounced improvements than ALLAM. While ALLAM reduced WER in several cases, the gains were generally smaller than those achieved by GPT, and in a few instances, semantic distance was not improved. Notably, Wav2Vec2 benefited most from GPT correction, with large relative WER reductions. However, even after LLM correction, Wav2Vec2 lagged behind Whisper-based systems.

Dialectal variation also influenced results. Across all systems, MSA achieved the lowest error rates (best WER: 0.4031 with Whisper-Large-GPT), while other dialects such as Khaliji and Najdi consistently recorded higher WER values. This aligns with prior work [32] indicating that dialectal variability and phonological divergence from MSA in addition to resource scarcity introduce greater transcription challenges. Egyptian dialect also performed better than the other Saudi dialects. This may indicate underrepresentation in training data of the ASR systems.

The findings from Phase 1 strongly indicated Whisper-Large with GPT correction as the optimal configuration, offering the lowest WER and Semantic Distance results across all dialects. This informed the system selection for Phase 2.

Table 3: Phase 2 results

	Egyptian	*Hijazi*	*Khaliji*	*MSA*	*Najdi*
WER for Whisper-lg (raw)	0.59	0.53	0.56	0.32	0.49
WER for Whisper-lg (GPT-corrected)	0.52	0.49	0.51	0.35	0.48
SemDist for Whisper-lg (raw)	0.27	0.26	0.27	0.13	0.28
SemDist for Whisper-lg (GPT-corrected)	0.26	0.25	0.25	0.13	0.26

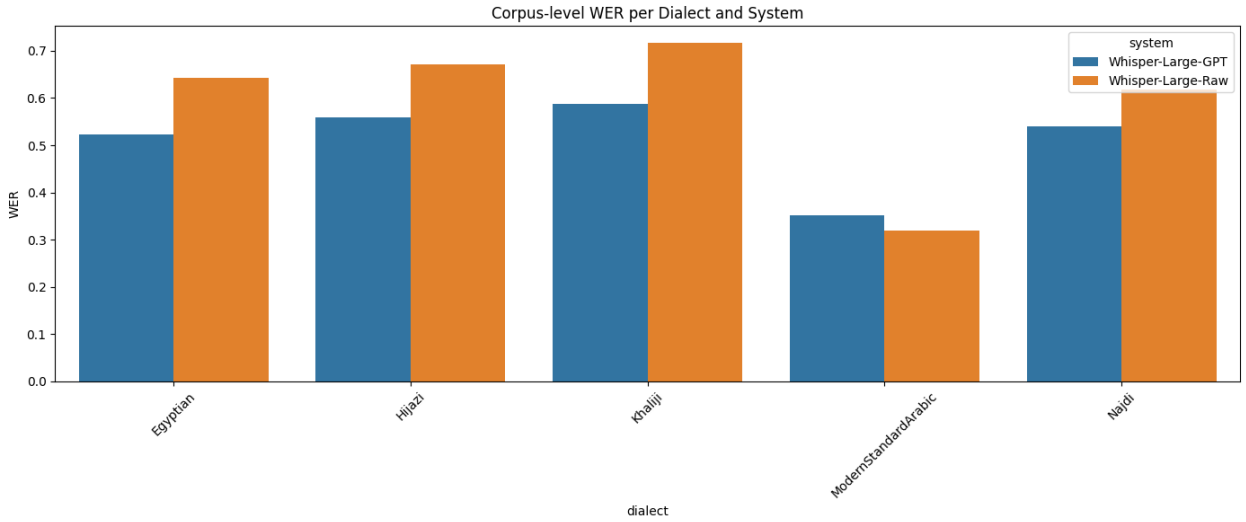


Fig. 4(a): WER for phase 2

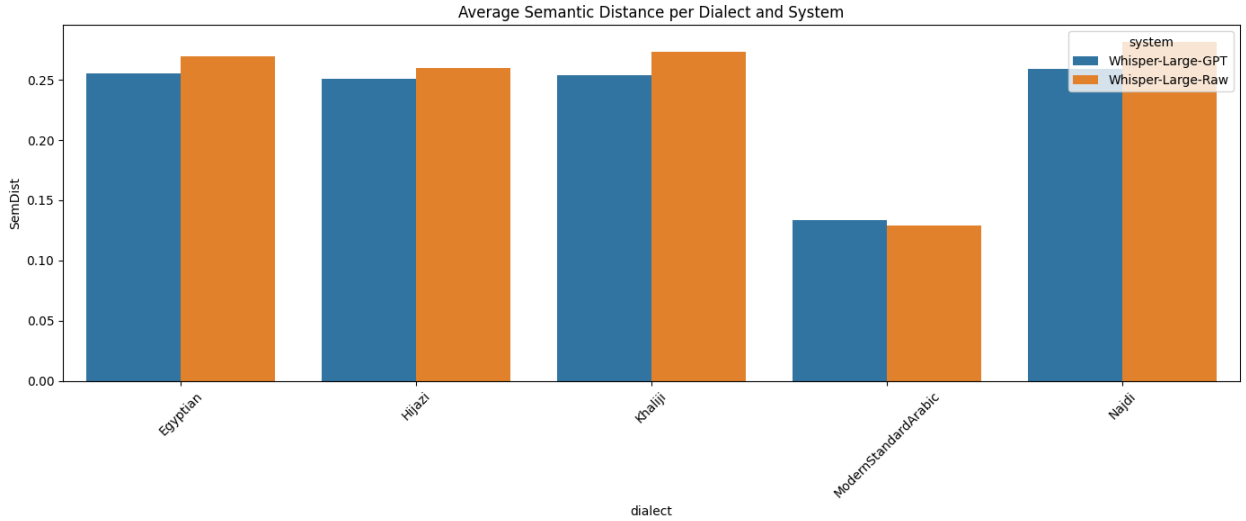


Fig. 4(b): SemDist for phase2

Phase 2 applied Whisper Large with GPT post-processing to the entire test set for each dialect, allowing for a more comprehensive performance evaluation, with Whisper-Large (raw) serving as a baseline. The initial results showed very high WER for some dialects datasets, and upon inspection, we found that a few audio segments were transcribed into repeated words which resulted into very

high WER (reaching over 2000 in some cases). We removed these outliers by setting a threshold of 20 for WER for all dialects.

To ensure statistical rigor, we evaluated system differences at the utterance level, reporting 95% bootstrap confidence intervals (CIs), conducting paired tests to avoid sample mismatch, and accompanying all p-values with effect sizes (Cliff's δ). We further applied Holm–Bonferroni correction across the set of dialect–metric comparisons to control for family-wise error.

The results confirmed the Phase 1 findings: GPT correction reduced WER consistently across all dialects except Modern Standard Arabic (MSA), while improvements in SemDist were smaller but statistically significant for Najdi and Khaliji.

- Egyptian dialect showed a substantial reduction in WER (mean difference = 0.183, CI [0.050, 0.400]), with a small effect size (δ = 0.25). The SemDist improvement was negligible and not statistically significant, indicating GPT mainly reduced literal transcription errors without altering semantic fidelity.
- Hijazi dialect demonstrated modest WER gains (mean difference = 0.084, CI [−0.026, 0.199]) that did not reach significance after Holm adjustment. SemDist showed negligible effects (δ = 0.05), consistent with Whisper-Large already producing semantically faithful transcripts.
- Najdi dialect benefited from consistent improvements in both metrics: WER decreased by 0.055 (CI [0.018, 0.127]) and SemDist by 0.022 (CI [0.010, 0.033]). Both differences were statistically significant with negligible-to-small effect sizes, suggesting GPT correction systematically cleaned up errors without altering meaning.
- Khaliji dialect also showed clear gains: WER decreased by 0.078 (CI [−0.016, 0.176]) with a small effect size (δ = 0.16), while SemDist improved by 0.020 (CI [0.007, 0.033]), significant after correction. This indicates GPT post-processing improved both literal accuracy and semantic similarity for this dialect.
- MSA diverged from the pattern: GPT slightly increased WER (−0.084, CI [−0.283, 0.092]) and produced negligible SemDist change. Effect sizes were small and non-significant, consistent with the intuition that GPT "correction" may introduce unnecessary modifications when the baseline system already performs well on standardized language.

Overall, these analyses strengthen the Phase 1 conclusion: GPT correction yields consistent WER reductions in most dialects, with statistically validated improvements in Najdi and Khaliji. However, SemDist effects remain small, suggesting that Whisper-Large already preserves meaning relatively well, and GPT mainly corrects surface-level transcription errors. Importantly, the paired tests and

effect size analysis show that gains are robust but modest, and their practical utility may depend on application context. For MSA, where baseline quality is already high, GPT may risk overcorrection—an observation consistent with Naderi et al. [33], who reported occasional semantic drift when applying generative LLM correction to already high-quality transcriptions.

To assess the robustness of the observed improvements, we stratified the evaluation by utterance duration and signal-to-noise ratio (SNR) bins. Figure 5 (duration) and figure 6 (SNR) report mean WER with 95% bootstrap confidence intervals for Whisper-Large (Raw) and Whisper-Large+GPT across all dialects.

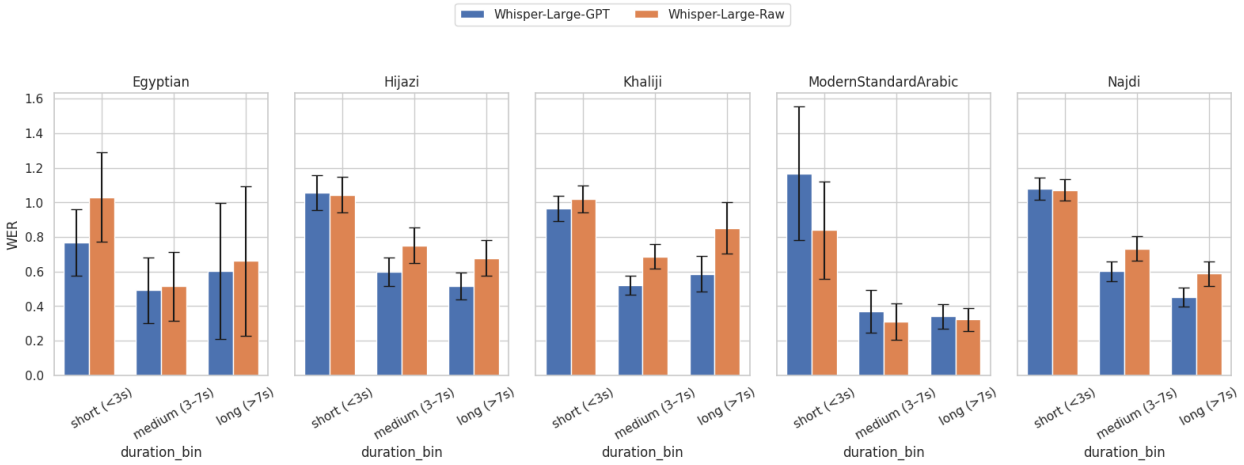


Fig. 5: WER by Duration Bin, Dialect, and System

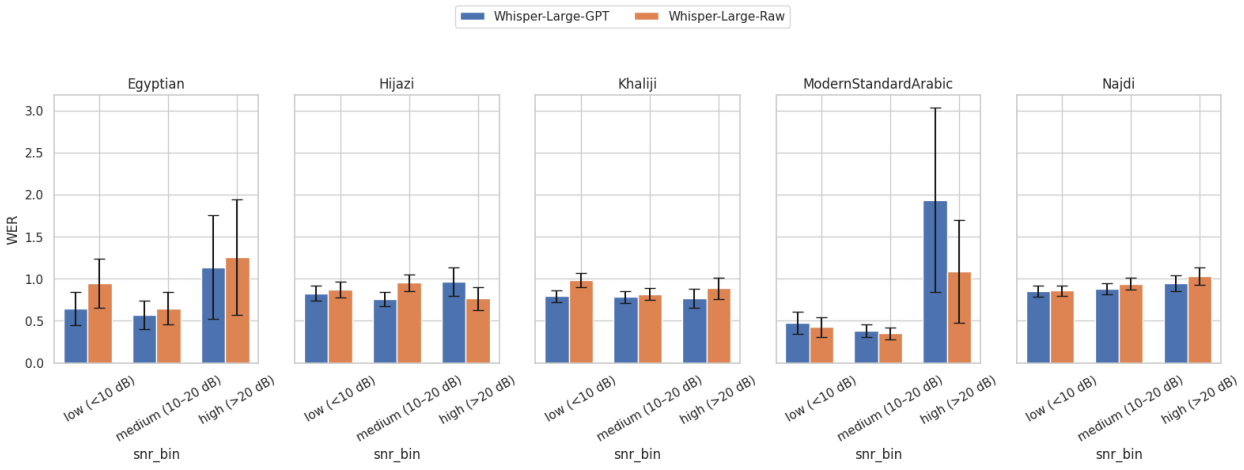


Fig. 6: WER by SNR Bin, Dialect, and System

When stratifying the results by utterance duration, we found that GPT post-processing did not provide uniform improvements across all bins. For short segments (<3s), gains were small and inconsistent: while Egyptian and Khaliji benefited modestly, MSA performance degraded, and Hijazi/Najdi showed negligible change. In contrast, for medium (3–7s) and long (>7s) segments, GPT consistently reduced

WER across dialects, with the largest relative improvements observed for Hijazi, Najdi, and Khaliji. This suggests that the LLM requires sufficient contextual information to apply meaningful corrections, and is less effective for short inputs where transcription errors may already be dominated by acoustic ambiguities rather than linguistic inconsistencies.

When stratifying by SNR, GPT also showed stable improvements in the low and medium ranges (<20 dB), confirming that it can mitigate some effects of noisy inputs. However, for high SNR (>20 dB), gains were smaller or inconsistent, with MSA showing degradation. This again indicates that LLM correction is most beneficial under challenging conditions, but can occasionally introduce unnecessary alterations when the base transcription is already relatively accurate.

The WER in all dialects apart from MSA can be considered relatively high, even after the possible gains from LLM post correction. Context-aware post-processing and few-shots LLM correction might increase the WER improvements. In addition, increasing the segments size could also provide more context to the LLM which may improve the results. Furthermore, as indicated in [34], dialect-pooled or fine-tuned models are needed to address resource scarcity.

5. Conclusion

This study presents a comprehensive evaluation of ASR performance for Arabic dialectal speech using Whisper and Wav2Vec2 models, supplemented by post-processing with large language models GPT and ALLAM. Our results demonstrate that Whisper, particularly in its larger variants, significantly outperforms Wav2Vec2 in zero-shot transcription across all tested dialects. However, both models show degradation in accuracy when transcribing non-standard or underrepresented dialects, such as Khaliji and Najdi, highlighting the limitations of current ASR systems in capturing the full linguistic diversity of Arabic.

Importantly, our findings confirm that large language models can meaningfully enhance ASR outputs. GPT, in particular, achieves reductions in both WER and Semantic Distance, indicating improvements not only in lexical correctness but also in semantic fidelity.

By combining quantitative and semantic evaluations, this study underscores the value of incorporating LLMs into ASR pipelines, especially for languages with rich dialectal variation. Future work could explore fine-tuning LLMs on dialectal corpora, integrating confidence-based editing mechanisms, and expanding evaluations to conversational or noisy speech scenarios. Ultimately, the integration of LLMs into ASR systems represents a promising direction for improving the accessibility and usability of Arabic speech technologies across dialects and domains.

Funding

The authors declare that this research received no external funding.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," OpenAI Technical Report, 2022.

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[3] Abdou and A. M. Moussa, "Arabic speech recognition: Challenges and state of the art," in *Computational Linguistics, Speech and Image Processing for Arabic Language*, 2019, pp. 1–27.

[4] A. Rahman, M. M. Kabir, M. F. Mridha, M. Alatiyyah, H. F. Alhasson, and S. S. Alharbi, "Arabic speech recognition: Advancement and challenges," *IEEE Access*, vol. 12, pp. 39689–39716, 2024.

[5] M. Afify, R. Sarikaya, H.-K. J. Kuo, L. Besacier, and Y. Gao, "On the use of morphological analysis for dialectal Arabic speech recognition," in *Proc. Interspeech 2006*, 2006, paper 1444-Mon2A2O.2, doi: 10.21437/Interspeech.2006-87.

[6] S. Alharbi *et al.*, "SADA: Saudi audio dataset for Arabic," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10286–10290.

[7] A. Ali *et al.*, "The MGB-2 challenge: Arabic multi-dialect broadcast media recognition," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 279–284, doi: 10.1109/SLT.2016.7846277.

[8] D. Wang, X. Wang, and S. Lv, "An overview of end-to-end automatic speech recognition," *Symmetry*, vol. 11, no. 8, p. 1018, 2019.

[9] N. Zerari, S. Abdelhamid, H. Bouzgou, and C. Raymond, "Bidirectional deep architecture for Arabic speech recognition," *Open Computer Science*, vol. 9, pp. 92–102, 2019.

[10] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Machine Learning (ICML)*, 2006, pp. 369–376.

[11] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.

[12] A. Waheed, B. Talafha, P. Sullivan, A. Elmadany, and M. Abdul-Mageed, "VoxArabica: A robust dialect-aware Arabic speech recognition system," in *Proc. ArabicNLP 2023 Workshop*, 2023, pp. 1–14.

[13] B. Talafha, A. Waheed, and M. Abdul-Mageed, "N-shot benchmarking of Whisper on diverse Arabic speech recognition," arXiv preprint, 2023.

[14] R. Sachdev, Z. Q. Wang, and C. H. H. Yang, "Evolutionary prompt design for LLM-based post-ASR error correction," arXiv preprint arXiv:2407.16370, 2024.

[15] R. Ma, M. Qian, M. Gales, and K. Knill, "ASR error correction using large language models," *IEEE Trans. Audio, Speech, Language Process.*, 2025.

[16] R. Ma, M. Qian, P. Manakul, M. Gales, and K. Knill, "Can generative large language models perform ASR error correction?" arXiv preprint arXiv:2307.04172, 2023.

[17] C. H. H. Yang, Y. Gu, Y. C. Liu, S. Ghosh, I. Bulyko, and A. Stolcke, "Generative speech recognition error correction with large language models and task-activating prompting," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8.

[18] Z. Min and J. Wang, "Exploring the integration of large language models into automatic speech recognition systems: An empirical study," in *Proc. Int. Conf. Neural Information Processing (ICONIP)*, Singapore, 2023, pp. 69–84.

[19] S. Li, C. Chen, C. Y. Kwok, C. Chu, E. S. Chng, and H. Kawai, "Investigating ASR error correction with large language model and multilingual 1-best hypotheses," in *Proc. Interspeech*, 2024, pp. 1315–1319.

[20] N. M. Matasyoh, R. A. Zeineldin, and F. Mathis-Ullrich, "Optimising speech recognition using LLMs: An application in the surgical domain," *Current Directions in Biomedical Engineering*, vol. 10, no. 1, pp. 45–48, Sep. 2024.

[21] S. Alharbi, R. Binmuqbil, A. Ali, R. Aloraini, S. Bari, A. Alowisheq, and Y. Alonaizan, "Leveraging LLM for augmenting textual data in code-switching ASR: Arabic as an example," in *Proc. SynData4GenAI*, 2024.

