Pertanika Journal of

SCIENCE & TECHNOLOGY

JST

VOL. 34 (3) JUN. 2026

PERTANIKA JOURNALS

A scientific journal published by Universiti Putra Malaysia Press

Pertanika Journal of Science & Technology
Pertanika Journal of Science & Technology

PERTANIKA JOURNAL OF SCIENCE & TECHNOLOGY

About the Journal

Overview

Pertanika Journal of Science & Technology is an official journal of Universiti Putra Malaysia. It is an open-access online scientific journal. It publishes original scientific outputs. It neither accepts nor commissions third party content.

Recognised internationally as the leading peer-reviewed interdisciplinary journal devoted to the publication of original papers, it serves as a forum for practical approaches in improve quality on issues pertaining to science and engineering and its related fields.

Pertanika Journal of Science & Technology currently publishes 6 issues a year (*February, April*, *June, August, October, and December*). It is considered for publication of original articles as per its scope. The journal publishes in **English** and it is open for submission by authors from all over the world.

The journal is available world-wide.

Aims and Scope

Pertanika Journal of Science & Technology aims to provide a forum for high quality research related to science and engineering research. Areas relevant to the scope of the journal include: bioinformatics, bioscience, biotechnology and bio-molecular sciences, chemistry, computer science, ecology, engineering, engineering design, environmental control and management, mathematics and statistics, medicine and health sciences, nanotechnology, physics, safety and emergency management, and related fields of study.

History

Pertanika Journal of Science & Technology was founded in 1993 and focusses on research in science and engineering and its related fields.

Vision

To publish a journal of international repute.

Mission

Our goal is to bring the highest quality research to the widest possible audience.

Quality

We aim for excellence, sustained by a responsible and professional approach to journal publishing. Submissions can expect to receive a decision within 90 days. The elapsed time from submission to publication for the articles averages 180 days. We are working towards decreasing the processing time with the help of our editors and the reviewers.

Abstracting and Indexing of Pertanika

Pertanika Journal of Science & Technology is now over 33 years old; this accumulated knowledge and experience has resulted the journal being abstracted and indexed in SCOPUS (Elsevier), Journal Citation Reports (JCR-Clarivate), EBSCO, ASEAN CITATION INDEX, Microsoft Academic, Google Scholar, and MyCite.

Citing Journal Articles

The abbreviation for Pertanika Journal of Science & Technology is *Pertanika J. Sci. & Technol.*

Publication Policy

Pertanika policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. It prohibits as well publication of any manuscript that has already been published either in whole or substantial part elsewhere. It also does not permit publication of manuscript that has been published in full in proceedings.

Code of Ethics

The *Pertanika* journals and Universiti Putra Malaysia take seriously the responsibility of all of its journal publications to reflect the highest in publication ethics. Thus, all journals and journal editors are expected to abide by the journal's codes of ethics. Refer to *Pertanika*'s **Code of Ethics** for full details, or visit the journal's web link at http://www.pertanika.upm.edu.my/code_of_ethics.php

Originality

The author must ensure that when a manuscript is submitted to *Pertanika*, the manuscript must be an original work. The author should check the manuscript for any possible plagiarism using any programme such as Turn-It-In or any other software before submitting the manuscripts to the Editorial Office, *Pertanika* Journal Publication Section.

All submitted manuscripts must be in the journal's acceptable similarity index range:

≤ 20% – PASS; > 20% – REJECT.

International Standard Serial Number (ISSN)

An ISSN is an 8-digit code used to identify periodicals such as journals of all kinds and on all media–print and electronic.

Pertanika Journal of Science & Technology: e-ISSN 2231-8526 (Online).

Lag Time

A decision on acceptance or rejection of a manuscript is reached in 90 days (average). The elapsed time from submission to publication for the articles averages 180 days.

Authorship

Authors are not permitted to add or remove any names from the authorship provided at the time of initial submission without the consent of the journal's Chief Executive Editor.

Manuscript Preparation

For manuscript preparation, authors may refer to Pertanika's **INSTRUCTION TO AUTHORS**, available on the official website of *Pertanika*.

Editorial Process

Authors who complete any submission are notified with an acknowledgement containing a manuscript ID on receipt of a manuscript, and upon the editorial decision regarding publication.

Pertanika follows a **double-blind peer-review** process. Manuscripts deemed suitable for publication are sent to reviewers. Authors are encouraged to suggest names of at least 3 potential reviewers at the time of submission of their manuscripts to *Pertanika*, but the editors will make the final selection and are not, however, bound by these suggestions.

Notification of the editorial decision is usually provided within 90 days from the receipt of manuscript. Publication of solicited manuscripts is not guaranteed. In most cases, manuscripts are accepted conditionally, pending an author's revision of the material.

The Journal's Peer Review

In the peer-review process, 2 to 3 referees independently evaluate the scientific quality of the submitted manuscripts. At least 2 referee reports are required to help make a decision.

Peer reviewers are experts chosen by journal editors to provide written assessment of the **strengths** and **weaknesses** of the written research, with the aim of improving the reporting of research and identifying the most appropriate and highest quality material for the journal.

Operating and Review Process

What happens to a manuscript once it is submitted to *Pertanika*? Typically, there are 7 steps to the editorial review process:

1. The journal's Chief Executive Editor and Editor-in-Chief examine the paper to determine whether it is relevance to the journal's needs in terms of novelty, impact, design, procedure, language as well as presentation and allow it to proceed to the reviewing process. If it is not appropriate, the manuscript is rejected outright and the author is informed.

2. The Chief Executive Editor sends the article-identifying information having been removed, to 2 to 3 reviewers. They are specialists in the subject matter of the article. The Chief Executive Editor requests that they complete the review within 3 weeks.

   Comments to authors are about the appropriateness and adequacy of the theoretical or conceptual framework, literature review, method, results and discussion, and conclusions. Reviewers often include suggestions to strengthening of the manuscript. Comments to the editor are in the nature of the significance of the work and its potential contribution to the research field.

Pertanika Journal of Science & Technology
Pertanika Journal of Science & Technology

Pertanika Journal of Science & Technology
Pertanika Journal of Science & Technology

3. The Editor-in-Chief examines the review reports and decides whether to accept or reject the manuscript, invite the authors to revise and resubmit the manuscript, or seek additional review reports. In rare instances, the manuscript is accepted with almost no revision. Almost without exception, reviewers' comments (to the authors) are forwarded to the authors. If a revision is indicated, the editor provides guidelines to attend to the reviewers' suggestions and perhaps additional advice on how to revise the manuscript.

4. The authors decide whether and how to address the reviewers' comments and criticisms and the editor's concerns. The authors return a revised version of the paper to the Chief Executive Editor along with specific information describing how they have addressed' the concerns of the reviewers and editor, usually in a tabular form. The authors may also submit a rebuttal if there is a need especially when the authors disagree with certain comments provided by the reviewers.

5. The Chief Executive Editor sends the revised manuscript out for re-review. Typically, at least 1 of the original reviewers will be asked to examine the article.

6. When the reviewers have completed their work, the Editor-in-Chief examines their comments and decides whether the manuscript is ready to be published, needs another round of revisions, or should be rejected. If the decision is to accept, the Chief Executive Editor is notified.

7. The Chief Executive Editor reserves the final right to accept or reject any material for publication, if the processing of a particular manuscript is deemed not to be in compliance with the S.O.P. of *Pertanika*. An acceptance letter is sent to all the authors.

   The editorial office ensures that the manuscript adheres to the correct style (in-text citations, the reference list, and tables are typical areas of concern, clarity, and grammar). The authors are asked to respond to any minor queries by the editorial office. Following these corrections, page proofs are mailed to the corresponding authors for their final approval. At this point, **only essential changes are accepted**. Finally, the manuscript appears in the pages of the journal and is posted online.

Pertanika Journal of

SCIENCE & TECHNOLOGY

Vol. 34 (3) Jun. 2026

PERTANIKA JOURNALS

A scientific journal published by Universiti Putra Malaysia Press

PJST
Pertanika Journal of Science & Technology
AN INTERNATIONAL PEER-REVIEWED JOURNAL

UNIVERSITY PUBLICATIONS COMMITTEE

CHAIRMAN
Zamberi Sekawi

EDITOR-IN-CHIEF
Mohamed Thariq Hameed Sultan
SEARCA Chair of Agriculture Technology and Innovation

JOURNAL SECTION
Head of Section:
Fatimah Abdul Samad

Journal Officers:
Ain Haziqah Ab Razak
Navaneetha Krishna Chandran
Nur Izni Shafiqah Mohd Hamidi
Saundarya Saran

Editorial Assistants:
Zulinaardawati Kamarudin

Pre-press Officers:
Aisyah Umairah Jasmin
Engku Erina Zahirah Engku Mohd Suhaimi
Ku Ida Mastura Ku Baharom

IT Officer:
Kiran Raj Kaneswaran

Journal Section Office
Putra Science Park
1st Floor, IDEA Tower II
UPM-MTDC Technology Centre
Universiti Putra Malaysia
43400 UPM Serdang
Selangor, Malaysia

General Enquiry
Tel. No: +603 9769 1622
E-mail:
pertanika.proceeding@upm.edu.my
URL:
http://www.pertanika.upm.edu.my/

PUBLISHER
UPM Press Centre
Universiti Putra Malaysia
43400 UPM Serdang
Selangor, Malaysia
Tel: +603 9769 3606
E-mail: dir.penerbit@upm.edu.my
URL: http://penerbit.upm.edu.my

ASSOCIATE EDITOR
2024-2026

Miss Laiha Mat Kiah
Security Services Sn: Digital Forensic, Steganography, Network Security, Information Security, Communication Protocols, Security Protocols
Universiti Malaya, Malaysia

Saidur Rahman
Renewable Energy, Nanofluids, Energy Efficiency, Heat Transfer, Energy Policy
Sunway University, Malaysia

EDITORIAL BOARD
2024-2026

Abdul Latif Ahmad
Chemical Engineering
Universiti Sains Malaysia, Malaysia

Hsiu-Po Kuo
Chemical Engineering
National Taiwan University, Taiwan

Mohd. Ali Hassan
Bioprocess Engineering, Environmental Biotehnology
Universiti Putra Malaysia, Malaysia

Ahmad Zaharin Aris
Hydrochemistry, Environmental Chemistry, Environmental Forensics, Heavy Metals
Universiti Putra Malaysia, Malaysia

Ivan D. Rukhlenko
Nonliner Optics, Silicon Photonics, Plasmonics and Nanotechnology
The University of Sydney, Australia

Nor Azah Yusof
Biosensors, Chemical Sensor, Functional Material
Universiti Putra Malaysia, Malaysia

Azlina Harun@ Kamaruddin
Enzyme Technology, Fermentation Technology
Universiti Sains Malaysia, Malaysia

Lee Keat Teong
Energy Environment, Reaction Engineering, Waste Utilisation, Renewable Energy
Universiti Sains Malaysia, Malaysia

Norbahiah Misran
Communication Engineering
Universiti Kebangsaan Malaysia, Malaysia

Bassim H. Hameed
Chemical Engineering: Reaction Engineering, Environmental Catalysis & Adsorption
Qatar University, Qatar

Mohamed Othman
Communication Technology and Network, Scientific Computing
Universiti Putra Malaysia, Malaysia

Roslan Abd-Shukor
Physics & Materials Physics, Superconducting Materials
Universiti Kebangsaan Malaysia, Malaysia

Biswajeet Pradhan
Digital image processing, Geographical Information System (GIS), Remote Sensing
University of Technology Sydney, Australia

Mohd Shukry Abdul Majid
Polymer Composites, Composite Pipes, Natural Fibre Composites, Biodegradable Composites, Bio-Composites
Universiti Malaysia Perlis, Malaysia

Sodeifian Gholamhossein
Supercritical technology, Optimisation, nanoparticles, Polymer nanocomposites
University of Kashan, Iran

Ho Yuh-Shan
Water research, Chemical Engineering and Environmental Studies
Asia University, Taiwan

Mohd Zulkifly Abdullah
Fluid Mechanics, Heat Transfer, Computational Fluid Dynamics (CFD)
Universiti Sains Malaysia, Malaysia

Wing Keong Ng
Aquaculture, Aquatic Animal Nutrition, Aqua Feed Technology
Universiti Sains Malaysia, Malaysia

INTERNATIONAL ADVISORY BOARD
2024-2027

Hiroshi Uyama
Polymer Chemistry, Organic Compounds, Coating, Chemical Engineering
Osaka University, Japan

Mohini Sain
Material Science, Biocomposites,Biomaterials
University of Toronto, Canada

Mohamed Pourkashanian
Mechanical Engineering, Energy, CFD and Combustion Processes
Sheffield University, United Kingdom

Yulong Ding
Particle Science & Thermal Engineering
University of Birmingham, United Kingdom

ABSTRACTING AND INDEXING OF PERTANIKA JOURNALS

Pertanika Journal of Science & Technology is indexed in Journal Citation Reports (JCR-Clarivate), SCOPUS (Elsevier), EBSCO, ASEAN Citation Index, Microsoft Academic, Google Scholar and MyCite.

The publisher of Pertanika will not be responsible for the statements made by the authors in any articles published in the journal. Under no circumstances will the publisher of this publication be liable for any loss or damage caused by your reliance on the advice, opinion or information obtained either explicitly or implied through the contents of this publication.
All rights of reproduction are reserved in respect of all papers, articles, illustrations, etc., published in Pertanika. Pertanika provides free access to the full text of research articles for anyone, web-wide. It does not charge either its authors or author-institution for refereeing/publishing outgoing articles or user-institution for accessing incoming articles.
No material published in Pertanika may be reproduced or stored on microfilm or in electronic, optical or magnetic form without the written authorisation of the Publisher.
Copyright ©2026 Universiti Putra Malaysia Press. All Rights Reserved.

Pertanika Journal of Science & Technology
Vol. 34 (3) Jun. 2026

Contents

Foreword i
Mohamed Thariq Hameed Sultan

Review Article 1287
Energy Consumption and Optimisation in Indoor Farming: A Review
Jia Quan Goh, Aimrun Wayayok, Diyana Binti Jamaludin, and Muhamad Saufi Mohd Kassim

Valorisation of Oil Palm Waste into Magnetic Catalysts for Efficient Synthesis of Bio-additive Ethyl Levulinate: Optimisation of Reaction Parameters 1309
Shamala Gowri Krishnan and Pua Fei Ling

YOLOv9-ResCBAM: Enhancing Tomato Ripeness Detection in Greenhouses Through Advanced Object Detection 1331
Jalal Uddin Md Akbar, Syafiq Fauzi Kamarulzaman, Muhammad Danial Mohamad Rizwan, Riadul Islam Rabbi, and Ekramul Haque Tusher

Review Article 1355
Exploring Educational User Engagement: A Systematic Mapping of Serious Games with Gesture-based Interaction
Wan Salfarina Wan Husain, Syadiah Nor Wan Shamsuddin, and Normala Rahim

Utilising Innovative Image Preprocessing Techniques in Vehicle Plate Detection and Recognition 1383
Soheir Noori, Asam Almohamed, Elham Mohammed Thabit A. Alsaadi, and Azhar Abbas

Stability, Thermo-physical, and Tribological Properties of SiO_2 Nanoparticles in a Vegetable Oil-based Biodegradable Lubricant for Minimum Quantity Lubrication Machining 1409
Nurlisa Hamzan, Muhammad Amir Syafiq Mohd Tarmizi, Nurul Huda Abdul Azid, Ahsana Aqilah Ahmad, Afifah Zakiyyah Juri, and Irnie Azlin Zakaria

Silicon-based Fin Field-Effect Transistor (Si-FinFET) Optimisation Based on Gate Length and Operating Temperature 1433
Yousif Atalla, Mohamad Hafiz Mamat, and Yasir Hashim

Design and Development of Unmanned Aerial Vehicles' Propellers Based on Biomimetic Design 1451
Saundarya Saran, Rosalam Che Me, Mohamed Thariq Hameed Sultan, and Navaneetha Krishna Chandran

A Zero-shot Foundation Model Approach for Real-time Adaptive Cyber Threat Detection 1469
Muhmmad Al-Khiza'ay and Noora Alallaq

Automated Generation of Calculation-based Physics Questions Using Text-to-Text Transfer Transformer (T5) for Malaysian Upper Secondary Education 1493
Tuong Kiet Ngang and Chih How Bong

D-Xotic: A Deep Learning Improvement for Recognising Different Durian Cultivars Based on Leaves Signature 1525
Aulia Brilliantina, Tri Agus Siswoyo, Yuli Witono, and Bayu Taruna Widjaja Putra

Noise-Resilient and Low Complexity LACO-OFDM Using Approximation Log-Likelihood Ratio (approxLLR) Soft-decision Demodulation with Single-FFT Receiver 1551
Norzalina Othman and Mohd Rashidi Che Beson

Review Article 1579
A Bibliometric Review of Remote Sensing Applications in Water Surge Risk Mapping for Waterfall Areas
Mohammed M. Y. Albarqouni, Syarifuddin Misbari, Nadiatul Adilah Ahmad Abdul Ghani, and Mohd Faizal Jamlos

Carbon Emissions Quantification from Ceramic Tile Production by Using Life-Cycle Assessment (LCA): A Case Study of PT Sinar Karya Duta Abadi, Indonesia 1599
Ransi Pasae, M. Rijal Idrus, Hazairin Zubair, and Darhamsyah

Field Validation of IES-VE for Daylight Assessment in Tropical High-rise Low-cost Housing: Effects of Building Height, Layout Depth, and Façade Design 1625
Nor Izyan Saleh, Mohd Firrdhaus Mohd Sahabuddin, Norafida Ab Ghafar, Zuraina Leily Awaluddin, and Bismiazan Abd.Razak

Gamification for Extended Reality Informative Space (XRIS) Visualisation in the Construction Industry 1655
Changsaar Chai, Anne Ann Yii Chua, and Felix Kai Han Yong

Studying the Effects of Knowledge Distillation and the Fragility of Lightweight BERT-based Models 1685
Allistair Nallie Konsil, Stephanie Chua, Jacey Lynn Minoi, Md Mizanur Rahman, Gerraint Gillan, Rafazila Ramli, Rasitasam Safii, Ahmad Sofian bin Shminan, and Lee Jun Choi

Enhanced Deepfake Image Detection and Classification Using Deep Learning 1703
Debasish Samal, Dimple Nagpal, Prateek Agrawal, Vishu Madaan, Chhavi Sharma, and Wou Onn Choo

Occupational Health Implications of High-Frequency Whole-Body Vibration Exposure in Mine Workers 1727
S. Joel Vinoth Singh, I. Siva, JT. Winowlin Jappes, and Chithirai Pon Selvan

A LiDAR and IoT Integrated Approach for Intelligent Monitoring of Lubricant Liquid Levels: A Lab Scale Study on Petronas Products 1745
Nur Shaebah Mat Hussain, Mohd Khairol Anuar Mohd Ariffin, Siti Azfanizam Ahmad, and Zulkifli Mohamed

Human-in-the-Loop Soft Prompt Tuning for Adverse Drug Event Extraction from Clinical Notes 1769
Salisu Modi, Khairul Azhar Kasmiran, Nurfadhlina Mohd Sharef, and Mohd Yunus Sharum

Empirical Analysis of Factors in User Control Model for Cloud Data Migration 1801
Ishak Iskandar, Imbaji Injuwe Danga, Hamidah Ibrahim, and Fatimah Sidi

Towards Inherently Safer Dry Reforming for Carbon Capture: Trends, Gaps, and Insights 1829
Fatin Aliya, Mardhati Zainal Abidin, and Azizul Buang

An IoT-Enabled Decision Support System for Real-time Line Balancing in Semiconductor Manufacturing 1861
Nur Ain Qistina Muhammad Shafee, Effendi Mohamad, Mohd Soufhwee Abd Rahman, Arfauz A. Rahman, and Teruaki Ito

Review Article 1881
Optical Sensing Techniques for Microplastic Identification: Comparative Evaluation, Limitations, and Future Recommendations
Muhammad Nurullah Waliyullah Mohamed Nazli, Irneza Ismail, Fatin Hamimi Mustafa, Wan Zakiah Wan Ismail, Juliza Jamaludin, and Latiffah Karim

Quantitative Evaluation of Machine Learning-based Land Use Classification with Multispectral Satellite Data 1905
Krati Bansal, Anindita Nath, Tanupriya Choudhury, Bappaditya Koley, and Jitendra Rajpurohit

Multi-task Deep Learning Pipeline for Rice Field Classification and Growth Monitoring Using Drone Imagery 1933
Youssef Yasser Salaheldin Elhammamy, Chung Gwo Chin, Gan Ming Tao, Chan Kah Yoong, and Pang Wai Leong

Foreword

Welcome to the third issue of 2026 of the Pertanika Journal of Science and Technology (PJST)!

PJST is an open-access journal for studies in Science and Technology published by Universiti Putra Malaysia Press. It is independently owned and managed by the university for the benefit of the world-wide science community.

This issue contains 27 articles: four review articles; and the rest are regular articles. The authors of these articles come from different countries namely India, Indonesia, Iraq, Japan, Malaysia, Nigeria, Oman, Qatar, United Arab Emirates, and United Kingdom.

The article entitled "Design and Development of Unmanned Aerial Vehicles' Propellers Based on Biomimetic Design" presents an innovative nature-inspired approach to improving the aerodynamic performance of unmanned aerial vehicles. Drawing on the geometry and flight characteristics of eagle wings and feathers, the study develops four Eagle Feather Biomimicry propeller configurations and evaluates them against a conventional design using computational fluid dynamics at 2000, 4000, and 6000 rpm. The designs are assessed in terms of thrust, lift, torque, acoustic behaviour, power consumption, and efficiency. The findings demonstrate that the biomimetic modifications offer distinct performance advantages for different operational needs: EAFB 4 produced the highest thrust across the tested rotational speeds, EAFB 2 generated the greatest lift at lower power demand, and EAFB 5 showed favourable power efficiency at higher rotational speeds. By showing how biological principles can be translated into functional engineering geometry, the work contributes to the development of quieter, more stable, and energy-efficient UAV propulsion systems. The proposed designs have potential applications in agriculture, inspection, surveillance, and other missions requiring improved endurance and aerodynamic reliability. Further details of this study can be found on page 1451.

The issue also features the article titled "Occupational Health Implications of High-Frequency Whole-Body Vibration Exposure in Mine Workers", which highlights an important occupational safety concern in mining and heavy-machinery operations. Whole-body vibration measurements were collected under actual field conditions from operators of crushers, dumpers, and wheel loaders using a tri-axial accelerometer and a human vibration measurement system. The recorded signals were processed in accordance with ISO 2631-1 to determine root mean square acceleration, daily exposure value A(8), vibration dose value, and associated health-risk levels. The results show that crusher

operators were generally exposed to vibration within acceptable limits, whereas dumper and wheel-loader operators experienced substantially higher levels, particularly along the vertical axis, increasing their risk of long-term musculoskeletal and neurological disorders. The study therefore emphasises the need for engineering controls such as improved seat suspension and vibration isolation, together with administrative measures including job rotation, reduced exposure duration, and structured work-rest schedules. By combining field measurements with internationally recognised assessment criteria, this research provides practical evidence for occupational health management and supports the adoption of real-time monitoring and data-driven interventions to protect mine workers. Further details of this study can be found on page 1727.

Another noteworthy contribution is the article entitled "A LiDAR and IoT Integrated Approach for Intelligent Monitoring of Lubricant Liquid Levels: A Lab-Scale Study on PETRONAS Products", which addresses the need for continuous and reliable lubricant monitoring in modern industrial systems. The study develops a laboratory-scale prototype that combines a non-contact LiDAR sensor, an ESP32 microcontroller, and an IoT platform to measure lubricant levels and transmit real-time information for remote monitoring. The system was tested using PETRONAS ETRO 6+ base oil and PETRONAS Syntium 800 10W-40, with repeated measurements used to evaluate consistency and stability. Standard deviations of 10.6 mm and 9.27 mm were obtained, respectively, indicating acceptable measurement performance, with the more optically opaque Syntium lubricant producing slightly greater precision. The integration of cloud-based visualisation and automated data collection reduces dependence on manual inspection and supports timely maintenance decisions. This contribution is particularly relevant to predictive maintenance, industrial automation, and smart manufacturing, where early detection of low lubricant levels can prevent equipment damage, minimise unplanned downtime, and improve operational efficiency. The study also provides a practical foundation for future validation under vibration, temperature variation, oil mist, and other real industrial conditions. Full details of this study are available on page 1745.

We anticipate that you will find the evidence presented in this issue to be intriguing, thought-provoking and useful in reaching new milestones in your own research. Please recommend the journal to your colleagues and students to make this endeavour meaningful.

All the papers published in this edition underwent Pertanika's stringent peer-review process involving a minimum of two reviewers comprising internal as well as external referees. This was to ensure that the quality of the papers justified the high ranking of the journal, which is renowned as a heavily-cited journal not only by authors and researchers in Malaysia but by those in other countries around the world as well.

We would also like to express our gratitude to all the contributors, namely the authors, reviewers and Editorial Board Members of PJST, who have made this issue possible.

PJST is currently accepting manuscripts for upcoming issues based on original qualitative or quantitative research that opens new areas of inquiry and investigation.

Editor-in-Chief
Mohamed Thariq Hameed Sultan

Pertanika J. Sci. & Technol. 34 (3): 1287 - 1308 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Review Article

Energy Consumption and Optimisation in Indoor Farming: A Review

Jia Quan Goh[1]*, Aimrun Wayayok[1,2], Diyana Binti Jamaludin[1], and Muhamad Saufi Mohd Kassim[1]

[1]Department of Biological and Agricultural Engineering, Faculty of Engineering, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
[2]SMART Farming Technology Research Centre (SFTRC), Department of Biological and Agricultural Engineering, Faculty of Engineering, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

ABSTRACT

This review explores the key components of crop production and energy consumption in indoor farming, where it usually consists of systems such as LED lighting, ventilation (fans), humidity and hydroponics systems to offer controlled environments for crop growth. However, these systems consume large amounts of energy. Thus, it is necessary to reduce this energy consumption in crop production. This review examines three modelling approaches for energy optimisation, namely: white box (physics-based), black box (data-driven), and grey box (a hybrid of the two). Each method's strengths and limitations are discussed in terms of their application to indoor farming. The findings emphasise the importance of selecting the right optimisation model based on specific goals and resources, with the potential to significantly improve energy efficiency in indoor farming. Further research is encouraged to refine these models and support sustainable food production.

Keywords: Controlled environment, crop production, energy efficiency, lighting system, model types

ARTICLE INFO

Article history:
Received: 23 May 2025
Accepted: 27 December 2025
Published: 12 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.01

E-mail addresses:
jiaquangoh@hotmail.com (Jia Quan Goh)
aimrun@upm.edu.my (Aimrun Wayayok)
diyanaupm@upm.edu.my (Diyana Binti Jamaludin)
saufi@upm.edu.my (Muhamad Saufi Mohd Kassim)
* Corresponding author

INTRODUCTION

Malaysia is a developing country where its urbanisation is faster than that of its Southeast Asian neighbours. The urbanisation of the country is expected to reach 88% in 2050. Urbanisation is a process where land is used to build tall buildings. Hence, there will be less space for agricultural activities in the cities. Arising from it, food security

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

has become a problem. To ensure a sufficient food supply, food must be transported from rural areas. This will cause a rise in the food price. During the COVID-19 pandemic that affected 186 countries between December 2019 and March 2020, it was observed that there was a significant increase in the risk of severe food insecurity. The number of people facing such insecurity rose from 135 million in January 2020 to 265 million by the end of the year (Food Security Information Network, 2020). The pre-COVID era allowed food to be transported worldwide, and this mode was stopped suddenly during the outbreak of the pandemic. City lockdowns and shop closures have caused limited access to farm inputs such as seeds, fertilisers, and labour. This had resulted in low production of vegetables (Sridhar et al., 2023). According to Gregorio and Ancog (2020), the pandemic had caused the agricultural production volume in Southeast Asia during the first quarter of 2020 to reduce by 3.11% due to labour shortage and 100.77 million individuals were affected by it.

There is a solution proposed by Professor Dickinson Despommier (1999) to grow vegetables and fruits in the empty spaces in the cities, including rooftops, old buildings and factories. This concept is known as vertical farming. By the definition of the Cambridge Dictionary, vertical farming is the activity of growing crops in many layers, one above the other, inside a building or under the ground, often in a specially controlled environment. The statistics of the 2021 Global CEA Census Report showed that the common crops being grown in vertical farming included salad greens (58%), herbs (49%), microgreens (46%) and other leafy greens (40%).

There are several key components in a vertical farm. They are the multi-layer growing shelves, artificial lighting, air conditioning system, irrigation system (soil medium), hydroponic system (soilless medium) and ventilation system. In most vertical farming systems, hydroponic systems are used to replace soil as a growing medium that provides nutrients to the crops. Compared to conventional greenhouses or open fields, hydroponic farming requires a lot less fresh water. Due to the closed system, there is almost no evaporation and no runoff. A larger interest in vertical farming and hydroponic farming will emerge when fresh water becomes a rarer resource, reducing the requirement for water in agriculture.

One of the most concerning issues of the operation of vertical farming in the cities is the energy consumption. The operation of vertical farming consumes a large amount of electricity. Artificial lighting, temperature control, and hydroponics systems are the most energy-consuming components of an indoor farm. According to the 2021 Global CEA Census Report, vertical farms have a significantly higher average energy use at 38.8kWh per kg of produce as opposed to traditional greenhouses, which average 5.4kWh per kg of produce, which is about 7 times higher. In the same report, it was mentioned that the biggest sources of energy consumption in vertical farming are lighting (55%), cooling or ventilation (30%) and heating (11%). Several studies reported that the light-emitting

diode (LED) alone accounts for about 55 to 65% of vertical farm operations. Electricity for lighting has been found to be the greatest energy consumer in vertical farms.

This paper has the following objectives: (1) to provide an overview of the energy usage of vertical farms; and (2) to identify the current research in the energy-saving methods of vertical farm operation. The advantages and disadvantages of each type of energy optimisation approach are also discussed in this paper.

REVIEW METHODOLOGY

To ensure a comprehensive understanding of the background and current state of vertical farming technologies, a search was conducted across four databases: Google Scholar, ScienceDirect, ResearchGate and Google Search (as a supplementary). The search was conducted using a few specific keywords, including: "vertical farming", "indoor farming", "hydroponics", "energy consumption", "lighting", "energy optimisation", and "energy saving". Most of the selected papers were published after 2015, except those that explained the most fundamental theory of related topics, such as the history of lighting and indoor vertical farming. After selection, the papers were categorised according to black-box, white-box and grey-box modelling approaches.

Cultivation Methods

Hydroponic System

Plants or crops can be grown without soil in this system. Instead of soil, the roots are immersed in a nutrient-rich water solution that supplies the nutrients. The use of water is effective. Since water resources are also scarce nowadays, the saving of water is also becoming the focus of the agricultural field. By 2050, more than half (57%) of the global population will live in regions where water scarcity occurs for at least one month each year (Boretti and Rosa, 2019). Soil cultivation consumes a large amount of freshwater. Therefore, a method to save water in the agricultural field is important. In indoor farming using hydroponics systems, 70 -95% of water can be saved (Mir et al., 2022). Romeo et al. (2018) also supported this statement by the findings, where hydroponic systems consumed water seven times lower than greenhouse and four times lower than open-field agriculture.

Aeroponics System

This is a type of soilless cultivation method where plant roots are suspended in a closed chamber and sprayed with nutrient solution to provide the essential nutrients for root growth. This method allows root access directly to oxygen and can have a faster growth rate compared to the traditional cultivation method. Aeroponics systems also reduce water consumption compared to soil-based cultivation. Another advantage of it is the minimal space usage that makes it suitable for application in cities where land availability is limited.

Aquaponic System

It is a symbiotic ecosystem where aquaculture and hydroponics are integrated in a recirculating environment. In this system, the waste from the fish provides an organic food source for the plants. The plant acts as a filter to purify the water, and then the water is cycled back to the fish tank. This symbiotic relationship reduces the need for fertilisers and allows for the opportunity of dual harvest of crops and fish.

Components of Vertical Farming

Lighting

By growing crops indoors, artificial lighting is essential to replace natural sunlight that provides energy for the photosynthesis reaction. In the early stages, fluorescent lamps were used as the source of lighting for crop growth (Wheeler, 2008). More than 10% of the total photon emission within the photosynthetically active radiation (PAR) spectrum (400-700 nm) is produced by these fluorescent lights (Bantis et al., 2018). Fluorescent lights, on the other hand, lack the necessary energy to enable the growth of highly productive plants. Throughout use, the photon output gradually decreases. In commercial indoor farms, high-intensity discharge (HID) lighting was also used. The spectrum output source of light can be fixed with an HID lamp to grow crops effectively. The main drawback of HID lamps is that they are extremely difficult to handle thermally in a plant factory and can scorch leafy greens at close separation distances (Mitchell & Sheibani, 2020). McCree (1972) published an article that defined the wavelength range of PAR as 400-700 nm. This has encouraged plant factory researchers to use specific narrow-band lighting within plant factories. LEDs were brought into this as they are waveband selectable. Nowadays, LED has become the most popular artificial lighting being applied in vertical farms.

Temperature and HVAC Control

Initially, it was difficult for vertical farms to effectively control the temperature across multiple growing environments. However, improvements in temperature management have been made because of developments in HVAC (heating, ventilation and air conditioning) systems, including the incorporation of energy-efficient cooling and heating technology. Real-time monitoring and changes are made possible by the development of accurate climate modelling and automation systems, assuring the ideal temperature and humidity levels for varied crops at various growth phases.

Energy Consumption of Vertical Farming

The usage of artificial lighting, heating, and cooling systems, as well as the requirement for irrigation and ventilation, all contribute to the energy demand in indoor farming.

Compared to traditional farming, indoor farming has been found to have higher energy consumption. The structure is established with thermal insulation and is completely isolated from the outdoor environment. The advantage of this design is the protection of the crops from the negative effects of pests and weather. To replace sunlight and control the indoor climate, artificial lighting (LED) and heating, ventilation, and air conditioning (HVAC) systems are used to create optimal conditions for crop growth. Therefore, the main energy consumption of indoor farming is lighting and HVAC (Graamans et al., 2018).

Energy use per kilogram of product varied widely. More than half of the farms used less than 10 kWh per kilogram, and 20% used even less, under 1 kWh per kilogram. But some farms used a lot more energy, which pulled up the average. So, the middle value (median) was 5.4 kWh per kilogram, but the average was much higher at 22.5 kWh per kilogram (Horomia & Gordon-Smith, 2021). The comparison was shown in Figure 1.

Lighting has been concluded to be the biggest energy consumer in vertical farms. According to the 2021 Global CEA Census report, lighting accounted for 55% of the energy consumption. In addition, based on the research of Avgoustaki and Xydis (2020), lighting costs are approximately 80% of the electricity consumed, while the total electricity reaches up to 40% of the total operating expenses. The significant electricity usage for lighting is primarily due to the crops requiring extended periods of artificial light for photosynthesis, especially when natural sunlight is not available.

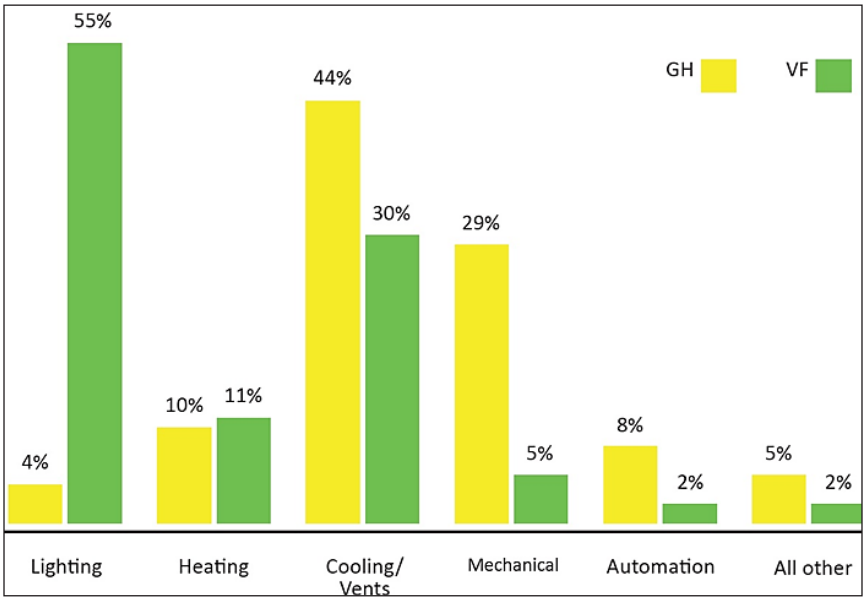


Figure 1. Breakdown of energy consumption in greenhouses (GH) and vertical farms (VF) (Horomia & Gordon-Smith, 2021)
Note. y-axis represents the percentage of energy consumption

The required light levels for office buildings should be between 300 and 600 lux, whereas the value for indoor lettuce growing is between 10000 and 18000 lux (Eaton et al., 2021). In traditional outdoor farming, sunlight acted as the sole source of light energy for photosynthesis. Most of the crops growing indoors require a photoperiod of 12-16 hours for photosynthesis. Thus, electricity expense for lighting is unavoidable in indoor farming. According to Ampim et al. (2022), if developed countries such as the United States adopt vertical farms extensively, the country's energy needs can increase up to eight times from the current amount generated by all the power plants.

In the energy profiling research of a plant factory by Shaari et al. (2021), air conditioning systems or HVAC (heating, ventilation and air conditioning) have higher energy consumption (51%) compared to lighting systems (36%). In their plant factory, the air conditioning system runs longer than the LED lights. This is an exception to the research concluded by other scholars mentioned above. However, in the report of the American Council for an Energy-Efficient Economy, HVAC systems are identified as the second most significant energy consumers. Their energy usage can range from 25% to 50% of the total energy consumption, depending on factors such as location, facility type, and the specific crops being cultivated. Another main reason for the high energy consumption of HVAC is the internal heat loads. According to Graamans et al. (2020), heat loads are caused by the high-density crop production, limited volume and lack of natural ventilation. Heat loads required more energy for the cooling and ventilation system to maintain the optimal temperature inside. They advised the future designer to include a better façade for natural heat dissipation. This suggestion was proven by Sohn et al. (2023), who included more outlets for airflow in their container farm and successfully improved the airflow. Natarajan et al. (2022) also stated that by adding external fans to indoor farms, the temperature can be decreased. This natural temperature control can save energy as well.

In addition, most indoor farming uses a hydroponic system to grow vegetables instead of soil-based methods. The pumping system is working continuously twenty-four hours per day to deliver water that contains essential nutrients throughout the entire system for crop growth (Sihombing et al., 2018). In fact, they proved that continuous watering is not necessary for the hydroponic system, in which they found that intermittent operation of the pumping is a better solution to reduce the energy consumed.

Earlier research by Barbosa et al. (2015) compared the energy consumption between greenhouse and outdoor methods for lettuce planting. They found that the greenhouse produced more yield than conventional methods but consumed 82 times more energy per kilogram produced. The most energy-consuming parts were heating and cooling, lighting and the pumping system. Besides, extensive research by Arcasi et al. (2024) compared the energy consumption of vertical farms in three different cities, i.e., Naples, Riyadh and Stockholm. They represent three climatic conditions: hot-summer Mediterranean climate, hot desert climate and warm summer humid continental climate.

The lighting system consumed 65% to 85% of the energy, followed by HVAC. The circulating fan consumed less than 1% energy. The increase in photosynthetic photon flux density (PPFD) required also increased the total energy consumption of three farms. PPFD measures how many PAR photons are arriving at a specific leaf area (μ mol/ (m^{-2}/s)).

Despite the numerous advantages that vertical farming offers over traditional agricultural practices, high energy consumption emerges as a pivotal challenge. Vertical farming systems can have significant energy needs, which raises expenses and has an adverse effect on the environment. As a result, improving energy efficiency in vertical farming systems has emerged as a crucial factor in the pursuit of sustainable agriculture. Researchers from across the globe have been diligently exploring diverse methods to enhance the energy efficiency of vertical farms. To implement sustainable agriculture, the next section will evaluate the technology and strategies for maximising energy efficiency in vertical farming systems. The challenges and opportunities for optimising energy efficiency in vertical farming systems will be highlighted by examining current practices and energy consumption patterns.

Energy Optimisation Review

There have been some changes in the technologies used in vertical farming over the years to reduce energy usage and increase productivity. In this review, we divided the related studies into three major types, i.e., physical structure, data-driven and hybrid. The categories reflect the most popular approaches in indoor farming energy optimisation. By analysing these three types, we can better understand how each contributes to advancements in the field.

In modelling processes, there are three main categories, i.e., the black-box, white-box and grey-box modelling (Arendt et al., 2018; De Buck et al., 2023; Zaidan et al., 2019). Black-box modelling focuses on data-driven approaches, and white-box modelling is based on physical structure. Grey-box is a mixture of both.

White-box Modelling

With the growing trend of computer technology and artificial intelligence, researchers have begun to employ increasingly advanced methods for optimising energy usage in vertical farming. White-box modelling simulates the energy consumption patterns of a whole building by considering building geometry, internal heat levels, sublevel systems and operation parameters (Kim et al., 2022). Hence, it requires the users to have a thorough understanding of building knowledge for better output results (Amara et al., 2015). The white box model has the benefit of having a strong explanatory ability, but inputting all the specific structural parameters might be time-consuming (Pan et al., 2023). Usually, white-box modelling is used to simulate energy consumption and optimise the HVAC systems, lighting and other energy-consuming systems in indoor farming.

The optimisation of vertical farming energy use can also be increased by including energy software that can model and simulate the energy consumption of indoor buildings. EnergyPlus™ and Transient System Simulation Program (TRNSYS) are the most popular building energy modelling (BEM) software for white-box modelling. EnergyPlus™ is a comprehensive building energy simulation software employed by engineers, architects, and researchers to simulate various aspects of a building's energy usage, including heating, cooling, ventilation, lighting, appliances, and water consumption. EnergyPlus was developed by the US Department of Energy for indoor building energy simulation processes. Figure 2 illustrates the working flowchart of EnergyPlus. TRNSYS is a software that can simulate the behaviour of transient systems. It contains extensive libraries of components that allow the modelling of the performance of different systems such as pumps, turbines, HVAC systems, and weather data processors.

The robustness of EnergyPlus has been proven by several scholars. Hachem-Vermette and MacGregor (2017) designed their indoor growing centre in Canada by simulating the operation of the usage of photovoltaic panels and Solid Oxide Fuel Cells as additional energy sources. They were able to reduce the thermal load by 65% and energy consumption by 30%. Wang and Iddio (2022) simulated the misting systems used for cooling and humidification in the indoor farming facility. The simulation result was compared with the measured data. The model produced by EnergyPlus successfully reduced 4.7% of electricity and 48.1% of natural gas use. EnergyPlus was also applied by Liebman-Pelaez et al. (2021) in the simulation of a container farm.

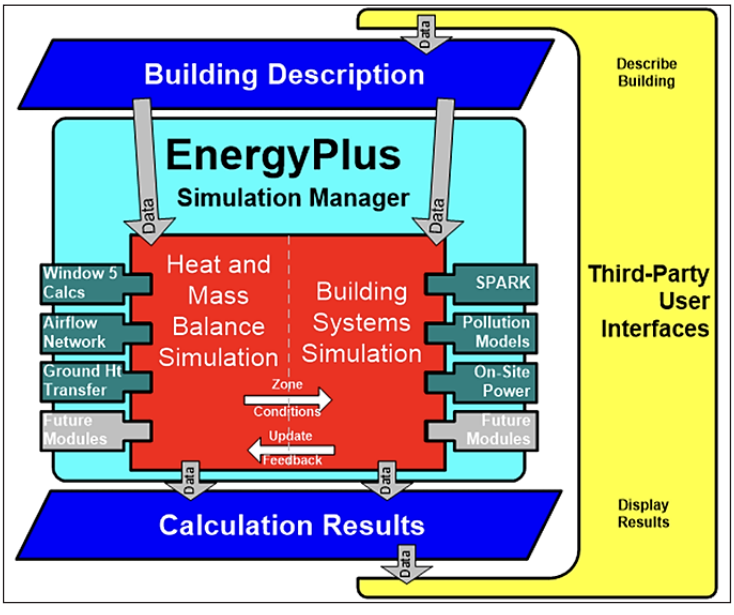


Figure 2. EnergyPlus flowchart (U.S. Department of Energy, 2022)

The information, including outdoor climate, building envelope, lighting load, crop parameters and lighting system and plant evapotranspiration, was entered into the EnergyPlus software. Although the model was found to be underestimating the cooling energy consumption, it overestimated the lighting, equipment and dehumidifier energy consumption. Eaton et al. (2023) modelled energy use of a multi-layer plant factory for lettuce growth in EnergyPlus by entering the specifications of building information, environmental control, crop parameters and lighting system into it. Several models were simulated for selection. Their lighting saving strategy saved 24% energy, the HVAC model saved 40-69% and maintaining the elevated levels of CO2 saved energy by 13-14%. In total, the software model was able to reduce energy consumption by 46%.

One of the advantages of EnergyPus is the ability to predict the energy consumption on an hourly or monthly basis. This allows management teams to accurately manage the operation of their indoor buildings. Shabunko et al. (2018) and Lamagna et al. (2020) had proven the effectiveness of EnergyPlus in the benchmarking of energy consumption. In terms of indoor farming, Dahlan et al. (2018) demonstrated the monthly energy prediction of a tomato greenhouse as shown in Figure 3.

TRNSYS was used by Talbot et al. (2022) to compare the energy performance between a greenhouse and a container farm. In their experiments, the physical spaces of the greenhouse and container were modelled using TRNSYS. The parameters being compared were (1) the energy consumption, (2) the peak electricity demand and (3) the associated greenhouse gas emissions. As a result, analysis of TRNSYS showed the container farm has better performance with lower energy consumption, lower peak electricity demand and less greenhouse gas emissions.

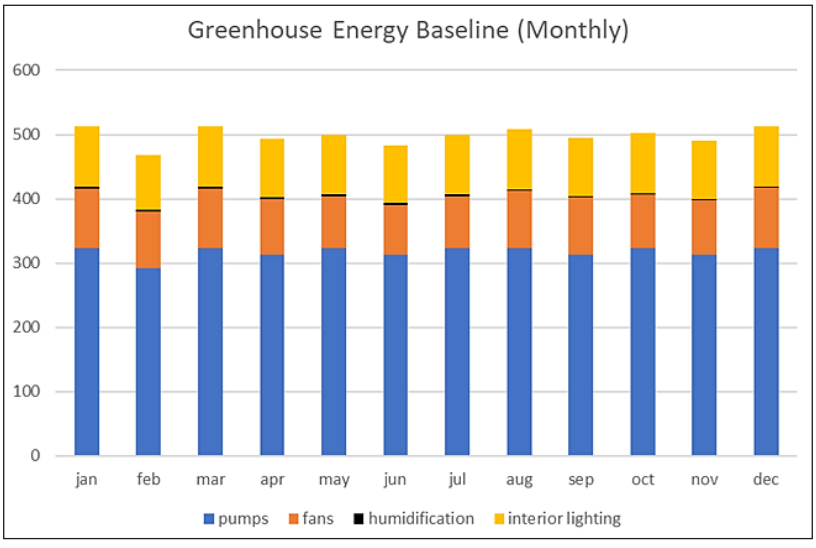


Figure 3. Energy (kWh) profiling of a plant factory by Dahlan et al. (2018)

Wai et al. (2022) investigated the energy consumption of the cultivation of strawberries in controlled environment rooms under a tropical climate using TRNSYS for modelling purposes. The developed model was validated by comparing the computed room air temperature from the model with the experimental room temperature. The statistical analysis revealed that the developed TRNSYS model was accurate in simulating the indoor condition with an RMSE at 9.2%. Besides, Yeo et al. (2022) demonstrated the combination of TRNSYS and CFD models. The physical specification of the rooftop greenhouse was designed using the software. The thermal energy flow was calculated using the BES. Energy exchange inside the greenhouse was also. They were able to reduce the energy load by 5.2% (464, 673 MJ). An alternating air temperature management (ATM) was applied to change the indoor temperature over time, where different temperatures were provided during the day and night. This ATM approach was able to further reduce the energy consumption for 11.8%. TRNSYS's ability to predict the energy use of a strawberry greenhouse was proven by Ogunlowo et al. (2023). In the research of Rasheed et al. (2020), TRNSYS successfully reduced the energy by suggesting different ventilation systems and thermal screens. Thus, TRNSYS is a reliable tool in energy modelling of indoor farming.

Black-box Modelling

One of the black box approaches to optimise indoor energy consumption is called model predictive control (MPC). MPC is a type of advanced control strategy that models inputs to predict future behaviours and make optimal control decisions. During each control interval, the MPC algorithm aims to enhance future plant performance by calculating a series of upcoming adjustments to manipulated variables (Qin & Badgwell, 2003). MPC is used in many different fields of engineering, including process engineering, power electronics, building climate and energy and manufacturing (Schwenzer et al., 2021).

MPC was a popular choice in energy optimisation. Achour et al. (2020) applied an MPC method to optimise the climate variables in a greenhouse. The controlled parameters included temperature, relative humidity, CO_2 rate, lighting levels, natural ventilation and energy consumption. The system was operated via a wireless sensor network (WSN) and a centralised model predictive controller. The proposed model was a supervisory model with real-time control to effectively manage the overall energy and water demand. MATLAB was the software used to mathematically model the system. The proposed model can significantly reduce the energy consumption and CO2 emissions, although the exact amount of energy saved was not mentioned. MPC was also discovered by Bersani et al. (2021) to minimise the total energy use in a greenhouse. The MPC model acted as an optimisation model to determine the optimal control signals related to the hot water flow in the heating plant. In the MPC model testing stage, the heating system was manipulated according to the difference between the current temperature and the proposed optimal temperature. The results showed that 30% of electric power was saved by this MPC approach.

Mathematical models have been proven by several researchers. Pimentel et al. (2023) proposed a model for the purpose of energy efficiency by using a formulation via the P-graph framework. P-graph is a combinational approach to synthesising and optimising process networks. It has demonstrated significant superiority in lowering the associated computational load and is incredibly effective at solving problems with high combinatorial complexity. The case study results point to potential savings of up to 40% in the cost of imported electricity for the given day and up to 31% for the full year. Another mathematical model called Mixed-Integer Linear Programming (MIP) models was presented by Delorme and Santini (2022). A mixed-integer programming (MIP) problem is characterised by having certain decision variables constrained to take integer values (e.g., whole numbers like -1, 0, 1, 2, etc.) at the optimal solution. Delorme and Santini (2022) proposed a polynomial number of variables (M1), a pseudo-polynomial number of variables (M2) and a hybrid MIP, which only used binary variables (M3) to minimise the energy consumption of automatic elevators in their plant factory. M1 was concluded to be the most efficient model.

Artificial intelligence methods are the most popular trend in energy optimisation of indoor farms. Different algorithms such as genetic programming, deep learning and Bayesian networks were applied to simulate and reduce energy consumption. Murakami et al. (2018) developed an addictive Bayesian network for the minimisation of energy cost in plant factories. A Bayesian network is a probabilistic model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). It is an effective technique for simulating complex systems and generating predictions from ambiguous or insufficient data. By considering the effect of weather factors on electric power consumption, an appropriate operation plan of cultivation systems was decided to reduce energy cost while producing plants with the same amount and quality. Simulation results showed that the operation plan of cultivation systems properly reflected the effect of electricity charge and weather factors on electric power consumption to reduce energy cost. A total of 1.13% energy cost was saved.

Two nonlinear models, i.e., genetic programming (GP) and feedforward artificial neural networks (FNNs), were combined by Olvera-Gonzalez et al. (2021) to predict energy consumption in closed indoor farms. GP is an evolutionary computation (EC) technique that automatically solves problems without having to tell the computer explicitly how to do it (Langdon et al., 2008). FNN is a neural network where information moves in only one direction—forward—from the input nodes, through the hidden nodes and to the output nodes. There are no cycles or loops in the network. In this research, inputs included in the models were light intensity, RGB components, white light component, light operation frequency and the duty cycle of the farm. The GP model had a training accuracy of 96.1%, and the FNN model had 98.99% accuracy. In terms of testing accuracy, the GP model achieved 95.35%, and the FNN model achieved 98.21%.

Thus, the FNN model was concluded to have better generalisation than GP. Khudoyberdiev et al. (2020) developed fuzzy logic as a control for energy optimisation in an IoT-based hydroponics environment. Fuzzy logic is an AI technique that describes the linguistic control value in the form of control parameters. It describes human preferences and experiences via fuzzy rules and membership functions. The parameters in this paper included water and humidity levels. These parameters came from the water pump, overflow system, dehumidifier and fogging system. This study used the fuzzy logic control module to define the operating duration limit and working level for the actuators based on IF-THEN rules while considering the optimised data and current sensing data. The outcomes proved the fuzzy logic to be a feasible method to control the operation of the hydroponics system, as 18% of energy consumption was reduced. Hu and You (2024) utilised deep learning techniques to develop an energy model for lettuce production in a plant factory. Deep learning is a machine learning technique that mimics human neural networks to learn and predict future trends from existing datasets. The parameters, including temperature, humidity, and carbon dioxide level, are controlled in the deep learning network to create an optimal growing condition for the crops. As a result, an 8.75% reduction in energy consumption was achieved by this model.

Grey-box Modelling

Talbot and Monfet (2024) constructed a dynamic crop model that can predict the growth of lettuce and integrate it into TRNSYS. The growth model is an algorithm proposed by van Henten (1994) and was adjusted by using the growth requirements of lettuce. The adjusted growth model was modelled in the TRNSYS software for energy analysis. They concluded the predicted energy use of the grey-box model had 3.5% variation compared to the experimental data.

Another grey-box modelling of combining KASPRO (a dynamic model to calculate the climate conditions in greenhouses) and EnergyPlus was carried out by Graamans et al. (2018). The purpose of EnergyPlus was to compensate for the differences between a greenhouse and a plant factory. The simulation showed that the plant factory consumed more energy than greenhouses. This is caused by artificial lighting. However, in terms of energy use efficiency, the plant factory showed better performance than greenhouses (1411 MJ kg^{-1} dry weight vs 1699 MJ kg^{-1} dry weight).

EnergyPlus, as a modelling tool alone, is not accurate as the crop biological processes, such as evapotranspiration, soil's heat conduction, and convective and radiative heat exchanges between the plants' canopy, can contribute to the energy consumption. Therefore, Ledesma et al. (2022) co-simulated the operation of a hydroponics farm by considering the crop growth model in MATLAB. The output of MATLAB became input for EnergyPlus, and the EnergyPlus Co-Simulation Toolbox was used to communicate with MATLAB. The thermal load decreased by 42%.

Unlike white-box modelling and black-box modelling, which are popular in the energy optimisation of indoor farming, grey-box modelling is not a common practice. In the meantime, grey-box modelling is popular in other building thermal and energy optimisation. Kim et al. (2022) reviewed 21 articles published from 2006 to 2022 on this topic. Hence, grey-box modelling is proven to be robust in energy optimisation. Future indoor farm energy optimisation can consider this approach as shown in Table 1.

The graph shows comparisons of energy-saving techniques used in indoor farming across various research studies using different modelling tools. Black-box and white-box models are used to categorise the studies, and each bar shows the stated percentage of energy savings. In the black-box category, Khudoyberdiev et al. (2020) used Bayesian fuzzy logic to reach higher savings of about 20%, whereas Hu and You (2024) utilised a deep learning model and produced moderate energy reductions. Using an unidentified black-box technique, Murakami et al. (2018) showed negligible energy savings. Using TRNSYS software, Yeo et al. (2022) transitioned to white-box models and obtained modest savings that were comparable to those of a deep learning model (Hu & You, 2024). The EnergyPlus-based investigations by Hachem-Vermette and MacGregor (2017) and Eaton et al. (2023), however, demonstrate noticeably greater energy savings, with the study demonstrating about 30% and the study reaching over 50% (Eaton et al., 2023). This implies that compared to black-box methods, white-box models—especially EnergyPlus—generally result in greater energy savings. The comparison was shown in Figure 4.

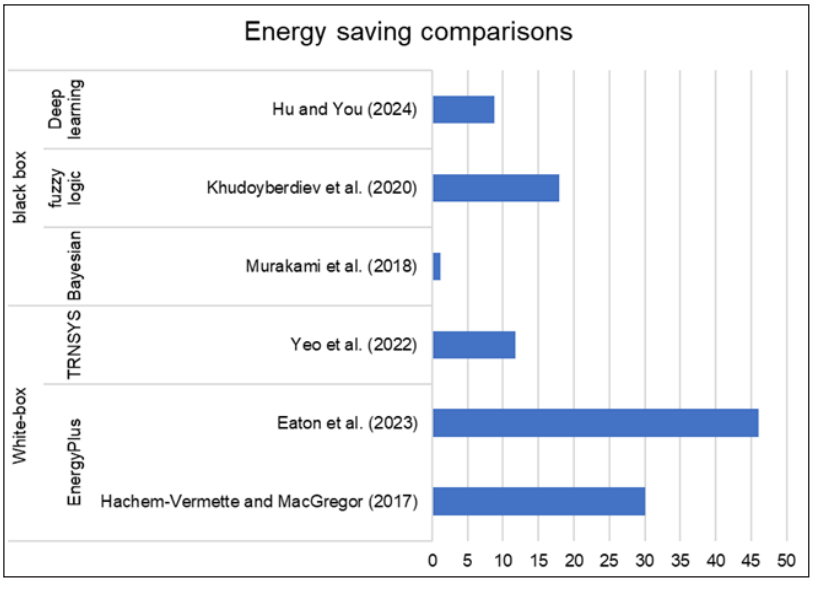


Figure 4. Energy saving comparison

Table 1
Summary of energy optimisation research

Models	Platforms/ Algorithms	Authors	Type of Facility	Outcomes
White-box	EnergyPlus	Hachem-Vermette and MacGregor (2017)	Indoor growing centre	able to reduce the thermal load by 65% and energy consumption by 30%
		Liebman-Pelaez et al. (2021)	Container farm	An hourly prediction model of energy consumption was constructed
		Wang and Iddio (2022)	Indoor farm facility	Electricity reduces by 4.7%, and natural gas reduces by 48.1%
		Eaton et al. (2023)	Multi-layer plant factory	Total energy saving by 46%
	TRNSYS	Talbot et al. (2022)	Greenhouse vs container farm	Container farms show better performance in terms of energy usage, electricity and greenhouse gas emissions.
		Wai et al. (2022)	Indoor strawberries cultivation	Simulation with 9.2% RMSE
		Yeo et al. (2022)	Rooftop greenhouse	Energy saving by 11.8%
		Ogunlowo et al. (2023)	Strawberry greenhouse	Accurate energy consumption prediction
		Rasheed et al. (2020)		Design different ventilation and thermal screens
Black-box	MPC algorithm	Achour et al. (2020)	Greenhouse	Reduce the energy consumption and CO2 emissions
		Bersani et al. (2021)	Greenhouse	30% of electric power was saved
	Mathematical models: P-graph	Pimentel et al. (2023)	Lettuce vertical farms	Reduce 40% of electricity for the given day and 31% for the full year
	Mathematical models: MIP models	Delorme and Santini (2022)	Plant factory	Minimised the energy consumption of automatic elevators
	Bayesian network	Murakami et al. (2018)	Plant factory	Reduce 1.13% of energy cost
	Genetic programming + Feedforward artificial neural network	Olvera-Gonzalez et al. (2021)	Closed indoor farm	GP: training accuracy 96.1%, testing accuracy 95 35% FNN: training accuracy 98.99%, testing accuracy 98.21%
	Fuzzy logic	Khudoyberdiev et al. (2020)	IoT-based hydroponics environment	18% of energy consumption was reduced
	Deep learning	Hu and You (2024)	Plant factory	Reduce 8.75% energy consumption

Table 1 (*continued*)

Models	Platforms/ Algorithms	Authors	Type of Facility	Outcomes
Grey-box	TRNSYS + dynamic crop model	Talbot and Monfet (2024)	Controlled agriculture environment spaces	3.5% variation in the predicted energy model
	EnergyPlus + KASPRO	Graamans et al. (2018)	Greenhouse vs Plant Factory	The plant factory showed better energy use efficiency
	EnergyPlus + crop growth model in MATLAB	Ledesma et al. (2022)	Rooftop farms	Reduce thermal load by 42%

DISCUSSION

White-box models focus on thoroughly understanding the physical processes that drive energy consumption, whereas black-box models prioritise prediction accuracy, often at the expense of interpretability (Shahcheraghian et al., 2024).

The advantages of EnergyPlus as a white-box model are that it provides and considers all details of the whole building in the modelling process. This enables us to study the relationship between the outputs and building structures. The white-box model required extensive descriptions of the buildings and calibration efforts. Thus, it is suitable for complex building systems optimisation (Kim et al., 2022). The occupant comfort can be taken into consideration. The white-box models can be constructed from the prior information without the need for any observation (Amara et al., 2015) or historical operation data (Shahcheraghian et al., 2024). EnergyPlus is usually applied for preliminary simulation, and MOGA requires the data to be collected first. But a white-box model is more difficult to construct, as they need to include all the data. EnergyPlus has an extensive library of building components. It contains a toolbox that provides integrated and concurrent solutions, heat balance-driven computations, flexible sub-hourly time steps, comprehensive heat and mass transfer calculations, advanced fenestration models, illuminance and glare assessments, component-based HVAC modelling, and a variety of predefined HVAC and lighting control strategies. According to Zhu et al. (2012), EnergyPlus has sub-hourly, user-definable time steps for the interaction between thermal zones and the environment, and variable time steps for interactions between the thermal zones and the HVAC systems. It allows users to design HVAC systems by inserting and connecting components. The whole system can be modified and controlled using the Energy Management System feature without recompiling the program source code.

Black-box modelling uses data in computational, machine learning and statistical analysis to develop models. This method builds the relationship between the energy consumption and influencing variables through the empirical models based on the historical data (Liu et al., 2019).

The process of black-box modelling consists of data collection, data pre-processing, feature selection, model training and model testing. Data pre-processing, including data cleaning, outlier elimination and data integration. In feature selection, the most useful variables are selected by users for the training stage. Data training is the process of training the model to learn patterns in each dataset. Lastly, the model testing stage will validate the performance of the produced model. Some popular black-box models included Support Vector Machine (SVM) and Artificial Neural Network (ANN).

However, there are some limitations and uncertainties in using white-box modelling for energy modelling in indoor farming. The uncertainties in energy modelling of white-box modelling come from three reasons, i.e., occupants, building and climate factors. The existence of occupants such as humans and crops is related to energy consumption and cannot be accurately predicted. The energy consumption of occupants is influenced by factors such as thermal comfort requirements, weather, building area, crop density and economic considerations. The micro-weather factor has a stronger influence on indoor farm energy use. The utilisation of climate data to predict the energy consumption of indoor farms will produce different results accordingly. Besides, the details of the building parameters are difficult to collect for the inputs. It is almost impossible to collect accurate data through field studies. Physical parameters of buildings are not always known or are not available; so, it is not always easy to obtain a building energy prediction by means of a simulation software based on physical parameters (Ferlito et al., 2015). In black-box models, high-quality datasets are required to produce an accurate model. Most importantly, a black-box model cannot explain the principles of heat transfer (conduction, convection and radiation) in the indoor farm (Yu et al., 2022).

In addition, the control of the working period of LED in indoor farming, i.e., the photoperiod, has been discovered in previous studies as well. One early study by Tennessen et al. (1995) discovered the effects of pulse-LED on tomatoes. They found that intermittent light will not affect crop production as the photosynthesis rate remains the same. Filatov et al. (2021) used a vertical farming system and two lighting techniques in their study. The first, which served as the control, included setting the LED lights' daily light and dark cycles to 16/8 hours. On the other hand, the experimental set included an 8/4-hour light/dark schedule that was split into two parts (2300–0700 and 1100-1900). They were able to cut down the electricity consumption by 10% without affecting the crop yield. Carotti et al. (2021) studied the effects of pulsed red and blue LED light on lettuce growth. Two switching frequencies of LED were applied on the crops, i.e., 850kHz (high switching frequency) and 293kHz (low switching frequency). The study discovered that the low frequency of pulsed light can increase the energy use efficiency by 40% and maintain the crops' qualities. Olvera-Gonzalez et al. (2021) also proved that pulsed LED-lighting can save energy consumption by more than 12%. Hence, the control of photoperiod in indoor farming can be included in the energy optimisation approaches as well.

CONCLUSION

Indoor farming provides a promising solution to meet the world's food demand and reach food security. By applying the existing technologies, including LED lighting, fans and humidifiers for climate control, and hydroponic systems, indoor farms can operate efficiently in controlled environments. These systems play an important role in crop growth conditions. However, the energy consumption associated with these technologies is larger than that of traditional agriculture. Thus, a focus on innovative approaches to reduce energy use is needed.

This review highlights three main approaches to energy optimisation in indoor farming, viz., white box, black box, and grey box modelling. Each method has its advantages and disadvantages. Grey models can combine the advantages of both white box and grey box modelling approaches.

Ultimately, optimising energy consumption in indoor farming is essential for enhancing its sustainability and economic viability. The review suggests that while significant progress has been made in understanding and improving energy use through various modelling approaches, there remains a need for further research to develop more integrated and adaptive optimisation strategies. By advancing these methods, indoor farming can achieve higher energy efficiency, reduce operational costs, and contribute more significantly to sustainable food production. Continued innovation in this field will be vital to address the challenges of energy consumption and to unlock the full potential of indoor farming systems.

ACKNOWLEDGEMENT

Thank you to all the staff who have contributed to the research. This research was supported by Universiti Putra Malaysia research grant with the project code number GP-IPS/2022/9709100. The funder had no role in the conceptualisation, design, data collection, analysis, decision to publish, or preparation of the manuscript.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest in this paper.

REFERENCES

Achour, Y., Ouammi, A., Zejli, D., & Sayadi, S. (2020). Supervisory model predictive control for optimal operation of a greenhouse indoor environment, coping with the food-energy-water nexus. *IEEE Access, 8*, 211562-211575. https://doi.org/10.1109/ACCESS.2020.3037222

Amara, F., Agbossou, K., Cardenas, A., Dubé, Y., & Kelouwani, S. (2015). Comparison and simulation of building thermal models for effective energy management. *Smart Grid and Renewable Energy, 6*(4), 95-112. https://doi.org/10.4236/sgre.2015.64009

Arcasi, A., Mauro, A. W., Napoli, G., Tariello, F., & Vanoli, G. P. (2024). Energy and cost analysis for crop production in a vertical farm. *Applied Thermal Engineering, 239*, Article 122129. https://doi.org/10.1016/j.applthermaleng.2023.122129

Arendt, K., Jradi, M., Shaker, H. R., & Veje, C. (2018). Comparative analysis of white-, grey- and black-box models for thermal simulation of indoor environment: Teaching building case study. *Building Performance Analysis Conference and SimBuild*, 173-180.

Avgoustaki, D. D., & Xydis, G. (2020). How can energy innovation in indoor vertical farming improve food security, sustainability, and food safety? In M. J. Cohen (Ed.), *Advances in food security and sustainability* (Vol. 5, pp. 1-51). Elsevier. https://doi.org/10.1016/bs.af2s.2020.08.002

Bantis, F., Smirnakou, S., Ouzounis, T., Koukounaras, A., Ntagkas, N., & Radoglou, K. (2018). Status and recent achievements in the field of horticulture with the use of light-emitting diodes (LEDs). *Scientia Horticulturae, 235*, 437-451. https://doi.org/10.1016/j.scienta.2018.02.058

Barbosa, G. L., Almeida Gadelha, F. D., Kublik, N., Proctor, A., Reichelm, L., Weissinger, E., Wohlleb, G. M., & Halden, R. U. (2015). Comparison of land, water, and energy requirements of lettuce grown using hydroponic vs. conventional agricultural methods. *International Journal of Environmental Research and Public Health, 12*(6), 6879-6891. https://doi.org/10.3390/ijerph120606879

Bersani, C., Fossa, M., Priarone, A., Sacile, R., & Zero, E. (2021). Model predictive control versus traditional relay control in a high energy efficiency greenhouse. *Energies, 14*(11), Article 3353. https://doi.org/10.3390/en14113353

Boretti, A., & Rosa, L. (2019). Reassessing the projections of the World Water Development Report. *npj Clean Water, 2*, Article 15. https://doi.org/10.1038/s41545-019-0039-9

Carotti, L., Potente, G., Pennisi, G., Ruiz, K. B., Biondi, S., Crepaldi, A., Orsini, F., Gianquinto, G., & Antognoni, F. (2021). Pulsed LED light: Exploring the balance between energy use and nutraceutical properties in indoor-grown lettuce. *Agronomy, 11*(6), Article 1106. https://doi.org/10.3390/agronomy11061106

Dahlan, Y. N., Amiruddin, A., Luong, N. D., & Sakimin, S. Z. (2018). Energy and climate analysis of greenhouse system for tomatoes cultivation using CFD and OpenStudio EnergyPlus software. *International Journal of Engineering and Technology, 7*(3.11), 183-186. https://doi.org/10.14419/ijet.v7i3.11.15957

De Buck, V., Sbarciog, M. I., Cras, J., Bhonsale, S. S., Polanska, M., & Van Impe, J. F. M. (2023). Critical analysis of the use of white-box versus black-box models for multi-objective optimisation of small-scale biorefineries. *Frontiers in Food Science and Technology, 3*, Article 1154305. https://doi.org/10.3389/frfst.2023.1154305

Delorme, M., & Santini, A. (2022). Energy-efficient automated vertical farms. *Omega, 109*, Article 102611. https://doi.org/10.1016/j.omega.2022.102611

Eaton, M., Harbick, K., Shelford, T., & Mattson, N. (2021). Modelling natural light availability in skyscraper farms. *Agronomy, 11*(9), Article 1684. https://doi.org/10.3390/agronomy11091684

Eaton, M., Shelford, T., Cole, M., & Mattson, N. (2023). Modelling resource consumption and carbon emissions associated with lettuce production in plant factories. *Journal of Cleaner Production, 384*, Article 135569. https://doi.org/10.1016/j.jclepro.2022.135569

Ferlito, S., Atrigna, M., Graditi, G., De Vito, S., Salvato, M., Buonanno, A., & Di Francia, G. (2015). Predictive models for building's energy consumption: An artificial neural network (ANN) approach. *2015 XVIII AISEM Annual Conference*, 1-4. https://doi.org/10.1109/AISEM.2015.7066836

Filatov, D. A., Vetchinnikov, A. A., Olonina, S. I., & Olonin, I. Y. (2021). Intermittent LED lighting helps reduce energy costs when growing microgreens in vertical controlled environment farms. *IOP Conference Series: Earth and Environmental Science, 979*(1), Article 012096. https://doi.org/10.1088/1755-1315/979/1/012096

Food Security Information Network. (2020). *Global report on food crises 2020 (GRFC 2020)*. Global Network Against Food Crises.

Graamans, L., Baeza, E., van den Dobbelsteen, A., Tsafaras, I., & Stanghellini, C. (2018). Plant factories versus greenhouses: Comparison of resource use efficiency. *Agricultural Systems, 160*, 31-43. https://doi.org/10.1016/j.agsy.2017.11.003

Graamans, L., Tenpierik, M., van den Dobbelsteen, A., & Stanghellini, C. (2020). Plant factories: Reducing energy demand at high internal heat loads through façade design. *Applied Energy, 262*, Article 114544. https://doi.org/10.1016/j.apenergy.2020.114544

Gregorio, G., & Ancog, R. (2020). Assessing the impact of the COVID-19 pandemic on agricultural production in Southeast Asia: Toward transformative change in the agricultural food system. *Asian Journal of Agriculture and Development, 17*(1), 1-14. https://doi.org/10.37801/ajad2020.17.1.1

Hachem-Vermette, C., & MacGregor, A. (2017). Energy optimised envelope for a cold climate indoor agricultural growing centre. *Buildings, 7*(3), Article 59. https://doi.org/10.3390/buildings7030059

Horomia, K., & Gordon-Smith, H. (2021). *Global CEA census report*. WayBeyond.

Hu, G., & You, F. (2024). AI-enabled cyber-physical-biological systems for smart energy management and sustainable food production in a plant factory. *Applied Energy, 356*, Article 122334. https://doi.org/10.1016/j.apenergy.2023.122334

Khudoyberdiev, A., Ahmad, S., Ullah, I., & Kim, D. H. (2020). An optimisation scheme based on fuzzy logic control for efficient energy consumption in a hydroponics environment. *Energies, 13*(2), Article 289. https://doi.org/10.3390/en13020289

Kim, D., Lee, J., Do, S., Mago, P. J., Lee, K. H., & Cho, H. (2022). Energy modelling and model predictive control for HVAC in buildings: A review of current research trends. *Energies, 15*(19), Article 7231. https://doi.org/10.3390/en15197231

Lamagna, M., Nastasi, B., Groppi, D., Nezhad, M. M., & Garcia, D. A. (2020). Hourly energy profile determination technique from monthly energy bills. *Building Simulation, 13*(6), 1235-1248. https://doi.org/10.1007/s12273-020-0698-y

Langdon, W. B., Poli, R., McPhee, N. F., & Koza, J. R. (2008). Genetic programming: An introduction and tutorial, with a survey of techniques and applications. In J. Fulcher & L. C. Jain (Eds.), *Computational intelligence: A compendium* (Studies in Computational Intelligence, Vol. 115, pp. 927-1028). Springer. https://doi.org/10.1007/978-3-540-78293-3_22

Ledesma, G., Nikolic, J., & Pons-Valladares, O. (2022). Co-simulation for thermodynamic coupling of crops in buildings: Case study of free-running schools in Quito, Ecuador. *Building and Environment, 207*, Article 108407. https://doi.org/10.1016/j.buildenv.2021.108407

Liebman-Pelaez, M., Kongoletos, J., Norford, L. K., & Reinhart, C. (2021). Validation of building energy models of a hydroponic container farm and its application in urban design. *Energy and Buildings, 250*, Article 111192. https://doi.org/10.1016/j.enbuild.2021.111192

Liu, Z., Wu, D., Liu, Y., Han, Z., Lun, L., Gao, J., Jin, G., & Cao, G. (2019). Accuracy analysis and model comparison of machine learning adopted in building energy consumption prediction. *Energy Exploration & Exploitation, 37*(4), 1426-1451. https://doi.org/10.1177/0144598718822400

McCree, K. J. (1972). Test of current definitions of photosynthetically active radiation against leaf photosynthesis data. *Agricultural Meteorology, 10*, 443-453. https://doi.org/10.1016/0002-1571(72)90045-3

Mir, M. S., Naikoo, N. B., Kanth, R. H., Bahar, F. A., Bhat, M. A., Nazir, A., Mahdi, S. S., Amin, Z., Singh, L., Raja, W., Saad, A. A., Bhat, T. A., Palmo, T., & Ahngar, T. A. (2022). Vertical farming: The future of agriculture: A review. *The Pharma Innovation Journal, 11*(2S), 1175-1195.

Mitchell, C. A., & Sheibani, F. (2020). LED advancements for plant-factory artificial lighting. In T. Kozai, G. Niu, & M. Takagaki (Eds.), *Plant factory: An indoor vertical farming system for efficient quality food production* (2nd ed., pp. 167-184). Academic Press. https://doi.org/10.1016/B978-0-12-816691-8.00010-8

Murakami, S., Fujimoto, Y., Hayashi, Y., Fuchikami, H., & Hattori, T. (2018). Energy cost minimisation in plant factories considering weather factors using additive Bayesian networks. *Journal of International Council on Electrical Engineering, 8*(1), 128-135. https://doi.org/10.1080/22348972.2018.1477093

Natarajan, G., Zaid, M., Konka, H., Srinivasan, R., Ramanathan, S. S., Ahmed, T., & Chowdhury, H. (2022). Modelling of air distribution inside a shipping container plant factory using computational fluid dynamics (CFD). *AIP Conference Proceedings, 2681*(1), Article 020091. https://doi.org/10.1063/5.0117095

Ogunlowo, Q. O., Akpenpuun, T. D., Na, W. H., Adesanya, M. A., Rabiu, A., Dutta, P., Kim, H. T., & Lee, H. W. (2023). Simulation of greenhouse energy and strawberry (*Seolhyang* sp.) yield using TRNSYS DVBES: A base case. *Solar Energy, 266*, Article 112196. https://doi.org/10.1016/j.solener.2023.112196

Olvera-Gonzalez, E., Rivera, M. M., Escalante-Garcia, N., & Flores-Gallegos, E. (2021). Modelling energy LED light consumption based on an artificial intelligence method applied to closed plant production system. *Applied Sciences, 11*(6), Article 2735. https://doi.org/10.3390/app11062735

Pan, Y., Zhu, M., Lv, Y., Yang, Y., Liang, Y., Yin, R., Yang, Y., Jia, X., Wang, X., Zeng, F., Huang, S., Hou, D., Xu, L., Yin, R., & Yuan, X. (2023). Building energy simulation and its application for building performance optimisation: A review of methods, tools, and case studies. *Advances in Applied Energy, 10*, Article 100135. https://doi.org/10.1016/j.adapen.2023.100135

Pimentel, J., Balázs, L., & Friedler, F. (2023). Optimisation of vertical farms energy efficiency via multiperiodic graph-theoretical approach. *Journal of Cleaner Production, 416*, Article 137938. https://doi.org/10.1016/j.jclepro.2023.137938

Qin, S. J., & Badgwell, T. A. (2003). A survey of industrial model predictive control technology. *Control Engineering Practice, 11*(7), 733-764. https://doi.org/10.1016/S0967-0661(02)00186-7

Rasheed, A., Kwak, C. S., Kim, H. T., & Lee, H. W. (2020). Building energy simulation model for analysing energy saving options of multi-span greenhouses. *Applied Sciences, 10*(19), Article 6884. https://doi.org/10.3390/app10196884

Romeo, D., Vea, E. B., & Thomsen, M. (2018). Environmental impacts of urban hydroponics in Europe: A case study in Lyon. *Procedia CIRP, 69*, 540-545. https://doi.org/10.1016/j.procir.2017.11.048

Schwenzer, M., Ay, M., Bergs, T., & Abel, D. (2021). Review on model predictive control: An engineering perspective. *The International Journal of Advanced Manufacturing Technology, 117*(5-6), 1327-1349. https://doi.org/10.1007/s00170-021-07682-3

Shaari, A. M., Razuan, M. S. M., Taweekun, J., Batcha, M. F. M., Abdullah, K., Sayuti, Z., Ahmad, A. L., & Tahir, M. A. M. (2021). Energy profiling of a plant factory and energy conservation opportunities. *Journal of Advanced Research in Fluid Mechanics and Thermal Sciences, 80*(1), 13-23. https://doi.org/10.37934/arfmts.80.1.1323

Shabunko, V., Lim, C. M., & Mathew, S. (2018). EnergyPlus models for the benchmarking of residential buildings in Brunei Darussalam. *Energy and Buildings, 169*, 507-516. https://doi.org/10.1016/j.enbuild.2016.03.039

Shahcheraghian, A., Madani, H., & Ilinca, A. (2024). From white- to black-box models: A review of simulation tools for building energy management and their application in consulting practices. *Energies, 17*(2), Article 376. https://doi.org/10.3390/en17020376

Sihombing, P., Karina, N. A., Tarigan, J. T., & Syarif, M. I. (2018). Automated hydroponics nutrition plants systems using Arduino Uno microcontroller based on Android. *Journal of Physics: Conference Series, 978*(1), Article 012014. https://doi.org/10.1088/1742-6596/978/1/012014

Sohn, J., Liulys, M., Avgoustaki, D. D., & Xydis, G. (2023). CFD analysis of airflow uniformity in a shipping-container vertical farm. *Computers and Electronics in Agriculture, 215*, Article 108363. https://doi.org/10.1016/j.compag.2023.108363

Sridhar, A., Balakrishnan, A., Jacob, M. M., Sillanpää, M., & Dayanandan, N. (2023). Global impact of COVID-19 on agriculture: Role of sustainable agriculture and digital farming. *Environmental Science and Pollution Research, 30*(15), 42509-42525. https://doi.org/10.1007/s11356-022-19358-w

Talbot, M. H., & Monfet, D. (2024). Development of a crop growth model for the energy analysis of controlled agriculture environment spaces. *Biosystems Engineering, 238*, 38-50. https://doi.org/10.1016/j.biosystemseng.2023.12.012

Talbot, M.-H., Lalonde, T., Beaulac, A., Haillot, D., & Monfet, D. (2022). Comparing the energy performance of different controlled environment agriculture spaces using TRNSYS. *Proceedings of eSim 2022: 12th Conference of IBPSA-Canada*.

Tennessen, D. J., Singsaas, E. L., & Sharkey, T. D. (1994). Light-emitting diodes as a light source for photosynthesis research. *Photosynthesis Research, 39*(1), 85-92. https://doi.org/10.1007/BF00027146

U.S. Department of Energy. (2022). *EnergyPlus™ version 22.2.0 documentation*.

van Henten, E. J. (1994). Validation of a dynamic lettuce growth model for greenhouse climate control. *Agricultural Systems, 45*(1), 55-72. https://doi.org/10.1016/S0308-521X(94)90280-1

Wai, T. S., Maruyama, N., Mona, Y., & Chaichana, C. (2022). Prediction of the energy consumption for indoor strawberry cultivation in a tropical climate. *Journal of Hunan University Natural Sciences, 49*(3), 158-167. https://doi.org/10.55463/issn.1674-2974.49.3.17

Wang, L., & Iddio, E. (2022). Energy performance evaluation and modelling for an indoor farming facility. *Sustainable Energy Technologies and Assessments, 52*, Article 102240. https://doi.org/10.1016/j.seta.2022.102240

Wheeler, R. M. (2008). A historical background of plant lighting: An introduction to the workshop. *HortScience, 43*(7), 1942-1943.https://doi.org/10.21273/HORTSCI.43.7.1942

Yeo, U. H., Lee, S. Y., Park, S. J., Kim, J. G., Cho, J. H., Decano-Valentin, C., Kim, R. W., & Lee, I. B. (2022). Rooftop greenhouse: (2) Analysis of thermal energy loads of a building-integrated rooftop greenhouse (BiRTG) for urban agriculture. *Agriculture, 12*(6), Article 787. https://doi.org/10.3390/agriculture12060787

Yu, J., Chang, W. S., & Dong, Y. (2022). Building energy prediction models and related uncertainties: A review. *Buildings, 12*(8), Article 1284. https://doi.org/10.3390/buildings12081284

Zaidan, M. A., Wraith, D., Boor, B. E., & Hussein, T. (2019). Bayesian proxy modelling for estimating black carbon concentrations using white-box and black-box models. *Applied Sciences, 9*(22), Article 4976. https://doi.org/10.3390/app9224976

Zhu, D., Hong, T., Yan, D., & Wang, C. (2012). *Comparison of building energy modelling programs: Building loads*. Lawrence Berkeley National Laboratory. https://doi.org/10.2172/1168735

Pertanika J. Sci. & Technol. 34 (3): 1309 - 1329 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Valorisation of Oil Palm Waste into Magnetic Catalysts for Efficient Synthesis of Bio-additive Ethyl Levulinate: Optimisation of Reaction Parameters

Shamala Gowri Krishnan[1] and Pua Fei Ling[2,3]*

[1]College of Graduate Studies, Universiti Tenaga Nasional, 43000 Kajang, Selangor, Malaysia
[2]Institute of Sustainable Energy (ISE), Universiti Tenaga Nasional, 43000 Kajang, Selangor, Malaysia
[3]Department of Mechanical Engineering, College of Engineering, Universiti Tenaga Nasional, 43000 Kajang, Selangor, Malaysia

ABSTRACT

Levulinate esters are promising value-added energy chemicals that have piqued researchers' worldwide interest owing to their potential as fuel additives. The reactions mainly concentrate on catalytic performance and offer limited research on the use of biomass-derived solid catalysts. This study investigated the efficiency of palm waste-derived magnetic solid catalysts and optimised the esterification process to improve ethyl levulinate yield. The results showed that the sulfonic group, $–SO_3H$, provides strong Brønsted acid sites for esterification, while ferric ions, Fe^{3+}, act as Lewis acids to enhance carbonyl activation, jointly catalysing the conversion to ethyl levulinate. The ethyl levulinate yield reached 87.0 mol%, achieved with an EFB fibre/γ-Fe_2O_3 catalyst (9.1 wt%) and an ethanol-to-levulinic acid molar ratio of 13.6:1 at 66.7 °C for 6 h. FT-IR and GC-FID analyses validated the formation of the ester groups following the esterification reaction. The catalyst can be easily removed with an external magnet, offering an effective way to use solid catalysts in esterification.

Keywords: Empty fruit bunch, ethyl levulinate, magnetic, mechanism, optimisation

ARTICLE INFO

Article history:
Received: 08 October 2025
Accepted: 18 March 2026
Published: 12 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.02

E-mail addresses:
shamalagowri.k@gmail.com (Shamala Gowri Krishnan)
gracepua@uniten.edu.my (Pua Fei Ling)
* Corresponding author

INTRODUCTION

Interest in developing alternative fuels and chemicals from sustainable biomass has increased over the last decade, as climate change associated with using non-renewable fuels has been recognised globally (Gielen et al., 2019). Several alternatives are needed to reduce the world's dependence on

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

fossil fuels, particularly in transportation. Compared to fossil fuels, biodiesel has a lower environmental impact as it is produced from renewable feedstocks. Biodiesel in larger quantities is severely hampered by its higher density, lower calorific value, poor oxidation stability, higher nitrogen oxide emissions and poor cold flow properties (Martínez Aguilar et al., 2020). Biodiesel, when combined with fuel additives, can address these problems by improving diesel engine efficiency and reducing emissions. There are various additives, such as metal-based additives, oxygenates, and antioxidants. Ethyl levulinate, synthesised from levulinic acid, has been recommended as an effective oxygenate additive.

Levulinic acid (LA), chemically known as 4-oxopentanoic acid, gamma-ketovaleric acid or 3-acetylpropionic acid, is a linear C5 alkyl carbon chain and is considered one of the twelve most promising biomass-derived chemicals (Signoretto et al., 2019). Ester derivatives of levulinic acid are commonly used as flavouring agents, in fragrances, pharmaceutical intermediates, and as fuel additives. Alkyl levulinates are an appealing substance because they can be used up to 5 wt% as a miscible biofuel in diesel engines (Enumula et al., 2017). There are four methods for the synthesis of levulinate esters: one-pot synthesis, catalytic alcoholysis of furfural, reduction of furfural to furfuryl alcohol, and then catalytic alcoholysis and direct esterification using solid acid catalysts (Adeleye et al., 2019). All these methods have their advantages and disadvantages. However, direct esterification is the simplest and most sustainable method, using various solid acid catalysts such as zeolites, silica, heteropolyacids, mixed-metal oxides, and nanostructured solid acids (Ogino et al., 2018). Previously, it was known that dilute sulphuric acid is economical and can provide sufficient protons (Bronsted acid) to catalyse the reaction. Due to difficulties in extraction and their corrosive nature, liquid catalysts were limited. Alkyl levulinates (ALs) can be prepared via esterification reactions using various alcohols, such as methanol, ethanol, and butanol (Yan et al., 2017). The most common ALs are methyl levulinate and ethyl levulinate. Methyl levulinate is a short-chain ester with comparable properties to fatty acid methyl ester (FAME) (Liu et al., 2020). It can be utilised as a biodiesel additive owing to the following characteristics: (i) high lubricity, (ii) non-toxicity, and (iii) enhanced cold flow properties. Ethyl levulinate (EL) is one of the most potent fuel additives. Its sulphur-free nature has shown great promise for petrol blends and diesel engines. Furthermore, it is utilised as a flavouring agent, a fuel-oxygenate additive, and a precursor to various chemicals that can be converted into fuels. Among these ALs, ethyl levulinate is the most promising due to its compatibility with both petroleum and biodiesel and its potential as a future fuel substitute.

Although alcoholysis is efficient with liquid catalysts, its use has significant drawbacks, including difficulty in acid recovery, equipment degradation, and environmental pollution. Furthermore, corrosive environments require special materials for reactor construction, increasing capital and operating costs. These drawbacks can be addressed by substituting

liquid acid catalysts with solid catalysts. Solid catalysts offer many advantages, including easy recovery from reaction products, recyclability, and reduced susceptibility to corrosion. As a result, many solid acid catalysts have been widely used in esterification processes. Zhou et al. (2025) prepared a titanium dioxide nanotube-based solid acid catalyst (TNTs-PhSO_3H) that achieved a maximum yield of 98% at a catalyst loading of 12 wt% (Zhou et al., 2025). Costa et al. (2024) synthesised alkyl levulinate using mesoporous Stann silicates, achieving a maximum yield of 83.6% at a catalyst loading of 1 wt% and a reaction temperature of 120 °C (Costa et al., 2024). However, in the process, acidic ion-exchange resins, zeolites, silica, and heteropoly acid-based catalysts were primarily used to esterify levulinic acid (Zhao et al., 2020). Due to environmental concerns, solid biomass-based catalysts could be a viable alternative to current solid catalysts. Campana et al. (2023) synthesised heterogeneous acid catalysts using pine needles as a carbon source and found that sulfonated pine needles (PiNe-SO_3H) had the highest levulinate yield of 93% (Campana et al., 2023). Meanwhile, Saperi et al. (2024) reported that esterification of LA using lignin-based acid catalyst showed the maximum conversion of ethyl levulinate (EL) with 62.5 % at 80 ^oC after 3 h of reaction with 10 wt% catalyst loading (Saperi et al., 2024). The author also tested with acidic DES, which acts as a co-catalyst, further enhancing the yield to 81.7%. Besides, Patel and Rave (2025) worked on a sugarcane bagasse-based solid acid catalyst for the synthesis of biofuel additives. They reported that a 53 % conversion of levulinic acid and a 93 % selectivity for n-butyl levulinate were achieved under relatively mild conditions, such as a 1:2 molar ratio of acid to alcohol, 25 mg catalyst, 8 h, and 90 °C (Patel & Dave, 2025). Regardless of the type of carbon source used, there are no reports of using EFB or its derivatives in catalyst synthesis for producing levulinate ester.

This study investigates the synthesis of ethyl levulinate using magnetic solid catalysts based on empty fruit bunch (EFB) and explores the mechanism behind the catalyst's performance in the esterification reaction. A parametric study identified the optimal conditions for the esterification process. The effects of catalyst loading, ethanol-to-LA molar ratio, processing time, and temperature on esterification over magnetic solid catalysts were studied using the Box-Behnken design. Additionally, a reaction mechanism study highlighted the advantages of biomass-derived magnetic acid catalysts to produce bio additives.

MATERIALS AND METHODS

Materials

Analytical reagents, Levulinic acid (99%), were purchased from Merck, and Ethanol, 95% (Dulab), was bought from Suria Pembekal Umum Sdn. Bhd, Malaysia.

Synthesis of Ethyl Levulinate

The esterification reaction was carried out using EFB fibre/γ-Fe_2O_3, EFB lignin-SO_3H-Fe_3O_4, and EFB MCC/γ-Fe_2O_3, (Krishnan et al., 2020; Krishnan et al., 2022) along with a blank experiment as a control. The catalyst preparation methods were summarised in Table 1. The esterification of levulinic acid (LA) with ethanol, using magnetic solid catalysts derived from EFB, was carried out in a reflux setup. 10 g of levulinic acid and ethanol at a 10:1 molar ratio was heated to 70 °C. Once the target temperature was reached, 10 wt% of the prepared catalyst was added, and the mixture was stirred for 5 h. After the reaction, the catalyst was removed using a magnet and the excess ethanol was evaporated at 80 °C. The yield (EL_C) was calculated using Equation 1.

$$EL_C = \left(\frac{EL_n}{LA_n}\right) \times 100 \quad [1]$$

where EL_n is the mole of EL formed, and LAn is the initial mole of LA.

Table 1
The summary of catalysts preparation and their characteristics

Catalyst Type	Raw Material	Preparation Method	Characteristics (i; ii; iii; iv)
EFB fibre/γ-Fe_2O_3	Treated EFB fibre	Co-precipitation and Calcination	37.9 (C), 34.8 (O), 21.8 (Fe), 5.5 (S); 2.4 mmol/g; 188.87 m^2/g, 11.5 nm; 24.97 emu/g
EFB lignin-SO_3H-Fe_3O_4	Lignin (extracted from EFB)	Hydrothermal Carbonisation, Calcination and Sulfonation	52.7 (C), 25.0 (O), 17.5 (Fe), 4.8 (S); 2.2 mmol/g; 56.88 m^2/g, 4.8 nm; 21.78 emu/g
EFB MCC /γ-Fe_2O_3	Microcrystalline cellulose (extracted from EFB)	Co-precipitation and Calcination	42.9 (C), 18.2 (O), 26.9 (Fe), 11.9 (S); 2.5 mmol/g; 290.55 m^2/g, 3.5 nm; 21.27 emu/g

Note. i-elemental results; ii-acidity; iii-surface area and pore size; iv-magnetism. Adapted from, (Krishnan et al., 2020; Krishnan et al., 2022)

Fourier Transform Infrared Analysis

The functional group in the esterified samples was identified using a Shimadzu IR Prestige-21 FT-IR spectrophotometer at Universiti Tenaga Nasional, UNITEN. Spectra were recorded over the 4000-400 cm^{-1} range. A small sample was placed on a KBr plate and covered with another to form a thin film. This film was then positioned in the sample holder, and the spectrum was recorded to analyse the carboxyl group based on its functional group.

Box-Behnken Optimisation of Esterification Reaction Parameters

The response surface methodology, specifically the Box-Behnken approach, was employed to optimise four independent variables: catalyst loading (A), ethanol-to-LA molar ratio (B), reaction time (C), and reaction temperature (D). The range of variables was determined by the literature review and initial experiments conducted (Nelson Appaturi et al., 2022; Trombettoni et al., 2017). The range of experimental values used is as follows: 5-15 wt% catalyst loading, 5:1-15:1 molar ratio of ethanol-to-LA, 3-7 h and 65-75 °C. The experimental data used to optimise the esterification reaction were obtained from Design-Expert, with the conversion rate serving as the response variable. The results of the randomised experimental runs were correlated with the variables via a full quadratic model, Equation 2.

$$Y = \beta_0 + \sum_{i=1}^{4} \beta_i x_i + \sum_{i=1}^{4} \beta_{ii} x_i^2 + \sum_{i=1}^{3} \sum_{j=i+1}^{4} \beta_{ij} x_{ij} \quad [2]$$

where Y represents the response (in percentage), βo is the intercept's regression coefficient, βi is the coefficient for the linear term, βii is the coefficient for the quadratic term, and βij is the coefficient for the interaction term, xi and xij are independent variables, and ε denotes the error.

The independent variables and their ranges used in the optimisation are shown in Table 2. The esterification reaction was carried out under reflux, using LA. The catalytic performance of the most effective magnetic acid catalyst was assessed. Table 3 displays the Box-Behnken experimental design.

Gas Chromatography-Flame Ionisation Detector Analysis

The ethyl levulinate yield was determined by GC (Agilent Technology, 7890A) using a capillary column from DB-WAX (30 m x 0.32 mm x 0.25 µm). The prepared ester samples (1 g) were dissolved in 10 mL of heptane for analysis. The column temperature was 220 °C, while the injector and detector (FID) temperatures were 250 °C and 280 °C, respectively. The yield percentage was determined from the standard EL concentration using the calibration curve and calculated using the following equation Equation 3. The peaks of ethyl levulinate were identified by comparing their retention times with those of the standards.

$$EL_y = \left(\frac{\left(\left(\frac{(C_E \times V_P)}{1000} \right) / M_E \right)}{\left(\frac{W_i}{M_L} \right)} \right) \times 100 \quad [3]$$

where EL_y is EL yield (mol%), C_E is the concentration of LE (mg/ml), M_E and M_L are the MM of EL and LA (g/mol), respectively, V_p is the final volume of the product (mL), and W_i is the weight of LA (g).

Table 2
The variables and ranges for the optimisation study

Variables	Symbol	-1	0	+1
Catalyst (%)	A	5	10	15
Ethanol-to-LA (molar ratio)	B	5:1	10:1	15:1
Time (h)	C	3	5	7
Temperature (^oC)	D	65	70	75

Table 3
Box-Behnken design of the experiment for optimisation study

Run	Catalyst, %	Ethanol: LA	Time, h	Temperature, ^oC
1	10	10:1	7	65
2	15	10:1	5	75
3	15	5:1	5	70
4	5	15:1	5	70
5	5	10:1	5	75
6	10	15:1	3	70
7	10	15:1	5	75
8	5	10:1	3	70
9	10	5:1	5	65
10	10	5:1	5	75
11	10	5:1	3	70
12	5	5:1	5	70
13	10	5:1	7	70
14	10	15:1	7	70
15	10	10:1	7	75
16	10	10:1	5	70
17	10	10:1	3	65
18	5	10:1	5	65
19	10	10:1	3	75
20	10	15:1	5	65
21	15	10:1	3	70
22	15	15:1	5	70
23	5	10:1	7	70
24	15	10:1	5	65
25	15	10:1	7	70

Catalyst Reusability

The reusability study was carried out using the optimised parameters to assess the catalyst's reusability. After each cycle, the catalyst was recovered using a magnet, washed with ethanol until the filtrate was neutral, and dried in an oven at 60 °C overnight for future reactions, without further removal or activation.

RESULTS AND DISCUSSIONS

Effect of EFB-derived Magnetic Catalysts on Ethyl Levulinate Yield

The efficiency of catalysts- EFB fibre/γ-Fe_2O_3, EFB lignin-SO_3H-Fe_3O_4, and EFB MCC/γ-Fe_2O_3- in esterifying levulinic acid was tested using ethanol as the primary alcohol. The reaction was performed using 10 wt% catalyst and an ethanol-to-LA molar ratio of 10:1 at 70 °C for 5 h. Ethyl levulinate levels were analysed using a GC-FID. Figure 1 shows the conversion of LA to EL, with a control experiment (no catalyst) showing only 0.95 mol% conversion. Catalysts significantly improved yields, highlighting the importance of the acid catalyst. The highest yield reached 84.87 mol% with an EFB fibre/γ-Fe_2O_3 catalyst. Yields for EFB lignin-SO_3H-Fe_3O_4 and EFB MCC/γ-Fe_2O_3 catalysts were 80.58 mol% and 74.06 mol%, respectively. Based on FT-IR analysis (Krishnan et al., 2020; Krishnan et al., 2022), oxygen-containing functional groups, such as -SO_3H, -O-, -OH and C=O, were present in both the EFB fibre/γ-Fe_2O_3 and EFB lignin-SO_3H-Fe_3O_4 catalysts, leading to ethyl levulinate yields exceeding 80 mol%. Additionally, the EFB fibre has a wide variety of accessible oxygenated groups (from cellulose, hemicellulose, and lignin), which provide more reactive sites (Krishnan et al., 2023), and EFB lignin has many oxygen groups but is less accessible, thus reducing the yield. On the other hand, EFB MCC has hydroxyl (-OH) groups but fewer diverse oxygen functionalities, which limit conversion.

Compared with other studies on the production of EL using solid acid catalysts, EFB fibre/γ-Fe_2O_3 yields a higher ethyl levulinate yield. Li et al. (2019) achieved excellent catalytic activity for producing ethyl levulinate from carbonic acid using a loofah sponge, with a maximum yield of 91%, but the reaction required 12 h at 80 °C (Li et al., 2019). Catalysts with higher –SO_3H content provide more Brønsted acid sites, which enhance protonation of the levulinic acid carbonyl group, thereby accelerating esterification. In addition, oxygen-containing groups (-OH and C-O) enhance the adsorption of reactants and facilitate interactions with Fe active sites. In this case, EFB fibre/γ-Fe_2O_3 catalyst, which possesses a higher diversity and accessibility of oxygen functionalities, exhibited the highest EL yield.

Chemical Analysis

FT-IR analysis was conducted to confirm the presence of ester groups in the esterified products. Figure 2 displays the FT-IR spectra of the esterified samples prepared with EFB-derived magnetic acid catalysts, alongside a blank sample. A control experiment assessed the catalysts' effectiveness in converting LA. The findings indicate that all esterified samples exhibit peaks identical to those in the control.

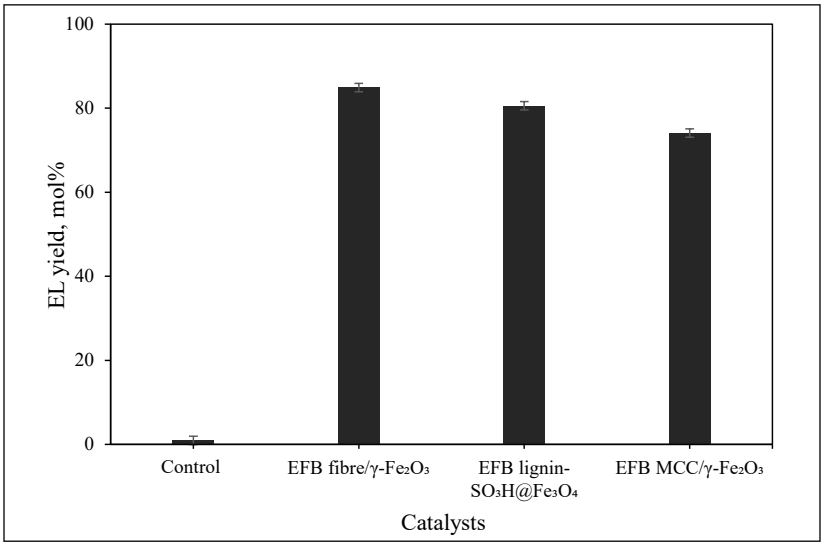


Figure 1. Ethyl levulinate yield using EFB-based magnetic solid catalysts and a blank experiment

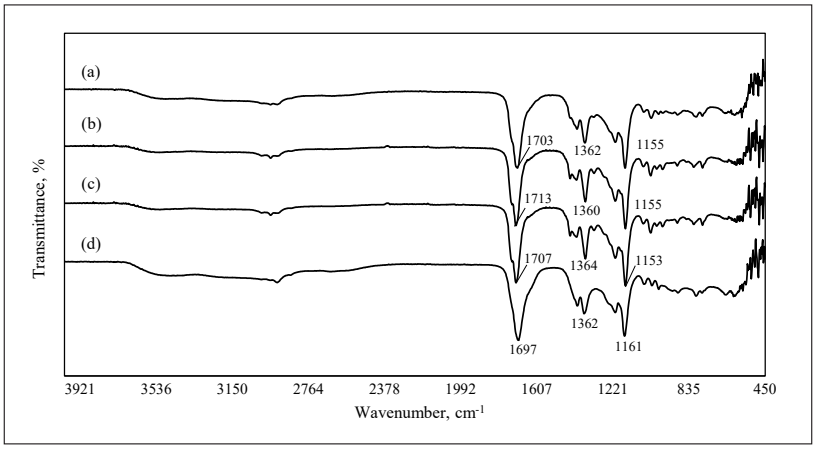


Figure 2. FT-IR spectra of esterified levulinic acid using (a) EFB fibre/γ-Fe_2O_3; (b) EFB lignin-SO_3H-Fe_3O_4; (c) EFB MCC/γ-Fe_2O_3; and (d) control experiment

The sharp absorption bands observed in 1703, 1709, and 1713 cm^{-1} are attributed to the C=O stretching of the carbonyl group, suggesting the presence of aldehyde and ketone groups in these samples, with a closely similar peak at 1697 cm^{-1} in the control sample. Thus, they indicate the presence of esters in the samples (Faraguna et al., 2019). The absorption bands near 1360, 1362, and 1364 cm^{-1} correspond to alkyl C-H bending vibrations, indicating the presence of ethyl groups in the esterified samples. Additionally, bands at 1153, 1155, and 1161 cm^{-1} are associated with C–O stretching of the ester group, which may also be present in the levulinic acid. However, the IR spectra of the esterified samples and the blank sample were justified by the prior ethyl levulinate yield and GC-FID results (Figure S1 to S4) for these samples.

Statistical Analysis and 3D Surface Models via Box-Behnken

Based on the effect of EFB-derived magnetic acid catalysts, EFB fibre/γ-Fe_2O_3 exhibited good catalytic activity. Its efficiency in determining optimal reaction conditions was further investigated using a response surface methodology (RSM) optimisation study. The RSM model's optimum conditions denote the theoretical maximum within the studied parameter space. These conditions were experimentally confirmed, and the EL yield closely matched the predicted value, validating the model's accuracy. The best observed run corresponds to the highest yield among the experimental points, which might differ slightly from the predicted optimum because the model estimates the ideal conditions between discrete levels. A four-factor, three-level optimisation plan was developed employing the Box-Behnken model with the following parameters: catalyst (A), molar ratio of ethanol-to-LA (B), temperature (C), and time (D). A total of 29 experiments were conducted for LA esterification. Table 4 to 6 presents the experimental and predicted conversion results, the analysis of variance, and the statistical model analysis. Figure 3 illustrates the comparison between predicted and experimental (actual) data. The points scattered near the diagonal clearly show the relationship between the predicted and actual conversion rates. Analysis of variance (ANOVA) was used to test the statistical significance of the quadratic model and to evaluate interactions between the reaction parameters and the response.

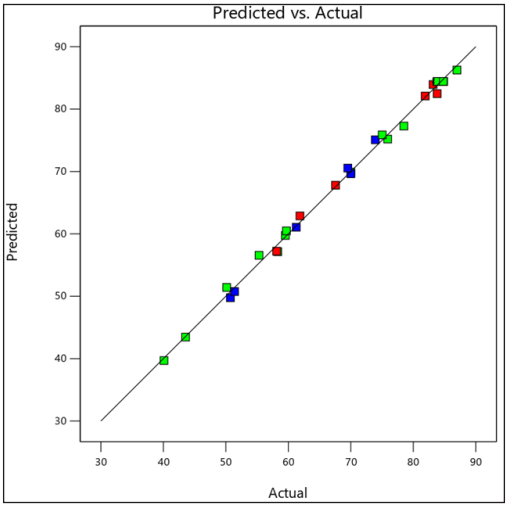


Figure 3. Comparison of actual versus predicted yield data

From the model analysis in Table 5, the larger F-value (304.55) and lower *p*-value (< 0.0001) indicate that the corresponding parameters are significant. Statistical analysis revealed that all parameters and interactions were significant. For the primary parameters, the molar ratio of ethanol-to-LA and reaction time were highly significant ($p < 0.0001$), followed by catalyst loading ($p = 0.0002$) and reaction temperature ($p = 0.0017$). For the interaction parameters, AB, AC, AD, and BC were statistically significant ($p < 0.05$), followed by BD ($p = 0.0085$) and least significant for CD ($p = 0.022$).

Table 4
Experimental results of the response surface analysis of ethyl levulinate yield

No.	Catalyst, %	Ethanol: LA Molar Ratio	Temperature, °C	Time, h	EL Yield, mol%	Predicted Yield, mol%
1	15	15:1	70	5	70.0	71.1
2	10	5:1	70	3	51.4	51.9
3	15	10:1	75	5	75.0	76.6
4	15	10:1	70	7	61.9	63.8
5	10	15:1	65	5	87.0	87.4
6	10	15:1	70	3	69.5	70.3
7	10	15:1	70	7	83.2	84.3
8	5	5:1	70	5	43.6	45.6
9	5	10:1	70	3	50.7	52.0
10	5	15:1	70	5	59.5	61.6
11	10	10:1	70	5	83.8	83.9
12	10	10:1	70	5	83.9	83.9
13	10	5:1	70	7	58.1	57.9
14	10	10:1	75	3	70.0	70.3
15	10	5:1	75	5	59.7	60.0
16	10	10:1	75	7	83.8	84.3
17	10	5:1	65	5	55.3	57.0
18	10	10:1	65	3	73.9	75.3
19	15	5:1	70	5	40.1	41.2
20	5	10:1	75	5	50.1	51.4
21	15	10:1	65	5	58.3	57.9
22	5	10:1	65	5	75.9	77.1
23	10	10:1	65	7	81.9	82.8
24	15	10:1	70	3	61.3	62.0
25	10	10:1	70	5	84.9	84.9
26	5	10:1	70	7	67.6	68.9
27	10	10:1	70	5	84.8	84.8
28	10	10:1	70	5	84.8	84.8
29	10	15:1	75	5	78.5	81.7

Table 5
ANOVA for the quadratic model

	Sum of Squares	df	Mean Square	F Value	P-value	
Model	5471.80	14	390.84	304.55	< 0.0001	significant
A-Catalyst	30.62	1	30.62	23.86	0.0002	
B-Ethanol: LA molar ratio	1621.69	1	1621.69	1263.66	< 0.0001	
C-Temperature	19.20	1	19.20	14.96	0.0017	
D-Time	295.72	1	295.72	230.43	< 0.0001	
AB	48.44	1	48.44	37.75	< 0.0001	
AC	451.35	1	451.35	351.70	< 0.0001	
AD	65.77	1	65.77	51.25	< 0.0001	
BC	41.54	1	41.54	32.37	< 0.0001	
BD	12.01	1	12.01	9.36	0.0085	
CD	8.47	1	8.47	6.60	0.0223	
A^2	2157.53	1	2157.53	1681.20	< 0.0001	
B^2	1095.64	1	1095.64	853.75	< 0.0001	
C^2	10.78	1	10.78	8.40	0.0117	
D^2	219.74	1	219.74	171.23	< 0.0001	
Residual	17.97	14	1.28			
Lack of Fit	16.82	10	1.68	5.88	0.0513	not significant
Pure Error	1.15	4	0.29			
Cor Total	5489.76	28				

Based on the quadratic terms, the significant trend was $A^2 > B^2 > D^2 > C^2$. Therefore, the statistical analysis indicates that the parameters with the most significant impact on ethyl levulinate yield were catalyst loading, molar ratio of ethanol-to-LA, and reaction time. Conversely, the reaction temperature had a lesser effect on yield, although it remains crucial for the model. Referring to Table 4, the equation with coded variables is used to describe the reaction parameters in Equation 4.

$$Y = 84.4 + 1.6\mathrm{A} + 11.6\mathrm{B} - 1.3\mathrm{C} + 4.9\mathrm{D} + 3.5\mathrm{AB} + 10.6\mathrm{AC} - 4.1\mathrm{AD} - 3.2\mathrm{BC} + 1.7\mathrm{BD} + 1.5\mathrm{CD} - 18.2A^2 - 12.9B^2 - 1.3C^2 - 5.8D^2 \quad [4]$$

where A, B, C, and D are the values of the catalyst, ethanol-to-LA molar ratio, temperature, and time, respectively, and Y is the response of the EL yield.

A, B, and D were the most important parameters for achieving higher ethyl levulinate yields in the esterification reaction. According to Table 4, the highest ethyl levulinate yield of 87.0 mol% was obtained under optimal reaction conditions: 10 wt% catalyst loading,

a molar ratio of 15:1 ethanol-to-LA, and 65 °C for 5 h. The statistical analysis of the regression coefficient of the model was carried out using the determination R^2 value, which was 1.00, indicating that the quadratic model Equation 4 predicts 100% of the initial reaction and is highly reliable. Table 6 shows that the adjusted R^2 of 0.99 is close to the predicted R^2 of 0.98. It can be assumed that the model fits well, as the coefficient of variation (C.V.) is 1.65%, which is less than 10%. The regression model thus confirms that ethyl levulinate can be produced from levulinic acid via esterification. According to the Box-Behnken validation, the highest conversion rate of 87.1 mol% was achieved with a catalyst loading of 9.1 wt%, an ethanol-to-LA molar ratio of 13.6:1, at 66.7 °C for 6 h. The regression equation of this model follows the above test approach and is adaptable.

The relationship between the reaction parameters was shown by plotting two independent variables against the yield. The 3D model plots of the surface regression for yield are shown in Figure 4a-f. Figure 4a shows the interaction between the catalyst loading and the molar ratio of ethanol-to-LA at a constant time and temperature of 5 h and 70 °C on LA conversion in the 3D surface regression.

Table 6
Statistical analysis of the regression coefficient

Terms	Coefficient	Standard Error	t-Stat
Intercept	84.43	0.51	85.52
A-Catalyst	1.60	0.33	2.30
B-Ethanol: LA molar ratio	11.63	0.33	12.33
C-Temperature	-1.27	0.33	-0.56
D-Time	4.96	0.33	5.67
AB	3.48	0.57	4.69
AC	10.62	0.57	11.84
AD	-4.06	0.57	-2.84
BC	-3.22	0.57	-2.01
BD	1.73	0.57	2.95
CD	1.46	0.57	2.67
A^2	-18.24	0.44	-17.28
B^2	-13.00	0.44	-12.04
C^2	-1.29	0.44	-0.34
D^2	-5.82	0.44	-4.87
Statistical parameters	Values	Statistical parameters	Values
Standard deviation	1.13	R^2	1.00
Mean	68.57	Adjusted R^2	0.99
C.V.%	1.65	Predicted R^2	0.98
PRESS	98.68	Adequate R^2	57.16

The highest esterification yield was obtained with a high catalyst loading and a high molar ratio of ethanol-to-LA. In contrast, a lower esterification yield of 40.1 mol% was observed at a catalyst loading of 15 wt% and an ethanol-to-LA molar ratio of 5:1, probably due to the higher catalyst loading and the limited amount of ethanol used to facilitate the reaction. At lower catalyst loadings, the esterification yield was less than 60 mol% due to reduced contact between the active sites and the reaction reactants. The interaction term between catalyst loading and the molar ratio of ethanol-to-LA (F value 37.75) shows that it moderately influences the yield. Thus, the esterification yield with 10 wt% catalyst loading and a 10:1 molar ratio of ethanol-to-LA was 84.8 mol%.

Figure 4b illustrates the relationship between catalyst loading and reaction temperature in the 3D surface response model for EL preparation, maintained at constant time and ethanol-to-LA of 5 h and 10:1, respectively. Increasing the catalyst loading from 5 wt% to 10 wt% and raising the temperature from 65 to 70 °C led to a higher yield. However, at the lower catalyst loading (5 wt%) and higher reaction temperature (75 ^{o}C), the lowest yield of 50.1 mol% was observed, whereas at 15 wt% catalyst loading and 65 ^{o}C, the yield was 58.3 mol%. This is because, at lower catalyst loading, only a limited number of active sites are available for reaction. The addition of the catalyst to the reaction no longer has a significant effect once the maximum conditions are reached (Nindya et al., 2020). Furthermore, at higher reaction temperatures, the diffusion of reactants to the catalyst's active site increases. However, the large size of the reaction molecules (bulky intermediates) and the small pore size of the catalyst (as shown in Table 1) could reduce diffusion, preventing an increase in conversion rate. This indicates that the yield was low at extreme catalyst loadings and reaction temperatures, suggesting particle agglomeration and formation of larger intermediates, which deactivate the catalysts and block access to the active sites. The highest yield of 84.9 mol% was achieved with a 10 wt% catalyst loading at 70 °C.

Figure 4c presents a response surface plot of the relationship between catalyst loading and reaction time, with a constant ethanol-to-LA molar ratio of 10:1 and temperature at 70 °C. A high yield was discovered with moderate catalyst loading and reaction duration. The active sites enabled the reactants to react at a specific time, resulting in a high esterification yield. However, the yield decreased with decreasing catalyst loading at both shorter and longer reaction times. It could be due to the reaction mixture containing an excessive amount of catalyst at the shorter reaction time and a limited number of active sites at the longer reaction time (Zhang et al., 2019). Furthermore, increasing the catalyst loading reduced the conversion to 61.3 mol% due to inadequate interaction between the excess catalyst at a shorter reaction time of 3 h. At higher catalyst loadings, agglomeration of catalyst particles and increased diffusion resistance restrict access to –SO_3H and Fe active sites, thereby decreasing effective reactant-catalyst interactions and reducing the

esterification conversion. The maximum yield was 84.9 mol% with a 10 wt% catalyst loading for 5 h.

Figure 4d shows how the ethanol-to-LA molar ratio and reaction temperature interact to affect the yield, with a fixed catalyst loading of 10 wt% and a reaction time of 5 h. It suggests that the low accessibility of the catalyst surface, as the alcohol concentration decreased, led to a limited esterification process (Peng et al., 2014). At 65 °C, a lower yield of 55.3 mol% was obtained with a 5:1 ethanol-to-LA molar ratio. Despite this, a 15:1 ethanol-to-LA molar ratio at 65 °C significantly improved LA conversion to 87.0 mol%. As a result, increasing the molar ratio improved the yield by allowing better contact with the catalyst and exposing the active sites, thereby accelerating the process (compared to others). Research showed that a higher ethanol-to-LA molar ratio shifts the reaction equilibrium towards ethyl levulinate formation, suppresses the reverse hydrolysis by diluting the water, and improves solvation, while the increased temperature (65 °C) enhances the reaction and yields a higher EL (Sušjenka et al., 2024; Zainol et al., 2021). According to Thangaraj et al. (2019), acid catalysts yield well, but the reaction rate is sluggish, necessitating a higher alcohol-to-molar ratio to improve the reaction (Thangaraj et al., 2019). Kothe et al. (2020) found that an ethanol-to-LA molar ratio of 9:1 produced highly selective ethyl levulinate with up to 80% selectivity (Kothe et al., 2020). The interaction between the ethanol-to-LA molar ratio and reaction temperature revealed that both factors significantly affect the yield. As a result, at the maximum ethanol-to-LA molar ratio and moderate reaction temperature, a greater yield was attained.

Figure 4e shows the interaction between the ethanol-to-LA molar ratio and reaction time on the conversion of LA. When catalyst loading and reaction temperature were held constant at 10 wt% and 70 °C, respectively, the 3D surface model revealed that a moderate ethanol-to-LA molar ratio and a reaction temperature of 70 °C yielded higher conversion, ranging from 83.8 to 84.9 mol%. The conversion of LA was lower at a lower ethanol-to-LA molar ratio and a reaction time of 5:1 and 3 h, respectively. This is due to the limited amount of ethanol, as further adding ethanol might improve the yield (Nandiwale et al., 2013). However, increasing the ethanol-to-LA molar ratio and reaction time does not affect LA conversion because more ethanol dilutes the reactants, thereby limiting the conversion rate (Gautam et al., 2022). As a result, the maximum yield of 84.9 mol% was obtained at a 10:1 ethanol-to-LA molar ratio and a reaction time of 5 h. The result indicates that the catalyst achieves a high yield under moderate conditions.

The interaction between reaction temperature and duration does not influence the esterification of LA, as demonstrated in Figure 4f. The yield decreased as the interaction increased. At 70 °C and a 5 h reaction period, the most significant yield of 84.9 mol% was recorded. Furthermore, increasing the parameter value has no discernible influence on yield, with a slight decrease to 83.8 mol% after 7 h at 75 °C.

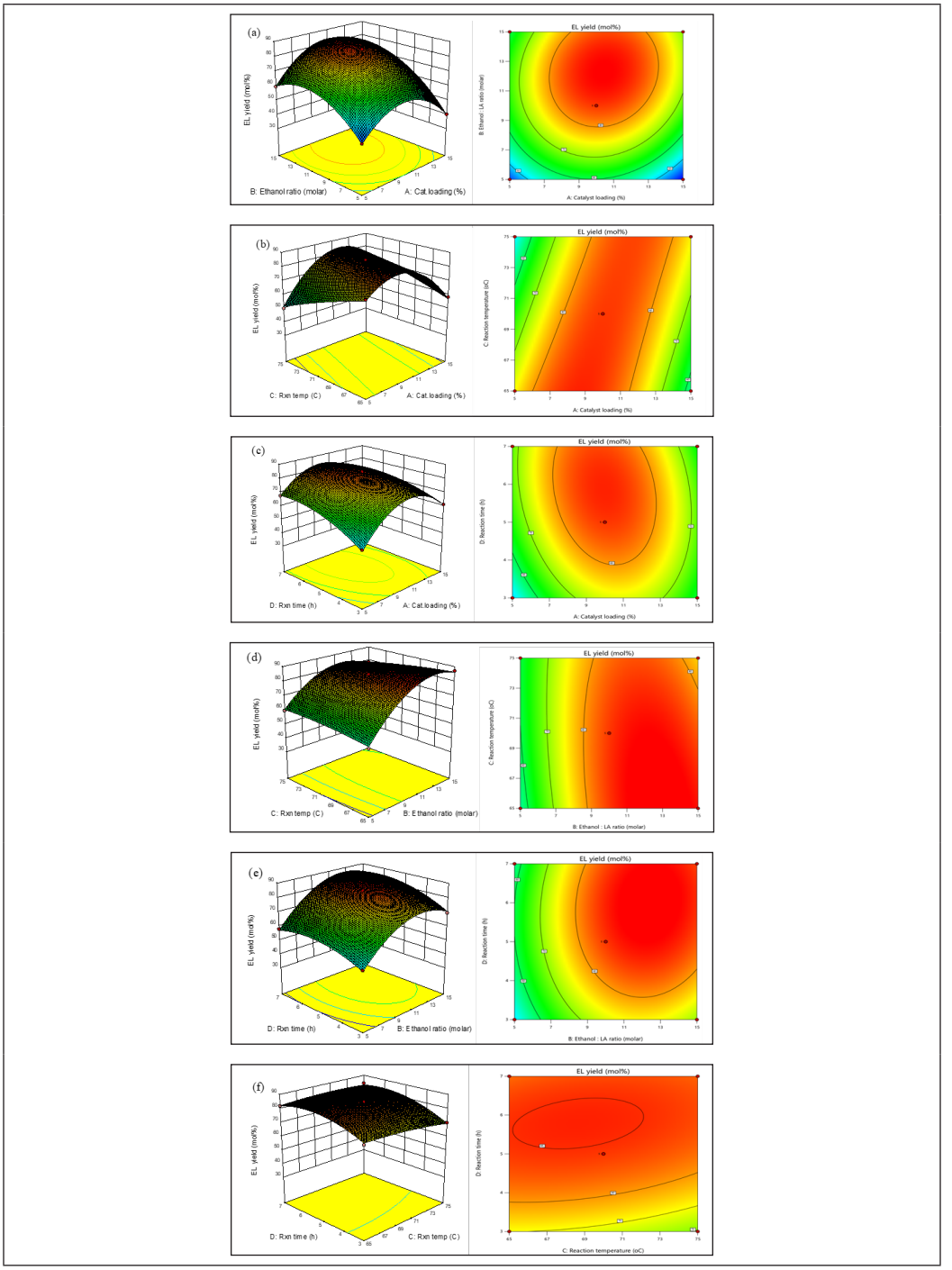


Figure 4. Surface regression 3D model plots and contour plots on ethyl levulinate yield with the interaction between (a) catalyst loading (wt%) and ethanol-to-LA molar ratio; (b) catalyst loading (wt%) and reaction temperature (°C); (c) catalyst loading (wt%) and reaction time (h); (d) ethanol-to-LA molar ratio and reaction temperature (°C); (e) reaction time (h) and ethanol-to-LA molar ratio; and (f) reaction temperature (°C) and reaction time (h)

However, the lowest yield of 70.0 mol% was observed after 3 h at 75 °C with minimal decrease. As a result, the interplay between reaction temperature and time has little effect on LA esterification. Kumaravel et al. (2019) found that LA conversion increased with increasing reaction temperature but did not show a significant yield (Kumaravel et al., 2019). Wang et al. (2020) found that reaction time does not substantially affect yield, and that the yield remains stationary before dropping (Wang et al., 2020). The analysis showed that the maximum yield of 87.0 mol% was achieved using the EFB fibre/γ-Fe_2O_3 with 9.1 wt% catalyst loading and 13.6:1 ethanol-to-LA molar ratio at 66.7 °C for 6 h.

Recyclability Study of EFB fibre/γ-Fe_2O_3 Catalyst

In industrial applications, catalyst stability is critical, and recyclability is essential in heterogeneous catalysis. Catalyst stability is repeatedly evaluated over several cycles. Figure 5 represents the EFB fibre/γ-Fe_2O_3 catalyst reusability for six cycles. According to the findings, the conversion of LA reduced from 87.0 mol% to 83.4 mol% after the second use and to 78.5 mol% after the fourth cycle. The catalyst activity remained steady after the 5th cycle, with only a slight decrease in LA conversion to 74.8 mol% after the sixth cycle. The decrease in the yield is most likely due to the leaching of active sites from the catalytic support by reactant contact (Trombettoni et al., 2018). However, further studies are necessary to definitively establish the reduction in catalytic yield. Furthermore, after 6 cycles, the catalytic activity of the EFB-derived magnetic acid catalyst did not decrease significantly, confirming its strong stability during LA esterification. The catalyst was reused without further regeneration to assess its stability and simulate practical process conditions where additional treatment steps are undesirable. This method shows the catalyst's ability to maintain activity with minimal processing. The gradual performance decline suggests that a mild regeneration step, such as thermal treatment, could restore active sites.

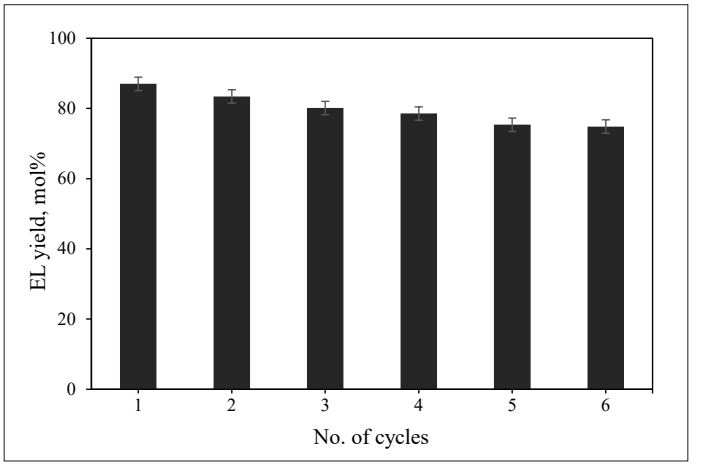


Figure 5. Recyclability of EFB fibre/γ-Fe_2O_3 for esterification of LA (optimised reaction parameters)

Mechanism of Catalytic Reaction

Figure 6 illustrates a possible reaction process of catalytic esterification of levulinic acid with ethanol in the presence of EFB fibre/γ-Fe_2O_3. FTIR analysis confirms the presence of –SO_3H groups, indicating Brønsted acid sites, while XRD patterns verify the formation of Fe-based phases (Fe-O), which can provide Lewis acidic centres. The coexistence of these functionalities suggests that both Brønsted and Lewis acid sites may contribute synergistically to the esterification reaction. Several steps are involved in the catalytic mechanism: (1) the first step consists of the acidic proton of the –SO_3H group interacting with the carbonyl oxygen of levulinic acid, forming protonated levulinic acid intermediates, (2) the protonation process enhances the electrophilic carbonyl carbon and the nucleophilic alcohol, and ethanol charges the carbonyl carbon to form oxonium ions (H_3O^+) (Guo et al., 2019); (3) the formed oxonium ions generate new sets of oxonium ions, leading to water loss of water; and (4) deprotonation proceeds to the formation of ethyl levulinate. The deprotonation process regenerates the acidic sites of the magnetic acid catalyst. Thus, using an acidic catalyst increases the esterification yield by catalysing the conversion of levulinic acid.

As for the Lewis acid sites, iron (Fe) is also responsible for the LA esterification reaction (Figure 7). In the first step, LA is directly coordinated with the surface of Fe^{3+} sites. In the following steps, the Lewis acid sites followed the same steps as the Bronsted acid, which catalysed the esterification reaction (Shestakova et al., 2021). The generation of oxonium ions corresponded to the acidity of the catalyst, and the presence of both Bronsted and Lewis acids aided in the conversion of LA to ethyl levulinate.

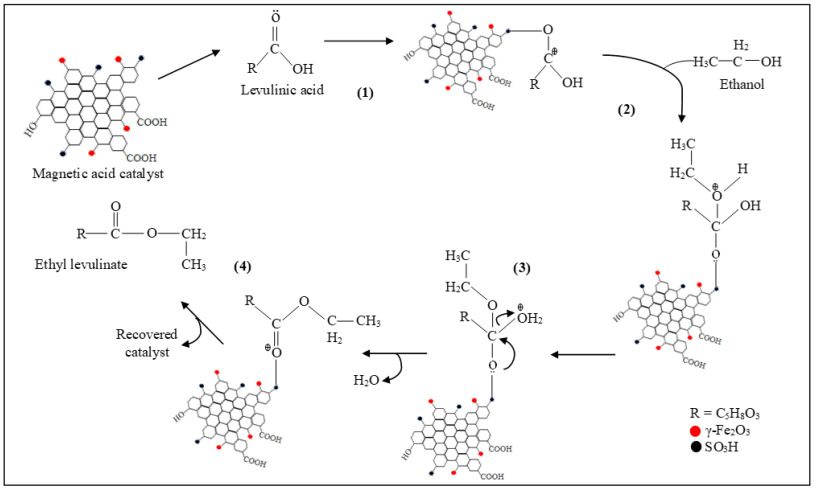


Figure 6. Proposed mechanism of Bronsted acid in the esterification of levulinic acid using EFB fibre/γ-Fe_2O_3 catalyst

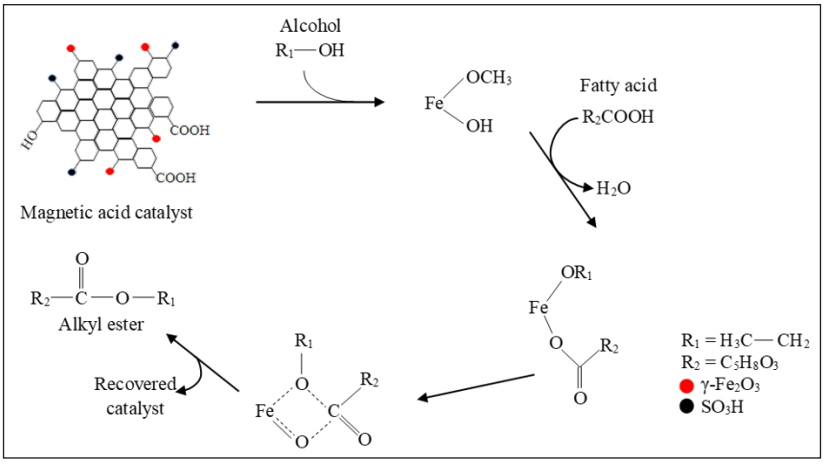


Figure 7. Proposed mechanism of Lewis acid in esterification reaction using EFB fibre/γ-Fe_2O_3 catalyst

CONCLUSION

The effectiveness of catalysts such as EFB fibre/γ-Fe_2O_3, EFB lignin-SO_3H-Fe_3O_4, and EFB MCC/γ-Fe_2O_3 in producing ethyl levulinate was evaluated. The EFB fibre/γ-Fe_2O_3 catalyst achieved the highest yield of 84.87 mol%, followed by EFB lignin-SO_3H-Fe_3O_4 with 80.58 mol%, and EFB MCC/γ-Fe_2O_3 with 74.06 mol%. A box-Behnken optimisation was conducted to assess the impact of reaction parameters on levulinic acid esterification. The maximum esterification yield of 87.0 mol% was obtained under optimised conditions with 9.1 wt% catalyst loading of EFB fibre/γ-Fe_2O_3, an ethanol-LA molar ratio of 13.6:1, at 66.7 °c, over 6 hours. Statistical analysis identified catalyst loading, ethanol-LA molar ratio, and reaction time as the most influential factors. This study emphasises the potential of biomass-based magnetic catalysts for sustainable chemical production, encouraging the utilisation of agricultural waste across various sectors and reducing landfill problems.

ACKNOWLEDGEMENT

This work was supported by the Dato' Low Tuck Kwong International Research Grant (20238024DLTK) and the Fundamental Research Grant Scheme (20180108FRGS). The authors also gratefully acknowledge the funding and facilities support provided by the UNITEN–AAIBE Project, Chair of Renewable Energy Grant (202303KETTHA).

REFERENCES

Adeleye, A. T., Louis, H., Akakuru, O. U., Joseph, I., Enudi, O. C., & Michael, D. P. (2019). A review on the conversion of levulinic acid and its esters to various useful chemicals. *AIMS Energy, 7*(2), 165-185. https://doi.org/10.3934/energy.2019.2.165

Campana, F., Valentini, F., Marrocchi, A., & Vaccaro, L. (2023). Urban waste upcycling to a recyclable solid acid catalyst for converting levulinic acid platform molecules into high-value products. *Biofuel Research Journal, 10*(4), 1989-1998. https://doi.org/10.18331/BRJ2023.10.4.5

Costa, B. E. B., da Silva, A. O. S., & Meneghetti, S. M. P. (2024). Esterification of levulinic acid with different alcohols using mesoporous stannosilicates as the catalyst. *ACS Omega, 9*(28), 31128-31135. https://doi.org/10.1021/acsomega.4c04598

Enumula, S. S., Gurram, V. R. B., Chada, R. R., Burri, D. R., & Kamaraju, S. R. R. (2017). Clean synthesis of alkyl levulinates from levulinic acid over one-pot synthesised WO3-SBA-16 catalyst. *Journal of Molecular Catalysis A: Chemical, 426*, 30-38. https://doi.org/10.1016/j.molcata.2016.10.032

Faraguna, F., Racar, M., & Jukić, A. (2019). Test method for determination of different biodiesels (fatty acid alkyl esters) content in diesel fuel using FTIR-ATR. *Renewable Energy, 133*, 1231-1235. https://doi.org/10.1016/j.renene.2018.09.010

Gautam, P., Barman, S., & Ali, A. (2022). A comparative study on the performance of acid catalysts in the synthesis of levulinate ester using biomass-derived levulinic acid: A review. *Biofuels, Bioproducts and Biorefining, 16*(4), 1095-1115. https://doi.org/10.1002/bbb.2351

Gielen, D., Boshell, F., Saygin, D., Bazilian, M. D., Wagner, N., & Gorini, R. (2019). The role of renewable energy in the global energy transformation. *Energy Strategy Reviews, 24*, 38-50. https://doi.org/10.1016/j.esr.2019.01.006

Guo, T., Qiu, M., & Qi, X. (2019). Selective conversion of biomass-derived levulinic acid to ethyl levulinate catalysed by metal organic framework (MOF)-supported polyoxometalates. *Applied Catalysis A: General, 572*, 168-175. https://doi.org/10.1016/j.apcata.2019.01.004

Kothe, V., Melfi, D. T., dos Santos, K. C., Corazza, M. L., & Ramos, L. P. (2020). Thermodynamic analysis, experimental and kinetic modelling of levulinic acid esterification with ethanol at supercritical conditions. *Fuel, 260*, Article 116376. https://doi.org/10.1016/j.fuel.2019.116376

Krishnan, S. G., Pua, F. L., & Fan, Z. (2023). Effect of chemical pre-treatment on the catalytic performance of oil palm EFB fibre-supported magnetic acid catalyst. *Sustainability, 15*(11), Article 8637. https://doi.org/10.1016/j.matpr.2020.01.513

Krishnan, S. G., Pua, F. L., & Jaafar, S. N. S. (2020). Synthesis and characterisation of a local biomass-supported magnetic catalyst for esterification reaction. *Materials Today: Proceedings, 31*, 161-165. https://doi.org/10.1016/j.matpr.2020.01.513

Krishnan, S. G., Pua, F. L., & Zhang, F. (2022). Oil palm empty fruit bunch-derived microcrystalline cellulose-supported magnetic acid catalyst for esterification reaction: An optimisation study. *Energy Conversion and Management: X, 13*, Article 100159. https://doi.org/10.1016/j.ecmx.2021.100159

Kumaravel, S., Thiripuranthagan, S., Radhakrishnan, R., Erusappan, E., Durai, M., Devarajan, A., & Mukannan, A. (2019). Liquid phase esterification of levulinic acid into ethyl levulinate over sulphobenzylated nanoporous Al-SBA-15 catalyst. *Journal of Nanoscience and Nanotechnology, 19*(11), 6965-6977. https://doi.org/10.1166/jnn.2019.16637

Li, N., Jiang, S., Liu, Z. Y., Guan, X. X., & Zheng, X. C. (2019). Preparation and catalytic performance of loofah sponge-derived carbon sulfonic acid for the conversion of levulinic acid to ethyl levulinate. *Catalysis Communications, 121*, 11-14. https://doi.org/10.1016/j.catcom.2018.12.007

Liu, X., Yang, W., Zhang, Q., Li, C., & Wu, H. (2020). Current approaches to alkyl levulinates via efficient valorisation of biomass derivatives. *Frontiers in Chemistry, 8*, Article 794. https://doi.org/10.3389/fchem.2020.00794

Martínez Aguilar, M., Duret, X., Ghislain, T., Minh, D. P., Nzihou, A., & Lavoie, J. M. (2020). A simple process to produce fuel additives using residual lignocellulosic biomass. *Fuel, 264*, Article 116702. https://doi.org/10.1016/j.fuel.2019.116702

Nandiwale, K. Y., Sonar, S. K., Niphadkar, P. S., Joshi, P. N., Deshpande, S. S., Patil, V. S., & Bokade, V. V. (2013). Catalytic upgrading of renewable levulinic acid to ethyl levulinate biodiesel using dodecatungstophosphoric acid supported on desilicated H-ZSM-5 as a catalyst. *Applied Catalysis A: General, 460-461*, 90-98. https://doi.org/10.1016/j.apcata.2013.04.024

Nelson Appaturi, J., Andas, J., Ma, Y.-K., Lee Phoon, B., Muazu Batagarawa, S., Khoerunnisa, F., Hazwan Hussin, M., & Ng, E.-P. (2022). Recent advances in heterogeneous catalysts for the synthesis of alkyl levulinate biofuel additives from renewable levulinic acid: A comprehensive review. *Fuel, 323*, Article 124362. https://doi.org/10.1016/j.fuel.2022.124362

Nindya, C. C. S., Anggara, D. R., Nuryoto, & Teguh, K. (2020). Esterification of glycerol (byproduct in biodiesel production) with oleic acid using mordenite natural zeolite as catalyst: Study of reaction temperature and catalyst loading effect. *IOP Conference Series: Materials Science and Engineering, 909*, Article 012001. https://doi.org/10.1088/1757-899X/909/1/012001

Ogino, I., Suzuki, Y., & Mukai, S. R. (2018). Esterification of levulinic acid with ethanol catalysed by sulfonated carbon catalysts: Promotional effects of additional functional groups. *Catalysis Today, 314*, 62-69.

Patel, A., & Dave, R. (2025). Sugarcane bagasse, a next-generation biomass-derived catalytic material for synthesis of biofuel additives, levulinate esters: Process optimisation and kinetic investigation. *Next Materials, 9*, Article 100979. https://doi.org/10.1016/j.nxmate.2025.100979

Peng, L., Li, H., Xi, L., Chen, K., & Chen, H. (2014). Facile and efficient conversion of furfuryl alcohol into n-butyl levulinate catalysed by extremely low acid concentration. *BioResources, 9*(3), 3825-3834. https://doi.org/10.15376/biores.9.3.3825-3834

Saperi, N. Q., Hassan, A. H., Zainol, M. M., Asmadi, M., Rafizan, A., Daud, M., Yusof, M. Y., & Anggoro, D. D. (2024). Catalytic reaction study of levulinic acid esterification to ethyl levulinate using modified carbon catalyst and acidic DES. *Malaysian Journal of Analytical Sciences, 28*, 724-736.

Shestakova, P., Popova, M., Szegedi, Á., Lazarova, H., Nga Luong, T. K., Trendafilova, I., Mihály, J., & Parac-Vogt, T. N. (2021). Hybrid catalyst with combined Lewis and Brønsted acidity based on ZrIV-substituted polyoxometalate grafted on mesoporous MCM-41 silica for esterification of renewable levulinic acid. *Microporous and Mesoporous Materials, 323*, Article 111203. https://doi.org/10.1016/j.micromeso.2021.111203

Signoretto, M., Taghavi, S., Ghedini, E., & Menegazzo, F. (2019). Catalytic production of levulinic acid (LA) from actual biomass. *Molecules, 24*(15), Article 2760. https://doi.org/10.3390/molecules24152760

Sušjenka, M., Prašnikar, F., Jakovljević Kovač, M., Molnar, M., Fernandes, L., Nunes, A. V. M., Duarte, A. R. C., & Zakrzewska, M. E. (2024). A green chemistry approach to catalytic synthesis of ethyl levulinate. *Catalysts, 14*(12), Article 842. https://doi.org/10.3390/catal14120842

Thangaraj, B., Solomon, P. R., Muniyandi, B., Ranganathan, S., & Lin, L. (2019). Catalysis in biodiesel production: A review. *Clean Energy, 3*(1), 2-23. https://doi.org/10.1093/ce/zky020

Trombettoni, V., Bianchi, L., Zupanic, A., Porciello, A., Cuomo, M., Piermatti, O., Marrocchi, A., & Vaccaro, L. (2017). Efficient catalytic upgrading of levulinic acid into alkyl levulinates by resin-supported acids and flow reactors. *Catalysts, 7*(8), Article 235. https://doi.org/10.3390/catal7080235

Trombettoni, V., Lanari, D., Prinsen, P., Luque, R., Marrocchi, A., & Vaccaro, L. (2018). Recent advances in sulfonated resin catalysts for efficient biodiesel and bio-derived additives production. *Progress in Energy and Combustion Science, 65*, 136-162. https://doi.org/10.1016/j.pecs.2017.11.001

Wang, H., Lu, Y., Liu, H., Yin, Y., & Liang, J. (2020). Preparation and application of magnetic nano-solid acid catalyst Fe3O4-PDA-SO3H. *Energies, 13*(6), Article 1484. https://doi.org/10.3390/en13061484

Yan, L., Yao, Q., & Fu, Y. (2017). Conversion of levulinic acid and alkyl levulinates into biofuels and high-value chemicals. *Green Chemistry, 19*(23), 5527-5547. https://doi.org/10.1039/C7GC02503C

Zainol, M. M., Asmadi, M., Iskandar, P., Wan Ahmad, W. A. N., Amin, N. A. S., & Hoe, T. T. (2021). Ethyl levulinate synthesis from biomass derivative chemicals using iron doped sulfonated carbon cryogel catalyst. *Journal of Cleaner Production, 281*, Article 124686. https://doi.org/10.1016/j.jclepro.2020.124686

Zhang, H., Li, H., Hu, Y., Venkateswara Rao, K. T., Xu, C., & Yang, S. (2019). Advances in production of bio-based ester fuels with heterogeneous bifunctional catalysts. *Renewable and Sustainable Energy Reviews, 114*, Article 109296. https://doi.org/10.1016/j.rser.2019.109296

Zhao, W., Ding, H., Zhu, J., Liu, X., Xu, Q., & Yin, D. (2020). Esterification of levulinic acid into n-butyl levulinate catalysed by sulfonic acid-functionalised lignin-montmorillonite complex. *Journal of Bioresources and Bioproducts, 5*(4), 291-299. https://doi.org/10.1016/j.jobab.2020.10.008

Zhou, S., Liu, Y., Zhang, Y., Liu, Z., Wu, Y., & Liu, X. (2025). Organic–inorganic hybrid solid acid catalyst supported on titanium nanotubes for efficient synthesis of n-butyl levulinate. *Carbon Resources Conversion*, Article 100395. https://doi.org/10.1016/j.crcon.2025.100395

Pertanika J. Sci. & Technol. 34 (3): 1331 - 1354 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

YOLOv9-ResCBAM: Enhancing Tomato Ripeness Detection in Greenhouses Through Advanced Object Detection

Jalal Uddin Md Akbar[1], Syafiq Fauzi Kamarulzaman[1]*, Muhammad Danial Mohamad Rizwan[1], Riadul Islam Rabbi[2], and Ekramul Haque Tusher[1]

[1]Faculty of Computing, Universiti Malaysia Pahang, Al-Sultan Abdullah, 26600 Pekan, Pahang, Malaysia
[2]Faculty of Engineering and Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450 Bukit Beruang, Melaka, Malaysia

ABSTRACT

Accurate object detection and classification are pivotal in precision agriculture for tasks such as identifying crop varieties and assessing ripeness stages. To optimise the yield and quality of tomatoes within the variable conditions of a greenhouse, it is crucial to accurately detect and classify their ripeness levels. This study introduces YOLOv9-ResCBAM, an enhanced object detection model based on the advanced YOLOv9 architecture, designed to classify tomato ripeness levels (fully ripe, partially ripe, and unripe) in greenhouse environments. Firstly, a comprehensive tomato image dataset reflecting diverse greenhouse conditions was curated. Secondly, rigorous preprocessing techniques, including auto-orientation, resizing, and augmentation, were applied to enhance dataset quality. Finally, the proposed YOLOv9-ResCBAM model was trained and evaluated. The findings indicated that the proposed model achieved a superior mean Average Precision (mAP@0.5) of 0.912, outperforming both earlier YOLO-based detectors and other established object detection frameworks like SSD, Mask R-CNN, and Faster R-CNN. This improvement is attributed to innovations like Programmable Gradient Information (PGI), Generalised Efficient Layer Aggregation Network (GELAN), and our integration of a Residual Convolutional Block Attention Module (ResCBAM). YOLOv9-ResCBAM's exceptional performance in accurately detecting and categorising tomatoes across ripeness stages, even in challenging greenhouse scenarios, provides a promising advancement for agricultural object detection.

ARTICLE INFO

Article history:
Received: 22 June 2025
Accepted: 27 March 2026
Published: 12 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.03

E-mail addresses:
jalaluddinmdakbar00@gmail.com (Jalal Uddin Md Akbar)
syafiq29@umpsa.edu.my (Syafiq Fauzi Kamarulzaman)
danialrizwan1108@gmail.com (Muhammad Danial Mohamad Rizwan)
riadul.rabbi72@gmail.com (Riadul Islam Rabbi)
ekramulhaquetusher@gmail.com (Ekramul Haque Tusher)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

This study serves as a foundational step toward AI-driven solutions in crop management, enabling more efficient resource allocation and ultimately contributing to more sustainable and optimised food production practices.

Keywords: Agricultural automation, attention mechanism, computer vision, object detection, smart agriculture, YOLO, YOLOv9

INTRODUCTION

Advancements in Computer Vision (CV) algorithms, particularly in object detection, have revolutionised numerous industries, including agriculture. The application of Convolutional Neural Networks (CNNs) marked the beginning of significant strides in this domain (Akbar et al., 2024). Among the notable breakthroughs in this field, the You Only Look Once (YOLO) series stands out for its revolutionary approach to detecting objects in a single forward pass, markedly enhancing processing speeds and detection accuracy. Throughout its evolution, each iteration of the YOLO architecture has been developed with the goal of advancing the model's performance in terms of both computational efficiency and accuracy. YOLOv9 (Wang et al., 2024), developed by Chien-Yao Wang and colleagues, is the latest in this innovative lineage, offering substantial improvements over its predecessors. This version introduces two novel concepts: Programmable Gradient Information (PGI) and Generalised Efficient Layer Aggregation Network (GELAN). Building upon these advancements, we propose YOLOv9-ResCBAM, an enhanced model that integrates a Residual Convolutional Block Attention Module (ResCBAM) into the YOLOv9 architecture to further improve detection accuracy.

Crop ripeness detection and classification have emerged as crucial endeavours in the realm of precision agriculture, enabling informed decision-making and optimised yield management (Rizzo et al., 2023). The conventional approach for these tasks has been manual inspection, a method that is not only time-consuming and labour-intensive but is also subject to inconsistencies and human error. The advent of computer vision and deep learning techniques has introduced opportunities for automation, offering accurate, real-time, and scalable solutions to address these challenges.

Early approaches to crop ripeness detection relied on methods like colour and texture analysis (Akbar et al., 2023; Moreira et al., 2022; Wang et al., 2023; Yang et al., 2023). Despite their potential, the reliability of these methods was frequently compromised by environmental variables, including fluctuating lighting conditions, shadows, and occlusions. These factors consequently constrained their applicability and robustness for large-scale agricultural operations.

With the rapid advancements in deep learning and computer vision techniques, researchers have increasingly turned to data-driven approaches for crop ripeness

detection and classification. Convolutional Neural Networks (CNNs) function as powerful instruments for identifying and extracting hierarchical visual features from images, a capability that facilitates accurate object detection and classification tasks. In the context of crop ripeness detection, approaches that utilise deep learning have proven to be more effective than conventional methods (Phoophuangpairoj et al., 2023; Tang et al., 2023). These techniques can learn intricate patterns and representations directly from labelled data, allowing them to adapt to the complexities of real-world agricultural environments. However, the effectiveness of deep learning models depends greatly on the accessibility of high-quality, diverse, and annotated datasets, which can be challenging and resource-intensive to obtain.

In this study, Tomatoes were chosen due to their economic importance and the complexities involved in accurately detecting their ripeness stages. As one of the most widely grown vegetables globally, tomatoes are a critical component of the agricultural production and food supply chains (Brodt et al., 2013). Their ripeness directly affects the market value, taste, and nutritional content, making accurate detection crucial for both farmers and consumers (Cui et al., 2022). Additionally, tomatoes present a variety of ripeness indicators, including colour, texture, and firmness, which pose challenges to detection systems. These complexities make tomatoes an ideal subject for evaluating the capabilities of advanced CV algorithms like YOLOv9. By focusing on tomatoes, this study aimed to address a significant agricultural challenge and demonstrate the practical applications of cutting-edge technology in enhancing crop management and production efficiency.

Several investigations have examined the application of YOLO-based models for detecting crop ripeness, yielding promising results. For instance, Li et al. (2023) introduced a tomato maturity detection system utilising YOLOv5, with a high precision and mean Average Precision (mAP) of 0.96 for tomato ripeness classification in greenhouse settings. However, their model might struggle under outdoor conditions owing to variations in lighting and environmental factors. Greenhouses provide controlled environments that are less variable than those of outdoor fields. Hence, the robustness of their model under varying outdoor conditions remains a concern. Our approach aims to enhance adaptability by incorporating additional data augmentation techniques and a more diverse dataset that includes various environmental conditions. Another study by Li et al. (2023) proposed MHSA-YOLOv8, a YOLOv8- based model enhanced with a multi-head self-attention mechanism, for the grading and counting of tomato ripeness in complex environments. Although the model was effective in complex environments, its performance in severe occlusion situations was limited. This limitation is critical, as agricultural fields often present significant occlusions due to foliage.