[22] T. Inoue, M. Elmahdy, A. Ali, and P. Bell, "Comparative study of end-to-end and hybrid ASR systems for Arabic broadcast speech," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 1085–1092.

[23] Y. Zhang, X. Wang, J. Xu, and H. Wang, "A comparative study of deep learning models for Arabic speech recognition," *IEEE Access*, vol. 10, pp. 10945–10953, 2022.

[24] S. Ding, A. Guo, A.-V. Nguyen, A. Vidon, J. Wang, and M. Zhou, "Data science capstone project: Robustness of Whisper and Wav2Vec2.0," 2023.

[25] Y. Wang, A. Alhmoud, and M. Alqurishi, "Open universal Arabic ASR leaderboard," arXiv preprint, 2024.

[26] A. C. Morris, V. Maier, and P. Green, "From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition," in *Proc. Interspeech*, 2004.

[27] A. Sahyoun and S. Shehata, "AraDiaWER: An explainable metric for dialectical Arabic ASR," in *Proc. 2nd Workshop on NLP Applications to Field Linguistics*, Dubrovnik, Croatia, 2023, pp. 64–73.

[28] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP/IJCNLP)*, 2019.

[29] M. Alian and A. Awajan, "Arabic semantic similarity approaches - review," in *Proc. Int. Arab Conf. Information Technology (ACIT)*, 2018, pp. 1–6, doi: 10.1109/ACIT.2018.8672665.

[30] T. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33, 2020, pp. 1877–1901.

[31] M. S. Bari *et al.*, "ALLaM: Large language models for Arabic and English," arXiv preprint, 2024.

[32] K. Almeman and M. Lee, "A comparison of Arabic speech recognition for multi-dialect vs. specific dialects," in *Proc. 7th Int. Conf. Speech Technology and Human-Computer Dialogue (SpeD)*, Cluj-Napoca, Romania, 2013.

[33] M. Naderi *et al.*, "Towards interfacing large language models with ASR systems using confidence measures and prompting," arXiv preprint arXiv:2407.19549, 2024.

[34] Ö. T. Özyilmaz, M. Coler, and M. Valdenegro-Toro, "Overcoming data scarcity in multi-dialectal Arabic ASR via Whisper fine-tuning," arXiv preprint arXiv:2506.02627, 2025.

Islamic University Journal of Applied Sciences (IUJAS)
https://journals.iu.edu.sa/jesc

Volume VIII, Issue I, July 2026, Pages 299-318

Approximation of a Common Solution to Split Equilibrium and Fixed Point Problems for a Nonexpansive Semigroup a Modified Viscosity Method

Shuja Haider Rizvi [1] *

[1] College of Science and Theoretical Studies, Saudi Electronic University, Riyadh 11673, Saudi Arabia
s.rizvi@seu.edu.sa, shujarizvi07@gmail.com

*Corresponding author: (Shuja Haider Rizvi), *Email Address*: s.rizvi@seu.edu.sa

Abstract

The main objective of this paper is to suggest and analyze a new iterative algorithm, grounded in modified viscosity techniques, for approximating a common solution of a split equilibrium problem and a fixed-point problem associated with a non-expansive semigroup in Hilbert spaces. We prove that the sequence generated by the proposed scheme converges strongly to a point that simultaneously solves the split equilibrium problem and the corresponding fixed-point problem for the non-expansive semigroup. Furthermore, the findings presented here offer improvements, generalizations, and refinements of results found in literature.

Keywords: Split equilibrium problem; Fixed-point problem; Non-expansive semigroup; Iterative algorithms.

https://doi.org/10.63070/jesc.2026.029
Received 15 February 2026; Revised 21 April 2026; Accepted 26 April 2026.
Available online 30 April 2026.
Published by Islamic University of Madinah on behalf of *Islamic University Journal of Applied Sciences*
This is an open access article published under the Creative Commons Attribution 4.0 International license (CC BY 4.0):https://creativecommons.org/licenses/by/4.0/

1 Introduction

In the sequel, let H_1 and H_2 be real Hilbert spaces equipped with the standard inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$. Let $C \subset H_1$ and $Q \subset H_2$ are nonempty closed and convex sets.

We consider the following classical equilibrium problem (in short EP) due to Blum and Oettli [1]: Find $x \in C$ such that

$$F(x,y) \geq 0, \ \forall y \in C, \tag{1.1}$$

where $F: C \times C \to \mathbb{R}$ is a bifunction.

The EP(1.1) encompasses a wide range of mathematical models, including variational inequality problems, optimization problems, Nash equilibrium problems, saddle point problems, fixed point problems, and complementarity problems as particular cases. Hence, EP(1.1) serves as a unified framework for modeling and analyzing many problems that arise in science, engineering, optimization, economics, and related fields.

Over the past two decades, the equilibrium problem EP(1.1) has been widely generalized and thoroughly investigated in various directions because of its fundamental importance. For results concerning the existence of solutions for different extensions of EP(1.1), we refer the reader to [2,4,3,5] and the references cited therein. Numerous iterative algorithms have also been developed for solving different classes of equilibrium problems; see, for instance, [6, 7, 14, 8, 9, 12, 13, 10, 11] and references cited therein.

However, in many practical situations, equilibrium problems in a system are defined in different spaces. Motivated by this, we focus in this paper on a pair of equilibrium problems formulated in distinct spaces, known as the split equilibrium problem (SEP), which was introduced and studied by Kazmi and Rizvi [15].

Let $F_1: C \times C \to \mathbb{R}$ and $F_2: Q \times Q \to \mathbb{R}$ be nonlinear bifunctions and $A: H_1 \to H_2$ be a bounded linear operator, then the *split equilibrium problem* SEP is to find $x^* \in C$ such that

$$F_1(x^*,x) \geq 0, \ \forall x \in C, \tag{1.2}$$

and such that

$$y^* = Ax^* \in Q \ \ solves \ F_2(y^*,y) \geq 0, \ \forall y \in Q. \tag{1.3}$$

EP(2.2) represents an equilibrium problem whose set of solutions is denoted by EP(F_1). The combined formulation (1.2)–(1.3) forms a pair of equilibrium problems that must be satisfied simultaneously. Specifically, one seeks a point x^* in the space H_1 that solves the equilibrium problem associated with (1.2), and whose image $y^* = Ax^*$, under a given bounded linear operator A solves another equilibrium problem defined in a second space H_2 as in (1.3). The solution set of the latter problem is denoted by EP(F_2). Accordingly, the set of solutions to the split equilibrium problem (1.2)–(1.3) is defined by $\Gamma = \{p \in \mathrm{E}P(F_1): Ap \in \mathrm{E}P(F_2)\}$.

The split equilibrium problem (1.2)–(1.3) can be viewed as a generalization of the multiple-sets split feasibility problem. Moreover, it encompasses, as a special case, the split variational inequality problem [16], which itself generalizes both split zero problems and split feasibility problems. For further details and related developments, we refer the reader to [20, 17, 21, 18, 19].

A family $S := \{T(s): 0 \leq s < \infty\}$ of mappings from C into itself is called *nonexpansive semigroup* on C if it satisfies the following conditions:

1. $T(0)x = x$ for all $x \in C$;
2. $T(s+t) = T(s)T(t)$ for all $s,t \geq 0$;
3. $\| T(s)x - T(s)y \| \leq \| x - y \|$ for all $x,y \in C$ and $s \geq 0$;

4. for all $x \in C$, $s \mapsto T(s)x$ is continuous.

The set of all the common fixed points of a family S is denoted by Fix(S), i.e.,

$$\mathrm{F}ix(S) := \{x \in C : T(s)x = x, 0 \le s < \infty\} = \bigcap_{0 \le s < \infty} \mathrm{F}ix(T(s)),$$

where $\mathrm{F}ix(T(s))$ is the set of fixed points of $T(s)$. It is well known that Fix(S) is closed and convex.

The *fixed point problem* (in short, FPP) for a nonexpansive semigroup S is:

$$\mathrm{F}ind\ x \in C\ such\ that\ x \in \mathrm{F}ix(S). \tag{1.4}$$

In 2006, Marino and Xu [24] considered the following implicit type iterative scheme for a nonexpansive mapping T:

$$x_t = t\gamma f(x_t) + (I - tP)Tx_t,$$

where f is a contraction mapping with constant α and $P: H_1 \to H_1$ is a strongly positive linear bounded operator, i.e., if there exists a constant $\bar{\gamma} > 0$ such that

$$\langle Px, x\rangle \ge \bar{\gamma} \parallel x \parallel^2, \quad \forall x \in H_1,$$

with $0 < \gamma < \frac{\bar{\gamma}}{\alpha}$ and $t \in (0,1)$ and proved that the net (x_t) converges strongly to the unique solution of the variational inequality

$$\langle (P - \gamma f)z, x - z\rangle \ge 0, \ \forall x \in \mathrm{F}ix(T),$$

which is the optimality condition for the minimization problem

$$\min_{x \in \mathrm{F}ix(T)} \frac{1}{2}\langle Px, x\rangle - h(x),$$

where h is the potential function for γf.

In 1997, Shimizu and Takahashi [25] introduced the following iterative method to prove the strong convergence theorem for FPP(1.4) in a real Hilbert space.

$$x_{n+1} = \alpha_n u + (1 - \alpha_n)\frac{1}{s_n}\int_0^{s_n} T(s)x_n ds, \quad \forall n \in \mathbb{N},$$

where $\{\alpha_n\}$ is a sequence in $(0,1)$ and $\{s_n\}$ is a sequence of positive real numbers which diverges to $+\infty$. Later, Chen and Song [26] introduced and studied the following iterative method to prove strong convergence theorem for FPP(1.4).

$$x_{n+1} = \alpha_n f(x_n) + (1 - \alpha_n)\frac{1}{s_n}\int_0^{s_n} T(s)x_n ds, \quad \forall n \in \mathbb{N}, \tag{1.5}$$

where f is a contraction mapping. Recently, Plubtieng and Punpaeng [27] introduced and studied the following iterative method to prove a strong convergence theorem for FPP(1.4).

$$x_{n+1} = \alpha_n f(x_n) + \beta_n x_n + (1 - \alpha_n - \beta_n)\frac{1}{s_n}\int_0^{s_n} T(s)x_n ds, \quad \forall n \in \mathbb{N}, \tag{1.6}$$

where $\{\alpha_n\}$ and $\{\beta_n\}$ are the sequences in $(0,1)$ and $\{s_n\}$ is a positive real divergent sequence.

In 2010, Cianciaruso *et al.* [28] introduced and studied the following explicit iterative method for finding the common element of EP(1.1) and FPP(1.4)

$$\begin{cases} F(u_n, y) + \frac{1}{r_n}\langle y - u_n, u_n - x_n\rangle, \forall y \in C; \\ x_{n+1} = \alpha_n \gamma f(x_n) + (I - \alpha_n P)\frac{1}{s_n}\int_0^{s_n} T(s)u_n ds, \end{cases} \tag{1.7}$$

where $\{\alpha_n\}$, $\{s_n\}$ and $\{r_n\}$ are the sequences in $(0,1)$, under the certain appropriate conditions on control sequences. They obtain the strong convergence theorem.

In 2014, Kazmi and Rizvi [15] introduced and studied an Halpern type iterative method to approximate the common solution of split equilibrium problem , variatinal inequality problem and FPP for a nonexpansive mapping T on C. They proved that under some mild conditions, the sequences generated by the proposed iterative algorithms converges strongly to a common solution of SEP, VIP and

FPP for a nonexpansive mapping.

In 2018, Rizvi [13] introduced and studied an iterative method by combining Mann and Halpern iterative method to approximate the common solution of split mixed equilibrium problem and FPP for a nonexpansive mapping T. He proved that under some mild conditions, the sequences generated by the proposed iterative algorithms converges strongly to a common solution of split mixed equilibrium problem and FPP for a nonexpansive mapping.

In 2023, Rizvi [29] introduced and studied an iterative method based on viscosity iterative method to approximate the common solution of split variational inclusion problem and FPP for a nonexpansive mapping T. Further, He proved that under some mild conditions, the sequences generated by the proposed iterative algorithms converges strongly to a common solution of split variational inclusion problem and FPP for a nonexpansive mapping. For related work, see [30] and references quoted therein.

In 2024, Rizvi and Sikander [31] introduced and studied an iterative method based on hybrid iterative method to approximate the common solution of split generalized mixed equilibrium problem and FPP for a nonexpansive mapping T. Further, He proved that under some mild conditions, the sequences generated by the proposed iterative algorithms converges strongly to a common solution of split generalized mixed equilibrium problem and FPP for a nonexpansive mapping.

It is noteworthy that the analysis of weak and strong convergence theorems for finding a common solution to the SEP $(1.2)-(1.3)$ and the FPP (1.4) associated with a semigroup of nonexpansive mappings remains largely unexplored in the existing literature. Consequently, establishing strong convergence results for a common of SEP $(1.2)-(1.3)$ and FPP (1.4) for a nonexpansive semigroup in this setting constitutes a compelling, interesting and significant research problem.

Motivated by the work of Shimizu and Takahashi [25], Chen and Song [26], Plubtieng and Punpaeng [27], Cianciaruso *et al.* [28], and by the ongoing research in this direction, we suggest and analyze an iterative method based on modified viscosity iterative method to approximate a common solution of SEP(1.2)-(1.3) and FPP(1.4) for a nonexpansive semigroup in real Hilbert spaces. Furthermore, we prove that the sequences generated by the iterative scheme converges strongly to a common solution of SEP(1.2)-(1.3) and FPP(1.4). The results presented in this paper can be treated as a generalization of the work of Plubtieng [27] and Cianciaruso *et al.* [28].

2 Preliminaries

We recall some definitions and results which will be useful for the presentation of subsequent sections.

Definition 2.1 [32] *For every point $x \in H_1$, there exists a unique nearest point in C denoted by $P_C x$ such that*

$$\| x - P_C x \| \leq \| x - y \|, \ \forall y \in C.$$

P_C is called the *metric projection* of H_1 onto C.

Remark 2.1 *It is well known fact that P_C is nonexpansive mapping and satisfies the following properties:*

$$\langle x - y, P_C x - P_C y \rangle \geq \| P_C x - P_C y \|^2, \ \forall x, y \in H_1. \tag{2.1}$$

Further, $P_C x$ is characterized by the fact that $P_C x \in C$ and

$$\langle x - P_C x, y - P_C x \rangle \leq 0, \tag{2.2}$$

and

$$\| x-y \|^2 \geq \| x-P_C x \|^2 + \| y-P_C x \|^2, \ \forall x \in H_1, y \in C. \quad (2.3)$$

Definition 2.2 [32] *In a real Hilbert space the following hold:*

1. $\| \lambda x+(1-\lambda)y \|^2 = \lambda \| x \|^2 + (1-\lambda) \| y \|^2 - \lambda(1-\lambda) \| x-y \|^2$, (2.4)
for all $x, y \in H_1$ and $\lambda \in [0,1]$.
2. (Opial's condition) [33]. For any sequence $\{x_n\}$ with $x_n \rightharpoonup x$ the inequality
$$\liminf_{n\to\infty} \| x_n - x \| < \liminf_{n\to\infty} \| x_n - y \|, \quad (2.5)$$
holds for every $y \in H_1$ with $y \neq x$.

Remark 2.2 *It is easy to observe that every nonexpansive operator $F: H_1 \to H_1$ satisfies, for all $(x, y) \in H_1 \times H_1$, the inequality*

$$\langle (x-F(x))-(y-F(y)), F(y)-F(x) \rangle \leq (1/2) \| (F(x)-x)-(F(y)-y \|^2, \quad (2.6)$$

therefore, we get, for all $(x, y) \in H_1 \times \mathrm{Fix}(F)$,

$$\langle x-F(x), y-F(x) \rangle \leq (1/2) \| F(x)-x \|^2, \quad (2.7)$$

see e.g., [[34], Theorem 2.1].

Lemma 2.1 [35] *Let $\{x_n\}$ and $\{y_n\}$ be bounded sequences in a Banach space X and $\{a_n\}$ be a sequence in $[0,1]$ with $0 < \liminf_{n\to\infty} a_n \leq \limsup_{n\to\infty} a_n < 1$. Suppose $x_{n+1} = (1-a_n)y_n + a_n x_n$, for all integers $n \geq 0$ and $\limsup_{n\to\infty} (\| y_{n+1}-y_n \| - \| x_{n+1}-x_n \|) \leq 0$. Then $\lim_{n\to\infty} \| y_n - x_n \| = 0$.*

Lemma 2.2 [32] *Assume that P is a strongly positive self-adjoint bounded linear operator on a Hilbert space H_1 with constant $\bar{\gamma} > 0$ and $0 < p \leq \| P \|^{-1}$. Then $\| I-\rho P \| \leq 1-\rho\bar{\gamma}$.*

Lemma 2.3 [25] *Let C be a nonempty, bounded, closed and convex subset of a Hilbert space H_1 and let $S := \{T(s): 0 \leq s < \infty\}$ be a nonexpansive semigroup on C. For each $x \in C$ and $t > 0$. Then for any $0 \leq h < \infty$,*

$$\limsup_{t\to\infty} \sup_{x\in C} \| \frac{1}{t}\int_0^t T(s)x\,ds - T(h)\left(\frac{1}{t}\int_0^t T(s)x\,ds\right) \| = 0.$$

Assumption 2.1 [1] *Let $F: C \times C \to \mathbb{R}$ be a bifunction satisfying the following assumptions:*

(i). $F(x,x) = 0, \ \forall x \in C$;
(ii) F is monotone, i.e., $F(x,y) + F(y,x) \leq 0, \ \forall x \in C$;
(iii) For each $x, y, z \in C$, $\limsup_{t\to 0} F(tz+(1-t)x, y) \leq F(x,y)$;
(iv) For each $x \in C$, $y \to F(x,y)$ is convex and lower semicontinuous.

Lemma 2.4 [14] *Assume that $F: C \times C \to \mathbb{R}$ satisfying Assumption* 3.1. *For $r > 0$ and for all $x \in H_1$, define a mapping $T_r^F: H_1 \to C$ as follows:*

$$T_r^F x = \{z \in C: F(z,y) + \frac{1}{r}\langle y-z, z-x \rangle \geq 0, \ \forall y \in C\}.$$

Then the following hold:

(i) T_r^F is single-valued;
(ii) T_r^F is firmly nonexpansive, i.e.,
$$\| T_r^F x - T_r^F y \|^2 \leq \langle T_r^F x - T_r^F y, x-y \rangle, \ \forall x, y \in H_1;$$
(iii) $\mathrm{Fix}(T_r^F) = Sol(EP(1.2))$;
(iv) Sol(EP(1.2)) is closed and convex.

Moreover, assume that $G: Q \times Q \to \mathbb{R}$ satisfying Assumption 2.1. For $s > 0$ and for all $w \in H_2$, define a mapping $T_s^G: H_2 \to Q$ as follows:

$$T_s^G w = \{d \in Q: \exists z \in T_r^F x \text{ such that } Az = d \text{ and } G(d,y) + \frac{1}{s}\langle d - e, e - w\rangle \geq 0, \forall e \in Q\}.$$

Then, we can easily observe that T_s^G is single-valued and firmly nonexpansive, Sol(EP(1.3)) is closed and convex and $\text{Fix}(T_s^G) = Sol(EP(1.3))$.

Lemma 2.5 [28] *Let $F: C \times C \to \mathbb{R}$ be a bifunction satisfying Assumption* 3.1 *hold and let T_r^F be defined as in Lemma* 2.4 *for $r > 0$. Let $x, y \in H_1$ and $r_1, r_2 > 0$. Then:*

$$\| T_{r_2}^F y - T_{r_1}^F x \| \leq \| y - x \| + \left|\frac{r_2 - r_1}{r_2}\right| \| T_{r_2}^F y - y \|.$$

Lemma 2.6 [36] *Let $(X, \langle\cdot,\cdot\rangle)$ be an inner product space, then for all $x, y \in X$ and $\alpha, \beta, \gamma \in [0,1]$ with $\alpha + \beta + \gamma = 1$, we have*

$$\| \alpha x + \beta y + \gamma z \|^2 = \alpha \| x \|^2 + \beta \| y \|^2 + \gamma \| z \|^2 - \alpha\beta \| x - y \|^2 - \alpha\gamma \| x - z \|^2 - \beta\gamma \| y - z \|^2.$$

Lemma 2.7 [37] *Let $\{a_n\}$ be a sequence of nonnegative real numbers such that*

$$a_{n+1} \leq (1 - \alpha_n)a_n + \delta_n, \quad n \geq 0,$$

where $\{\alpha_n\}$ is a sequence in $(0,1)$ and $\{\delta_n\}$ is a sequence in $\mathbb{R}$ such that

1. $\sum_{n=1}^{\infty} \alpha_n = \infty$;
2. $\limsup_{n\to\infty} \frac{\delta_n}{\alpha_n} \leq 0$ or $\sum_{n=1}^{\infty} |\delta_n| < \infty$.

Then $\lim_{n\to\infty} a_n = 0$.

3 Main Result

Theorem 3.1 *Let H_1 and H_2 be two real Hilbert spaces and let $C \subseteq H_1$ and $Q \subseteq H_2$ be nonempty closed convex subsets. Let $A: H_1 \to H_2$ be a bounded linear self adjoint operator. Assume that $F_1: C \times C \to \mathbb{R}$ and $F_2: Q \times Q \to \mathbb{R}$ are the bifunctions satisfying Assumption* 2.1 *and F_2 is upper semicontinuous in first argument. Let $S = \{T(s): 0 \leq s < \infty\}$ be a nonexpansive semigroup on C such that $Fix(S) \cap \Gamma \neq \emptyset$. Let $f: C \to C$ be a contraction mapping with constant $\alpha \in (0,1)$ and P be a strongly positive bounded linear operator on H_1 with constant $\bar{\gamma} > 0$ such that $0 < \gamma < \frac{\bar{\gamma}}{\alpha} < \gamma + \frac{1}{\alpha}$. Let $\{s_n\}$ is a positive real sequence which diverges to $+\infty$. For a given $x_0 \in C$ arbitrarily, let the iterative sequences $\{u_n\}$ and $\{x_n\}$ be generated by*

$$\begin{cases} u_n = T_{r_n}^{F_1}(x_n + \delta A^*(T_{r_n}^{F_2} - I)Ax_n); \\ x_{n+1} = \alpha_n \gamma f(x_n) + (I - \alpha_n P)\frac{1}{s_n}\int_0^{s_n} T(s)u_n ds, \end{cases} \tag{3.1}$$

where $r_n \subset (0, \infty)$ and $\delta \in (0, 1/L)$, L is the spectral radius of the operator A^*A and A^* is the adjoint of A and $\{\alpha_n\}$ and $\{\beta_n\}$ are the sequences in $(0,1)$ satisfying the following conditions:

1. $\lim_{n\to\infty} \alpha_n = 0$, $\sum_{n=0}^{\infty} \alpha_n = \infty$; and $\sum |\alpha_{n+1} - \alpha_n| < \infty$;
2. $\liminf r_n > 0$ and $\sum_{n=0}^{\infty} |r_{n+1} - r_n| < \infty$;
3. $\lim_{n\to\infty} \frac{|s_{n+1} - s_n|}{s_{n+1}} = 0$.

Then the sequence $\{x_n\}$ converges strongly to $z \in \text{F}ix(S) \cap \Gamma$, where $z = P_{\text{F}ix(S)\cap\Gamma}(I - P + \gamma f)z$.

Proof. We note that from condition (i), we may assume without loss of generality that $\alpha_n \leq (1-\beta_n) \parallel P \parallel^{-1}$ for all n. From Lemma 2.2, we know that if $0 < \rho \leq \parallel P \parallel^{-1}$, then $\parallel I - \rho P \parallel \leq 1 - \rho\bar{\gamma}$. We will assume that $\parallel I - P \parallel \leq 1 - \bar{\gamma}$.

Since P is positive linear bounded self-adjoint operator on H_1, then

$$\parallel P \parallel = \sup\{|\langle Pu, u\rangle| : u \in H_1, \parallel u \parallel = 1\}.$$

Observe that

$$\begin{aligned}\langle (I-\alpha_n P)u, u\rangle &= 1 - \alpha_n\langle Pu, u\rangle \\ &\geq 1 - \alpha_n \parallel P \parallel \\ &\geq 0,\end{aligned}$$

which implies that $(1-\alpha_n P)$ is positive. It follows that

$$\begin{aligned}\parallel (I-\alpha_n P \parallel &= \sup\{\langle((1-\alpha_n P)u, u\rangle : u \in H_1, \parallel u \parallel = 1\} \\ &= \sup\{1 - \alpha_n\langle Pu, u\rangle : u \in H_1, \parallel u \parallel = 1\} \\ &\leq 1 - \alpha_n\bar{\gamma}.\end{aligned}$$

Let $q = P_{\mathrm{Fix}(S)\cap\Gamma}$. Since f is a contraction mapping with constant $\alpha \in (0,1)$. It follows that

$$\begin{aligned}\parallel q(I-P+\gamma f)(x) - q(I-P+\gamma f)(y) \parallel &\leq \parallel (I-P+\gamma f)(x) - (I-P+\gamma f)(x) \parallel \\ &\leq \parallel I-P \parallel\parallel x-y \parallel + \gamma \parallel f(x) - f(y) \parallel \\ &\leq (1-\bar{\gamma}) \parallel x-y \parallel + \gamma\alpha \parallel x-y \parallel \\ &\leq (1-(\bar{\gamma}-\gamma\alpha)) \parallel x-y \parallel,\end{aligned}$$

for all $x, y \in H_1$. Therefore the mapping $q(I-P+\gamma f)$ is a contraction mapping from H_1 into itself. It follows from Banach contraction principle that there exists an element $z \in H_1$ such that $z = q(I-P+\gamma f)z = P_{\mathrm{Fix}(S)\cap\Gamma}(I-P+\gamma f)(z)$.

Let $p \in Fix(S) \cap \Gamma$, i.e., $p \in \Gamma$, we have $p = T_{r_n}^{F_1} p$ and $Ap = T_{r_n}^{F_2}(Ap)$.

We estimate

$$\begin{aligned}\parallel u_n - p \parallel^2 &= \parallel T_{r_n}^{F_1}(x_n + \delta A^*(T_{r_n}^{F_2} - I)Ax_n) - p \parallel^2 \\ &= \parallel T_{r_n}^{F_1}(x_n + \delta A^*(T_{r_n}^{F_2} - I)Ax_n) - T_{r_n}^{F_1} p \parallel^2 \\ &\leq \parallel x_n + \delta A^*(T_{r_n}^{F_2} - I)Ax_n - p \parallel^2 \\ &\leq \parallel x_n - p \parallel^2 + \delta^2 \parallel A^*(T_{r_n}^{F_2} - I)Ax_n \parallel^2 \\ &+ 2\delta\langle x_n - p, A^*(T_{r_n}^{F_2} - I)Ax_n\rangle. \qquad (3.2)\end{aligned}$$

Thus, we have

$$\begin{aligned}\parallel u_n - p \parallel^2 &\leq \parallel x_n - p \parallel^2 + \delta^2\langle(T_{r_n}^{F_2} - I)Ax_n, AA^*(T_{r_n}^{F_2} - I)Ax_n\rangle \\ &+ 2\delta\langle x_n - p, A^*(T_{r_n}^{F_2} - I)Ax_n\rangle. \qquad (3.3)\end{aligned}$$

Now, we have

$$\begin{aligned}&\delta^2\langle(T_{r_n}^{(F_2,h_2)} - I)Ax_n, AA^*(T_{r_n}^{F_2} - I)Ax_n\rangle \\ &\leq L\delta^2\langle(T_{r_n}^{F_2} - I)Ax_n, (T_{r_n}^{F_2} - I)Ax_n\rangle \\ &= L\delta^2 \parallel (T_{r_n}^{F_2} - I)Ax_n \parallel^2. \qquad (3.4)\end{aligned}$$

$Denoting\ \mathrm{F} = 2\delta\langle x_n - p, A^*(T_{r_n}^{F_2} - I)Ax_n\rangle$ and using (2.7), we have

$$\begin{aligned}\Lambda &= 2\delta\langle x_n - p, A^*(T_{r_n}^{F_2} - I)Ax_n\rangle \\ &= 2\delta\langle A(x_n - p), (T_{r_n}^{F_2} - I)Ax_n\rangle\end{aligned}$$

$$= 2\delta\langle A(x_n - p) + (T_{r_n}^{F_2} - I)Ax_n - (T_r^{F_2} - I)Ax_n, (T_{r_n}^{F_2} - I)Ax_n\rangle$$
$$= 2\delta\{\langle T_{r_n}^{F_2}Ax_n - Ap, (T_{r_n}^{F_2} - I)Ax_n\rangle - \| (T_{r_n}^{F_2} - I)Ax_n \|^2\}$$
$$\leq 2\delta\{\tfrac{1}{2} \| (T_{r_n}^{F_2} - I)Ax_n \|^2 - \| (T_{r_n}^{F_2} - I)Ax_n \|^2\}$$
$$\leq -\delta \| (T_{r_n}^{F_2} - I)Ax_n \|^2. \quad (3.5)$$

Using (3.3), (3.4) and (3.5), we obtain

$$\| u_n - p \|^2 \leq \| x_n - p \|^2 + \delta(L\delta - 1) \| (T_{r_n}^{F_2} - I)Ax_n \|^2. \quad (3.6)$$

Since, $\delta \in (0, \frac{1}{L})$, we obtain

$$\| u_n - p \|^2 \leq \| x_n - p \|^2. \quad (3.7)$$

Now setting $t_n := \frac{1}{s_n}\int_0^{s_n} T(s)u_n ds$ and since $p \in Fix(S) \cap \Gamma$, we obtain

$$\| t_n - p \| = \| \frac{1}{s_n}\int_0^{s_n} T(s)u_n ds - p \|$$
$$\leq \frac{1}{s_n}\int_0^{s_n} \| T(s)u_n - T(s)p \| ds$$
$$\leq \| u_n - p \|$$
$$\leq \| x_n - p \|. \quad (3.8)$$

Further, we estimate

$$\| x_{n+1} - p \| = \| \alpha_n\gamma f(x_n) + (I - \alpha_n P)t_n - p \|$$
$$= \| \alpha_n(\gamma f(x_n) - Pp) + (I - \alpha_n P)(t_n - p) \|$$
$$\leq \alpha_n \| \gamma f(x_n) - Pp \| + (1 - \alpha_n\bar{\gamma}) \| t_n - p \|$$
$$\leq \alpha_n\gamma \| f(x_n) - f(p) \| + \alpha_n \| \gamma f(p) - Pp \| + (1 - \alpha_n\bar{\gamma}) \| x_n - p \|$$
$$\leq \alpha_n\gamma\alpha \| x_n - p \| + \alpha_n \| \gamma f(p) - Pp \| + (1 - \alpha_n\bar{\gamma}) \| x_n - p \|$$
$$= (1 - (\bar{\gamma} - \gamma\alpha)\alpha_n) \| x_n - p \| + \alpha_n \| \gamma f(p) - Pp \|$$
$$\leq max\{\| x_n - p \|, \frac{1}{\bar{\gamma} - \gamma\alpha} \| \gamma f(p) - Pp \|\}, \quad n \geq 0$$
$$\vdots$$
$$\leq max\{\| x_0 - p \|, \frac{1}{\bar{\gamma} - \gamma\alpha} \| \gamma f(p) - Pp \|\}. \quad (3.9)$$

Hence $\{x_n\}$ is bounded and consequently, we deduce that $\{u_n\}$, $\{t_n\}$ and $\{f(x_n)\}$ are bounded.

Next, we estimate

$$\| t_{n+1} - t_n \| = \| \frac{1}{s_{n+1}}\int_0^{s_{n+1}} T(s)u_{n+1} ds - \frac{1}{s_n}\int_0^{s_n} T(s)u_n ds \|$$
$$= \| \frac{1}{s_{n+1}}\int_0^{s_{n+1}} [T(s)u_{n+1} - T(s)u_n] ds + (\frac{1}{s_{n+1}} - \frac{1}{s_n}) \times$$
$$\int_0^{s_n} T(s)u_n ds + \frac{1}{s_{n+1}}\int_{s_n}^{s_{n+1}} T(s)u_n ds \|$$
$$= \| \frac{1}{s_{n+1}}\int_0^{s_{n+1}} [T(s)u_{n+1} - T(s)u_n] ds + (\frac{1}{s_{n+1}} - \frac{1}{s_n}) \times$$
$$\int_0^{s_n} [T(s)u_n - T(s)p] ds + \frac{1}{s_{n+1}}\int_{s_n}^{s_{n+1}} [T(s)u_n - T(s)p] ds \|$$
$$\leq \| u_{n+1} - u_n \| + \frac{|s_{n+1} - s_n|s_n}{(s_{n+1})s_n} \| u_n - p \| + \frac{|s_{n+1} - s_n|}{s_{n+1}} \| u_n - p \|$$
$$\leq \| u_{n+1} - u_n \| + 2\frac{|s_{n+1} - s_n|}{s_{n+1}} \| u_n - p \|. \quad (3.10)$$

Since $u_n = T_{r_n}^{F_1}(x_n + \delta A^*(T_{r_n}^{F_2} - I)Ax_n)$ and $u_{n+1} = T_{r_{n+1}}^{F_1}(x_{n+1} + \delta A^*(T_{r_{n+1}}^{F_2} - I)Ax_{n+1})$.

It follows from Lemma 2.5 that

$$\begin{aligned}
&\| u_{n+1} - u_n \| \leq \| x_{n+1} - x_n + \delta[A^*(T_{r_{n+1}}^{F_2} - I)Ax_{n+1} - A^*(T_{r_n}^{F_2} - I)Ax_n] \| \\
&+ \left|1 - \frac{r_n}{r_{n+1}}\right| \| T_{r_{n+1}}^{F_1}(x_{n+1} + \delta A^*(T_{r_{n+1}}^{F_2} - I)Ax_{n+1}) \\
&-(x_{n+1} + \delta A^*(T_{r_{n+1}}^{F_2} - I)Ax_{n+1}) \| \\
&\leq \| x_{n+1} - x_n + \delta A^*A(x_{n+1} - x_n) \| + \delta \| A \| \| T_{r_{n+1}}^{F_2} Ax_{n+1} - T_{r_n}^{F_2,} Ax_n \| \\
&+\delta_n, \\
&\leq \{\| x_{n+1} - x_n \|^2 - 2\delta \| Ax_{n+1} - Ax_n \|^2 + \delta^2 \| A \|^4 \| x_{n+1} - x_n \|^2\}^{\frac{1}{2}} \\
&+\delta \| A \| \{\| Ax_{n+1} - Ax_n \| + \left|1 - \frac{r_n}{r_{n+1}}\right| \| T_{r_{n+1}}^{F_2)} Ax_{n+1} - Ax_{n+1} \|\} + \delta_n \\
&\leq (1 - 2\delta \| A \|^2 + \delta^2 \| A \|^4)^{\frac{1}{2}} \| x_{n+1} - x_n \| + \delta \| A \|^2 \| x_{n+1} - x_n \| \\
&+\delta \| A \| \sigma_n + \delta_n \\
&= (1 - \delta \| A \|^2) \| x_{n+1} - x_n \| + \delta \| A \|^2 \| x_{n+1} - x_n \| + \delta \| A \| \sigma_n + \delta_n, \\
&= \| x_{n+1} - x_n \| + \delta \| A \| \sigma_n + \delta_n, \qquad (3.11)
\end{aligned}$$

where

$$\sigma_n = \left|1 - \frac{r_n}{r_{n+1}}\right| \| T_{r_{n+1}}^{F_2} Ax_{n+1} - Ax_{n+1} \|$$

and

$$\delta_n = \left|1 - \frac{r_n}{r_{n+1}}\right| \| T_{r_{n+1}}^{F_1}(x_{n+1} + \delta A^*(T_{r_{n+1}}^{F_2} - I)Ax_{n+1}) - (x_{n+1} + \delta A^*(T_{r_{n+1}}^{F_2} - I)Ax_{n+1}) \|.$$

Using (3.10) and (3.11), we have

$$\| t_{n+1} - t_n \| \leq \| x_{n+1} - x_n \| + \delta \| A \| \sigma_n + \delta_n + 2\frac{|s_{n+1} - s_n|}{s_{n+1}} \| u_n - p \|. \qquad (3.12)$$