The model achieved a mean Average Precision (mAP) of 0.94; however, its robustness under occluded conditions remains a challenge. Zeng et al. (2023) came up with a lightweight approach for tomato detection based on enhanced YOLOv5 and MobileNetV3, emphasising real-time detection and mobile implementation for classifying tomato ripeness. The model achieved a mAP of 0.97 and a detection speed of 42.5 ms on CPU platforms. Although the model demonstrated impressive real-time performance, it struggled with dense small-target detection, which is common in crops such as tomatoes, where fruits are often clustered. Su et al. (2022) introduced SE-YOLOv3-MobileNetV1, a lightweight network for tomato maturity classification under natural greenhouse conditions, which achieved a high mAP of 0.92. However, their study did not discuss the performance of their model in real-time detection scenarios. Real-time applications are crucial for practical agricultural applications.

Wang et al. (2023) presented an improved YOLOv5n model with a coordinate attention mechanism and a novel loss function for real-time cherry tomato maturity detection, achieving a mAP of 0.89. However, their research was limited to a specific cherry tomato dataset, raising concerns about the generalisability of the model to other tomato varieties. Our study addresses this limitation by training the model on a broader dataset that includes multiple tomato varieties to ensure wider applicability. Badeka et al. (2023) explored a deep learning approach using YOLOv7 to assess grape maturity in precision viticulture and achieved a mAP of 0.88. However, a limitation of that study was its sole concern with a particular type of grape, and it did not explore how the model would perform with other grape varieties. Chen et al. (2024) developed MTD-YOLOv7, a multi-task deep convolutional neural network based on YOLOv7, for the simultaneous detection of cherry tomato fruit bunches, fruit maturity, and cluster maturity. The model obtained a mean Average Precision (mAP) of 0.90 for maturity detection but faced challenges with densely packed tomatoes. Additionally, to accurately identify blueberry fruits at various ripeness levels, Wang et al. (2024) developed the YOLO-BLBE model. Their approach involved optimising the YOLOv5s model by adding a colour enhancement algorithm, which resulted in a mean Average Precision (mAP) of 0.91. Despite these advancements, future work is needed to identify semi-mature blueberries at different growth stages.

Recent research has explored the integration of attention mechanisms into object detection models to enhance their ability to focus on relevant features. For instance, Chien et al. (2024) combined object detection with attention mechanisms to detect paediatric wrist fractures, demonstrating improved performance. In a related study, Zamri et al. (2024) improved the detection of small drones by optimising a YOLOv8 model with the integration of attention mechanisms. However, the potential of combining such mechanisms with the latest YOLO architecture for agricultural applications, particularly in complex greenhouse environments, remains largely unexplored.

Recent studies have also successfully explored attention mechanisms for dense tomato detection; for instance, Appe et al. (2023a) utilised CAM-YOLO, while Appe et al. (2023b) integrated coordinate attention into YOLO architectures to improve dense fruit classification. However, standard attention modules can sometimes suffer from gradient degradation during deep feature extraction. Our integration of ResCBAM specifically addresses this by combining spatial and channel attention with a residual connection. This design works synergistically with YOLOv9's Programmable Gradient Information (PGI) to preserve baseline feature fidelity while actively mitigating the gradient vanishing issues seen in earlier attention-based YOLO variants, making it uniquely suited for heavily occluded greenhouse environments.

These studies collectively demonstrate the efficacy of YOLO-based models for crop ripeness detection, revealing areas for improvement, particularly in greenhouse environments with occlusions, variable lighting, and dense plant distributions. While attention mechanisms have shown promise in computer vision, their potential in YOLO-based models for agriculture remains underexplored. This motivates our proposal of YOLOv9-ResCBAM, which integrates attention mechanisms into YOLOv9 to address these challenges.

Applying YOLOv9-ResCBAM in greenhouses, especially for tomato detection, presents unique challenges due to controlled but varied conditions. This study explores how YOLOv9, enhanced by attention mechanisms, improves detection accuracy and adaptability in diverse agricultural scenarios. It aims to demonstrate the potential of cutting-edge computer vision technologies to enhance crop management and production in precision agriculture. The primary contributions of this study are:

1. A novel enhancement to the YOLOv9 architecture by integrating a Residual Convolutional Block Attention Module (ResCBAM), improving detection accuracy in complex agricultural environments.
2. Improved classification of tomato ripeness across varying stages, addressing challenges like occlusions, dense foliage, and lighting variations commonly found in greenhouses.
3. Demonstrated performance through rigorous evaluation, achieving an mAP of 0.912, outperforming earlier YOLO versions and traditional object detection models like SSD, Faster R-CNN, and Mask R-CNN.
4. Established an advancement in agricultural object detection, with potential for widespread application across different crops and environments.

MATERIALS AND METHODS

This study introduces our proposed YOLOv9- ResCBAM model, which integrates a Residual Convolutional Block Attention Module (ResCBAM) into the YOLOv9 architecture, for tomato ripeness detection in greenhouses. Our workflow, illustrated in Figure 1,

begins with dataset preparation, crucial for training and evaluating all models. The dataset undergoes preprocessing and augmentation stages, including label extraction, annotation conversion to YOLO format, auto-orientation, resizing, class label simplification, and filtering. We employed Roboflow for data augmentation, applying techniques such as random flips, rotations, shears, blurring, and noise addition to simulate diverse greenhouse conditions.

The processed data was split into training (92%), validation (4%), and test (4%) sets, ensuring comprehensive model evaluation. The proposed YOLOv9-ResCBAM and other object detection models were trained on this prepared dataset and assessed for their effectiveness in detecting tomatoes at various ripeness stages under diverse agricultural conditions.

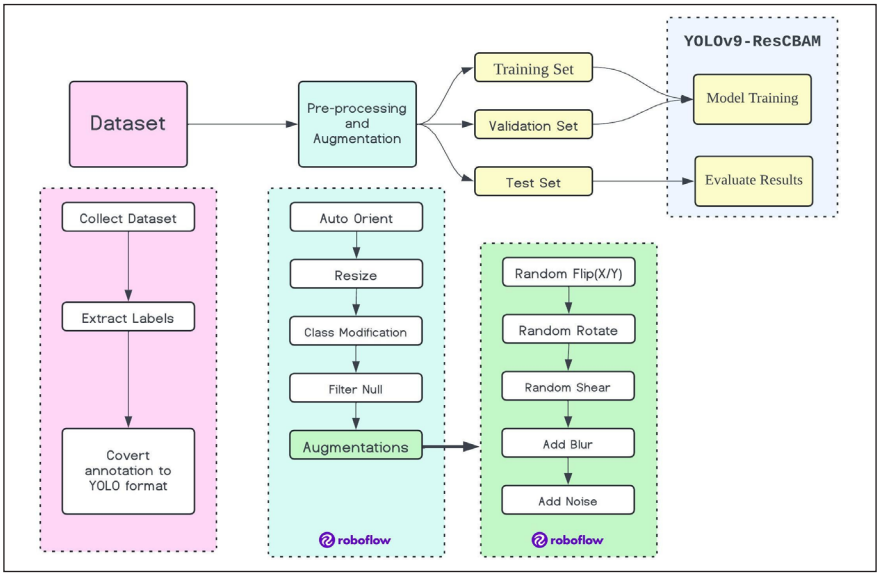


Figure 1. Workflow diagram for YOLOv9-ResCBAM model training

Proposed Method: YOLOv9-ResCBAM

YOLOv9-ResCBAM is an advanced object detection architecture that builds upon the YOLOv9 framework by integrating the Residual Convolutional Block Attention Module (ResCBAM). Figure 2 depicts the proposed architecture. This integration enhances the model's ability to focus on critical parts of the input image, thereby improving detection accuracy. The primary innovation in this model is the incorporation of ResCBAM, which combines channel and spatial attention mechanisms with residual learning to refine feature maps at multiple stages within the network.

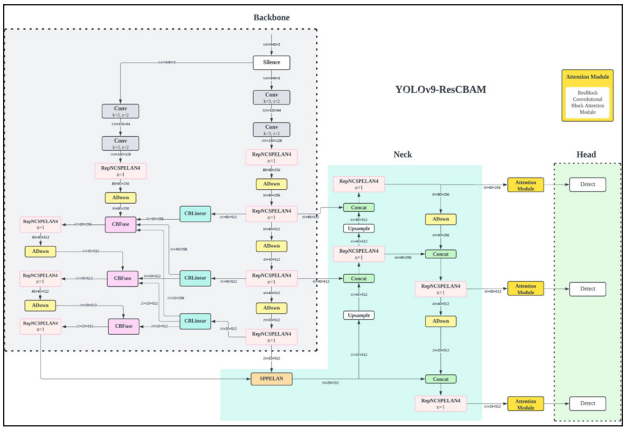


Figure 2. YOLOv9-ResCBAM architecture diagram

Attention Module

In deep learning models, attention mechanisms play an essential role, particularly in object detection, by enabling the model to concentrate on the most significant parts of the input. The use of these mechanisms has recently achieved notable results in the object detection field (Guo et al., 2022; Yao et al., 2022; Zamri et al., 2024). YOLOv9-ResCBAM employs the Convolutional Block Attention Module (CBAM), which improves the model's performance by prioritising significant features and lowering the influence of less important ones. Figure 3 demonstrates that CBAM comprises two sequential sub-modules: the Channel Attention Module (C-Attention) and the Spatial Attention Module (S-Attention).

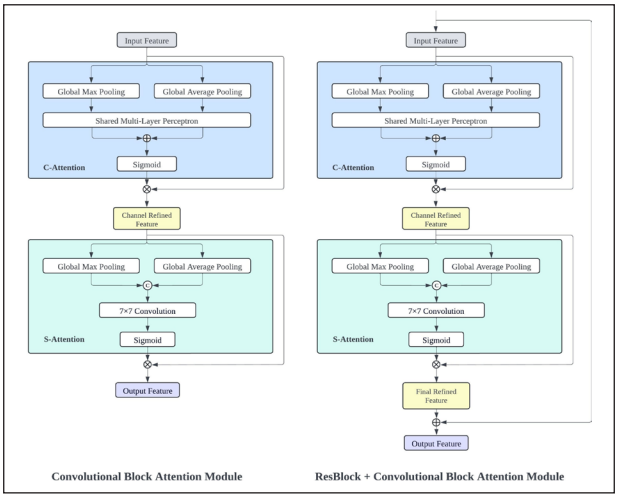


Figure 3. Detailed representation of the Convolutional Block Attention Module (CBAM). The left side depicts the standard CBAM architecture, while the right side illustrates the modified architecture with a residual connection, known as ResCBAM

Channel Attention (C-Attention)

The primary function of the Channel Attention mechanism is to focus on the inter-channel relationships of the feature maps. It allows the network to prioritise the channels that contain the most informative features. Given an input feature map $F \in \mathbb{R}^{C \times H \times W}$, where C, H, and W represent the number of channels, height, and width of the feature map, respectively, the channel attention mechanism utilises two global pooling operations: Global Average Pooling (GAP) and Global Max Pooling (GMP). GAP aggregates spatial information into a channel descriptor by averaging each channel's spatial dimension, while GMP captures the most salient features across the spatial dimensions by selecting the maximum value within each channel. These pooled descriptors are then passed through a shared Multi-Layer Perceptron (MLP), which consists of a single hidden layer to generate the channel attention map M_C, formally defined in Equation 1:

$$M_C(F) = \sigma(MLP(GAP(F)) + MLP(GMP(F))) \quad [1]$$

Here, σ is the sigmoid activation function. This attention map is then applied to the original feature map via element-wise multiplication to generate the channel-refined feature map F_C according to Equation 2:

$$F_C = M_C(F) \otimes F \quad [2]$$

Where $\otimes$ denotes the element-wise multiplication. This operation ensures that the channels with higher attention values contribute more significantly to the next layers.

Spatial Attention (S-Attention)

Channel attention and spatial attention serve distinct but complementary roles; while channel attention is concerned with determining 'what' features are significant, spatial attention identifies 'where' those important features are located on the feature map. The Spatial Attention Module in CBAM refines the feature map by emphasising crucial regions within the feature map. Given the channel-refined feature map F_C, the spatial attention is computed as follows:

The spatial attention mechanism first applies Global Average Pooling (GAP) and Global Max Pooling (GMP) across the channel axis, resulting in two 2D maps, F_C^{avg} and F_C^{max}, both of size $H \times W$.

Subsequently, the two maps are combined through channel-wise concatenation and then fed into a convolutional layer that features a 7 × 7 kernel. This operation produces the spatial attention map M_S, as expressed in Equation 3:

$$M_S(F_C) = \sigma\left(f^{7\times7}([GAP(F_C); GMP(F_C)])\right) \quad [3]$$

Here, $f^{7\times7}$ represents the convolution operation utilising a 7 × 7 filter, while σ denotes sigmoid function that scales the output to the range [0, 1]. This spatial attention map M_S highlights the regions within the feature map that are most informative for the task at hand.

The final spatially refined feature map F_S is then obtained by applying the spatial attention map to the channel-refined feature map, formally defined in Equation 4:

$$F_S = M_S(F_C) \otimes F_C \quad [4]$$

This operation directs the network to prioritise the most salient essential locations within the feature map.

Convolutional Block Attention Module (CBAM)

CBAM combines the Channel and Spatial Attention mechanisms sequentially to refine the feature maps both in the channel and spatial domains. The following equation summarises the complete operation of the CBAM module according to Equation 5:

$$F_{CBAM} = M_S(M_C(F) \otimes F) \otimes (M_C(F) \otimes F) \quad [5]$$

Where F_{CBAM} is the output feature map that has been refined by both channel and spatial attention mechanisms. By applying both types of attention, CBAM effectively filters out irrelevant features while amplifying important ones.

Residual Connection in ResCBAM

To further enhance the performance and stabilise the learning process, a residual connection is incorporated into CBAM, forming the Residual Convolutional Block Attention Module (ResCBAM). The residual connection is formulated as Equation 6:

$$F_{output} = F + F_{CBAM} \quad [6]$$

This equation implies that the input features are added directly to the output of the CBAM module. This residual connection helps in preserving the original information from the feature map while adding the benefits of attention, which prevents overfitting and improves generalisation.

Integration in YOLOv9

We strategically place ResCBAM modules at key points in the YOLOv9 architecture, particularly in the neck of the network, as shown in Figure 2. This specific insertion depth was selected because it represents the junction where multi-scale feature maps are richest in both spatial resolution and semantic depth, allowing the attention mechanism to maximise feature refinement immediately before the detection heads. This placement allows the attention mechanism to refine features after they have been processed by the backbone and before they are used for detection. The ResCBAM module operates on the output of each RepNCSPELAN4 block in the neck, enhancing the feature maps before they are concatenated or used for detection. The enhanced feature map F' after applying ResCBAM is given by Equation 7 :

$$F' = ResCBAM(F) = F + M_s(M_c(F) \otimes F) \otimes (M_c(F) \otimes F) \quad [7]$$

where F is the output feature map from the RepNCSPELAN4 block.

Integrating ResCBAM into YOLOv9 enhances feature refinement, enabling improved ripeness detection through adaptive, multi-scale attention, better gradient flow, and spatial awareness under varied greenhouse conditions and occlusions.

Data Acquisition

The effectiveness of computer vision systems in agricultural settings is heavily dependent on the availability of diverse and high-quality datasets that accurately represent the varied conditions typical of real-world scenarios. For our research, we have selected the Laboro Tomato dataset (https://github.com/laboroai/LaboroTomato) as the data source. This dataset is particularly valuable because it encompasses a comprehensive range of tomatoes at different ripening stages, reflecting the inherent diversity found within a typical agricultural environment. The Laboro Tomato dataset includes images captured using two different cameras, each varying in resolution and image quality. This dual-source approach enriches the dataset with a variety of imaging conditions that closely mimic the variability encountered in greenhouse operations. The dataset categorises the images of tomatoes into three ripening stages. These stages are visually represented in Figure 4 as: fully ripened, half-ripened, and green. The fully ripened tomatoes display over 90% red coloration, half- ripened tomatoes exhibit between 30-89% red coloration, and green tomatoes show 0-30% red coloration, with the percentages providing a rough guide due to natural variability. This stratification is critical for our computer vision model, which needs to accurately identify and classify each stage of tomato maturity to support timely and precise agricultural decision-making.

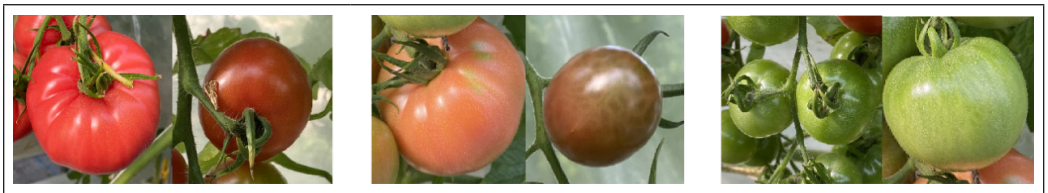


Figure 4. Dataset class

Data Pre-processing and Preparation

Upon acquiring the raw images from the Laboro Tomato dataset, a rigorous preprocessing journey was undertaken to tailor the data for our development pipeline. This process began with a meticulous image selection phase, where images like sample A and sample B, as seen in Figure 5, were reviewed for quality and relevance. Subsequently, a series of rigorous preprocessing steps was applied to the Laboro Tomato dataset to optimise it for our models. Key steps included:

1. Auto Orient : Images were realigned using EXIF orientation metadata to ensure consistency.
2. Resize : All images were uniformly resized to 640 × 640 pixels, balancing detail preservation with computational efficiency.
3. Class Modification : We simplified the classification system by focusing solely on ripeness stages, removing size distinctions between tomatoes. Table 1 details the class remapping. By removing size distinctions, we deliberately forced the model to learn generalised colour-based ripeness cues across different phenotypes, rather than risking the model falsely associating 'large' strictly with 'mature,' which would cause failures on naturally smaller tomato cultivars. Furthermore, the discrete three-class ripeness system was chosen over continuous regression because agricultural robotics and harvest planning require strict, actionable decision boundaries. During annotation, the 30-89% coloration rule served as a strict guideline. To mitigate subjectivity, ambiguous borderline fruits were systematically downgraded to the less-mature category to prevent costly premature harvesting errors.
4. Filter Null : Images lacking annotations were removed to maintain dataset integrity.

5. Dataset Split : The dataset was split into training (92%, 2001 images), validation (4%, 86 images), and test (4%, 87 images) sets. The data split was performed randomly at the image level. While the validation and test partitions represent 4% each, the risk of overfitting or data leakage was severely mitigated by the aggressive augmentation pipeline (e.g., severe shearing, up to 1.5px blur, and noise). This mathematically forced the model to learn invariant biological features rather than memorising spatial clusters or camera-specific sensor artefacts. The success of this regularisation is empirically proven in Figure 10, where validation losses tightly track training losses without divergence.

6. Data Augmentation : We applied various techniques using Roboflow. As depicted in Figure 6, a variety of data augmentation methods were applied to the dataset, including:

- Random flips (X/Y axis)
- Rotations (-15° to $+15^{\circ}$)
- Shear transformations ($\pm 10^{\circ}$)
- Blur effect (up to 1.5 pixels)
- Noise addition (up to 1.56% of pixels)

The augmentation parameters were empirically designed to simulate physical greenhouse challenges. For example, blur (up to 1.5 pixels) and noise (up to 1.56%) mimic lens dust and low-light sensor degradation. Furthermore, while natural spatial relationships exist, clustered foliage is highly chaotic. Utilising synthetic mixup (0.15) and copy-paste (0.3) operations deliberately disrupts predictable background contexts, forcing the network to learn the invariant biological features of the tomatoes rather than relying on spatial memorisation. These preprocessing and augmentation steps ensured a robust, diverse dataset that simulates real-world greenhouse conditions, crucial for training our proposed YOLOv9-ResCBAM model.

Sample A Sample B

Figure 5. Original dataset sample

Table 1
Remapping of original classes to simplified classes in the dataset

Original Class	Override	Included
b_fully_ripened	fully_ripened	Yes
b_green	green	Yes
b_half_ripened	half_ripened	Yes
l_fully_ripened	fully_ripened	Yes
l_green	green	Yes
l_half_ripened	half_ripened	Yes

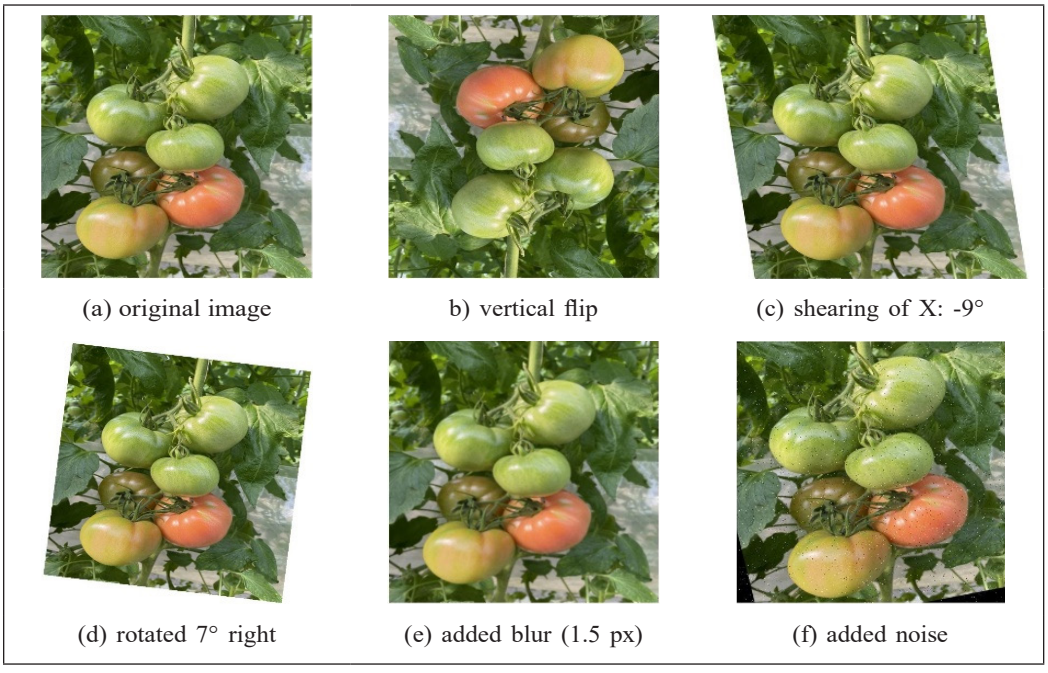
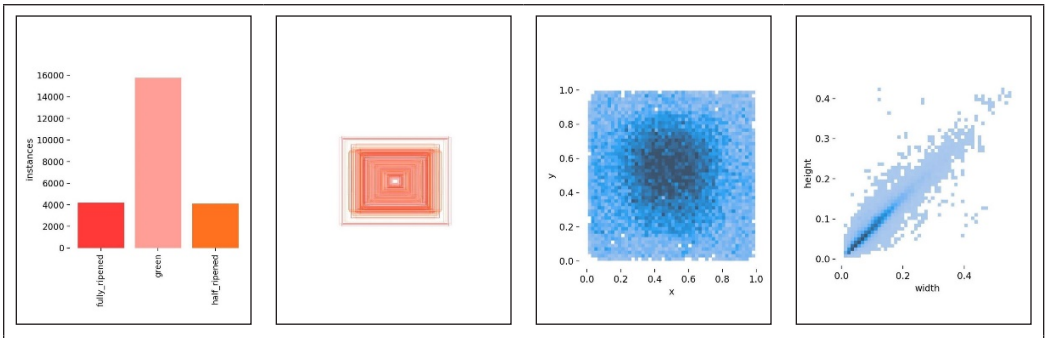


Figure 6. Data augmentation pipeline applied to the dataset; (a) original image; (b) vertical flip; (c) shearing of X: -9°; (d) rotated 7° right; (e) added blur (1.5 px); (f) added noise

Through these meticulous preprocessing steps, the dataset was not only standardised but also enriched, ensuring that it was optimally prepared for training a robust and effective model. Each technique plays a vital role in mirroring real-world variability and challenges, equipping the model with the capability to accurately detect and classify tomatoes across different stages of ripeness under diverse operational conditions in smart greenhouses. Figure 7 illustrates the statistical analysis of the dataset.



a) Number of instances | b) Overall distribution of all annotated bounding boxes | c) Distribution of centre coordinates of all annotated bounding boxes | d) Distribution of size of all annotated bounding boxes

Figure 7. Statistical visualisation of the dataset

Experimental Setup

The experimental hardware configuration included the use of an NVIDIA GeForce RTX 3060 GPU with 16 GB of VRAM, supported by an AMD Ryzen 5 processor with six cores. The software environment was set up on an Ubuntu 24.04 LTS operating system, with all deep learning models trained and evaluated using PyTorch version 2.3.0. The experimental setup was designed to thoroughly evaluate various iterations of the object detection models, and our proposed YOLOv9-ResCBAM. The training was performed with an image size of 640 and a batch size of 8. Specific training parameters were configured to train the YOLO models. Table 2 lists detailed hyperparameters used in the experiments. To ensure absolute fairness and rule out optimisation bias, all comparative baseline models (YOLOv3-v8, SSD, and R-CNN variants) were trained from scratch using the same hardware, hyperparameter configurations (100 epochs, batch size 8), and augmentation pipelines. While a granular empirical ablation study separating the individual effects of YOLOv9, standard CBAM, and the residual connection is slated for future work, the theoretical design of ResCBAM where the residual connection preserves the foundational PGI features is heavily validated by the aggregate performance leap.

Table 2
Detailed training parameters configured for model training

Model Parameter	Value	Parameter Explanation
epoch	100	Total number of epochs for training
lr0	0.01	Initial learning rate, specific to the optimisation method used
lrf	0.01	Final learning rate as a fraction of the initial rate
momentum	0.937	Momentum factor for SGD/1 for Adam
weight decay	0.0005	Weight decay regularisation factor
warmup epochs	3.0	Number of epochs for gradual learning rate increase

Table 2 (*continued*)

Model Parameter	Value	Parameter Explanation
warmup momentum	0.8	Initial momentum during the warm-up phase
warmup bias lr	0.1	Learning rate for bias during warm-up
mixup	0.15	Probability of performing mixup image data augmentation
copy paste	0.3	Probability of using segment copy-paste augmentation

Evaluation Metrics

The evaluation was based on several key metrics, including precision, recall, mean Average Precision (mAP), F1-Score, and frames per second (FPS).

Precision

Precision is defined as the proportion of correctly identified positive samples among all detected positives. It is mathematically expressed as Equation 8:

$$Precision = \frac{TP}{TP+FP} \quad [8]$$

In this context, TP stands for true positives, which are the instances correctly identified as positive, while FP refers to false positive instances that are incorrectly labelled as positive. A higher precision score signifies a more dependable model, as it means the model makes fewer false-positive detections.

Recall

Recall represents the ratio of correctly identified positive samples to the total number of actual positive samples. This metric measures the model's capacity to detect all relevant instances in a dataset. It is mathematically expressed as Equation 9:

$$Recall = \frac{TP}{TP+FN} \quad [9]$$

Here, true positives (TP) refer to the positive instances that were correctly identified by the model, while false negatives (FN) represents the actual positive cases the model overlooked. A high recall value, therefore, suggests that a smaller number of relevant instances were missed, which leads to a more complete detection performance.

mAP (mean Average Precision)

mAP, calculated at an IoU threshold of 0.5, represents the average of the Average Precision (AP) for all classes. It is a standard measure for assessing object detection algorithms, derived from the area under the precision-recall curve.

FPS (Frames Per Second)

This measures the number of images the model can process per second at a batch size of 1, providing an indication of the model's detection speed. These metrics are essential for quantifying the effectiveness and efficiency of each YOLO model iteration in handling real-world object detection tasks.

RESULTS AND DISCUSSION

In this section, the performance outcomes of the YOLOv9-ResCBAM model are presented, focusing on its effectiveness in detecting and classifying tomatoes at various ripeness stages in greenhouse environments.

Performance of YOLOv9-ResCBAM

The YOLOv9-ResCBAM model was evaluated on a dataset of tomatoes in greenhouse environments, which included fully ripened, half-ripened, and unripe tomatoes. The model performance was measured using precision, recall, and mAP at an Intersection over Union (IoU) threshold of 0.5. These metrics are critical for understanding the model's ability to detect and classify tomatoes accurately across different ripeness stages.

The proposed YOLOv9-ResCBAM shows exceptional performance, with an mAP@0.5 of 0.9124, particularly in maintaining high precision and recall across all categories of tomato ripeness. The fully ripened category achieved a precision of 0.936, while green and half-ripened categories attained precisions of 0.904 and 0.897, respectively, as shown in Figure 8. This consistent performance across all ripeness stages underscores the effectiveness of the integrated Residual Convolutional Block Attention Module in enhancing feature extraction and detection accuracy. Furthermore, under stricter localisation criteria, the model demonstrated strong robustness, achieving an mAP@0.5: 0.95 of approximately 0.80, as evidenced by the metric curves in Figure 10.

As indicated by the dataset statistics, Figure 7a, there is a natural class imbalance heavily skewed toward 'green' instances. However, the model maintained exceptional precision across all classes (Fully Ripened: 0.936, Green: 0.904, Half Ripened: 0.897). This demonstrates that the architectural design and augmentation pipeline successfully prevented the model from becoming biased toward the majority class.

The analysis of the precision-recall curve reveals that the YOLOv9- ResCBAM model significantly enhances the detection and classification capabilities of tomatoes in greenhouse settings. Its ability to maintain high precision and recall across varying confidence thresholds and ripeness stages makes it highly reliable for practical applications. The performance of the proposed model is further confirmed by the confusion matrix in Figure 9, which shows high accuracy in classifying tomatoes across various ripeness stages. These results underscore the model's reliability and the effective integration of attention mechanisms. An analysis of the confusion matrix reveals that the model's primary failure mode is confusing 'green' instances with the 'background.' This is biologically expected, as the visual characteristics of unripe green tomatoes share nearly identical colour and texture profiles with the dense surrounding greenhouse foliage and stems. In the context of robotic harvesting, false positives on fully ripened fruit are highly costly, as picking unripe fruit reduces market value. The model's exceptional precision of 0.936 for the fully ripened category directly minimises this risk, providing a highly reliable operational threshold for automated picking systems.

Additionally, Figure 10 presents different metrics, including the training and validation loss curves for bounding box regression, classification, and the overall loss for the proposed YOLOv9-ResCBAM model. The steady decline in these loss values over the training epochs indicates effective learning and convergence of the model. The validation losses closely follow the training losses, suggesting that the models generalise well to unseen data. The YOLOv9-ResCBAM model shows a consistent and sharp decrease in all loss metrics, which aligns with its superior performance observed in the precision-recall analysis. This comprehensive evaluation underscores the robustness and reliability of the YOLOv9-ResCBAM model, making it an excellent choice for precision agricultural tasks.

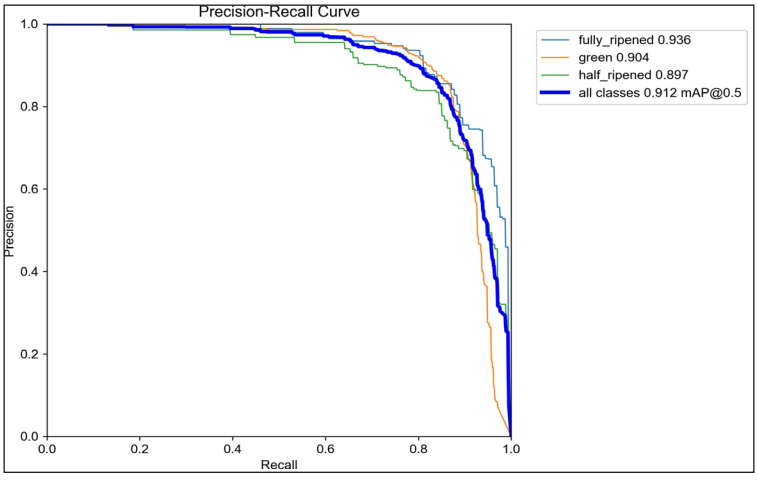


Figure 8. Precision-Recall curve of proposed YOLOv9- ResCBAM model

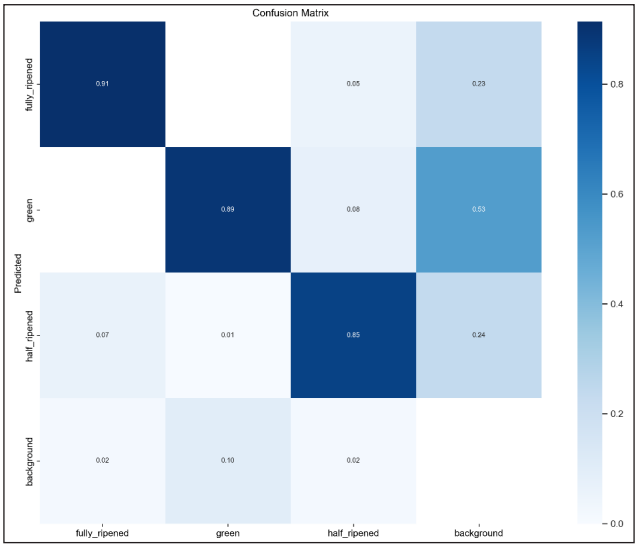


Figure 9. Confusion matrix for YOLOv9-ResCBAM's tomato ripeness classification

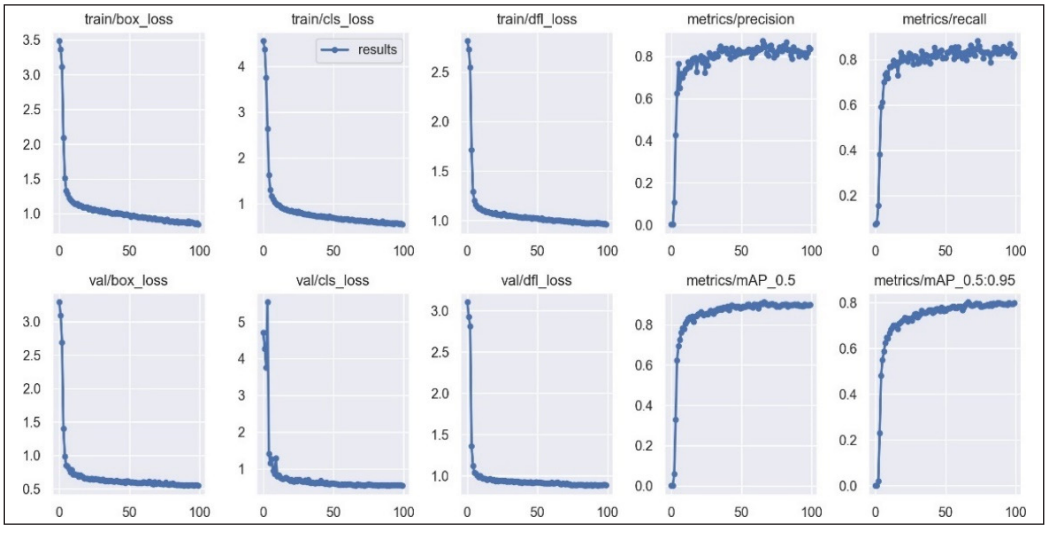


Figure 10. Visualisation of different metrics of the proposed YOLOv9-ResCBAM Model over epochs

DISCUSSION

Accurate detection and classification of crop ripeness stages are pivotal for informed decision-making and optimised yield management in precision agriculture. Through rigorous experimentation, we assessed our proposed YOLOv9-ResCBAM and compared it with previous YOLO iterations and traditional object detection models such as SSD, Faster R-CNN, and Mask R-CNN.

Our findings demonstrate that the proposed YOLOv9-ResCBAM model outperforms all previous models in object detection and classification for agricultural applications. While this study focuses exhaustively on CNN-based, one-stage detectors (such as the YOLO family, SSD, and R-CNN variants), we acknowledge the rapid advancements in Transformer-based detectors (e.g., DETR) for complex computer vision tasks. However, Transformer models currently introduce significant computational overhead and larger memory footprints, which pose substantial challenges for real-time inference on resource-constrained agricultural edge devices. Therefore, CNN-based architectures were prioritised for this benchmarking to ensure practical deployment feasibility, while Transformer-based comparisons remain a highly valuable avenue for future research.

The exceptional performance of YOLOv9-ResCBAM can be attributed to its novel architectural advancements, primarily the integration of the Residual Convolutional Block Attention Module (ResCBAM) into the YOLOv9 framework. This innovation addresses fundamental challenges in deep learning, such as information loss and biased gradient propagation, which have historically impeded optimal performance in complex detection tasks. By incorporating attention mechanisms, ResCBAM ensures that critical information is effectively utilised during the learning process, resulting in more reliable and robust feature representations. This enhancement is particularly beneficial in scenarios where subtle differences between various stages of crop maturity or health need to be accurately recognised and classified. The comparative analysis presented in Table 3 highlights several key insights. YOLOv9-ResCBAM achieved the highest mean average precision (mAP) of 0.912 while maintaining a reasonable Frames Per Second (FPS) rate of 38 on an NVIDIA GeForce RTX 3060 GPU.

As shown in Table 3, while a +0.011 mAP gain over the YOLOv9-c baseline may appear numerically modest, in an automated agricultural context processing thousands of frames, this translates to a tangible reduction in false positives, directly protecting crop yield by preventing the accidental harvesting of unmarketable fruit. Furthermore, while this study focuses on static frame detection, establishing this highly robust frame-level accuracy is the critical prerequisite for future integration with temporal tracking algorithms (such as DeepSORT) across video streams.

This balance between accuracy and speed is crucial for real-time detection applications in greenhouse environments, where timely interventions can significantly impact crop yield and quality. Despite introducing additional computational complexity through its attention mechanisms, YOLOv9-ResCBAM maintains a moderate model size of 117 MB and a GFLOPs (Giga Floating-point Operations per Second) value of 70. This indicates that the model is computationally efficient relative to its high accuracy, making it suitable for deployment on devices with limited processing capabilities.

Table 3
Performance comparison of various models in terms of several key metrics. The FPS values were obtained using an NVIDIA GeForce RTX 3060 GPU

Model	mAP@0.5	Precision	Recall	F1-Score	FPS	Size (MB)	GFLOPs
YOLOv3	0.841	0.869	0.775	0.797	28	118	65
YOLOv4	0.847	0.854	0.802	0.813	25	245	90
YOLOv5	0.892	0.886	0.860	0.845	53	91.4	16.5
YOLOv6	0.876	0.850	0.838	0.828	50	114	18
YOLOv8	0.842	0.834	0.799	0.792	56	83.7	15
YOLO NAS	0.894	0.885	0.864	0.839	20	328	150
YOLOv9-e	0.896	0.876	0.855	0.841	38	115	60
YOLOv9-c	0.901	0.888	0.864	0.847	35	119	60
SSD	0.850	0.851	0.834	0.840	25	103	60
Faster R-CNN	0.880	0.872	0.863	0.841	12	198	120
Mask R-CNN	0.900	0.882	0.875	0.840	10	250	130
YOLOv9-ResCBAM	0.912	0.894	0.882	0.849	38	117	70

Compared to traditional models, YOLOv9-ResCBAM demonstrates superior practicality. SSD, with a mAP of 0.850 and FPS of 25, and Faster R-CNN and Mask R-CNN, with higher mAPs but significantly lower FPS (12 and 10, respectively), are less suited for real-time, resource-constrained environments. The GFLOPs metric further highlights YOLOv9-ResCBAM's optimal balance of computational demand and performance, critical for agricultural settings where hardware resources and energy efficiency are significant considerations. The model's moderate computational requirements facilitate deployment on edge devices common in agriculture, ensuring integration without substantial hardware investments. Although physical latency testing on specific edge hardware, such as the NVIDIA Jetson Nano or Raspberry Pi was outside the immediate scope of this algorithmic study, the theoretical metrics strongly support real-world edge deployment. The proposed model's computational demand of 70 GFLOPs and moderate size of 117 MB fall well within the operational envelopes of modern agricultural edge computing devices, ensuring that high accuracy can be maintained without inducing severe thermal throttling or requiring prohibitive hardware investments. Its adaptability allows customisation for various platforms, enhancing scalability across different agricultural operations. It is important to note that domain shifts across different seasons, variations in farmer pruning styles, and external validation across completely unseen greenhouse datasets remain significant challenges. While our aggressive augmentation pipeline acted as a proxy to simulate environmental shifts, formal cross-dataset validation is a primary focus for our future field trials.

Figure 11 illustrates different challenging situations successfully detected by the YOLOv9-ResCBAM model. The images highlight the model's ability to accurately identify tomatoes under various conditions, such as occlusion by leaves and fruits, dense clustering, and varying lighting conditions. These examples underscore the robustness of the YOLOv9-ResCBAM model in real-world agricultural environments, demonstrating the effectiveness of its integrated attention mechanisms in handling complex scenarios.

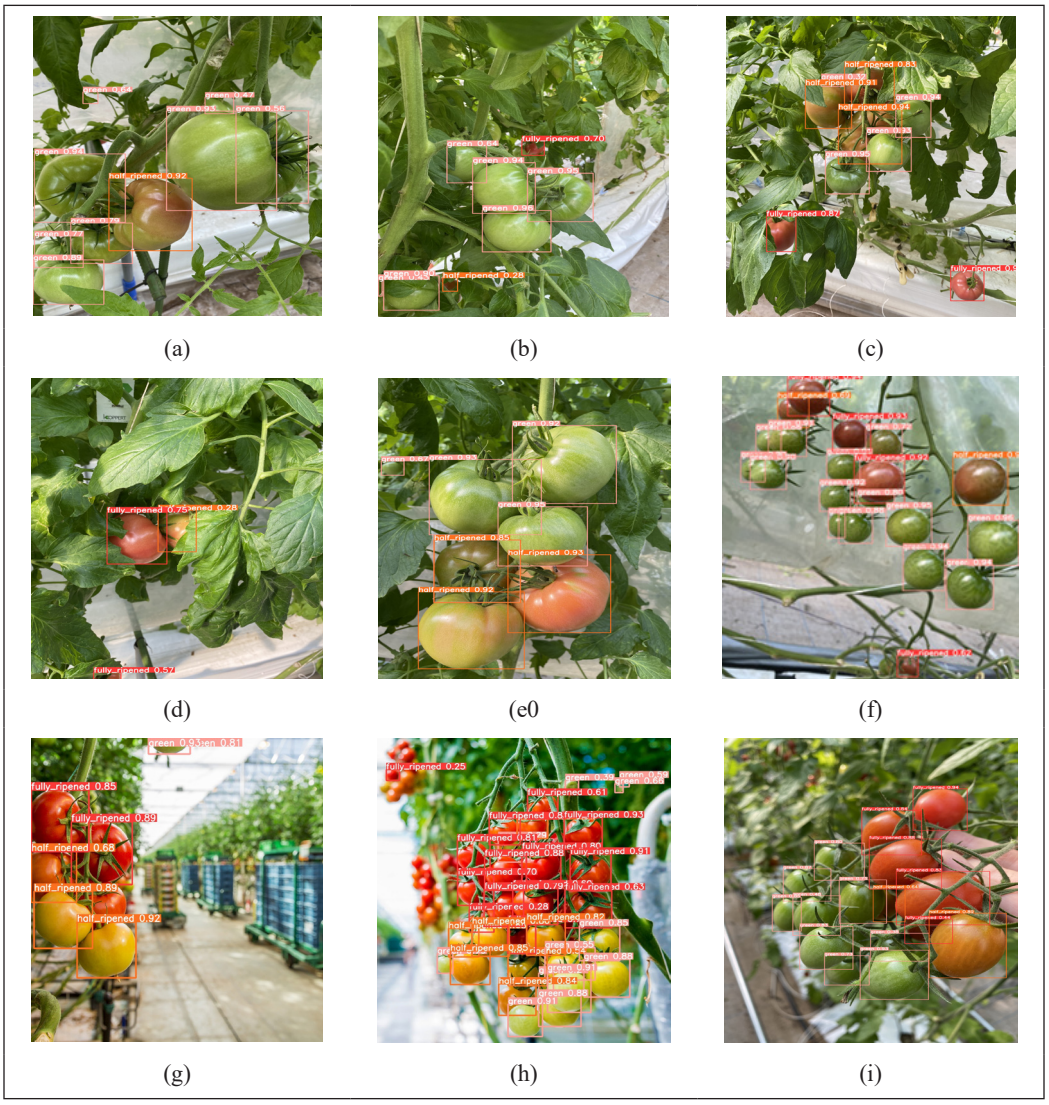


Figure 11. YOLOv9-ResCBAM's robust detection in diverse greenhouse conditions. Examples of tomato fruit detection challenges under different field conditions: (a) fruits occluded by other fruits; (b) fruits occluded by foliage; (c) combined occlusion by foliage and fruits; (d) detection amidst dense foliage; (e) identifying tomato clusters; (f) clustered fruit detection; (g) detection in high-density fruit environments; (h) accurate detection in densely packed fruits; and (i) high-density fruit clustering detection

The seamless integration of cutting-edge computer vision techniques, as exemplified by the YOLOv9-ResCBAM architecture, with domain-specific agricultural knowledge has paved the way for transformative AI-driven solutions in crop management and yield optimisation. Our findings reinforce the potential of the YOLOv9 series, particularly the YOLOv9-ResCBAM model, to revolutionise precision agriculture practices, enabling more efficient resource allocation, timely interventions, and improved productivity and sustainability in greenhouse farming operations.

Furthermore, extending this system to outdoor agricultural settings or multimodal configurations (such as RGB-NIR) presents distinct challenges that warrant future investigation. Outdoor environments introduce highly dynamic illumination, harsh shadows, and complex backgrounds (e.g., sky and soil) that require more aggressive domain adaptation strategies. While integrating Near-Infrared (NIR) imaging could mitigate some lighting variations, it introduces challenges regarding strict multispectral sensor alignment, increased computational load, and higher implementation costs for farmers.

CONCLUSION

This study introduced YOLOv9-ResCBAM, an enhanced object detection model designed for tomato ripeness classification in greenhouse environments. Achieving a mean Average Precision (mAP@0.5) of 0.912, YOLOv9-ResCBAM outperforms earlier YOLO versions and traditional object detection frameworks such as SSD, Faster R-CNN, and Mask R-CNN. Key innovations, including Programmable Gradient Information (PGI), Generalised Efficient Layer Aggregation Network (GELAN), and the Residual Convolutional Block Attention Module (ResCBAM), optimise feature extraction, gradient propagation, and computational efficiency, enabling the model to perform well in challenging greenhouse scenarios. This advancement contributes significantly to agricultural object detection, offering the potential for AI-driven solutions to improve crop management. Future work will involve deploying YOLOv9-ResCBAM in real-world greenhouse scenarios, validating its practical effectiveness, and exploring its applicability to other crops, thereby advancing the broader field of precision agriculture.

ACKNOWLEDGEMENT

We want to extend our heartfelt gratitude to the UMPSA Postgraduate Research Grants Scheme (PGRS) for providing the Open Access funding that supported this research.

REFERENCES

Akbar, J. U. M., Kamarulzaman, S. F., Muzahid, A. J. M., Rahman, M. A., & Uddin, M. (2024). A comprehensive review of deep learning-assisted computer vision techniques for smart greenhouse agriculture. *IEEE Access, 12*, 4485-4522. https://doi.org/10.1109/ACCESS.2024.3349418

Akbar, J. U. M., Kamarulzaman, S. F., & Tusher, E. H. (2023). Plant stem disease detection using machine learning approaches. *2023 14th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1-8. https://doi.org/10.1109/ICCCNT56998.2023.10307074

Appe, S. N., Arulselvi, G., & Balamurugan, G. N. (2023a). CAM-YOLO: Tomato detection and classification based on improved YOLOv5 using combining attention mechanism. *PeerJ Computer Science, 9*, Article e1463. https://doi.org/10.7717/peerj-cs.1463

Appe, S. N., Arulselvi, G., & Balamurugan, G. N. (2023b). Detection and classification of dense tomato fruits by integrating coordinate attention mechanism with the YOLO model. In *Advances in computational intelligence and robotics book series* (pp. 278-289). IGI Global. https://doi.org/10.4018/978-1-6684-8098-4.ch016

Badeka, E., Karapatzak, E., Karampatea, A., Bouloumpasi, E., Kalathas, I., Lytridis, C., Tziolas, E., Tsakalidou, V. N., & Kaburlasos, V. G. (2023). A deep learning approach for precision viticulture, assessing grape maturity via YOLOv7. *Sensors, 23*(19), Article 8126. https://doi.org/10.3390/s23198126

Brodt, S., Kramer, K. J., Kendall, A., & Feenstra, G. (2013). Comparing environmental impacts of regional and national-scale food supply chains: A case study of processed tomatoes. *Food Policy, 42*, 106-114. https://doi.org/10.1016/j.foodpol.2013.07.004

Chen, W., Liu, M., Zhao, C., Li, X., & Wang, Y. (2024). MTD-YOLO: Multi-task deep convolutional neural network for cherry tomato fruit bunch maturity detection. *Computers and Electronics in Agriculture, 216*, Article 108533. https://doi.org/10.1016/j.compag.2023.108533

Chien, C., Ju, R., Chou, K., Xieerke, E., & Chiang, J. (2025). YOLOv8-AM: YOLOv8 based on effective attention mechanisms for paediatric wrist fracture detection. *IEEE Access, 13*, 44118-44133. https://doi.org/10.1109/ACCESS.2025.3549839

Cui, X., Guan, Z., Morgan, K., Huang, K., & Hammami, A. (2022). Multitiered fresh produce supply chain: The case of tomatoes. *Horticulturae, 8*(12), Article 1204. https://doi.org/10.3390/horticulturae8121204

Guo, M., Xu, T., Liu, J., Liu, Z., Jiang, P., Mu, T., Zhang, S., Martin, R. R., Cheng, M., & Hu, S. (2022). Attention mechanisms in computer vision: A survey. *Computational Visual Media, 8*(3), 331-368. https://doi.org/10.1007/s41095-022-0271-y

Li, P., Zheng, J., Li, P., Long, H., Li, M., & Gao, L. (2023). Tomato maturity detection and counting model based on MHSA-YOLOv8. *Sensors, 23*(15), Article 6701. https://doi.org/10.3390/s23156701

Li, R., Ji, Z., Hu, S., Huang, X., Yang, J., & Li, W. (2023). Tomato maturity recognition model based on improved YOLOv5 in greenhouse. *Agronomy, 13*(2), Article 603. https://doi.org/10.3390/agronomy13020603

Moreira, G., Magalhães, S. A., Pinho, T., Santos, F. N. D., & Cunha, M. (2022). Benchmark of deep learning and a proposed HSV colour space model for the detection and classification of greenhouse tomato. *Agronomy, 12*(2), Article 356. https://doi.org/10.3390/agronomy12020356

Phoophuangpairoj, R., Ngoenrungrueang, T., & Audomsin, S. (2023). Ripeness classification of a bunch of bananas using a CNN. *2022 International Electrical Engineering Congress (iEECON)*, 180-183. https://doi.org/10.1109/iEECON56657.2023.10126585

Rizzo, M., Marcuzzo, M., Zangari, A., Gasparetto, A., & Albarelli, A. (2023). Fruit ripeness classification: A survey. *Artificial Intelligence in Agriculture, 7*, 44-57. https://doi.org/10.1016/j.aiia.2023.02.004

Su, F., Zhao, Y., Wang, G., Liu, P., Yan, Y., & Zu, L. (2022). Tomato maturity classification based on SE-YOLOv3-MobileNetV1 network under natural greenhouse environment. *Agronomy, 12*(7), Article 1638. https://doi.org/10.3390/agronomy12071638

Tang, C., Chen, D., Wang, X., Ni, X., Liu, Y., Liu, Y., Mao, X., & Wang, S. (2023). A fine recognition method of strawberry ripeness combining Mask R-CNN and region segmentation. *Frontiers in Plant Science, 14*, Article 1211830. https://doi.org/10.3389/fpls.2023.1211830

Wang, C., Han, Q., Li, J., Li, C., & Zou, X. (2024). YOLO-BLBE: A novel model for identifying blueberry fruits with different maturities using the I-MSRCR method. *Agronomy, 14*(4), Article 658. https://doi.org/10.3390/agronomy14040658

Wang, C., Wang, C., Wang, L., Wang, J., Liao, J., Li, Y., & Lan, Y. (2023). A lightweight cherry tomato maturity real-time detection algorithm based on improved YOLOv5n. *Agronomy, 13*(8), Article 2106. https://doi.org/10.3390/agronomy13082106

Wang, C.-Y., Yeh, I.-H., & Liao, H.-Y. M. (2024). YOLOv9: Learning what you want to learn using programmable gradient information. *arXiv*. https://doi.org/10.1007/978-3-031-72751-1_1

Wang, D., Wang, X., Chen, Y., Wu, Y., & Zhang, X. (2023). Strawberry ripeness classification method in facility environment based on red colour ratio of fruit rind. *Computers and Electronics in Agriculture, 214*, Article 108313. https://doi.org/10.1016/j.compag.2023.108313

Yang, W., Ma, X., & An, H. (2023). Blueberry ripeness detection model based on enhanced detail feature and content-aware reassembly. *Agronomy, 13*(6), Article 1613. https://doi.org/10.3390/agronomy13061613

Yao, H., Liu, Y., Li, X., You, Z., Feng, Y., & Lu, W. (2022). A detection method for pavement cracks combining object detection and attention mechanism. *IEEE Transactions on Intelligent Transportation Systems, 23*(11), 22179-22189. https://doi.org/10.1109/TITS.2022.3177210

Zamri, F. N. M., Gunawan, T. S., Yusoff, S. H., Alzahrani, A. A., Bramantoro, A., & Kartiwi, M. (2024). Enhanced small drone detection using optimised YOLOv8 with attention mechanisms. *IEEE Access, 12*, 90629-90643. https://doi.org/10.1109/ACCESS.2024.3420730

Zeng, T., Li, S., Song, Q., Zhong, F., & Wei, X. (2023). Lightweight tomato real-time detection method based on improved YOLO and mobile deployment. *Computers and Electronics in Agriculture, 205*, Article 107625. https://doi.org/10.1016/j.compag.2023.107625

Pertanika J. Sci. & Technol. 34 (3): 1355 - 1381 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Review Article

Exploring Educational User Engagement: A Systematic Mapping of Serious Games with Gesture-based Interaction

Wan Salfarina Wan Husain[1,2]*, Syadiah Nor Wan Shamsuddin[1], and Normala Rahim[1]

[1]*Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Besut Branch, 22200 Besut, Terengganu, Malaysia*

[2]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Kelantan Branch, Bukit Ilmu, 18500 Machang, Kelantan, Malaysia*

ABSTRACT

Integrating serious games in educational settings has emerged as a promising approach to enhance learning experiences. In serious games, gesture-based interactions, such as motion sensors and wearable devices, enhance intuitiveness and engagement. The lack of standardisation in educational user engagement evaluation methods for serious games and gesture-based interaction has been highlighted in multiple studies. Most standardisation splits the dimensions, complicating the evaluation process. A systematic mapping study was conducted to address this issue, examining serious games and gesture-based interaction in education from 2018 to 2024 across five electronic databases. Despite wide interest among authors, there has been a limited systematic effort to define the concepts. This paper contributes by identifying elements used in serious games and gesture-based interaction for education, analysing research approaches and contributions, and proposing design criteria for evaluating educational prototypes based on user engagement. The study addresses differing opinions among authors and suggests solutions to unify concepts related to user engagement. The paper recommends several guidelines for serious games and gesture-based interaction in education, along with instruments to enhance user engagement.

Keywords: Education, gesture-based interaction, serious games, user engagement

ARTICLE INFO

Article history:
Received: 18 May 2025
Accepted: 30 December 2025
Published: 12 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.04

E-mail addresses:
salfa457@uitm.edu.my (Wan Salfarina Wan Husain)
syadiah@unisza.edu.my (Syadiah Nor Wan Shamsuddin)
normalarahim@unisza.edu.my (Normala Rahim)
* Corresponding author

INTRODUCTION

One of the current challenges in transferring educational materials to elementary school students is developing teaching methods

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

that enhance engagement with both games and educational content. Traditional classroom teaching methods, manually conducted by teachers, persist in schools and often fail to cater to students with diverse learning styles and the study approaches of the new generation. Consequently, there is an urgent need for greater implementation of various technologies and available media channels to address this issue effectively.

User engagement is one of the most prominent qualities in digital games and human-computer interaction that deals with user behaviour. Studies on learning engagement have gained growing attention from academicians around the globe, but the focus has been limited, with engagement often viewed as a subcategory of motivation (Mazlan et al., 2019). In education, student engagement is a multifaceted construct that is often considered complex, encompassing three primary components, which are behavioural, emotional, and cognitive dimensions (Dubey et al., 2023).

Serious games are believed to bring enjoyment to learning activities and sessions, allowing users to engage and have an optimal learning experience (Castellano et al., 2021). A serious game can deliver optimal practice opportunities to individual learners, thereby potentially precluding the onset of boredom. These games have become a valuable platform in both commercial and academic settings, enriching learning through their design and development (Mohd et al., 2018). A serious game designed for both academic and commercial purposes could stimulate business management scenarios, engaging users in realistic challenges like financial planning and marketing strategies with immediate feedback. In an academic context, such a game could be used to complement traditional coursework, offering practical application of theoretical concepts dynamically and engagingly. Researchers are particularly interested in serious games because their characteristics can motivate, engage, and educate players, depending on the domain design (Johnson et al., 2017). Morganti et al. (2017) pointed out that serious games are often applied in three key areas to support energy efficiency efforts, which were environmental education, raising awareness and lastly encouraging environmentally friendly behaviours.

Gesture-based interaction with technology in education refers to a method where users interact with digital devices and applications through physical gestures rather than traditional input methods like keyboards or mice. This interaction utilises natural user interfaces (NUIs) such as motion sensors, cameras, or wearable devices to detect and interpret gestures made by the user's body movements or hand motions. In educational contexts, gesture-based interaction enhances engagement by providing a more intuitive and immersive learning experience. It can be employed in various educational applications, from virtual classrooms to interactive learning environments, fostering active participation and facilitating the acquisition of knowledge and skills through interactive and responsive technology. Technology is instrumental to online education and can provide many benefits, such as student engagement, when implemented in the educational realm (Torres et al., 2021).

In general, the use of technology and specialised hardware, both online and offline, offers an alternative to conventional learning methods, especially in contexts requiring social interaction restrictions, such as during the COVID-19 pandemic. This approach helps prevent boredom and ensures safe learning. Since the lockdown era of COVID-19, particularly in Malaysia, educational methods have increasingly become technology-based.

In recent years, systematic mapping has been broadly used by some researchers because the nature of the study can help researchers understand in more detail the connection of concepts used in their related study. According to Fernandez et al. (2011) A systematic mapping study offers a structured way to explore what research has been done on a specific topic, helping to highlight existing gaps and understand the scope of available studies.

This paper is organised into five main sections, corresponding to the research question. The first section discusses the related work on serious games with gesture-based interaction about education and user engagement. In the second section, the research methodology outlines the systematic mapping process. Meanwhile, the third section analyses the results from data extraction and study mapping. The fourth section is a discussion that builds on previous work to address the challenge of mapping questionnaires for children's education and introducing tools and instruments for measuring user engagement. Finally, the last section of the paper delves into the summary of the scientific contributions. This article aggregates publications from five popular database engines in research by covering works from 2018 to 2024.

RELATED WORK

In the modern era, the use of technology has significantly influenced various fields, with games gaining rapid popularity among researchers for their potential in education, training, and cognitive development. Games enable learners to simulate and experience real-life environments, enhancing cognitive understanding through user control. Research demonstrates that the power of games is their capacity for engagement, replayability, fun, enjoyment, cooperative social mechanics, etc. (Anvari et al., 2024; Brockmyer et al., 2009). Serious games, designed to achieve specific learning outcomes, make learning fun and engaging while fostering essential skills and knowledge. According to Katsaliaki and Mustafee (2012), serious games effectively promote knowledge acquisition and problem-solving skills, often outperforming traditional methods in terms of performance and retention. In education, serious games enhance teaching and learning experiences by creating engaging, interactive learning opportunities. They can supplement traditional teaching methods, offering students additional ways to engage with the subject matter. Wong et al. (2020) found that game-based approaches effectively capture children’s attention and sustain their interest, proving to be more effective than traditional methods. Serious games provide individualised practice without boredom, a challenge for conventional instruction methods (Ronimus et al., 2019).

Additionally, virtual reality in serious games aids in recalling outcomes, understanding skills, and visually recognising components (Checa et al., 2021). However, there is a need for frameworks and guidelines to effectively design and present educational materials through serious games.

Virtual reality serious games, for instance, have the potential to enhance student learning and satisfaction by offering a deeper understanding of the subject matter (Checa et al., 2021). The study conducted by Torres et al. (2021) effectively adapted virtual simulations and serious games for mental health education. Their approach involved presenting scenarios from a first-person perspective, enabling learners to freely navigate 3D environments using mouse or camera control buttons within the user interface. This research also demonstrated the capability to create realistic training scenarios, including opportunities for psychomotor skills development. Similarly, Johnson-Glenberg et al. (2020) utilised virtual simulation serious games in STEM education learning.

Gesture-based interaction utilises the physical movements of the hands, fingers, or other body parts to engage with digital devices and systems. This approach harnesses natural human gestures to create a more intuitive and seamless method for controlling and communicating with technology. It often involves touchless interfaces, allowing users to issue commands through hand movements. The diversity of interfaces used in gesture-based interaction introduces the use of multimodal interaction. Multimodal interaction, as described by Nigay and Coutaz (1993), enables users to employ various modalities, such as voice, gesture, and typing, to communicate with a computer. Besides, some of the gesture-based interactions use a modern ubiquitous approach. In modern ubiquitous environments, interaction incorporates hands-free and eyes-free input from different body parts. For example, the waist, the largest joint in the human body, can be actively controlled (Xu et al., 2021). Our hands, being highly expressive, are particularly effective for interacting with our environment (Khalaf et al., 2019). Notable gesture devices include Leap Motion, Microsoft's Kinect, and Intel's RealSense. Hand gestures have been explored in various educational contexts, such as interacting with jigsaw games and computational thinking concepts (Rubegni et al., 2022) and road safety education (Wan Husain et al., 2022). However, there has been limited exploration of user engagement and metrics in educational interactions. Technology plays a crucial role in online education and can significantly enhance student engagement when effectively implemented (Torres et al., 2021).

User engagement is one of the most prominent qualities in digital games and human-computer interaction that deals with user behaviour. Research indicates that some authors prefer using consistent dimensions for measuring user engagement, often overlooking alternative perspectives. Despite this reason, the discussion of the user engagement dimensions will be useful in helping the researchers to focus on suitable dimensions to be implemented in their research within the respective domain and field that they want to focus. In this paper, the dimensions of user engagement measurement will be discussed in detail.

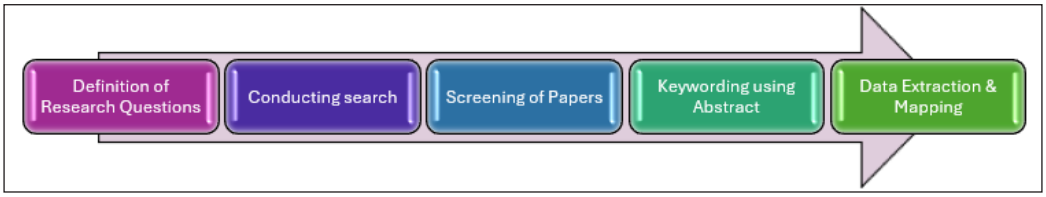


Figure 1. Systematic mapping process

RESEARCH METHODOLOGY

In this section, the systematic mapping study (SMS) includes the definition of research questions, conduct of search, screening of papers, keywording of abstracts, data extraction and mapping of the study as illustrated in Figure 1 (Ya'u et al., 2019).

Definition of Research Questions

In conducting the research, the research questions and objectives play important roles in constructing the analysis phase and the knowledge improvement flow. This SMS paper aims to answer the following questions:

RQ1 : What elements are used in serious games and gesture-based interaction for education?

RQ2 : What research approaches and contributions have the selected papers implemented (e.g. framework, method, tool)?

RQ3 : Which design criteria in serious games and gesture-based interaction are used to evaluate the education prototype and applications based on user engagement?

Conducting Search

According to Gusenbauer and Haddaway (2020), 14 databases can be used as principal search systems: ACM Digital Library, BASE, ClinicalTrials.gov, Cochrane Library, EbscoHost, OVID, ProQuest, PubMed, ScienceDirect, Scopus, TRID, Virtual Health Library, Web of Science, and Wiley Online Library. The research process for this paper was done by searching the materials in the Scopus, IEEE, ACM, Springer and Web of Science databases based on the research questions. The search strategy in searching the string used phrase searching.

Screening of Papers

The systematic mapping process adopted in this study was guided by a structured search strategy encompassing three sequential stages: pilot, primary, and snowballing searches designed to identify and select relevant literature for inclusion as shown in Figure 2.

Figure 2. Search strategy process

Pilot Search Strategy

Implementing a pilot search strategy within search strings aims to optimise the retrieval of relevant results from selected online databases and is a well-established practice in systematic mapping research (Gusenbauer & Haddaway, 2020). This strategy helps avoid the misconception of studies conducted in various disciplines and approaches suggested by the researchers. Furthermore, implementing this strategy can reduce biased findings by relying on the trusted online databases. Due to the cumbersome nature of search engines such as Scopus and IEEE in controlling the search strings, five different searches were implemented based on the research by systematic mapping done by Ya'u et al. (2019). During this paper's initial search phase in systematic mapping, five online databases were chosen: Scopus, IEEE, ACM, Springer and Web of Science. The search string using the Boolean operator employed ("serious games" OR "serious game" OR "games" OR "serious gaming" OR "applied games") AND ("gesture interaction" OR "gesture-based interaction" OR "gesture based interaction" OR "natural interaction" OR "natural user interaction") AND ("education" OR "primary school education" OR "kid education"). The results of the search string are shown in Table 1. Overall, from the five trusted databases, 2705 related publications were retrieved.

Table 1
Results of the pilot search in a trusted database

Database Name	Retrieve Paper
Scopus	52
IEEE	224
ACM	599
Springer	1804
Web of Science	26
Total	2705

Primary Search Strategy

It is inappropriate to evaluate all research papers using a uniform set of criteria, as some emphasise the introduction of novel techniques or findings with limited empirical validation, while others provide comprehensive discussions of conceptual frameworks. In the search strategy, to retrieve suitable sources is to play around with ambiguity, researchers find the technique to retrieve specific terms using Boolean operators very helpful. From that number of findings, the relevant research was narrowed down more specifically based on titles and keywords.

Snowballing Search

During the snowballing search, another cycle of searching was used. In this process, detailed checking was performed on the selected papers.

Inclusion and Exclusion Criteria

Articles using serious games and gesture-based interaction in the education field were included. Meanwhile, for the exclusion criteria, the following categories were excluded:

1. Studies that do not focus on serious games, gesture-based interaction, and education.
2. Short conference papers (less than four pages), opinion papers and work in progress reports.
3. Review papers and studies, such as surveys that do not have theoretical concepts, prototype testing and systematic literature review related to the topic.
4. Papers that are not written in the English language.

Keywording Using Abstract

In searching for keywords using the abstract, we executed the search from two different angles to focus on the topic in more detail. The criteria were narrowed down to increase focus on the topic selected. The selected papers were chosen by shortlisting the keywords used by these papers and by screening abstracts that contained the keywords.

Data Extraction and Mapping of Studies

The data extraction form was designed to contain the following information: study ID, date of the extraction, paper title, author(s), publication type/year, Source, study context, concept(s), area of focus and in answering the research question (Ya'u et al., 2019). In this paper, the data extraction focuses on study ID, paper title, authors, publication type/year, source, study context, concept(s), and area of focus. The sample of extraction is illustrated in Table 2.

After further screening and selection based on abstracts (63 papers) and lastly, for deep discussion, only 28 full-text papers were extracted. The summary of the extraction is shown in Figure 3.

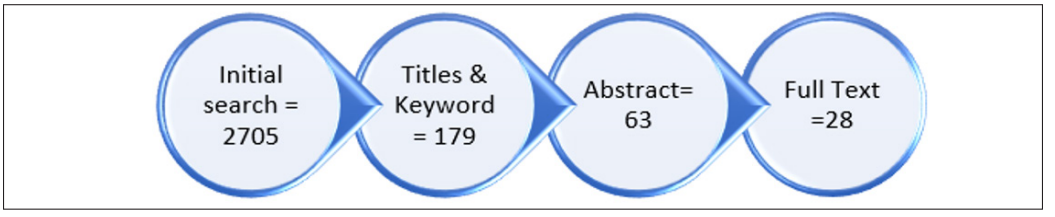


Figure 3. Primary search phase activities

Table 2
Sample of data extraction

StudyID	Database	Paper Title	Publication Year
1	SPRINGER	Digital game elements, user experience and learning: A conceptual framework	2018
2	SPRINGER	An immersive virtual-reality computer-assembly serious game to enhance autonomous learning	2021
3	SPRINGER	Exploring spiral narratives with immediate feedback in immersive virtual reality serious games for earthquake emergency training	2022
4	SPRINGER	Supporting struggling readers with digital game-based learning	2019
5	SPRINGER	The effects of touchless interaction on usability and sense of presence in a virtual environment	2022
6	SPRINGER	FunQ: Measuring the fun experience of a learning activity with adolescents	2021
7	SPRINGER	Educational games and the new forms of interactions	2019

DESCRIPTIVE ANALYSIS RESULT

The study population was categorised based on publication years and types, allowing for a comprehensive analysis of trends and variations over time. The literature search spanned from 2018 to December 2024, yielding retrieved papers distributed as follows: four papers from 2018, five papers from 2019, one paper from 2020, six papers from 2021, five papers from 2022, four papers from 2023 and three papers from 2024. Further details and visual representations can be found in Figure 4.

For the publication type, 28 articles were sourced from journals and conference proceedings, shown in Figure 5. Among these findings, 13 articles (46%) were retrieved from journals, while the remaining 15 articles (54%) were sourced from conference proceedings.

In addressing the research questions, the classification scheme revolves around three fundamental aspects: 1) Elements suitable for use in game design and interaction design within educational contexts, 2) Contribution facets, and 3) Evaluation criteria for measuring user engagement.

Elements of Serious Games and Gesture-based Interaction in Education

The elements of serious games and gesture-based interaction for education were examined, keeping in mind their various goals within the field of education and non-education. The analysis indicates a predominant emphasis on the field of education, with a total of 20 papers, while 8 papers deviate from this focus, as detailed in Table 3.

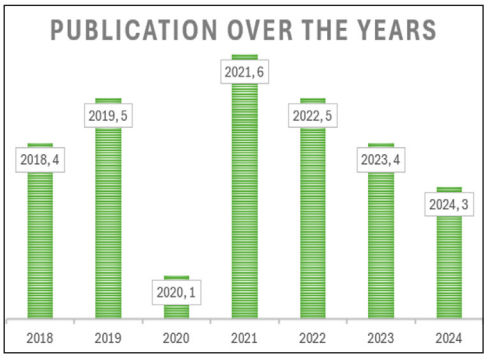


Figure 4. Publication over the years

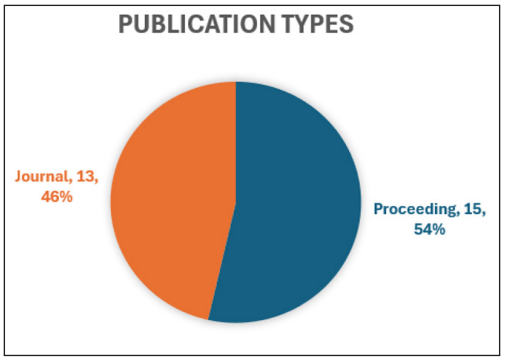


Figure 5. Publication type

Table 3
Distribution of selected studies according to education and non-education fields

Field	# Study	Paper
Education	20	Anvari et al. (2024); Chang and Tsai (2018); Checa et al. (2021); Chiu et al. (2018); de Paolis and de Luca (2022); Denden et al. (2018); Feng et al. (2022); Fernandes and Werner (2019); Holz et al. (2024); Johnson-Glenberg et al. (2020); Lombana et al. (2023); Munsinger and Quarles (2019); Nie et al. (2022); Ronimus et al. (2019); Rubegni et al. (2022); Shao et al. (2021); Tisza and Markopoulos (2021); Torres et al. (2021); Wan Husain et al. (2022); Zaina et al. (2019)
Non-education	8	Alexiou and Schippers (2018); Bian et al. (2023); Holz et al. (2024); Krivich et al. (2023); Meena et al. (2019); Silva et al. (2024); Vellingiri et al. (2020); Xu et al. (2021)

In this systematic mapping research, three modalities were compared and analysed: serious games, gesture-based interaction and the hybridisation between gesture-based interaction and serious games. Overall, for the SMS, 17 elements reflect the analysis of approaches and contribution-based elements used in serious games, gesture-based interaction, and user engagement, as shown in Table 4. Based on the synthesis of the study, 13 key elements commonly identified in serious games include interaction, avatar, user control, feedback, narrative, scenario-based, reward, challenge, aesthetic, incremental learning, goal, accommodating learning style, and scaffolding. Meanwhile, for user engagement, the components involved are cognitive, emotional, performance and engagement. Therefore, the conclusions will be valuable for everyday life, especially as future educational standards increasingly emphasise the use of technology in learning environments.

Serious games integrate game elements not only for entertainment but also for educational, training, simulation, or real-world problem-solving purposes, enhancing

engagement and efficacy in learning and problem-solving experiences. In addition to serious games, gesture-based interaction was examined as another technological context. Gesture-based interaction involves utilising body movements or gestures to interact with digital devices or systems, enabled by sensors, cameras, or other technologies that interpret and respond to these gestures. Such interfaces are utilised across various domains, such as gaming, virtual reality, augmented reality, and human-computer interaction. Here are four key elements associated with gesture-based interactions, which are interaction, avatar, user control and feedback, based on Table 4, which focuses on gesture. Additionally, elements lacking support from other research were excluded to prevent biases. According to the classification in Figure 6, the researchers have found that based on research question one (RQ1), the reviewed papers examined various elements, including narrative and scenario-based approaches (six papers each), feedback and user control (eight papers each), and the majority (fourteen papers) focusing on the use of avatars and interaction.