Next, we show that the sequence $\{x_n\}$ is asymptotically regular

$$\begin{aligned}
\| x_{n+1} - x_n \| &= \| \alpha_n \gamma f(x_n) + (1 - \alpha_n P)t_n - [\alpha_{n-1}\gamma f(x_{n-1}) + (1 - \alpha_{n-1}P)t_{n-1}] \| \\
&= \| \alpha_n \gamma f(x_n) - \alpha_n \gamma f(x_{n-1}) + \alpha_n \gamma f(x_{n-1}) - \alpha_{n-1}\gamma f(x_{n-1}) \\
&+(1 - \alpha_n P)t_n - (1 - \alpha_n P)t_{n-1} + (1 - \alpha_n` P)t_{n-1} - (1 - \alpha_{n-1}P)t_{n-1} \| \\
&\leq \alpha_n \gamma \alpha \| x_n - x_{n-1} \| + \gamma|\alpha_n - \alpha_{n-1}| \| f(x_{n-1}) \| + \| I - \alpha_n)P \| \| t_n - t_{n-1} \| \\
&+|\alpha_n - \alpha_{n-1}| \| Pt_{n-1} \|, \\
&\leq \alpha_n \gamma \alpha \| x_n - x_{n-1} \| \\
&+(1 + \gamma)|\alpha_n - \alpha_{n-1}|K_1 + (1 - \alpha_n \bar{\gamma}) \| t_n - t_{n-1} \| \qquad (3.13)
\end{aligned}$$

where $K_1 = \sup\{\| f(x_n) \| + \| Pt_n \| : n \in N\} < \infty$. It follows from (3.12) and (3.13) that

$$\begin{aligned}
\| x_{n+1} - x_n \| &\leq \alpha_n \gamma \alpha \| x_n - x_{n-1} \| + (1 + \gamma)|\alpha_n - \alpha_{n-1}|K_1 + (1 - \alpha_n \bar{\gamma})[\| x_n - x_{n-1} \| \\
&+\delta \| A \| \sigma_n + \delta_n + 2\frac{|s_n - s_{n-1}|}{s_n} \| u_{n-1} - p \|], \\
&= (1 - \alpha_n(\bar{\gamma} - \gamma\alpha)) \| x_n - x_{n-1} \| + (1 + \gamma)|\alpha_n - \alpha_{n-1}|K_1 \\
&+(1 - \alpha_n \bar{\gamma}) \left[\delta \| A \| \sigma_n + \delta_n + 2\frac{|s_n - s_{n-1}|}{s_n} \| u_{n-1} - p \|\right] \\
&\leq (1 - \alpha_n(\bar{\gamma} - \gamma\alpha)) \| x_n - x_{n-1} \| + (1 + \gamma)|\alpha_n - \alpha_{n-1}|K_1 \\
&+\delta \| A \| \sigma_n + \delta_n + 2\frac{|s_n - s_{n-1}|}{s_n} \| u_{n-1} - p \| \\
&= (1 - \alpha_n(\bar{\gamma} - \gamma\alpha)) \| x_n - x_{n-1} \| + (1 + \gamma)|\alpha_n - \alpha_{n-1}|K_1 \\
&+\delta \| A \| \left|1 - \frac{r_n}{r_{n+1}}\right| K_2 + \left|1 - \frac{r_n}{r_{n+1}}\right| K_3 + 2\frac{|s_n - s_{n-1}|}{s_n} K_4, \qquad (3.14)
\end{aligned}$$

where $K_2 = \sup\{\| (T_{r_n}^{F_2} - I)Ax_n \| : n \in N\}, K_3 = \sup\{\| u_n - x_n + \delta A^* K_2 \| : n \in N\}$ and $K_4 =$

$\sup\{\| u_n - p \|: n \in N\}$. Setting $K = \max\{K_1, K_2, K_3, K_4\}$. It follows from (3.14) that

$$\| x_{n+1} - x_n \| \leq (1 - \alpha_n(\bar{\gamma} - \gamma\alpha)) \| x_n - x_{n-1} \| + K\left[\left|1 - \frac{r_n}{r_{n+1}}\right| + 2\frac{|s_n - s_{n-1}|}{s_n} + (1+\gamma)|\alpha_n - \alpha_{n-1}|\right] \quad (3.15)$$

Hence by Lemma 2.7 and (i)-(iii), we have

$$\lim_{n\to\infty} \| x_{n+1} - x_n \| = 0. \quad (3.16)$$

Now, $x_{n+1} - x_n = \alpha_n\gamma f(x_n) + (I - \alpha_n P)t_n - x_n$
$= \alpha_n(\gamma f(x_n) - x_n) + (I - \alpha_n P)(t_n - x_n)$.

Since $\| x_{n+1} - x_n \| \to 0$ as $\alpha_n \to 0$ as $n \to \infty$, we obtain

$$\lim_{n\to\infty} \| t_n - x_n \| = 0. \quad (3.17)$$

Next, we have

$$\begin{aligned}
&\| T(s)x_n - x_n \| = \| T(s)x_n - T(s)\tfrac{1}{s_n}\textstyle\int_0^{s_n} T(s)u_n ds + T(s)\tfrac{1}{s_n}\int_0^{s_n} T(s)u_n ds \\
&-\tfrac{1}{s_n}\textstyle\int_0^{s_n} T(s)u_n ds + \tfrac{1}{s_n}\int_0^{s_n} T(s)u_n ds - x_n \| \\
&\leq \| T(s)x_n - T(s)\tfrac{1}{s_n}\textstyle\int_0^{s_n} T(s)u_n ds \| \\
&+ \| T(s)\tfrac{1}{s_n}\textstyle\int_0^{s_n} T(s)u_n ds - \tfrac{1}{s_n}\int_0^{s_n} T(s)u_n ds \| \\
&+ \| \tfrac{1}{s_n}\textstyle\int_0^{s_n} T(s)u_n ds - x_n \| \\
&\leq \| x_n - \tfrac{1}{s_n}\textstyle\int_0^{s_n} T(s)u_n ds \| \\
&+ \| T(s)\tfrac{1}{s_n}\textstyle\int_0^{s_n} T(s)u_n ds - \tfrac{1}{s_n}\int_0^{s_n} T(s)u_n ds \| \\
&+ \| \tfrac{1}{s_n}\textstyle\int_0^{s_n} T(s)u_n ds - x_n \| \\
&\leq 2 \| x_n - \tfrac{1}{s_n}\textstyle\int_0^{s_n} T(s)u_n ds \| \\
&+ \| T(s)\tfrac{1}{s_n}\textstyle\int_0^{s_n} T(s)u_n ds - \tfrac{1}{s_n}\int_0^{s_n} T(s)u_n ds \|. \quad (3.18)
\end{aligned}$$

Since $\{x_n\}$ and $\{f(x_n)\}$ is bounded. Let $K := \{w \in C: \| w - p \| \leq \| x_0 - p \|, \frac{1}{\bar{\gamma} - \gamma\alpha} \| \gamma f(p) - Pp \|\}$, then K is nonempty bounded closed convex subset of C which is $T(s)$-invariant for each $0 \leq s < \infty$ and contains $\{x_n\}$. So without loss of generality, we may assume that $S := \{T(s): 0 \leq s < \infty\}$ is nonexpansive semigroup on K. By Lemma 2.3, we have

$$\lim_{n\to\infty} \| T(s)\tfrac{1}{s_n}\textstyle\int_0^{s_n} T(s)u_n ds - \tfrac{1}{s_n}\int_0^{s_n} T(s)u_n ds \| = 0. \quad (3.19)$$

Using (3.10)-(3.12), we obtain

$$\lim_{n\to\infty} \| T(s)x_n - x_n \| = 0. \quad (3.20)$$

It follows from (3.6) and Lemma 2.6 (i) that

$$\begin{aligned}
&\| x_{n+1} - p \|^2 = \| \alpha_n\gamma f(x_n) + (I - \alpha_n P)t_n - p \|^2 \\
&= \| \alpha_n(\gamma f(x_n) - Pp) + (I - \alpha_n P)(t_n - p) \|^2 \\
&\leq \| (I - \alpha_n P)(t_n - p) \|^2 + 2\alpha_n\langle \gamma f(x_n) - Pp, x_{n+1} - p\rangle \\
&\leq (1 - \alpha_n\bar{\gamma})^2 \| t_n - p \|^2 + 2\alpha_n \| \gamma f(x_n) - Pp \| \| x_{n+1} - p \|
\end{aligned}$$

$$= (I - \alpha_n\bar{\gamma})^2 \| u_n - p \|^2 + 2\alpha_n \| \gamma f(x_n) - Pp \| \| x_{n+1} - p \| \quad (3.21)$$
$$\leq (1 - \alpha_n\bar{\gamma})^2[\| x_n - p \|^2 + \delta(L\delta - 1) \| (T_{r_n}^{F_2} - I)Ax_n \|^2]$$
$$+2\alpha_n \| \gamma f(x_n) - Pp \| \| x_{n+1} - p \|$$
$$\leq [1 - 2\alpha_n\bar{\gamma} + (\alpha_n\bar{\gamma})^2] \| x_n - p \|^2 + (1 - \alpha_n\bar{\gamma})^2\delta(L\delta - 1) \| (T_{r_n}^{F_2} - I)Ax_n \|^2$$
$$+2\alpha_n \| \gamma f(x_n) - Pp \| \| x_{n+1} - p \|$$
$$\leq \| x_n - p \|^2 + \alpha_n\bar{\gamma}^2 \| x_n - p \|^2 + (1 - \alpha_n\bar{\gamma})^2\delta(L\delta - 1) \| (T_{r_n}^{F_2} - I)Ax_n \|^2$$
$$+2\alpha_n \| \gamma f(x_n) - Pp \| \| x_{n+1} - p \|.$$

Therefore,

$$(1 - \alpha_n\bar{\gamma})^2\delta(1 - L\delta) \| (T_{r_n}^{F_2} - I)Ax_n \|^2$$
$$\leq \| x_n - p \|^2 - \| x_{n+1} - p \|^2 + \alpha_n\bar{\gamma}^2 \| x_n - p \|^2$$
$$+2\alpha_n \| \gamma f(x_n) - Pp \| \| x_{n+1} - p \|$$
$$\leq (\| x_n - p \| + \| x_{n+1} - p \|) \| x_n - x_{n+1} \| + \alpha_n\bar{\gamma}^2 \| x_n - p \|^2$$
$$+2\alpha_n \| \gamma f(x_n) - Pp \| \| x_{n+1} - p \|.$$

Since $\delta(1 - L\delta) > 0$, $\alpha_n \to 0$, $\| x_n - t_n \| \to 0$ and $\| x_{n+1} - x_n \| \to 0$ as $n \to \infty$, we obtain

$$\lim_{n\to\infty} \| (T_{r_n}^{F_2} - I)Ax_n \| = 0. \quad (3.22)$$

Next, we show that $\| x_n - u_n \| \to 0$ as $n \to \infty$. Since $p \in \mathrm{F}ix(S) \cap \Gamma$, we obtain

$$\| u_n - p \|^2 = \| T_{r_n}^{F_1}(x_n + \delta A^*(T_{r_n}^{F_2} - I)Ax_n) - p \|^2$$
$$= \| T_{r_n}^{F_1}(x_n + \delta A^*(T_{r_n}^{F_2} - I)Ax_n) - T_{r_n}^{F_1}p \|^2$$
$$\leq \langle u_n - p, x_n + \delta A^*(T_{r_n}^{F_2} - I)Ax_n - p \rangle$$
$$= \frac{1}{2}\{\| u_n - p \|^2 + \| x_n + \delta A^*(T_{r_n}^{F_2} - I)Ax_n - p \|^2$$
$$- \| (u_n - p) - [x_n + \delta A^*(T_{r_n}^{F_2} - I)Ax_n - p] \|^2\}$$
$$= \frac{1}{2}\{\| u_n - p \|^2 + \| x_n - p \|^2 - \| u_n - x_n - \delta A^*(T_{r_n}^{F_2} - I)Ax_n \|^2\}$$
$$= \frac{1}{2}\{\| u_n - p \|^2 + \| x_n - p \|^2 - [\| u_n - x_n \|^2$$
$$+\delta^2 \| A^*(T_{r_n}^{F_2} - I)Ax_n \|^2 - 2\delta\langle u_n - x_n, A^*(T_{r_n}^{F_2} - I)Ax_n \rangle]\}.$$

Hence, we obtain

$$\| u_n - p \|^2 \leq \| x_n - p \|^2 - \| u_n - x_n \|^2 + 2\delta \| A(u_n - x_n) \| \| T_{r_n}^{F_2} - I)Ax_n \|.$$

It follows from (3.21) and (3.22) that

$$\| x_{n+1} - p \|^2 \leq (1 - \alpha_n\bar{\gamma})^2[\| x_n - p \|^2 - \| u_n - x_n \|^2$$
$$+2\delta \| A(u_n - x_n) \| \| T_{r_n}^{F_2} - I)Ax_n \|]$$
$$+2\alpha_n \| \gamma f(x_n) - Pp \| \| x_{n+1} - p \|$$
$$\leq [1 - 2\alpha_n\bar{\gamma} + (\alpha_n\bar{\gamma})^2] \| x_n - p \|^2 - (1 - \alpha_n\bar{\gamma})^2 \| u_n - x_n \|^2$$
$$+2(1 - \alpha_n\bar{\gamma})^2\delta \| A(u_n - x_n) \| \| T_{r_n}^{F_2} - I)Ax_n \|$$
$$+2\alpha_n \| \gamma f(x_n) - Pp \| \| x_{n+1} - p \|$$
$$\leq \| x_n - p \|^2 + \alpha_n\bar{\gamma}^2 \| x_n - p \|^2 - (1 - \alpha_n\bar{\gamma})^2 \| u_n - x_n \|^2$$
$$+2(1 - \alpha_n\bar{\gamma})^2\delta \| A(u_n - x_n) \| \| T_{r_n}^{F_2} - I)Ax_n \|$$
$$+2\alpha_n \| \gamma f(x_n) - Pp \| \| x_{n+1} - p \|.$$

Therefore,

$$\begin{aligned}(1-\alpha_n\bar{\gamma})^2 \parallel u_n - x_n \parallel^2 &\leq \parallel x_n - p \parallel^2 - \parallel x_{n+1} - p \parallel^2 \\ &+2(1-\alpha_n\bar{\gamma})^2\delta \parallel A(u_n - x_n) \parallel \parallel T_{r_n}^{F_2} - I)Ax_n \parallel \\ &+\alpha_n\bar{\gamma}^2 \parallel x_n - p \parallel^2 \\ &+2\alpha_n \parallel \gamma f(x_n) - Pp \parallel \parallel x_{n+1} - p \parallel \\ &\leq (\parallel x_n - p \parallel + \parallel x_{n+1} - p \parallel) \parallel x_n - x_{n+1} \parallel \\ &+2(1-\alpha_n\bar{\gamma})^2\delta \parallel A(u_n - x_n) \parallel \parallel T_{r_n}^{F_2} - I)Ax_n \parallel \\ &+\alpha_n\bar{\gamma}^2 \parallel x_n - p \parallel^2 + 2(1-\alpha_n\bar{\gamma})\beta_n \parallel u_n - p \parallel \parallel x_n - t_n \parallel \\ &+2\alpha_n \parallel \gamma f(x_n) - Pp \parallel \parallel x_{n+1} - p \parallel.\end{aligned}$$

Since $\alpha_n \to 0$, $\parallel x_n - t_n \parallel \to 0$, $\parallel T_{r_n}^{F_2} - I)Ax_n \parallel \to 0$ and $\parallel x_{n+1} - x_n \parallel \to 0$ as $n \to \infty$, we obtain

$$\lim_{n\to\infty} \parallel u_n - x_n \parallel = 0. \tag{3.23}$$

Since, we can write

$$\begin{aligned}\parallel T(s)t_n - x_n \parallel &\leq \parallel T(s)t_n - T(s)x_n \parallel + \parallel T(s)x_n - x_n \parallel \\ &\leq \parallel t_n - x_n \parallel + \parallel T(s)x_n - x_n \parallel \\ &\to 0 \text{ as } n \to \infty.\end{aligned}$$

Also, we have

$$\begin{aligned}\parallel T(s)t_n - t_n \parallel &\leq \parallel T(s)t_n - T(s)x_n \parallel + \parallel T(s)x_n - x_n \parallel + \parallel x_n - t_n \parallel \\ &\leq \parallel t_n - x_n \parallel + \parallel T(s)x_n - x_n \parallel + \parallel x_n - t_n \parallel \\ &\to 0 \text{ as } n \to \infty.\end{aligned}$$

Next, we show that $\limsup_{n\to\infty}\langle (P-\gamma f)z, x_n - z\rangle \leq 0$, where $z = P_{\mathrm{F}ix(S)\cap\Gamma}(I - P + \gamma f)z$. To show this inequality, we choose a subsequence $\{t_{n_i}\}$ of $\{t_n\} \subseteq K$ such that

$$\limsup_{n\to\infty}\langle (P-\gamma f)z, t_n - z\rangle = \lim_{i\to\infty}\langle (P-\gamma f)z, t_{n_i} - z\rangle.$$

Since $\{t_{n_i}\}$ is bounded, there exists a subsequence $\{t_{n_{i_j}}\}$ of $\{t_{n_i}\}$ which converges weakly to some $w \in C$. Without loss of generality, we can assume that $t_{n_i} \rightharpoonup w$.

Now, we prove that $w \in \mathrm{F}ix(S) \cap \Gamma$. Let us first show that $w \in \mathrm{F}ix(S)$. Assume that $w \notin \mathrm{F}ix(S)$. Since $t_{n_i} \rightharpoonup w$ and $T(s)w \neq w$. From Opial's condition (2.5), we have

$$\begin{aligned}\liminf_{i\to\infty} \parallel t_{n_i} - w \parallel &< \liminf_{i\to\infty} \parallel t_{n_i} - T(s)w \parallel \\ &\leq \liminf_{i\to\infty}\{\parallel t_{n_i} - T(s)t_{n_i} \parallel + \parallel T(s)t_{n_i} - T(s)w \parallel\} \\ &\leq \liminf_{i\to\infty} \parallel t_{n_i} - w \parallel,\end{aligned}$$

which is a contradiction. Thus, we obtain $w \in \mathrm{F}ix(S)$.

Next, we show that $w \in \mathrm{E}P(F_1)$. Since $u_n = T_{r_n}^{F_1}x_n$, we have

$$F_1(u_n, y) + \frac{1}{r_n}\langle y - u_n, u_n - x_n\rangle \geq 0, \ \forall y \in C.$$

It follows from monotonicity of F_1 that

$$\frac{1}{r_n}\langle y - u_n, u_n - x_n\rangle \geq F_1(y, u_n)$$

and hence

$$\left\langle y - u_{n_i}, \frac{u_{n_i} - x_{n_i}}{r_n}\right\rangle \geq F_1(y, u_{n_i}).$$

Since $\parallel u_n - x_n \parallel \to 0$ we get $u_{n_i} \rightharpoonup w$ and $\frac{u_{n_i} - x_{n_i}}{r_n} \to 0$. It follows by Assumption 2.1 (iv) that $0 \geq F_1(y, w), \forall w \in C$. For t with $0 < t \leq 1$ and $y \in C$, let $y_t = ty + (1-t)w$. Since $y \in C$, $w \in C$,

we get $y_t \in C$ and hence $F_1(y_t, w) \leq 0$. So from Assumption 2.1 (i) and (iv) we have

$$0 = F_1(y_t, y_t) \leq tF_1(y_t, y) + (1-t)F_1(y_t, w)$$
$$\leq tF_1(y_t, y).$$

Therefore $0 \leq F_1(y_t, y)$. From Assumption 2.1 (iii), we have $0 \leq F_1(w, y)$. This implies that $w \in \mathrm{E}P(F_1)$.

Next, we show that $w \in \mathrm{E}P(F_2)$. Since $\| u_n - x_n \| \to 0, u_n \rightharpoonup w$ as $n \to \infty$ and $\{x_n\}$ is bounded, there exists a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ such that $x_{n_k} \rightharpoonup w$ and since A is a bounded linear operator so that $Ax_{n_k} \rightharpoonup Aw$.

Now setting $v_{n_k} = Ax_{n_k} - T_{r_{n_k}}^{F_2} Ax_{n_k}$. It follows that from (3.22) that $\lim_{k\to\infty} v_{n_k} = 0$ and $Ax_{n_k} - v_{n_k} = T_{r_{n_k}}^{F_2} Ax_{n_k}$.

Therefore from Lemma 2.4, we have

$$F_2(Ax_{n_k} - v_{n_k}, z) + \frac{1}{r_{n_k}} \langle z - (Ax_{n_k} - v_{n_k}), (Ax_{n_k} - v_{n_k}) - Ax_{n_k} \rangle \geq 0, \forall z \in Q.$$

Since F_2 is upper semicontinuous in first argument, taking $\liminf$ to above inequality as $k \to \infty$ and using condition (iv), we obtain

$$F_2(Aw, z) \geq 0, \forall z \in Q,$$

which means that $Aw \in \mathrm{E}P(F_2)$ and hence $w \in \Gamma$.

Next, we claim that $\limsup_{n\to\infty} \langle f(z) - z, x_n - z \rangle \leq 0$, where $z = P_{\mathrm{F}ix(S)\cap\Gamma}(I - P + \gamma f)z$. Now from (2.2), we have

$$\limsup_{n\to\infty} \langle (P - \gamma f)z - z, x_n - z \rangle = \limsup_{n\to\infty} \langle (P - \gamma f)z - z, t_n - z \rangle$$
$$\leq \limsup_{i\to\infty} \langle (P - \gamma f)z - z, t_{n_i} - z \rangle$$
$$= \langle (P - \gamma f)z - z, w - z \rangle$$
$$\leq 0. \qquad (3.24)$$

Finally, we show that $x_n \to z$.

$$\| x_{n+1} - z \|^2 = \| \alpha_n \gamma f(x_n) + [(I - \alpha_n P)t_n - z \|^2$$
$$= \| \alpha_n(\gamma f(x_n) - Pz) + (I - \alpha_n P)(t_n - z) \|^2$$
$$\leq \| (I - \alpha_n P)(t_n - z) \|^2 + 2\alpha_n \langle \gamma f(x_n) - Pz, x_{n+1} - z \rangle$$
$$\leq (1 - \alpha_n \bar{\gamma})^2 \| t_n - z \|^2 + 2\alpha_n \gamma \langle f(x_n) - f(z), x_{n+1} - z \rangle$$
$$+ 2\alpha_n \langle \gamma f(z) - Pz, x_{n+1} - z \rangle$$
$$\leq (1 - \alpha_n \bar{\gamma})^2 \| x_n - z \|^2 + 2\alpha_n \gamma \alpha \| x_n - z \| \| x_{n+1} - z \|$$
$$+ 2\alpha_n \langle \gamma f(z) - Pz, x_{n+1} - z \rangle$$
$$\leq (1 - \alpha_n \bar{\gamma})^2 \| x_n - z \|^2 + \alpha_n \gamma \alpha \{ \| x_n - z \|^2 + \| x_{n+1} - z \|^2 \}$$
$$+ 2\alpha_n \langle \gamma f(z) - Pz, x_{n+1} - z \rangle$$
$$\leq (1 - \alpha_n \bar{\gamma})^2 \| x_n - z \|^2 + \alpha_n \gamma \alpha \| x_n - z \|^2 + \gamma \alpha \| x_{n+1} - z \|^2$$
$$+ 2\alpha_n \langle \gamma f(z) - Pz, x_{n+1} - z \rangle.$$

This implies that

$$\| x_{n+1} - z \|^2 \leq \frac{1 - 2\alpha_n \bar{\gamma} + (\alpha_n \bar{\gamma})^2 + \alpha_n \gamma \alpha}{1 - \gamma \alpha} \| x_n - z \|^2$$
$$+ \frac{2\alpha_n}{1 - \gamma \alpha} \langle \gamma f(z) - Pz, x_{n+1} - z \rangle$$
$$= \left[1 - \frac{2(\bar{\gamma} - \gamma \alpha)\alpha_n}{1 - \gamma \alpha}\right] \| x_n - z \|^2 + \frac{(\alpha_n \bar{\gamma})^2}{1 - \gamma \alpha} \| x_n - z \|^2$$

$$+\frac{2\alpha_n}{1-\gamma\alpha}\langle \gamma f(z)-Pz, x_{n+1}-z\rangle$$
$$\leq [1-\frac{2(\bar{\gamma}-\gamma\alpha)\alpha_n}{1-\gamma\alpha}] \parallel x_n-z \parallel^2+\frac{2(\bar{\gamma}-\gamma\alpha)\alpha_n}{1-\gamma\alpha}$$
$$\times \{\frac{(\alpha_n\bar{\gamma}^2)M}{2(\bar{\gamma}-\gamma\alpha)}+\frac{1}{\bar{\gamma}-\gamma\alpha}\langle \gamma f(z)-Pz, x_{n+1}-z\rangle\}$$
$$=(1-\delta_n) \parallel x_n-z \parallel^2+\delta_n\sigma_n, \quad (3.25)$$

where $M:=\sup\{\parallel x_n-z \parallel^2: n\geq 1\}$, $\delta_n=\frac{2(\bar{\gamma}-\gamma\alpha)\alpha_n}{1-\gamma\alpha}$ and $\sigma_n=\frac{(\alpha_n\bar{\gamma}^2)M}{2(\bar{\gamma}-\gamma\alpha)}+\frac{1}{\bar{\gamma}-\gamma\alpha}\langle \gamma f(z)-Pz, x_{n+1}-z\rangle$. Since $\lim_{n\to\infty}\alpha_n=0$ and $\sum_{n=0}^{\infty}\alpha_n=\infty$, it is easy to see that $\lim_{n\to\infty}\delta_n=0$, $\sum_{n=0}^{\infty}\delta_n=\infty$ and $\limsup_{n\to\infty}\sigma_n\leq 0$. Hence from (3.24), (3.25) and Lemma 2.7, we deduce that $x_n\to z$. This completes the proof.

We have the following consequences of Theorem 3.1.

Corollary 3.1 [28] *Let H be a real Hilbert space and let $C\subseteq H$ be nonempty closed convex subset. Assume that $F: C\times C\to\mathbb{R}$ be a bifunction satisfying Assumption* 2.1 *for F only. Let $S=\{T(s): 0\leq s<\infty\}$ be a nonexpansive semigroup on C such that $Fix(S)\cap EP(F)=\emptyset$. Let $f: C\to C$ be a contraction mapping with constant $\alpha\in(0,1)$ and P be a strongly positive bounded linear operator on H with constant $\bar{\gamma}>0$ such that $0<\gamma<\frac{\bar{\gamma}}{\alpha}<\gamma+\frac{1}{\alpha}$. Let $\{s_n\}$ is a positive real sequence which diverges to $+\infty$. For a given $x_0\in C$ arbitrarily, let the iterative sequences $\{u_n\}$ and $\{x_n\}$ be generated by*

$$\begin{cases} u_n=T_{r_n}^F x_n; \\ x_{n+1}=\alpha_n\gamma f(x_n)+(1-\alpha_n P)\frac{1}{s_n}\int_0^{s_n} T(s)u_n ds, \end{cases}$$

where $r_n\subset(0,\infty)$ and $\{\alpha_n\}$ is a sequences in $(0,1)$ satisfying

(i) $\lim_{n\to\infty}\alpha_n=0$ and $\sum_{n=0}^{\infty}\alpha_n=\infty$;

(ii) $\liminf_{n\to\infty} r_n>0$, $\sum_{n=1}^{\infty}|r_{n+1}-r_n|<\infty$;

(iii) $\lim_{n\to\infty}\frac{|s_{n+1}-s_n|}{s_{n+1}}=0$.

Then the sequence $\{x_n\}$ converges strongly to $z\in \mathrm{F}ix(S)\cap \mathrm{E}P(F)$, where $z=P_{\mathrm{F}ix(S)\cap \mathrm{E}P(F)}(I-P+\gamma f)z$.

Proof. Taking $F_1=F_2=F$, $H_1=H_2=H$ and $A=0$ in Theorem 3.1 then the conclusion of Corollary 3.1 is obtained.

Corollary 3.2 *Let H be a real Hilbert space and let $C\subseteq H$ be a nonempty closed convex subset. Let $S=\{T(s): 0\leq s<\infty\}$ be a nonexpansive semigroup on C such that $Fix(S)\neq\emptyset$. Let $f: C\to C$ be a contraction mapping with constant $\alpha\in(0,1)$. Let $\{s_n\}$ is a positive real sequence which diverges to $+\infty$. For a given $x_0\in C$ arbitrarily, let the iterative sequence $\{x_n\}$ be generated by*

$$x_{n+1}=\alpha_n\gamma f(x_n)+(1-\alpha_n P)\frac{1}{s_n}\int_0^{s_n} T(s)x_n ds,$$

where $\{\alpha_n\}$ be a sequence in $(0,1)$ satisfying the following conditions:

1. $\lim_{n\to\infty}\alpha_n=0$ and $\sum_{n=0}^{\infty}\alpha_n=\infty$;
2. $\lim_{n\to\infty}\frac{|s_{n+1}-s_n|}{s_{n+1}}=0$.

Then the sequence $\{x_n\}$ converges strongly to $z\in \mathrm{F}ix(S)$, where $z=P_{\mathrm{F}ix(S)}f(z)$.

Proof. Taking $H_1 = H_2 = H$, $u_n = x_n$, and $F_1 = F_2 = 0$ in Theorem 3.1 then the conclusion of Corollary 3.2 is obtained.

Corollary 3.3 *Let H_1 and H_2 be two real Hilbert spaces and let $C \subseteq H_1$ and $Q \subseteq H_2$ be nonempty closed convex subsets. Let $A: H_1 \to H_2$ be a bounded linear self adjoint operator. Assume that $F_1: C \times C \to \mathbb{R}$ and $F_2: Q \times Q \to \mathbb{R}$ are the bifunctions satisfying Assumption* 2.1 *and F_2 is upper semicontinuous in first argument. Let T be a nonexpansive mapping on C such that $Fix(T) \cap \Gamma \neq \emptyset$. Let $f: C \to C$ be a contraction mapping with constant $\alpha \in (0,1)$ and P be a strongly positive bounded linear operator on H_1 with constant $\bar{\gamma} > 0$ such that $0 < \gamma < \frac{\bar{\gamma}}{\alpha} < \gamma + \frac{1}{\alpha}$. For a given $x_0 \in C$ arbitrarily, let the iterative sequences $\{u_n\}$ and $\{x_n\}$ be generated by*

$$\begin{cases} u_n = T_{r_n}^{F_1}(x_n + \delta A^*(T_{r_n}^{F_2} - I)Ax_n); \\ x_{n+1} = \alpha_n \gamma f(x_n) + (1 - \alpha_n P)Tu_n, \end{cases}$$

where $r_n \subset (0, \infty)$ and $\delta \in (0, 1/L)$, L is the spectral radius of the operator A^*A and A^* is the adjoint of A and $\{\alpha_n\}$ is a sequence in $(0,1)$ satisfying the following conditions:

(i) $\lim_{n\to\infty} \alpha_n = 0$, $\sum_{n=0}^{\infty} \alpha_n = \infty$; and $\sum_{n=1}^{\infty} |\alpha_{n+1} - \alpha_n| < \infty$;.

(ii) $\liminf_{n\to\infty} r_n > 0$, $\sum_{n=1}^{\infty} |r_{n+1} - r_n| < \infty$;.

Then the sequence $\{x_n\}$ converges strongly to $z \in \mathrm{F}ix(T) \cap \Gamma$, where $z = P_{\mathrm{F}ix(T)\cap\Gamma}(I - P + \gamma f)z$.

Proof. Taking $T(s) := T$, a nonexpansive mapping in Theorem 3.1 then the conclusion of Corollary 3.3 is obtained.

Corollary 3.4 [26] *Let H be a real Hilbert space and let $C \subseteq H$ be a nonempty closed convex subset. Assume that $F: C \times C \to \mathbb{R}$ be the bifunctions satisfying Assumption* 2.1. *Let T be a nonexpansive mapping on C such that $Fix(T) \cap EP(F) \neq \emptyset$. Let $f: C \to C$ be a contraction mapping with constant $\alpha \in (0,1)$ and P be a strongly positive bounded linear operator on H with constant $\bar{\gamma} > 0$ such that $0 < \gamma < \frac{\bar{\gamma}}{\alpha} < \gamma + \frac{1}{\alpha}$. For a given $x_0 \in C$ arbitrarily, let the iterative sequences $\{u_n\}$ and $\{x_n\}$ be generated by*

$$x_{n+1} = \alpha_n f(x_n) + (1 - \alpha_n P)\frac{1}{s_n}\int_0^{s_n} T(s)x_n ds,$$

where $\{\alpha_n\}$ is a sequence in $(0,1)$ satisfying $\lim_{n\to\infty} \alpha_n = 0$, $\sum_{n=0}^{\infty} \alpha_n = \infty$; and $\sum_{n=1}^{\infty} |\alpha_{n+1} - \alpha_n| < \infty$;. Then the sequence $\{x_n\}$ converges strongly to $z \in \mathrm{F}ix(T) \cap \mathrm{E}P(F)$, where $z = P_{\mathrm{F}ix(T)\cap\mathrm{E}P(F)}(I - P + \gamma f)z$.

Proof. Taking $F_1 = F_2 = 0$, $H_1 = H_2 = H$, $\gamma = 1$, and $P = I$ in Theorem 3.1 then the conclusion of Corollary 3.4 is obtained.

Corollary 3.5 [11] *Let H be a real Hilbert space and let $C \subseteq H$ be a nonempty closed convex subset. Assume that $F: C \times C \to \mathbb{R}$ be the bifunctions satisfying Assumption* 2.1. *Let T be a nonexpansive mapping on C such that $Fix(T) \cap EP(F) \neq \emptyset$. Let $f: C \to C$ be a contraction mapping with constant $\alpha \in (0,1)$. For a given $x_0 \in C$ arbitrarily, let the iterative sequences $\{u_n\}$ and $\{x_n\}$ be generated by*

$$\begin{cases} u_n = T_{r_n}^{F} x_n; \\ x_{n+1} = \alpha_n f(x_n) + (1 - \alpha_n)Tu_n, \end{cases}$$

where $r_n \subset (0, \infty)$ and $\{\alpha_n\}$ is a sequence in $(0,1)$ satisfying the following conditions:

(i) $\lim_{n\to\infty} \alpha_n = 0$, $\sum_{n=0}^{\infty} \alpha_n = \infty$; and $\sum_{n=1}^{\infty} |\alpha_{n+1} - \alpha_n| < \infty$;.

(ii) $\liminf_{n\to\infty} r_n > 0,\ \sum_{n=1}^{\infty} |r_{n+1} - r_n| < \infty;$.

Then the sequence $\{x_n\}$ converges strongly to $z \in \mathrm{Fix}(T) \cap \mathrm{EP}(F)$, where $z = P_{\mathrm{Fix}(T)\cap \mathrm{EP}(F)}(I - P + \gamma f)z$.

Proof. Taking $F_1 = F_2 = F$, $H_1 = H_2 = H$, $\gamma = 1$, $P = I$ and $T(s) := T$, a nonexpansive mapping in Theorem 3.1 then the conclusion of Corollary 3.5 is obtained.

4 Numerical Example

Now, we give a numerical example which justify the main result Theorem 3.1.

Example 4.1 *Let $H_1 = H_2 = \mathbb{R}$, the set of all real numbers, with the inner product defined by $\langle x, y\rangle = xy, \forall x, y \in \mathbb{R}$, and induced usual norm $|.|$. Let $C = [0, +\infty)$ and $Q = (-\infty, 0]$; let $F_1 : C \times C \to \mathbb{R}$ and $F_2 : Q \times Q \to \mathbb{R}$ be defined by $F_1(x, y) = x(y - x), \forall x, y \in C$ and $F_2(u, v) = (u - 2)(v - u), \forall u, v \in Q$; let for each $x \in \mathbb{R}$, we define $f(x) = \frac{1}{2}x$, $A(x) = \frac{-1}{2}x$, and let, for each $x \in C$, $T(s)x = \frac{e^{-s}}{1000}x$.*

Then it is easy to prove that the bifunctions F_1 and F_2 satisfy the Assumption 2.1 and F_2 is upper semicontinuous and $T(s)x = \frac{e^{-s}}{1000}x$ is a nonexpansive semigroup and A is a bounded linear operator on $\mathbb{R}$ with adjoint operator A^* and A is a bounded linear operator on $\mathbb{R}$ with adjoint operator A^* and $\| A \| = \| A^* \| = \frac{1}{4}$. Furthermore, take $\gamma = 1$, $\alpha_n = \frac{1}{n+1}$ and following the steps of proof of Theorem 3.1, we obtain that $\{x_n\}$ converge strongly to $w = 0 \in \mathrm{Fix}(T) \cap \Gamma$. Furthermore, the following figure shows the convergence behavior and comparison of our proposed iterative method with the iterative method of (1.6) and (1.7) with the stopping criteria is $\| x_n - p \| \leq 10^{-4}$.

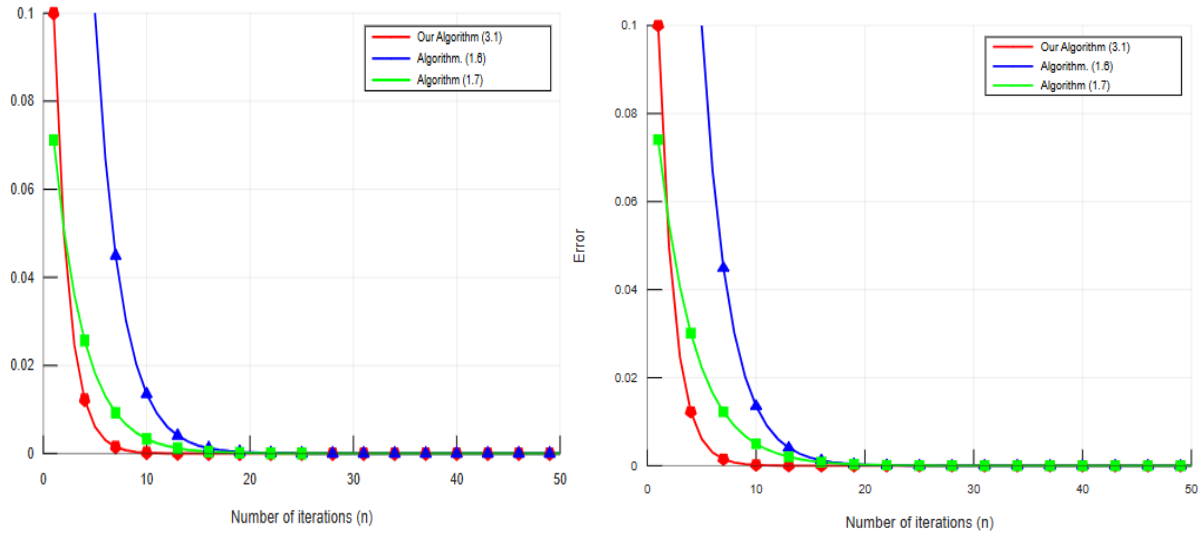


Figure: Convergence of {xn} for different initial values

Conclusion

In this paper, we have considered and studied the split equilibrium problem and fixed point problem which includes the previously known one as special cases. The aim of present work is to study the strong convergence theorem based on the modified viscosity iterative method for split equilibrium and fixed point problems for a nonexpansive semigroup. The main result extends and generalize the result of Plubtieng [27] and Cianciaruso *et al.* [28] for equilibrium problem into a split equilibrium problem and fixed point problem for a nonexpansive semigroup.