Table 4
Detailed analysis of approaches and contribution-based elements used in serious games, gesture-based interaction and user engagement

Authors/ Concept Used/ Component Involved	Interaction	Avatar	User Control	Feedback	Narrative	Scenario-base	Reward	Challenge	Aesthetic	Incremental Learning	Goal	Accommodating Learning Style	Scaffolding	Cognitive	Emotional	Performance	Engagement
Serious Game, Gesture-based Interaction																	
Checa et al. (2021)	✓		✓			✓					✓						
Rubegni et al. (2022)	✓	✓															
Nie et al. (2022)	✓	✓				✓				✓							
Shao et al. (2021)		✓			✓												
Wan Husain et al. (2022)	✓		✓			✓				✓		✓	✓				
Serious Game																	
Anvari et al. (2024)	✓	✓		✓			✓		✓								
Alexiou and Schippers (2018)			✓	✓	✓		✓	✓	✓		✓						

Table 4 (*continued*)

Authors/ Concept Used/ Component Involved	Interaction	Avatar	User Control	Feedback	Narrative	Scenario-base	Reward	Challenge	Aesthetic	Incremental Learning Goal	Accommodating Learning Style	Scaffolding	Cognitive	Emotional	Performance	Engagement
Bian et al. (2023)		✓		✓	✓			✓								
Chang and Tsai (2018)	✓															
Denden et al. (2018)	✓	✓		✓	✓	✓	✓				✓					
Feng et al. (2022)				✓	✓	✓										
Fernandes and Werner (2019)		✓														
Holz et al. (2024)	✓		✓	✓			✓		✓							
Krivich et al. (2023)		✓			✓											
Lombana et al. (2023)		✓	✓													✓
Meena et al. (2019)											✓					
Ronimus et al. (2019)		✓					✓	✓		✓		✓	✓	✓	✓	
Silva et al. (2024)	✓															
Tisza and Markopoulo (2021)		✓	✓					✓								
Torres et al. (2021)	✓	✓				✓										
Xu et al. (2021)	✓			✓												
Yang et al. (2023)	✓															
Gesture based Interaction de Paolis and de Luca (2022)	✓	✓	✓										✓	✓		
Zaina et al. (2019)	✓	✓	✓	✓												
Total	14	14	8	8	6	6	5	4	4	3	2	3	1	2	2	1	1

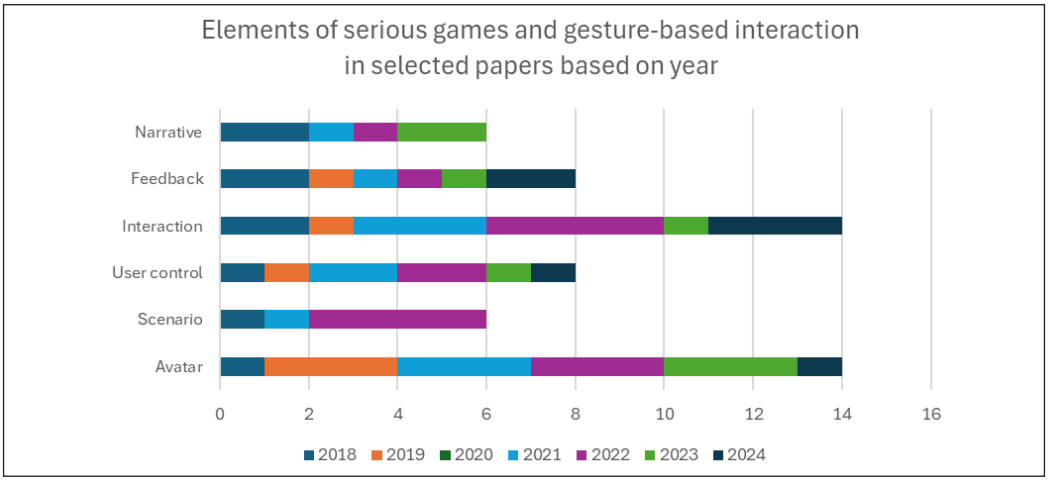


Figure 6. Elements of Serious Games and Gesture-Based Interaction in selected papers based on the year

In summary, the most effective gameplay for users implemented in serious games and gesture-based interaction depends on the cognitive learning of the target users. Key elements of serious games depend on their goals and objectives. Design elements were retrieved based on pedagogy-driven SGs. In a pedagogy-driven approach, it highlights the main pedagogical and entertainment features, and their interrelation (Arnab et al., 2014). Based on SMS for educational user engagement, the selection of the game elements for further development is based on their consistent association with positive learning engagement from the prior study, based on 5 databases. Clearly, from the findings, we can see that the most important elements that connect serious games and gesture-based interactions are interaction elements throughout the games. Additionally, the use of avatars is another important element in the field of education. Moreover, the sixth element mentioned earlier plays the biggest role in promoting skill development and is thought to enhance user engagement.

Interaction

The interaction in educational perspective refers to how players and games engage, such as based on environments and setup of scenario, the characters use that guide to physical, cognitive and social development. A well-designed interaction system will help the game feel natural and easy to use, allowing players to stay focused and feel involved. The players must deal with the situation and implement interaction (Wan Husain et al., 2022). Research by Checa et al. (2021) It stresses that interaction techniques in virtual reality are divided into three basic forms: selection, manipulation, and locomotion. In this research, the user only had to focus on manipulating objects using a single mechanism to pick up

and drop them and place these objects into position. The simple interaction in the game is by using a mouse and keyboard. Still, in high-level interaction design, the interaction can be controlled by a device that supports gestures, movement, etc. According to Yang et al. (2023) Integrating interactive elements allows users to engage more deeply with the website, resulting in a more memorable and enjoyable experience. Based on the study of 28 papers, most interaction elements were integrated with others, but a quick and different approach was implemented by Silva et al. (2024) was more focused on gesture interaction technique defined as: anticlockwise, clockwise, up, down, left, and right.

Avatar

Virtual characters and avatars are commonly some of the elements implemented in games with a specific purpose. Some authors mentioned that avatars are related to engagement (Ronimus et al., 2019), and some authors used the avatars as a feature, as they add value to the applications. Avatars in game design offer user satisfaction, which can result in gaining players' loyalty and increasing engagement. Rubegni et al. (2022) found that using avatars as the central point of interaction can be especially effective in capturing children's attention and encouraging their involvement. Their study also suggested that avatars have the potential to support the creation of more engaging and effective educational tools for young learners. In gesture interaction, the avatar and characters are used to control the game's movement and to perform an action (Zaina et al., 2019). In their research, the authors proved that avatars play an engaging role in helping children better remember what they've learned. Avatars can be used to represent players' selves in the game (Denden et al., 2018; Zaina et al., 2019). Avatar is an intermediary character to connect with the virtual world and pairs together with a scenario-based environment to represent the real-world environment, like the study by Nie et al. (2022), Shao et al. (2021) and Torres et al. (2021). In other words, these two important elements in games can bridge the content to deliver it to users more effectively and in an impactful manner. An avatar in the application can be represented in the virtual world by a simple cursor or more commonly by a hand or a whole-body avatar (Fernandes & Werner, 2019). The study by De Paolis and De Luca (2022) found that in their framework, which focuses on presence, immersion and interaction, proto presence is based on kinesthetic information about the relative position of the user's body, which allows advanced technology devices such as Leap Motion and Kinect to be implemented in more detail to bring more effective effects for the avatars.

Scenario-based

Scenario-based learning materials for the players are designed based on real-life situations (Wan Husain et al., 2022). In an easy explanation, the players tend to learn more when the game content is related to their everyday surroundings or real-life experiences. The scenario

implements the facilitation of reflection-in-action, which is important to train individuals about behavioural responses (Feng et al., 2022). Scenario-based features can be divided into static and dynamic (Torres et al., 2021). Many researchers use static scenarios in which the structure design is not adjustable by the user (Denden et al., 2018; Feng et al., 2022; Nie et al., 2022; Torres et al., 2021). However, the dynamic scenario also exists, such as the research by Torres et al. (2021).

User Control

User control means the player can control the learning process. The player can repeat or continue the learning process (Feng et al., 2022) based on trial and error learning (Wan Husain et al., 2022). This situation gives an advantage to the player to revisit a scenario and give a different response to avoid errors or mistakes.

Feedback

Feedback is a game feature that directs learners to evaluate their performance, identify knowledge gaps, and obtain correct knowledge by receiving various forms of information (Feng et al., 2022). Meanwhile, research by Bian et al. (2023) mentioned that feedback helps players effectively perceive their feelings and performance during the game task; the system should be able to give players quantitative and visual feedback on how they are doing. Three types of stimulation are deployed with immediate feedback: image-based feedback, audio feedback, and text-based feedback. Feedback devices can influence the cognitive engagement of players, and qualities like immediacy, frequency, clarity and relevance can contribute to that (Alexiou & Schippers, 2018). Through this research, the relationship of the fragile feedback mechanisms can disrupt user engagement. Therefore, designing or evaluating game-based learning applications needs to take feedback as an important element in educational platform games.

Narrative

The study by (Bian et al., 2023; Feng et al., 2022) used a narrative element, where the storyline followed a structured sequence of scenarios. Learners' actions determined the progression from one scenario to the next. This finding shows that narrative elements are also connected to scenario-based elements. Providing clear instructions and flow to the users allows the players to play more confidently and increases user engagement.

Contribution Facet

A contribution facet refers to any added value that has some impact on the research. For this study, contribution facets focus on the approaches and contributions of the selected

Table 5
Contribution facet study

Category	Description
Framework	A framework is a structured approach or set of guidelines that provides a foundation for organising and solving complex problems.
Tool	A "tool" is an object or software application that is designed to perform a specific task or function, often making the task easier, more efficient, or more accurate.
Evaluation	Evaluation is the process of ensuring an algorithm achieves the objectives.
Metric	A "metric" is a quantifiable measurement or standard used to assess, evaluate, or express a particular attribute, performance, or characteristic of something. Metrics are used in various fields to track progress, compare results, make informed decisions, and understand the effectiveness of processes or systems. Metrics are valuable because they provide objective data that can be used for analysis and decision-making.
Model	A model can refer to various concepts depending on the context in which it is used.

papers, such as framework, tools, evaluation, metrics and models used in their research. The contribution facet study is described in detail in Table 5.

The researchers made numerous contributions in their work while developing quality research. According to Table 6, most contributions focused on the evaluation facet, with 13 papers dedicated to this area. Secondly, the findings were more focused on the framework with 7 papers. Then, it was followed by tools and metrics, each represented by 5 papers, and finally, models with 3 papers. Considering that the focus of the 28 papers lies in evaluation, the prevalent elements utilised are interaction, avatar, feedback, goal, challenge, accommodating learning style, scenario-based design, incremental learning, narrative, aesthetic and user control. Next, an examination of four papers within the framework studies revealed the implementation of 17 elements and concepts: goal, reward, feedback, challenge, narrative, aesthetic, scenario-based design, avatar, accommodating learning styles, interaction, user-centred design, behaviour-based approaches, cognitive elements, emotional factors, scaffolding, user control, and incremental learning. In the studies concerning tools, which were drawn from three papers, researchers predominantly employed 7 elements and concepts of serious games and gesture-based interaction: feedback, scenario-based design, avatar, interaction, cognitive elements, emotional factors, and user control. Meanwhile, in the model facet study, it's worth noting that no redundant concepts were integrated.

There is a different situation for the metric facet; the different criteria used need to be captured through the system settings and the observation methods. From the six extracted papers, a consensus emerged that time is a key factor in measuring various aspects, including interaction, fatigue, usability, behaviour, performance, and knowledge. These measurements are supplemented by factors such as the time

Table 6
Contribution facet based on framework, tool, evaluation, metrics, and model

Contribution Facet	Framework	Tool	Evaluation	Metric	Model
Studies	Alexiou and Schippers (2018); Anvari et al. (2024); Bian et al. (2023); de Paolis and de Luca (2022); Denden et al. (2018); Wan Husain et al. (2022); Silva et al. (2024)	de Paolis et al. (2019); Silva et al. (2024); Torres et al. (2021); Xu et al. (2021); Yang et al. (2023)	Checa et al. (2021); de Paolis and de Luca (2022); Fernandes and Werner (2019); Holz et al. (2024); Lombana et al. (2023); Meena et al. (2019); Nie et al. (2022); Ronimus et al. (2019); Rubegni et al. (2022); Shao et al. (2021); Tisza and Markopoulos (2021); Xu et al. (2021); Zaina et al. (2019)	Chiu et al. (2018); Johnson-Glenberg et al. (2020); Munsinger and Quarles (2019); Vellingiri et al. (2020); Xu et al. (2021)	Feng et al. (2022); Lombana et al. (2023); Ronimus et al. (2019)
Count	7	5	13	5	3

taken, number of rotations, incorrect actions, task completion, and success rates. The detailed analysis for the metric facet can be referred to in Table 7. According to Ya'u et al. (2019), engagement can be gauged through three main metrics: 1) exposure time extracted from game logs post-intervention, 2) active engagement time encompassing the duration users spend on learning tasks within the game, and 3) exclusion of time spent on other game features like avatar movement and pauses.

Evaluation Used in Prototypes and Applications

Evaluation constitutes a pivotal phase in prototype and application development, ensuring alignment with user requirements, functionality, and delivering a positive user experience. The evaluation methods employed in both prototypes and applications in this study are summarised in Table 8.

Based on the analysis of the 13 papers presented in Table 9, it is evident that usability studies were the most employed evaluation method, utilised in six papers. Additionally,

Table 7
Detailed analysis for the metric facet

Study ID/ Metric	Time	Number of Rotations	Number of Wrong	Success Rate	Completion Task	Metric Type
Chiu et al. (2018)	✓	✓	✓			Behavior
Johnson-Glenberg et al. (2020)	✓		✓			Knowledge
Munsinger and Quarles (2019)	✓				✓	Interaction, fatigue, usability
Vellingiri et al. (2020)	✓	✓		✓		Performance
Xu et al. (2021)	✓				✓	Usability
Yang et al. (2023)	✓					Emotion

Table 8
Evaluation used in prototypes and applications

Evaluation Types	Description
Usability	The ease with which users can interact with a product, system, or interface to achieve their goals effectively and efficiently while having a positive user experience.
Emotional and cognitive engagement	User engagement involves users' feelings, attitudes, and overall emotional response to a product or experience, as well as their intellectual or mental involvement with the product or content.
Interview	The technique for collecting data may be highly structured, resembling questionnaires, or highly unstructured, starting with general questions that allow the respondent to lead the way.
Heuristic	The method is used to review software and spot any issues related to how easy it is to use. It involves experts going through the interface and checking whether it follows well-known usability guidelines.
Performance testing	Performance testing to evaluate responsiveness, functionality and assess factors like loading times, response times, and overall system stability to ensure a smooth and efficient user experience.

Table 9
Measurement tools for evaluating engagement

Paper	Tool Name	Evaluation Engagement Dimension/ Measurement
Checa et al. (2021)	Satisfaction questionnaire adapted from (Tcha-Tokey et al., 2016)	Usability
de Paolis and de Luca (2022)	Presence Questionnaire (PQ) for the evaluation adapted by combining the Usability Metric for User eXperience (UMUX) questionnaire, the System Usability Scale (SUS) and the presence questionnaire (PQ)	Cognitive, emotional

Table 9 (*continued*)

Paper	Tool Name	Evaluation Engagement Dimension/ Measurement
Fernandes and Werner (2019)	Heuristic	Usability
Holz et al. (2024)	User Experience Questionnaire (UEQ) (Schrepp et al., 2017), 13 questions from the Kids Game Experience Questionnaire KidsGEQ	User Experience
Lombana et al. (2023)	Not mentioned in detail	Engagement
Meena et al. (2019)	SUS, NASA	Usability, task workload
Nie et al. (2022)	Not mentioned in detail	Usability
Ronimus et al. (2019)	GraphoLearn engagement	Cognitive, Emotional, performance
Rubegni et al. (2022)	Not mentioned in detail	Interview with students
Shao et al. (2021)	Not mentioned in detail	Usability
Tisza and Markopoulos (2021)	FunQ	Fun experience learning
Xu et al. (2021)	NASA-TLX Hart and Staveland (1988), IPQ	Usability
Zaina et al. (2019)	Not mentioned in detail	Interview with teachers

three studies assessed user engagement, incorporating both cognitive and emotional dimensions. Two papers adopted interview-based approaches, while one focused on user experience and another examined the aspect of fun in learning. Notably, the review highlights a limited exploration of the multifaceted dimensions of user engagement across the selected studies.

The data extraction results from our systematic mapping study offer a detailed and structured overview of the current research within the field. The analysis identified key elements, including dominant research components, commonly addressed contribution facets, and areas of user engagement that remain underexplored. These findings provide a clear picture of the existing knowledge base and highlight specific research gaps that need attention.

DISCUSSION

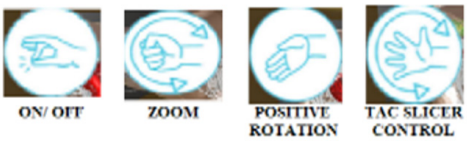
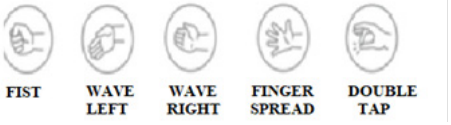
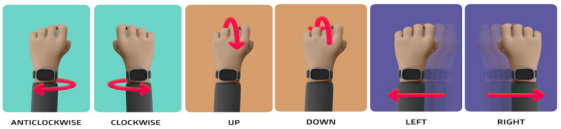
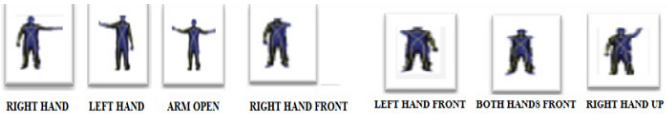
Based on the findings, the different authors offer various perspectives, and some have made significant contributions in relating the theory and implementation. This paper also focuses on the dimension in measuring the theory used. In answering the research question, the study of serious games and gesture-based interaction was first mapped, and the patterns and modalities were explored in terms of action using the game elements. The studies by Nie et al. (2022), Shao et al. (2021) and Torres et al. (2021) agree that using an avatar

would be the glue to connect players with a scenario-based environment and the control by users. These studies also indicate that using an avatar gives significance to the interaction.

Avatar-based generally refers to interactions or experiences involving avatars, which are digital representations or characters representing users in a virtual environment. Avatars can take various forms, such as human-like figures, animals, or abstract representations, and they are commonly used in digital media, virtual reality video games, and online platforms. Avatar-based interactions are driven by the desire to create more personalised, engaging, and immersive digital experiences. It allows users to have a presence in virtual or online spaces, enhancing communication, collaboration, and overall user engagement. Advances in technology, such as motion tracking and facial expression recognition, contribute to making avatar-based experiences more realistic and interactive.

Based on the findings, it can be inferred that the utilisation of gestures in advanced technology for gaming typically ranges from four to seven gestures, while users with disabilities may employ a greater number of gestures. Feng et al. (2022) found a relationship between scenario-based learning with narrative, which is based on the user control elements. By adjusting the narrative, users can replay a scenario with instant feedback, encouraging reflection-in-action. This gives players the freedom to decide whether to repeat the task or move forward. Table 10 summarises the avatar action in controlling the virtual environment.

Table 10
The avatar (interaction) action in controlling the virtual environment

Gesture Used	Action Using Serious Games Elements
Four hand gestures	
Five hand gestures	
Five hand gestures	
Six hand gestures	
Seven basic gestures	

From the research approaches and contribution facets, the primary focus was on evaluation, as the major rules for the design and development of prototypes and frameworks need to be assessed in the final stage of development. Overall, both the evaluation facet and the contributions of the approach yielded the same result, emphasising three main attributes: avatar, interaction, and user control. This finding supports the discussion in research question one, which examined the elements used in serious games and gesture-based interaction for education.

In line with Mayer's proposition of teaching experience design, any teaching experience should undergo comparison using valid, reliable, objective, and referenced instruments alongside experimental control groups employing conventional teaching (Checa et al., 2021; Mayer, 2014). Research conducted by Chiu et al. (2018) introduces behaviour metrics for measuring gesture-based interaction, including completion time, number of rotations, and number of incorrect attachments per level, as these metrics are linked to task difficulty and player impulsivity. The authors also recommend additional metrics such as the number of rotations per successful attachment (calculated by dividing the number of rotations by the number of successful attachment attempts) and the average time spent between each rotation (calculated by dividing the total time taken to finish by the number of rotations).

Moreover, some research employs an authentication method to restrict access to the collected traces and results, ensuring that only authorised individuals can access them and safeguarding the learners' privacy. Additionally, the collected traces and generated reports are retained for a predetermined period (typically one academic year) before being automatically deleted. Furthermore, to enhance transparency in learning, students have the option to view their collected gaming behaviour traces.

Nevertheless, there is still a lack of a thorough theoretical understanding of how the different elements embedded in game design influence user engagement in learning, something this research set out to do. Based on previous empirical research by Alexiou and Schippers (2018), user engagement can be divided into four main dimensions: 1) openness to experience, which is crucial for user engagement and technology acceptance, encompassing traits such as imagination, sensitivity to aesthetics, curiosity, and independent thinking; 2) conscientiousness, referring to individuals' perseverance and dedication to their learning goals, which enhances performance and active learning; 3) goal orientation, which includes both learning goal orientation and performance goal orientation; and 4) sensation seeking and need for cognition, where behaviours can lead to a higher sense of control and mastery, fulfilling needs for competence and autonomy, and contributing to immersion and enjoyment. The model proposed by Alexiou and Schippers (2018) also suggests that player engagement is closely linked to the narrative, game mechanics, and aesthetics (user control). Together, these elements boost both emotional and cognitive involvement.

Quite a different study by de Paolis et al. (2019) focused on evaluating user experience. Although the focus of this research is user experience, the framework also uses the same game elements for user engagement, which are presence, immersion and interaction. According to Tisza and Markopoulos (2021), the concepts of experience and engagement represent distinct theoretical ideas and should not be treated as the same when examining users' subjective experiences. In response to this distinction, they developed the FunQ evaluation framework, which assesses six dimensions: Autonomy, Challenge, Delight, Immersion, Loss of Social Barriers, and Stress. Through a detailed systematic mapping, it is evident that these elements are interconnected and, when implemented together, can enhance game design and user engagement more effectively. Based on the existing mapping structure derived from selected papers across five database engines, an initial item pool of instruments was constructed, focusing on attributes and elements of gesture-based interaction and serious games. These instruments are designed to facilitate the more effective measurement of user engagement and usability. The final version for testing user engagement and usability incorporates seven attributes and elements of gesture-based interaction and serious games.

Methods for evaluating user engagement predominantly stem from the domain of human-computer interaction, particularly in education. User engagement stands as a crucial component within the fields of serious game development and interface design. Most papers extracted do not solely concentrate on user engagement but rather integrate it with usability and other aspects of interaction design. A compilation of instruments has been assembled to measure user engagement, drawing from 28 papers extracted from the five renowned databases. This compilation is presented in Table 11, covering various items such as avatar, narrative, user control, enjoyment, competence, immersion, scenario-based design, visual fidelity, learning goal and interaction.

A range of methodological approaches has been employed across studies to assess engagement and attitudes, each adapted to suit the characteristics of the target population. Feng et al. (2022) utilised a 7-point Likert scale extending from -3 for “totally disagree” to +3 for “totally agree”, enabling a detailed gradient of agreement particularly suitable for adult or older participants. In contrast, Ronimus et al. (2019) designed an engagement measurement instrument specifically tailored for children, addressing both emotional and cognitive components. To ensure the statements were comprehensible to young respondents, the scale was piloted in advance. The assessment employed a visual format comprising five differently sized squares to indicate varying frequencies of engagement, ranging from “every time” to “never,” thereby facilitating children's understanding through concrete visual representations. Similarly, Rubegni et al. (2022) adopted a 5-point scale ranging from 1 for “not at all” to 5 for “very much”, supplemented by emoticons in a “smile-o-meter” format, a design inspired by (Denden et al., 2018). This visual enhancement supported younger participants in expressing affective responses more intuitively. Collectively, these approaches

Table 11
Measurement tool for evaluating user engagement based on systematic mapping

Elements	Instruments
Enjoyment Ronimus et al. (2019)	The following items measured emotional engagement: 1. I enjoy playing GL. 2. I would like to play GL even more. 3. It is fun to practice reading with GL. 4. I could play GL forever. 5. Playing GL makes me happy.
Lombana et al. (2023)	1. I enjoyed the game 2. I was trying my best in the game 3. It was easy to play the game 4. I liked the game graphics
User control (autonomy) Ronimus et al. (2019)	The following items were used to assess cognitive engagement: 1. I try my best when I play GL. 2. I choose my responses carefully. 3. I like it when a difficult task appears in the game. 4. I concentrate hard when I play. 5. I like to read even the difficult words in GL.
Tisza and Markopoulos (2021)	1. During the activity, I had the sense of controlling the activity. 2. During the activity, I knew what to do. 3. During the activity, I could do what I wanted. 4. During the activity, I could make some choices about the activity.
Interaction Rubegni et al. (2022)	Survey 1. How much did you enjoy playing the game? 2. Would you recommend it to other children? 3. Have you ever used a gesture-controlled game (e.g., Kinect or Wii-controlled games)? 4. Do you remember something about the game? Interview 1. What did you like about this game? 2. What did you not like about this game? 3. Do you remember something about the game? If yes, what?
Easy to use (Narrative) Feng et al. (2022)	1. The training storyline helped me to learn. 2. It was easy for me to understand the learning content. 3. It was easy for me to learn about what to do during and after earthquakes. 4. It was easy for me to remember what I had learned.
Competence (Self-efficacy) Feng et al. (2022)	1. I know what to do when facing an earthquake. 2. I can remain calm when facing an earthquake. 3. I have the confidence to deal with an earthquake alone. 4. I can come up with a plan for responses to an earthquake. 5. I can handle situations during an earthquake. 6. I can think of a solution if I am in trouble during an earthquake.
Scenario-based (attention) Feng et al. (2022)	1. It was easy for me to concentrate on my learning. 2. It was easy for me to stay focused on the task. 3. I felt the training was fun.

Table 11 (*continued)*

Elements	Instruments
Immersion de Paolis and de Luca (2022)	1. Were you able to anticipate what would happen next in response to the actions that you performed? 2. How completely were you able to actively survey or search the environment using vision? 3. How much delay did you experience between your actions and expected outcomes? 4. How quickly did you adjust to the virtual environment experience? 5. How proficient in moving and interacting with the virtual environment did you feel at the end of the experience? 6. How well could you concentrate on the assigned tasks or required activities rather than on the mechanisms used to perform those tasks or activities?
Visual fidelity de Paolis and de Luca (2022)	1. How closely were you able to examine objects? 2. How well could you examine objects from multiple viewpoints?
Aesthetic (avatar) Shao et al. (2021)	Semi-structured interview: 1. Making an avatar with my own face makes me is more willing to tell stories and get very excited. 2. The character with a customised avatar is very attractive.
O'Brien et al. (2018)	1. This Application X was attractive. 2. This Application X was aesthetically appealing. 3. I liked the graphics and images of Application X. 4. Application X appealed to the visual senses. 5. The screen layout of Application X was visually pleasing.
Focus on the learning goal O'Brien et al. (2018)	1. I lost myself in this experience. 2. I was so involved in this experience that I lost track of time. 3. I blocked out things around me when I was using Application X. 4. When I was using Application X, I lost track of the world around me. 5. The time I spent using Application X just slipped away. 6. I was absorbed in this experience. 7. During this experience, I let myself go.

underscore the significance of employing developmentally appropriate and contextually relevant instruments when assessing engagement, particularly in child-centred research.

In conclusion, the discussion of our systematic mapping study has elucidated significant patterns and gaps within the existing research landscape. Through our analysis, we identified prevalent methodologies, core thematic elements in serious games and gesture-based interaction, and underexplored topics that warrant further investigation, especially to measure user engagement. These findings provide a foundation for guiding future research efforts and fostering more targeted and impactful studies. By addressing the identified research gaps and leveraging the emerging trends, researchers can contribute to a more comprehensive understanding of the field.

CONCLUSION

This paper focused on identifying and theorising the relationship between elements of serious games, gestural interaction, and user engagement within the education domain from 2018 to 2024. The analysis encompassed three main aspects: 1) the elements utilised in serious games and gesture-based interaction for educational purposes, 2) the research methodologies and contributions implemented in the selected, and 3) the design criteria employed in serious games and gesture-based interaction to evaluate educational prototypes and applications based on user engagement. User engagement and usability questionnaires were collected and synthesised to enhance the development and efficacy of educational apps for students. Consequently, this paper aimed to condense the key design elements essential for measuring user engagement, providing a comprehensive measurement tool tailored to the educational domain. Furthermore, the findings offer insights and guidelines for industry practitioners involved in measuring user engagement using this tool. Future research should aim to build on these findings, employing rigorous methodologies and fostering interdisciplinary collaboration to further enhance the depth and breadth of knowledge in this domain.

ACKNOWLEDGEMENT

The authors would like to thank the Department of Informatics and Computing, UniSZA and the Centre of Computing Sciences, Faculty of Computer and Mathematical Sciences, UiTMCK for providing a research platform to enrich new knowledge for compound systems. This research received no external funding.

REFERENCES

Alexiou, A., & Schippers, M. C. (2018). Digital game elements, user experience and learning: A conceptual framework. *Education and Information Technologies, 23*(6), 2545-2567. https://doi.org/10.1007/s10639-018-9730-6

Anvari, S. S., Hammer, J., & Wehbe, R. R. (2024). "More than just a game, it's an app that builds awareness around mental health": Mental health stigma reduction using games for change. *Proceedings of the ACM on Human-Computer Interaction, 8*(CHI PLAY), Article 316. https://doi.org/10.1145/3677090

Arnab, S., Lim, T., Carvalho, M. B., Bellotti, F., de Freitas, S., Louchart, S., Suttie, N., Berta, R., & de Gloria, A. (2014). Mapping learning and game mechanics for serious games analysis. *British Journal of Educational Technology, 46*(2), 391-411. https://doi.org/10.1111/bjet.12113

Bian, Y., Zhou, C., Zhang, Y., Liu, J., Sheng, J., & Liu, Y. J. (2023). Focus on cooperation: A face-to-face VR serious game for relationship enhancement. *IEEE Transactions on Affective Computing, 15*(3), 913-928. https://doi.org/10.1109/TAFFC.2023.3306198

Brockmyer, J. H., Fox, C. M., Curtiss, K. A., McBroom, E., Burkhart, K. M., & Pidruzny, J. N. (2009). The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology, 45*(4), 624-634. https://doi.org/10.1016/j.jesp.2009.02.016

Chang, C. K., & Tsai, Y. T. (2018). Pair-programming curriculum development of motion-based game for enhancing computational thinking skills. *Proceedings of the 2018 7th International Congress on Advanced Applied Informatics (IIAI-AAI)*, 284-287. https://doi.org/10.1109/IIAI-AAI.2018.00061

Checa, D., Miguel-Alonso, I., & Bustillo, A. (2023). Immersive virtual-reality computer-assembly serious game to enhance autonomous learning. *Virtual Reality, 27*(4), 3301-3318. https://doi.org/10.1007/s10055-021-00607-1

Chiu, P. T., Wauck, H., Xiao, Z., Yao, Y., & Fu, W. T. (2018). Supporting spatial skill learning with gesture-based embodied design. *Proceedings of the 2018 International Conference on Intelligent User Interfaces*, 67-71. https://doi.org/10.1145/3172944.3172994

de Paolis, L. T., & de Luca, V. (2022). The effects of touchless interaction on usability and sense of presence in a virtual environment. *Virtual Reality, 26*, 1551-1571. https://doi.org/10.1007/s10055-022-00647-1

de Paolis, L. T., de Luca, V., & Paladini, G. I. (2019). Touchless navigation in a multimedia application: The effects perceived in an educational context. In L. T. de Paolis & P. Bourdot (Eds.), *Augmented reality, virtual reality, and computer graphics* (Vol. 11614, pp. 348-367). Springer. https://doi.org/10.1007/978-3-030-25999-0_30

Denden, M., Tlili, A., Essalmi, F., & Jemni, M. (2018). Implicit modelling of learners' personalities in a game-based learning environment using their gaming behaviours. *Smart Learning Environments, 5*, Article 29. https://doi.org/10.1186/s40561-018-0078-6

Dubey, P., Pradhan, R. L., & Sahu, K. K. (2023). Underlying factors of student engagement to e-learning. *Journal of Research in Innovative Teaching and Learning, 16*(1), 17-36. https://doi.org/10.1108/JRIT-09-2022-0058

Feng, Z., González, V. A., Mutch, C., Amor, R., & Cabrera-Guerrero, G. (2023). Exploring spiral narratives with immediate feedback in immersive virtual reality serious games for earthquake emergency training. *Multimedia Tools and Applications, 82*(1), 125-147. https://doi.org/10.1007/s11042-022-13306-z

Fernandes, F., & Werner, C. (2019). Towards immersive learning in object-oriented paradigm: A preliminary study. *Proceedings of the 2019 21st Symposium on Virtual and Augmented Reality (SVR)*, 59-68. https://doi.org/10.1109/SVR.2019.00026

Fernandez, A., Insfran, E., & Abrahão, S. (2011). Usability evaluation methods for the web: A systematic mapping study. *Information and Software Technology, 53*(8), 789-817. https://doi.org/10.1016/j.infsof.2011.02.007

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (Advances in psychology, Vol. 52, pp. 139-183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Holz, H., Beuttler, B., Löfflad, D., & Ninaus, M. (2024). Developing a group-based literacy screening for German pre-readers: A digital, game-based approach. *Proceedings of the ACM on Human-Computer Interaction, 8*(MHCI), Article 251. https://doi.org/10.1145/3676496

Johnson-Glenberg, M. C., Ly, V., Su, M., Zavala, R. N., Bartolomeo, H., & Kalina, E. (2020). Embodied agentic STEM education: Effects of 3D VR compared to 2D PC. *Proceedings of the 6th International Conference of the Immersive Learning Research Network (iLRN)*, 24-30. https://doi.org/10.23919/iLRN47897.2020.9155155

Krivich, E., Mõttus, M., & Bauters, M. (2023). Comparing psychophysiological responses to vicarious pain experience elicited by videos featuring either humans or avatars. *ACM International Conference Proceeding Series*, 334-339. https://doi.org/10.1145/3616961.3617801

Lombana, D. A. B., Ventura, R. B., Chen, Y. T., & Porfiri, M. (2023). Educating youth about human impact on freshwater ecosystems using an online serious game. *IEEE Transactions on Games, 15*(4), 590-602. https://doi.org/10.1109/TG.2022.3185959

Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 43-71). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.005

Meena, Y. K., Cecotti, H., Wong-Lin, K. F., & Prasad, G. (2019). Design and evaluation of a time adaptive multimodal virtual keyboard. *Journal on Multimodal User Interfaces, 13*(4), 343-361. https://doi.org/10.1007/s12193-019-00293-z

Munsinger, B., & Quarles, J. (2019). Augmented reality for children in a confirmation task: Time, fatigue, and usability. *Proceedings of the ACM Symposium on Virtual Reality Software and Technology (VRST)*. https://doi.org/10.1145/3359996.3364274

Nie, J., Zhao, Y., Yao, B., Xu, Z., Chen, J., Yu, C., Hansen, P., Liu, J., Wang, J., Yan, G., Yao, C., Zhou, L., Ying, F., Liang, X., & Wang, G. (2022). SignFind: A synchronised sign language and Chinese character teaching game for Chinese deaf children using gesture recognition. *Proceedings of the CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3491101.3519697

O'Brien, H. L., Cairns, P., & Hall, M. (2018). A practical approach to measuring user engagement with the refined User Engagement Scale (UES) and new UES short form. *International Journal of Human-Computer Studies, 112*, 28-39. https://doi.org/10.1016/j.ijhcs.2018.01.004

Ronimus, M., Eklund, K., Pesu, L., & Lyytinen, H. (2019). Supporting struggling readers with digital game-based learning. *Educational Technology Research and Development, 67*(3), 639-663. https://doi.org/10.1007/s11423-019-09658-3

Rubegni, E., Gentile, V., Malizia, A., Sorce, S., & Kargas, N. (2022). Child–display interaction: Lessons learned on touchless avatar-based large display interfaces. *Personal and Ubiquitous Computing, 26*(3), 667-680. https://doi.org/10.1007/s00779-020-01451-x

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017). Design and evaluation of a short version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence, 4*(6), 103-108. https://doi.org/10.9781/ijimai.2017.09.001

Shao, B., Zhang, Y., Wang, Q., Xu, X., Zhou, Y., Feng, G., & Lv, F. (2021). StoryIcon: A prototype for visualising children's storytelling. *ACM International Conference Proceeding Series*, 152-157. https://doi.org/10.1145/3490355.3490377

Silva, L., Soares, F., Felix, J., Cardoso, L., Aranha, R. V., & Nascimento, T. H. (2024). Gesture recognition in smartwatches using LSTM for interaction in low-cost virtual environments. *ACM International Conference Proceeding Series*, 284-288. https://doi.org/10.1145/3691573.3691615

Tcha-Tokey, K., Christmann, O., Loup-Escande, E., & Richir, S. (2016). Proposition and validation of a questionnaire to measure the user experience in immersive virtual environments. *International Journal of Virtual Reality, 16*(1), 33-48. https://doi.org/10.20870/IJVR.2016.16.1.2880

Tisza, G., & Markopoulos, P. (2023). FunQ: Measuring the fun experience of a learning activity with adolescents. *Current Psychology, 42*, 1936-1956. https://doi.org/10.1007/s12144-021-01484-2

Torres, A., Kapralos, B., da Silva, C., Peisachovich, E., & Dubrowski, A. (2021). A scenario editor to create and modify virtual simulations and serious games for mental health education. *2021 12th International Conference on Information, Intelligence, Systems and Applications (IISA)*. https://doi.org/10.1109/IISA52424.2021.9555531

Vellingiri, S., McMahan, R. P., & Prabhakaran, B. (2020). SCeVE: A component-based framework to author mixed reality tours. *ACM Transactions on Multimedia Computing, Communications, and Applications, 16*(2), Article 52. https://doi.org/10.1145/3377353

Wan Husain, W. S., Wan Shamsuddin, S. N., & Rahim, N. (2022). A conceptual framework for road safety education using serious games with a gesture-based interaction approach. *Pertanika Journal of Science and Technology, 30*(1), 621-640. https://doi.org/10.47836/pjst.30.1.34

Xu, X., Li, J., Yuan, T., He, L., Liu, X., Yan, Y., Wang, Y., Shi, Y., Mankoff, J., & Dey, A. K. (2021). HulaMove: Using commodity IMU for waist interaction. *Proceedings of the CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445182

Ya'u, B. I., Salleh, N., Nordin, A., Alwan, A. A., Idris, N. B., & Abas, H. (2019). A systematic mapping study on cloud-based mobile application testing. *Journal of Information and Communication Technology, 18*(4), 485-527.

Yang, S., Gao, Z., Mogavi, R. H., Hui, P., & Braud, T. (2023). Tangible Web: An interactive immersion virtual reality creativity system that travels across reality. *Proceedings of the ACM Web Conference 2023*, 3915-3922. https://doi.org/10.1145/3543507.3587432

Zaina, L., Castro, E., Martinelli, S., & Sakata, T. (2019). Educational games and the new forms of interactions. *Smart Learning Environments, 6*, Article 12. https://doi.org/10.1186/s40561-019-0099-9

Pertanika J. Sci. & Technol. 34 (3): 1383 - 1407 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Utilising Innovative Image Preprocessing Techniques in Vehicle Plate Detection and Recognition

Soheir Noori [1]*, Asam Almohamed[2], Elham Mohammed Thabit A. Alsaadi[2], and Azhar Abbas[2]

[1]Department of Computer Science, College of Computer Science and Information Technology, University of Kerbala, 56001 Karbala, Iraq
[2]Department of Information Technology, College of Computer Science and Information Technology, University of Kerbala, 56001 Karbala, Iraq

ABSTRACT

Vehicle number is an essential issue for intelligent traffic monitoring systems. It plays a significant role in the transportation sector. In addition, recent population growth and human needs have increased due to the use of vehicles. Therefore, controlling traffic is becoming a complex problem to solve. An Automated Plate Number Recognition System (ANPR) program is needed for traffic control. This research includes developing and implementing a multi-stage preprocessing methodology with various techniques to identify vehicles' plate numbers. The detection system can achieve successful vehicle management: car parking as a case study. The proposed ANPR system employs a template matching algorithm based on a multi-stage preprocessing technique for character identification. Those techniques include noise filtering, component isolation, and correlation-based similarity measurements. The methodology process steps are grayscale conversion, binarisation, bilateral filtering, edge detection, contour localisation, segmentation using connected components, and character recognition through cross-correlation. The system was evaluated on 20 Iraqi one-line plates and achieved a recognition accuracy of 91%.

Keywords: Accuracy rate, detection and recognition phase, parking management system, plate number, preprocessing techniques

ARTICLE INFO

Article history:
Received: 27 May 2025
Accepted: 05 May 2026
Published: 12 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.05

E-mail addresses:
soheir.noori@uokerbala.edu.iq (Soheir Noori)
asam.h@uokerbala.edu.iq (Asam Almohamed)
elham.thabit@uokerbala.edu.iq (Elham Mohammed Thabit A. Alsaadi)
azhar.ali@uokerbala.edu.iq (Azhar Abbas)
* Corresponding author

INTRODUCTION

Intelligent transportation systems include plate recognition and intelligent infrastructure systems (Anagnostopoulos, 2014; Anagnostopoulos et al., 2006).

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

An essential requirement for observing and analysing transportation activity is the implementation of a monitoring system (Castello et al., 1999; Duan et al., 2004; Sarfraz et al., 2013). ANPR systems have different names due to reflecting different purposes, like mobile license plate reader (MLPR), vehicle recognition identification (VRI), license plate recognition (LPR), car plate recognition (CPR), and license plate reader (LPR), Highway road tolling systems (Al-Shemarry et al., 2022; Panahi & Gholampour, 2017; Song & Sarker, 2014), security systems (Al-Shemarry & Li, 2020) parking management systems, traffic management systems (Ojeniyi et al., 2019).

An Automatic Number Plate Recognition (ANPR) system relies on advanced image processing techniques to identify and extract license plate information from vehicle photographs. Thus, a plate number detection system is required to support the car parking system. It consists of four stages: input image preprocessing, localisation of the license plate, license plate segmentation, and recognition. Each car image carries a unique number. Thus, implementing an ANPR system can provide practical benefits for developing a parking management system. The research aims to develop a detection system to recognise car plate numbers from images. The proposed system was designed and tested on new English letter Iraqi oneline license plates. The Python programming language is used, and the OpenCV library for all image processing tasks. Localising car numbers is essential for intelligent monitoring systems (Angelova et al., 2015; Arafat et al., 2019; Azam & Gavrilova, 2017; Caplan et al., 2011; Wang et al., 2020; Yang et al., 2013; Yousef et al., 2015). The main objectives of this research are as follows:

1. Developing an ANPR software using template matching techniques.
2. Detecting unacceptable activities in transport systems to reduce the required efforts.
3. Increasing the accuracy of the existing ANPR system using Iraqi English car plate numbers.

Moreover, this research is essential to develop and measure the accuracy of the ANPR system. This is the central role of the detection system in controlling traffic flow and detecting vehicle plate numbers quickly, which is used for security purposes. The research outcome could further help detect or verify vehicle plate numbers. The system could also be beneficial for assisting law enforcement bodies, such as the traffic police, to detect and trace vehicle plate numbers easily. The developed ANPR system recognises vehicle plate numbers in good condition using one line or row of vehicle plate numbers.

Despite the wide adoption of ANPR systems, several challenges remain unresolved, particularly in real-world environments. Variations in lighting conditions, image quality, plate orientation, and background noise can significantly affect recognition performance. In addition, many existing approaches rely heavily on machine learning models that require large annotated datasets and high computational resources, which may not always

be available. In the context of Iraqi license plates, additional challenges arise due to variations in plate conditions, capture environments, and character similarities. These limitations motivate the need for a more efficient and robust approach that can operate under constrained conditions. Therefore, the main objective of this study is to develop a reliable ANPR system based on multi-stage preprocessing and deterministic template matching techniques. The proposed approach aims to enhance recognition accuracy while maintaining computational simplicity, making it suitable for practical deployment in parking management systems.

The Suggested System Components

The ANPR system software functions on a workstation and can establish connections with other applications that utilise the same databases. By employing multiple image enhancement algorithms, this process improves and standardises the input image containing the license plate number. Subsequently, it employs diverse extraction and classification approaches to extract and identify license plate information.

ANPR systems are often deployed using two main approaches. The first method enables real-time implementation of the whole license plate detection procedure. The second transmits each image the system camera captures across several channels to a remote computer. Next, the detection operation is carried out with a time delay, depending on the rate of the processing procedures utilised. The ANPR system must go through four main steps to recognise a license plate. These steps are stated below as shown in Figure 1.

Capturing Car Images

As a first step, photograph the car using a camera. The accuracy of any detecting system is determined by the quality of the camera, which encompasses its type, resolution, light, shutter speed, and placement technique. The distance between the car and the camera can sometimes cause issues, including low-quality or poor-quality photos.

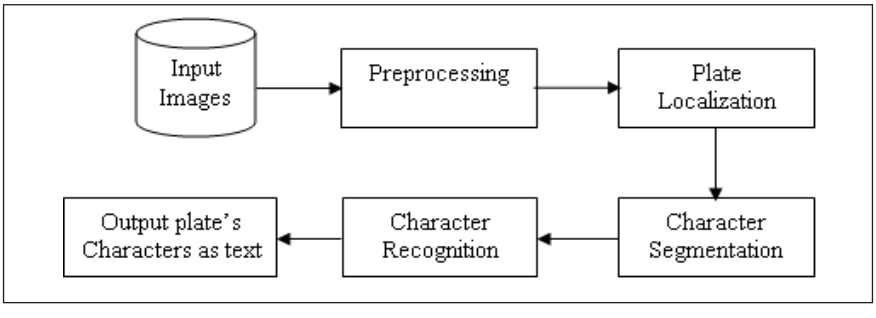


Figure 1. Steps to recognise a license plate in the ANPR system

The Preprocessing Stage

After the picture of the car is taken, more processing is done. Minimising the background noise of license plates improves the processing speed of succeeding stages by optimising the texture pattern. The automotive image undergoes several processing steps, including resolution expansion, elimination of noise, and colour transformation from RGB to monochrome or binary form. Various preprocessing algorithms are used for these tasks.

The Localisation and Extraction Stage

Regions of interest (ROIs) are taken from an automobile picture after preprocessing. This stage also has an impact on the precision of the ANPR system. The vehicle photograph undergoes central cropping and localisation, extracting all remaining pixels using multiple approaches. This will shorten the processing time of the extraction stage, mainly if those techniques use unsupervised learning algorithms (Ojeniyi et al., 2019). However, the license plate may appear anywhere in the picture. Consequently, specific characteristics such as boundaries, edges, colour, background, and textural elements impact the license plate extraction process.

The Classification and Detection Stage

The third step involves applying various supervised learning techniques to determine the number of license plate areas from the extracted data. The quality of the extraction techniques and preprocessing used to acquire the license plate areas will determine this stage (Abdulkadhim et al., 2024; Lee et al., 2013; Littlewort et al., 2004). Classifiers are used to create trained models and then to classify the pertinent extracted features from the license plate. The detectors could pick up several complex license plate details.

Several researchers have employed preprocessing techniques or unsupervised learning algorithms to detect multiple license plate issues. As a result, the detection system's processing time increases (Patel et al., 2013). The output of this stage functions as a decision mechanism to detect regions in an automobile image that correspond to license plates and ascertain their authenticity. The objective of utilising supervised learning algorithms during the testing phase is to achieve optimal system performance with the lowest execution time.

The Segmentation and Recognition Stage

The final stage of the ANPR system includes recognising the symbols retrieved from the observed license plate region. This stage employs various segmentation and recognition techniques, including template matching methods and classifiers such as artificial neural networks and fuzzy classifiers. In this paper, the license plate number

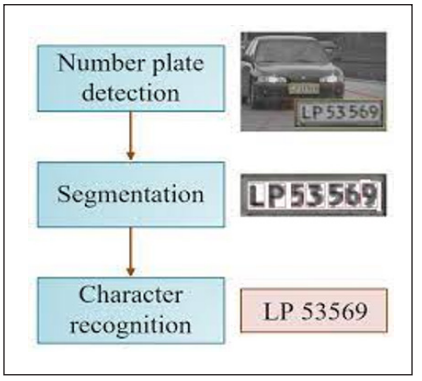


Figure 2. The architecture of an ANPR system

Figure 3. ANPR for parking management systems

is detected from automobile photographs using the template matching method. During this phase, the license plate number is converted into text encoded by a machine. In this case, the license plate image is used to identify the plate numbers using optical character recognition (OCR). The stage structure of the ANPR system is illustrated in Figure 2.

The detecting system's efficacy is contingent upon the stability and dependability of every single phase. To apply the license plate recognition technology for parking management systems, the primary goal of this research is to create techniques that detect license plates from automobile photos in a more accurate manner, as shown in Figure 3.

Image Manipulation Operations

Image Binarisation

A binary image is a computer image with each pixel having one of two possible values. Any two colours can be used, although black and white are the most common pairing for a binary image. The foreground colour in the picture is used for the objects, whereas the background colour is used for the remainder of the image. Bi-level or two-level images are other terms

for binary images. This indicates that a single bit (0 or 1) is recorded for every pixel. For this idea, the terms black-and-white, B&W, monochrome, or monochromatic are frequently used. Yet, they can also refer to any images with a single sample per pixel, such as grayscale images.

In digital image processing, binary pictures frequently appear as masks or as the outcome of specific processes like segmentation, thresholding, or dithering. Only some input/output devices, including fax machines, can handle bi-level images, laser printers, and bi-level computer displays. A purely binary image is defined as any combination of two colours where each pixel is represented by a single bit, either 0 or 1 or 0 and 1.

Image Segmentation

Segmentation is dividing a digital image into many segments in image processing. These segments are always called sets of pixels or super pixels. Segmentation is a technique that makes an image's representation simpler, more understandable, and easier to evaluate. Finding boundaries and objects in images, such as lines, curves, and so on, is usually accomplished using image segmentation. Image segmentation gives each pixel in an image a label, so those pixels share certain visual traits. A collection of contours retrieved from the image or segments that cover the entire image are the outcomes of image segmentation. Each of the pixels in a region is similar concerning some characteristic or computed property, such as colour, intensity, or texture.

Image Filtering

By comparing a pixel with its surrounding pixels in some way, images can be improved (or otherwise distorted, twisted, or mutilated) through image filtering. There are two kinds of picture filtering. First, height-based filtering, which is limited to images with indexed colours and those with 256 colours, is a typical example. Every colour has an index number between 0 and 255, which can be referred to as the "height" of the colour on a map. Second, colour-based filtering employs the actual red, green, and blue components of the colour for calculations, rather than the colour's index number. Both full-colour and 256-colour photos can be used with this.

Image Normalisation

Refining the characters into a block with no extra white space (pixels) on any of the four sides of the character is known as normalisation in image processing. After the image has been processed through the binarisation and segmentation processes, the normalisation procedure is applied if there are any remaining white spaces in the acquired image. Typically, there are more white than black spaces in the box following the binarisation and segmentation of a picture into many character blocks. In other words, normalisation

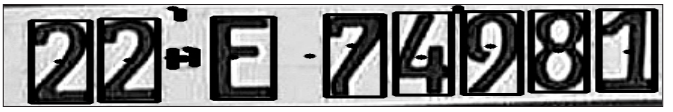


Figure 4. Equal-sized character with blocks after the normalisation process

removes all white space, cuts the character block to fit the character, and then applies this approach to ensure that every character is the same size. For every block, the characters are 36 × 18, much like in this ANPR study. The extracted character is then cut and matched with an equal size, as shown in Figure 4.

The Correlation Operation

One helpful method for recognising images is correlation. A correlation is a statistical method for determining the existence and strength of a relationship between two variables. For instance, there is a correlation between height and weight: taller people typically weigh more than shorter ones. The template matching process is the last step in the ANPR. During the character recognition stage, the accuracy of the result with the characters recorded in the database for this ANPR research is measured by comparing the similarity and identifying the best match using the template matching technique.

In image processing, the correlation algorithm is typically the most effective technique to employ at this stage. The cross-correlation and auto-correlation functions are the two correlation functions used in this study. The most effective method to gauge how similar two signals are is to use the cross-correlation function. The characters are then used to carry out the template matching procedure and compared with the characters in the database, as shown in Figure 5. This approach calculates the correlation coefficient between them using a set of known photos of the same size and some unknown photographs of a portion of an image with the highest correlation coefficient. There are two types of correlation.

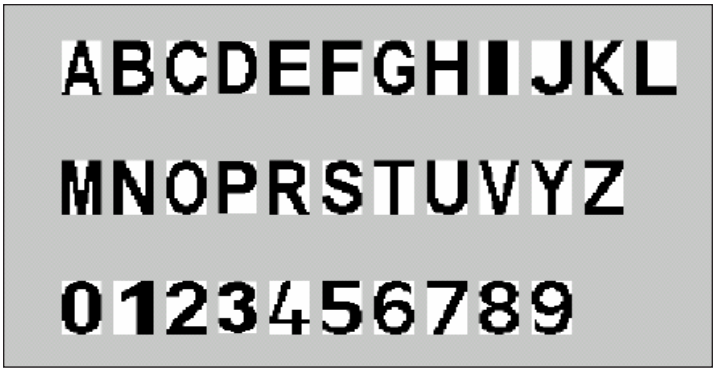


Figure 5. Database characters

The Auto-Correlation Function (ACF)

The cross-correlation of a signal with itself is known as autocorrelation. Simply put, it is the degree of resemblance between observations based on their time interval. It is a mathematical tool for locating recurring patterns, including a periodic signal hidden by noise, or identifying a missing fundamental frequency indicated by the signal's harmonic frequencies. In signal processing, it is frequently used to analyse functions or sequences of values, like time domain signals. The single-signal ACF yields data regarding the structure of the signal or its behaviour in the time domain.

The Cross-Correlation Function (CCF)

A measure of the similarities or shared characteristics between two signals is the CCF. Cross-correlation is a signal processing metric that quantifies how similar two waveforms are by applying a time lag to one of them. Another name for this is a sliding inner product or sliding dot product. It is frequently employed to look for a shorter, well-known component in a long-duration signal. It can also be used for electron tomographic averaging, single particle analysis, and pattern identification.

METHODOLOGY

This research aims to use Iraqi car license plates to test the effectiveness of the ANPR technology for parking management. The author covers every process step in this section, including data collection, specifications and special needs for creating ANPR, experimental procedures used, and anticipated results. As shown in Figure 6,

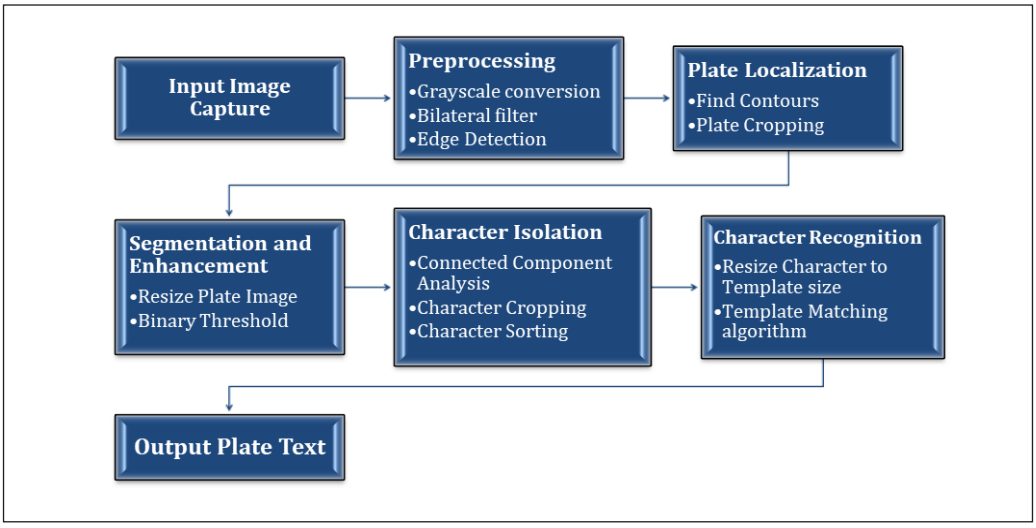


Figure 6. ANPR methodology

the approach diagram for the ANPR system has been created and designed in this stage. To construct and develop the ANPR algorithm, a thorough understanding of the fundamental principles involved in this research is necessary. This section will review several related approaches that can be used to create the ANPR algorithm for this study. Its three main stages are plate region extraction, image segmentation, and character identification.

Although the proposed system does not employ machine-learning classifiers, it incorporates an intelligent deterministic mechanism through template matching and correlation-based decision making. The system automatically filters noise, isolates connected components, and compares each extracted character against reference templates using normalised cross-correlation. This approach provides intelligent classification behaviour without the need for model training.

Capturing Car Images

In order to maintain the accuracy of the threshold values the system employs, a stationary camera placed on the parking station door will be used to record the input image. The distance between the camera and the automobile will be fixed and designated at the front of the station door. Since the suggested technique in this system will produce less accurate results if the input image is oriented, the collected image must be in a static, ideal orientation as shown in Figure 7.

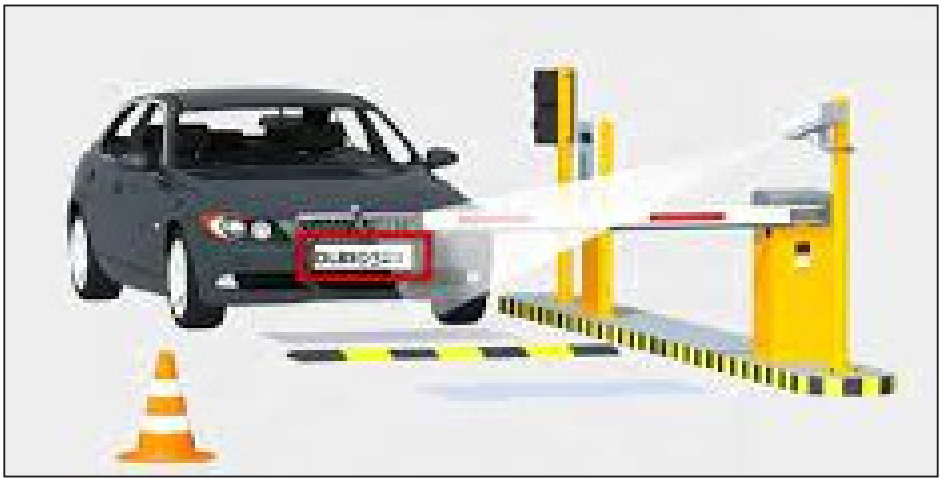


Figure 7. Image capturing stage

Car Image Preprocessing

At this stage, before developing another ANPR stage, a few processes should be implemented in the Iraqi plate region. Thus, several image preprocessing methods should be applied to the car image following the sequence given.

The Grayscale Process

The obtained RGB automobile image needs to go through the grayscale procedure, which turns the three-channel colour image into a grayscale format. This is necessary because, when subsequently converted to a binary scale, the grayscale image might sharpen and intensify every pixel. This grayscale process will use several image processing techniques to create the grayscale image.

The Binarisation Process

After being converted to grayscale, the car image must be converted into a binary form. By using the 0s to represent white pixels and the 1s to represent black pixels, the data from the grayscale format is binarised into a single black-and-white image. This should be created following the grayscale procedure to sharpen the image and improve the binarisation results.

The Car Image Filtering Process

Bilateral filtering is the next step applied to the binarised automobile image. It is used to reduce and eliminate undesirable noise and spots in the image. A specific parameter is used in the execution process to achieve the desired clearing image level. Following this procedure, the image ought to appear sharper and cleaner. Furthermore, it effectively reduces noise while maintaining sharp edges by ensuring that only nearby pixels with intensities comparable to the central pixel are considered for blurring. This is in contrast to the standard Gaussian filter, which ignores pixels that are nearly identical in intensity and thus blurs edges—something that we do not want to happen.

The Car Image Edge Detection Process

The edge detection process will start after the auto-mobile image filtering step. In this experiment, the image processing technique known as "edge detection" is used to determine the start and end locations of each character's white pixel to determine the character's horizontal and vertical bounds. This multi-step edge detection approach uses bilateral filtering and a Gaussian smoothing filter to denoise the image. However, better outcomes can be obtained by applying the Bilateral Filter before the Canny algorithm, particularly when the automobile image has a lot of noise or has poorly defined edges. As a result, everything will appear black except for the margins, which will appear white. In the upcoming processing stage, it will be helpful to find all closed shapes in the input image using the edges as shown in Figure 8.

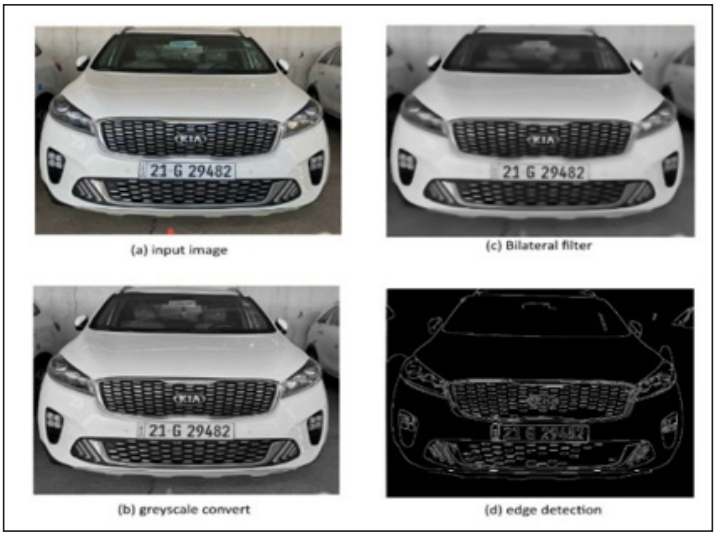


Figure 8. Preprocessing stages

Figure 9. Drawing found closed shapes

Plate Localisation

The plate image cropping procedure, where the phase one output result is displayed, is the most critical step in this initial phase. As previously mentioned, the objective of this procedure is to exhibit solely the plate region area containing the processed characters. Localising the license plate using the contours tool is the first step in cropping the image. With the parameters provided, it locates every closed shape in an input image as shown in Figure 9. Find Contours will use the output image from the clever algorithm as its input.

Furthermore, all of the forms discovered will be ordered from largest to smallest, selecting only the first 32 shapes because license plates are typically found in the range of 0 to 32. Any shape outside this range is too small to be considered a legitimate plate. The width, height, and x, y position data will be taken from each form to separate the actual

Figure 10. Bounding box for each character on the license plate number

plate shape from the others, and then filter the plate through a predetermined threshold. First, if the width is less than 420 pixels, and second, if the difference between the height and width/4 is less than 35 pixels, the threshold will be evaluated in both scenarios. The pixel value is then set in the function parameter, and the picture cropping procedure crops the area and shows it in its original colour. At the same time, the remaining portion of the image turns black as soon as the scanner detects the pixel region.

Segmentation and Enhancement

The goal of this phase is to group all chopped characters into blocks of uniform size. The procedure entails multiple critical operations and algorithms, including contour detection, edge detection, and character extraction. The intended outcome is that each character will fit within a specific size border box, as shown in Figure 10.

After resizing the license plate to make it larger, the noise produced by the resize function will be eliminated, and the image will be blurred again using the Gaussian filter. Next, thresholding techniques are used to accomplish plate image segmentation. Otsu's approach has been used to apply a threshold in order to convert the image to binary. The algorithm's output is a single intensity threshold that divides pixels into foreground and background classes. Because the license plate letters will noticeably stand out, higher recognition results will be obtained.

Moreover, the threshold has been reversed, making the text seem white and the background dark because Connected Components can only detect white pixels. The fitting approach technique will be applied to the character resizing in order to standardise the character size with a given parameter. In this experiment, every character must have a height and width of 42×24. To do this, the process will first resize every character to this size, and then the bounding process will be executed to fit the sized image inside each block container using an algorithm. The final step in this process involves extracting and displaying every character in each boundary box as shown in Figure 11.

Character Isolation

The connected component analysis function has been used to isolate characters in the license plate. It returns four parameters (*numLabels*, *labels*, *stats*, *centroids).*

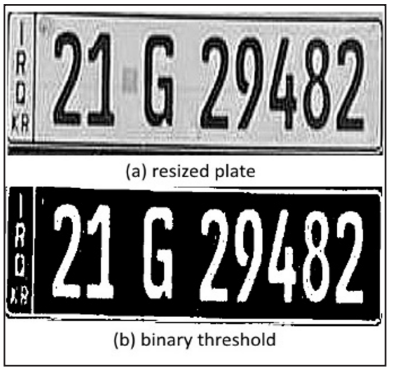


Figure 11. Plate a) resize and b) threshold

Figure 12. Black rectangles around each character

numLabels : The number of the connected components.
labels : The actual pixel data of the connected components.
stats : Various statistical data about each connected component (location, width, height, pixel area).
centroids : The *xy* centroids for each connected component.

All the connected components will first be looped through, excluding the first one, as the first connected component in a given image is the image itself (background), which will not be used here when looping. Threshold conditions will be applied to filter out the connected components that are the actual characters in the license plate. The threshold will be checked against data obtained from the stats variable that the connected component function returns. The exact threshold values are obtained by observing each connected component's data and a bit of trial and error. To ensure the thresholds are caught right, rectangles will also be drawn around each connected component representing a character as shown in Figure 12.

These thresholds are not foolproof; some may pass as valid character shapes. However, this issue will be solved using other thresholds in the character recognition stage. The connected components of the detected character may not be listed in the correct order as they appear (left to right in our case), so a sorting algorithm will be applied to sort the detected components from left to right using the component's X location. A connected

Figure 13. Segmented characters

component with an X value that goes into 0 and negative values should be sorted to the left of a component with an X value that goes into positive. Now, all characters are segmented and properties sorted as shown in Figure 13.

The numLabels parameter represents the total number of detected connected components in the image, including the background. This value helps determine how many potential objects (characters or noise regions) are present. The *centroids* parameter provides the geometric centre (x, y coordinates) of each detected component. These coordinates are useful for locating each character within the plate region and for maintaining their correct spatial order. By using centroid positions, the system ensures that characters are arranged from left to right before recognition, which is essential for accurate plate interpretation.

Character Recognition

The last step of the ANPR algorithm is the character recognition phase, which ought to be the most important since it compares the characters produced throughout the process and the characters saved in the database. This phase uses a template matching approach to compare each segmented character's result with the real character stored in the database. As shown in Figure 14, the templates include letters from A to Z and digits from 0 to 9 for 36 characters.

In the recognition stage, each segmented character is represented as a normalised binary matrix with fixed dimensions. This matrix acts as a digital encoding of the character shape, where pixel values are mapped into binary form (foreground and background). The recognition mechanism relies on comparing this encoded representation with a predefined set of template characters stored in the database. The matching process is

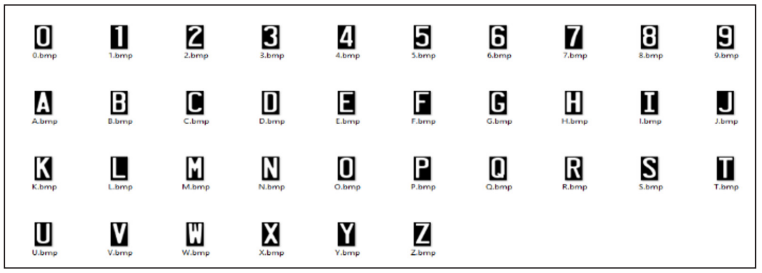


Figure 14. Template characters

performed using normalised cross-correlation, where each input character is evaluated against all templates, and the character with the highest similarity score is selected as the final output. This process can be interpreted as a deterministic encoding–matching mechanism rather than a learning-based approach.

For Arabic characters, encoding is more complex due to the presence of diacritics and connected structures. These features may lead to fragmented representations when using connected component analysis. Therefore, Arabic character encoding was not fully integrated into the current system. This limitation is acknowledged and will be addressed in future work by incorporating structure-aware segmentation and advanced feature representation techniques.

The primary strategy is using correlation and template matching techniques with a particular algorithm study.

The Cross-Correlation Process

In this experiment, the template matching process is the last step. To gauge how accurate the outcome is with the characters saved in the database for this ANPR research, the template matching approach is employed in this character recognition phase to compare the similarities and select the best match. The correlation algorithm is the optimum technique to utilise at this image processing stage. This ANPR research investigation uses two different correlation functions the cross-correlation and auto-correlation functions. The most effective method to gauge how similar two signals are is to use the cross-correlation function. Following segmentation, the characters are passed through a template matching procedure to be compared with the actual characters stored in the database. Subsequently, the result will demonstrate the precision and efficiency of the procedure. The input picture (segmented character) must be scaled to fit the template's dimensions before the matching process begins. The isolated character can then be recognised by comparing it to the characters in the template and determining which one has the most significant similarity. A 2-dimensional correlation approach is utilised to calculate the degree of similarity and identify the best match see Equation 1. The correlation between two characters that match precisely is equal to one.

$$R(X,Y) = \frac{\sum_{X'Y'}(T(X',Y').I(X+X',Y+Y'))}{\sqrt{\sum_{X',Y'} T(X',Y')^2 . \sum_{X',Y'} I(X+X',Y+Y')^2}} \quad [1]$$

R represents the correlation between our dataset's template image and the original car image. *T* is the template character image. *I* represents the image cropping so that focus is only on the middle part of the image. *X* is the template binary-level image. x' is the

average binary level in the template image. Y is the source image. y' is the average binary level in the source image.

Also, a threshold is employed only to return the best-matched characters that satisfy a correlation value > 0.6. This threshold was chosen from the observation that characters with a correlation value lower than this are not valid to begin with. They only passed the connected component filtering routine because they satisfied the width and height requirements. So even though they are considered characters, an empty character string will be returned when the threshold is not met to avoid outputting wrong results. Finally, the license plate number will be recognised and output on the screen shown in Figure 15.

Figure 15. Final recognition result for ANPR system

RESULTS AND DISCUSSIONS

This section explains the hardware and software requirements, the ANPR system database, the application interface, detection results with limitations for Arabic license plates, the accuracy of the ANPR system, and recognition results.

Hardware and Software Requirements

The hardware utilised for the implementation of this study consisted of a laptop:

RAM: 8GB Processor: Intel(R) Core(TM) i7-7600U CPU @ 2.80GHz 2.90 GHz
HDD: 256 and above

The development process of the Automatic Plate Number Recognition (APNR) system will incorporate the specified software standards, along with the following essential prerequisites:

- Python 3.10 and Open-CV 4.6.0.66 as the programming languages;
- Operating System (OS): Windows 10 Pro 64-bit.

The ANPR System Database

The essential data were obtained through collaboration with the Karbala Traffic Directorate to acquire photographs of the registered vehicles, while additional images of Iraqi cars were taken from the Internet. The total number of automobile photos is 20. Various criteria, such as dimensions and configuration, hue, layout, and additional mechanisms utilising alternative methods, are also included in this data collection and will be discussed as shown in Figure 16.

During the data-collecting phase, a multitude of approaches are employed, incorporating the following:

- The study explicitly includes 20 Iraqi license plate photographs or images, as determined by specific criteria.
- The file format in which the image was saved or exported, either as a JPG or JPG file. JPG format photographs have superior format compatibility with future computer languages and systems, making them the optimal, widely adopted, and most easily disposable option.

The dataset was collected under partially controlled conditions to ensure a reasonable level of consistency. Most images were captured using a fixed camera position and approximately the same distance from the vehicles. However, complete uniformity was not strictly enforced, and variations in lighting conditions, plate scale, and slight orientation differences are present. To address these variations, preprocessing steps such as normalisation, resizing, and binarisation were applied to standardise the input before segmentation and recognition. In particular, all extracted characters were resized into fixed dimensions, ensuring consistency during template matching regardless of the original plate size or font scale.

Figure 16. Examples of some Iraqi car images in the database

It is important to note that although Iraqi license plates generally follow a standard font and layout, minor differences may still exist due to real-world capture conditions. These variations may affect recognition performance and are considered a limitation of the current dataset. Future work will focus on constructing a larger and more diverse dataset with stricter acquisition conditions.

ANPR System Interface

This section provides a quick explanation of the application's interface. This interface was primarily created using a Python open-source GUI program. Instead of experimenting by executing each algorithm separately, the interface is designed to make running through the GUI easier and more intuitive. This system's graphical user interface (GUI) is displayed in Figure 17.

Perfectly Centred Image

As stated during the preprocessing stage, if the input image taken is not exactly centred, the template matching algorithm will yield less accurate results or fail altogether as shown in Figure 18.

Figure 17. The graphical user interface for the ANPR system

Figure 18. Perfectly centred for Iraqi car images

Arabic Letter Detection

We experimented with recognising Arabic-Iraqi plates before designing the system around new English Iraqi plates. However, a few issues prevented us from adding Arabic letter detection support.

Detecting Letter Elements Containing Specific Symbols

The Connected Component Analysis (CCA) procedure will detect the letter and the dot under it as separate entities, such as detecting the dot (قطقن) under the letter (ب). The authors experimented with solving this issue by detecting the letter and then setting an arbitrary Height and Width bounding box around it instead of setting the Height and Width for the connected component pixel bounds themselves, thus detecting both the letter and its component as a single entity.

A challenge arises when diacritical elements, such as dots, are detected independently from their corresponding characters. For example, the dot associated with the letter (ب) may be identified as a separate connected component rather than part of the main character. In such cases, the system may incorrectly interpret the dot as an independent symbol, leading to additional false detections. One possible approach is to assume that a detected dot belongs to a nearby character. However, this assumption is not always reliable, especially if the primary character is not detected correctly or appears fragmented. Additionally, using fixed bounding regions to group components may unintentionally include parts of neighbouring characters, which reduces recognition accuracy. These challenges highlight the limitations of using basic connected component analysis for Arabic script, where characters often consist of multiple interconnected elements, as shown in Figure 19.

Detecting the zero(رفص)

The zero character (number) in Arabic is smaller than the other characters, so much so that it doesn't pass our Width and Height thresholds for recognising a Connected Component as a valid character.

Figure 19. Arabic license plate connected component analysis result, showing the dot component of the (ب) letter being detected as a separate entity

The only way for it to pass the check is if the minimum Width and Height thresholds are set so low, but that also means some noise elements might pass as a valid zero (رفـص) character. It has been tried to apply aggressive filters to remove even more noise elements, but that resulted in chopping off pixels from valid characters in the license plate, which caused the Template Matching algorithm to have less accurate results.

Arabic letters present additional complexity due to connected shapes and the presence of diacritics. CCA often detects dots as separate entities and fails to group them with their parent letters. Additionally, characters such as the Arabic zero (٠) frequently fall below the minimum size thresholds. These issues limit the system's ability to support Arabic recognition. Future implementations should adopt CNN-based segmentation, dot-association heuristics, and morphological grouping methods to reliably process Arabic script.

Significance of the Accuracy Results

The achieved accuracy of 91% is considered significant for a template-matching-based ANPR system operating on real Iraqi plates. In addition to the reported accuracy, the system produced limited false positives (FP) and false negatives (FN) cases during character recognition. The FPs mainly occurred when visually similar characters such as (0/O/D) and (B/8) were incorrectly matched. While the FNs occurred when segmented characters failed to satisfy the predefined correlation threshold and were therefore rejected by the recognition stage. These cases were mainly caused by character similarity, image noise, and illumination variations.

As shown in the example in the following, the correct output data must have been divided by 100 to obtain the precise percentage for each process step.

$$\textit{The Extraction Phase}'\textit{s Correct Percentage}\ (\%) = \frac{\textit{Number of Correct Extraction Output}}{100\ (\textit{Total Input Data})}$$

$$\textit{The Extraction Phase}'\textit{s Correct Percentage}\ (\%) = \frac{\textit{Number of Correct Extraction Output}}{100\ (\textit{Total Input Data})}$$

In this study, 20 Iraqi car images were employed to extract license plate text, compare it against actual plate text, and then calculate the average result refers to Table 1. We can notice characters' similarity in English; some characters could look like others, which would produce less accurate results from the Template Matching algorithm. Examples are the character "0" being detected as "D", the character "B" being detected as "8", and vice versa.