Acknowledgement

This work is supported by the Deanship of Scientic Research (DSR), Saudi Electronic University, Riyadh. The author, therefore gratefully acknowledge the DSR technical support.

Funding

The authors declare that this research received no external funding

Data Availability Statement

Not applicable

Conflicts of Interest

The authors declare no conflict of interest / or state any conflict if any.

References

[1] E. Blum and W. Oettli, From optimization and variational inequalities to equilibrium problems, Math. Stud., 63 (1994) 123-145.

[2] F. Giannessi, Vector variational inequalities and vector equilibria. Mathematical theories. Nonconvex Optimization and its Applications, 38, Kluwer Academic Publishers, Dordrecht, 2000.

[3] S.D. Flam and A.S. Antipin, Equilibrium programming using proximal-like algorithms, Math. Program., 78 (1997) 29-41.

[4] Moudafi, A, Théra, M: *Proximal and Dynamical Approaches to Equilibrium Problems, Lecture Notes in Economics and Mathematical Systems*, Springer-Verlag, New York 477 (1999) 187-201.

[5] K.R. Kazmi and S.H. Rizvi, An iterative algorithm for generalized mixed equilibrium problem, Afrika Mathematika, 25(4) (2014) 857-867.

[6] L.C. Ceng and J.C. Yao, A hybrid iterative scheme for mixed equilibrium problems and fixed point

problems, J. Comput. Appl. Math., 214 (2008) 186-201.

[7] P. Daniele, F. Giannessi and A. Mougeri (Eds), Equilibrium Problems and Variational Models, Nonconvex Optimization and its Application, 68, Kluwer Academic Publications, Norwell, 2003.

[8] K.R. Kazmi and S.H. Rizvi, A hybrid extragradient method for approximating the common solutions of a variational inequality, a system of variational inequalities, a mixed equilibrium problem and a fixed point problem, Appl. Math. Comput., 218 (2012) 5439-5452.

[9] K.R. Kazmi and S.H. Rizvi, Iterative algorithms for generalized mixed equilibrium problems, J. Egyptian Math. Soc., 21(3) (2013) 340-345.

[10] A. Tada and W. Takahashi, Strong convergence theorem for an equilibrium problem and and a nonexpansive mapping in W. Takahashi, T. Tanaka (eds), Nonlinear Analysis and Convex Analysis, Yokohama, 2006, 609-617.

[11] S. Takahashi and W. Takahashi, Viscosity approximation method for equilibrium problems and fixed point problems in Hilbert space, J. Math. Anal. Appl., 331 (2007) 506-515.

[12] Rizvi, SH, Sikander, F: Iterative approximation of a common solution of split equilibrium, split variational inequality, and fixed point problem for a nonexpansive semigroup, *Math. Meth. Appl. Sci.,* 45(8) (2022) 4343-4364.

[13] Rizvi, SH: A strong convergence theorem for split mixed equilibrium and fixed point problems for nonexpansive mappings, *J. Fixed Theory Appl.,* 20(1) (2018), 8 Pages.

[14] P.L. Combettes and S.A. Hirstoaga, Equilibrium programming in Hilbert spaces, J. Nonlinear Convex Anal., 6 (2005) 117-136.

[15] K.R. Kazmi and S.H. Rizvi, Iterative approximation of a common solution of a split equilibrium problem, a variational inequality problem and a fixed point problem, J. Egyptian Math. Soc., 21(1) (2013) 44-51.

[16] Y. Censor, A. Gibali and S. Reich, Algorithms for the split variational inequality problem, Numer. Algorithms, 59(2) (2012) 301–323.

[17] A. Moudafi, Split monotone variational inclusions, J. Optim. Theory Appl., 150 (2011) 275-283.

[18] Y. Censor, T. Bortfeld, B. Martin and A. Trofimov, A unified approach for inversion problems in intensity modulated radiation therapy, Physics in Medicine and Biology, 51 (2006) 2353-2365.

[19] Y. Censor and T. Elfving, A multiprojection algorithm using Bregman projections in product space, Numer. Algorithms, 8 (1994) 221-239.

[20] C. Byrne, Iterative oblique projection onto convex sets and the split feasibility problem. Inverse Problem, 18 (2002) 441-453.

[21] C. Byrne, A unified treatment for some iterative algorithms in signal processing and image reconstruction, Inverse Problem, 20 (2004) 103-120.

[22] P.L. Combettes, The convex feasibility problem in image recovery, Adv. Imaging Electron Physics 95 (1996) 155-453.

[23] H.H. Bauschke and P.L. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Springer, New York, 2011.

[24] G. Marino and H.K. Xu, A general iterative method for nonexpansive mappings in Hilbert spaces, J. Math. Anal. Appl. 318 (2006) 43-52.

[25] T. Shimizu and W. Takahashi, Strong convergence to common fixed points of families of nonexpansive mappings, J. Math. Anal. Appl., 211 (1997) 71-83.

[26] R. Chen and Y. Song, Convergence to common fixed point of nonexpansive semigroups, J. Comput. Appl. Math., 200 (2007) 566-575.

[27] S. Plubtieng and R. Punpaeng, Fixed point solutions of variational inequalities for nonexpansive semigroups in Hilbert spaces, Math. Comput. Model., 48 (2008) 279-286.

[28] F. Cianciaruso, G. Marino, L. Muglia and Y. Yao, A hybrid projection algorithm for finding solutions of mixed equilibrium problem and variational inequality problem, Fixed Point Theory Appl., (2010), Article ID 383740, 19 pages.

[29] Rizvi, SH: Viscosity approximation for split monotone variational inclusions and fixed point problem, *Fixed Point Theory,* 24(2) (2023) 701-720.

[30] Rizvi, SH: A Mann-Halpern algorithm for solving split null point, split mixed equilibrium and fixed point problems for a nonexpansive mapping, *Numerical Algorithms,* (2026) 1-27, https://doi.org/10.1007/s11075-026-02317-8.

[31] Rizvi, SH, Sikander, F: A strongly convergent hybrid method to unify split generalized mixed equilibrium and fixed point problems, *Int. J. Anal. Appl.,* 22(168) (2024), https://doi.org/10.28924/2291-8639-22-2024-168.

[32] Goebel, K, Reich, S: *Uniform Convexity, Hyperbolic Geometry, and Nonexpansive Mappings*, Dekker, New York, 1984.

[33] Z. Opial, Weak convergence of the sequence of successive approximations for nonexpansive mappings. Bull. Amer. Math. Soc., 73(4) (1967) 595-597.

[34] G. Crombez, A hierarchical presentation of operators with fixed points on Hilbert spaces, Numerical Funct. Anal. Optim., 27 (2006) 259-277.

[35] T. Suzuki, Strong convergence of Krasnoselskii and Mann's type sequences for one parameter nonexpansive semigroups without Bochner integrals, J. Math. Anal. Appl., 305 (2005) 227-239.

[36] M. O. Osilike and D.I. Igbokwe, Weak and strong convergence theorems for fixed points of pseudocontractions and solutions of monotone type operator equations, Comput. Math. Appl., 40 (2000) 559-567.

[37] H. K. Xu, Viscosity approximation method for nonexpansive mappings, J. Math. Anal. Appl., 298 (2004) 279-291.

Islamic University Journal of Applied Sciences (IUJAS)

https://journals.iu.edu.sa/jesc

Volume VIII, Issue I, July 2026, Pages 319-342

Autonomous Drones for Precision Pest Control in Desert Agriculture: Technologies, Applications, and Future Prospects

Amal Bedewy[1] *, Reda Ibrahim[1]

[1] Department of Economic Entomology, Kafrelsheikh University, Kafrelsheikh 33516, Egypt

*Corresponding author (Amal Bedwey), *Email Address*: Amalbedewy8@gmail.com

Abstract

Autonomous drone technologies are increasingly recognized as transformative tools for precision pest management, particularly in desert agricultural systems characterized by harsh environmental conditions and limited resources. Conventional pest control approaches, largely dependent on uniform pesticide application, often result in inefficient input use, environmental contamination, and reduced sustainability. This review synthesizes recent advancements in unmanned aerial vehicles (UAVs), remote sensing, and artificial intelligence (AI)-driven analytics for targeted pest detection, monitoring, and control. Drone-based applications, including precision spraying and biological control delivery, have demonstrated the potential to reduce pesticide use by approximately 20–30% while improving spatial accuracy and operational efficiency. Integration with AI and Internet of Things (IoT) systems enables real-time data acquisition, automated pest detection, and adaptive decision-making, enhancing system performance in heterogeneous and remote desert landscapes. Despite these advantages, challenges such as high initial costs, technical limitations, regulatory constraints, and environmental factors remain significant barriers to widespread adoption. Future developments, including drone swarms and predictive modelling, are expected to further improve efficiency. Overall, autonomous drones offer a promising pathway toward sustainable, data-driven pest management in desert ecosystems.

Keywords:

Autonomous drones; Precision agriculture; Desert agriculture; Smart pest management; UAV; Artificial intelligence; Drone swarms; Sustainable agriculture

https://doi.org/10.63070/jesc.2026.030

Received 11 March 2026; Revised 24 April 2026; Accepted 28 April 2026.

Available online 02 May 2026.

Published by Islamic University of Madinah on behalf of *Islamic University Journal of Applied Sciences*

This is an open access article published under the Creative Commons Attribution 4.0 International license (CC BY 4.0):https://creativecommons.org/licenses/by/4.0/

1. Introduction

Agricultural production in arid and semi-arid regions plays a critical role in ensuring food security for rapidly growing populations, particularly across North Africa and the Middle East. However, desert agriculture is inherently constrained by extreme environmental conditions, including high temperatures, limited and erratic precipitation, poor soil fertility, and water scarcity. These stressors not only reduce crop productivity but also profoundly influence pest dynamics, often leading to increased infestation risks and unstable agroecosystems [1], [2].

In such environments, pest populations frequently exhibit rapid growth, irregular spatial distribution, and enhanced adaptability to control measures. Conventional pest management strategies—primarily based on uniform pesticide application—are often inefficient under these conditions. They can result in excessive chemical use, environmental contamination, and adverse effects on non-target organisms, while also accelerating the development of pesticide resistance [3], [4]. These limitations highlight the need for more precise, efficient, and sustainable pest management approaches.

Precision agriculture has emerged as a transformative paradigm that enables site-specific crop and pest management through the integration of advanced technologies such as remote sensing, geographic information systems (GIS), and data analytics [5], [6]. By capturing spatial and temporal variability within agricultural systems, this approach allows for optimized input use, improved efficiency, and reduced environmental impact. Within this framework, unmanned aerial vehicles (UAVs), commonly known as drones, have gained prominence as versatile tools for both monitoring and intervention.

Recent advancements in UAV technologies have significantly expanded their capabilities. Modern drones are equipped with high-resolution and multispectral sensors that enable early detection of crop stress associated with pest infestations, often before visible symptoms appear [7], [8]. In parallel, UAV-based spraying systems facilitate targeted pesticide application, improving treatment precision while reducing chemical usage [9], [10]. These developments mark a shift from conventional blanket treatments toward data-driven, site-specific pest control.

The integration of artificial intelligence (AI) and machine learning (ML) further enhances the functionality of drone-based systems. These technologies enable automated image analysis, accurate pest detection, and predictive modelling, supporting real-time decision-making and adaptive management strategies [11], [12]. Additionally, the incorporation of Internet of Things (IoT) technologies allows continuous monitoring and seamless data exchange between sensors, drones, and decision-support platforms, thereby improving system efficiency and responsiveness.

Despite these technological advancements, the application of autonomous drone systems in desert agriculture remains relatively underexplored. Most existing studies focus on temperate agricultural

systems, with limited attention to the unique ecological and operational challenges of arid environments. Factors such as harsh climatic conditions, limited infrastructure, high operational costs, and regulatory constraints continue to hinder widespread adoption, particularly in developing regions. In this context, the present review aims to provide a comprehensive and critical evaluation of autonomous drone technologies for precision pest control in desert agriculture. Specifically, it seeks to: (i) examine pest management challenges in arid agroecosystems; (ii) review recent advancements in drone-based monitoring and control systems; (iii) assess the advantages and limitations of these technologies; and (iv) identify future research directions for sustainable implementation. By synthesizing current knowledge and highlighting emerging trends, this review contributes to advancing data-driven and environmentally sustainable pest management strategies tailored to desert ecosystems.

2. Autonomous Drones in Precision Pest Management: Concepts and Applications

2.1. Desert Agriculture: Challenges and Pest Dynamics

Desert agriculture represents one of the most environmentally constrained forms of crop production, where extreme climatic conditions and limited natural resources impose significant pressures on agroecosystem stability. High temperatures, low relative humidity, intense solar radiation, and chronic water scarcity directly influence plant physiology, soil properties, and pest population dynamics [1], [2]. These environmental stressors create conditions that often favour rapid pest adaptation, increased reproductive rates, and altered seasonal activity patterns.

In arid and semi-arid regions, agricultural expansion through irrigation and land reclamation has led to the establishment of newly cultivated ecosystems characterized by reduced biodiversity and weakened ecological regulation. The limited presence of natural enemies in such systems frequently results in increased pest outbreak frequency and severity [13], [14]. Key pest groups include sap-sucking insects such as aphids and whiteflies, lepidopteran larvae (e.g., *Spodoptera* spp.), and various soil and stored-product pests, all of which can cause substantial yield losses under favourable climatic conditions.

In Egypt, particularly in newly reclaimed desert lands and the Nile Delta fringes, pest infestations by whiteflies (*Bemisia tabaci*) and cotton leafworm (*Spodoptera littoralis*) have been widely reported as major constraints to crop productivity under high-temperature conditions. Similarly, in Gulf countries such as Saudi Arabia and the United Arab Emirates, greenhouse and irrigated desert farming systems frequently experience outbreaks of aphids, thrips, and mite species, driven by continuous cropping and favourable microclimatic conditions. These regional examples highlight the growing vulnerability of desert agricultural systems to pest pressures under intensified production practices. Temperature plays

a critical role in shaping insect development and population dynamics. Elevated temperatures commonly observed in desert environments accelerate insect metabolism, shorten generation time, and increase the number of life cycles per season, thereby promoting rapid population build-up [15]. Concurrently, water-stressed crops often exhibit reduced defensive capacity, making them more susceptible to pest attacks and amplifying yield losses.

Another defining characteristic of pest infestations in desert agriculture is their pronounced spatial heterogeneity. Pest populations are often unevenly distributed across fields due to microclimatic variability, irrigation patterns, and soil heterogeneity. This patchy distribution reduces the effectiveness of conventional uniform pesticide applications, leading to over-application in some areas and insufficient control in others [5], [6]. The heavy reliance on chemical pesticides in these systems further exacerbates pest management challenges. Repeated and indiscriminate use contributes to resistance development, environmental contamination, and negative impacts on beneficial organisms, including pollinators and natural enemies [4], [16]. In fragile desert ecosystems, where ecological resilience is inherently limited, such effects can have long-lasting consequences.

In addition to ecological constraints, logistical challenges hinder effective pest management in desert regions. Large field sizes, difficult terrain, and limited infrastructure complicate routine monitoring and timely intervention using traditional ground-based methods. Labour shortages and increasing operational costs further reduce the feasibility of manual scouting and conventional spraying approaches. Collectively, these challenges underscore the urgent need for innovative, adaptive, and resource-efficient pest management strategies tailored to desert agroecosystems. Approaches based on precision agriculture and advanced technologies, particularly autonomous drone systems, offer significant potential to address these limitations by enabling targeted monitoring and site-specific pest control interventions.

2.2. Precision Agriculture and Smart Pest Management

Precision agriculture has emerged as a transformative approach to modern farming, enabling the optimization of agricultural inputs through site-specific management practices. By integrating advanced technologies such as remote sensing, global positioning systems (GPS), geographic information systems (GIS), and data analytics, precision agriculture facilitates the monitoring of spatial and temporal variability within fields, thereby improving productivity while minimizing environmental impacts [5], [6]. Within this framework, pest management has evolved from conventional uniform control strategies toward more targeted and adaptive approaches, commonly referred to as smart pest management. This paradigm relies on real-time data acquisition, predictive modelling, and automated intervention systems to detect, monitor, and control pest populations with high precision and efficiency.

A fundamental component of smart pest management is early pest detection using advanced sensing technologies. Remote sensing platforms, including satellites and unmanned aerial vehicles (UAVs), enable the collection of high-resolution data on crop health. One of the most widely used indicators derived from such data is the Normalized Difference Vegetation Index (NDVI). NDVI is a simple numerical index that reflects the "greenness" or health status of vegetation by comparing the amount of light reflected in the near-infrared (which healthy plants strongly reflect) and red wavelengths (which plants absorb during photosynthesis). The index typically ranges from −1 to +1, where higher positive values indicate healthy, actively growing vegetation, while lower values may signal plant stress, damage, or poor growth conditions. In the context of pest management, a localized decline in NDVI values can serve as an early warning sign of pest infestation, often before visible symptoms appear in the field [8], [17].

Hyperspectral imaging further enhances detection accuracy by capturing detailed spectral information across a wide range of wavelengths, allowing for more precise identification of plant stress factors. Although highly informative, this technology is often more complex and costlier than multispectral approaches, which may limit its widespread application. Artificial intelligence (AI) and machine learning (ML) techniques play a crucial role in processing and interpreting the large datasets generated by remote sensing systems. Advanced algorithms, particularly convolutional neural networks (CNNs), can analyse aerial imagery to detect pest presence, classify infestation severity, and generate spatial distribution maps with high accuracy [11], [12]. These capabilities support automated decision-making and enable timely, targeted pest control interventions.

Another key element of precision pest management is variable-rate technology (VRT), which allows for the site-specific application of pesticides, fertilizers, and biological control agents. By applying inputs only where needed and in appropriate quantities, VRT reduces chemical usage, lowers operational costs, and minimizes environmental impact [5]. This approach is especially important in desert agriculture, where efficient resource use is essential. The integration of the Internet of Things (IoT) further enhances smart pest management systems by enabling continuous monitoring and real-time communication between sensors, drones, and decision-support platforms. IoT devices, such as soil sensors, weather stations, and automated insect traps, provide complementary data that improve the accuracy of pest detection and forecasting models [18]. These interconnected systems enable dynamic, data-driven responses to changing field conditions.

Autonomous drones serve as a critical link between sensing and intervention within precision agriculture systems. They enable rapid data acquisition, high-resolution mapping, and immediate implementation of control measures, such as targeted pesticide application or biological agent release. Compared to conventional approaches, drone-based systems offer superior spatial accuracy, reduced

input use, and enhanced operational efficiency. Overall, the integration of precision agriculture technologies with smart pest management strategies represents a significant shift toward sustainable, data-driven crop protection. This approach is particularly well suited to desert agricultural systems, where environmental constraints and resource limitations demand efficient and adaptive management solutions. The integration of these technologies into pest management systems is conceptually illustrated in Figure 1, which highlights the interaction between sensing, analytics, and intervention components.



Figure 1. Conceptual framework of smart drone-based pest management systems.

A comparative overview between conventional and precision pest management approaches is presented in Table 1, demonstrating the efficiency gains and sustainability advantages of data-driven systems.

Table 1. Comparison between conventional and precision pest management.

Aspect	Conventional Pest Management	Precision Pest Management
Application Method	Uniform spraying	Site-specific application
Pesticide Use	High and often excessive	Optimized and reduced
Environmental Impact	High contamination risk	Lower environmental footprint

Aspect	Conventional Pest Management	Precision Pest Management
Efficiency	Low efficiency	High efficiency
Monitoring	Manual scouting	Automated sensing (UAV, IoT)
Decision Making	Experience-based	Data-driven (AI, analytics)
Cost (Long-term)	High due to waste	Reduced through optimization
Adaptability	Limited	Highly adaptive and dynamic
Suitability for Desert Areas	Poor	Highly suitable

2.3. Autonomous Drone Technology in Agriculture

Unmanned aerial vehicles (UAVs), commonly referred to as drones, have become integral components of modern precision agriculture, offering advanced capabilities for data acquisition, monitoring, and field-level intervention. Recent technological advancements have enabled the evolution of UAVs from remotely piloted systems to semi-autonomous and fully autonomous platforms capable of executing complex agricultural tasks with minimal human intervention [6], [19].

Agricultural UAVs are generally classified into three main categories: fixed-wing, rotary-wing (multirotor), and hybrid systems. Each category differs in flight characteristics, operational capabilities, and suitability for specific agricultural applications. Fixed-wing drones are well suited for large-scale field monitoring due to their long flight endurance and wide coverage capacity. In contrast, multirotor drones provide superior manoeuvrability and hovering capabilities, making them ideal for precision spraying and localized pest control operations. Hybrid UAVs combine the advantages of both systems by integrating vertical take-off and landing (VTOL) capabilities with extended flight range, offering flexibility in complex and infrastructure-limited environments [19].

To provide a clearer comparison of these UAV types and their agricultural applications, a summary is presented in Table 2.

Table 2. Comparative Characteristics of UAV Types Used in Agriculture

UAV Type	Key Features	Advantages	Limitations	Typical Applications
Fixed-wing	Airplane-like structure; long flight endurance	Large area coverage; energy-efficient; suitable for mapping	Requires runway/launch system; limited hovering capability	Large-scale crop monitoring, mapping, surveillance
Rotary-wing (Multirotor)	Multiple rotors; vertical take-off and landing (VTOL)	High manoeuvrability; precise hovering; easy deployment	Short flight time; limited coverage area	Precision spraying, localized pest control, detailed inspection
Hybrid (VTOL)	Combines fixed-wing and multirotor features	Extended range + vertical take-off; flexible operation	More complex design; higher cost	Mixed applications (monitoring + spraying), remote/desert farming

The effectiveness of autonomous UAV systems relies on the integration of advanced hardware and software components. Key onboard sensors include high-resolution RGB cameras, multispectral and hyperspectral sensors, thermal imaging devices, and, in some cases, LiDAR systems. These sensors enable the collection of detailed spatial and spectral data necessary for detecting crop stress and pest infestations [8], [20]. In addition, Global Navigation Satellite Systems (GNSS), often enhanced by Real-Time Kinematic (RTK) positioning, provide high-precision geolocation required for accurate navigation and targeted interventions.

Autonomous navigation and flight control systems represent a major advancement in UAV technology. These systems utilize pre-programmed flight paths, sensor fusion, and real-time environmental feedback to ensure accurate and efficient operation. Features such as obstacle avoidance, terrain-following capabilities, and adaptive route optimization enhance operational safety and coverage efficiency, particularly in complex agricultural landscapes [21]. The integration of artificial intelligence (AI) and machine learning (ML) further enhances UAV functionality by enabling real-time data processing and decision-making. AI-powered drones can analyse imagery during flight to identify pest-infested areas, assess crop health, and generate actionable insights for targeted intervention. This capability significantly reduces response time and improves the precision of pest management strategies [11], [12].

In desert agriculture, the advantages of autonomous UAVs are particularly significant. Their ability to operate in remote, inaccessible, and environmentally harsh conditions makes them well suited for arid

regions where conventional machinery is often impractical. UAVs can rapidly cover large areas while minimizing water and chemical usage, thereby supporting sustainable agricultural practices under resource-limited conditions. Despite these advantages, several technical limitations remain. Restricted battery life and payload capacity constrain flight duration and operational efficiency, especially in large-scale farming systems. Additionally, UAV performance can be adversely affected by extreme environmental conditions such as high temperatures, strong winds, and dust, which are common in desert regions. These challenges highlight the need for continued innovation in energy storage systems, lightweight materials, and climate-resilient UAV design. Overall, autonomous drone technology represents a critical enabler of next-generation precision pest management systems, effectively bridging the gap between data acquisition and targeted field intervention.

2.4. Drone-Based Pest Detection and Monitoring Systems

Effective pest management depends on the timely and accurate detection of infestations. Conventional field scouting methods are often labour-intensive, time-consuming, and limited in their ability to capture the spatial variability of pest populations. In contrast, drone-based detection systems provide a rapid, non-destructive, and high-resolution approach to monitoring crop health and pest dynamics, forming a cornerstone of precision pest management. Drone-based pest detection primarily relies on remote sensing technologies that capture spectral and thermal information reflecting plant physiological status. Multispectral imaging is among the most widely used techniques, enabling the calculation of vegetation indices such as the Normalized Difference Vegetation Index (NDVI), which serves as a reliable indicator of plant health. Variations in NDVI values can reveal early signs of stress associated with pest infestations before visible symptoms appear [17], [22]. This early detection capability is critical for timely intervention and minimizing crop damage.

Hyperspectral imaging offers enhanced detection precision by capturing a large number of narrow spectral bands, allowing for detailed characterization of plant biochemical and structural changes. This technology enables the differentiation between various stress factors, including pest damage, nutrient deficiencies, and water stress, although its high cost and data complexity may limit widespread adoption [23]. Thermal imaging provides an additional layer of diagnostic capability, particularly in arid environments. Pest infestations often disrupt plant transpiration processes, leading to changes in canopy temperature that can be detected using thermal sensors. When combined with multispectral and hyperspectral data, thermal imaging improves the accuracy of stress detection and supports more reliable pest identification [24].

The integration of artificial intelligence (AI) and deep learning techniques has significantly enhanced the effectiveness of drone-based detection systems. Machine learning models, particularly

convolutional neural networks (CNNs), can analyse high-resolution aerial imagery to automatically identify pest species, classify infestation severity, and map affected areas with high accuracy [11], [12]. These systems can be trained on large datasets and continuously improved through adaptive learning, enabling robust performance across diverse agricultural conditions. Another important advancement is the use of real-time data transmission and cloud-based processing platforms. Drone-acquired data can be transmitted to cloud systems for immediate analysis, allowing farmers and decision-makers to access actionable insights through user-friendly interfaces. This capability significantly reduces response time and enhances decision-making efficiency. The architecture of drone-based pest detection and monitoring systems is summarized in Figure 2, illustrating the integration of sensing platforms, AI analytics, and data communication networks.

The integration of drone-based sensing with Internet of Things (IoT) devices further strengthens monitoring systems. Ground-based sensors, such as weather stations and pheromone traps, provide complementary data on environmental conditions and pest activity. When combined with aerial data, these systems enable comprehensive and continuous monitoring of pest populations and crop health [18]. Importantly, drone-based detection systems facilitate the generation of high-resolution pest distribution maps, which are essential for implementing site-specific control strategies. These maps identify infestation hotspots and guide targeted interventions, reducing unnecessary pesticide applications and improving overall management efficiency. Overall, drone-based pest detection and monitoring systems represent a significant advancement over traditional methods, enabling early detection, precise mapping, and data-driven decision-making—key components of sustainable pest management in desert agriculture

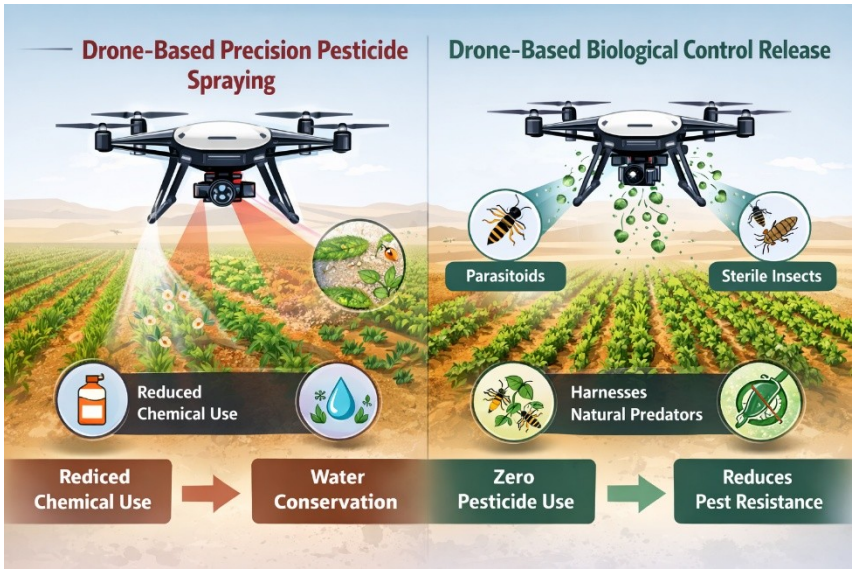


Figure 2. Architecture of Drone-Based Pest Detection and Monitoring System

2.5. Drone-Based Pest Control Applications

Autonomous drones have evolved beyond monitoring tools to become active platforms for pest control, enabling precise, efficient, and scalable intervention strategies. By integrating sensing, decision-making, and application capabilities, UAVs provide a unified approach to crop protection that is particularly well suited to precision agriculture systems.

Drone-based pest control applications can be broadly categorized into chemical, biological, and integrated approaches. One of the most widely adopted applications is precision pesticide spraying. UAV-based spraying systems utilize pre-programmed flight paths, GNSS positioning, and variable-rate technologies to apply pesticides directly to identified infestation zones. Compared to conventional ground-based spraying, drone-based systems have been shown to reduce pesticide use by approximately 20–30%, while improving droplet deposition uniformity and spatial accuracy. In addition, field studies report that UAV spraying can increase operational efficiency by up to 40% and significantly reduce labour requirements, particularly in large or difficult-to-access areas [9], [25], [31], [32].

In addition to chemical control, drones are increasingly used for the delivery of biological control agents. These include the aerial release of parasitoids, predators, and sterile insects for environmentally sustainable pest suppression. For example, the deployment of *Trichogramma* spp. via UAVs has demonstrated effective control of lepidopteran pests, with reported parasitism rates exceeding 70% under field conditions. Similarly, drone-assisted implementation of the Sterile Insect Technique (SIT) enables large-scale release of sterilized insects, leading to measurable reductions in pest population density and reproductive success [26], [27].

Another emerging application involves the deployment of smart pest control devices, such as pheromone traps, attract-and-kill systems, and sensor-based monitoring tools. UAVs can precisely position these devices in infestation hotspots, improving their effectiveness while minimizing labour inputs. This approach enhances integrated pest management (IPM) strategies by combining monitoring and control within a coordinated framework.

Drone-based systems also support real-time adaptive pest control. By integrating AI-driven analytics with onboard sensors, UAVs can dynamically adjust spraying parameters—such as dosage, droplet size, and flight altitude—based on pest density and environmental conditions. This adaptive capability improves treatment efficacy while reducing resource use and environmental impact.

Importantly, drone applications can be scaled from individual fields to landscape-level pest management. Coordinated multi-drone operations enable synchronized interventions across large agricultural areas, which is particularly valuable in desert environments where pest outbreaks can spread rapidly under favourable climatic conditions. Such scalability enhances regional pest management efficiency and system resilience.

Despite these advantages, operational challenges remain, including payload limitations, regulatory constraints, and the need for specialized technical expertise. However, ongoing advancements in UAV technology, automation, and integration with smart agricultural systems are expected to further expand the role of drones in sustainable pest management.

Overall, drone-based pest control applications represent a significant advancement in modern agriculture, offering precise, efficient, and environmentally responsible alternatives to conventional pest control methods. The major categories of drone-based pest control applications and their associated benefits are summarized in Table 3.

Table 3. Major Drone-Based Pest Control Applications and Their Benefits

Application Type	Description	Key Benefits	Suitability in Desert Agriculture
Precision Pesticide Spraying	Targeted application using UAV sprayers	Reduced chemical use, high accuracy	Very high
Biological Control Release	Release of parasitoids/predators	Eco-friendly, no residues	High
Sterile Insect Technique (SIT)	Release of sterile males	Population suppression	High
Smart Device Deployment	Placement of traps/sensors	Monitoring + control integration	Moderate–High
AI-Based Adaptive Spraying	Real-time adjustment of spraying	Optimized efficiency	Very high
Landscape-Level Control	Coordinated multi-drone operations	Large-scale management	Very high

2.6. Advantages of Autonomous Drones in Pest Management

The integration of autonomous drones into pest management systems offers significant advantages over conventional approaches, particularly in resource-constrained and environmentally sensitive agroecosystems such as desert agriculture. These advantages are primarily associated with improved precision, operational efficiency, and sustainability. One of the most notable benefits is the substantial reduction in pesticide usage. Drone-based systems enable site-specific application through precise targeting of infestation hotspots, thereby minimizing unnecessary chemical inputs. This targeted approach not only reduces input costs but also mitigates environmental contamination and human exposure to hazardous substances [9], [25]. Moreover, optimized droplet size and controlled application techniques improve pesticide deposition efficiency while reducing off-target drift.

Autonomous drones also enhance operational efficiency by significantly reducing labour requirements and time associated with pest monitoring and control. UAVs can rapidly cover large agricultural areas, even in difficult terrain or inaccessible regions where conventional machinery is impractical. This capability is particularly valuable in desert environments, where labour shortages and logistical constraints often limit effective pest management [19]. Water conservation represents another critical advantage, especially in arid regions. Precision spraying systems integrated into UAV platforms optimize liquid usage by delivering minimal yet effective volumes of pesticides or biological agents. This contributes to sustainable water management practices and improves resource-use efficiency in water-scarce agricultural systems. From an environmental perspective, drone-based pest management supports ecological sustainability by reducing the negative impacts associated with excessive pesticide use. Lower chemical inputs help preserve beneficial organisms, including pollinators and natural enemies, and reduce the risk of soil and water contamination [4]. Additionally, the integration of biological control agents through drone delivery systems further enhances environmentally friendly pest management strategies. A comparative operational assessment between conventional and drone-based pest management systems is provided in Table 4.

Table 4. Comparison Between Conventional and Drone-Based Pest Management

Parameter	Conventional Pest Management	Drone-Based Pest Management
Application Method	Uniform spraying	Site-specific targeted spraying
Precision	Low	High
Pesticide Use	High	Reduced
Labour Requirement	High	Low
Time Efficiency	Low	High

Parameter	Conventional Pest Management	Drone-Based Pest Management
Environmental Impact	High	Low
Cost (Long-term)	High	Moderate
Accessibility	Limited in remote areas	High (especially in difficult terrain)

Another important advantage is the enhancement of decision-making through data-driven approaches. By integrating real-time data from sensors, remote sensing platforms, and AI-based analytics, drones enable timely and informed interventions. This reduces uncertainty and improves the effectiveness of pest control strategies, particularly in dynamic environments where pest populations fluctuate rapidly [11], [18]. Furthermore, autonomous drones provide scalability and flexibility in pest management operations. They can be adapted to different crop types, field sizes, and environmental conditions, allowing for broad applicability across diverse agricultural systems. The potential use of coordinated multi-drone systems (drone swarms) further enhances scalability, enabling efficient management of large agricultural landscapes.

Despite these advantages, it is important to recognize that the successful implementation of drone-based pest management depends on factors such as technological accessibility, regulatory support, and user expertise. Nevertheless, the demonstrated benefits strongly position autonomous drones as a key component of sustainable and precision-based crop protection systems.

2.7. Challenges and Limitations of Autonomous Drone-Based Pest Management

Despite the significant potential of autonomous drones in precision pest management, several technical, environmental, economic, and regulatory challenges continue to limit their widespread adoption, particularly in desert agricultural systems. Addressing these constraints requires targeted mitigation strategies and continued technological innovation. One of the primary technical limitations is restricted battery life and flight endurance. Most multirotor UAVs can operate for relatively short durations, often less than one hour depending on payload and environmental conditions, which reduces coverage efficiency in large-scale farming systems [19], [21]. To mitigate this limitation, ongoing research is focusing on advanced battery technologies, rapid battery-swapping systems, hybrid power solutions (e.g., fuel-assisted or solar-supported UAVs), and optimized flight path planning to maximize coverage efficiency.

Payload capacity represents another constraint, limiting the volume of pesticides or biological agents that can be carried per flight. This necessitates multiple flights for large fields, increasing operational time**.** Potential solutions include the development of lightweight materials, improved payload-to-

weight ratios, and coordinated multi-drone (swarm) operations that distribute workload across several UAVs to enhance overall efficiency. Environmental conditions in desert regions pose additional challenges. High temperatures, strong winds, and dust can negatively affect drone performance, flight stability, and sensor accuracy. Extreme heat may reduce battery efficiency, while dust can interfere with imaging systems. Mitigation strategies include the design of climate-resilient UAVs with heat-resistant components, protective sensor enclosures, improved cooling systems, and the use of robust flight control algorithms capable of adapting to variable environmental conditions.

From an economic perspective, the high initial investment cost associated with UAV acquisition, maintenance, and supporting technologies remains a significant barrier, particularly for smallholder farmers. To address this issue, service-based models such as “drone-as-a-service” (DaaS), cooperative ownership schemes, government subsidies, and cost-sharing initiatives can improve accessibility and reduce financial burden. Another major limitation is the complexity of data processing and interpretation. Drone-based systems generate large volumes of high-resolution data that require advanced analytical tools and technical expertise. This challenge can be mitigated through the development of user-friendly software platforms, cloud-based processing systems, and AI-driven automated analytics that simplify data interpretation and reduce reliance on specialized expertise. Training programs and capacity-building initiatives are also essential to enhance user adoption.

Regulatory and policy-related constraints further restrict UAV deployment. In many countries, drone operations are subject to strict aviation regulations, including licensing requirements, restricted flight zones, and altitude limitations [21]. Addressing these barriers requires the development of clear, agriculture-specific UAV regulations, streamlined licensing procedures, and collaboration between policymakers, researchers, and industry stakeholders to create enabling regulatory frameworks.