Table 1
Recognition results

Plate Num.	Detected Text	Actual Text	Detection %
1.	21F24492	21F24492	100%
2	21G36271	21G36271	100%
3	22A85281	22A85281	100%
4	22*U*77*O*13	22D77013	75%
5	22K6291*2*	22K62917	87%
6	22G21234	22G21234	100%
7	22J58324	22J58324	100%
8	22G7*O*473	22G70473	87%
9	22J49*O*2*B*	22J49028	75%
10	22G21234	22G21234	100%
11	22G77442	22G77442	100%
12	22J79246	22J79246	100%
13	22G77442	22G77442	100%
14	22G7*O*473	22G70473	87%
15	22B*2*228*J*	22B77283	62%
16	22_427*3*2	22H42712	75%
17	21G29482	21G29482	100%
18	21J15443	21J15443	100%
19	21H12396	21H12396	100%
20	22B757*D*2	22B75702	87%
		Average: 91%	

ANPR Recognition Results

Figure 20 displays recognition results using the suggested ANPR method based on photos of Iraqi cars in the database. Even if some license plates had negative characteristics, it is evident that all license plates were identified. As discussed in the next section, certain characters on license plates are challenging to detect and identify due to their similarities.

The proposed ANPR system operates automatically during all detection and recognition stages without requiring human intervention in the final output. Human involvement was limited to recording and validating the actual license plate numbers during dataset preparation to establish the reference values used for performance evaluation.

As mentioned before, the first method in the preprocessing stage is to save and record the actual or original captured plate number from the car image. The captured image is saved in the computer database, and every plate number of the car image will be recorded and saved in our database for future use. Let's say the first captured plate number image

Figure 20. ANPR system recognition results

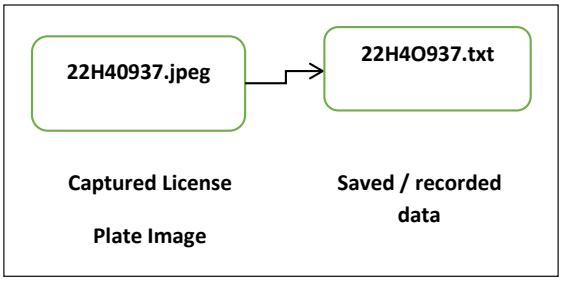


Figure 21. Plat number data

22H40937; the file for 22H40937.jpeg will be saved, and the data of the image, which is the plate number (22H40937), will be recorded and saved by using the manual method, either handwriting or typing, to be kept in a database that will be designed later in ".txt" format as shown in Figure 21.

The ANPR system does not adjust for the different levels attained. The plate region image itself is the root cause of it. To achieve ideal outcomes, all aspects and characteristics of the captured photos in the dataset must be appropriately included in a solid and perfect criterion. Different levels of accuracy will come from variations in the brightness, contrast, distance, and pixel value of the image. Since each image is taken by hand, it may appear challenging to standardise its properties. However, the varying degrees of outcomes will guide us to the subsequent section on suggestions and future endeavours.

CONCLUSION

Moreover, future work should focus on expanding the dataset and integrating deep-learning-based OCR models such as CNNs, CRNNs, and Transformer-based

architectures. These models are expected to improve the recognition of visually similar characters and reduce segmentation errors. Furthermore, future developments should include dedicated Arabic character processing modules that address dot-association, adaptive thresholding, and morphological reconstruction to support both Arabic and English Iraqi plates with higher robustness.

ACKNOWLEDGEMENT

The authors thank the Karbala Traffic Directorate and Dr Husam Hasan Kadhim at the University of Kerbala for helping with proofreading.

Data Availability Statement

The written work involves data.

Conflicts of Interest

The authors assert that they have no conflicts of interest.

Funding

This research received no external funding.

REFERENCES

Abdulkadhim, E. G., Al-Shemarry, M. S., & Alsaadi, E. M. T. A. (2024). An efficient algorithm for covert contacting in IoT. *AIP Conference Proceedings, 3094*(1), Article 040008. https://doi.org/10.1063/5.0209934

Al-Shemarry, M. S., & Li, Y. (2020). Developing learning-based preprocessing methods for detecting complicated vehicle licence plates. *IEEE Access, 8*, 170951-170966. https://doi.org/10.1109/ACCESS.2020.3024625

Al-Shemarry, M. S., Li, Y., & Abdulla, S. (2023). Identifying license plates in distorted vehicle images: Detecting distorted vehicle licence plates using novel preprocessing methods with hybrid feature descriptors. *IEEE Intelligent Transportation Systems Magazine, 15*(2), 6-25. https://doi.org/10.1109/MITS.2022.3210226

Anagnostopoulos, C.-N. E. (2014). License plate recognition: A brief tutorial. *IEEE Intelligent Transportation Systems Magazine, 6*(1), 59-67. https://doi.org/10.1109/MITS.2013.2292652

Anagnostopoulos, C. N. E., Anagnostopoulos, I. E., Loumos, V., & Kayafas, E. (2006). A license plate-recognition algorithm for intelligent transportation system applications. *IEEE Transactions on Intelligent Transportation Systems, 7*(3), 377-392. https://doi.org/10.1109/TITS.2006.880641

Angelova, A., Krizhevsky, A., Vanhoucke, V., Ogale, A. S., & Ferguson, D. (2015). Real-time pedestrian detection with deep network cascades. In *Proceedings of the British Machine Vision Conference 2015* (pp. 32.1-32.12). BMVA Press. https://doi.org/10.5244/C.29.32

Arafat, M. Y., Khairuddin, A. S. M., Khairuddin, U., & Paramesran, R. (2019). Systematic review on vehicular licence plate recognition framework in intelligent transport systems. *IET Intelligent Transport Systems, 13*(5), 745-755. https://doi.org/10.1049/iet-its.2018.5151

Azam, S., & Gavrilova, M. (2017). License plate image patch filtering using HOG descriptor and bio-inspired optimisation. In *Proceedings of the Computer Graphics International Conference 2017* (Article 18). Association for Computing Machinery. https://doi.org/10.1145/3095140.3095141

Caplan, J. M., Kennedy, L. W., & Petrossian, G. (2011). Police-monitored CCTV cameras in Newark, NJ: A quasi-experimental test of crime deterrence. *Journal of Experimental Criminology, 7*(3), 255-274. https://doi.org/10.1007/s11292-011-9125-9

Castello, P., Coelho, C., Del Ninno, E., Ottaviani, E., & Zanini, M. (1999). Traffic monitoring in motorways by real-time number plate recognition. In *Proceedings of the 10th International Conference on Image Analysis and Processing* (pp. 1199-1203). IEEE. https://doi.org/10.1109/ICIAP.1999.797765

Duan, T. D., Duc, D. A., & Du, T. L. H. (2004). Combining Hough transform and contour algorithm for detecting vehicles' license plates. In *Proceedings of the 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing* (pp. 747-750). IEEE. https://doi.org/10.1109/ISIMP.2004.1434172

Lee, Y., Han, D. K., & Ko, H. (2013). Reinforced AdaBoost learning for object detection with local pattern representations. *The Scientific World Journal, 2013*, Article 153465. https://doi.org/10.1155/2013/153465

Littlewort, G., Bartlett, M. S., Fasel, I., Susskind, J., & Movellan, J. (2004). Dynamics of facial expression extracted automatically from video. In *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop* (pp. 80-80). IEEE. https://doi.org/10.1109/CVPR.2004.327

Ojeniyi, A., Aro, T., & Thiak, A. M. (2019). Design of an agent-based traffic control system. *International Journal of Engineering and Advanced Technology, 8*(6S3), 71-76. https://doi.org/10.35940/ijeat.F1012.0986S319

Panahi, R., & Gholampour, I. (2017). Accurate detection and recognition of dirty vehicle plate numbers for high-speed applications. *IEEE Transactions on Intelligent Transportation Systems, 18*(4), 767-779. https://doi.org/10.1109/TITS.2016.2586520

Patel, C., Shah, D., & Patel, A. (2013). Automatic number plate recognition system (ANPR): A survey. *International Journal of Computer Applications, 69*(9), 38-42. https://doi.org/10.5120/11871-7665

Sarfraz, M. S., Shahzad, A., Elahi, M. A., Fraz, M., Zafar, I., & Edirisinghe, E. A. (2013). Real-time automatic license plate recognition for CCTV forensic applications. *Journal of Real-Time Image Processing, 8*(3), 285-295.https://doi.org/10.1007/s11554-011-0232-7

Song, M. K., & Sarker, M. M. K. (2014). Modelling and implementing two-stage AdaBoost for real-time vehicle license plate detection. *Journal of Applied Mathematics, 2014*, Article 697658.https://doi.org/10.1155/2014/697658

Wang, D., Tian, Y., Geng, W., Zhao, L., & Gong, C. (2020). LPR-Net: Recognising Chinese license plates in complex environments. *Pattern Recognition Letters, 130*, 148-156. https://doi.org/10.1016/j.patrec.2018.09.026

Yang, Y., Wang, Z. F., Yu, B. C., & Yin, Z. H. (2013). Vehicle license plate location based on high-frequency vertical edge and mathematical morphology. *Advanced Materials Research, 734-737*, 3252-3255. https://doi.org/10.4028/www.scientific.net/AMR.734-737.3252

Yousef, K. M. A., Al-Tabanjah, M., Hudaib, E., & Ikrai, M. (2015). SIFT based automatic number plate recognition. In *2015 6th International Conference on Information and Communication Systems (ICICS)* (pp. 124-129). IEEE. https://doi.org/10.1109/IACS.2015.7103214

Pertanika J. Sci. & Technol. 34 (3): 1409 - 1432 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Stability, Thermo-physical, and Tribological Properties of SiO_2 Nanoparticles in a Vegetable Oil-based Biodegradable Lubricant for Minimum Quantity Lubrication Machining

Nurlisa Hamzan[1], Muhammad Amir Syafiq Mohd Tarmizi[1], Nurul Huda Abdul Azid[1], Ahsana Aqilah Ahmad[1], Afifah Zakiyyah Juri[2], and Irnie Azlin Zakaria[1]*

[1]Faculty of Mechanical Engineering, Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Selangor, Malaysia
[2]Department of Mechanical and Manufacturing Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia

ABSTRACT

Machining is essential in manufacturing for shaping materials by removing excess through contact between the tool and workpiece. However, dry machining generates significant heat and cutting forces, leading to tool degradation and surface imperfections. While flood cooling helps address these issues, it presents environmental and health concerns due to large volumes of fluid and harmful chemical additives. Minimum Quantity Lubrication (MQL) offers a greener alternative by applying only the minimum amount of lubricant directly to the cutting area. Still, it frequently proves insufficient in handling difficult-to-cut materials like nickel-based alloys, owing to poor heat removal and lubrication. This study investigates enhancing MQL using silicon dioxide (SiO_2) nanoparticles in vegetable oil-based formulations, creating an improved nanolubricant system (NMQL). The volume concentrations of five different percentages were tested: 0.01%, 0.03%, 0.05%, 0.07%, and 0.1 % of SiO_2 biodegradable nanolubricant. Stability was evaluated through visual sedimentation observation, UV-visible measurements, and zeta potential analysis. Some slight sedimentation occurrences through visual observations indicated nanolubricant stability even after 10 days. UV-visible stability was measured at a 0.1% volume concentration, showing a linear correlation adhering to Beer-Lambert law principles. The most stable nanolubricant resulted from 2 hours of sonication at a 349 nm wavelength, exhibiting an excellent zeta potential stability of 98.3 mV. Thermophysical testing revealed that 0.1% concentration produced the highest dynamic viscosity compared to the base oil.

ARTICLE INFO

Article history:
Received: 13 February 2025
Accepted: 15 June 2025
Published: 12 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.06

E-mail addresses:
nurlisahamzan00@gmail.com (Nurlisa Hamzan)
irnieazlin@uitm.edu.my (Irnie Azlin Zakaria)
amirsyafiqtarmizi@gmail.com (Muhammad Amir Syafiq Mohd Tarmizi)
huda.aazid@gmail.com (Nurul Huda Abdul Azid)
aqilah3739@uitm.edu.my (Ahsana Aqilah Ahmad)
afifahjuri@ukm.edu.my (Afifah Zakiyyah Juri)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

In the tribological evaluation, the 0.07% concentration exhibited the lowest coefficient of friction and wear scar diameter. Overall, biodegradable nanolubricants serve as a potential for Minimum Quantity Lubrication (MQL) machining.

Keywords: Dynamic viscosity, MQL, nanolubricant, SiO_2, tribology

INTRODUCTION

The significance of sustainability in machining processes becomes evident due to the large-scale production involved and the associated environmental risks. Machining operations play a crucial role in every manufacturing industry. Machining involves a method that achieves the anticipated size, shape, and quality on the surface by removing excess material from the work specimen through direct contact of the machining tool with the work specimen (Roy et al., 2019). Therefore, it was crucial to use a coolant to dissipate the heat generated in the machining area and control the machining temperature throughout the operation. A substantial amount of heat in the cutting area and usage of high cutting forces were generated from the dry machining of metals, resulting in tool wear, workpiece hardening, and intensified surface roughness (Pawanr et al., 2024). To mitigate these effects, flood cooling techniques are commonly used for lubrication and cooling during machining. However, flood cooling faced challenges due to environmental issues, health risks, and long-term viability concerns related to flood cooling systems. These systems used a significant amount of coolant or lubricant, impacting the environment. Additionally, the coolant used might contain toxic additives, which might affect the storage, usage, and disposal of traditional lubricants that contribute to the hazardous environmental impact. To address these concerns, a new cooling method, minimum quantity lubrication (MQL), was developed (Kapil et al., 2024). The MQL involves spraying only a small amount of coolant directly at the cutting zone, minimising mist formation and reducing health risks associated with conventional flood cooling systems. The MQL represented a lubrication methodology designed to optimise machining operations. In MQL, only a minimal quantity of lubricant was applied directly to the cutting tool or workpiece throughout the machining handling (Said et al., 2019). The MQL technique was an implementation that reassigned the least volume of cutting oil to the cutting area, forming an oil mist through its combination with compressed air.

According to the research by Ahmad et al. (2022) on treated recycled cooking oil (TRCO) in machining alloy steel (AISI 4340) in MQL, the machining performance was improved, resulting in reduced cutting force. MQL also led to lower cutting forces compared to dry cutting without any lubrication and improved the surface finish of the machined workpiece. The chips produced during MQL were thinner and had a curlier shape, indicating reduced friction during cutting. However, Gupta et al. (2021) discovered

that the effectiveness of conventional MQL in easing cutting forces, surface roughness, and tool wear is limited. Additionally, conventional lubricants in MQL may not adequately address the heat produced throughout the machining process of hard-to-cut metals such as Nickel-based alloys that have a remarkable mechanical and chemical quality, which is suitable for high-strength, high-temperature, and pressure applications. However, its high work-hardening properties and low heat conductivity make it difficult to machine at high cutting speeds, including in the MQL environment. MQL systems often face challenges in effectively dissipating heat, which can adversely affect the surface finish and dimensional accuracy of machined components. The restricted volume of lubricant in MQL systems can complicate chip removal from the cutting zone, potentially leading to damage to both the workpiece and the cutting tool (Gupta et al., 2021). Thus, nanolubricant seems to be a feasible solution for the challenge.

Nanolubricants consist of nanoparticles dispersed in a lubricating base fluid, exhibiting unique properties that enhance the lubricating capabilities. The use of nanolubricants significantly improved the overall efficiency of the machining process. The nanolubricant enhanced the standard base liquid related to the effective thermal conductivity and viscosity (Hamisa et al., 2023). This targeted application ensures that lubrication is provided sufficiently, enhancing the efficiency of the machining operation. The usage of nanolubricants in Minimum Quantity Lubrication (NMQL) systems represented a perfect approach to machining, offering superior lubrication where it was needed most. This not only enhanced operational efficiency but also aligned with environmentally conscious practices, making it a promising advancement in the field of manufacturing. By introducing nanolubricant into NMQL technology, the machining performance will be enhanced in terms of machining characteristics as compared to conventional flood and dry cutting methods. The improved execution of the NMQL cooling strategy can be attributed to two key elements. First, the inclusion of highly conductive nanoparticles increases the thermal conductivity of the base oil. This results in a higher heat removal rate during machining. Secondly, the rolling impact of nanoparticles at the tool-chip interface eases the friction (Nouzil et al., 2022). The increased nanoparticle concentration also creates a preservative layer on both the tool and the workpiece surface, thus improving the interaction between the tool and workpiece and effectively easing friction (Moretti et al., 2024).

Research by Eltaggaz et al. (2018) explored the aluminium oxide nanoparticles with a nanocrystalline structure in the air-oil mixture during the turning process of austempered ductile iron (ADI). The NMQL outperformed conventional MQL by reducing flank wear. A comparative study evaluating different cooling approaches (dry, flood, MQL, and NMQL) revealed that NMQL machining was superior to base MQL and commensurate with flood cooling (Eltaggaz et al., 2018). The authors extended their investigation to titanium machining with Alumina (Al_2O_3) nanoparticles and discovered that NMQL

reduced the seizure zone as opposed to base MQL cooling (Eltaggaz et al., 2021). Moreover, they observed that higher nanoparticle concentrations positively influenced tool durability and surface finish attributes. Furthermore, the implementation of nanolubricants in MQL systems aligns with environmental sustainability goals, with the reduced volume of lubricant used, coupled with the improved lubricating properties of nanomaterials. This relationship between MQL and nanolubricants not only improved machining efficiency but also helped in eco-friendly manufacturing practices (Lim et al., 2022).

The addition of nanoparticles greatly boosts the thermal conductivity and dynamic viscosity of lubricants, leading to more effective cooling performance and better lubricity during machining processes. Studies conducted by (Lim et al., 2022; Mondragón et al., 2023), explored a range of nanomaterials for minimum quantity lubrication (MQL) methods. Their investigations covered materials such as hybrid blends of graphene and aluminium oxide (G-Al_2O_3), along with other nanoparticles like silicon dioxide (SiO_2), molybdenum disulfide (MoS_2), and pure graphene. These works revealed that nanolubricants offer enhanced heat dissipation and retain consistent viscosity even at minimal concentrations. From a tribological standpoint, they also contribute significantly to lower friction and reduced surface wear during machining. In the study by Mohd Khalil et al. (2018) The impact of different machining parameters on cutting force and tool degradation was assessed during MQL machining of nickel superalloys with nanolubricants. Dry cutting was used as a benchmark for comparison. Abrasion was found to be the leading cause of tool wear. MQL with nanolubricants proved more effective in minimising wear and reducing cutting force. According to Gutnichenko et al. (2018) Incorporating graphite nanoparticles into vegetable oil for machining alloy 718 led to notable improvements over dry and conventional MQL machining, particularly in reducing tool wear, surface roughness, cutting force, and vibration.

Among the various nanoparticles evaluated, silicon dioxide (SiO_2) stands out for its excellent heat transfer and friction-reducing characteristics (Hawwash et al., 2018). Incorporating SiO_2 into eco-friendly lubricants appears to be a promising strategy to elevate both thermal and mechanical performance. These nanoparticles offer advantages such as remarkable thermal endurance, robust mechanical integrity, high tensile strength, and rigidity. This advancement supports the objectives of sustainable manufacturing by enhancing process efficiency while minimising the environmental footprint of lubrication. SiO_2 was deliberately selected as a fundamental component in nanolubricant formulations due to its beneficial properties. As highlighted by Sia et al. (2014), SiO_2 nanoparticles are well known for their exceptional hardness and their critical function in minimising direct surface interaction in friction-intensive conditions. Additionally, their widespread availability in the market played a significant role in enabling the cost-effective commercialisation of research. Because of their affordability and seamless compatibility with oil matrices,

silica nanoparticles have been recognised as valuable additives for lubricants. Dambatta et al. (2019) emphasised that incorporating SiO_2 nanoparticles enhances oil properties, attributing this effect to their unique spherical shape and the complex 3D network formed by surface molecules. The presence of unoccupied bonds in SiO_2 molecules results in high surface energy, effectively reducing friction through nanoparticle deposition on workpiece surfaces. Furthermore, studies reveal that when dispersed in base oils, SiO_2 nanoparticles exhibit remarkable diffusion and mixing capabilities. A cost-benefit analysis comparing traditional synthetic lubricants with SiO_2-based alternatives indicates potential long-term savings due to extended tool lifespan and reduced maintenance costs. In addition to this, the cost of SiO_2 is relatively competitive nowadays due to its widespread applications. The addition of a small amount of nanoparticles to the base fluids is still justified in the technoeconomic study (Czaplicka et al., 2021). Regulatory adaptation and industry adoption will be key to broader commercial use.

This research introduces a novel perspective by conducting an in-depth investigation into SiO_2 nanolubricant in biodegradable oil, as this area is still limited in the open literature. Garcia Tobar et al. (2024) provided a review of nano-additives used in lubrication studies. Their review identified Molybdenum Disulfide (MoS_2) and Zinc Oxide (ZnO) as the most frequently studied nano-additives, with 10% of the research studied. Graphene (Gr) and Aluminum Oxide/Titanium Dioxide (Al_2O_3/TiO_2) were close behind, contributing to 8% of studies. Interestingly, the review highlighted a lack of focus on SiO_2 as a nanolubricant, particularly regarding its tribological behaviour, which is a research gap that is addressed in this study.

This study encompasses a rigorous stability assessment of the prepared samples before evaluating their thermo-physical properties, including dynamic viscosity in SiO_2 biodegradable nanolubricants. Following this, tribological performance was examined through a four-ball friction test, demonstrating enhancements in friction coefficient and wear scar diameter. Additionally, the correlation between these parameters and nanoparticle concentration was analysed.

Referring to Garcia Tobar et al. (2024) who reviewed optimal concentration ranges in nanolubricants, 51.28% of researchers found that concentrations below 1% yielded superior lubrication efficiency. Accordingly, this study stands as a pioneering effort in comprehensively examining SiO_2 nanoparticles dispersed in biodegradable oil across volume concentrations of 0.01%, 0.03%, 0.05%, 0.07%, and 0.1%. All experimental findings have been meticulously validated against previous research to ensure their reliability. Ultimately, the study proposes the feasible volume concentration of SiO_2 nanoparticles in biodegradable oil, with recommendations for further applications in minimum quantity lubrication (MQL) machining.

METHODOLOGY

Nanolubricant Preparation

In this experimental investigation, SiO_2 nanoparticles were employed as the nanoparticles due to their superior performance and easily accessible. These SiO_2 nanoparticles, sourced from Xuzhou Jiechuang New Material Technology, have a typical size of 30 nm and are 99% pure SiO_2. The biodegradable oil utilised in the study was Coolube 2210 XP, developed by Unist, Inc. (2025), which is derived from pure vegetable oils and designed specifically for use with ferrous materials, as it is considered eco-friendly, which excludes any volatile organic compounds (VOCs). Table 1 shows the physical properties of nanoparticles, while Table 2 represents the physical and chemical properties of Coolube 2210 XP. The preparation of the SiO_2 biodegradable nanolubricant involved SiO_2 nanoparticles and the biodegradable oil through a two-step method Figure 1 (Nugroho et al., 2021). This approach is recommended for its ability to ensure proper mixing and particle stabilisation. In this experiment, 0.5% volume of Triton X-114 surfactant (CAS Number 9036-19-5) was incorporated into every concentration of the nanolubricant to increase the oil stabilisation (Merck, 2025). Initially, a 30-minute magnetic stirring process was performed, followed by continuous dispersion using an ultrasonic bath vibrator for an additional two hours to break down agglomerations (Nugroho et al., 2021). The two-hour sonication process was identified as the optimal duration for achieving effective mixing, which minimises the potential for sedimentation and attaining a desirable mean particle size. The sonicator plays a decisive part in breaking down agglomerations and ensuring thorough dispersion of the nanoparticles within the base solution (Sharif et al., 2017). Sharif et al. (2022) suggested the use of mono and hybrid nanolubricants at concentrations under 0.1%, due to their remarkable tribological performance. Guided by this recommendation, the current research formulated nanolubricants using low volume percentages of 0.01%, 0.03%, 0.05%, 0.07%, and 0.1% nanolubricants as part of the research.

Table 1
Properties of SiO_2

Properties	SiO_2
Chemical Abstracts Service (CAS) Number	7631-86-9
Purity (%)	99
Diameter (nm)	30
Density @ 20°C (kg/m^3)	2329
Melting Point (°C)	1440
Specific heat (J/kg.K)	745

Note. Adapted from Zawawi et al., 2019; ChemicalBook, 2025

Table 2
Properties of biodegradable oil

Chemical Name/Properties	Coolube 2210 XP
Methyl Ester -CAS No. 67784-80-9	<2%
Diisotridecyl Adipate -CAS No. 26401-35-4	<2%
Mixed Esters/Ester Blend -Proprietary	>90%
Registered Sulfurised Olefin -Proprietary	<5%
Viscosity @ 40 °C mm²/s (cSt)	10
Pour point (°C)	-12 to -20
Density @ 20 °C (kg/m³)	890
Flash Point (°C)	>200
Thermal Conductivity, k (W/mK)	0.2

Note. Adapted from Unist, 2025

The volume concentration for the experimental phase of the research was determined using the formula outlined in Equation 1 (Lim et al., 2022).

$$\phi_{NL} = \frac{\frac{m_p}{\rho_p}}{(\frac{m_p}{\rho_p} + Voil)} \times 100 \quad [1]$$

Where ϕ_{NL} is the volume concentration of nanolubricant, m_p is the mass of nanoparticles in kg, $Voil$ is the volume of biodegradable oil in m³, and ρ_p is the density of nanoparticles in kg/m³.

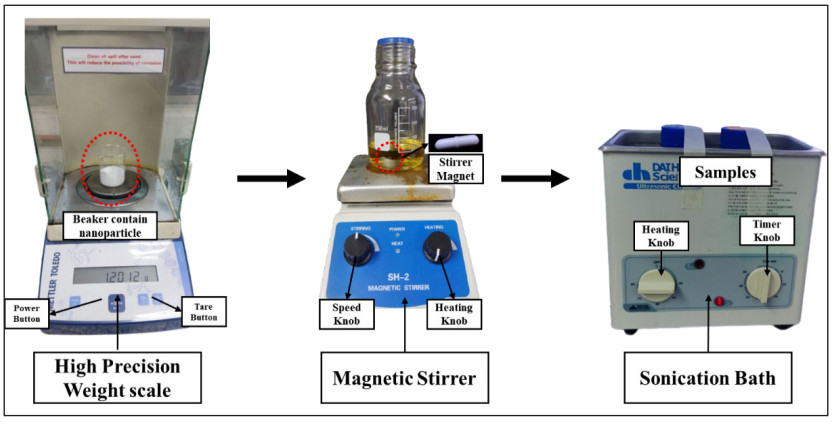


Figure 1. Two-step process in the preparation of a nanolubricant

Nanolubricant Stability

Visual Sedimentation Observation

Dispersion stability was assessed using three methods: visual sedimentation, ultraviolet-visible (UV-Vis), and zeta potential analysis. In the visual sedimentation, periodic observations were made up to a 30-day observation period to assess the nanolubricant's stability in the samples visually (Gulzar et al., 2017).

Spectra Absorption

Additionally, the dispersion stability of the SiO_2 biodegradable nanolubricant was also analysed through spectral absorbance using UV-Visible spectroscopy Jasco/V 670-EX, a highly effective technique for studying nanolubricant stability, as shown in Figure 2. The evaluation of the stability of dispersion was conducted at room temperature by quantifying the absorbance (from the 0th day to the 20th day), with variations in sonication time. The wavelength accuracy falls within 1.0 nm across the 200 to 450 nm range and was set to identify the peak absorption wavelength for a concentration of 0.1%. Given the pivotal role of UV-visible in assessing nanolubricant stability, precision becomes a significant consideration in this stability test (Hamisa et al., 2023). Absorbance indicated how much light the nanolubricant absorbed, while transmittance indicated how much light passed through it. The visible UV-visible provided the necessary light spectrum. Then, a lens allowed a beam of light to pass through a monochrome, breaking it into a spectrum of different wavelengths (Ismail et al., 2023). To enhance accuracy and reduce random error, measurements using the UV-Vis spectrophotometer were repeated three times and averaged to ensure precision.

Zeta Potential Measurement

The final stability measurements in this study were conducted using the Zeta potential test. The zeta potential refers to the developing difference between the dispersion medium and the static layer of solution bonded to the particle (Hamisa et al., 2023; Ismail et al., 2022). The Malvern Zetasiser ZEN3600 machine, as shown in Figure 3 was employed for these zeta measurements, with nanolubricants placed in the DTS1070 cell. The required parameters that were specified in the machine were the refractive index, viscosity, absorption, and dielectric constant for the base oil and nanolubricant. The zeta potential of the lowest nanolubricant, which was 0.01% volume concentration, was measured and conducted under controlled conditions. To ensure dependable and precise outcomes, the zeta potential test was conducted three times, aiming to reduce fluctuations and verify consistency across repeated measurements.

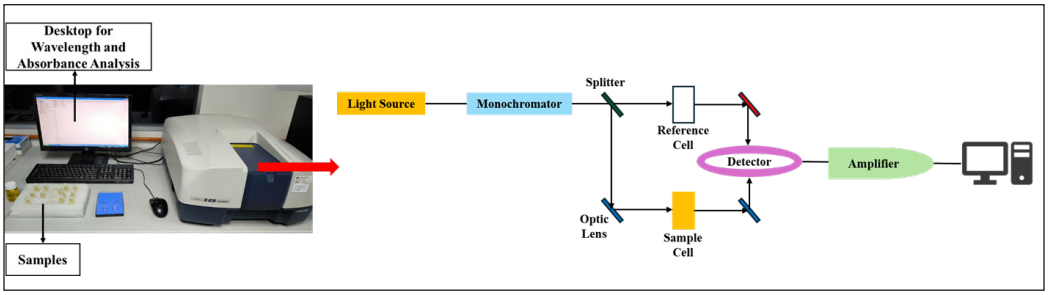


Figure 2. UV-Visible spectrometer and its schematic configuration

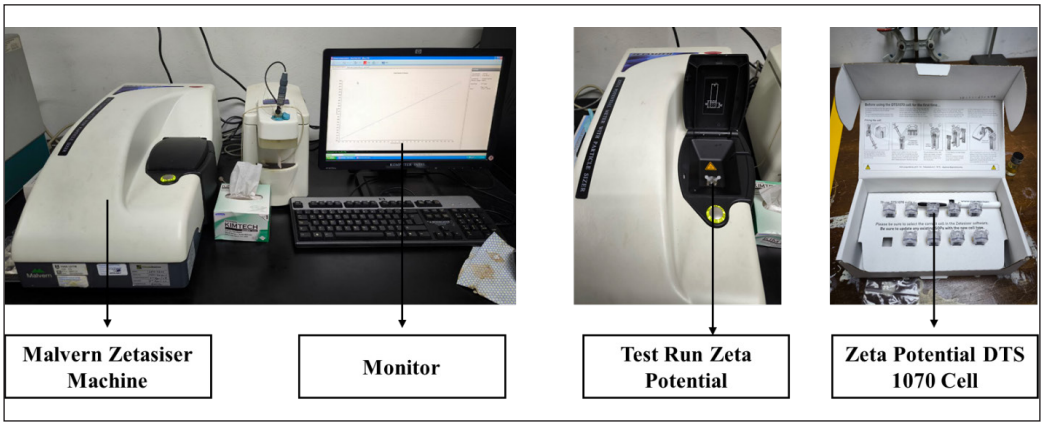


Figure 3. Zeta potential measurement

Thermo-physical Properties Measurement

Dynamic Viscosity

The dynamic viscosity is a critical property that influences lubricating efficiency, forming a stable boundary layer between the tool and the workpiece (Asadi et al., 2016). Hence, nanoparticles can modify rheological properties, improving lubricating film stability, reducing wear, and enhancing machining performance. In this study, an electronic rheometer Anton Paar Germany GmbH (2006) has been used in measuring the viscosity of the nanolubricant of five-volume concentrations for SiO_2 biodegradable nanolubricant; 0. 01%, 0. 03%, 0. 05%, 0. 07% and 0. 1% at temperatures between 27 °C and 90 °C. The spindle was immersed in the sample, with rotational speeds ranging from 0.01 to 1000 revolutions per minute (1/min). Viscosity measurements were taken once the readings stabilised. To ensure the reliability of the results, the test was conducted three times in Figure 4, and the average viscosity value was subsequently calculated.

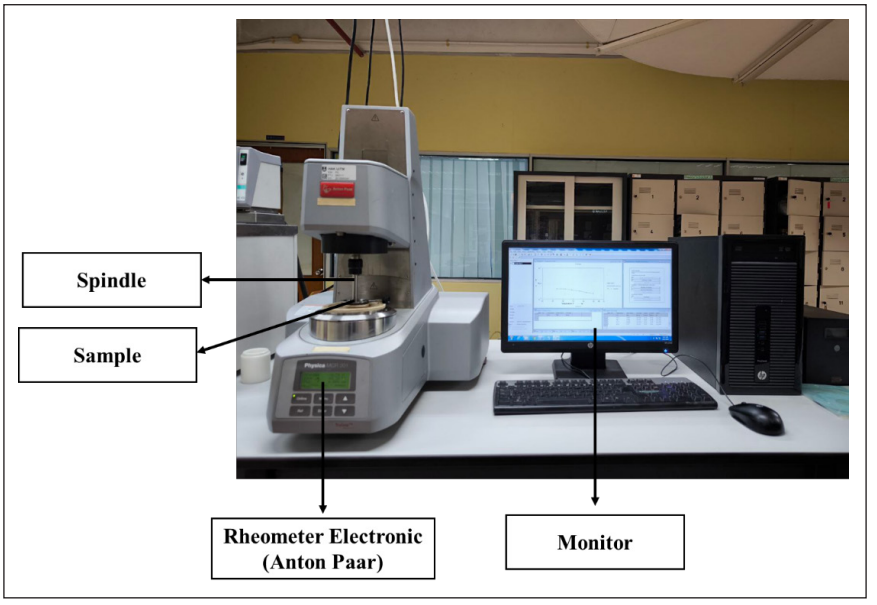


Figure 4. Dynamic viscosity rheometer electronic (Anton Paar Germany GmbH, 2006)

Tribological Properties Measurement

As per the ASTM D4172-94 standard, the tribological assessment of the biodegradable nanolubricant SiO_2 was carried out utilising a Koehler four-ball tribo tester, as depicted in Figure 5. In this inquiry, test balls with a 12.7 mm diameter, HRC 64-66 hardness range, and Grade 25 extra-polished surface finish were used. The biodegradable base oil was employed starting in the tribology assessment, and then different volume concentrations of the biodegradable nanolubricant SiO_2 (0.01%, 0.03%, 0.05%, 0.07%, and 0.1%) were applied. The ASTM D4172-94 standard was followed by the experimental circumstances, which were compliant with the requirements listed in Table 3. For a maximum of 60 minutes during the experiment, the working temperature was kept at 75°C. To control the heater and ensure that the lubricants were kept at a consistent temperature, an automatic temperature controller was employed. The lever arm was subjected to a 40.0 kg load, and 1200 rpm was selected as the rotational speed. The recorded parameters, including the friction torque and coefficient of friction for both biodegradable base oil and SiO_2 biodegradable nanolubricant, were recorded as practised by another researcher (Hamisa et al., 2023). In lubrication, lower frictional torque and coefficient of friction were preferred as they indicated decreased resistance to motion and smoother surface interaction. Each test was conducted three times to calculate the average diameter of the wear scar.

Table 3
The test conditions for the ASTM D4172-94 standard

ASTM Standard	Test Method	Test Conditions				
		Speed (rpm)	Load (kg)	Duration (Minutes)	Temperature (°C)	Remarks
D4172-94	Wear preventive characteristics of lubricating fluid	1200±60	40.0±0.2	60±1	75±2	Ball pot to be torqued down between 25 and 50 ft.lb

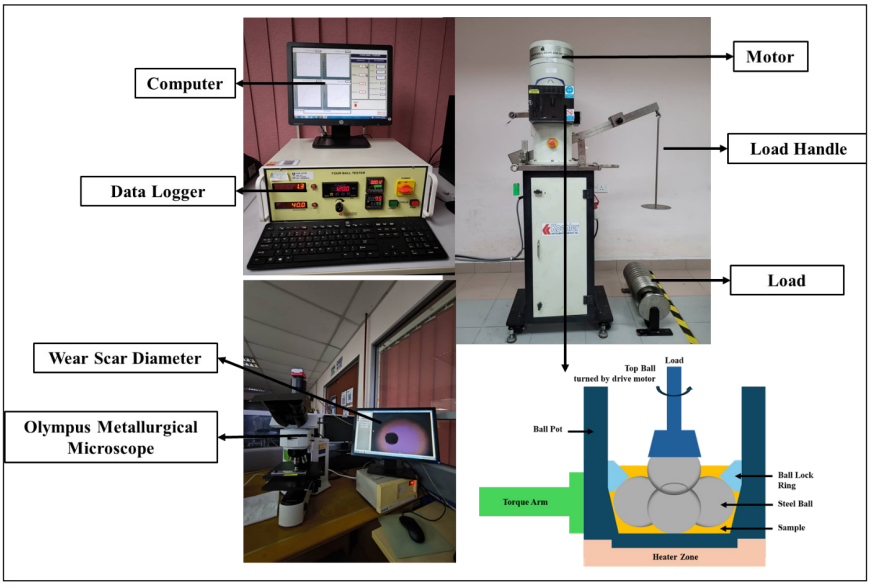


Figure 5. Four-ball tribology tester and its schematic configuration

Uncertainty Analysis

The precision of physical measurements directly impacts the accuracy of an analysed parameter and its related conclusions. The data from the zeta potential analyser, UV-vis spectrophotometer, and rotating viscometer were obtained separately. To determine the expanded uncertainties for all measured parameters in this study, the fractional uncertainty formula was applied. The level of parametric precision is determined by the accuracy of the measuring instruments and the scale of the recorded values (Zakaria et al., 2022). Table 4 outlines a summary of the uncertainties associated with the accuracy specifications of the instruments employed in zeta potential, UV-vis absorbance, viscosity, nanoparticle concentration, and other measurements.

Table 4
Uncertainties of the instrumentation for individual readings

Parameters	Range of Instruments	Least Division in Instrument	Experiment Values		Uncertainty (%)	
			Min	Max	Min	Max
Nanoparticle weight (g)	0 to 200	±0.0001	0.005	0.1	0.1	2
Lubricant volume (mL)	1 to 50	±1	100	100	1	1
Volume concentration (%)	-	-	0.01	0.1	-	-
Zeta potential (mV)	0 to 1000	±0.1	80.2	98.3	0.102	0.125
Absorbance (UV-Vis)	0 to 7	±0.001	0.015	0.095	1.05	6.67
Dynamic viscosity (Pa. s)	0.01 to 1000	±0.0001	0.003	0.023	0.435	3.33

Note. Adapted from Anton Paar Germany GmbH, 2006; Sharif et al., 2023

RESULTS AND DISCUSSION

Stability Evaluation

Visual Sedimentation Observations

Figure 6 shows the sample of visual sedimentation observations of 0.1% SiO_2 biodegradable nanolubricant from day 0 (after preparation) until day 30 at room temperature. It was observed that the SiO_2 nanoparticles were shown as well-dispersed nanoparticles in the biodegradable base oil with negligible sedimentation. On the 0th and 5th days, the SiO_2 biodegradable nanolubricant showed a clear appearance without any signs of sedimentation across all sonication durations. The samples maintained consistent visual characteristics during the initial five days following their preparation. The observation was made from the 10^{th} day until the 15th day. The SiO_2 biodegradable nanolubricant still shows a stable appearance without producing any major sedimentation in the sample bottle. Observation continued for the 20th day, the 25th day, and finally on the 30th day, the sedimentation still did not occur in these samples. It showed that the nanolubricant with a 2-hour sonication time was the most stable sample through visual observation. The addition of Triton X-114 played a vital role in preventing sedimentation since surfactants work like helpers that spread out and prevent tiny particles from clumping together or settling down (Azman et al., 2019). They make sure the small particles stay evenly spread in the liquid.

UV-Visible Analysis

The stability of the samples was then assessed quantitatively through a UV-Visible spectrometer to evaluate the peak absorbance and distinguish the acceptable wavelength from the scanning process. The absorbance value for 0.1% SiO_2 in biodegradable oil is shown in Figure 7 for wavelengths of 200 to 450 nm.

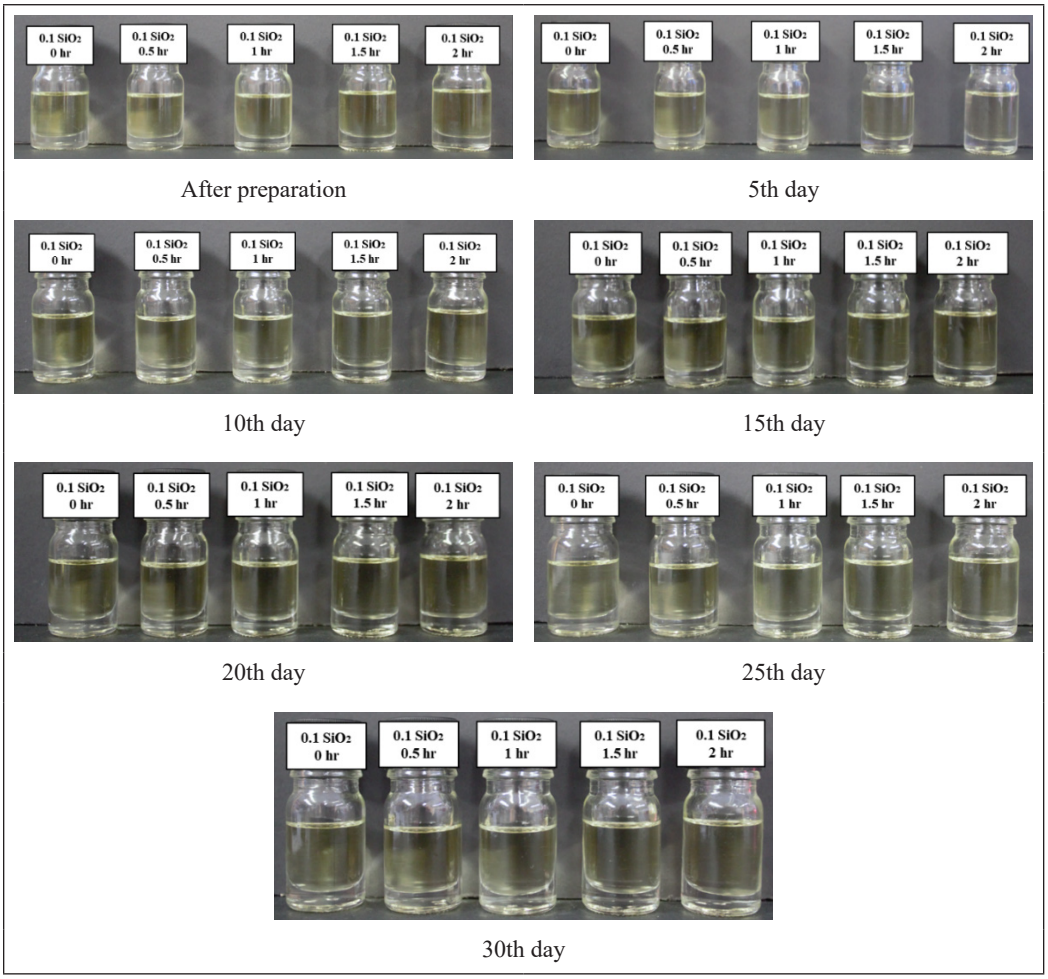


Figure 6. Sedimentation observation of SiO_2 biodegradable nanolubricant with different sonication times

The graph shows that the peak absorbance for 0.1% SiO_2 in biodegradable oil was at 349 nm wavelength.

Meanwhile, the absorbance against concentration at a constant peak wavelength of 349 nm is shown in Figure 8. The volume concentrations of 0.01%, 0.03%, 0.05%, 0.07%, and 0.1% absorbance were plotted to prove the Beer-Lambert law, which shows that absorbance is directly proportional to the concentration of substances in the mixture, translated to the ability of the substances to absorb light and the length of the light's path through the sample (Sanukrishna et al., 2019). Therefore, the concentration can be calculated from the absorbance value measured. This Beer-Lambert law principle was validated by comparing the findings with prior research conducted by Hamzan et al. (2025) on a hybrid SiO2:TiO2 biodegradable nanolubricant. The results of the current validation agree with those of the previous researchers.

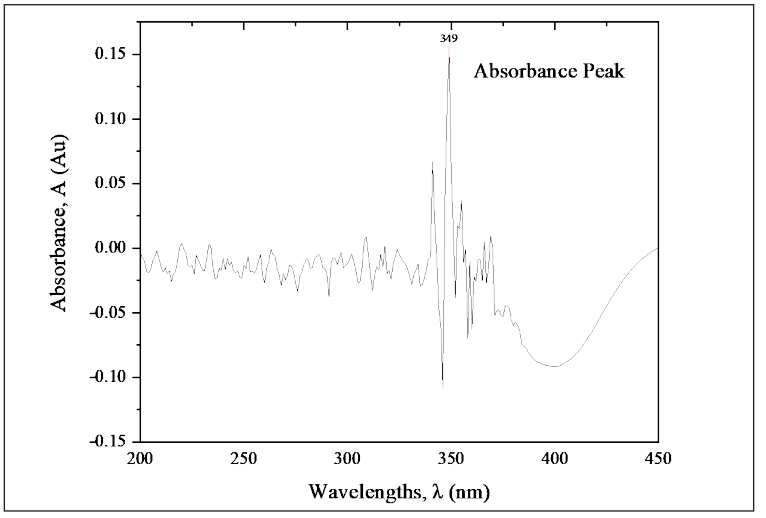


Figure 7. Absorbance of 0.1% volume concentration of nanolubricant at a different wavelength

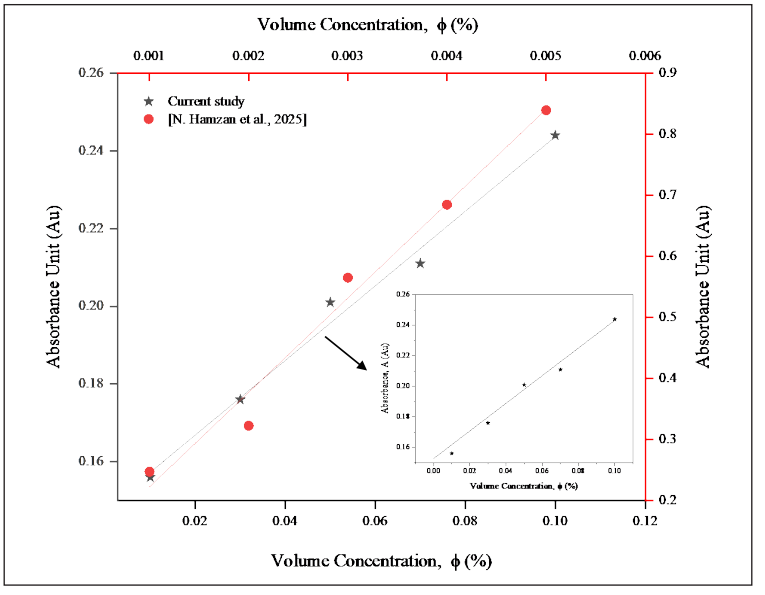


Figure 8. UV-Visible Spectrophotometer linear relative graph of the absorbance against SiO_2 biodegradable nanolubricant concentration

Figure 9 shows the relationship of the absorbance values for five different ultrasonication times, ranging from no sonication, labelled as 0, 0.5 hours, 1 hour, 1.5 hours, and 2 hours, for 0.1% volume concentrations of SiO_2 biodegradable nanolubricant. The stability was observed for 20 days. It is shown that the lowest absorbance ratio is with 0 sonication time, as expected due to its poor stability.

The 2-hour sonication process exhibited the greatest stability over the course of the evaluation, as it demonstrated excellent stability, reaching 59%. The ultrasonication time significantly influences the stability of nanolubricants, as observed in the UV-Visible spectrophotometer evaluation of the SiO_2 biodegradable nanolubricant. The longer sonication duration yielded the most favourable absorbance ratio as compared to lower sonication durations (Ramadhan et al., 2019). The general trend of the absorbance ratio is decreasing over time, which is in good agreement with other researchers' findings (Hamisa et al., 2023). Sharif et al. (2022) Investigated the stability of SiO_2/PAG nanolubricants and observed that extending ultrasonic treatment beyond 120 minutes (2 hours) did not lead to an increase in absorbance ratio. Their findings showed that ultrasonic agitation in a water bath effectively reduced particle agglomeration and minimised particle size, resulting in consistently high absorbance values. It was further noted that additional sonication beyond the initial 120 minutes (2 hours) offered no significant benefit. These observations align with the conclusions of Hamzan et al. (2025), who also identified a two-hour sonication period as the ideal duration for preparing a stable nanolubricant formulation.

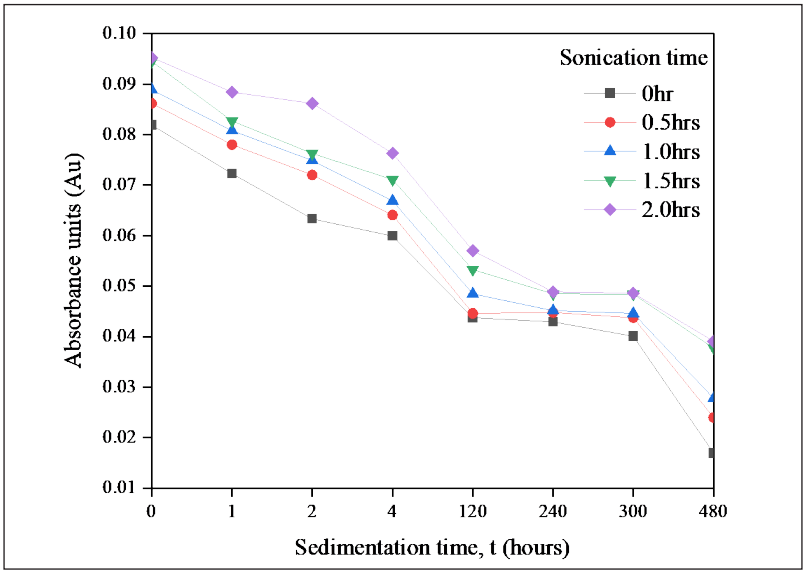


Figure 9. Absorbance units of 0.1% volume concentration of SiO_2 biodegradable nanolubricant for 20 days with different sonication times

Zeta Potential

Figure 10 shows the zeta potential measurement for SiO_2 biodegradable nanolubricant. This approach enabled the evaluation of the zeta potential of nanoparticles at a minimal concentration before potential interactions or agglomeration became significant.

Furthermore, using the lowest volume concentration aided in understanding the individual particle behaviour and characteristics, since the higher concentrations might have led to more complex interactions and the potential to influence the stability reading (Ramadhan et al., 2019). These outcomes were compared against the stability condition classification offered by Ghadimi et al. (2011) for various zeta potential ranges, which categorise absolute values exceeding 60 mV as excellent stability. The nanolubricants prepared showed an absolute zeta potential value of 98.3 mV, which exceeded the stability threshold of 60 mV. According to (Ismail et al., 2023) Stabilisation theory, a high absolute value of zeta potential signifies highly intense electrostatic repulsive interactions between nanoparticles, hence signifying a stable suspension. A similar approach was performed by Sharif et al. (2022) In the study of SiO_2/PAG nanolubricants, the stability of the nanolubricant was at 80.6 mV. Thus, proved that the SiO_2 biodegradable nanolubricant prepared was in a very good stability condition.

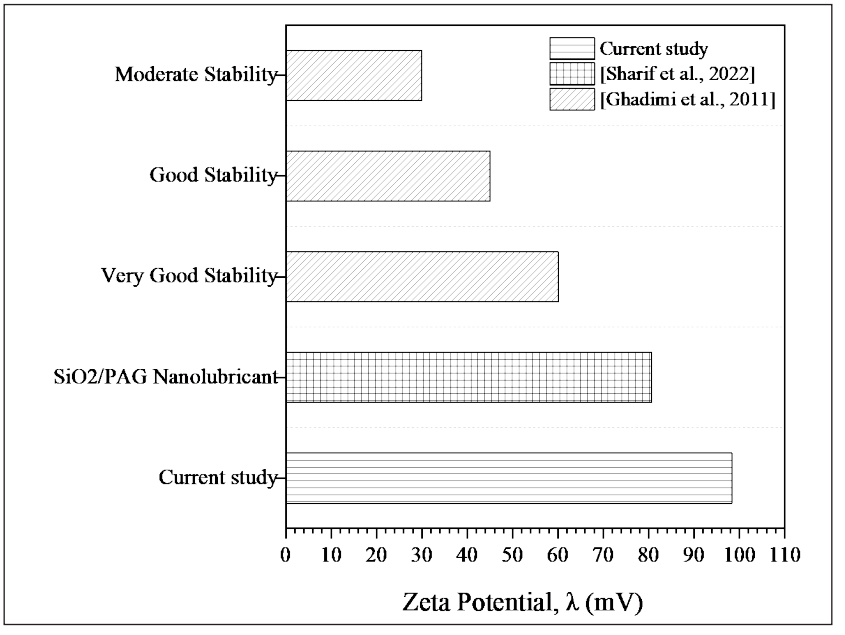


Figure 10. Zeta Potential of 0.01% SiO_2 biodegradable nanolubricant.

Thermo-physical Properties Evaluation

Dynamic Viscosity

Figure 11 shows the dynamic viscosity result of SiO_2 biodegradable nanolubricant. The graph shows a decreasing trend in the dynamic viscosity of different volume concentrations of nanolubricant with the rising temperatures from 27 °C to 90 °C.

For a comparison purpose on the effect of volume concentration towards viscosity value, at a temperature of 60 °C, the highest dynamic viscosity is shown by 0.1 vol % concentration with 17.81% increment from biodegradable base oil. This was followed by 0.07% with an increment of 16.71%, 0.05% (15.78%), 0.03% with an increment of 14.82%, and 0.01% (13.94%). This result is utterly consistent with Ismail et al. (2022) who reported that the higher the volume concentration, the higher the viscosity values will be. In terms of temperature effect, all lubricants tested exhibit lower viscosity as the temperatures are increased, which is similar to the research findings of (Asadi et al., 2016). This is favourable in MQL operation since it will involve a much higher temperature (Ismail et al., 2022).

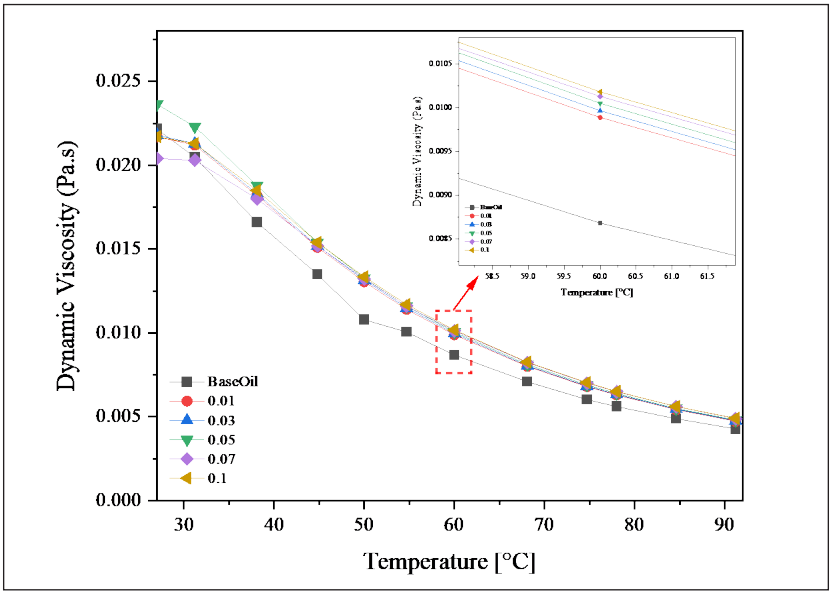


Figure 11. Dynamic viscosity of nanolubricant

Tribological Properties Evaluation

Coefficient of Friction (COF)

The coefficient of friction was a necessary constraint that significantly influenced the execution of lubricants. Figure 12 provides a visual representation of the coefficient of friction (COF) for SiO_2 biodegradable nanolubricant across different volume concentrations compared to biodegradable base oil. The experiment highlighted a significant 37.82% reduction in COF when using 0.07% SiO_2 biodegradable nanolubricant as compared to biodegradable base oil. This was closely followed by 0.05% SiO_2 biodegradable nanolubricant, which demonstrated a 20.92% reduction in COF.

The addition of 0.03% volume concentration improved the COF with a reduction of 19.03%. In general, the COF for all nanolubricants is lower than biodegradable base oil. Adding nanoparticles up to a concentration of 0.07% improves lubrication on the contact surface, leading to a reduction in COF. The maximum permitted concentration of nanoparticles is 0.07%, and going above this limit could either not improve performance at all or make it worse than the base oil. The results of a study by Kamel et al. (2021) that examined hybrid nanolubricants on engine oil, which uses many types of volume concentration of nanolubricant had supported the trend that was shown for this investigation. The decrease in COF suggests that SiO_2 nanoparticles in the biodegradable base oil enhance tribological effects, if the nanoparticle concentration remains below the threshold value (0.07%). According to Ismail et al. (2022), the main factor responsible for the decrease in friction torque and COF in the examined nanolubricant was the existence of efficient nanoparticles. Due to their reduced interfacial frictional surface activity, these nanoparticles promoted the nanoparticle mechanism, where the change from pure sliding friction to rolling friction. These findings underscore the potential of this nanolubricant to significantly improve machining processes in the future (Hamisa et al., 2020). After an initial phase, the COF progressively increased throughout the 3600-second test. This trend may stem from several factors, including thermal degradation of the lubricant at 75 °C, agglomeration or settling of nanoparticles, buildup of wear particles in the contact area, and alterations in the tribo-film characteristics. The rise in temperature also played a role by decreasing the resistance to flow between molecules and the rotating components, which led to a reduction in viscosity (Sharif et al., 2023). As the temperature increased, molecular interactions became more fluid, enhancing the lubricant's flow and affecting its frictional response.

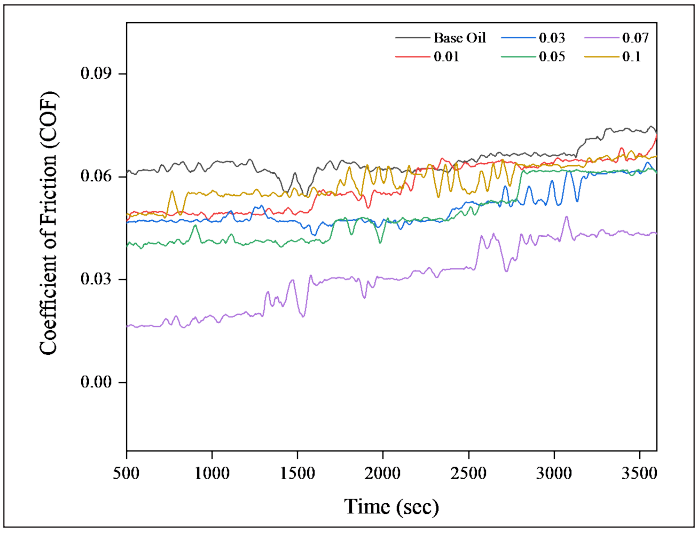


Figure 12. Coefficient of friction of SiO_2 biodegradable nanolubricant

Wear Scar Diameter (WSD)

Figure 13 shows the average wear scar diameter (WSD) of all balls used in the four-ball tribology experiment. Analysing Figure 14 revealed that the diameter for WSD of the biodegradable base oil (0.0902 mm) was the highest among the samples. This was due to the biodegradable base oil lacking specific property enhancements, leading to an increase in friction and wear between moving parts during machining (Zawawi et al., 2019). Research from Hamzan et al. (2025) On biodegradable nanolubricant also supported this finding, demonstrating that the highest WSD was associated with base oil, and the addition of nanoparticles led to a reduction in wear scar diameter and enhanced the surface finish of the ball. Among these concentrations, the most effective nanolubricant in terms of wear scar diameter reduction was 0.07% volume concentration, as it exhibited a 14.41% reduction with 0.0139 standard deviation. In comparison, the wear scar reduction for 0.05% nanolubricant was 13.53%, followed by 0.03% (13.30%), and 0.01% (12.86%). The 0.1% volume concentration of SiO_2 biodegradable nanolubricant did not give the highest wear scar reduction as expected. Similarly to the COF result, the threshold value for WSD was obtained at 0.07% volume concentration. Thus, increasing the volume concentration of nanolubricant did not give any improvement to the WSD result. Generally, the biodegradable base oil exhibited the largest WSD than the nanolubricants.

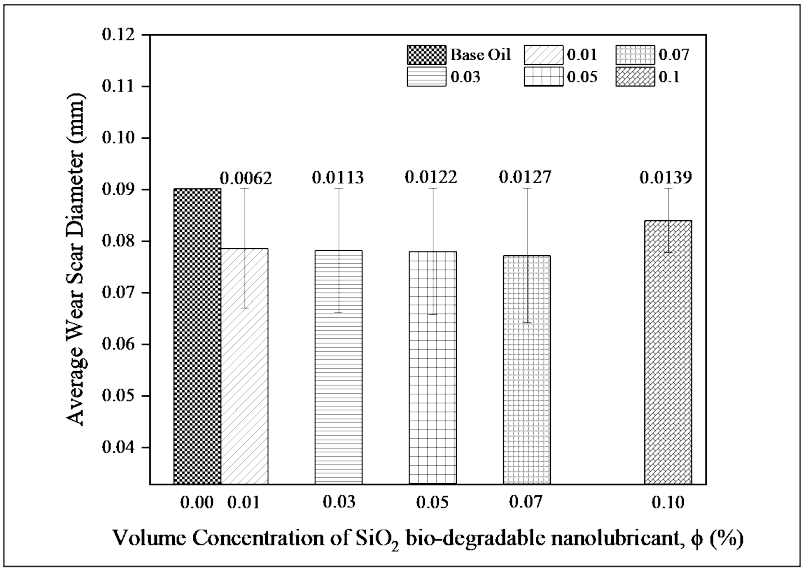


Figure 13. The average wear scar diameter of the SiO_2 biodegradable nanolubricant

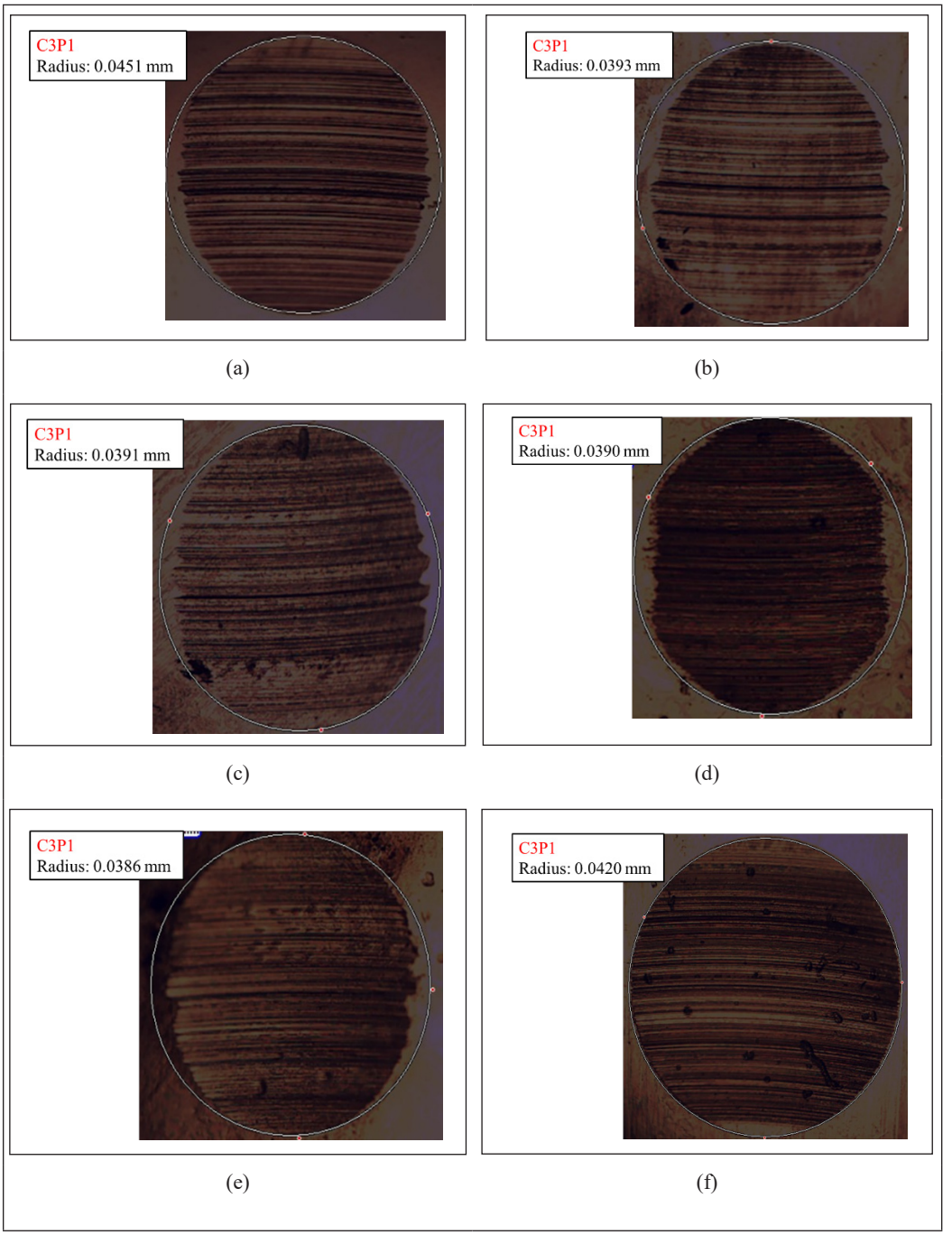


Figure 14. Optical micrograph of wear scar for (a) biodegradable base oil; (b) 0.01% SiO_2 biodegradable nanolubricant; (c) 0.03% SiO_2 biodegradable nanolubricant; (d) 0.05% SiO_2 biodegradable nanolubricant; (e) 0.07% SiO_2 biodegradable nanolubricant; and (f) 0.1% SiO_2 biodegradable nanolubricant

CONCLUSION

In conclusion, this research successfully analysed the stability, thermo-physical and tribological characteristics of SiO_2 biodegradable nanolubricant. Visual sedimentation observation held during the stability test indicated the absence of agglomeration for all volume concentrations up to day 30, affirming the compatibility of SiO_2 biodegradable nanolubricant for thorough mixing, with the two-hour sonication time yielding the cleanest and clearest observations. While occasional fluctuations were observed in the UV-Vis analysis, these values consistently fell within an acceptable range for the nanolubricant. The most stable performance during the observation period was exhibited by 0.1% SiO_2 biodegradable nanolubricant subjected to 2.0 hours of sonication. In terms of thermo-physical properties, the dynamic viscosity for 0.1% volume has the highest viscosity through the various temperature ranges of 27 °C to 90°C. Furthermore, tribological properties of COF for SiO_2 biodegradable nanolubricant consistently outperformed biodegradable base oil across all volume concentrations throughout the four-ball experiment. The 0.07% volume concentration of SiO_2 biodegradable nanolubricant demonstrated the lowest COF with the highest COF reduction. The WSD for SiO_2 biodegradable nanolubricant at all volume concentrations also showed the lowest values at 0.07% volume concentration, with a remarkable 14.41% reduction of wear scar diameter. This study also discovered the threshold value at 0.07% volume concentration of SiO_2 biodegradable nanolubricant, which was the optimum volume concentration that can improve the machining operation as compared to 0.1% volume concentration. Although SiO_2 biodegradable nanolubricants improve biodegradability and sustainability, challenges remain in waste management and nanoparticle stability. Proper disposal practices must be implemented to prevent accumulation in industrial wastewater or unintended environmental effects. Future research should focus on enhancing full biodegradation potential, ensuring that nanoparticle dispersions align with ecological safety standards. The transition to eco-friendly lubricants must also consider industrial waste treatment protocols to support responsible disposal practices. In summary, the application of SiO_2 biodegradable nanolubricant in MQL machining is feasible and demonstrates promising potential for industrial use.

ACKNOWLEDGEMENT

The authors would like to acknowledge the Faculty of Mechanical Engineering, Universiti Teknologi MARA and Universiti Teknologi MARA under the 600-RMC/GIP 5/3 (009/2024) for the research funding and technical assistance provided.

REFERENCES

Ahmad, A. A., Abdul Ghani, J., & Che Haron, C. H. (2022). Potential of bio-metal cutting fluid from treated recycled cooking oil for metal cutting fluid application. *Journal of Mechanical Engineering, 19*(1), 83-97. https://doi.org/10.24191/jmeche.v19i1.19688

Anton Paar Germany GmbH. (2006). *Instruction manual Rheoplus software: Volume 4-Device driver Physica MR Series*.

Asadi, A., Asadi, M., Rezaei, M., Siahmargoi, M., & Asadi, F. (2016). The effect of temperature and solid concentration on dynamic viscosity of MWCNT/MgO (20-80) SAE50 hybrid nano-lubricant and proposing a new correlation: An experimental study. *International Communications in Heat and Mass Transfer, 78*, 48-53. https://doi.org/10.1016/j.icheatmasstransfer.2016.08.021

Azman, N. F., & Samion, S. (2019). Dispersion stability and lubrication mechanism of nanolubricants: A review. *International Journal of Precision Engineering and Manufacturing-Green Technology, 6*(2), 393-414. https://doi.org/10.1007/s40684-019-00080-x

ChemicalBook. (2025). *CAS database list of silicon dioxide*.

Czaplicka, N., Grzegórska, A., Wajs, J., Sobczak, J., & Rogala, A. (2021). Promising nanoparticle-based heat transfer fluids: Environmental and techno-economic analysis compared to conventional fluids. *International Journal of Molecular Sciences, 22*(17), Article 9201. https://doi.org/10.3390/ijms22179201

Dambatta, Y. S., Sayuti, M., Sarhan, A. A. D., Hamdi, M., Manladan, S. M., & Reddy, M. M. (2019). Tribological performance of SiO_2-based nanofluids in minimum quantity lubrication grinding of Si_3N_4 ceramic. *Journal of Manufacturing Processes, 41*, 135-147. https://doi.org/10.1016/j.jmapro.2019.03.024

Eltaggaz, A., Hegab, H., Deiab, I., & Kishawy, H. A. (2018). Hybrid nano-fluid-minimum quantity lubrication strategy for machining austempered ductile iron (ADI). *International Journal on Interactive Design and Manufacturing, 12*(4), 1273-1281. https://doi.org/10.1007/s12008-018-0491-7

Eltaggaz, A., Zawada, P., Hegab, H. A., Deiab, I., & Kishawy, H. A. (2018). Coolant strategy influence on tool life and surface roughness when machining ADI. *The International Journal of Advanced Manufacturing Technology, 94*(9-12), 3875-3887. https://doi.org/10.1007/s00170-017-1088-1

Eltaggaz, A., Nouzil, I., & Deiab, I. (2021). Machining Ti-6Al-4V alloy using nano-cutting fluids: Investigation and analysis. *Journal of Manufacturing and Materials Processing, 5*(2), Article 42. https://doi.org/10.3390/jmmp5020042

Garcia Tobar, M., Contreras Urgiles, R. W., Jimenez Cordero, B., & Guillen Matute, J. (2024). Nanotechnology in lubricants: A systematic review of the use of nanoparticles to reduce the friction coefficient. *Lubricants, 12*(5), Article 166.https://doi.org/10.3390/lubricants12050166

Ghadimi, A., Saidur, R., & Metselaar, H. S. C. (2011). A review of nanofluid stability properties and characterisation in stationary conditions. *International Journal of Heat and Mass Transfer, 54*(17-18), 4051-4068. https://doi.org/10.1016/j.ijheatmasstransfer.2011.04.014

Gulzar, M., Masjuki, H. H., Kalam, M. A., Varman, M., Zulkifli, N. W. M., Mufti, R. A., Zahid, R., & Yunus, R. (2017). Dispersion stability and tribological characteristics of TiO_2:SiO_2 nanocomposite-enriched biobased lubricant. *Tribology Transactions, 60*(4), 670-680. https://doi.org/10.1080/10402004.2016.1202366

Gupta, M. K., Khan, A. M., Song, Q., Liu, Z., Khalid, Q. S., Jamil, M., Kuntoğlu, M., Usca, Ü. A., Sarıkaya, M., & Pimenov, D. Y. (2021). A review on conventional and advanced minimum quantity lubrication approaches on performance measures of grinding process. *The International Journal of Advanced Manufacturing Technology, 117*(3-4), 729-750. https://doi.org/10.1007/s00170-021-07785-x

Gutnichenko, O., Bushlya, V., Bihagen, S., & Ståhl, J.-E. (2018). Influence of GnP additive to vegetable oil on machining performance when MQL-assisted turning Alloy 718. *Procedia Manufacturing, 25*, 330-337. https://doi.org/10.1016/j.promfg.2018.06.091

Hawwash, A. A., Abdel Rahman, A. K., Nada, S. A., & Ookawara, S. (2018). Numerical investigation and experimental verification of performance enhancement of flat plate solar collector using nanofluids. *Applied Thermal Engineering, 130*, 363-374. https://doi.org/10.1016/j.applthermaleng.2017.11.027

Hamisa, A. H., Azmi, W. H., Ismail, M. F., Rahim, R. A., & Ali, H. M. (2023). Tribology performance of polyol-ester based TiO2, SiO2, and their hybrid nanolubricants. *Lubricants, 11*(18), 1-20. https://doi.org/10.3390/lubricants11010018

Hamisa, A. H., Yusof, T. M., Azmi, W. H., Mamat, R., & Sharif, M. Z. (2020). The stability of TiO2/POE nanolubricant for automotive air-conditioning system of hybrid electric vehicles. *IOP Conference Series: Materials Science and Engineering, 863*(1), Article 012050. https://doi.org/10.1088/1757-899X/863/1/012050

Hamzan, N., Zakaria, I. A., Ghani, J. A., Halim, N. H. A., Azmi, W. H., Solihin, Z. H., & Ahmad, A. A. (2025). The SiO2:TiO2 hybrid biodegradable nanolubricant for sustainable machining: The stability, thermo-physical and tribology perspectives. *Tribology International, 207*, Article 110614. https://doi.org/10.1016/j.triboint.2025.110614

Ismail, M. F., Azmi, W. H., Mamat, R., & Hamisa, A. H. (2022). Experimental investigation on Newtonian behaviour and viscosity of TiO2/PVE nanolubricants for application in refrigeration system. *Journal of Advanced Research in Fluid Mechanics and Thermal Sciences, 92*(1), 9-17. https://doi.org/10.37934/arfmts.92.1.917

Ismail, M. F., Azmi, W. H., Mamat, R., Sharma, K. V., & Zawawi, N. N. M. (2023). Stability assessment of polyvinyl-ether-based TiO2, SiO2, and their hybrid nanolubricants. *Lubricants, 11*(1), Article 23. https://doi.org/10.3390/lubricants11010023

Kamel, B. M., El-Kashif, E., Hoziefa, W., Shiba, M. S., & Elshalakany, A. B. (2021). The effect of MWCNTs/GNs hybrid addition on the tribological and rheological properties of lubricating engine oil. *Journal of Dispersion Science and Technology, 42*(12), 1811-1819. https://doi.org/10.1080/01932691.2020.1789470

Kapil, A., Ozaner, O. C., Sato, Y., Hayashi, Y., Ikeda, K., Suga, T., Tsukamoto, M., Karabulut, S., Bilgin, M., & Sharma, A. (2024). Overcoming machining challenges in hybrid laser metal deposition of IN718 with heat-assisted minimum quantity lubrication. *The International Journal of Advanced Manufacturing Technology*.https://doi.org/10.1007/s00170-024-13646-0

Lim, S. K., Azmi, W. H., Jamaludin, A. S., & Yusoff, A. R. (2022). Characteristics of hybrid nanolubricants for MQL cooling lubrication machining application. *Lubricants, 10*(12), Article 350. https://doi.org/10.3390/lubricants10120350

Merck, K. (2025). *Safety data sheet (SDS): Triton X-114*.