Operational challenges also include limited technical expertise among end-users and insufficient infrastructure, such as charging stations and reliable internet connectivity, particularly in remote desert regions. Mitigation efforts should focus on expanding rural digital infrastructure, establishing technical support networks, and implementing targeted training programs to build local capacity. Finally, concerns related to data privacy, cybersecurity, and system reliability must be considered. Autonomous drone systems rely on continuous data exchange, making them potentially vulnerable to cyber threats or system failures. To address these concerns, robust cybersecurity protocols, encrypted data transmission systems, and fail-safe operational mechanisms should be integrated into UAV platforms to ensure secure and reliable performance. Despite these challenges, rapid advancements in UAV technology, artificial intelligence, energy systems, and regulatory frameworks are expected to

progressively overcome many of these limitations. Continued research, supportive policies, and investment in infrastructure and capacity building will be essential for enabling the large-scale adoption of drone-based pest management systems, particularly in desert agriculture.

2.8. Integration with Artificial Intelligence, IoT, and Big Data Analytics

The transformative potential of autonomous drone-based pest management systems is significantly enhanced through their integration with artificial intelligence (AI), the Internet of Things (IoT), and big data analytics. This convergence underpins the development of intelligent, adaptive, and automated agricultural systems, often referred to as Agriculture 4.0, which aim to optimize productivity while minimizing environmental impact. Artificial intelligence plays a central role in enabling advanced data analysis and decision-making within drone-based systems. Machine learning (ML) and deep learning algorithms can process large volumes of high-resolution aerial imagery to detect pest infestations, classify species, and estimate infestation severity with high accuracy [11], [12]. In addition, AI-driven predictive models can forecast pest outbreaks by integrating historical data, climatic variables, and crop conditions, thereby supporting proactive pest management strategies [28].

The integration of IoT technologies further enhances system functionality by enabling real-time data acquisition and communication across multiple components of the agricultural ecosystem. IoT devices, including soil moisture sensors, weather stations, and automated insect traps, continuously monitor environmental and biological parameters. These data streams are transmitted to centralized platforms, where they are combined with drone-acquired data to provide a comprehensive and dynamic representation of field conditions [18], [29]. This interconnected framework facilitates rapid decision-making and enables automated responses to emerging pest threats. Big data analytics provides the computational infrastructure required to manage, process, and interpret the vast datasets generated by UAVs and IoT systems. Advanced data processing techniques allow for the identification of patterns, trends, and correlations that may not be detectable through conventional analysis. Cloud computing platforms further enhance accessibility by enabling real-time data storage, processing, and sharing across multiple users and devices [18].

The integration of AI, IoT, and big data enables the development of closed-loop pest management systems, in which data collection, analysis, decision-making, and intervention are seamlessly interconnected. In such systems, drones function as both sensing and action units, while AI algorithms provide analytical intelligence and IoT networks ensure continuous connectivity. This closed-loop architecture significantly improves system responsiveness, efficiency, and adaptability. Another

emerging advancement is the deployment of drone swarms, where multiple UAVs operate collaboratively using decentralized AI algorithms. Swarm intelligence enables coordinated behaviour, optimized field coverage, and efficient resource utilization, thereby enhancing scalability and operational efficiency in large agricultural landscapes [30].

In desert agriculture, the integration of these technologies is particularly valuable due to the need for efficient resource management and rapid adaptation to environmental variability. Predictive models can anticipate pest outbreaks under fluctuating climatic conditions, enabling proactive interventions that reduce crop losses and input costs. Despite these advantages, challenges related to data interoperability, infrastructure requirements, and user accessibility remain. Ensuring seamless integration between different data sources and platforms is essential for maximizing system efficiency. Additionally, investments in digital infrastructure, capacity building, and training programs are necessary to support the effective implementation of these technologies.

Overall, the integration of AI, IoT, and big data analytics with autonomous drone systems represents a paradigm shift toward intelligent, data-driven pest management, offering significant potential for improving sustainability and resilience in desert agriculture.

2.9. Case Studies and Global Applications of Autonomous Drone-Based Pest Control

Representative global applications of drone-based pest management across different regions are summarized in Table 5. The application of autonomous drone technologies in pest management has progressed rapidly from experimental research to large-scale implementation across diverse agricultural systems worldwide. These real-world case studies demonstrate the effectiveness of drone-based solutions in improving pest control efficiency, reducing environmental impact, and enhancing agricultural sustainability.

Table 5: Global Case Studies on Drone-Based Pest Management Applications

Region	Application	Key Outcome
China	UAV pesticide spraying	Reduced chemical use, improved efficiency
Japan	Aerial spraying (rice fields)	High precision, reduced labour
USA	AI-based pest detection	Improved decision-making
Africa	Fall armyworm monitoring	Enhanced surveillance and control
Europe	IPM and biological control	Reduced pesticide dependency
Middle East	Desert agriculture monitoring	Improved resource efficiency

2.9.1. Asia: Large-Scale Adoption in China and Japan

China represents one of the most advanced adopters of agricultural drone technology, with widespread deployment of UAVs for crop protection in rice, wheat, and cotton production systems. Commercial platforms, such as those developed by DJI, are extensively used for precision pesticide application, enabling site-specific spraying and reduced chemical usage. Field studies have reported that UAV-based spraying can reduce pesticide consumption by up to 30% while improving deposition uniformity and operational efficiency [31], [32].

In Japan, unmanned aerial systems have been used for decades in rice cultivation, particularly through the deployment of Yamaha's RMAX helicopters. These systems enable efficient pesticide application in small and fragmented fields, demonstrating high precision and reduced labour requirements. The Japanese model highlights the long-term feasibility of integrating aerial technologies into national pest management programs [33].

2.9.2. North America: Data-Driven Precision Agriculture

In the United States, drone-based pest management is integrated within broader precision agriculture frameworks that emphasize data-driven decision-making. UAVs equipped with multispectral sensors are widely used to detect pest infestations, monitor crop health, and generate prescription maps for targeted pesticide application.

Applications include monitoring invasive pests such as the fall armyworm (*Spodoptera frugiperda*) and implementing site-specific control strategies. The integration of UAV data with farm management platforms enhances decision support systems, allowing for optimized input use and improved sustainability [22], [34].

2.9.3. Africa: Emerging Applications in Pest Surveillance and Control

In several African countries, drone technologies are increasingly being adopted to address major pest challenges, particularly in maize production systems affected by fall armyworm outbreaks. UAVs are used for mapping infestation zones, monitoring pest spread, and guiding targeted pesticide application in large or inaccessible fields.

Collaborative initiatives involving international organizations have demonstrated the feasibility of drone-assisted pest management in resource-limited settings. These efforts highlight the potential of

UAVs to enhance food security and reduce crop losses in developing regions, despite ongoing challenges related to infrastructure and cost [35].

2.9.4. Europe: Sustainable and Regulation-Driven Applications

In Europe, drone-based pest management is closely aligned with sustainability goals and strict environmental regulations. UAVs are increasingly used in vineyards, orchards, and specialty crops for precision spraying and biological control applications.

Countries such as France and Germany have adopted UAV technologies to support integrated pest management (IPM) programs, including the release of beneficial insects and targeted pesticide application. These practices are consistent with the European Union's "Farm to Fork" strategy, which aims to significantly reduce pesticide use while maintaining agricultural productivity [36].

2.9.5. Middle East: Applications in Desert Agriculture

In arid regions of the Middle East, including Saudi Arabia and the United Arab Emirates, drone technologies are being adapted for use in desert agriculture. UAVs are employed for monitoring pest populations in irrigated systems, conducting precision spraying under extreme environmental conditions, and integrating with smart irrigation technologies. In Egypt, emerging research initiatives are exploring the application of drone-based pest management in the Nile Delta and newly reclaimed desert lands. These efforts are aligned with national strategies aimed at improving agricultural productivity and sustainability under water-limited conditions [1], [37]. Overall, these global case studies demonstrate that autonomous drone technologies are adaptable to a wide range of agricultural systems and environmental conditions. While adoption levels vary across regions, the consistent benefits of improved efficiency, reduced chemical use, and enhanced decision-making highlight the significant potential of UAV-based pest management as a scalable and sustainable solution for modern agriculture. These case studies collectively demonstrate that drone-based pest management systems are highly adaptable across diverse agroecological conditions. Despite variations in scale, infrastructure, and technological maturity, consistent outcomes include reduced pesticide use, improved operational efficiency, and enhanced decision-making. However, the extent of adoption remains strongly influenced by economic capacity, regulatory frameworks, and technical expertise.

2.10. Future Perspectives and Research Gaps

The rapid advancement of autonomous drone technologies, combined with developments in artificial intelligence, sensor systems, and data analytics, is expected to significantly transform pest

management strategies in the coming years. Future research and innovation will focus on enhancing system efficiency, scalability, and adaptability, particularly in challenging environments such as desert agriculture. One of the most promising directions is the development of fully autonomous, closed-loop pest management systems. These systems will integrate real-time sensing, AI-driven analysis, and automated intervention, enabling drones to detect pest infestations, make decisions, and execute control measures without human intervention. Such systems have the potential to greatly improve response time and operational precision while reducing labour dependency.

Advancements in swarm robotics represent another key area of development. Coordinated fleets of drones operating collaboratively can cover large agricultural areas more efficiently than individual UAVs. Swarm intelligence, supported by decentralized AI algorithms, allows for dynamic task allocation, optimized flight paths, and enhanced resilience in large-scale pest management operations [30]. This approach is particularly suitable for desert environments, where fields are often extensive and heterogeneous. Energy efficiency and flight endurance remain critical research priorities. Future innovations are expected to focus on improved battery technologies, lightweight materials, and hybrid energy systems, including solar-powered UAVs. These developments will extend flight duration, increase payload capacity, and improve the feasibility of large-scale operations under harsh environmental conditions.

Sensor technologies are also expected to evolve significantly. The development of low-cost, high-resolution hyperspectral and thermal sensors will enhance pest detection accuracy while improving accessibility. In addition, the integration of biosensors and chemical detection systems may enable the direct identification of pest-related biomarkers, further improving detection capabilities. Another important research direction involves the integration of predictive modelling and climate-based forecasting. By combining historical pest data with real-time environmental information, AI-driven models can anticipate pest outbreaks under changing climatic conditions. This proactive approach is particularly relevant in the context of climate change, which is expected to alter pest distribution, behaviour, and impact on agricultural systems [2], [15]. From a sustainability perspective, future developments will emphasize the integration of eco-friendly pest control strategies. These include the drone-based delivery of biological control agents, the expansion of sterile insect technique (SIT) programs, and the adoption of reduced-risk pesticides. Such approaches align with global efforts to minimize environmental impact and promote sustainable agriculture.

Policy and regulatory frameworks will also play a crucial role in shaping the future of drone-based pest management. Harmonization of UAV regulations, development of standardized operational

protocols, and investment in infrastructure and training programs are essential to support widespread adoption. In addition, improving accessibility for smallholder farmers through cost reduction and service-based models (e.g., drone-as-a-service) will be critical for equitable technology dissemination. Finally, interdisciplinary research will be essential to fully realize the potential of autonomous drone systems. Collaboration among agronomists, entomologists, engineers, data scientists, and policymakers will drive innovation and ensure that technological solutions are both scientifically robust and practically applicable.

Therefore, the future of autonomous drone-based pest management lies in the integration of intelligent technologies, sustainable practices, and supportive policy frameworks. Continued research and development will be key to overcoming existing challenges and unlocking the full potential of these systems in enhancing agricultural productivity and resilience, particularly in desert ecosystems. Despite significant progress, several critical research gaps remain, (i) Limited field validation under extreme desert environmental conditions, (ii) Lack of standardized protocols for UAV-based pesticide application, (iii) Challenges in integrating multi-source data (UAV, IoT, climate models) and (iv) Insufficient economic feasibility studies, particularly for smallholder systems. Addressing these gaps will be essential for enabling large-scale adoption and maximizing the effectiveness of drone-based pest management technologies.

3. Conclusion

Autonomous drones are fundamentally redefining precision pest management in desert agriculture, facilitating a critical transition from conventional uniform spraying to intelligent, data-driven systems. By integrating multispectral sensing with AI and machine learning, these platforms enable the early detection and site-specific treatment of infestations. This precision is particularly vital for fragile arid ecosystems, as it drastically reduces chemical inputs, lowers operational costs, and minimizes non-target environmental impacts, effectively aligning pest control with the principles of climate-smart farming. While the potential is significant, the path to widespread adoption is shaped by the need to overcome technical and environmental barriers. Future efforts must address hardware resilience against extreme heat and dust, while research should pivot toward the delivery of biological control agents and the evaluation of long-term impacts on soil health. Furthermore, the evolution of this field will likely see drones operating in synergy with satellite platforms and ground-based sensors, creating a fully autonomous, interconnected ecosystem for real-time predictive modelling and intervention. To ensure this technology becomes a cornerstone of global food security, policy frameworks must evolve alongside the hardware. Governments and regulatory bodies should prioritize the harmonization of

autonomous flight standards and provide economic incentives to reduce the high entry barriers for end-users in developing regions. By fostering interdisciplinary collaboration between agronomists, engineers, and data scientists, drone-based pest management can move beyond its role as a specialized tool to become a transformative, sustainable standard for resilient crop protection in the world's most challenging environments.

Funding

The authors declare that this research received no external funding.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] FAO, "The State of Food and Agriculture 2022: Leveraging Automation in Agriculture," Food and Agriculture Organization, Rome, 2022.

[2] IPCC, "Climate Change 2022: Impacts, Adaptation and Vulnerability," Intergovernmental Panel on Climate Change, Cambridge University Press, 2022.

[3] V. Popp, K. Peto, and J. Nagy, "Pesticide productivity and food security," Agronomy for Sustainable Development, vol. 33, pp. 243–255, 2013.

[4] P. Nicolopoulou-Stamati, S. Maipas, C. Kotampasi, P. Stamatis, and L. Hens, "Chemical pesticides and human health: The urgent need for a new concept in agriculture," Frontiers in Public Health, vol. 4, 2016.

[5] R. Gebbers and V. I. Adamchuk, "Precision agriculture and food security," Science, vol. 327, pp. 828–831, 2010.

[6] C. Zhang, J. M. Kovacs, "The application of small unmanned aerial systems for precision agriculture: a review," Precision Agriculture, vol. 13, pp. 693–712, 2012.

[7] A.-K. Mahlein, "Plant disease detection by imaging sensors," Annual Review of Phytopathology, vol. 54, pp. 135–153, 2016.

[8] L. Pádua et al., "UAS, sensors, and data processing in agroforestry: a review," Remote Sensing, vol. 12, 2020.

[9] Y. Huang, W. C. Hoffmann, Y. Lan, W. Wu, and B. K. Fritz, "Development of a spray system for unmanned aerial vehicles," Applied Engineering in Agriculture, vol. 36, pp. 1–9, 2020.

[10] X. Xue and Y. Lan, "Agricultural aviation applications in China," Outlooks on Pest Management, vol. 28, pp. 56–59, 2017.

[11] K. G. Liakos et al., "Machine learning in agriculture: A review," Sensors, vol. 18, 2018.

[12] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," Computers and Electronics in Agriculture, vol. 147, pp. 70–90, 2018.
[13] M. A. Altieri and C. I. Nicholls, "Biodiversity and pest management in agroecosystems," 2nd ed., CRC Press, 2004.

[14] D. Pimentel, "Environmental and economic costs of the application of pesticides primarily in the United States," Environment, Development and Sustainability, vol. 7, pp. 229–252, 2005.

[15] C. A. Deutsch et al., "Increase in crop losses to insect pests in a warming climate," Science, vol. 361, pp. 916–919, 2018.

[16] M. E. Whalon, D. Mota-Sanchez, and R. M. Hollingworth, "Global pesticide resistance in arthropods," CABI Publishing, 2008.

[17] J. Xue and B. Su, "Significant remote sensing vegetation indices: A review of developments and applications," Journal of Sensors, 2017.

[18] S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, "Big Data in smart farming – A review," Agricultural Systems, vol. 153, pp. 69–80, 2017.

[19] D. C. Tsouros, S. Bibi, and P. G. Sarigiannidis, "A review on UAV-based applications for precision agriculture," Information, vol. 10, no. 11, p. 349, 2019.

[20] A. Matese and S. F. Di Gennaro, "Technology in precision viticulture: A state of the art review," International Journal of Wine Research, vol. 7, pp. 69–81, 2015.

[21] P. Radoglou-Grammatikis, P. Sarigiannidis, T. Lagkas, and I. Moscholios, "A compilation of UAV applications for precision agriculture," Computer Networks, vol. 172, 2020.

[22] E. R. Hunt Jr., C. S. T. Daughtry, "What good are unmanned aircraft systems for agricultural remote sensing and precision agriculture?" International Journal of Remote Sensing, vol. 39, pp. 5345–5376, 2018.

[23] P. Thenkabail, J. G. Lyon, and A. Huete, "Hyperspectral remote sensing of vegetation," CRC Press, 2012.

[24] S. Sankaran, A. Mishra, R. Ehsani, and C. Davis, "A review of advanced techniques for detecting plant diseases," Computers and Electronics in Agriculture, vol. 72, pp. 1–13, 2010.

[25] H. Qin, X. Xue, and Y. Lan, "Effect of spray parameters of unmanned aerial vehicle on droplet deposition and drift in plant protection," International Journal of Agricultural and Biological Engineering, vol. 9, no. 5, pp. 38–46, 2016.

[26] A. Hassan, "Aerial release of Trichogramma for biological control: A review," Biocontrol Science and Technology, vol. 10, pp. 1–12, 2000.

[27] J. Hendrichs, M. J. B. Vreysen, M. J. Robinson, and J. P. Cayol, "Strategic options in using sterile insects for area-wide integrated pest management," Insects, vol. 12, 2021.

[28] J. van Klompenburg, G. Kassahun, and C. Catal, “Crop yield prediction using machine learning: A systematic literature review,” Computers and Electronics in Agriculture, vol. 177, 2020.

[29] M. Zhang, Q. Wang, and X. Zhao, “Applications of IoT in smart agriculture: A review,” Information Processing in Agriculture, vol. 7, pp. 1–16, 2020.

[30] L. Chung, J. Dayoub, and N. Sünderhauf, “A survey on aerial swarm robotics,” IEEE Transactions on Robotics, vol. 34, pp. 837–855, 2018.

[31] Y. Lan and J. Chen, “Current status and trends of UAV application in agriculture,” International Journal of Agricultural and Biological Engineering, vol. 11, pp. 1–17, 2018.

[32] X. Xue, Y. Lan, et al., “Droplet deposition and efficacy of UAV-based spraying in crop protection,” Pest Management Science, vol. 72, pp. 201–210, 2016.

[33] S. Yamada, “Development of unmanned helicopter for agriculture in Japan,” Journal of Robotics and Mechatronics, vol. 24, pp. 577–582, 2012.

[34] J. L. Peña et al., “Weed and pest detection using UAV imagery,” Sensors, vol. 13, pp. 141–155, 2013.

[35] FAO, “Fall Armyworm Monitoring and Early Warning System (FAMEWS),” Food and Agriculture Organization, 2019.

[36] European Commission, “Farm to Fork Strategy for a fair, healthy and environmentally-friendly food system,” 2020.

[37] World Bank, “Harvesting Prosperity: Technology and Productivity Growth in Agriculture,” 2020.