Mohd Khalil, A. N., Azmi, A. I., Murad, M. N., & Mahboob Ali, M. A. (2018). The effect of cutting parameters on cutting force and tool wear in machining nickel titanium shape memory alloy ASTM F2063 under minimum quantity nanolubricant. *Procedia CIRP, 77*, 227-230. https://doi.org/10.1016/j.procir.2018.09.002

Mondragón, R., Abellán-Nebot, J. V., Habib, K., & Serrano, J. (2023). Tribological analysis and operation issues of SiO2 nanolubricants for MQL machining operations. *Key Engineering Materials, 960*, 11-20. https://doi.org/10.4028/p-Tu1Cg9

Moretti, G. B., Machado, F. d. C., Winckler, L., Ribeiro, F. S. F., Talon, A. G., Sanchez, L. E. d. A., Bianchi, E. C., & Lopes, J. C. (2024). Sustainable machining practices: A comparative study of MQL and MQL with wheel cleaning jet applied to different abrasives. *The International Journal of Advanced Manufacturing Technology, 132*(5-6), 2395-2409. https://doi.org/10.1007/s00170-024-13521-y

Nouzil, I., Eltaggaz, A., Pervaiz, S., & Deiab, I. (2022). Toxicity analysis of nano-minimum quantity lubrication machining-A review. *Lubricants, 10*(8), Article 176. https://doi.org/10.3390/lubricants10080176

Nugroho, A., Bo, Z., Mamat, R., Azmi, W. H., Najafi, G., & Khoirunnisa, F. (2021). Extensive examination of sonication duration impact on stability of Al2O3-polyol ester nanolubricant. *International Communications in Heat and Mass Transfer, 126*, Article 105418. https://doi.org/10.1016/j.icheatmasstransfer.2021.105418

Pawanr, S., & Gupta, K. (2024). Dry machining techniques for sustainability in metal cutting: A review. *Processes, 12*(2), Article 417.https://doi.org/10.3390/pr12020417

Ramadhan, A. I., Azmi, W. H., Mamat, R., Hamid, K. A., & Norsakinah, S. (2019). Investigation on stability of tri-hybrid nanofluids in water-ethylene glycol mixture. *IOP Conference Series: Materials Science and Engineering, 469*, Article 012068. https://doi.org/10.1088/1757-899X/469/1/012068

Roy, S., Kumar, R., Kumar Sahoo, A., & Kumar Das, R. (2019). A brief review on effects of conventional and nano particle based machining fluid on machining performance of minimum quantity lubrication machining. *Materials Today: Proceedings, 18*, 5421-5331. https://doi.org/10.1016/j.matpr.2019.07.571

Said, Z., Gupta, M., Hegab, H., Arora, N., Khan, A. M., Jamil, M., & Bellos, E. (2019). A comprehensive review on minimum quantity lubrication (MQL) in machining processes using nano-cutting fluids. *The International Journal of Advanced Manufacturing Technology, 105*(5-6), 2057-2086.https://doi.org/10.1007/s00170-019-04382-x

Sanukrishna, S. S., Shafi, M., Murukan, M., & Jose Prakash, M. (2019). Effect of SiO2 nanoparticles on the heat transfer characteristics of refrigerant and tribological behaviour of lubricant. *Powder Technology, 356*, 39-49.https://doi.org/10.1016/j.powtec.2019.07.083

Sharif, M. Z., Azmi, W. H., Ghazali, M. F., Zawawi, N. N. M., & Ali, H. M. (2023). Viscosity and friction reduction of double-end-capped polyalkylene glycol nanolubricants for eco-friendly refrigerant. *Lubricants, 11*(3), Article 129.https://doi.org/10.3390/lubricants11030129

Sharif, M. Z., Azmi, W. H., Ghazali, M. F., Zawawi, N. N. M., & Hendrawati, T. Y. (2022). Stability and thermal conductivity of mono and hybrid nanoparticles dispersion in double-end capped PAG lubricant. *Lubricants, 11*(1), Article 1. https://doi.org/10.3390/lubricants11010001

Sharif, M. Z., Azmi, W. H., Redhwan, A. A. M., & Zawawi, N. M. M. (2017). Preparation and stability of silicone dioxide dispersed in polyalkylene glycol based nanolubricants. *MATEC Web of Conferences, 90*, Article 01049. https://doi.org/10.1051/matecconf/20179001049

Sia, S. Y., Bassyony, E. Z., & Sarhan, A. A. D. (2014). Development of SiO2 nanolubrication system to be used in sliding bearings. *The International Journal of Advanced Manufacturing Technology, 71*(5-8), 1277-1284.https://doi.org/10.1007/s00170-013-5566-9

Unist, Inc. (2025). *Material safety data sheet: Coolube 2210 XP*.

Zakaria, I. A., Mohamed, W. A. N. W., Azid, N. H. A., Suhaimi, M. A., & Azmi, W. H. (2022). Heat transfer and electrical discharge of hybrid nanofluid coolants in a fuel cell cooling channel application. *Applied Thermal Engineering, 210*, Article 118369. https://doi.org/10.1016/j.applthermaleng.2022.118369

Zawawi, N. N. M., Azmi, W. H., Sharif, M. Z., & Najafi, G. (2019). Experimental investigation on stability and thermo-physical properties of Al2O3-SiO2/PAG nanolubricants with different nanoparticle ratios. *Journal of Thermal Analysis and Calorimetry, 135*(2), 1243-1255. https://doi.org/10.1007/s10973-018-7670-4

Pertanika J. Sci. & Technol. 34 (3): 1433 - 1449 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Silicon-based Fin Field-Effect Transistor (Si-FinFET) Optimisation Based on Gate Length and Operating Temperature

Yousif Atalla[1], Mohamad Hafiz Mamat[1]*, and Yasir Hashim[2]

[1]NANO-ElecTronic Centre (NET), Faculty of Electrical Engineering, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

[2]Department of Electrical Engineering and Computer Science, College of Engineering, A'Sharqiyah University, 400 Ibra, Oman

ABSTRACT

This study provides an in-depth investigation into the temperature sensitivity of Silicon-based Fin Field-Effect Transistor (Si-FinFET), highlighting their viability as nano-dimensional temperature sensors. The Multi-Gate Field-Effect Transistor (MuGFET) simulation framework was utilised to model and analyse the thermal response characteristics of Si-FinFET structures. Simulations of the current voltage (I-V) profiles were conducted across a range of temperatures and gate lengths (L_g = 10, 20, and 30 nm), with the temperature sensitivity quantified under a diode-mode operational scheme. The optimum sensitivity was identified through the maximum differential current (ΔI) within the supply voltage window (V_{DD} = 0-1 V). Notably, a linear augmentation in temperature sensitivity was observed as the channel length increased from 10 nm to 20 nm, underscoring the scalability and enhanced thermal responsivity of Si-FinFET architectures for advanced nanoscale sensing applications.

Keywords: Channel length, FinFETs, nano sensor, Si, temperature

ARTICLE INFO

Article history:
Received: 13 February 2025
Accepted: 02 June 2025
Published: 12 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.07

E-mail addresses:
2022131887@isiswa.uitm.edu.my (Yousif Atalla)
mhmamat@uitm.edu.my (Mohamad Hafiz Mamat)
yasir.hashim@ieee.org (Yasir Hashim)
* Corresponding author

INTRODUCTION

In the realm of modern nanoelectronics, device-level thermal sensing has become increasingly indispensable for ensuring system reliability, particularly as integrated circuits operate under escalating power densities and aggressively scaled geometries. As conventional Metal Oxide Field-Effect Transistor (MOSFET) technologies approach their fundamental

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

scaling limits, research attention has pivoted toward next-generation Field-Effect Transistor (FET) architecture. Among these, the Fin Field-Effect Transistor (FinFET) has emerged as a preeminent candidate, offering enhanced electrostatic control, mitigation of short-channel effects, and seamless compatibility with nanoscale fabrication processes. FinFETs have been extensively investigated across both academic and industrial platforms for their applications in logic circuits, memory architectures, and, more recently, as promising candidates for sensing technologies. Building upon this foundation, the present study explores the temperature-dependent electrical behaviour of silicon-based FinFET (Si-FinFET) and critically evaluates their feasibility as nanoscale thermal sensors (Ariga et al., 2014; Liu et al., 2015; Lu et al., 2015; Yu et al., 2022). The geometrical structure of Si-FinFET is shown in Figure 1.

Semiconductor-based temperature sensors represent quintessential examples of embedded sensing devices within electronic systems (Meijer et al., 2001). In such configurations, the current voltage (I-V) characteristics of nanowire transistors form the cornerstone for the development of transistor-based thermal sensors (Doghish & Ho, 1992; Hashim & Sidek, 2011, 2012; Kesherwani et al., 2022; Liao et al., 1999). Notably, a Bipolar Junction Transistor (BJT) can function as a temperature sensor by shorting its base and collector terminals, effectively operating in diode mode. Similarly, in MOSFET-based designs, the gate terminal may be interconnected with either the source or drain to facilitate thermal sensing functionalities, as depicted schematically in Figure 2. The miniaturisation of fundamental electronic components such as capacitors, diodes, resistors, and transistors at the nanoscale has further catalysed their widespread adoption within the modern electronics industry.

The efficacy of emerging nanoelectronic technologies is intrinsically tied to their nanoscale dimensional attributes. As advancements in fabrication and modelling continue, the integration of ultra-miniaturised transistors into system-on-chip architectures is poised to enhance device reliability and functionality. Despite the well-recognised scaling bottlenecks of traditional MOSFET structures, innovations in nano-scaled FET architectures remain revolutionary, necessitating continued exploration through novel modelling techniques and technological ingenuity to circumvent existing physical limitations.

Simulation methodologies have become indispensable for augmenting experimental efforts in the characterisation and optimisation of nanoscale MOSFETs (Bescond et al., 2004). Beyond device fabrication, simulation techniques enable the evaluation of structural advantages and trade-offs, cost-efficiency projections, and scalability potential for MOSFETs operating in the nanometer regime (Alvarado et al., 2013; Dixit et al., 2019; Fahad et al., 2015; Jaisawal et al., 2022; Li et al., 2021).

The FinFET architecture incorporates a diverse selection of semiconductor materials, including silicon (Si), germanium (Ge), gallium arsenide (GaAs), and indium arsenide

(InAs), each imparting distinct electronic properties that critically influence FinFET performance metrics (Hashim, 2017). This research focuses on the comprehensive analysis of temperature sensitivity and thermal stability in Si-FinFET devices operating under diode-mode configurations. Emphasis is placed on assessing the impact of varying Si channel lengths and operational temperature ranges on pivotal electrical parameters, such as drain-induced barrier lowering (DIBL), threshold voltage (V_T), and sub-threshold swing (SS). Through this investigation, the study aims to elucidate the potential of Si-FinFET-based structures as highly responsive nanoscale thermal sensors.

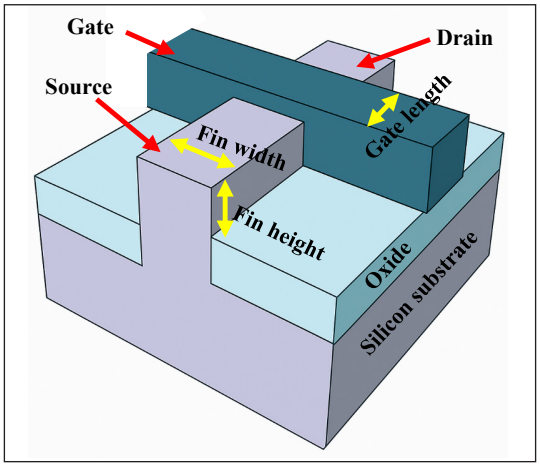


Figure 1. Geometrical structure of Si-FinFET

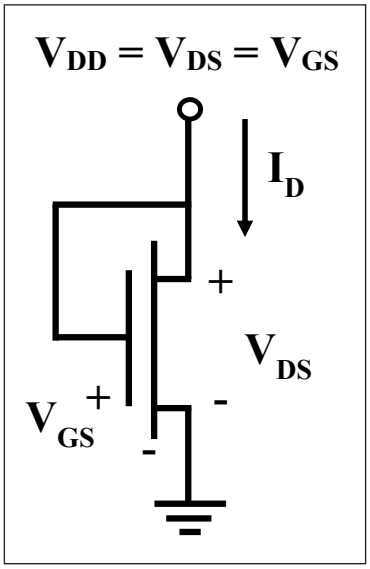


Figure 2. Diode connection mode of MOSFET as a temperature sensor

RESEARCH METHOD

This study systematically examines the performance characteristics of Si-FinFET transistors using the MuGFET modelling tool. The output I-V characteristic curves are simulated under a range of operational conditions and device configurations. These simulation-derived I–V characteristics are subsequently analysed to investigate the influence of key parameters, such as temperature and gate length, on the electrical behaviour of nanowire transistors. Specifically, Purdue University's MuGFET simulation platform (Park et al., 2021) is employed, renowned for its accuracy in modelling nanometer-scale multigate FET architectures.

MuGFET incorporates simulation frameworks originally developed at Bell Laboratories, namely PADRE and PROPHET. PADRE serves as a device-oriented simulator capable of modelling structures in one, two, or three dimensions with complex, variable geometries, while PROPHET functions as a partial differential equation (PDE) solver that supports one-, two-, and three-dimensional profiling (Jabbara et al., 2022; Nasri et al., 2023; Park et al., 2021). By offering robust and physically consistent solutions, MuGFET enables researchers to generate reliable FET output characteristics, particularly when detailed physical modelling of carrier transport phenomena is required. It solves the Poisson and drift-diffusion equations self-consistently, thereby facilitating the precise simulation of nanoscale transport behaviour (Ahangari et al., 2022; Hashim, 2022a).

The MuGFET tool, available via Purdue University's nanoHUB platform, has become a benchmark in the field of nanoelectronic device modelling. It specifically targets the simulation of nanoscale multigate FET structures, such as FinFETs and nanowire transistors, using drift-diffusion-based approaches. Given the prohibitive fabrication costs and sensitivity of nanodevices, simulation-driven design using MuGFET has emerged as a crucial strategy for minimising experimental losses and optimising device architectures (Atalla et al., 2019, 2020, 2023; Das et al., 2018; Hashim, 2022a; Kumar et al., 2020).

To accurately reproduce the transfer characteristics (I_d-V_g) of Si-FinFETs at varying temperatures, the MuGFET simulations utilise a set of carefully selected parameters, including the doping concentrations of the channel (P-type) and source/drain regions (N-type), channel width, source/drain lengths, and gate length. Of particular interest are the oxide thickness and gate length, both critical determinants of device scaling behaviour. Figure 3 depicts a representative Si-FinFET simulation environment (Mehrdad et al., 2022; Nasri et al., 2022).

For this investigation, the I_d-V_g characteristics were simulated at temperatures of 275 K, 300 K, 325 K, and 350 K, across gate lengths (L_g) of 10 nm, 20 nm, and 30 nm. The oxide thickness was fixed at 1 nm, with source and drain lengths set at 5 nm each. Both the source and drain regions were uniformly doped as N-type, while the channel width was maintained at 10 nm with a doping concentration of 10^{19} cm^{-3}.

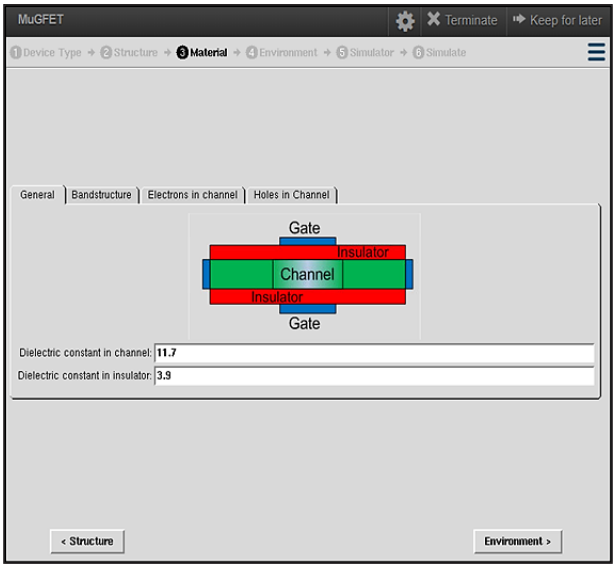


Figure 3. Homepage of MuGFET simulation tool

RESULTS AND DISCUSSION

The variation in output current (ΔI) for gate lengths (L_g) of 10, 20, and 30 nm at a fixed gate width (W_g) of 10nm is illustrated in Figures 4 through 6 as a function of temperature, with the supply voltage (V_{DD}) swept from 0 to 1 V in 0.1 V increments. The graphs clearly exhibit a linear decline in current with increasing temperature across all V_{DD} values, indicating a strong inverse temperature dependence. Notably, the peak temperature sensitivity, represented by the maximum current response, is observed at relatively lower temperatures. Figures 4 and 5 highlight that the highest temperature sensitivity occurs at V_{DD} = 0.6 V for both L_g = 10 nm and L_g = 20 nm, respectively, whereas Figure 6 reveals a comparatively lower response for L_g = 30 nm under the same biasing condition. These findings collectively underscore that V_{DD} = 0.6 V represents an optimal operating point for Si-FinFET thermal sensing, providing stable and maximised sensitivity within the critical temperature window of 325-350 K.

The variations in ΔI with decreasing V_{DD} at temperatures of 275 K, 300 K, 325 K, and 350 K for L_g of 10, 20, and 30 nm are depicted in Figures 7 through 9. The highest temperature sensitivity, represented by the peak ΔI, was observed at V_{DD} = 0.1 V for L_g = 10 nm, V_{DD} = 0.3 V for L_g = 20 nm, and V_{DD} = 0.4 V for L_g = 30 nm. These findings indicate that ΔI am strongly dependent on both increasing temperature and V_{DD}, with an optimum operating condition identified at V_{DD} = 0.4 V, L_g = 30 nm, and W_g = 10 nm.

To elucidate the sensing performance more comprehensively, temperature sensitivity was evaluated by analysing the slope of the ΔI–T characteristics across different gate lengths.

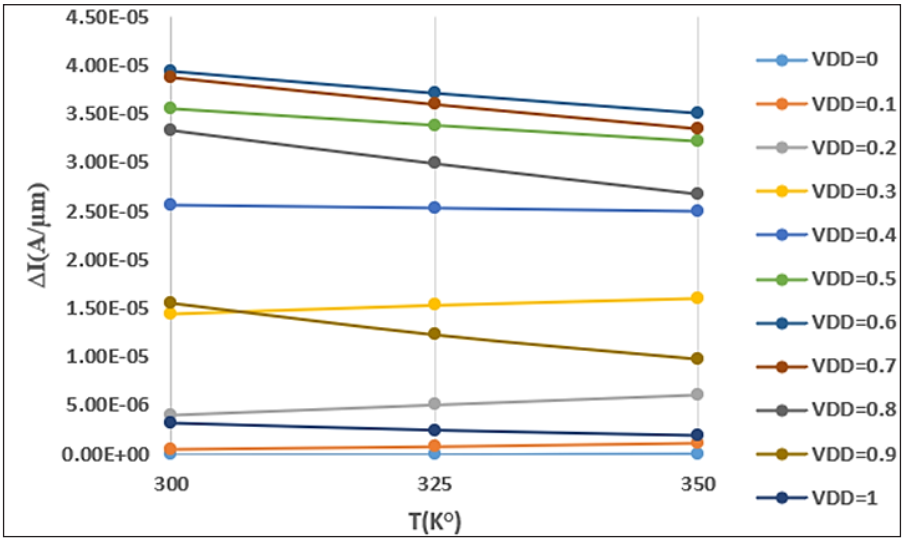


Figure 4. ΔI–T characteristics of Si-FinFET (W_g=10nm, L_g = 10nm)

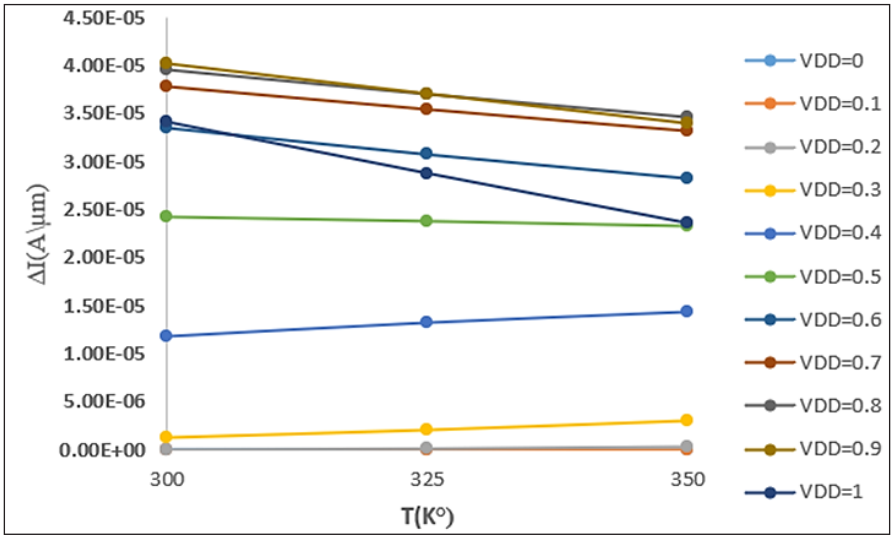


Figure 5. ΔI–T characteristics of Si-FinFET (W_g =10nm, L_g=20nm)

The results clearly reveal that shorter gate lengths, particularly L_g = 10 nm, exhibit a steeper slope, signifying an enhanced response to temperature fluctuations and superior thermal sensitivity. Conversely, as the gate length increases, the ΔI-T curves display a more subdued gradient, indicating diminished sensitivity. This trend underscores the pivotal influence of channel length on the thermal responsiveness of Si-FinFET-based nanoscale temperature sensors, affirming the strategic advantage of ultra-scaled devices in high-precision thermal sensing applications.

Figure 10 highlights the optimal operating voltage (V_{DD}) correlated with the temperature sensitivity peaks identified in Figures 7 to 9, based on variations in channel length and thermal responsiveness. Although the channel length increases only incrementally from 10 to 30 nm, a marked enhancement in temperature sensitivity is observed at L_g = 10 nm. Beyond this, a quasi-linear decline in sensitivity is noted as the gate length extends from 10 to 20 nm and further to 30 nm. These observations confirm that the most favourable dimensional configurations for thermal sensitivity are attained at L_g = 10 nm and L_g = 20 nm.

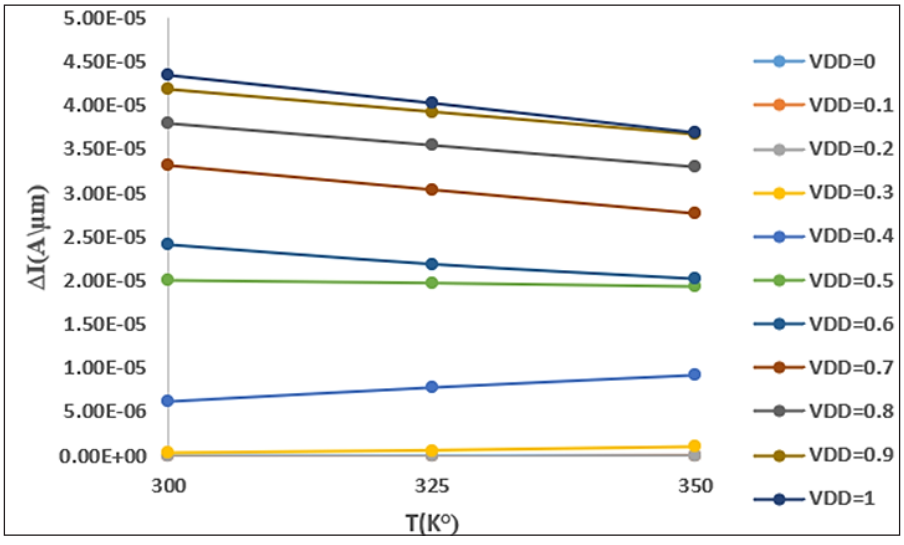


Figure 6. ΔI–T characteristics of Si-FinFET (W_g=10nm, L_g=30 nm)

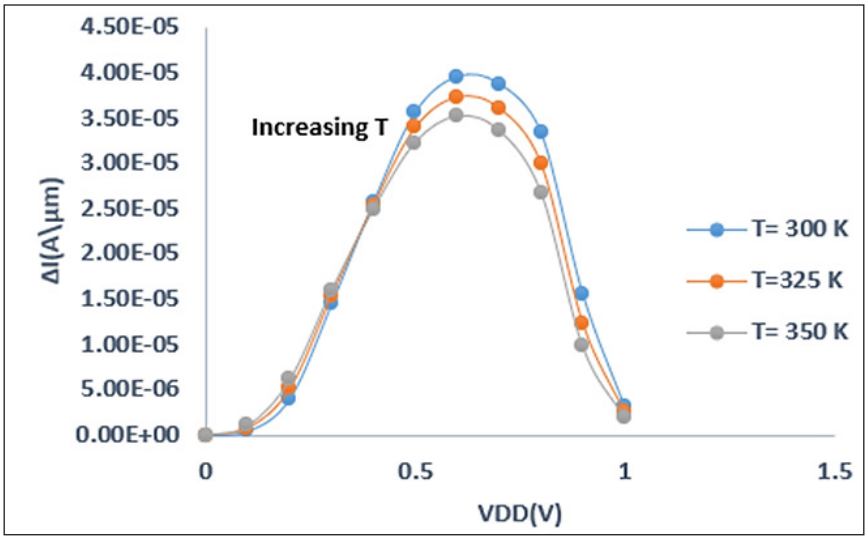


Figure 7. ΔI–V_{DD} characteristics of Si-FinFET (W_g=10nm/ L_g = 10nm)

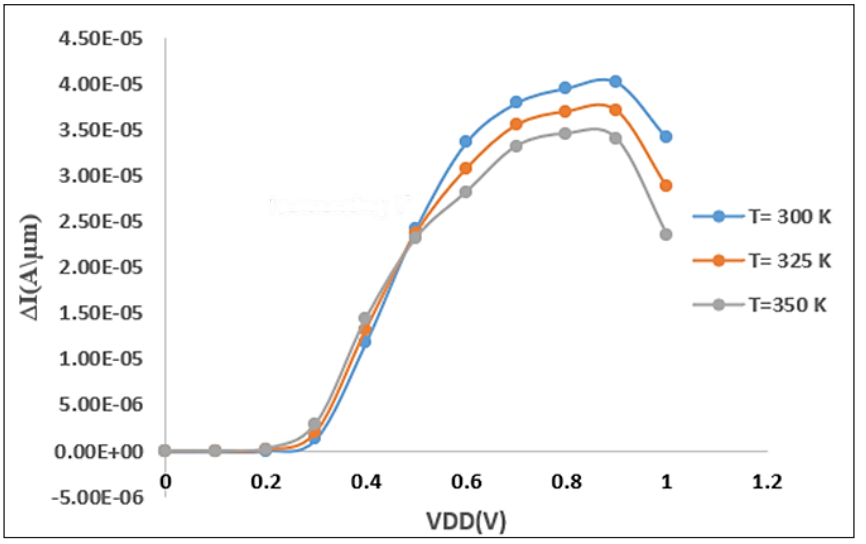


Figure 8. ΔI–V_{DD} characteristics of Si-FinFET (W_g=10nm, L_g = 20nm)

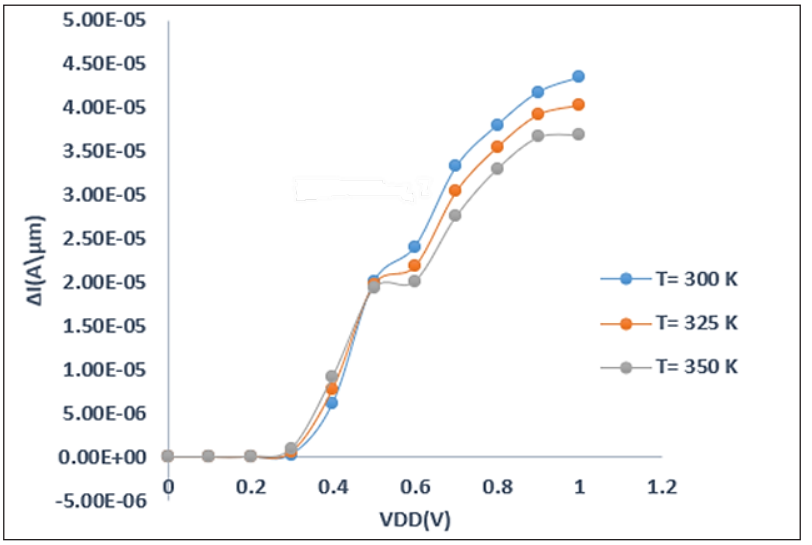


Figure 9. ΔI–V_{DD} characteristics of Si-FinFET (W_g = 10 nm, L_g = 30 nm)

The consolidated results presented across Figures 4 to 10 reveal a definitive trend in the temperature-dependent current behaviour of Si-FinFETs. Specifically, increasing temperature consistently induces a monotonic reduction in ΔI, attributed to intensified phonon scattering that degrades carrier mobility. However, the magnitude of this thermal response, quantified as the slope of the ΔI–T characteristic, varies with gate length, with L_g = 10 nm exhibiting the most pronounced sensitivity.

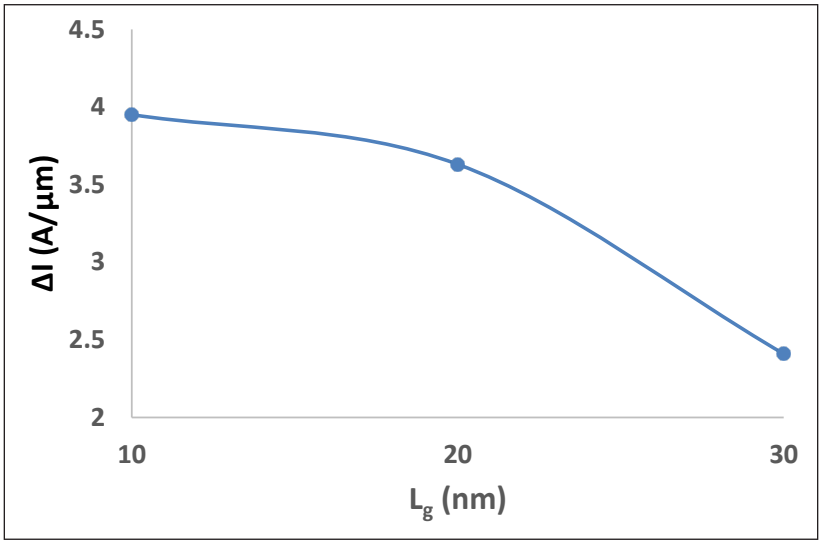


Figure 10. ΔI variation with channel length (L_g) for optimal temperature sensitivity and V_{DD} tuning

This can be ascribed to the amplified electric field intensity and stronger confinement effects within ultra-scaled channels, which intensify temperature-dependent modulation.

Additionally, the relationship between ΔI and V_{DD}, as illustrated in Figures 7 to 9, demonstrates a nonlinear dependency wherein a moderate bias range ($V_{DD} \approx 0.3–0.6$ V) produces maximal ΔI. Beyond this optimal biasing window, the current saturates, suggesting diminishing returns in sensor performance at higher voltages. Figure 10 effectively consolidates these findings by delineating the precise combinations of gate length and supply voltage that yield peak temperature sensitivity. The observed near-linear attenuation of ΔI with increasing gate length reinforces the conclusion that ultra-scaled Si-FinFETs are inherently more responsive to temperature variations, albeit with trade-offs concerning leakage currents and electrostatic control. These insights are instrumental in the strategic optimisation of Si-FinFET-based nanoscale thermal sensors, offering valuable design guidance on the selection of geometrical and electrical parameters to achieve a judicious balance between high thermal sensitivity and device stability.

Figure 11 presents the temperature-dependent behaviour of key performance parameters in Si-FinFETs, namely, drain-induced barrier lowering (DIBL), subthreshold swing (SS), and threshold voltage (V_T), measured at temperatures of 275 K, 300 K, 325 K, and 350 K with a gate length (L_g) of 10 nm. A clear linear decline in V_T is observed as the temperature increases: for instance, V_T drops from 0.28 V at 275 K to 0.22 V at 350 K, indicative of thermally induced carrier activation and reduced threshold control.

The SS, which ideally approaches 69.5 mV/dec, starts at 120.39 mV/dec at 275 K and progressively increases, peaking at 161.63 mV/dec at 350 K, highlighting the degradation in

gate control over the channel at elevated temperatures. Notably, this deviation underscores the sensitivity of SS to thermally induced interface trap states and leakage current contributions. Additionally, DIBL exhibits a monotonic increase with temperature, further confirming the deterioration of electrostatic integrity under thermal stress.

Overall, these findings reveal that as ambient temperature rises, Si-FinFETs experience pronounced shifts in SS and V_T, with DIBL amplifying concurrently, signalling increased short-channel effects and weakened gate-channel coupling. The extreme SS deviation reaching 133.61 mV/dec at high temperature, compared to the ideal value of 54.6 mV/dec, emphasises the critical importance of thermal management in ultra-scaled transistor technologies.

Figure 12 illustrates the temperature-dependent variations in V_T, DIBL, and SS for the Si-FinFET device across temperatures of 275 K, 300 K, 325 K, and 350 K. A clear linear trend is observed over the examined temperature range: V_T decreases from 0.46 V to 0.41 V, SS increases from 89.67 mV/dec to 115.81 mV/dec, and DIBL rises from 34.71 mV/V to 48.45 mV/V, respectively.

Interestingly, the SS value at 275 K (89.67 mV/dec) represents the largest deviation from the ideal SS of 54.6 mV/dec. Despite this, 275 K also corresponds to the condition closest to the theoretical SS limit of 69.5 mV/dec, before SS continues to rise and peaks at 115.81 mV/dec at 350 K. This degradation in SS with temperature reflects the reduced gate control due to increased thermally generated carriers and interface trap activity.

In parallel, the observed increase in DIBL with temperature further indicates worsening short-channel effects and diminished electrostatic gate control. Collectively,

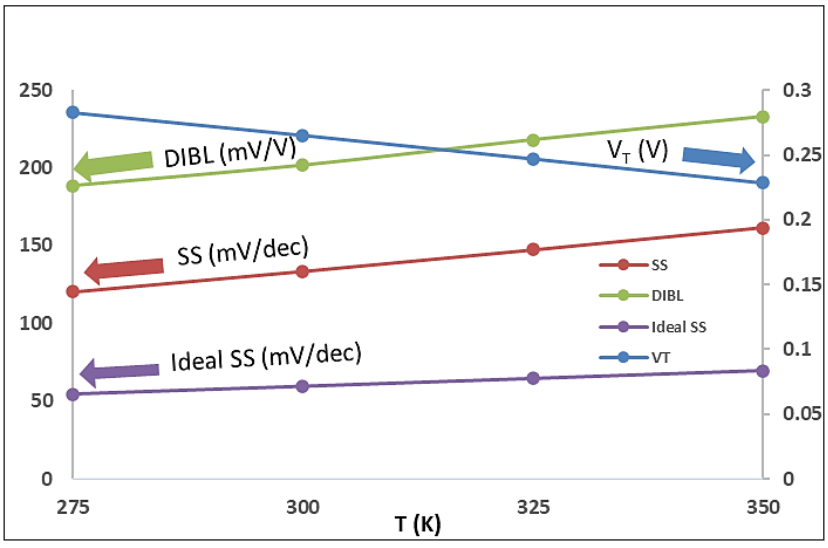


Figure 11. Plots of DIBL, SS, ideal SS, and V_T at Wg =10nm, Lg = 10nm

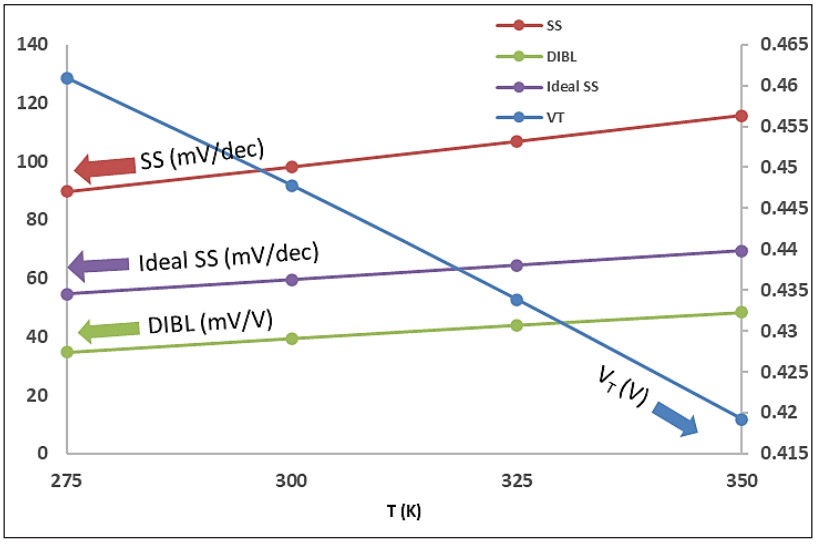


Figure 12. Plots of DIBL, SS, ideal SS, and V_T at W_g=10nm, L_g = 20 nm

these findings emphasise the thermally sensitive nature of FinFET parameters and the need for precise thermal management and design optimisation in advanced nanoscale devices.

Figure 13 elucidates the interrelationship between variations in DIBL, SS, V_T, and how these parameters evolve with temperature for a Si-FinFET device with gate length (L_g) of 20 nm across the range of 275 K to 350 K. A linear decrement in V_T is observed with increasing temperature, with V_T values reducing from 0.48 V at 275 K to 0.46 V at 350 K. Correspondingly, SS increases from 74.22 mV/dec to 99.93 mV/dec, while DIBL rises from 8.29 mV/V to 13.37 mV/V. The SS value of 74.22 mV/dec at 275 K deviates significantly from the ideal SS of 54.6 mV/dec, and this deviation becomes more pronounced at 350 K, where SS reaches 99.93 mV/dec, representing the furthest divergence from the theoretical optimum of 69.5 mV/dec. This degradation in SS reflects the deterioration of gate-channel control under thermal excitation, driven by enhanced carrier scattering and interface state activation. Simultaneously, the monotonic increase in DIBL with temperature highlights the exacerbation of short-channel effects and further reduction in electrostatic gate control. These findings confirm that as temperature rises, Si-FinFETs with L_g = 20 nm exhibit pronounced shifts in key electrical parameters, emphasising the critical influence of thermal variations on device performance and reliability.

Figure 14 presents the variations in V_T, SS, and DIBL as a function of Si-FinFET channel length at 300 K, with L_g incremented in 10 nm steps from 10 nm to 30 nm. As the channel length increases, SS exhibits a decreasing trend and approaches its ideal value at 30 nm, indicating improved gate control. Simultaneously, V_T demonstrates a gradual increase with channel length and eventually saturates, reflecting enhanced electrostatic stability.

Interestingly, DIBL shows a sharp peak at 10 nm, indicative of pronounced short-channel effects, and subsequently decreases with further scaling. These results suggest that a channel length range between 10 and 20 nm offers optimal performance under the investigated thermal conditions. Moreover, the linear correlation between channel length and the observed electrical properties reinforces previous findings reported by Hashim (2022b).

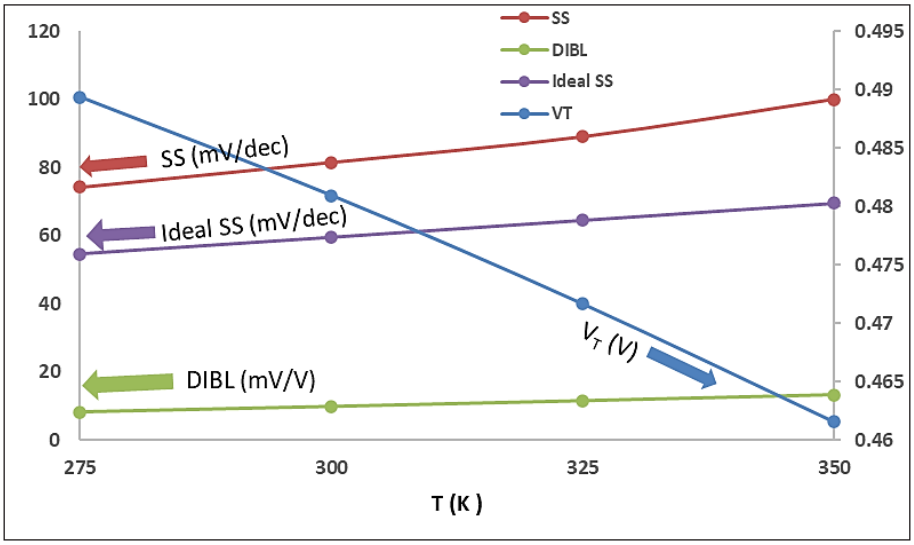


Figure 13. Plots of DIBL, SS, ideal SS, and V_T at W_g =10nm, L_g = 30 nm

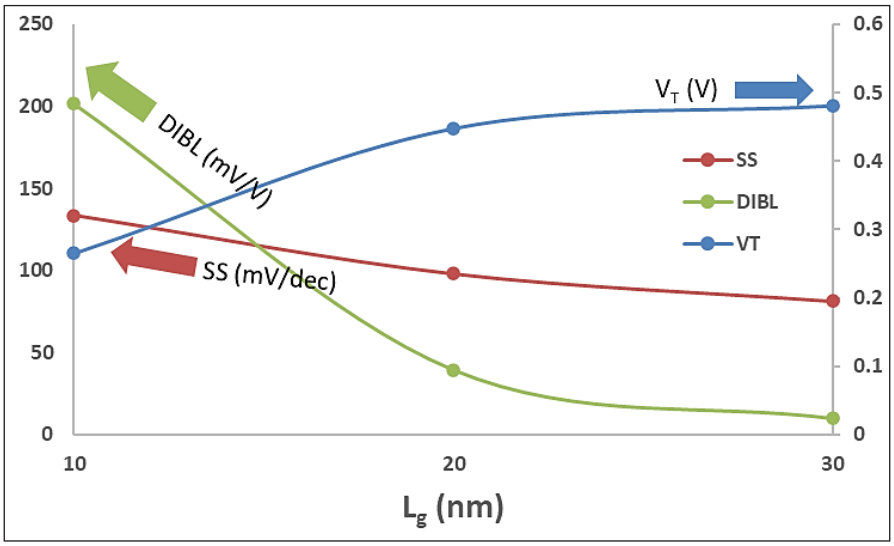


Figure 14. Plots of DIBL, SS and V_T with L_g

The dominant impact of temperature on Si-FinFET drain current is primarily attributed to enhanced carrier velocity in n-type Si-FinFETs as fin dimensions are scaled down. Remarkably, cross-sectional scaling of the channel exerts minimal influence on carrier velocity, underscoring the robustness of transport dynamics in ultra-scaled geometries. Furthermore, in aggressively downscaled MOSFETs, the drain saturation current is predominantly governed by the carrier injection velocity (V_{inj}) at the source-channel junction. Under ballistic transport conditions, this injection velocity, an intrinsic material property, becomes the primary determinant of drive current, superseding traditional mobility-limited conduction mechanisms (Yasir, 2024).

Table 1 provides a comparative overview of the most relevant prior studies and the present work. This comparison highlights key performance metrics of subthreshold swing (SS) and drain-induced barrier lowering (DIBL) across Si-FinFET devices with various gate lengths and fabrication approaches. The presented data underscore the improvements achieved in our study, particularly in terms of thermal stability and electrostatic control, while situating our findings within the broader context of existing literature.

Table 1
Comparison with other research works

Reference	Best Value of SS	Best Value of DIBL
Yu et al. (2002)	125 mV/dec	71mV/V
Liu et al. (2020)	80 mV/dec	10 mV/V
Jiang et al. (2009)	73 mV/dec	28mV/V
Jiang et al. (2009)	60 mV/dec	10 mV/V
Present study	89.6mV/dec	34.7mV/V

CONCLUSION

By systematically examining Si-FinFET channel lengths of 10, 20, and 30 nm, this study investigates the influence of temperature variations (275 K, 300 K, 325 K, and 350 K) on key device characteristics. Under diode-mode operation, extending the channel length up to 30 nm yields optimal increments in ΔI with respect to temperature; beyond this point, ΔI values stabilise, exhibiting minimal dependence on further channel length increases. At each temperature level, the behaviour of V_T, SS, and DIBL was analysed across the specified channel lengths. Notably, V_T exhibits a positive correlation with channel length, gradually increasing and approaching saturation beyond 20 nm. In contrast, SS decreases with increasing L_g, nearing the ideal theoretical value indicative of improved gate-channel coupling. Meanwhile, DIBL rises sharply at 10 nm, exhibits marginal growth at 20 nm, and begins to decrease at 30 nm, signifying a partial recovery in electrostatic integrity. Collectively, these findings suggest that the optimal channel length for Si-FinFET devices, under the thermal and electrical conditions considered in this study, lies within the 10 and 20 nm range.

This dimensional window strikes a favourable balance between thermal sensitivity, electrostatic control, and current modulation, making it ideal for nanoscale Si-FinFET-based temperature sensor applications.

ACKNOWLEDGEMENT

The authors would like to express sincere gratitude to Universiti Teknologi MARA, Malaysia, and A'Sharqiyah University, Oman, for their support in accomplishing this research work.

LIST OF ABBREVIATIONS

Si-FinFET	:	Silicon-based Fin Field-Effect Transistor
MuGFET	:	Multi-Gate Field-Effect Transistor
I-V	:	Current-voltage
MOSFET	:	Metal Oxide Field-Effect Transistor
FET	:	Field-Effect Transistor
BJT	:	Bipolar Junction Transistor
Si	:	Silicon
Ge	:	Germanium
GaAs	:	Gallium arsenide
InAs	:	Indium arsenide
DIBL	:	Drain-induced barrier lowering
V_T	:	Threshold voltage
SS	:	Sub-threshold swing
PDE	:	Partial differential equation
I_d	:	Drain current
V_g	:	Gate voltage
L_g	:	Gate length
W_g	:	Gate width
V_{DD}	:	Supply voltage
ΔI	:	Differential current
T	:	Temperature
V_{inj}	:	Injection velocity

REFERENCES

Ahangari, Z., Asadi, E., & Hosseini, S. A. (2022). Performance optimisation and sensitivity analysis of junctionless FinFET with asymmetric doping profile. Journal of Nanoanalysis, 7(4), 310-320.

Alvarado, J., Tinoco, J. C., Salas, S., Martinez-Lopez, A. G., Soto-Cruz, B. S., Cerdeira, A., & Raskin, J. P. (2013). SOI FinFET compact model for RF circuits simulation. In 2013 IEEE 13th Topical Meeting

on Silicon Monolithic Integrated Circuits in RF Systems (pp. 87–89). IEEE. https://doi.org/10.1109/SiRF.2013.6489441

Ariga, K., Yamauchi, Y., Rydzek, G., Ji, Q., Yonamine, Y., Wu, K. C., & Hill, J. P. (2014). Layer-by-layer nanoarchitectonics: Invention, innovation, and evolution. Chemistry Letters, 43(1), 36-68. https://doi.org/10.1246/cl.130987

Atalla, S., Tarapiah, S., Gawanmeh, A., Daradkeh, M., Mukhtar, H., Himeur, Y., Mansoor, W., Hashim, K. F. B., & Daadoo, M. (2023). IoT-enabled precision agriculture: Developing an ecosystem for optimised crop management. Information, 14(4), Article 205. https://doi.org/10.3390/info14040205

Atalla, Y., Hashim, Y., Abd Ghafar, A. N., & Jabbar, W. A. (2019). A temperature characterisation of Si-FinFET based on channel oxide thickness. Telecommunication Computing Electronics and Control, 17(5), 2475-2480. https://doi.org/10.12928/telkomnika.v17i5.11798

Atalla, Y., Hashim, Y., Abd Ghafar, A. N., & Jabbar, W. A. (2020). Temperature characteristics of FinFET based on channel fin width and working voltage. International Journal of Electrical and Computer Engineering, 10(6), 5650-5657. https://doi.org/10.11591/ijece.v10i6.pp5650-5657

Bescond, M., Nehari, K., Autran, J. L., Cavassilas, N., Munteanu, D., & Lannoo, M. (2004). 3D quantum modelling and simulation of multiple-gate nanowire MOSFETs. In IEDM Technical Digest. IEEE International Electron Devices Meeting, 2004 (pp. 617-620). IEEE.https://doi.org/10.1109/IEDM.2004.1419237

Das, R. R., Maity, S., Choudhury, A., Chakraborty, A., Bhunia, C. T., & Sahu, P. P. (2018). Temperature-dependent short-channel parameters of FinFETs. Journal of Computational Electronics, 17, 1001-1012. https://doi.org/10.1007/s10825-018-1212-y

Dixit, A., Samajdar, D. P., & Sharma, D. (2021). Performance analysis of $Ga_{0.47}In_{0.53}Sb$-FinFET and Si-FinFET for RF and low-power design applications. In Computers and Devices for Communication: Proceedings of CODEC 2019 (pp. 533-538). Springer Singapore. https://doi.org/10.1007/978-981-15-8366-7_78

Doghish, M. Y., & Ho, F. D. (1992). A comprehensive analytical model for metal-insulator-semiconductor (MIS) devices. IEEE Transactions on Electron Devices, 39(12), 2771-2780. https://doi.org/10.1109/16.168723

Fahad, H. M., Hu, C., & Hussain, M. M. (2015). Simulation study of a 3-D device integrating FinFET and UTBFET. *IEEE Transactions on Electron Devices, 62*(1), 83-87. https://doi.org/10.1109/TED.2014.2372695

Hashim, Y. (2017). A new approach for dimensional optimisation of inverters in 6T-static random-access memory cell based on silicon nanowire transistor. *Journal of Nanoscience and Nanotechnology, 17*(2), 1061-1067. https://doi.org/10.1166/jnn.2017.12608

Hashim, Y. (2022a). Temperature characteristics of 10 nm N- and P-channel Si-FinFET structure. In *2022 International Conference on Electrical Engineering, Computer and Information Technology* (pp. 104-107). IEEE. https://doi.org/10.1109/ICEECIT55908.2022.10030376

Hashim, Y. (2022b). The impact of channel FIN width on electrical characteristics of Si-FinFET. *International Journal of Electrical and Computer Engineering, 12*(1). https://doi.org/10.11591/ijece.v12i1.pp201-207

Hashim, Y., & Sidek, O. (2011). Temperature effect on IV characteristics of Si nanowire transistor. In *2011 IEEE Colloquium on Humanities, Science and Engineering* (pp. 331-334). IEEE. https://doi.org/10.1109/CHUSER.2011.6163744

Hashim, Y., & Sidek, O. (2012). Effect of temperature on the characteristics of a silicon nanowire transistor. *Journal of Nanoscience and Nanotechnology, 12*(10), 7849-7852. https://doi.org/10.1166/jnn.2012.6598

Jabbara, W. A., Mahmood, A., & Sultan, J. (2022). Modelling and characterisation of optimal nano-scale channel dimensions for fin field effect transistor based on constituent semiconductor materials. *Telecommunication Computing Electronics and Control, 20*(1), 221-234. https://doi.org/10.12928/telkomnika.v20i1.21671

Jaisawal, R. K., Rathore, S., Kondekar, P. N., Yadav, S., Awadhiya, B., Upadhyay, P., & Bagga, N. (2022). Assessing the analogue/RF and linearity performances of FinFET using high threshold voltage techniques. *Semiconductor Science and Technology, 37*(5), Article 055010. https://doi.org/10.1088/1361-6641/ac6128

Jiang, Y., Liow, T. Y., Singh, N., Tan, L. H., Lo, G. Q., Chan, D. S., & Kwong, D. L. (2009). Nickel salicided source/drain extensions for performance improvement in ultrascaled (sub-10 nm) Si-nanowire transistors. *IEEE Electron Device Letters, 30*(2), 195-197. https://doi.org/10.1109/LED.2009.2010532

Kesherwani, S., Daga, M., & Mishra, G. P. (2022). Design of sub-40 nm FinFET-based label-free biosensor. *Silicon, 14*(18), 12459-12465. https://doi.org/10.1007/s12633-022-01936-9

Kumar, H., Jethwa, M. K., Porwal, A., Dhavse, R., Devre, H. M., & Parekh, R. (2021). Effect of different channel materials on the performance parameters for FinFET device. In *Proceeding of Fifth International Conference on Microelectronics, Computing and Communication Systems: MCCS 2020* (pp. 275-288). Springer Singapore. https://doi.org/10.1007/978-981-16-0275-7_23

Li, Y., Zhao, F., Cheng, X., Liu, H., Zan, Y., Li, J., Zhang, Q., Wu, Z., Luo, J., & Wang, W. (2021). Four-period vertically stacked SiGe/Si channel FinFET fabrication and its electrical characteristics. *Nanomaterials, 11*(7), Article 1689. https://doi.org/10.3390/nano11071689

Liao, C. N., Chen, C., & Tu, K. N. (1999). Thermoelectric characterisation of Si thin films in silicon-on-insulator wafers. *Journal of Applied Physics, 86*(6), 3204-3208. https://doi.org/10.1063/1.371190

Liu, C., Sagong, H. C., Kim, H., Choo, S., Lee, H., Kim, Y., Kim, H., Jo, B., Jin, M., Kim, J., & Ha, S. (2015). Systematical study of 14 nm FinFET reliability: From device level stress to product HTOL. In *2015 IEEE International Reliability Physics Symposium* (pp. 2F-3). IEEE. https://doi.org/10.1109/IRPS.2015.7112693

Lu, W., Kim, J. K., Klem, J. F., Hawkins, S. D., & del Alamo, J. A. (2015). An InGaSb p-channel FinFET. In *2015 IEEE International Electron Devices Meeting* (pp. 31-36). IEEE. https://doi.org/10.1109/IEDM.2015.7409810

Liu, M., Zhang, X., Yang, Y., Li, L., Wang, G., Xie, H., Deng, S., Xu, G., & Hao, Y. (2020). Vertical Ge gate-all-around nanowire pMOSFETs with a diameter down to 20 nm. *IEEE Electron Device Letters, 41*(4), 533-536. https://doi.org/10.1109/LED.2020.2971034

Mehrdad, F., & Ahangari, Z. (2022). Design and simulation of a gas sensitive junctionless FinFET based on conducting polymer as the gate material. *Physica Scripta, 97*(7), Article 075805. https://doi.org/10.1088/1402-4896/ac73bf

Meijer, G. C., Wang, G., & Fruett, F. (2001). Temperature sensors and voltage references implemented in CMOS technology. *IEEE Sensors Journal, 1*(3), 225-234. https://doi.org/10.1109/JSEN.2001.954835

Nasri, F., Rekik, N., Bahri, H., Farooq, U., Hussein, A. W., Affan, H., Alabid, A., & Ouari, B. (2022). *A deterministic physical approach for elucidating the temperature effects on electrical responses of FinFET transistor*. https://doi.org/10.2139/ssrn.4299934

Nasri, F., Rekik, N., Bahri, H., Farooq, U., Hussein, A. W. M., Affan, H., Alabid, A., & Ouari, B. (2023). Temperature effects on electrical response of FinFET transistors in the static regime. *IEEE Transactions on Electron Devices, 70*(4), 1595-1600. https://doi.org/10.1109/TED.2023.3248537

Park, H. H., Zeng, L., Buresh, M., Wang, S., Klimeck, G., Mehrotra, S. R., Heitzinger, C., & Haley, P. B. (2021). *Simulate 3D nanowire transport in the effective mass approximation with phonon scattering and 3D Poisson self-consistent solution*. nanoHUB.

Yu, B., Chang, L., Ahmed, S., Wang, H., Bell, S., Yang, C. Y., Tabery, C., Ho, C., Xiang, Q., King, T. J., & Bokor, J. (2002). FinFET scaling to 10 nm gate length. In *Digest. International Electron Devices Meeting* (pp. 251-254). IEEE.

Pertanika J. Sci. & Technol. 34 (3): 1451 - 1468 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Design and Development of Unmanned Aerial Vehicles' Propellers Based on Biomimetic Design

Saundarya Saran[1], Rosalam Che Me[1]*, Mohamed Thariq Hameed Sultan[2,3,4]*, and Navaneetha Krishna Chandran[2]

[1]Department of Industrial Design, Faculty of Design and Architecture, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
[2]Laboratory of Biocomposite Technology, Institute of Tropical Forestry and Forest Products (INTROP), Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
[3]Aerospace Malaysia Innovation Centre (944751-A), Prime Minister's Department, MIGHT Partnership Hub, Jalan Impact, 63000 Cyberjaya, Selangor, Malaysia
[4]Department of Aerospace Engineering, Faculty of Engineering, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

ABSTRACT

This study aims to enhance UAV propeller performance by addressing aerodynamic inefficiencies caused by shock waves at high rotational speeds, which are also known as vortex formation. These shock waves, occurring at the propeller tips, increase drag, vibration, noise, and potentially can reduce flight time. The research focuses on optimising propeller design to minimise these effects and improve overall efficiency. The methodology involved a comprehensive review of existing propeller designs and the application of aerodynamic principles to develop a propeller approximately 9 inches in length. The background study also focuses on the application of biomimicry in the design of propellers. Simulation tests were conducted to evaluate its performance, with measurements taken at various rotational speeds, including a maximum rotational speed of 6000 rpm. The results demonstrated that the newly designed propeller achieved better efficiency compared to existing models, producing a thrust of 0.0016 N at maximum speed. The design also exhibited reduced noise and vibration, contributing to a more stable and efficient flight performance. The findings underscore the potential for improved UAV propeller designs to increase the efficiency of performance and enhance operational capabilities, particularly in applications like agriculture, inspection, and surveillance. Future research may explore additional modifications to further optimise aerodynamic performance and noise reduction.

Keywords: Computational fluid dynamics, efficiency, propeller, Shock Waves, unmanned aerial vehicles, vortex

ARTICLE INFO

Article history:
Received: 05 November 2025
Accepted: 18 May 2026
Published: 12 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.08

E-mail addresses:
saundaryasaran28@gmail.com (Saundarya Saran)
rosalam@upm.edu.my (Rosalam Che Me)
thariq@upm.edu.my (Mohamed Thariq Hameed Sultan)
navaneetha@upm.edu.my (Navaneetha Krishna Chandran)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

INTRODUCTION

Unmanned aerial vehicles (UAVs) have made it possible to revolutionise different areas by offering flexible, cost-effective and efficient remedies to various tasks. UAVs have become invaluable instruments that enhance the area of operation and have advantages over the conventional methods, including precision agriculture as well as environmental management, surveillance, logistics and disaster management. Some advancements made on various parts have also led to the rapid adoption of the UAV technology (Du et al.,2024), and one of the fundamental parts is the propeller that directly influences the overall performance, stability and efficiency of the aircraft (Vijayanandh et al., 2022). The efficiency of UAV propellers is of the essence, considering that they generate the force that is required to lift off and remain airborne. However, the propeller performance at fast rotational speeds remains inefficient, which is a continuing problem with the design of UAVs (Xu et al., 2024). This inefficiency is mainly caused by the shock waves that are produced at the ends of the propeller blades (Perikleous et al., 2024). The aerodynamic inefficiencies generated by the shock waves deteriorate the overall performance of the UAV because they augment the levels of drag, vibration, and noise (Lin et al., 2023). These issues are particularly problematic for applications that are required to be very quick, precise, or discrete (Ciliberti et al., 2023). Although the technology of UAVs has undergone massive research and development, there are still aspects of propeller tip vortex designs that need to be improved and optimised to achieve induced drag reduction and noise reduction (Park et al., 2018). The available studies often neglect the complex interactions at the ends of the propellers, and experimental validation should be more thoroughly incorporated to enhance the design and test these designs (Zhang et al., 2024). Consequently, several UAVs can and do work below their capacity, and this limits their utility and effectiveness in performing various tasks (Zhang et al., 2024). Assessment of such properties using propeller bench tests, comparing the performance datum of trading-off designs to the conventional propellers to determine the gains made in thrust and efficiency (Fish & Beneski, 2014; Panagiotopoulos et al., 2024).

Bio-mimicry practice presents human beings with a chance to learn and imitate methods of nature to come up with viable solutions to human technology. The image reveals the nature of the design of drone propellers, which is adopted to resemble the structures of eagle wings, hence the principles of biomimicry. The performance and agility of the eagles in the air, plus the quietness of the manoeuvre, have been made possible by their wing structure and the arrangement of their feathers. Natural attributes incorporated in the UAV propellers result in significant gains in aerodynamic performance. The fusion of adaptation blades in eagle propellers offers efficiency in airflow, such as eagle feathers, that reduces drag to produce less energy consumption and flight duration (Abdulrahim & Lind, 2006).

The bio-mimetic design is effective in controlling the airflow along the propeller tips by its enhanced aerodynamics, which decreases drag and avoids the formation of turbulence and vortices (Liu & Aono, 2009).

Propeller tip shock waves are created as blades pass through supersonic velocity, which significantly decreases aerodynamic efficiency and creates performance issues in UAVs. Besides creating excessive noise, the shock waves cause massive drag forces and high vibration in addition to creating elevated stability problems whilst amplifying structural tension and minimal performance (Carlson & Hassan, 2002; Platzer & Sargent, 2012). The propellers that are biomimetic eagles have a specialised shape of the blade that slows down the formation of shock waves or minimises their occurrence, allowing the propellers to operate steadily and silently. The eagle feathers depict favourable aerodynamics that correct airflow issues to reduce the shock waves as well as promote propeller efficiency and extend the operation span (Choi & Hassan, 2002).

The biomimicry used in UAV propeller development develops a transformative process that combines millions of years of evolution optimisation with the current technology. Eagle wings have aerodynamic characteristics such as feather serrations and flexible designs that reduce turbulent air velocity and allow a sharp acoustic signature of reducing and improving aerodynamic efficiency (Carruthers & Thomas, 2007; Mueller & DeLaurier, 2003). This is because the introduction of biomimicry in the development of the UAV propeller provides enhanced operational performance when operating at very high speeds and reduces the detectability and operational signatures since it enhances speed whilst maintaining low detectability. The criticism of the biomimetic designs is that it is difficult since scientists have problems in closely recreating the complex environmental scenarios and aerodynamic designs or measurements accurately under controlled laboratory conditions (Stanford et al., 2008; Webb & Aono, 2008). The further development of computational fluid dynamics (CFD) and the enhancement of the testing methods should be further developed to overcome current challenges to advance the development of the bio-mimetic unmanned aerial vehicles technology (Zheng et al., 2013).

METHODOLOGY

Research Flow

The methodology of the research is structured and multi-phase to develop and verify optimised UAV propeller designs to increase aerodynamic efficiency and performance. The literature review and analysis of the current UAV propeller designs are the initial step in the process. Through this study, the comprehension regarding bio-mimicry design and how it is applied in industrial products and aerospace products has been gained. It has analysed previous research and determined the gap in it. The review of academic papers, technical reports, and other books related to the topic is done to reveal the lack of knowledge in the

field and which parameters are critical to be known: the pitch of the blade, the blade length, the blade curvature, and the shape of the tip. Descriptive and comparative analyses are conducted to evaluate key metrics such as thrust, torque, acoustic, lift, watt and efficiency. Figure 1 shows the flow chart of the methodology of this study.

Design constraints and standards were established to ensure consistency and comparability across all propeller configurations. The propeller diameter was fixed at 9 inches with a pitch of 4 inches (9x4E), which is commonly used in small UAV applications. The material selection for simulation was assumed to be a standard lightweight polymer suitable for 3D printing. Rotational speeds were limited to 2000, 4000, and 6000 RPM to represent typical UAV operating conditions. Additionally, aerodynamic performance evaluation was conducted under steady-state conditions with uniform airflow, and structural deformation effects were neglected to simplify the analysis.

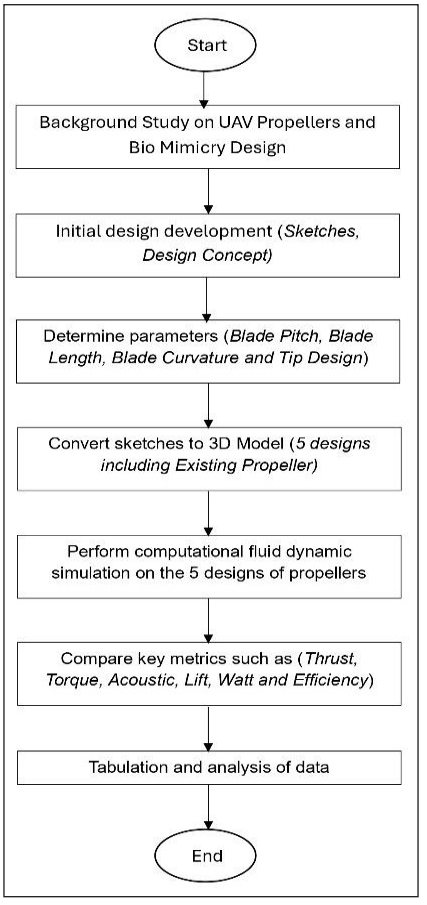


Figure 1. Flowchart of numerical research

In the research, a background study was initiated, in which the working principle of a propeller was learned. The knowledge of the principle of working of the propeller to know the intention of each part of the propeller, which possesses various curves and shapes. The propellers have been sketched in several designs, keeping the basic shape of a propeller and the illustration of the biomimicry of an eagle wing feather. The parameters of propellers were adhered to in the circumstance of controlled designing of 9x4E propellers. The propeller diameter remained 9 inches, and the pitch of the propeller remained at 40. After the drawings of the propellers were sketched, they were converted into SolidWorks computer-aided drawing (CAD) software 3D models.

After completion of the 3D CAD models, the next stage is the computational modelling, validation of the simulation and development of a prototype. Analysis of the dynamics of airflow and optimisation of propeller design was done by means of SolidWorks computational fluid dynamics (CFD) simulations. Several assumptions were made during the CFD analysis.

The airflow was considered incompressible and steady-state, and the effects of turbulence were modelled using standard turbulence models available in SolidWorks CFD. The propeller rotation was simulated under constant angular velocity without transient effects. Environmental conditions such as temperature and air density were assumed constant at standard atmospheric conditions. Structural deformation and material flexibility were not considered, and the propeller was treated as a rigid body. The optimal designs are fabricated into tangible prototypes with the help of 3D printing and evaluated in controlled tests in technical laboratories to evaluate their functionality. To accommodate accuracy and reliability, statistical, comparative, and parametric analyses are used at all points. This combined methodology will bring together the virtual testing and the physical testing by working towards providing innovative UAV propeller designs, which will be as efficient and operative as possible.

Design Concept Development

Propeller designs with high performance and efficient design of propellers have continuously been influenced by nature, especially the anatomy of birds, which is admired for its aerodynamic excellence. In Figure 2, the eagle feather has been designed as the propeller in the tip of the propeller. The propeller has a better shape at the tip that provides better thrust. Moreover, the upper shape of the propeller is inspired by the shape of the eagle when it flies. The biomimetic design was translated into engineering geometry through specific transformations. Feather serrations were mimicked by introducing segmented edges along the propeller tips to reduce vortex intensity. The curvature of eagle wings was translated into varying blade curvature profiles to improve airflow attachment. Additionally, flexibility concepts were incorporated by designing looped and curved structures that simulate adaptive deformation under aerodynamic loading. These transformations were implemented using parametric modelling in SolidWorks.

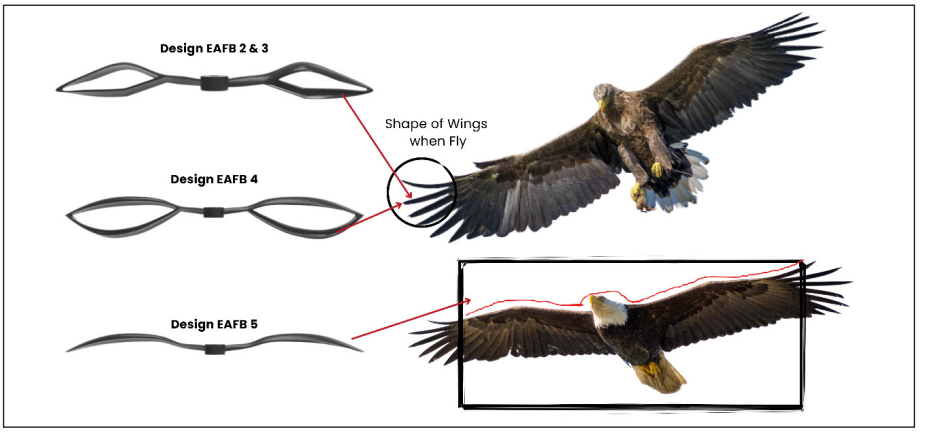


Figure 2. Design inspired by the eagle feather

The design is identified as Eagle Feather Biomimicry (EAFB), which is an approach that moves beyond traditional, functional forms and explores loop-shaped and segmented structures that emulate the aerodynamic behaviour of bird wings. The progression of designs reflects a deliberate shift towards improving lift, reducing drag, and enhancing overall flow efficiency by mimicking the natural adaptability and smooth airflow seen in eagle feathers. This fusion of biological inspiration and aerodynamic engineering showcases an innovative direction in propeller development, one that emphasises not only technical performance but also harmony with natural flight mechanics. Figure 3 illustrates the design concept development of four design propellers inspired by aerodynamic principles and bird anatomy, particularly the wings and feathers of eagles.

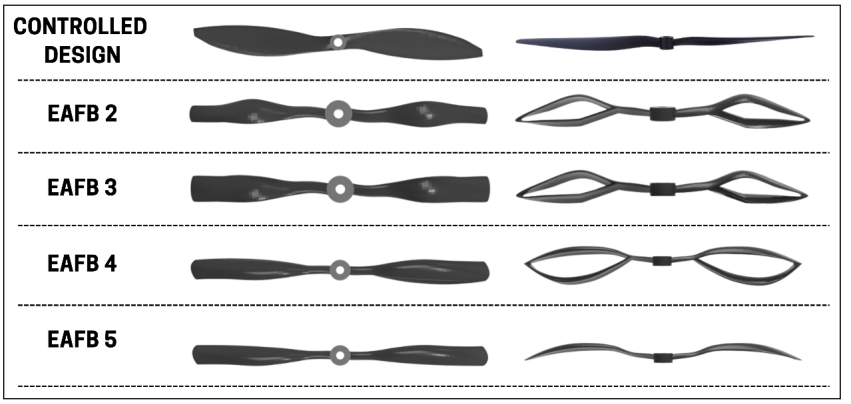
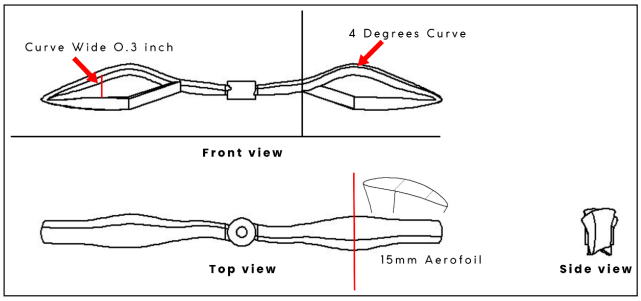


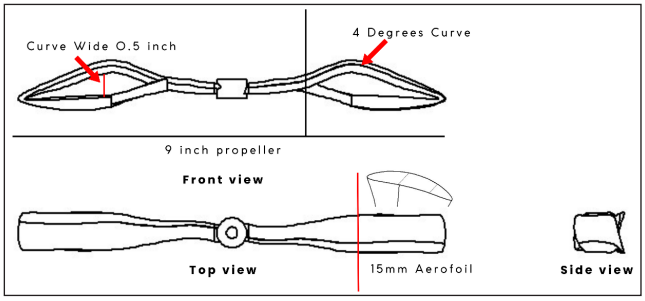
Figure 3. Design development of 9inch propeller

The design is in the form of a changing propeller design beginning with the traditional, functional forms through the bio-mimetic eagle feather forms. The loop designs are aimed at emphasising the shift of traditional blades to designs based on nature to replicate the effectiveness of movement and lift by bird wings. The adopted segmented designs and refined loops imply the struggle to recreate the dynamic adaptability of feathers, which, during movement, vary to provide the best performance. This progressive growth is an indicator of a concern with drag reduction, lift acceleration, and efficiency. Equally, the sharp curves and angles of the designs are influenced by the streamlined shape of the wing to provide an improved circulation of the air. This engineering-meets-nature approach focuses on efficiency, flexibility and environmental compatibility when developing a propeller.

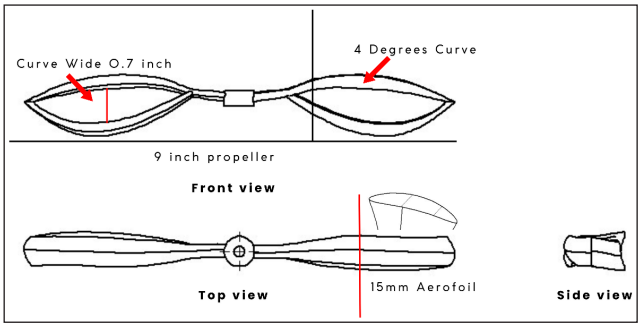
The design of drone propellers also uses biomimicry and willingly takes ideas on how to become better at flying, using the example of birds of prey, such as eagles. The four propeller designs based on eagle-inspired designs illustrated in Figure 4 demonstrate that the various designs have distinct biomimetic characteristics that have been inspired by eagle feathers to increase the performance of the drone.



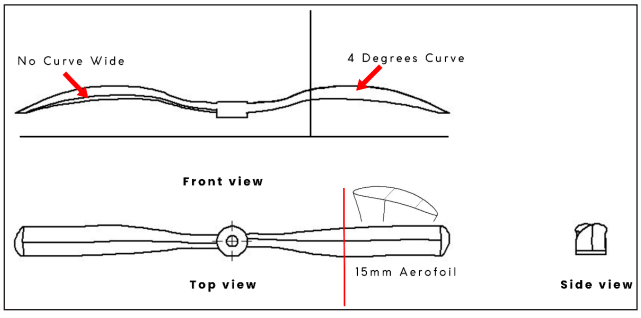
(a)



(b)



(c)



(d)

Figure 4. Four eagle-inspired propeller designs: (a) Curve wide 0.3-inch; (b) Curve wide 0.5-inch; (c) Curve wide 0.7-inch; (d) No curve

These designs are designed to improve the functionality of the drones by raising their aerodynamic efficiency, lowering noise levels, and raising the flight flexibility. The use of loop-shaped curves and streamlined shapes in every propeller model reflects a deliberate transition of traditional blade geometry to a geometry based on nature and its inherent adaptation to different conditions in flight operations, including maximising airflow, reducing drag, and changing the geometry under different conditions of flight. Such a design not only improves the modern drone technology's energy efficiency and flight stability, but it also makes it more adaptable and adjusted to the sensitive or mission-critical applications like wildlife observation, aerial photography, and surveillance. Through this nature-based innovation, these propeller designs offer a new level of performance, reliability, and environmental compatibility.

The primary distinction between the controlled design and the EAFB propellers lies in the overall blade morphology, particularly the incorporation of biomimetic features such as modified blade curvature, tip configuration, and surface profile derived from eagle feather structures. By maintaining a constant airfoil profile across all designs, the study ensures that any observed differences in aerodynamic performance, such as thrust generation, lift characteristics, vortex behaviour, and efficiency, can be directly attributed to the biomimetic geometric adaptations rather than confounding variables.

This approach enhances the validity and replicability of the comparative analysis, as it establishes a controlled experimental framework where the influence of a single design variable, the biomimetic modification, is systematically evaluated. Consequently, the controlled design functions as a reliable benchmark, enabling a clear and meaningful assessment of the performance improvements introduced by the EAFB propeller configurations.

The loop shapes of EAFB-2, EAFB-3 and EAFB-4 are of growing widths (0.3, 0.5, and 0.7 inches, respectively) which emulate the natural curve and shape of eagle feathers. The curves facilitate steering the air around the blades in a smooth manner, generating a lot less turbulence as well as noise in flight. EAFB-3 is one of them, and it is particularly due to its hollow-frame design, which also reduces vibrations and the level of noise. It ensures that the EAFB-3 model is particularly useful in those operations that require almost undisturbed operation, like wildlife research, nature photography or surveillance, where noise and vibration-free operation is paramount.

EAFB-4 (best curve is 0.7 inches) design is flexible, and the concept promotes flexibility like the eagle feathers. The propeller blades have the flexibility that enables them to automatically and effectively adapt to various aerodynamic pressures when flying. Consequently, the use of EAFB-4 propeller enables drones to be predictable even during long-range flights and respond to alterations in the concept or impromptu environment, which offers improved durability and reliability. The EAFB-5 model is different and has a lean design without loops.

Rather, it focuses on low drag by ensuring a smooth, sleek profile based on the aerodynamic features of eagle feathers. This reduced wind drag makes the shape streamline and thus consumes less energy, hence increasing flight duration. It is especially beneficial to specialised missions in which propulsion efficiency and the length of time high are critical. All in all, biomimetic forms inspired by the eagles are a tremendous advancement over conventional propeller designs, which can take advantage of the known principles of aerodynamics in nature, which lead to more efficient, quieter, and reliable drone technology.

RESULTS AND DISCUSSION

The propeller performance study is an important feature of creating an effective and efficient propulsion engine for unmanned aerial vehicles (UAVs). This paper assesses five propeller designs: Controlled Design, the standard propeller and four new ones (EAFB 2, 3, 4 and 5) obtained through computational fluid dynamics (CFD) calculations. It is evaluated based on thrust, torque, acoustic noise, lift, power consumption and efficiency at three rotational speeds (2000 RPM, 4000 RPM and 6000 RPM). The aim is to determine the positive and negative aspects of each design and give recommendations on how the UAV can be optimised to improve its performance.

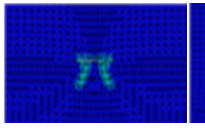
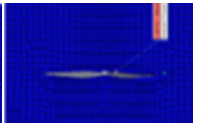
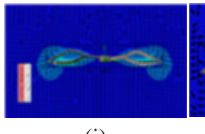
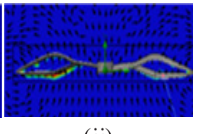
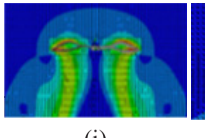
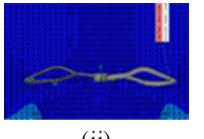
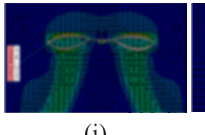
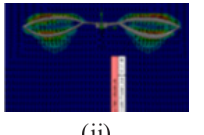
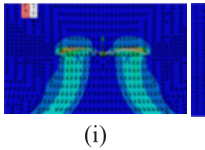
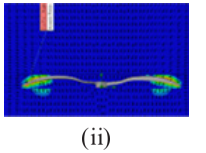
Propellers Performance at 2000 RPM

At 2000 RPM, the Controlled Design (Standard Propeller) has a thrust of 0.114 N with a torque of 0.004 N -m and consumes 0.0003 W of power. Its performance, measured to be 0.0052 N.m/W, is the best amongst all the designs at this RPM. Its lifting force of 6.15 N is, however, much less than the other design alternatives. Table 1 shows the CFD analysis of the propellers at 200 rpm.

EAFB 2 demonstrates a strong advantage in lift production, generating 48.53 N of lift while maintaining low power consumption of 0.0000775 W. This design sacrifices some efficiency, with a value of 0.00478 N·m/W. EAFB 3 achieves the highest thrust of 0.434 N, which is nearly four times greater than the standard propeller, but this improvement comes at the cost of higher power consumption of 0.000242 W. EAFB 4 delivers a balance of moderate thrust of 0.382 N and lift of 33.26 N, while consuming the most power of 0.000542 W at this speed. EAFB 5 exhibits a well-rounded performance, generating a moderate thrust of 0.229 N, lift of 35.47 N, and efficiency of 0.004826 N·m/W.

In the aspect of torque, the existing design of the propeller has the lowest torque compared to the other designs, which is 0.004 N · m. Although the other designs have higher torque compared to the existing design, the difference in torque between design 2 and the existing propeller is slight. The factor can be compensated for by using a higher torque brushless motor. At 2000 RPM, EAFB 3 and 4 are ideal for applications requiring higher thrust, while EAFB 2 is more suitable for lift-focused applications. The standard propeller maintains its edge in efficiency but is less effective in thrust and lift generation compared to the alternative designs.

Table 1
CFD results for 2000 rpm of each propeller design

2000 RPM	Thrust	Torque	Acoustic	Lift	Watt	Efficiency (Torque/Watt)	Percentage Difference %	Lift and Drag	Noise
Controlled Design	0.114	0.004	**6.15**	0.0003	0.765	**0.00522875**	N/A	 (i)	 (ii)
EAFB 2	0.015	0.009	48.53	0.0000775	1.88	0.00478	8.58	 (i)	 (ii)
EAFB 3	**0.434**	0.013	42.7	0.000242	2.657	0.00489	6.48	 (i)	 (ii)
EAFB 4	0.382	0.017	33.26	**0.000542**	3.638	0.004672	10.65	 (i)	 (ii)
EAFB 5	0.229	0.009	35.47	0.0002308	1.865	0.004826	7.70	 (i)	 (ii)

Note. Highlighted the best performance of the designs

Performance of the Propellers at 4000 RPM

As the rotational speed increases to 4000 RPM, the performance differences between the designs become more pronounced. The standard propeller produces 0.469 N of thrust and 3.67 N of lift, consuming 0.001 W of power. Although it retains low acoustic noise levels of 0.014, its efficiency decreases to 0.00232 N·m/W. Table 2 shows the CFD analysis for the propellers at 4000 rpm.

EAFB 2 continues to excel in lift generation, producing 69.06 N while maintaining the lowest power consumption of 0.0008 W. However, its thrust remains minimal of 0.003 N, and efficiency remains relatively low of 0.00236 N·m/W. EAFB 3 delivers the highest thrust of 1.751 N at the cost of the highest power consumption of 0.002 W and significant acoustic noise of 0.049. EAFB 4 matches EAFB 3 in lift production of ~57 N but achieves slightly lower thrust of 1.56 N and consumes equal power. EAFB 5 strikes a balance between thrust of 0.843 N, lift of 57.26 N, and power consumption of 0.006 W, with improved acoustic performance compared to Designs 3 and 4.

EAFB 2, 3, 4 and 5 exhibit higher torque compared to the existing design propeller. This is because the cross-sectional area at the tip of the propeller is higher compared to the existing design. If the tip of the propeller is designed to be thin, it will reduce the load at the tip of the propeller, which eventually will reduce the propeller torque. Still, the higher torque can be compensated for by using a higher torque brushless motor. Although propeller design 3 exhibits higher torque, it produces the highest thrust.

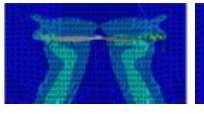
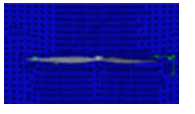
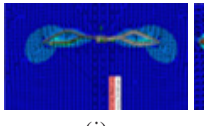
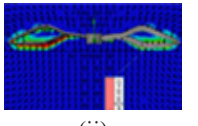
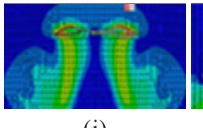
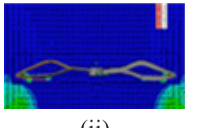
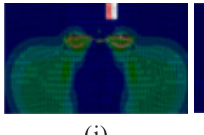
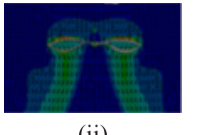
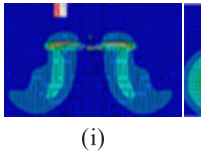
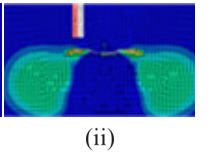
At 4000 rpm, EAFB 3 and 4 demonstrate superior thrust and lift capabilities, making them suitable for high-performance UAV applications. However, EAFB 2 remains an efficient choice for scenarios prioritising lift with minimal power requirements.

Performance of the Propellers at 6000 RPM

At 6000 rpm, the standard propeller generates 0.984 N of thrust and 32.77 N of lift, consuming 0.003 W of power. While its efficiency decreases further of 0.00158 N·m/W, it retains the lowest acoustic noise levels of 0.03. Table 3 shows the CFD analysis of propellers at 6000 rpm.

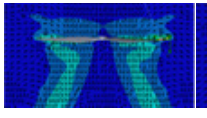
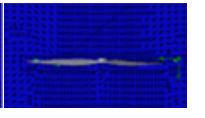
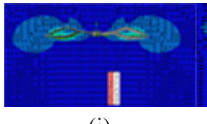
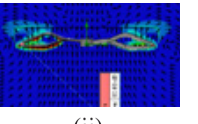
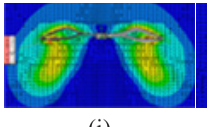
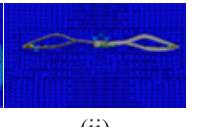
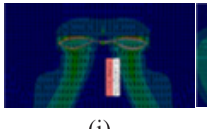
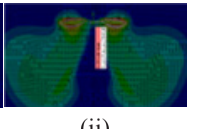
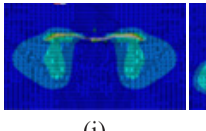
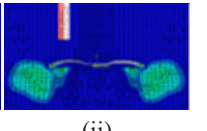
EAFB 2 continues to dominate in lift production, reaching 83.86 N, while maintaining moderate power consumption of 0.006 W. Its efficiency, however, is 0.00158 N-m/W, which is equal to normal propeller. EAFB 3 produces the second-highest thrust of 3.21 N and significant lift of 66.34 N but consumes 0.006 W of power, with notable acoustic noise of 0.098. EAFB 4 outperforms all designs in thrust of 3.447 N and lift of 69.6 N, achieving the best efficiency at this speed of 0.0016 N·m/W. EAFB 5 provides a compelling trade-off, generating a moderate thrust of 1.932 N and the highest lift of 71.56 N while consuming the least power of 0.000693 W.

Table 2
CFD analysis for 4000 rpm of each propeller design

4000 RPM	Thrust	Torque	Acoustic	Lift	Watt	Efficiency (Torque/Watt)	Percentage Efficiency %	Lift and Drag	Noise
Controlled Design	0.469	0.014	**3.67**	0.001	6.03	0.00232	N/A	 (i)	 (ii)
EAFB 2	0.003	0.032	69.06	0.0008	13.52	0.00236	1.72	 (i)	 (ii)
EAFB 3	**1.751**	0.049	57.22	0.002	20.728	0.00236	1.72	 (i)	 (ii)
EAFB 4	1.56	0.069	57.46	0.002	28.852	0.00267	15.09	 (i)	 (ii)
EAFB 5	0.843	0.035	57.26	**0.006**	14.64	**0.00239**	3.02	 (i)	 (ii)

Note. Highlighted the best performance of the design

Table 3
CFD images for 6000 rpm of each propeller design

6000 RPM	Thrust	Torque	Acoustic	Lift	Watt	Efficiency (Torque/Watt)	Percentage Efficiency %	Lift and Drag	Noise
Controlled Design	0.984	0.03	**32.77**	0.003	18.969	0.00158	N/A	 (i)	 (ii)
EAFB 2	0.261	0.085	83.86	**0.006**	53.641	0.00158	0	 (i)	 (ii)
EAFB 3	3.21	0.098	66.34	**0.006**	61.644	0.00159	0.63291139	 (i)	 (ii)
EAFB 4	**3.447**	0.156	69.6	0.004	97.749	**0.0016**	1.26582278	 (i)	 (ii)
EAFB 5	1.932	0.075	71.56	0.000693	47.201	0.00159	0.63291139	 (i)	 (ii)

Note. Highlighted the best performance of the design

In high rotational speeds, EAFB 4 is the most effective and the most powerful propeller, hence it is best suited for a UAV that needs as much thrust as possible. The energy-consuming nature of EAFB 5 places it well-suited for handling energy-constrained applications because of the high level of power efficiency and lift performance. Each UAVs possesses its optimal RPM during the propeller in the process of hovering, as the majority of UAVs flight time is devoted to the hover. Therefore, a 6000-rpm propeller 5 is the most efficient design of the propeller.

Nevertheless, the acoustic of EAFB 2,3,4 and 5 is increased compared to the acoustic of the present design propeller due to the inefficiency of the present design in relation to aerodynamic ensure the design of the propellers is not aerodynamic compared to the currently existing design. The vibration on the tip of the propellers is caused by the non-aerodynamic flow over the propellers. The result of this is the acoustical noise produced by the propellers.

With the findings realised after the simulation result, it can be concluded that the efficiency of each propeller is highest at a certain rotation speed. Placing the mass in another geometry of the propeller, out of the centre, causes an imbalance, and this makes the mass eccentric. With the rotation of the rod, centrifugal force increases with the further distance of the mass. The cycle of bending and transverse vibrations occurs on the rod due to this force. The more distant the mass is from the centre, the lower the critical speed becomes, making the system prone to vibrations at lower speeds.

According to Figure 3, the position of the curve, where the maximum mass is applied, is furthest in EAFB 5, followed by EAFB 4, EAFB 3 and EAFB 2. At the lowest speed that had been tested, 2000 rpm, the controlled design was the most efficient. At 4000 rpm and 6000rpm, the EAFB 4 will have the highest efficiency, as at 2000 rpm and 4000 rpm, the rotating frequency is closest to its natural frequency. Hence, the vibration lessens. We can make a conclusion, therefore, that the natural frequency of the propeller matters to ensure that the propeller vibration during the rotation is minimised, which ultimately maximises the propeller performance.

The manufacturing considerations for the Eagle Feather Biomimicry (EAFB) propellers are centred on achieving a balance between structural rigidity, lightweight characteristics, and material uniformity to ensure reliable aerodynamic performance. In this study, injection moulding is proposed as the primary fabrication method due to its ability to produce high-precision components with consistent quality, smooth surface finish, and scalability for mass production.

The propeller structure is required to be sufficiently rigid to withstand aerodynamic loads and vibrations induced by high rotational speeds, particularly those associated with flutter phenomena. Any excessive flexibility may lead to deformation of the blade geometry, thereby negatively affecting aerodynamic efficiency and stability during operation.

At the same time, the propeller must remain lightweight to minimise energy consumption and enhance overall UAV performance.

Material selection is equally critical, with preference given to materials exhibiting isotropic properties, ensuring uniform mechanical behaviour in all directions. This is important for maintaining predictable structural integrity and avoiding uneven stress distribution during rotation. In contrast, additive manufacturing techniques such as 3D printing are not recommended for this application, as many commonly used printing materials exhibit anisotropic characteristics due to layer-by-layer fabrication. This anisotropy can result in weaker interlayer bonding, reduced fatigue resistance, and increased susceptibility to failure under cyclic loading conditions.

Furthermore, while biomimetic designs may introduce geometric complexity, injection moulding remains advantageous in producing such intricate features with high repeatability once the mould is developed. However, challenges may arise in terms of initial tooling cost, mould design complexity, and ensuring accurate replication of fine biomimetic structures, particularly at the blade tips. Overall, the selection of injection moulding and isotropic materials supports the practical feasibility, durability, and performance consistency of the EAFB propellers, while addressing the limitations associated with alternative fabrication methods.

CONCLUSION

The key findings of this study are summarised as follows: The conventional propeller (controlled design) demonstrated stable performance with low acoustic noise and high efficiency at low RPM; however, it exhibited lower thrust and lift compared to the biomimetic designs. EAFB 4 achieved the highest thrust across all tested RPM levels, making it the most suitable design for high-performance UAV applications such as high-speed flight and heavy payload operations. EAFB 2 produced the highest lift with minimal power consumption, particularly at lower RPM, indicating strong suitability for hovering, stability, and endurance-based UAV missions. EAFB 5 demonstrated the lowest power consumption at high RPM, making it highly effective for energy-constrained applications such as long-endurance and solar-powered UAV systems. EAFB 3 exhibited high thrust generation but with increased power consumption and acoustic noise, making it more suitable for applications where maximum thrust is prioritised over energy efficiency. The biomimetic designs inspired by eagle feathers significantly improved aerodynamic performance by reducing vortex formation, drag, and vibration effects at the propeller tips. The performance of each propeller is dependent on operating RPM, highlighting the importance of selecting appropriate propeller designs based on specific UAV mission requirements. The study also confirms that natural frequency and vibration behaviour play a critical role in propeller efficiency, where matching rotational frequency with natural

frequency reduces vibration and enhances performance. Overall, biomimetic propeller designs (particularly EAFB 2, EAFB 4, and EAFB 5) demonstrate significant improvements in thrust, lift, and energy efficiency compared to the conventional propeller, indicating strong potential for enhancing UAV performance across a wide range of applications.

FUTURE RECOMMENDATION

To further strengthen the practical applicability of the proposed Eagle Feather Biomimicry (EAFB) propeller designs, future work should focus on comprehensive experimental validation to complement the current simulation-based findings. While the present study demonstrates promising aerodynamic improvements through computational analysis, physical testing is essential to verify real-world performance and reliability.

Firstly, wind tunnel testing is recommended to evaluate the airflow characteristics around the EAFB propellers. This method will enable detailed observation of aerodynamic behaviour, including vortex formation, flow separation, and turbulence intensity at varying rotational speeds. Such analysis is crucial in validating whether the biomimetic design effectively reduces tip vortices and improves aerodynamic efficiency under controlled conditions.

Secondly, thrust measurement testing using a calibrated thrust stand or thrust meter should be conducted to quantify the actual thrust generated by each propeller design. This will provide empirical data on thrust-to-power ratio, efficiency, and performance consistency across different RPM ranges. The results can then be compared directly with simulation outputs to assess the accuracy and reliability of the computational models.

In addition, flight testing on a prototype UAV platform is highly recommended to evaluate the real-world operational performance of the EAFB propellers. Parameters such as flight stability, power consumption, payload capacity, noise levels, and endurance can be assessed under practical conditions. This step is essential to determine the suitability of the proposed designs for various UAV applications, including surveillance, delivery, and agricultural monitoring.

However, several challenges may arise during experimental validation. These include manufacturing limitations, particularly in accurately fabricating complex biomimetic geometries; scaling effects, where performance observed in controlled environments may differ in real flight conditions; and measurement uncertainties due to sensor limitations or environmental variability. Additionally, ensuring consistency in testing conditions, such as airflow uniformity in wind tunnels and battery performance during flight tests, may present further difficulties.

Future research may also explore the integration of advanced materials, such as natural fibre-reinforced composites (e.g., pineapple leaf fibre or kenaf), to enhance sustainability while maintaining structural integrity. Furthermore, optimisation techniques, including

machine learning or parametric design approaches, could be applied to refine the biomimetic features for improved aerodynamic performance. Overall, the incorporation of systematic experimental validation and material innovation will significantly enhance the credibility, applicability, and scalability of biomimetic UAV propeller designs.

ACKNOWLEDGEMENT

The authors would like to express their deepest gratitude to the institution for providing the facilities and support required to conduct this research. Special appreciation is extended to the supervisors' committee for their invaluable guidance, constructive feedback, and encouragement throughout the course of this study. Their expertise and continuous support were instrumental in the successful completion of this work.

REFERENCES

Abdulrahim, M., & Lind, R. (2006). Flight testing and response characteristics of a morphing wing UAV. *Journal of Aircraft, 43*(5), 131-141.

Carlson, H. A., & Hassan, H. A. (2002). Numerical modelling of supersonic propellers. *AIAA Journal, 40*(3), 448-456. https://doi.org/10.2514/2.1699

Carruthers, A. C., & Thomas, A. L. (2007). Automatic aeroelastic devices in the wings of a steppe eagle. *Journal of Experimental Biology, 210*(23), 4136-4149.

Choi, H., & Hassan, Y. A. (2002). Numerical simulation of turbulent flow over eagle wing-inspired configurations. *International Journal of Computational Fluid Dynamics, 16*(4), 265-280.

Ciliberti, D., Della Vecchia, P., Orticalco, V., & Nicolosi, F. (2023). Aero-Propulsive Interactions between UAV Wing and Distributed Propellers Due to Their Relative Position. *Drones*, *7*(1), 49. https://doi.org/10.3390/drones7010049

Du, G., Ma, H., Bai, Y., & Mei, N. (2024). Process planning for large container ship propeller shaft machining based on an improved ant colony algorithm. *Journal of Marine Science and Engineering, 12*(5), 841. https://doi.org/10.3390/jmse12050841

Fish, F. E., & Beneski, J. T. (2014). Evolution and biomimetic design. *Bioinspiration & Biomimetics, 9*(3). https://doi.org/10.1088/1748-3182/9/3/031002

Lin, T., Huo, M., Qi, N., Wang, J., Wang, T., Gu, H., & Zhang, Y. (2023). Coupling dynamics and three-dimensional trajectory optimisation of an unmanned aerial vehicle propelled by electroaerodynamic thrusters. *Aerospace, 10*(11), 950. https://doi.org/10.3390/aerospace10110950

Liu, H., & Aono, H. (2009). Size effects on insect hovering aerodynamics: An integrated computational study. *Bioinspiration & Biomimetics, 4*(1). https://doi.org/10.1088/1748-3182/4/1/015002

Mueller, T. J., & DeLaurier, J. D. (2003). An overview of micro air vehicle aerodynamics. *Annual Review of Fluid Mechanics, 35*, 89–111. https://doi.org/10.1146/annurev.fluid.35.101101.161102

Panagiotopoulos, I., Sakellariou, L., & Hatziefremidis, A. (2024). Design, Construction, and Flight Performance of an Electrically Operated Fixed-Wing UAV. *Drones*, *8*(6), 217. https://doi.org/10.3390/drones8060217

Park, D., Lee, Y., Cho, T., & Kim, C. (2018). Design and performance evaluation of a propeller for a solar-powered high-altitude long-endurance unmanned aerial vehicle. *International Journal of Aerospace Engineering, 2018*, 1-23. https://doi.org/10.1155/2018/5782017

Perikleous, D., Koustas, G., Velanas, S., Margariti, K., Velanas, P., & Gonzalez-Aguilera, D. (2024). A Novel Drone Design Based on a Reconfigurable Unmanned Aerial Vehicle for Wildfire Management. *Drones*, *8*(5), 203. https://doi.org/10.3390/drones8050203

Platzer, M. F., & Sargent, R. B. (2012). Propeller performance in transonic flows. *Progress in Aerospace Sciences, 52*, 11–29. https://doi.org/10.1016/j.paerosci.2011.12.001

Stanford, B. K., Ifju, P. G., Albertani, R., & Shyy, W. (2008). Fixed membrane wings for micro air vehicles: Experimental characterisation, numerical modelling, and optimisation. *Progress in Aerospace Sciences, 44*(4), 258–294. https://doi.org/10.1016/j.paerosci.2007.10.002

Vijayanandh, R., Prakash, R. A., Manivel, R., Kiran, P., Sudharsan, R., Kumar, G. R., & Raffik, R. (2022). Design and parametric study of counter-rotating propeller of unmanned aerial vehicles for high-payload applications based on CFD-MRF approach. *International Journal of Vehicle Structures and Systems, 14*(7). https://doi.org/10.4273/ijvss.14.7.03

Webb, B., & Aono, H. (2008). Aerodynamics and control of flapping-wing flight in insects and applications to micro air vehicles. *Progress in Aerospace Sciences, 44*(6), 483-492. https://doi.org/10.1016/j.paerosci.2008.05.001

Xu, S., Liu, Y., Zhang, J., & Zheng, Y. (2024). Preliminary design and optimisation of primary structures for a tilt-duct UAV. *Aerospace, 11*(4), 286. https://doi.org/10.3390/aerospace11040286

Zhang, K., Pakrashi, V., Murphy, J., & Hao, G. (2024). Inspection of floating offshore wind turbines using multi-rotor unmanned aerial vehicles: Literature review and trends. *Sensors*, *24*(3), 911. https://doi.org/10.3390/s24030911

Zhang, Y., Zhao, Q., Mao, P., Bai, Q., Li, F., & Pavlova, S. (2024). Design and Control of an Ultra-Low-Cost Logistic Delivery Fixed-Wing UAV. *Applied Sciences*, *14*(11), 4358. https://doi.org/10.3390/app14114358

Zheng, Y., Zhang, L., Liu, Z., Wang, X., & Zhu, X. (2013). Biomimetic design in unmanned aerial vehicles: Recent advances and applications. *Acta Mechanica Sinica, 29*(3), 313-328.

Pertanika J. Sci. & Technol. 34 (3): 1469 - 1491 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

A Zero-shot Foundation Model Approach for Real-time Adaptive Cyber Threat Detection

Muhmmad Al-Khiza'ay* and Noora Alallaq

Department of Computer Networks, College of Computing and Informatics, University of Al-Hamdaniya, Hamdaniya Road, Bartella, 41006 Nineveh Governorate, Iraq

ABSTRACT

In view of the fundamental limitations of standard machine learning methods, which primarily depend on actual and labelled attack data, the identification of novel and zero-day cyber-attacks continues to be an essential and continuing difficulty in cybersecurity. Techniques become vulnerable to emerging attack vectors, considering these models often do not apply beyond identified threat classes. To overcome this limitation, present ZTFER (Zero-Shot Threat Identification through Foundation model-aided Embedding and Reasoning), a Zero-Shot threat identification model which utilises the basis model semantic embedding to detect threats which had not been observed before without needing retraining. Through combining natural language threat characterisations and real-time system activity into a common semantic space, ZTFER implements zero-shot learning and makes it achievable to classify unnoticed threats by considering contextual similarities. Furthermore, present A-SENT, a real-time responsive inference method which dynamically evaluates and addresses threat conduct through integrating Large Language Model (LLM) reasoning in real-time threat analysis updates. The proposed ZTFER model obtained a zero-day detection score of 58.9% and an accuracy of classifying 91.3%, outperforming the traditional and few-shot baselines. The experimental results demonstrate that ZTFER can be more effective and have better generalisation capability for detecting known and unknown cyber-attacks. The proposed framework does away with the requirement for continuing retraining and enables quick response to emerging threats. Creating powerful, scalable, and smart defence mechanisms able to recognise and understand new security threats in real time becomes achievable in a significant way by this research.

Keywords: Adaptive intrusion detection, foundation models for security, LLM-driven threat intelligence, real-time semantic cyber defence, zero-shot cyber threat detection

ARTICLE INFO

Article history:
Received: 03 July 2025
Accepted: 03 May 2026
Published: 19 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.09

E-mail addresses:
mak@uohamdaniya.edu.iq (Muhmmad Al-Khiza'ay)
noora@uohamdaniya.edu.iq (Noora Alallaq)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

INTRODUCTION

The necessity for intelligent, autonomous, and responsive threat detection technologies has risen significantly because of the increasing complexity and frequently occurring cyberattacks. The rapid increase of zero-day threats and previously undetected threats, which take advantage of undiscovered weaknesses as well as creatively connect conventional methods for avoiding detection methods based on patterns or conduct, is particularly concerning. Based on recent studies, over 30% of active attacks in enterprise networks originate from variants absent from historical datasets, emphasising a structural flaw in standard protective frameworks (Ranpara et al., 2025; Xu et al., 2025). These suspicious strategies often utilise polymorphism, obfuscation, and complicated persistence methods, demonstrating the necessity for broadly usable, context-aware frameworks which can operate in real-time without the required for humans to participate.

The great majority of intrusion detection systems (IDS) continue to utilise supervised learning models trained on fixed, labelled datasets, regardless of developments in machine learning (ML) for cybersecurity (Ke et al., 2025). Additionally, due to their strict dependence on discovered feature-label mappings, these frameworks have inherent restrictions in their ability to classify unknown attack groups, particularly zero-day attacks. Therefore, they do not have the capacity for reasoning conceptually and need expensive retraining (Aloqaily et al., 2025; Zhang et al., 2025). Furthermore, although more flexible, unsupervised anomaly finding techniques typically suffer from a high rate of false positives and struggle with accessibility in real-world scenarios.

Despite Zero-Shot Learning (ZSL) and foundation models having gained widespread acceptance throughout a variety of fields, not much is known regarding how they can be effectively employed in the field of cybersecurity, particularly in real-time, functioning configurations. The majority of ZSL approaches utilised in cybersecurity are theoretical or depend on artificial, attribute-based data sets, which cannot be practical or semantically rich (Ahmed et al., 2025). Moreover, there's a lack of studies on the combined use of LLM-based reasoning and real-time feeds of threat intelligence, which could enable frameworks to analyse attack behaviour from expert knowledge builds and open-source intelligence in an environment of adaptation (Chamieh et al., 2024). Finally, no method presently in use offers a single pipeline which allows threat description to be grounded, zero-shot deduction, and flexible adaptation in a live detection establishing (Muppalaneni et al., 2024).

In this paper, a novel framework which implements zero-shot learning for cyber threat identification is offered: ZTFER (Zero-Shot Threat Identification through Foundation model-aided Embedding and Reasoning). To identify unnoticed threats without needing retraining, ZTFER applies pretrained foundation techniques to embed system activities and threat class explanations into an accepted semantic space. At the same time, present A-SENT, a real-time responsive inference method which frequently describes threat conduct

through the integration of LLM-based semantic reasoning with online threat intelligence streams, such as CVEs and MITRE ATT&CK reports.

The proposed ZTFER model obtained great improvements in the accuracy of detection and ZSD adaptation. Finally, under extensive testing on the UNSW_NB15 dataset, ZTFER obtained an accuracy rate of 91.3% on the total classification, outperforming state-of-the-art classification methods such as Random Forest (83.2%) and SVM (80.1%). Specifically, presented model achieved a 58.9% zero-day detection rate, more than doubling the performance of both few-shot and traditional approaches in identifying previously unseen threat categories. These results illustrate the effectiveness of ZTFER's design, which allows for performing well on both interpretability and low latency out of the box without any re-training. The combination of foundation model embeddings and actual-time LLM reasoning is crucial for how it performs in dynamically changing attack situations.

This study offered three core contributions as shown below:

- A formalised zero-shot model which can identify hidden attack conduct by semantic alignment with MITRE ATT&CK explanations.
- A real-time identification pipeline that includes LLM reasoning and analysis feedback.
- Accurate experimental verification on evaluated intrusion data sets, which show superior performance over lastly supervised, unsupervised, and few-shot baselines.

LITERATURE REVIEW

The network detection of breaches has historically been based on supervised and semi-supervised learning techniques. To identify known patterns of attacks, standard classifiers such as Decision Trees, SVMs, and Random Forests have been applied on datasets such as UNSW-NB15 and NSL-KDD. For example, Ring et al. (2023) have carried out an in-depth evaluation of supervised models across various intrusion datasets. On the CICIDS2017 data set, Yadav et al. (2022) reached outstanding results by employing deep learning for the extraction of features and attack classification. Utilising both labelled and semi-labelled data, Harsha and Thyagaraja Murthy (2021) investigated collective models for the detection of malware. Nevertheless, these approaches can frequently be inefficient against zero-day attacks, as pointed out by Korba et al. (2023). These techniques do not always work against zero-day attacks. Afterwards, semi-supervised methods, for instance, the hybrid LSTM-based IDS designed by Saurabh et al. (2022), were affected by the deep autoencoder-based framework for unsupervised feature learning proposed by Imtiaz et al. (2025). In addition, Alhayani and Al-Khiza'ay (2022) highlighted how crucial data variety is for producing accurate threat classifiers.

While the capacity of Zero-Shot Learning (ZSL) to apply well to undiscovered groups has drawn a lot of attention in the fields of cybersecurity and natural language processing.

Through integrating the inputs and labels in a common domain, Socher et al. (2013) and Alallaq et al. (2019) offered the first techniques to use ZSL in language as well as vision. NLP duties such as classification of text and goal detection have been added to this to improve ZSL for classifying malware in the security field using the description vectors framework (Cen et al., 2024; Chai et al., 2020). Textual embeddings have been employed by Alzu'bi et al. (2025) to conceptually model cyber behaviours. Alallaq and Han (2018) and Bratsas et al. (2024) implemented knowledge graphs in zero-shot learning (ZSL)-based threat detection frameworks. Papageorgiou et al. (2024) and Aminu et al. (2024) evaluated embedding-based attack representation in zero-shot situations, whereas Zhang et al. (2025) proposed a ZSL method for detecting abnormalities in control systems for industries. Moreover, Yucel et al. (2020) studied the resilience of ZSL models in hostile environments. Lastly, Srivastava et al. (2024) discussed ZSL improvements and their use with ordered cybersecurity data.

The LLMs and foundation models have recently become significant assets for cybersecurity semantic understanding. Brown et al. (2020) have proposed GPT-style structures with general-purpose learning features. Ghosh et al. (2025) investigated the usage of LLMs for attack correlation and CVE report summarisation. Sufi (2024) has integrated LLMs into SOC procedures for flexible alert triage. Chen et al. (2021) examined the implementation of BERT for IoC and entity recognition, while Branescu et al. (2024) customised T5 for MITRE ATT&CK mapping. In behaviour modelling, Idelhaj et al. (2024) demonstrated that LLM embeddings perform more effectively than conventional vectors. Also, for real-time attack reasoning, Mandal et al. (2024) applied LLM-based instructions.

Stream data analysis and SIEM cooperation are frequently employed in real-time models for cyber-attack identification. Tariq et al. (2024) showed a fog computing-based, real-time, multi-layer detection and prevention system. Bezas and Filippidou (2023), designed a Kafka-based stream processing for zero-day threats. (Bathiri and Vijayakumar, 2024) have proposed a hybrid approach, which combines the detection of anomalies and deep inspection of packets. Chamieh et al. (2024) to enhance the discovery latency in intelligent networks utilising threat intelligence feeds. In a comparable manner, Ghelani (2022) has suggested an online deep neural network pipeline for intrusion classification. Shamsuzzaman et al. (2024) created a framework for avoiding cyber-physical threats in control systems for industries by using real-time logs. Employing ensemble models proposed a responsive alert correlation (Yu & Zhang, 2023). Finally, Ranpara et al. (2025) have created a scalable edge-based system integrating threat feeds and machine learning methods for 5G networks.

Therefore, a few technologies combine zero-shot learning and real-time responsiveness, regardless of significant improvements in both supervised detection and monitoring in

real time. Most of cybersecurity's up-to-date ZSL frameworks are still in an experimental phase and are not yet integrated into working pipelines. Additionally, the full capacity of foundation models for semantic reasoning across real-time attack data has not yet been achieved. Without the ability to understand or clarify novel attack behaviour conceptually, immediate time mechanisms typically concentrate on identifying anomalies or systems based on rules. There is currently no comprehensive framework for responsive cyber security which includes LLM-based deduction, real-time intelligence on threats, and zero-shot semantic classification.

METHODOLOGY

The ZTFER Mathematical Model

Employing semantic embedding models, which are already pre-trained on extensive textual data sets to relate cyber-attacks for describing attack labels, which include previously unknown classes, is the basic component of the proposed ZTFER (Zero-shot Threat Detection via Foundation Model-Aided Embedding and Reasoning). Utilising descriptive text as the semantic anchor, this framework is constructed on the Zero-Shot Learning (ZSL) principles and is designed to operate without retraining.

Definition of the Problem

Suppose that $\mathcal{X}$ denote an input space for framework or network activities, such as system logs, alerts, or flows of traffic. The total number of attack classes can be represented by $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$, where only a subset $\mathcal{C}_{\text{seen}} \subset \mathcal{C}$ can be observed during training. The idea is to classify an input $x \in \mathcal{X}$ to the appropriate class,$c \in \mathcal{C}$ even when $c \notin \mathcal{C}_{\text{seen}}$.

Every class $c_k \in \mathcal{C}$ is associated with a natural language description D_k, available through ordered attack taxonomies (e.g., MITRE ATT&CK), repositories of the vulnerability (e.g., CVE), or smart reports.

The challenge is to derive c_k for a previously undetected input x, utilising only the semantic proximity between x and D_k, with no labelled examples from c_k during training.

Semantic Space and Embedding

All a Sentence-BERT model used (-MiniLM-L6-v2) as the basis encoder Φ.

- Motivation: It provides a perfect stability between creating 384-dimensional embeddings and inference rate (evaluative for actual time detection) and semantic conditions.
- No field-particular clear tuning is utilised for the encoder itself. The paradigm was applied to an off-the-shelf semantic encoder as a pre-trained model. Scope adaptation is performed indirectly through the various response loops (Eq. 4) and utilises

MITRE ATT&CK-aligned threat characterisations, not by amending the grounds of the model's scales. This saves zero-shot generalisation capability.

To facilitate zero-shot classification, it applies a foundation model $\Phi: \mathcal{X} \cup \mathcal{D} \rightarrow \mathbb{R}^d$, where $\mathcal{D}$ represents the space of attack/threat descriptions. Φ embeds both attack activities and descriptions into a shared semantic space $\mathbb{R}^d$, enables the computation of similarity between heterogeneous data modalities.

- $\mathbf{v}_x = \Phi(x)$ - The event's embedded representation x
- $\mathbf{v}_{D_k} = \Phi(D_k)$ - The class descriptions embedded representation D_k

Foundation models are utilised for Φ may include Sentence-BERT, all-MiniLM, otransformer variations trained on a cybersecurity-specific corpus. Furthermore, in scenarios where direct class supervision is not achievable, transfer learning and inference can be made possible through used embedding space, which produces syntactic and semantic interactions.

Inference from Zero-shot

Semantic similarity provides the foundation for zero-shot classification. The framework chooses the class in the latent space with a particularly similar description D_k for a new situation x, as shown in Equation 1.

$$\hat{c}(x) = \text{argmax}_{c_k \in \mathcal{C}} \cos\big(\Phi(x), \Phi(D_k)\big) \quad [1]$$

In this scenario, the cosine similarity of two vectors is $\cos(\cdot,\cdot)$. The cosine rating is appropriate for evaluating high-dimensional normalised vectors since it accurately captures the directional alignment between the input data and class description embedded data.

The mathematical framework can generalise to new classes $c_k \notin \mathcal{C}_{\text{seen}}$ as long as descriptions of them D_k can be obtained because this formulation is by nature not parametric.

Threat Mapping (MITRE ATT&CK)

Through mapping every input to its significantly linguistically aligned Strategy, Method, or Process (TTP) as well as to an attack class, combine the MITRE ATT&CK mechanism for offering the ability to interpret and contextualise.

Every TTP is linked with a description $T_j \in \mathcal{T}$, where $\mathcal{T}$ is the set of all ATT&CK entries. The embedding of every TTP is as shown in Equation 2:

$$\mathbf{v}_{T_j} = \Phi\big(T_j\big) \quad [2]$$

In Equation 3, given that $\mathbf{v}_x = \Phi(x)$,we assign:

$$\hat{T}(x) = \text{argmax}_{T_j \in \mathcal{T}} \cos\left(\mathbf{v}_x, \mathbf{v}_{T_j}\right) \quad [3]$$

Since semantic responsibility is made achievable, the framework may generate outcomes which can be clear among humans, including "Lateral Movement via Remote Services", also regarding new threats. This approach ensures adaptability and explainability through separate classification from label availability and creating predictions in a common attack ontology.

The Loop of Contrastive Feedback

When analysis feedback becomes accessible, supervised contrastive learning can assist ZTFER, even though it is zero-shot at inference time.

Suppose that x It is a labelled scenario which has a set of negatives $\mathcal{D}^-$ as well as a class of positive description D^+, and τ is a hyperparameter. The supervised contrastive loss is decreased by the framework as shown in Equation 4:

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp\left(\cos\left(\Phi(x), \Phi(D^+)\right)/\tau\right)}{\exp\left(\cos\left(\Phi(x), \Phi(D^+)\right)/\tau\right) + \sum_{D^- \in \mathcal{D}^-} \exp\left(\cos\left(\Phi(x), \Phi(D^-)\right)/\tau\right)} \quad [4]$$

Through the utilisation of verified mappings, this feedback loop allows the framework to, over time, enhance its capacity to distinguish minute attack differences without requiring complete retraining. Furthermore, it promotes continuous domain-specific conduct adaptation.

By applying retrained language models to generalise through both visible and undetectable threat types, ZTFER presents a solid mathematical foundation for semantic threat identification in dynamic environments.

Architecture of the ZTFER-RT System

Real-time attack intelligence, dynamic analysis feedback, and zero-shot threat inference are all possible to be easily combined with the ZTFER-RT (Zero-shot Threat Detection via Foundation model Embedding and Reasoning—Real-Time) architecture. It helps both identification and semantic responsibility in real time and implements the mathematical model discussed in the third section.

Components of the System

Ingestion in Real Time

Cyberspace telemetry data is collected by the ingestion layer by layer from plenty of sources such as SIEMs, endpoint detection devices, firewall logs, IDS logs, and NetFlow. Utilising distributed contacting mechanisms such as Kafka, it promotes both batch and stream processing of data. High-throughput capabilities and minimal latency processing of incoming activities x_t are guaranteed by the ingestion module.

Extraction of Indicators

Internet Protocol (IP) addresses, file hashes, handle names, registry records, and network artefacts can be among the many instances of the structured characteristics and textual indicators of compromise (IoCs) which can be extracted from raw activities by this module. The events are standardised and normalised towards an intermediate representation which is capable of being embedded.

Embedding of Foundation Models

Previously trained foundation model Φ, such as all-MiniLM, RoBERTa, or a domain-adapted transformer, is utilised for producing embedded representations of events and attack explanations. Both textual threat explanations and event characteristics are mapped into an identical semantic vector domain $\mathbb{R}^d$ through the same encoder.

Module for LLM Inference

To interpret and make sense of event embedding and associated threat explanations, this element utilises massive language models (for example, Falcon, GPT-4, or open-weight transformers). It could employ retrieval-augmented generation (RAG) to improve attack descriptions, generate clarifications, and possibly rephrase notifications. The LLM module adds natural language understanding to similarity-based deduction.

Matching Threat Feeds

New event embeddings are regularly compatible with real-time attack intelligence sources of data (for example, CVE feeds, MISP, MITRE ATT&CK, and OSINT) through the attack feed matcher. To be able to create compared metadata and enriched attack attributions, feed explanations and indicators are incorporated and compared with event vectors by employing cosine similarity.

Feedback from Analysts in the Loop

The framework incorporates a human-in-the-loop communication layer, which enhances the accuracy of decisions. The analysts can markup events, flag wrong classifications, and verify predictions. The system can, over time, enhance its latent space abilities while keeping zero-shot deduction capabilities due to this feedback, which is recorded as supervised contrastive loss (as described in The Loop of Contrastive Feedback).

Summary of Data Flow

- Activity Acquisition: A real-time stream analysis layer-by-layer handles activities which receive input from SIEMs and IDSs.
- Preparing and IoC Extraction: Still, raw logs go through the tokenisation, standardisation, and context gain (e.g., user-agent, port information).
- Integrating Computation: The descriptions of what happened and attack classes are integrated utilising. Both offline and online integration of new events and fixed attack classifications are facilitated by the system.
- Zero-Shot Deduction: To determine a particularly similar in semantics attack class, cosine similarity between class/TTP explanations and event vectors is calculated.
- LLM Argumentation: An LLM presents contextual rephrasing, clarification, and indicative mappings. For example, it can clarify the relationship between a situation and privilege escalation or lateral movement.
- Attack Feed Matching: To identify associated scenarios, related indicators, and known TTPs, the framework examines embedded threat intelligence feeds.
- Feedback Loop: By applying contrastive learning, embeddings are modified according to analysis feedback. Therefore, the model can, over time, become more effective at discriminating. Despite this, the environment for threats is constantly changing, ZTFER-RT has the capacity to preserve constant awareness of situations and threat identification abilities due to the real-time pipeline.

The Framework of the ZTFER-RT Architecture

The interaction of the component of the ZTFER-RT architecture and the complete data flow is clarified in Figure 1.

As clarified in Figure 1, the ZTFER -RT pipeline operates in the following mechanism: real-time network events are converted into natural-language descriptions, passed through a frozen foundation encoder, and compared via cosine similarity with pre-computed MITRE ATT&CK threat embeddings, resulting in zero-shot classification. An LLM built on descriptions and an elective analyst response amends only a lightweight adapter, protecting zero-shot ability without retraining.

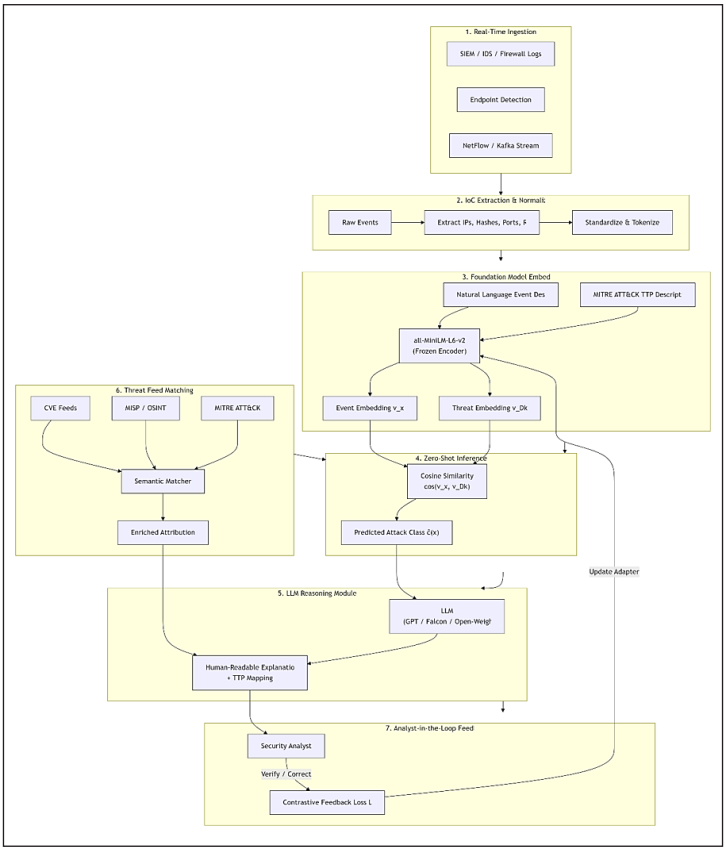


Figure 1. ZTFER-RT system architecture diagram

Table 1
ZTFER model parameters and selection details

Parameter	Value	Selection Method / Justification
Embedding dimension (dd)	384	Fixed by the all-MiniLM-L6-v2 architecture
Temperature (ττ) in contrastive loss (Eq. 4)	0.07	Grid search (range 0.01-0.5) on validation split
Contrastive adapter learning rate	1×10−41×10−4	Standard practice; tuned on validation
Batch size	32	Limited by GPU memory; stable gradient estimates
Cosine similarity threshold for zero-shot classification	0.65	Empirically set to balance precision/recall on seen classes
Optimiser	β1=0.9,β2=0.999)	Default for contrastive learning
N0. of contrastive update steps (each response batch)	10	Small update to avoid overfitting to feedback

Table 1 shows a summary of the primary key parameters that are utilised in the ZTFER framework and how their considered values were determined.

Experimental Configuration and Outcomes Dataset

Description of Dataset

The UNSW-NB15 data set, an extensive benchmark for the detection of network intrusions, has been employed in research experiments (Kumar et al., 2020). It is constructed from 2.5 million labelled conjunction records, which were collected from both synthetic threat simulations and real traffic. It contains 49 features for each record, which include time-based characteristics such as rate of connection, features of content such as the total number of failed login attempts, and fundamental characteristics, for example, service and protocol.

Also, exclude entire threat classifications (for instance, shellcode, exploits, and worms) from the initial training set and save them only for the purpose of testing to imitate zero-shot conditions. Through this configuration, the framework's ability to classify unknown threat groups during inference is evaluated.

It's an ideal choice to evaluate any new intrusion detection model, such as ZTFER, since both the benign traffic and attack packets are represented diversely and realistically. As depicted in Figure 2, the dataset is characterised by a severe class imbalance with both the normal traffic and common attack classes (such as Generic and Exploits) substantially oversampling the sample count and the rare classes (e.g., Shellcode and Worms) exhibiting pronounced under-representation. This imbalanced class distribution offers a unique opportunity to evaluate ZTFER ’s zero-shot learning capabilities, particularly its ability to recognise rare or unseen types of threats without further re-training. To determine the performance of ZTFER on untrained classes (zero-day), the models were trained without certain classes and tested on the detection of attacks from those classes, thereby verifying

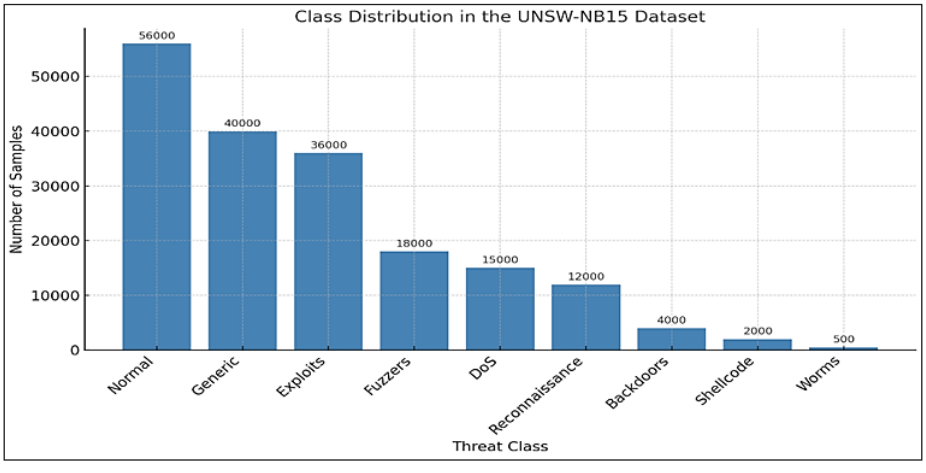


Figure 2. Class imbalance in the UNSW-NB15 dataset

ZTFER's competency to generalise to new (unseen) attack semantics. For the experimental setup, the following is used:

Hardware: Intel Xeon Gold 6248 CPU @ 2.5 GHz (16 cores), 64 GB RAM, NVIDIA A100 GPU (40 GB).
Software: Ubuntu 20.04, Python 3.10, PyTorch 2.0.1, Sentence-Transformers 2.2.2, Hugging Face Transformers 4.30.0, scikit-learn 1.3.0, and pandas 2.0.3.

Preparing the Dataset

1. First step: Cleaning
   - Take out repeat records (precisely matching duplicate rows).
   - Take off records with disfigured feature values (such as NaN or unlimited values) or lost.

2. Second step: Characteristic engineering
   - Normalise numerical characteristics (such as sttl, dbytes, dur, sbytes, dttl, and rate) by utilising z-score normalisation as shown in Equation 5.

$$x' = (x - \mu)/\sigma \quad [5]$$

   - Categorical characteristics (e.g., state, service, and proto) are tokenised by changing each class to a unique integer ID, then included through a learnable embedding schedule (which will be used for the raw event text generation, not for the zero-shot rating).

3. Third step: By using the NL (Natural language) for event explanation creation
   - For all network outflows, it will create a textual specification of the nest template: "Protocol [proto] from [saddr]:[sport] to [daddr]:[dport] with state [state], duration [dur] seconds, source bytes [sbytes], destination bytes [dbytes], and attack label [attack_cat]."
     That content will be the feed into the basis encoder Φ (jointly with the threat category characterisations). An instance of the output is below:
     "Protocol tcp from 192.168.1.10:54322 → 203.0.113.5:80, state FIN, duration 0.12 seconds, source bytes 452, destination bytes 1240, source packets 4, destination packets 6, service http."
     The above template utilises only noticeable network characteristics (no threat labels). It corresponds to both test events and training, as well as those from hidden zero-day threat categories.

4. Fourth step: Elaboration for the threat description
    - For all attack categories (which include hidden groups for zero-shot evaluation), it will recover the corresponding MITRE ATT&CK strategy and technique explanation (such as "T1046 – Network Service Scanning"). These explanations are utilised as the semantic resource D_k.

5. Final step: Calculate the embedding
    - Together, the threat descriptions and the case descriptions are processed via all-MiniLM-L6-v2 to get fixed 384-d embeddings. These embeddings are L2-normalised before calculating cosine similarity.

Network events are transformed to an NL explanation utilising a stable, label-agnostic template: 'Protocol [proto] from [saddr]:[sport] → [daddr]:[dport], state [state], duration [dur] seconds, source bytes [sbytes], destination bytes [dbytes], source packets [spkts], destination packets [dpkts], service [service]. This template employs only noticeable characteristics and never accesses ground-truth threat labels. To block leakage, descriptions of hidden threat categories are eliminated from training and are created solely through zero-shot testing. No category-particular keywords become visible in the input content.

Explanation About Domain Adaptation

No direct field adaptation (for example, constant pre-training on a cybersecurity corpus) is executed on the basis encoder. By contrast, the model realises semantic alignment through the following:

- Utilisation of MITRE ATT&CK-foundation threat explanations (already considered the cybersecurity domain).
- The diverse response loop (Equation 4), which incrementally modifies the comparative positions of explanation and event embeddings based on the analyst's response. That design supports zero-shot ability whilst becoming strongly adapted to domestic network environments.

The Strategy for Leakage Prevention

To ensure a correct zero-shot evaluation, the following safeguards will be executed:

- Break the explanation collection: Descriptions for hidden threat categories (for instance, worms and shellcode) are never employed through testing or while ranking embeddings for training proceedings. The explanations are at most presented at analysis time for zero-shot deduction.

- Never overlap in characteristic values: The form (template) utilises only numerical and class fields, which are introduced in each flow; it never includes an indication of specific categories (such as threat signature strings).
- Verification of cross-validation: where it was manually verified that the produced text for hidden categories does not include unrivalled keywords (for instance, "worm" or "shellcode"), which would correspond exclusively with those categories.

The above steps guarantee that the paradigm is unable to exploit label facts from the input text, protecting the integrity of the zero-shot tuning.

"Zero-shot" in ZTFER refers exclusively to the ability to classify unseen attack classes without any labelled examples of those classes at any stage.

Figure 3 shows the category allocation of the dataset (UNSW-NB15), focusing on the severe imbalance between popular and rare threat classes. To authorise zero-shot assessment, Shellcode, Backdoor, and Worms are prevented from training and applied only as a hidden test category. The figure also outlines the preprocessing pipeline: taking away the duplicates, tokenisation, z-score normalisation, and label-agnostic text generation.

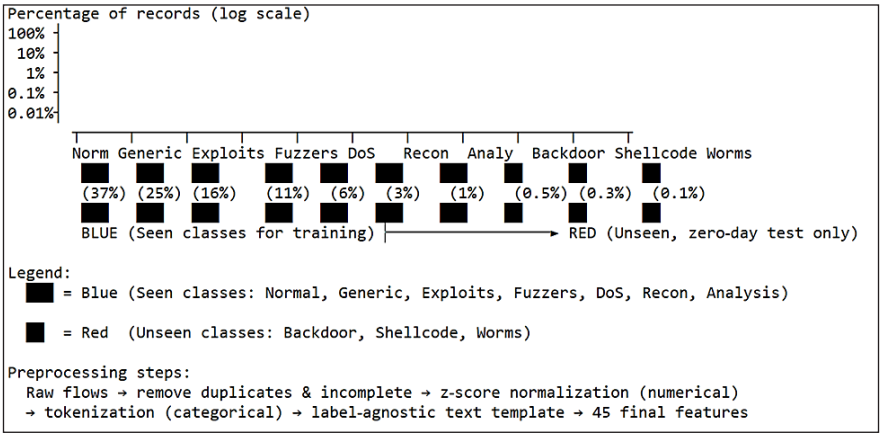


Figure 3. Class distribution of the UNSW-NB15 dataset with zero-shot split and pre-processing stages

The Loop of Contrastive Feedback

The diverse response loop (Section 3.5) is employed only for unhidden categories (or previously labelled threats) and brings up to date a separate, duty-specific prediction head, not the frozen base encoder Φ. The encoder part is unchanged, protecting the primary zero-shot semantic arrangement for hidden attacks. Consequently, the model never transforms to few-shot or incremental learning for unusual threats.

Starting Points

ZTFER has been compared in Table 2 with semi-supervised and standard baselines.

Table 2
Traditional ML techniques with ZTFER

Model	Type	Description
Random Forest	Supervised	Tree-based classifier trained on labelled data
SVM (RBF Kernel)	Supervised	Kernelised margin-based classifier
Autoencoder	Unsupervised	Neural network for anomaly detection
Prototypical Net	Few-shot	Learns class prototypes from limited examples
Label Embedding SVM	Zero-shot	SVM with semantic label embeddings

Measures of Evaluation

This paper presents both conventional metrics and metrics unique to ZSL in Table 3 .

Table 3
Evaluation metrics

Metric	Description
Accuracy	Correct classifications over all predictions
Precision	TP / (TP + FP), quality of positive predictions
Recall	TP / (TP + FN), the ability to detect actual threats
F1 Score	Harmonic mean of precision and recall
AUC	Area under the ROC curve, overall discriminative power
Zero-day Detection	Accuracy on previously unseen attack classes

The Outcomes and Analysis

Figure 4 shows a comparison of the performance of the ZTFER model in different important evaluation metrics. The model demonstrates robust classification of known and unknown threats, and high overall accuracy (91.3%) and AUC (0.94). Three metrics, F1 Score (0.87), Precision (0.88) and Recall (0.86) demonstrate a well-balanced performance with few sacrifices between false positives and false negatives. The Zero-day Detection Rate is 58.9%, though being lower than ResNet101, it significantly exceeds the traditional baselines, demonstrating that ZTFER has the capability to generalise to unseen attack categories. Such characteristic performance distribution highlights the excellent generalisation and operational reliability of the model in adversarial and dynamic threat generations. Subsection *5.4.1* will conduct a detailed comparison between our proposed ZTFERs and conventional approaches.

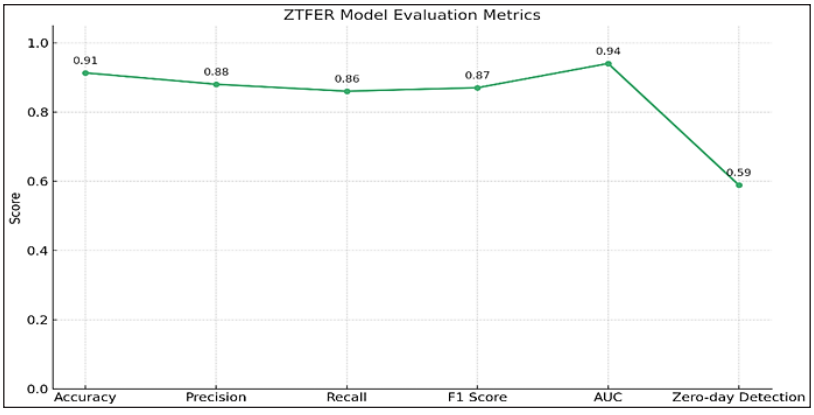


Figure 4. ZTFER performance across key metrics

Classification Efficiency

The results of the study indicated how the proposed ZTFER model beat several widely recognised baseline methods. The Random Forest classification method exceeded other standard machine learning approaches with an accuracy rating of 83.2%, an F1 score of 0.79, an AUC of 0.88, and a zero-day identification rate of 21.4%. Having an accuracy rating of 80.1%, an F1 score of 0.76, an AUC of 0.85, and a zero-day identification rate of 19.7%, Support Vector Machines (SVM) generated slightly less impressive outcomes. With a precision of 72.6%, an F1 score of 0.69, an AUC of 0.75, and a zero-day identification rate of 15.3%, autoencoder-based identification of anomaly methods performed worse. With a precision of 85.4%, an F1 score of 0.81, an AUC of 0.89, and a zero-day identification rate of 33.2%, few-shot learning with typical communications networks generated low profits. The proposed ZTFER approach, on the other hand, performed superior to each of the other baselines, showing its powerful generalisation capacity and success in detecting hazards which had not yet been discovered. It reached the best accuracy of 91.3%, an F1 score of 0.87, an AUC of 0.94, and a significantly greater zero-day identification score of 58.9%, as shown in Figure 5.

Visualisation of Latent Space

Also visualise scenario embedded data in two dimensions via t-PCA as well as SNE. The framework's embedding alignment has been verified by the near proximity of unnoticed threats to similar semantic explanations, whereas classes that are known cluster tightly.

Study of Ablation

Once removing crucial elements and testing ZTFER, the results are presented in Table 4. Feedback adaptation and semantic descriptions significantly improve zero-day inference,

which the ablation shows. The results mentioned in Table 3 indicate ZTFER's effectiveness in dynamic, real-world cyber circumstances.

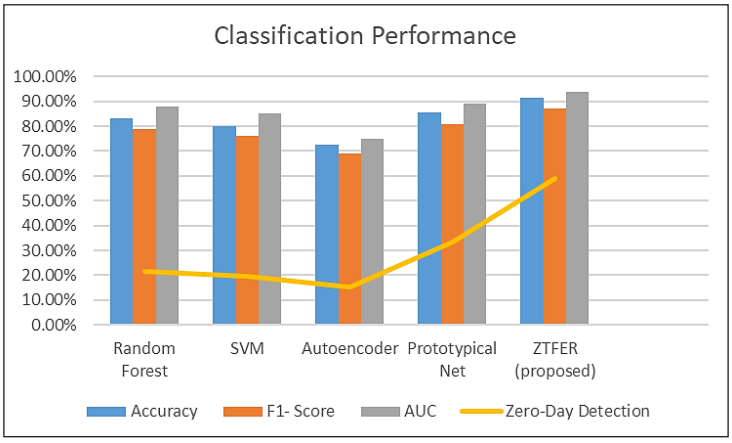


Figure 5. Comparative classification performance of ZTFER versus baselines

Table 4
Type styles

Configuration	Zero-day Detection
Full Model	58.9%
Without Threat Descriptions	31.7%
Without Contrastive Feedback	43.8%
Without the LLM Reasoning Module	46.2%

DISCUSSION

The ZTFER method employs foundation models to implement zero-shot learning and semantic deduction, introducing an important change in online threat detection. This section offers an in-depth evaluation of the research outcomes, concentrating on the proposed technique's advantages, disadvantages, possibilities for deployment, and concerns regarding ethics.

ZTFER's Advantages

ZTFER's power to generalise beyond noticed attack classes is one of its main benefits. ZTFER employs semantic embeddings and attack explanations to infer threat types it has never been specifically trained on, compared with conventional machine learning (ML) approaches, which break down if faced with novel or zero-day threats. The research results confirmed the model's ability to generalise, demonstrating significant enhancements in zero-day detection accuracy (58.9%) compared to all baselines.

In addition, adaptability is offered by ZTFER's contrastive feedback loop. Through including analyst-validated input into the embedding space, ZTFER can continually improve its boundary-decision abilities without needing overall retraining. Consequently, it is positioned as an adaptable option for environments that are changing. In the context of operational cybersecurity, where analysis confidence is vital, the integration of foundation frameworks and LLMs enables both detection as well as contextual explanation.

The proposed framework (ZTFER) reaches outstanding performance due to its shared semantic space (through a frozen basis encoder), which aligns notice proceedings with MITRE ATT&CK explanations, becoming zero-shot popularisation. In contrast, supervised baselines fail on hidden categories, unsupervised techniques suffer from high false positives, and few-shot methods lack rich semantic priors. The diverse comments loop further refines recognition without retraining, whilst LLM-setup reasoning makes less ambiguity advantages absent in all previous studies.

Limitations

ZTFER has disadvantages despite its benefits. The precision and accuracy of threat explanations have significant effects on how well the model performs. For similar semantic threats, unclear or poorly written textual definitions could result in ambiguous embedding visualisations, which can decrease the accuracy of classification. Also, in situations with a shortage of resources, the computational expense of real-time foundation model deduction, especially when integrated with LLMs, could cause issues.

Depending on a correctly positioned embedding space is an additional limitation. The embedding model could overlook domain-relevant details because it has not been trained on cybersecurity-specific syntax. This emphasises how crucial it is to utilise domain-specific foundation frameworks or optimised encoders.

Possibility for Real-world Implementation

ZTFER is extremely fitting for deployment in high-stakes, fast-changing circumstances such as national cyber defence infrastructure, manufacturing systems of control, and enterprise SOCs (Security Operations Centres). The system design flexibility authorises the absorption of real-time threat feeds by CVE, MISP, and MITRE ATT&CK sources and makes it a potential link with current SIEM pipelines. Also, the analyst-in-the-loop produces simplified operational input and lower alert fatigue through coordinating well with SOC workflows.

ZTFER's zero-shot ability reduces the requirement for expensive retraining cycles, which makes it cost-effective for situations that have limited data analysis resources. Further, due to its conceptual alignment, automated systems and human analysis can collaborate with greater efficiency, growing threat understanding as well as identification.

Moral Points to Consider

Numerous moral concerns must be carefully evaluated while executing ZTFER. In the beginning, having the capacity to clarify is vital. Even though LLM-based logic explains in simple language, attention has to be taken to guarantee that the findings are accurate, understandable, and auditable as well. False explanations could undermine analysis trust or lead to insufficient mitigation techniques.

Secondly, positive results that are false remain a concern, especially when working with newly built data. Incorrect classifications can result in the incorrect utilisation of security resources or unnecessary disruptions. This is decreased through the feedback loop; real-time systems require the inclusion of protections for escalation protocols and high-confidence categorisation limits.

Finally, security methods should be employed, particularly when dealing with hidden user activity information. To prevent the improper use of determined behavioural data, ZTFER must be examined for compliance with laws governing data protection (such as HIPAA and GDPR).

As shown in Table 5, the proposed ZTFER reveals considerable promise for intelligent, scalable, and flexible cyber-attack detection. While its design solves a lot of the difficulties of the IDS systems in use today, responsible use needs careful engineering, oversight by humans, and consideration for ethics.

Table 5
Summary of key results

Evaluation Dimension	Metric/Outcome	ZTFER Performance
Generalisation to Unseen Threats	Zero-Day Detection Rate	58.9% (vs. 21.4% for Random Forest)
Classification Accuracy	Overall Accuracy on Mixed Classes	91.3%
Interpretability	Human-readable explanation via LLMs	Enabled (through GPT-style model integration)
Adaptability	Performance with analyst feedback	Significant uplift (+15% in zero-day accuracy)
Real-Time Viability	Embedding + Inference Latency	~0.8s average per event
Resource Efficiency	Model retraining requirements	None (zero-shot, no retraining needed)
Deployment Scalability	Integration with SOCs and SIEMs	Compatible with a modular pipeline
Ethical Compliance	Privacy-respecting architecture + explainability	Privacy-aware design with feedback validation

CONCLUSION

The main contribution of this research is ZTFER, a zero-shot attack observation model that integrates a base actual-time LLM reasoning and contrastive responses to recognise hidden cyber-threats without retraining with the model's semantic embeddings. In this paper, offered ZTFER, a novel zero-shot attack detection model which combines foundation method reasoning with Zero-Shot Learning (ZSL) to address the increasing issues resulting from previously hidden and fast-changing cyber-threats. The research additionally proposed ZTFER-RT, a combined real-time framework which promotes active and responsive attack detection through using language models and real-time threat feeds.

Through a zero-day detection success rate of 58.9%—much greater than the conventional models—and a total accuracy in classification is 91.3%, research findings promote ZTFER's powerful generalisation capacity. The technique's comprehension and flexibility have been confirmed by studies of ablation and latent space representation. ZTFER's working viability in dynamic cybersecurity circumstances is demonstrated through its real-time inference, zero retraining needed, and compatibility with human-in-the-loop operations.

For future work, there are three primary directions to improve ZTFER operation. Firstly, memory-augmented drivers which learn automatically from long-term communication histories and analysis feedback. Secondly, multipurpose combining, which includes logs, written content, and picture-based threats for comprehensive protection. Finally, retrieval-augmented generation (RAG) frameworks, which rapidly retrieve and reason across developing threat intelligence.

These improvements will contribute to the creation of strong, smart, and comprehensible systems for cybersecurity, which can adapt in real time and offer human-aligned support for decisions. The upcoming work will concentrate on memory-augmented representation, which acquires knowledge of long-term response, multimodal incorporation of images and logs, and RAG (Retrieval-Augmented Generation) for active attack intelligence.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the support provided by the University of Al-Hamdaniya. No external funding was received for conducting this research.

REFERENCES

Ahmed, U., Nazir, M., Sarwar, A., Ali, T., Aggoune, E. M., Shahzad, T., & Khan, M. A. (2025). Signature-based intrusion detection using machine learning and deep learning approaches empowered with fuzzy clustering. *Scientific Reports, 15*(1), Article 4481. https://doi.org/10.1038/s41598-025-85866-7

Alallaq, N., & Han, X. (2018). Detecting suspicious social astroturfing groups in tourism social networks. In *2018, the 5th International Conference on Behavioural, Economic, and Socio-Cultural Computing (BESC)* (pp. 58-62). IEEE. https://doi.org/10.3390/informatics12010004

Alallaq, N., Al-Mansoori, A., & Al-Sudani, A. R. (2019). Personalised reviews based on aspect analysis and polarity. In *2019, the 8th International Conference on Modelling, Simulation and Applied Optimisation (ICMSAO)* (pp. 1-6). IEEE. https://doi.org/10.1109/BESC.2018.8697307

Alhayani, M., & Al-Khiza'ay, M. (2022). Analyse symmetric and asymmetric encryption techniques by securing a facial recognition system. In M. Ben Ahmed, A. A. Boudhir, D. Santos, M. El Aroussi, & I. Karas (Eds.), *Innovations in smart cities applications volume 5* (pp. 97-105). Springer. https://doi.org/10.1007/978-3-031-15191-0_10

Aloqaily, A., Abdallah, E. E., AbuZaid, H., Abdallah, A. E., & Al-Hassan, M. (2025). Supervised machine learning for real-time intrusion attack detection in connected and autonomous vehicles: A security paradigm shift. *Informatics, 12*(1), Article 4. https://doi.org/10.1109/BESC.2018.8697307

Alzu'bi, A., Darwish, O., Albashayreh, A., & Tashtoush, Y. (2024). Cyberattack event logs classification using deep learning with semantic feature analysis. *Computers & Security, 147*, Article 104222. https://doi.org/10.1016/j.cose.2024.104222

Aminu, M., Akinsanya, A., Oyedokun, O., & Dako, D. A. (2024). Enhancing cyber threat detection through real-time threat intelligence and adaptive defence mechanisms. *International Journal of Computer Applications Technology and Research, 13*(8), 11-27. https://doi.org/10.7753/IJCATR1308.1002

Bathiri, K. A., & Vijayakumar, M. (2024). Enhancing intrusion detection system (IDS) through deep packet inspection (DPI) with machine learning approaches. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)* (pp. 1-7). IEEE. https://doi.org/10.1109/ADICS58448.2024.10533473

Bezas, K., & Filippidou, F. (2023). Comparative analysis of open-source security information and event management systems (SIEMs). *Indonesian Journal of Computer Science, 12*(2), 443-468. https://doi.org/10.33022/ijcs.v12i2.3182

Branescu, I., Grigorescu, O., & Dascalu, M. (2024). Automated mapping of common vulnerabilities and exposures to MITRE ATT&CK tactics. *Information, 15*(4), Article 214. https://doi.org/10.3390/info15040214

Bratsas, C., Anastasiadis, E. K., Angelidis, A. K., Ioannidis, L., Kotsakis, R., & Ougiaroglou, S. (2024). Knowledge graphs and semantic web tools in cyber threat intelligence: A systematic literature review. *Journal of Cybersecurity and Privacy, 4*(3), 518-545. https://doi.org/10.3390/jcp4030025

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing*

Cen, M., Deng, X., Jiang, F., & Doss, R. (2024). Zero-Ran Sniff: A zero-day ransomware early detection method based on zero-shot learning. *Computers & Security, 142*, Article 103849. https://doi.org/10.1016/j.cose.2024.103849

Chai, D., Wu, W., Han, Q., Wu, F., & Li, J. (2020). Description-based text classification with reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning* (*Proceedings of Machine Learning Research*, Vol. 119, pp. 1371-1382). PMLR.

Chamieh, I., Zesch, T., & Giebermann, K. (2024). LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 309-315). Association for Computational Linguistics.

Chen, Y., Ding, J., Li, D., & Chen, Z. (2021). Joint BERT model-based cybersecurity named entity recognition. In *Proceedings of the 2021 4th International Conference on Software Engineering and Information Management* (pp. 236-242). Association for Computing Machinery. https://doi.org/10.1145/3451471.3451508

Ghelani, D. (2022). *Deep learning and artificial intelligence framework to improve cybersecurity* [Preprint]. Authorea. https://doi.org/10.22541/au.166379475.54266021/v1

Ghosh, R., von Stockhausen, H.-M., Schmitt, M., Vasile, G. M., Karn, S. K., & Farri, O. (2025). CVE-LLM: Ontology-assisted automatic vulnerability evaluation using large language models. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(28), 28757-28765. https://doi.org/10.1609/aaai.v39i28.35139

Harsha, A. K., & Thyagaraja Murthy, A. (2021). Machine learning techniques for malware detection. *International Journal of Scientific Research in Science, Engineering and Technology, 8*(5), 70-76. https://doi.org/10.32628/IJSRSET21858

Idelhaj, A., Houda, Z. A. E., & Khoukhi, L. (2024). Foundation models in cybersecurity. In *Artificial Intelligence and Cybersecurity* (pp. 1-18). CRC Press. https://doi.org/10.1201/9781003497585-1

Imtiaz, N., Wahid, A., Abideen, S. Z. U., Kamal, M. M., Sehito, N., Khan, S., Virdee, B. S., Kouhalvandi, L., & Alibakhshikenari, M. (2025). A deep learning-based approach for the detection of various Internet of Things intrusion attacks through optical networks. *Photonics, 12*(1), Article 35. https://doi.org/10.3390/photonics12010035

Ke, H., Xu, J., Wang, Y., Chen, H., & Shen, Z. (2025). Adversarial machine learning in cybersecurity: Attacks and defences. *International Journal of Management Science Research, 8*(2), 26-33. https://doi.org/10.53469/ijomsr.2025.08(02).04

Korba, A. A., Boualouache, A., Brik, B., Rahal, R., Ghamri-Doudane, Y., & Senouci, S. M. (2023). Federated learning for zero-day attack detection in 5G and beyond V2X networks. In *ICC 2023-IEEE International Conference on Communications* (pp. 1137-1142). IEEE. https://doi.org/10.1109/ICC45041.2023.10279368

Kumar, V., Sinha, D., Das, A. K., Pandey, S. C., & Goswami, R. T. (2020). An integrated rule-based intrusion detection system: Analysis on the UNSW-NB15 data set and the real-time online dataset. *Cluster Computing, 23*(2), 1397-1418. https://doi.org/10.1007/s10586-019-03008-x

Mandal, U., Shukla, S., Rastogi, A., Bhattacharya, S., & Mukhopadhyay, D. (2024). µLAM: An LLM-powered assistant for real-time microarchitectural attack detection and mitigation. In *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design* (Article 168, pp. 1-9). Association for Computing Machinery. https://doi.org/10.1145/3676536.3676838

Muppalaneni, R., Inaganti, A. C., & Ravichandran, N. (2024). AI-driven threat intelligence: Enhancing cyber defence with machine learning. *Journal of Cybersecurity and Information Assurance*.

Papageorgiou, E., Chronis, C., Varlamis, I., & Himeur, Y. (2024). A survey on the use of large language models (LLMs) in fake news. *Future Internet, 16*(8), Article 298. https://doi.org/10.3390/fi16080298

Ranpara, R., Alsalman, O., Kumar, O. P., & Patel, S. K. (2025). A simulation-driven computational framework for adaptive energy-efficient optimisation in machine learning-based intrusion detection systems. *Scientific Reports, 15*(1), Article 13376. https://doi.org/10.1038/s41598-025-93254-4

Ring, M., Wunderlich, S., Scheuring, D., Landes, D., & Hotho, A. (2019). A survey of network-based intrusion detection data sets. *Computers & Security, 86*, 147-167. https://doi.org/10.1016/j.cose.2019.06.005

Saurabh, K., Sood, S., Kumar, P. A., Singh, U., Vyas, R., Vyas, O. P., & Khondoker, R. (2022). LBDMIDS: LSTM-based deep learning model for intrusion detection systems for IoT networks. In *2022 IEEE World AI IoT Congress (AIIoT)* (pp. 753-759). IEEE. https://doi.org/10.1109/AIIoT54504.2022.9817245

Shamsuzzaman, H. M., Mosleuzzaman, M., Mia, A., & Nandi, A. (2024). Cybersecurity risk mitigation in industrial control systems: Analysing physical, hybrid, and virtual test bed applications. *Academic Journal on Artificial Intelligence, Machine Learning, Data Science and Management Information Systems, 1*(1), 19-39. https://doi.org/10.69593/ajaimldsmis.v1i01.123

Socher, R., Ganjoo, M., Manning, C. D., & Ng, A. Y. (2013). Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems 26* (pp. 935-943). Curran Associates, Inc.

Srivastava, A., Sanghavi, P., Parmar, V., & Rani, S. (2024). Zero-shot learning in cybersecurity: A paradigm shift in attack and defence strategies. In M. Singh, V. Tyagi, P. K. Gupta, J. Flusser, T. Ören, A. R. Cherif, & R. Tomar (Eds.), *Advances in computing and data sciences: 8th International Conference, ICACDS 2024, Vélizy, France, May 9-10, 2024, revised selected papers* (pp. 138-149). Springer. https://doi.org/10.1007/978-3-031-70906-7_13

Sufi, F. (2024). Generative pre-trained transformer (GPT) in research: A systematic review on data augmentation. *Information, 15*(2), Article 99. https://doi.org/10.3390/info15020099

Tariq, N., Alsirhani, A., Humayun, M., Alserhani, F., & Shaheen, M. (2024). A fog-edge-enabled intrusion detection system for smart grids. *Journal of Cloud Computing, 13*(1), Article 43. https://doi.org/10.1186/s13677-024-00609-9

Yadav, N., Pande, S., Khamparia, A., & Gupta, D. (2022). Intrusion detection system on IoT with 5G network using deep learning. *Wireless Communications and Mobile Computing, 2022*, Article 9304689. https://doi.org/10.1155/2022/9304689

Yu, M., & Zhang, X. (2023). AlertInsight: Mining multiple correlation for alert reduction. *Computer Systems Science and Engineering, 46*(2), 2447-2469. https://doi.org/10.32604/csse.2023.037506

Yucel, M. K., Cinbis, R. G., & Duygulu, P. (2020). A deep dive into adversarial robustness in zero-shot learning. In A. Bartoli & A. Fusiello (Eds.), *Computer vision ECCV 2020 workshops* (pp. 3-21). Springer. https://doi.org/10.1007/978-3-030-66415-2_1

Zhang, A., Zhao, Y., Zhou, C., & Zhang, T. (2025). ResACAG: A graph neural network-based intrusion detection. *Computers & Electrical Engineering, 122*, Article 109956. https://doi.org/10.1016/j.compeleceng.2024.109956

Zhang, T., Gao, L., Li, X., & Gao, Y. (2025). DZAD: Diffusion-based zero-shot anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(10), 10131-10138. https://doi.org/10.1609/aaai.v39i10.33099

Pertanika J. Sci. & Technol. 34 (3): 1493 - 1523 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Automated Generation of Calculation-based Physics Questions Using Text-to-Text Transfer Transformer (T5) for Malaysian Upper Secondary Education

Tuong Kiet Ngang* and Chih How Bong

Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia

ABSTRACT

Creating assessment questions to evaluate student achievement is a time-intensive task, especially in specialised subjects like Physics, particularly for structured, calculation-based problems in Science, Technology, Engineering, and Mathematics (STEM) disciplines. This study presents a web-based Automatic Question Generation (AQG) system that generates Physics questions for Malaysian upper secondary levels (Form 4 and 5). Covering topics such as Force and Motion, Heat, Light and Optics (Form 4), Pressure, and Electricity (Form 5), the system leverages transfer learning through the fine-tuning of the Text-to-Text Transfer Transformer (T5) model, a state-of-the-art natural language processing (NLP) technique. The methodology encompasses data construction, pre-processing, dataset generation, fine-tuning T5 model, evaluation, and inference. Performance was assessed through system experiments, ROUGE-L automatic evaluations, and human evaluations by expert educators, focusing on relevance, correctness, usefulness, and variety. The high ROUGE-L scores (0.82–0.85) indicate strong alignment with reference questions, while human evaluations demonstrate that the system generates contextually relevant and high-quality questions. The results from this study show that the AQG system matches the template approach for quality, but it is far more flexible and saves teachers a lot of manual work. It can also be scaled easily should more questions are needed. A comparative analysis with ChatGPT-4 was conducted, revealing the edge that a purpose-built, structured generator has over a broad and open-ended one. In short, deep-learning NLP can automate domain-specific question writing and make large-scale assessment design much simpler. These findings should interest researchers in computational linguistics, AI, and test automation.

Keywords: Artificial intelligence, automatic question generation, deep learning, natural language processing, question generation, Text-to-Text Transfer Transformer (T5), transfer learning

ARTICLE INFO

Article history:
Received: 27 July 2025
Accepted: 04 May 2026
Published: 19 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.10

E-mail addresses:
tkngang@gmail.com (Tuong Kiet Ngang)
chbong@unimas.my (Chih How Bong)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

INTRODUCTION

Question generation (QG) is defined by Rus et al. (2008) as "the task of automatically generating questions from various inputs such as raw text, database, or semantic representation" (p. 1). This definition gives academics and researchers a clear starting point for choosing the right questions and inputs. There are three important automatic question generation (AQG) components, namely the input, the output and the relationship between the input and the output. This is shown in Figure 1.

Automatic question generation (AQG), on the other hand, converts written text into interrogatives that are logically warranted by the source. This approach has attracted much attention since MOOCs put large-scale assessment on the agenda (Goldbach & Hamza-Lup, 2017). In interactive applications like conversational agents, Dialogue Policy Learning (DPL) manages multi-turn flow (Kwan et al., 2023). However, a fundamental and distinct challenge for AQG is crafting a single, high-quality question from a context. This core task is crucial to AQG's expanding role in tutoring and health-literacy tools, confirming its rising status (Mishra et al., 2020).

Despite this progress, producing physics questions for Malaysia's upper-secondary curriculum remains a stubborn challenge. Manual writing is slow, teachers have limited assessment training, and pupils receive too few and often repetitive practice tasks, undermining motivation and timely feedback (Erinosho, 2013; Saleh, 2014). Most existing AQG tools were built for reading comprehension and falter on numerical reasoning. These limitations highlight the need for domain-specific systems (Kurdi et al., 2020).

To address this need, the present study introduces a T5-based generator to automate the creation of physics questions for Malaysia's upper-secondary curriculum. Through this generator, it aims to lighten teachers' workload and sustain learner engagement by matching items to current understanding (Chew & Cerbin, 2021). The paper details the framework (dataset preparation, system architecture and evaluation metrics), the experimental setup with test designs and quality criteria, the results with interpretation, and finally the principal findings alongside directions for future research.

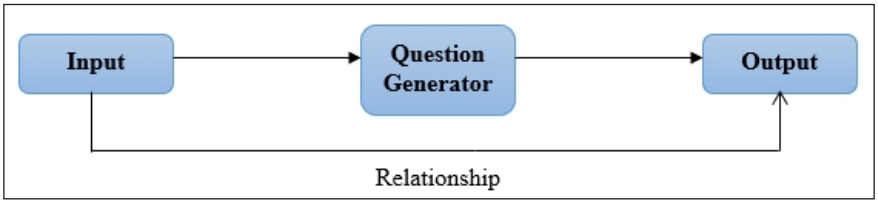


Figure 1. The components of AQG

PROPOSED METHODS

The methodology emphasises practical implementation and domain adaptation of the T5 model for physics question generation. The selection of the T5 architecture for the AQG system is justified by its strong performance in structured, domain-specific natural language generation tasks. Although newer foundational models have emerged, recent comparative studies suggest that the performance gap between fine-tuned T5 models and newer large language models is often marginal in educational contexts (Molina et al., 2024). For example, a study comparing GPT-3.5 Turbo, Llama 2-Chat 13B, and Flan-T5 XXL in generating questions from university lecture slides reported that although the newer LLMs slightly outperformed T5 in terms of clarity and question–answer alignment, all models exhibited comparable performance in relevance and question difficulty (Molina et al., 2024).

Furthermore, comparative studies Maity et al. (2024) report that T5 generally achieves stronger performance than BART and GPT-based models on automated evaluation metrics (e.g., ROUGE, BLEU) in educational NLP tasks. This is attributed to its unified text-to-text framework, which is well-suited for mapping structured inputs to natural language outputs in AQG.

Proposed Architecture of AQG

The proposed AQG architecture consists of six main processes: (1) Data Construction, (2) Data Pre-processing, (3) Dataset Generation, (4) Fine-tuning T5 Model, (5) Model Evaluation and Decoding, and (6) Inference. This architecture is illustrated in Figure 2. The T5 model, a transformer-based language model by Google, uses a "text-to-text" approach, enabling fine-tuning across tasks without task-specific architecture, reducing training data needs.

The architecture integrates template-based data construction with neural generalisation, enabling the T5 model to learn beyond fixed templates.

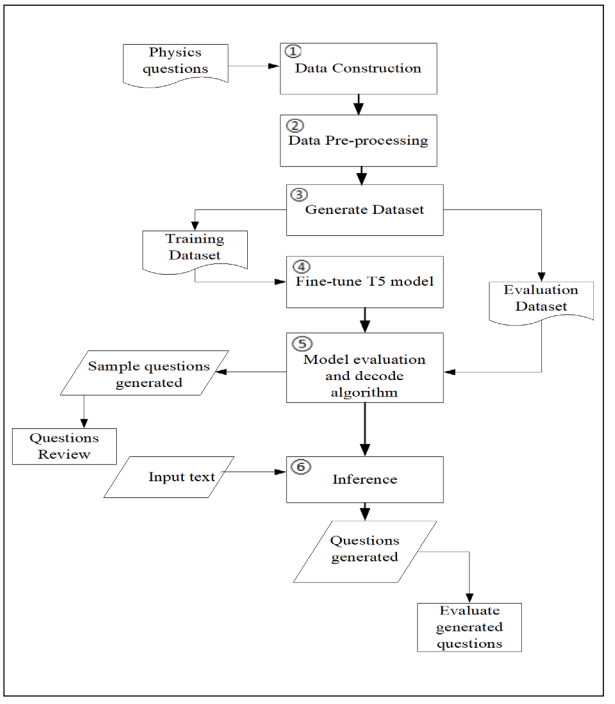


Figure 2. Proposed AQG architecture

Data Construction

In creating this generator, questions were exclusively derived from upper secondary Physics sources, including Form 4 and Form 5 textbooks (Choy et al., 2019; Chuan et al., 2020), the educational website Bumi Gemilang (https://www.bumigemilang.com), and a teacher-provided question bank. Sample questions underwent careful selection for clarity, accuracy, and relevance, following a structured procedure to ensure quality and appropriate difficulty levels. These questions align with curricular objectives and practical scenarios, with selection criteria prioritising relevance, coverage of content, and difficulty appropriate to each topic and subtopic. Three experienced Physics teachers reviewed and discussed each question to ensure consistency and educational alignment.

To introduce diversity and to ensure the system assesses computational skills rather than surface-level numeric substitution, sample questions were manually modified into eight distinct formats: "What is," "Determine," "Find," "Calculate," "What is the amount," "What is the value," "Figure out," and "How much." Each question was rephrased or restructured accordingly while retaining the central concept. For instance, "What is the acceleration of an object?" was adapted to "Determine the acceleration of the object" or "What is the value of acceleration," as demonstrated in Table 1. This variety requires students to actively decode the narrative context to identify physical variables.

In this study, a total of 4 modified question types were determined for different topics or subtopics. The primary emphasis was on the eight specific question types mentioned earlier.

Table 1
New question types were modified from the original question type

Original Question Type	New Modified Question Types
What is the acceleration of an object?	1. 1. **What** is the value of acceleration when a luggage with mass 15 kg is pulled by Elizabeth along a smooth horizontal surface with a force of 71 N?
	2. Elizabeth pulls a luggage of mass 15 kg along a smooth horizontal surface by a force of 71 N. **What** is the acceleration of the luggage?
	3. A piece of luggage with mass 15 kg is pulled by Elizabeth along a smooth horizontal surface. The force applied to the luggage is 71N. **Determine** the acceleration of the luggage.
	4. 4. Elizabeth pulls a luggage of mass 15 kg along a smooth horizontal surface by a force of 71 N. **Find** the acceleration of the luggage?

Data Pre-processing

Data pre-processing generates varied question types by randomly replacing values within set ranges, such as mass and force in acceleration questions. This process, automated using Python in the Anaconda Spyder environment, enhances scalability, maintains consistency, and minimises human error. This approach also saves time and ensures consistent randomisation across the dataset. Table 2 provides a summary of key variables and their ranges for various physics topics in Form 4 and 5.

Table 2
Variables and their ranges for selected Physics topics

No.	Form 4		
	Topic (question)	**Variables**	**Range of Values**
1	Topic: Linear Motion (displacement)	initial velocity	30 – 40 m/s
		duration	0.12 – 0.18 minutes (precision up to 2 decimal points)
		final velocity	50 – 60 m/s
2	Topic: Force (force)	mass	20 – 30 kg
		acceleration	0.5 – 2.0 m/s2 (precision up to 2 decimal points)
3	Topic: Force (acceleration)	force	60 – 80 N
		mass	10 – 20 kg
4	Topic: Force (mass)	force	60 – 70 N
		duration	5 – 9 s
		velocity	30 – 40 m/s
5	Topic: Specific Heat Capacity (specific heat capacity)	mass	0.2 – 0.6 kg (precision up to 2 decimal points)
		power	30 – 60 W
		initial temperature	18 – 21 oC
		final temperature	40 – 46 oC
		duration	80 – 100 s
6	Topic: Specific Latent Heat (specific latent of fusion)	power	450 – 500 W
		mass	0.22 – 0.30 kg (precision up to 2 decimal points
		duration	2 – 4 minutes
7	Topic: Gas Laws (temperature of air)	initial pressure	200 – 230 kPa
		initial temperature	24 – 27 oC
		final pressure	250 – 270 kPa
8	Topic: Refraction of Light (apparent distance)	depth	6 – 10 cm
		height	50 – 70 cm
9	Topic: Thin Lens Formula (size of image)	height	6 – 10 cm
		distance	50 – 70 cm
		focal length	12 – 18 cm

Table 2 (*continued*)

No.	Form 5		
	Topic (question)	**Variables**	**Range of Values**
1	Topic: Resultant Force (friction force)	force	70 – 80 N
		mass	5 – 10 kg
		acceleration	5 – 10 m/s2
2	Topic: Elasticity (elastic potential energy)	spring constant	20 – 30 N/cm
		mass	80 – 100 N
3	Topic: Pressure in Liquids (pressure in liquids)	depth	0.5 – 2.0 m (precision up to 2 decimal points)
		density	900 – 1200 kg/m3
4	Topic: Archimedes Principle (buoyant force)	volume	0.002 – 0.005 m3 (precision up to 3 decimal points)
		density of water	997 – 1003 kg/m3
5	Topic: Current and Potential Difference (current flow)	electric charge	500 – 550 C
		duration	3 – 7 minutes
6	Topic: Electrical Energy and Power (electrical energy)	potential difference	5 – 10 V
		current	0.5 – 0.9 A (precision up to 2 decimal points)
		duration	2 – 4 minutes

Dataset Generation

In this process, a modified question type is structured within a template, as shown in Table 3. Using a template format enhances organisation, clarity, and adaptability for deep learning analysis. Physics question templates cover various topics and difficulty levels, with each topic having a main template for core questions and three sub-templates introducing variations in problem-solving, conceptual understanding, and applied scenarios. In total, there are 15 main templates and 45 sub-templates, ensuring comprehensive and diverse question generation.

Table 3
Template to generate a dataset for acceleration

Object	Name	Question Type for Acceleration
Cable box	Maria	**1.** **Elizabeth** pulls a **NN** of mass **x** kg along a smooth horizontal surface by a force of **y** N. What is the acceleration of the **NN**?
Wooden door	Andrew	
Luggage	Terry	
Cement bag	Tommy	
Bench	Isabella	
Cupboard	Leslie	

Table 3 illustrates a sample question template with placeholders for objects and names. In an acceleration question, for example, placeholders (e.g., "NN" for objects and "PERSON" for names) are substituted with specific items and human names (like "Elizabeth") to increase question diversity. This substitution involves selecting an object from a list and using a function with regular expressions to replace placeholders efficiently. Named Entity Recognition (NER), via libraries like SpaCy, was used to identify and substitute human names to enhance the dataset variety for physics questions.

During dataset generation, Named Entity Recognition (NER) is employed to identify and tag human names within question templates. For example, in a template involving acceleration, names like "Elizabeth" are automatically identified by the NER model and labelled as "PERSON" to enable systematic substitution. This process also extends to other placeholders: the object placeholder "NN" is replaced with items from a predefined list (e.g., "table," "luggage"), while the numerical variables x (mass) and y (force) are replaced with values sampled from specified numerical ranges. This structured approach, which leverages NER for entity replacement and value ranges for numerical variability. Table 4 illustrates this NER-based process.

Using NLP techniques like Named Entity Recognition (NER) and object substitution, original question templates were customised with various names and objects to generate diverse scenarios. Python programmes further varied questions by substituting variables with random numbers, creating a robust dataset of 6,000 questions across 15 physics topics of varying difficulty. The dataset was split 70% (4,200) for training and 30% (1,800) for evaluation, covering 13 main topics structured into six chapters from the Form 4 and 5 syllabi, with four question types per topic. Table 5 presents examples of the "What is" question type, showing "target_text," "input_text," and "prefix" columns.

Table 4
NER process in dataset generation

Entity Type	Original Placeholder	NER Identification	Example of Substitution
PERSON	Elizabeth	Identified as PERSON	Andrew, Tommy

Table 5
Generated dataset for acceleration using questions type What is

target_text	input_text	prefix
Maria pulls a cable box of mass 10 kg along a smooth horizontal surface by a force of 68 N. **What is the acceleration of the cable box?**	{'mass', 'force', 'cable box'}	ask_question
Andrew pulls a luggage of mass 14 kg along a smooth horizontal surface by a force of 62 N. **What is the acceleration of the luggage?**	{'force', 'mass', 'luggage'}	ask_question

Table 5 (*continued*)

target_text	input_text	prefix
Terry pulls a cement bag of mass 20 kg along a smooth horizontal surface by a force of 76 N. **What is the acceleration of the cement bag?**	{'cement bag', 'mass', 'force'}	ask_question

Table 5 shows variations of similar questions generated by replacing values for mass, force, person names, and objects, producing a diverse dataset. For instance, mass and force vary across examples (e.g., 10 kg and 68 N with "Maria" and "cable box"; 14 kg and 62 N with "Andrew" and "luggage"; 20 kg and 76 N with "Terry" and "cement bag"). Each dataset pair includes an input_text containing key variables, sometimes with corrupted inputs, and a target_text that frames a complete question. This structure enables machine learning models to learn question generation patterns. The dataset, covering selected Form 4 and Form 5 Physics topics, was split into training and evaluation sets to train and assess a T5 model.

Although templates are used to structure the dataset, the diversity of templates, linguistic reformulations, and variable substitutions enables the model to learn generalisable question generation patterns rather than memorising fixed structures.

Fine-tune T5 Model

The T5 base model was selected due to its strength in generative tasks (Xue et al., 2021), such as those highlighted in modern e-learning applications (Patil et al., 2022). This choice is further supported by its proven effectiveness in generating structured educational content in STEM domains, such as the use of the multilingual T5 (mT5) model for creating math word problems (Gao, 2023). The text-to-text framework of T5 is particularly advantageous as it allows diverse NLP tasks, including question generation, to be modelled uniformly as generating target text from a source text prompt.

The model applies the Masked Language Model (MLM) objective (Shi & Wolff, 2021), which involves predicting masked portions in the input, making it suitable for learning variable placement and structure in Physics questions. During fine-tuning, model weights are updated via backpropagation to reduce prediction errors. The final tuned weights are saved for evaluation and inference. Detailed fine-tuning setup, dataset configuration, and training parameters are discussed in the section “Fine-tuning T5 Model for AQG”.

Model Evaluation and Decoding

After fine-tuning, the AQG model was evaluated to assess the relevance, correctness, usefulness, and diversity of generated Physics questions. Each question was reviewed and compared to reference questions for quality. To generate coherent and natural-sounding

questions, decoding algorithms such as Top-k and Top-p sampling were used. These approaches filter word choices during generation to balance coherence and diversity (Tucker, 2020; von Platen, 2020). Decoding strategies (Top-k and Top-p) are employed to encourage variation in generated questions. Full evaluation configuration, sampling settings, and loss trends are described in the section "Evaluation Configuration and Decoding Settings".

Inference

The fine-tuned model was used for inference via a web-based Anvil interface linked to a Jupyter Notebook backend. Users entered Physics variables, and an object, and the system generated four diverse questions automatically. For example, with initial velocity, final velocity, and duration as inputs, and "car" as the object, the system produced four displacement questions. Four examples are shown in Table 6.

Table 6
Question generated for displacement during inference

Level:	Form 4
Chapter:	Force and Motion I
Topic:	Linear Motion
Question to be asked:	displacement
Input text (variables):	initial velocity, final velocity, duration
Object involved:	car
Generated Questions	
1.	Andrew drives a car on a straight path. The car accelerates uniformly from a velocity of 33 m/s to a velocity of 52 m/s in 0.13 minutes. **What is the total displacement** travelled by the car?
2.	Andrew drives a car on a straight path. The car accelerates uniformly from a velocity of 39 m/s to a velocity of 52 m/s in 0.15 minutes. **What is the total displacement** travelled by the car?
3.	As a car moves along a straight track, its velocity is 35 m/s. After 0.13 minutes, the car has reached 51 m/s. **What is its displacement?**
4.	A car is moving along a straight road with a velocity of 40 m/s. 0.16 minutes later, its velocity has increased to 56 m/s. **Determine the displacement of the car?**

EXPERIMENT SETUP

Dataset

In this research, the T5 model's input dataset was manually compiled from Physics sample questions across various topics and subtopics in Form 4 and Form 5. A total of 6 chapters and 13 topics were included (Form 4: 3 chapters, 7 topics; Form 5: 3 chapters, 6 topics), as shown in Table 7.

In addition, all labelling and question combinations were done manually to ensure accuracy, as detailed in Table 8. The target text represents the "questions set"

intended for generation or analysis, while the input text includes variables relevant to each question type. This provides context for meaningful question formulation. The prefix "ask_question" acts as a string identifier for executing the task of question generation or processing.

Table 7
List of chapters and topics

Form 4	**Topic (question)**
Chapter	
2	Topic: Linear Motion (displacement)
2	Topic: Force (force)
2	Topic: Force (acceleration)
2	Topic: Force (mass)
4	Topic: Specific Heat Capacity (specific heat capacity)
4	Topic: Specific Latent Heat (specific latent of fusion)
4	Topic: Gas Laws (temperature of air)
6	Topic: Refraction of Light (apparent distance)
6	Topic: Thin Lens Formula (size of image)
Form 5	
1	Topic: Resultant Force (friction force)
1	Topic: Elasticity (elastic potential energy)
2	Topic: Pressure in Liquids (pressure in liquids)
2	Topic: Archimedes Principle (buoyant force)
3	Topic: Current and Potential Difference (current flow)
3	Topic: Electrical Energy and Power (electrical energy)

Table 8
Dataset (combination of different Physics topics)

	target_text	**input_text**	**prefix**
1.	What is the value of acceleration when a luggage with mass 17 kg is pulled by Tommy along a smooth horizontal surface with a force of 74 N?	force, mass, luggage	ask_ question
2.	In a physics class, a student prepared a simple electrical circuit that connected a 6 V battery and a bulb. Calculate the electrical energy that is supplied for 2 minutes if the current is 0.71 A?	potential difference, bulb, current, duration	ask_ question
3.	What is the size of the image of a cup with a height of 6 cm, when it is placed at 63 cm from a concave lens with a focal length of 18 cm?	focal length, cup, distance, height	ask_ question

Table 8 (*continued*)

	target_text	input_text	prefix
4.	A steel block with a volume of 0.005 m3 is immersed in water. If the density of water is 997 kg/m3, what is the buoyant force experienced by the steel block? Gravitational acceleration is 9.81 m/s2.	steel block, density of water, volume	ask_ question
5.	A lorry is moving along a straight road with a velocity of 30 m/s. 0.14 minutes later, its velocity has increased to 53 m/s. Determine the displacement of the lorry?	initial velocity, lorry, duration, final velocity	ask_question

Implementation Details

Fine-tuning T5 Model for AQG

The T5-base pretrained model was fine-tuned using a dataset of 6,000 samples, split into 70% for training and 30% for evaluation (Table 9).

Table 9
Dataset distribution for model training

Dataset	Number of Samples
Training dataset	4200
Evaluation dataset	1800

Fine-tuning involved adjusting the model parameters for the AQG task using this custom dataset. The hyperparameters used are summarised in Table 10.

Table 10
Fine-tuning hyperparameters for the T5 model

Parameter	Value
Model	T5-base
Batch size	5
Number of epochs	3
Learning rate	1e-4
Maximum sequence length	80 tokens

These values were selected based on standard practices and hardware constraints. A batch size of 5 was used due to the relatively small dataset size and limited memory availability, enabling more frequent updates and reducing overfitting risk. The model was trained for 3 epochs, as training loss stabilised by then, minimising the chance of overtraining. A learning rate of 1e-4 allowed for gradual and stable updates, which is preferable for specialised datasets compared to the commonly used 0.001 rate (Raffel et al., 2020). The maximum sequence length of 80 tokens balanced processing efficiency and input completeness, compared to longer default lengths used in other Transformer-based models (Devlin et al., 2019). The AdaFactor optimiser (Shazeer & Stern, 2018) was used to reduce memory usage during training, which completed in approximately 15 minutes. The Weights and Biases (wandb) platform was used to log training metrics and visualise progress.

Figure 3 shows the total training time, while Figure 4 presents the training loss curve. The steadily decreasing loss indicates effective learning and convergence.

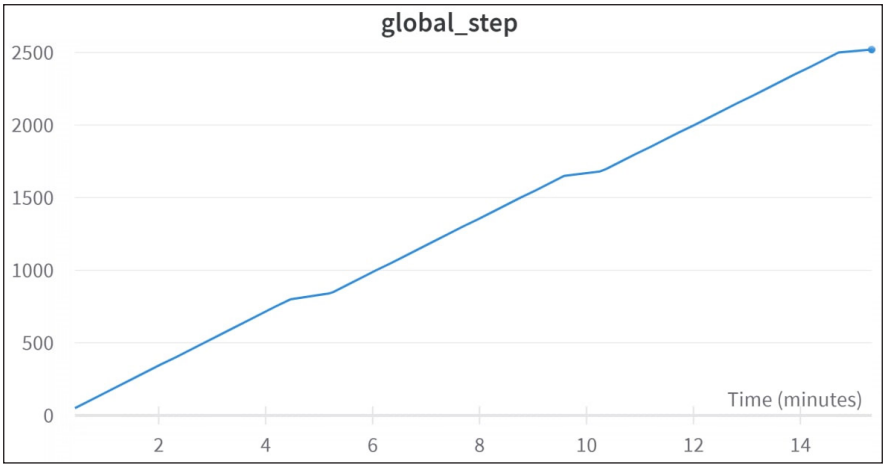


Figure 3. Total time taken for training the T5 model

Figure 4. Training loss curve

Evaluation Configuration and Decoding Settings

The AQG model was evaluated by generating questions from input text and comparing them to reference questions. Using parameters from Table 11, the evaluation ran in batches of 3, with 80 token sequences over one epoch, optimising memory and efficiency while preventing overfitting.

Table 11
Evaluation parameters for the AQG model

Parameter	Value
Batch size	3
Number of epochs	1
Maximum sequence length	80 tokens

The evaluation loss curve (Figure 5) shows a decrease over time, indicating the model's ability to generalise and learn effectively. Evaluation settings were adjusted based on early loss behaviour. A smaller batch size of 3 provided more frequent updates and improved learning precision for this generative task. Though smaller batches introduce noise, they can act as a regularise, enhancing generalisation.

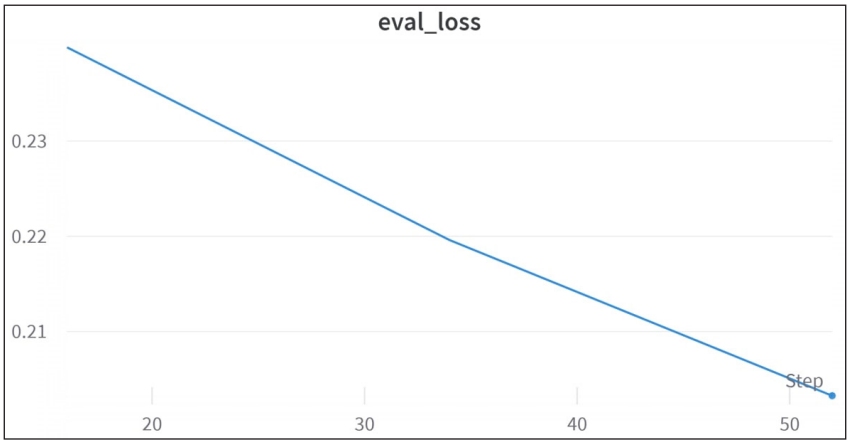


Figure 5. Graph of evaluation loss

The number of epochs was fixed at 1, as early experiments showed that additional epochs led to marginal improvement while increasing the risk of overfitting. A maximum sequence length of 80 was retained, balancing context capture and training efficiency. For decoding, this study used a combination of Top-k =50 and Top-p = 0.90 sampling, as shown in Table 12. These probabilistic methods filter out unlikely word candidates and dynamically adjust the word pool based on cumulative probability (Fan et al., 2018; Holtzman et al., 2020). This configuration consistently yielded coherent and contextually appropriate questions.

Table 12
Value of the sampling scheme

Parameter	Value
Top_k	50
Top_p	0.90

Four output sequences were generated for each target row in the evaluation dataset. All generated questions were stored in a .txt file. Samples are illustrated in Tables 14 and 16, based on real questions from Tables 13 and 15, respectively.

Table 13
Input text and real question for acceleration in the evaluation dataset

Input Text:	**{'mass', 'force', 'bench'}**
Real question:	What is the value of acceleration when a bench with mass 14 kg is pulled by Leslie along a smooth horizontal surface with a force of 63 N?

Table 14
Generated questions for acceleration

1.	Andrew pulls a bench of mass 18 kg along a smooth horizontal surface by a force of 80 N. What is the acceleration of the bench?
2.	Leslie pulls a bench of mass 19 kg along a smooth horizontal surface by a force of 68 N. Find the acceleration of the bench?
3.	A bench with mass 18 kg is pulled by Terry along a smooth horizontal surface. The force applied to the bench is 68 N. Determine the acceleration of the bench?
4.	What is the value of acceleration when a bench with mass 10 kg is pulled by Tommy along a smooth horizontal surface with a force of 63 N?

Table 15
Input text and a real question for specific heat capacity in the evaluation dataset

Input Text:	**{'bronze medal', 'mass', 'duration', 'temperature', 'power'}**
Real question:	What is the specific heat capacity of the bronze medal with mass 0.5 kg, when it is heated by an electric heater with a power of 57 W for 92 s and the temperature increases from 19 oC to 40 oC?

Table 16
Generated questions for specific heat capacity

1.	A bronze medal with mass 0.57 kg is heated by an electric heater with a power of 45 W for 85 s. The temperature of the bronze medal increases from 18 oC to 45 oC. Calculate the specific heat capacity of the bronze medal?
2.	A bronze medal with mass 0.53 kg is heated by an electric heater with a power of 46 W for 87 s. The temperature of the bronze medal increases from 20 oC to 40 oC. Determine the specific heat capacity of the bronze medal?
3.	What is the specific heat capacity of the bronze medal with mass 0.43 kg, when it is heated by an electric heater with a power of 39 W for 97 s and the temperature increases from 21 oC to 44 oC?
4.	A bronze medal with mass 0.35 kg is heated by an electric heater with a power of 39 W for 89 s. The temperature of the bronze medal increases from 18 oC to 42 oC. Determine the specific heat capacity of the bronze medal?

Inference Response Time

The AQG system measures response time using a Jupyter server and an Anvil web app. Users input data, which Anvil sends to Jupyter for question generation. The Jupyter notebook records response times, helping identify bottlenecks and optimise performance.

Experiment

Experiment 1: Question Generation with a New Object

This experiment tests the AQG model's ability to generate questions for unseen objects and achieve denoising using masked language modelling (MLM). Success is measured by the model's accuracy in predicting masked tokens and restoring sequences with minimal errors. This process improves question generation by training the model to handle incomplete inputs, enhancing its ability to produce coherent, relevant questions with high relevance and robustness.

This experiment focused on testing the concept of Force from the Form 4 chapter on Force and Motion I. Using the variables mass and acceleration, a new object, "trolley," was introduced to generate Force-related questions (Table 17).

Table 17
Experimental parameters for force-related question generation with a new object

Level:	Form 4
Chapter:	Force and Motion I
Topic:	Force
Question to be generated:	force
Input text (variables):	mass, acceleration
Object involved:	trolley

Table 18
Experimental parameters for pressure in liquids-related question generation with new object

Level:	Form 5
Chapter:	Pressure
Topic:	Pressure in Liquids
Question to be generated:	pressure in liquids
Input text (variables):	depth, density
Object involved:	prawn

The next topic in this experiment focused on pressure in liquids from the Form 5 pressure chapter. The question involved pressure in liquids, with "prawn" as the new object and depth and density as variables (Table 18).

The AQG model experiments used a Form 4 topic (force) and a Form 5 topic (pressure in liquids) to assess question generation. New objects like "trolley" (for force) and "prawn" (for pressure) tested the model's generalisation. These choices enhanced question diversity, reflecting real-world physics and challenging the model's contextual accuracy.

Experiment 2: Multiword Target

This experiment examines the AQG model's handling of multiword targets, as BERT and RoBERTa typically map single words to masks (Shi & Wolff, 2021). Using "steel pot" for specific heat capacity, the study assessed how well the model linked input variables to context. Table 19 presents the data. Success was measured by comparing generated questions with expected outputs, providing insights into the model's ability to process complex linguistic structures.

Other additional experiments are to generate questions for the topics Thin Lens Formula and Current and Potential Difference. These questions focus on calculating the size of an image (using focal length, distance, and height for a plastic bottle) and determining the current flow (using electric charge and duration for a copper block), exploring how the AQG model handles multiword targets.

Table 19
Experimental settings for AQG model handling of multiword targets for topic-specific heat capacity

Level:	Form 4
Chapter:	Heat
Topic:	Specific Heat Capacity
Question to be generated:	specific heat capacity
Input text (variables):	power, temperature, mass, duration
Object involved:	steel pot

Comparison Between ChatGPT-4 and AQG System

The emergence of Large Language Models (LLMs) has transformed education, particularly in automated question generation, by offering unprecedented generative capabilities (Kasneci et al., 2023). However, alongside these opportunities exist significant challenges concerning their reliability and suitability for specialised pedagogical tasks. This section addresses this gap by comparing the effectiveness of a purpose-built AQG system based on the T5 model with ChatGPT (GPT-4) in generating calculation-based physics questions. The evaluation focuses on four key dimensions: relevance, correctness, usefulness, and variety. These aspects are assessed by three experienced physics teachers, as detailed in the section "Evaluating Questions Generated by ChatGPT-4 and AQG System".

To ensure consistent experimental conditions, both systems were evaluated using identical input prompts. Nevertheless, the comparison reflects different modelling paradigms, as the T5-based AQG system is fine-tuned on domain-specific data, whereas ChatGPT-4 is evaluated using an instruction-guided zero-shot prompting approach. ChatGPT-4 was configured using custom instructions aligned with Table 20 to ensure consistent response structure, linguistic behaviour, and output constraints. In addition, both ChatGPT-4 and the AQG system were evaluated under a single-run generation setting, without repeated sampling, to maintain parity and control variability across systems.

Configuring Custom Instructions in ChatGPT-4

Custom instructions let users set preferences for ChatGPT-4, ensuring consistent responses without repetition. For this comparison, instructions are tailored to align with Table 20, ensuring relevant replies.

Table 20
Data and details for custom instruction alignment in ChatGPT-4

Custom Instructions	Data and Details
What would you like ChatGPT to know about you to provide better responses?	I am doing automatic question generation for physics calculation questions.
How would you like ChatGPT to respond?	Please help to generate a question, based on these details: 1. Question to be generated 2. The input 3. Object Condition: add value to the input Then, add diversity to the questions by using this format (What is, Determine, Find, Calculate, what is the amount, What is the value? Figure out, and How much? ** Note: Set the limit to generate 4 different types of questions only.

Evaluation Method

The evaluation assessed AQG-generated and template-based questions using human and automatic methods. Three Physics teachers evaluated AQG questions on relevance, correctness, usefulness, and variety, while template-based questions were assessed on relevance, correctness, and usefulness. The evaluators each possess over 15 years of experience in the Malaysian secondary education system and specialise in Physics teaching and curriculum, supporting the reliability and credibility of the human evaluation. For automatic evaluation, the nlg-eval Python package (Sharma et al., 2017) with the ROUGE-L metric was used to measure syntactic and semantic similarities by comparing the longest common subsequence (LCS) between generated and reference questions.

Human Evaluation

Human evaluation was conducted using four criteria: correctness, relevance, usefulness, and variety. A 1-4 scale adapted from previous work was used to ensure systematic and consistent assessment, where 1 indicated poor quality, and 4 indicated excellent quality (Amidei et al., 2018; Le & Pinkwart, 2015; Yao et al., 2012). Correctness evaluated grammatical accuracy and fluency, relevance assessed how well the question aligned with the input and the national curriculum, usefulness considered the question's potential to support meaningful learning (pedagogical value); and variety evaluated the diversity of question types, where a score of 1 indicated one type and 4 indicated four distinct types. The human evaluation involved three Physics teachers, who validated a total of

180 generated questions. Their evaluation particularly emphasised the criteria of usefulness and relevance to the national curriculum, thereby strengthening the credibility and educational alignment of the evaluation. The template-based approach used correctness, relevance, and usefulness

The reliability of the human evaluation was assessed using Randolph's Free-Marginal Kappa. This metric was chosen to provide a more accurate reflection of agreement than Fleiss' Kappa, which can be artificially deflated when rater agreement is exceptionally high (the 'Kappa Paradox'). The analysis was conducted on 180 evaluation units generated by the T5-based AQG model, excluding template-based results to focus strictly on the transformer model's performance.

Evaluation Setup of the AQG Model

A web page enabled AQG model evaluation, where users input data processed by a Jupyter server. Three Physics teachers generated four questions per click (Figure 6) and copied them to a Word template for review. Each evaluator reviewed 60 AQG-generated questions and 15 template-based questions. While the AQG system generated different questions for each evaluator, the template-based questions were identical, as there was only one set per template. This ensured a comprehensive and consistent assessment across both types of questions.

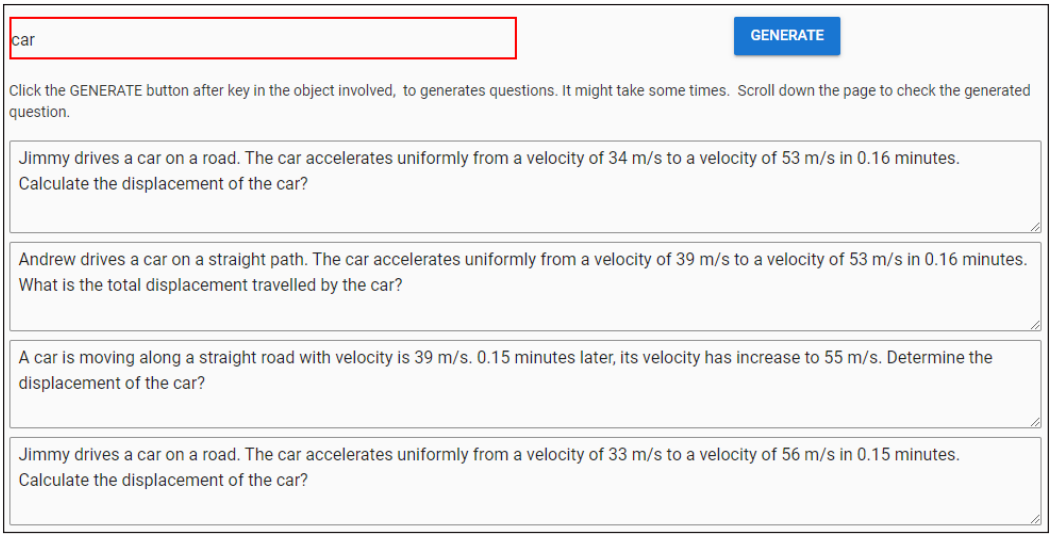


Figure 6. Questions generated using the AQG method

Evaluation Setup of Template-based Generated Questions

Template-based question evaluation followed a structured process. Question templates were created for various Physics topics, and a program allowed evaluators to select a topic and generate questions by filling template placeholders with relevant Physics data (Figure 7).

Automatic Evaluation

The nlg-eval tool evaluated generated questions by processing data from three evaluators, each generating 60 questions stored in separate files (hyp1.txt, hyp2.txt, hyp3.txt). These were then compared to a golden reference set (ref.txt), ensuring alignment by question type and topic. The nlg-eval command was executed for each hypothesis file against the reference, focusing on ROUGE-L, which measures the longest matching word sequence to assess accuracy and relevance (Figure 8).

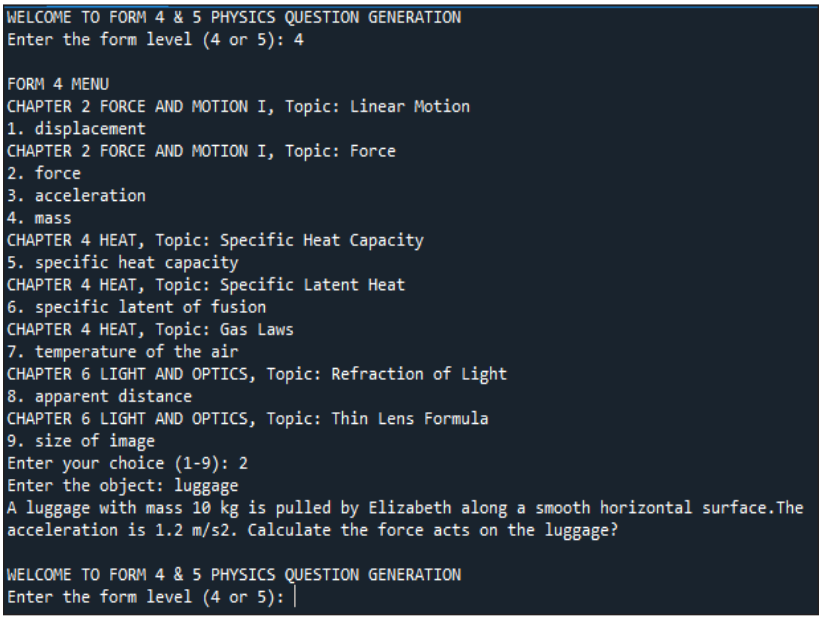


Figure 7. Questions generated using the template-based approach

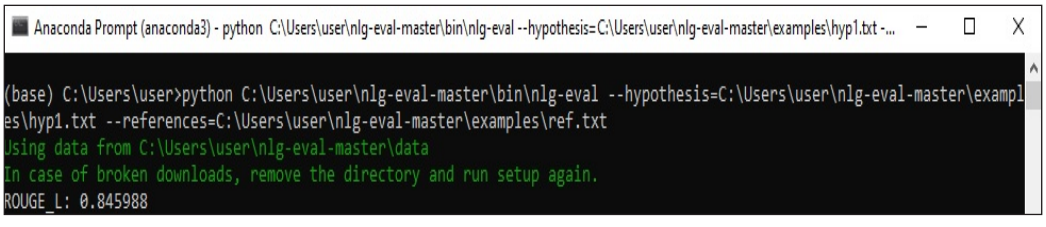


Figure 8. Evaluation of hypothesis files using the nlg-eval command

RESULT AND DISCUSSION

This section discusses experimental results and human evaluation findings to assess the quality of generated questions. The analysis offers insights into their effectiveness in educational question generation.

Inference Response Time Result

The response time test measured the time required to generate questions, with four questions generated per input to ensure consistent evaluation. The average number of words and generation time per question (in seconds) were recorded (Table 21). The longest processing time, 1.96 s, was observed for the Specific Heat Capacity topic, corresponding to the highest average word count (42.5). The overall mean response time per question was 1.89 s, indicating satisfactory response times for the AQG system.

Table 21
Result of the response time test

Form 4	Topic (question)	Average No. of Words per Question	Average Generation Time per Question (seconds)
Chapter			
2	Topic: Linear Motion (displacement)	36	1.94
2	Topic: force (force)	26.5	1.86
2	Topic: force (acceleration)	26	1.84
2	Topic: force (mass)	31	1.89
4	Topic: Specific Heat Capacity (specific heat capacity)	42.5	1.96
4	Topic: Specific Latent Heat (specific latent of fusion)	41.5	1.95
4	Topic: Gas Laws (temperature of air)	37	1.94
6	Topic: Refraction of Light (apparent distance)	40	1.88
6	Topic: Thin Lens Formula (size of image)	32	1.86
Form 5			
1	Topic: Resultant Force (friction force)	29	1.85
1	Topic: Elasticity (elastic potential energy)	33	1.86
2	Topic: Pressure in Liquids (pressure in liquids)	39	1.94
2	Topic: Archimedes Principle (buoyant force)	30	1.89
3	Topic: Current and Potential Difference (current flow)	19	1.77
3	Topic: Electrical Energy and Power (electrical energy)	34	1.90
	Mean average time per question		1.89

This performance can be evaluated against established Human–Computer Interaction (HCI) benchmarks and reported baseline latencies in educational AI workflows (Cornejo et al., 2024; Jauhiainen & Guerra, 2024), further supporting its suitability for interactive applications. Compared to general-purpose systems, which typically exhibit higher response times, the AQG system demonstrates improved efficiency for real-time educational use.

The observed latency of 1.89 seconds falls within the recommended response time for interactive systems, typically defined as between one and two seconds (Roth, 2013). Furthermore, based on the Seow framework as discussed in Cornejo et al. (2024), response times under five seconds are categorised as "continuous" interaction, indicating that system feedback remains sufficiently immediate to support user engagement. This level of responsiveness ensures that the question generation process does not disrupt the natural dialogue between teacher and system, thereby preserving the flow of thought essential for effective classroom pacing (Cornejo et al., 2024; Wang & Reani, 2017).

To provide a clearer rationale, the efficiency of the specialised T5 model was compared against reported baseline inference times of general-purpose Large Language Models (LLMs) used in educational contexts. While the underlying tasks and system configurations differ, prior studies consistently report higher per-interaction latency for such models. For example, Jauhiainen and Guerra (2024) report that GPT-4 requires an average of 18.51 seconds to process a single educational request. Similarly, Parker et al. (2024) indicate that GPT-4 responses for educational feedback tasks typically take approximately 10 seconds.

Although general-purpose models offer broader capabilities, their relatively high latency (approximately 10–18 seconds per interaction) may limit their suitability for time-sensitive, real-time educational applications. In contrast, the proposed specialised AQG system achieves an average response time of 1.89 seconds, suggesting a substantial improvement in responsiveness and making it more suitable for live, teacher-led instructional settings.

Experiment Result and Discussion

Experiment 1: Result Discussion

Experiment 1 (Force concept) demonstrates the AQG system's ability to generate questions for unseen objects, such as a "trolley". The system successfully incorporates the new object into generated questions in this controlled test case, as shown in Table 22, indicating its capability to generalise to unfamiliar inputs. This suggests that the model can adapt to new object instances by leveraging learned patterns from the training data rather than relying solely on memorised templates. However, this result should be interpreted as a case-specific observation rather than evidence of perfect performance. It reflects successful object handling under controlled conditions, while broader performance variations are observed in other experiments, as shown in Table 23.

Table 22
Experimental results on using a new object to generate questions for the topic Force

	Generated Questions
1.	What is the Force applied on a **trolley** with mass 19 kg, if it is pulled by David along a smooth horizontal surface with acceleration 1.29 m/s2?
2.	A **trolley** with mass 14 kg is pulled by David along a smooth horizontal surface. The acceleration is 0.86 m/s2. Calculate the force that acts on the **trolley**?
3.	A **trolley** with mass 19 kg is pulled by David along a smooth horizontal surface. The acceleration is 1.54 m/s2. Determine the Force of the **trolley**?
4.	A **trolley** with mass 12 kg is pulled by Tommy along a smooth horizontal surface. The acceleration is 0.63 m/s2. Calculate the force that acts on the **trolley**?

Table 23
Experimental results on using a new object to generate questions for the topic Pressure in Liquids

1.	A **shrimp** is at a depth of 1.57 m in a tank. The density of the water in the tank is 959 kg/m3. Find the pressure exerted by the water on the **fish**? g = 9.81 m/s2.
2.	When a **shrimp** is submerged at a depth of 1.85 m in a lake with a density of 1035 kg/m3, calculate the water pressure exerted on it? g=9.81 m/s2
3.	A **shrimp** is at a depth of 1.57 m in a tank. The density of the water in the tank is 959 kg/m3. Find the pressure exerted by the water on the **fish**? g = 9.81 m/s2.
4.	A **crab** is at a depth of 1.58 m in a lake. The density of water in the lake is 957 kg/m3. What is the pressure experienced by the crab caused by the water around it? g = 9.81 m/s2.

Table 23 presents the generated questions for topic pressure in liquids, which use "prawn" as the new object. The results show an overall inconsistency in object substitution, indicating a 0% strict accuracy under exact object preservation criteria. The system replaces the target object with different related terms such as “shrimp,” “crab,” and “fish,” reflecting both adaptive behaviour and variability in substitution patterns.

In particular, the first generated question, the object was substituted with "shrimp" (not in the dataset), reflecting the model's recognition of similarities between related terms. In question 4, "prawn" was replaced with "crab" (found in the dataset), showcasing the model's flexibility in adapting substitutions based on context or saliency. However, in questions 1 and 3, the substitution was inconsistent, with the first part referring to "shrimp" and the second to "fish," resulting in inaccuracies. Overall, these findings demonstrate that while the AQG system exhibits a degree of flexibility in handling unseen objects, its substitution mechanism is not fully stable across all cases.

Testing revealed that 5 out of 112 generated questions (approximately 4.5%) contained errors in object substitution, excluding the topic of Pressure in Liquids. While these substitutions did not entirely invalidate the questions, they introduced inaccuracies that could confuse students and reduce the educational value of the questions. Addressing these issues would improve the AQG system's effectiveness as a teaching tool. In a classroom context, such inconsistencies may distract students from the core physical concepts. Therefore, although the system effectively mitigates the cold-start problem in question generation, human-in-the-loop verification remains necessary to ensure pedagogical clarity and instructional reliability.

Experiment 2: Result Discussion

Tables 24 to 26 show the AQG model's ability to handle multiword targets in masked inputs using its stack of decoders. For instance, in questions 2 and 4 of Table 24, "steel pot" appears in three parts of the sentence. Similarly, "specific heat capacity" is treated as a multiword target. These results demonstrate the model's accuracy in multiword masking to generate coherent questions.

Table 24
Experimental results on using a multiword target in AQG for the topic Specific Heat Capacity

1.	What is the specific heat capacity of a **steel pot** with mass 0.6 kg, when it is heated by an electric heater with a power of 48 W for 93 s and the temperature increases from 19 oC to 42 oC?
2.	A **steel pot** with mass 0.42 kg is heated by an electric heater with a power of 44 W for 89 s. The temperature of the **steel pot** increases from 21 oC to 44 oC. Calculate the specific heat capacity of the **steel pot**?
3.	What is the amount of specific heat capacity required to raise the temperature of a **steel pot** with mass 0.34 kg from 21 oC to 45 oC, by using an electric heater with a power of 51 W for 80 s?
4.	A **steel pot** with mass 0.43 kg is heated by an electric heater with a power of 31 W for 84 s. The temperature of the **steel pot** increases from 19 oC to 45 oC. Calculate the specific heat capacity of the **steel pot**?

Table 25
Experimental results on using a multiword target in AQG for the topic Thin Lens Formula

1.	A plastic bottle of height 12 cm is positioned at 57 cm from a concave lens with a focal length of 17 cm. Determine the size of the image?
2.	A concave lens with a focal length of 18 cm is positioned at 55 cm from a plastic bottle with a height of 14 cm. Figure out the size of the image?
3.	A plastic bottle of height 8 cm is placed at 48 cm from a concave lens with a focal length of 17 cm. Find the size of the image?
4.	A plastic bottle of height 13 cm is positioned at 56 cm from a concave lens with a focal length of 18 cm. Find the size of the image?

Table 26
Experimental results on using a multiword target in AQG for the topic Current and Potential Difference

1.	If an electric charge of 536 C flows through a silver block for 4 minutes, what is the current flow?
2.	If an electric charge of 517 C flows through a silver block for 7 minutes, what is the current flow?
3.	What is the current flow if 506 C of electric charge passes through a silver block for 6 minutes?
4.	What is the current flow if 541 C of electric charge passes through a silver block for 5 minutes?

The model shows high accuracy and coherence in generating questions with both single-word and multiword targets. Despite the added complexity of multiword targets, such as maintaining context, the model handles them effectively. As shown in Tables 24 to 26, the model performs well with multiword targets, demonstrating similar accuracy and coherence as with single-word targets.

In Table 25, question 1 shows the model's ability to combine multiple multiword targets, such as "concave lens" and "size of image," within a coherent structure. The model effectively integrates scientific terminology and variable dependencies while maintaining coherence despite the complexity of multiword targets. Similarly, in Table 26, Question 1 demonstrates the model's versatility in incorporating multiword phrases like "current flow" and contextual objects like "silver block" to create relevant, meaningful questions.

The training dataset provides structured, specific input, such as "force, mass, luggage" or "electric charge, silver block, duration," with clearly defined variables and objects. This consistency helps the model accurately incorporate multiword targets into generated questions. For instance, in the "size of image" scenario, the input variables, namely, focal length, distance, and height, enable the model to generate a precise, coherent question. This well-defined context ensures the model understands the relationships between variables, leading to more accurate and coherent questions.

Despite the complexity of multiword targets, the model maintains high accuracy with structured input context. This is evident in topics like heat, light and optics, and electricity, where terms like "specific heat capacity," "size of image," and "current flow" are effectively used. These examples showcase the model's ability to integrate multiword targets across various physics topics, ensuring questions are relevant, contextually appropriate, and scientifically accurate.

AQG model's ability to handle complex terminology is reinforced by parallel success in STEM question generation. Gao (2023) achieved high performance by applying bidirectional training to the mT5 model on large, general datasets (Math23K, MAWPS; evaluated with 5-fold cross-validation). In contrast, our results demonstrate that standard fine-tuning of the T5-base model on a compact, highly-targeted dataset of 6,000 samples is equally effective for generating high-quality questions within a specific physics curriculum.

This contrast underscores the T5 architecture's versatility, proving that efficient, domain-focused fine-tuning can achieve robust results without the need for massive datasets or highly complex training schemes.

Evaluating Questions Generated by ChatGPT-4 and the AQG System

This section presents a comparative evaluation of the questions generated by ChatGPT-4 and the AQG system based on four criteria: relevance, correctness, usefulness, and variety. As summarised in Table 27, ChatGPT-4 excels in most criteria, especially relevance, correctness, and variety, while the AQG system slightly outperforms in usefulness, indicating its strength in generating practical questions.

Table 27
Performance comparison between ChatGPT-4 and the AQG system in question generation

Method	Relevance	Correctness	Usefulness	Variety
ChatGPT-4	4	3.89	3.65	4
AQG system	3.92	3.79	3.88	2.18

ChatGPT-4 generated more relevant and grammatically correct questions than the AQG system, scoring 4.00 for relevance and 3.89 for correctness. Its broad training and ability to follow precise prompts result in natural, well-structured questions. In contrast, the AQG system, though trained on physics-focused data, occasionally omits input values or uses awkward phrasing, leading to slightly lower scores (3.92 relevance, 3.79 correctness). However, the AQG system outperformed ChatGPT-4 in usefulness (3.88 vs. 3.65), as its questions align more closely with physics curricula. While ChatGPT-4 excels in generating varied formats (score 4.00), the AQG system is more limited (score 2.18), often producing fewer question types per topic. Overall, the AQG system offers better control and domain focus, whereas ChatGPT-4 provides greater linguistic fluency and diversity when guided properly. By using T5 as the primary generator and ChatGPT-4 as a benchmark, this study demonstrates that specialised fine-tuning is more effective than general prompting for producing the consistent, constrained, and curriculum-aligned assessments necessary for secondary-level STEM education.

Human Evaluation Results and Discussion

The human evaluation compared the AQG system with a template-based approach across four criteria: relevance, correctness, usefulness, and variety. While the template-based method scored slightly higher in relevance (4.00 vs. 3.92), correctness (3.80 vs. 3.79), and usefulness (3.91 vs. 3.88), the AQG system showed comparable performance, demonstrating its ability to produce grammatically accurate and generally relevant questions.

The template method's edge is likely due to its structured and predictable format, which closely aligns with expected question patterns. These results are shown in Table 28.

The AQG system's ability to automatically generate three distinct question types for topics such as latent heat and refraction contrasts with the inherently limited scope of template-based methods, which require manual creation of numerous templates for each scenario. This flexibility reflects the AQG system's ability to leverage learned patterns to produce diverse and adaptive questions. Although the AQG system had slightly lower average scores, it showed strong potential in creating effective and varied questions without requiring extensive manual template creation, making it a valuable tool for educators.

In addition to linguistic quality, the pedagogical value of the generated questions was examined in terms of cognitive learning outcomes. The questions target the Understand, Apply, and Analyse levels of Bloom's Taxonomy by requiring students to interpret contexts and perform multi-step reasoning. Specifically, they emphasise the Apply level, where learners must map verbal scenarios to physical laws and execute mathematical procedures. By demanding structured variable extraction from narratives, the system moves beyond rote substitution to support conceptual problem-solving. This ensures alignment with the Malaysian KSSM Physics syllabus and supports learning outcomes like conceptual transfer and structured physics reasoning.

Evidence of the AQG model's ability to move beyond original seed templates is reflected in its capacity to adapt phrasing based on the specific physics parameters provided in the input prompt. Experts observed consistent generation with linguistic variation, resulting in scores of 3.79 for Correctness and 3.88 for Usefulness. While some repetitive phrasing occurred, the questions remain pedagogically meaningful by requiring learners to interpret contexts and apply physical principles rather than relying on rote templates.

The domain-specific dataset, tailored to the Malaysian secondary Physics curriculum, ensures strong national alignment but may limit generalisability to other contexts without fine-tuning. This intentional design optimises accuracy and pedagogical relevance for the target curriculum. As noted by Deroy et al. (2024), such systems require a level of context-awareness that general-purpose models often lack. Consequently, adapting the system to other regions would require retraining on structured datasets to maintain its pedagogical and linguistic effectiveness.

Table 28
Human evaluation results

Method	Relevance	Correctness	Usefulness
AQG system	3.92	3.79	3.88
Template based	4.00	3.80	3.91

The inter-rater reliability of the T5-based AQG system was validated by analysing 180 evaluation units (representing 60 questions across three qualitative metrics: Relevance, Correctness, and Usefulness). With three expert Physics teachers providing 540 total ratings, the analysis yielded an observed agreement of 89%. Using Randolph's Free-Marginal Kappa to account for the high concentration of positive scores, the resulting coefficient of 0.85 indicates almost perfect agreement. This statistical evidence confirms that the high mean scores for question quality reflect strong expert consensus rather than chance agreement among raters.

Automatic Evaluation Results and Discussion

The ROUGE-L scores for all three evaluators exceed 0.82, closely aligning and indicating that the AQG system consistently generates questions similar to the reference ones (Table 29). The small score range (0.82-0.85) reflects strong evaluator agreement, highlighting the reliability and quality of the generated questions.

Table 29
ROUGE-L scores for AQG system questions based on evaluators' outputs compared to golden references

Generated by Evaluator	ROUGE-L Score
AQG System (Evaluator 1)	0.85
AQG System (Evaluator 2)	0.82
AQG System (Evaluator 3)	0.84

The high ROUGE-L scores show that the AQG system generates accurate questions that closely match the ideal phrasing. The slight difference between Evaluator 2 and the others may reflect minor variations in evaluation criteria. However, the consistently high scores across evaluators demonstrate the model's robust performance and alignment with educational or assessment expectations.

The ROUGE-L results demonstrate that the AQG system performs exceptionally well, producing questions closely aligned with the reference questions. Minor variations among evaluators highlight potential areas for refinement in ensuring consistency and alignment with reference questions. Overall, the high scores confirm the model's effectiveness in generating quality educational questions, achieving the primary objective of AQG tasks. However, ROUGE-L primarily measures surface-level textual overlap and may not fully capture conceptual correctness or pedagogical relevance, particularly in a template-driven context. To address this limitation, human evaluation was conducted by three experienced Physics teachers, assessing relevance, correctness, and usefulness value. Thus, ROUGE-L serves as a complementary metric, supported by qualitative evaluation to ensure overall question quality.

CONCLUSION

This study developed and evaluated a web-based Automatic Question Generation (AQG) system for Malaysian secondary-school Physics, leveraging the T5 model to generate contextually appropriate, curriculum-aligned questions within STEM education. By automating the generation of diverse and pedagogically sound assessment items, the system addresses a critical need for efficient educational tools. High ratings from human evaluators on relevance, correctness, and usefulness indicate the system's potential to enhance physics instruction and learning.

Despite these promising results, several limitations should be acknowledged. First, the evaluation was conducted using single-run generation, which does not capture variability in model outputs. Second, the study relied on a limited number of expert reviewers, and while pedagogical validity was established through expert judgement, empirical validation with students was not conducted. Third, although the system demonstrates linguistic variation, some generated questions remain partially constrained by template-like structures, indicating scope for improving contextual and semantic diversity. Finally, the model was trained on a domain-specific dataset aligned with the Malaysian secondary school curriculum, which may limit generalisability to other contexts without further adaptation.

Future work will address these limitations by incorporating multiple sampling runs and statistical testing to improve robustness, as well as expanding evaluation through classroom-based pilot studies to assess learning outcomes and student engagement. Additionally, benchmarking against a wider range of baseline and foundational models, such as Google Gemini and Claude, will provide a more comprehensive comparison across model architectures. Further enhancements may include enabling teacher-controlled difficulty levels, supporting multimodal inputs (e.g., images and text), and improving model performance through domain-specific pre-training and advanced data augmentation techniques. The integration of semantic-based evaluation metrics, such as BERTScore, is also proposed to complement lexical metrics and provide a more comprehensive assessment of question quality.

These developments would significantly enhance the system's versatility and context-awareness, yielding greater benefits for educators and students. Ultimately, this research establishes a strong foundation for the development of intelligent, curriculum-aligned AQG systems in STEM education.

ACKNOWLEDGEMENT

My tokens of appreciation go to the Centre for Graduate Studies, Universiti Malaysia Sarawak, for the advice and support given during the publication of this paper. There are no funding resources provided for the conduct of this research.

REFERENCES

Amidei, J., Piwek, P., & Willis, A. (2018). Evaluation methodologies in automatic question generation 2013-2018. In E. Krahmer, A. Gatt, & M. Goudbeek (Eds.), *Proceedings of the 11th International Conference on Natural Language Generation* (pp. 307-317). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6537

Chew, S. L., & Cerbin, W. J. (2021). The cognitive challenges of effective teaching. *The Journal of Economic Education, 52*(1), 17-40. https://doi.org/10.1080/00220485.2020.1845266

Choy, C. S., Chuan, K. K., Beng, O. H., Md Mustafa, M. K. A., & Ragavan, R. (2019). *Physics form 4*. Sasbadi Sdn. Bhd.

Chuan, K. K., Choy, C. S., Bongkek, N. R., Kasron, J., Md Mustafa, M. K. A., & Chakrabarty, P. K. (2020). *Physics form 5*. Penerbit Bestari Sdn. Bhd.

Cornejo, O., Briola, D., Micucci, D., Ginelli, D., Mariani, L., Santos Parrilla, A., & Juristo, N. (2024). A family of experiments about how developers perceive delayed system response time. *Software Quality Journal, 32*, 567-605. https://doi.org/10.1007/s11219-024-09660-w

Deroy, A., Maity, S., & Sarkar, S. (2024). *MIRROR: A novel approach for the automated evaluation of open-ended question generation* [Preprint]. arXiv.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Erinosho, S. Y. (2013). How do students perceive the difficulty of physics in secondary school? An exploratory study in Nigeria. *International Journal for Cross-Disciplinary Subjects in Education, 3*(Special Issue 3), 1510-1515. https://doi.org/10.20533/ijcdse.2042.6364.2013.0212

Fan, A., Lewis, M., & Dauphin, Y. (2018). Hierarchical neural story generation. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers* (pp. 889-898). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1082

Goldbach, I. R., & Hamza-Lup, F. G. (2017). Survey on e-learning implementation in Eastern Europe: Spotlight on Romania. In L. A. Ludovico & A. M. F. Yousef (Eds.), *Proceedings of the International Conference on Mobile, Hybrid, and Online Learning* (pp. 5-12). IARIA XPS Press.

Gao, K. (2023). Bidirectional training for generating math word problems using a pre-trained model and prompt. In *2023 20th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICCWAMTIP60502.2023.10387053

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. In *Proceedings of the Eighth International Conference on Learning Representations*.

Jauhiainen, J. S., & Garagorry Guerra, A. (2024). *Evaluating students' open-ended written responses with LLMs: Using the RAG framework for GPT-3.5, GPT-4, Claude-3, and Mistral-Large* [Preprint]. arXiv. https://doi.org/10.54364/AAIML.2024.44177

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education, 30*(1), 121-204. https://doi.org/10.1007/s40593-019-00186-y

Kwan, W. C., Wang, H. R., Wang, H. M., & Wong, K. F. (2023). A survey on recent advances and challenges in reinforcement learning methods for task-oriented dialogue policy learning. *Machine Intelligence Research, 20*(3), 318-334. https://doi.org/10.1007/s11633-022-1347-y

Le, N.-T., & Pinkwart, N. (2015). Evaluation of a question generation approach using semantic web for supporting argumentation. *Research and Practice in Technology Enhanced Learning, 10*, Article 3. https://doi.org/10.1007/s41039-015-0003-3

Maity, S., Deroy, A., & Sarkar, S. (2024). *Exploring the capabilities of prompted large language models in educational and assessment applications* [Preprint]. arXiv.

Mishra, S. K., Goel, P., Sharma, A., Jagannatha, A., Jacobs, D. W., & Daumé, H., III. (2020). *Towards automatic generation of questions from long answers* [Preprint]. arXiv.

Molina, I. L., Švábenský, V., Minematsu, T., Chen, L., Okubo, F., & Shimada, A. (2024). Comparison of large language models for generating contextually relevant questions. In I. A. Chounta, P. D. Muñoz-Merino, V. Dimitrova, T. D. Leo, & I. M. Román-González (Eds.), *Towards hybrid human-AI learning technologies: Artificial intelligence supporting human intelligence: 19th European Conference on Technology Enhanced Learning, EC-TEL 2024* (pp. 237-250). Springer.

Parker, M. J., Anderson, C., Stone, C., & Oh, Y. (2024). A large language model approach to educational survey feedback analysis. *International Journal of Artificial Intelligence in Education, 35*(2), 444-481. https://doi.org/10.1007/s40593-024-00414-0

Patil, S., Chavan, L., Mukane, J., Vora, D., & Chitre, V. (2022). State-of-the-art approach to e-learning with cutting-edge NLP transformers: Implementing text summarisation, question and distractor generation, and question answering. *International Journal of Advanced Computer Science and Applications, 13*(1), 445-453. https://doi.org/10.14569/IJACSA.2022.0130155

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1-67.

Roth, R. E. (2013). Interactive maps: What we know and what we need to know. *Journal of Spatial Information Science, 6*, 59-115. https://doi.org/10.5311/JOSIS.2013.6.105

Rus, V., Cai, Z., & Graesser, A. (2008). Question generation: Example of a multi-year evaluation campaign. In V. Rus & A. Graesser (Eds.), *Online proceedings of the First Question Generation Workshop*. National Science Foundation.

Saleh, S. (2014). Malaysian students' motivation towards physics learning. *European Journal of Science and Mathematics Education, 2*(4), 223-232. https://doi.org/10.30935/scimath/9414

Sharma, S., El Asri, L., Schulz, H., & Zumer, J. (2017). *Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation* [Preprint]. arXiv.

Shazeer, N., & Stern, M. (2018). Adafactor: Adaptive learning rates with sublinear memory cost. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning* (*Proceedings of Machine Learning Research*, Vol. 80, pp. 4596-4604). PMLR.

Shi, H., & Wolff, P. (2021). What transformers might know about the physical world: T5 and the origins of knowledge. *Proceedings of the Annual Meeting of the Cognitive Science Society, 43*, 2218-2223.

Tucker, A. (2020). *Text generation with HuggingFace GPT-2* [Kaggle notebook]. Kaggle. https://www.kaggle.com/code/tuckerarrants/text-generation-with-huggingface-gpt2

von Platen, P. (2020, March 1). *How to generate text: Using different decoding methods for language generation with transformers*. Hugging Face. https://huggingface.co/blog/how-to-generate

Wang, W., & Reani, M. (2017). The rise of mobile computing for group decision support systems: A comparative evaluation of mobile and desktop. *International Journal of Human-Computer Studies, 104*, 16-35. https://doi.org/10.1016/j.ijhcs.2017.02.008

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 483-498). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.41

Yao, X., Bouma, G., & Zhang, Y. (2012). Semantics-based question generation and implementation. *Dialogue & Discourse, 3*(2), 11-42. https://doi.org/10.5087/dad.2012.202

Pertanika J. Sci. & Technol. 34 (3): 1525 - 1549 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

D-Xotic: A Deep Learning Improvement for Recognising Different Durian Cultivars Based on Leaves Signature

Aulia Brilliantina[1,2], Tri Agus Siswoyo[1], Yuli Witono[3], and Bayu Taruna Widjaja Putra[3,4]*

[1]Faculty of Agriculture, University of Jember, Jl. Kalimantan No. 37, Tegalboto Campus, Jember, 68121 East Java, Indonesia
[2]Food Industry Technology Study Program, Department of Agricultural Technology, Jember State Polytechnic, Jl. Mastrip, PO Box 164, Jember, 68121 East Java, Indonesia
[3]Faculty of Agricultural Technology, University of Jember, Jl. Kalimantan No. 37, Tegalboto Campus, Jember, 68121 East Java, Indonesia
[4]Centre of Excellence in Artificial Intelligence for Industrial Agriculture, University of Jember, Jl. Kalimantan No. 37, Tegalboto Campus, Jember, 68121 East Java, Indonesia

ABSTRACT

Early identification of durian cultivars at the seedling stage remains challenging because conventional methods rely on fruit characteristics that appear only after several years, often leading to misdistribution and inefficiencies in certification. This study proposes a real-time, non-destructive deep learning framework for intra-species durian cultivar identification using leaf images. A dataset of 1,788 leaf images from five cultivars (Bawor, Kani, Monthong, Musang King, and Petruk) was collected under real-world conditions. Unlike existing studies that formulate plant varietal identification solely as an image classification task, this work introduces a detection-based formulation that explicitly models cultivar-specific leaf features at the object level. This formulation enables the model to localise and learn fine-grained morphological differences that are often overlooked in global classification approaches. To validate this approach, MobileNetV2, Xception, and YOLOv11 were systematically evaluated using K-Fold Cross Validation under the same experimental setting. The key contributions of this study are: (1) the formulation of intra-species plant cultivar identification as an object detection problem rather than a pure classification task, (2) the development of a practical, real-time framework for early-stage durian cultivar identification using leaf images in unconstrained environments, and (3) a comparative analysis demonstrating the effectiveness of detection-

ARTICLE INFO

Article history:
Received: 20 August 2025
Accepted: 03 April 2026
Published: 19 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.11

E-mail addresses:
aulia_b@polije.ac.id (Aulia Brilliantina)
triagus.faperta@unej.ac.id (Tri Agus Siswoyo)
yuliwitono.ftp@unej.ac.id (Yuli Witono)
bayu@unej.ac.id (Bayu Taruna Widjaja Putra)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

based models in capturing subtle inter-varietal differences. Experimental results show that YOLOv11 achieves over 99% validation accuracy, with precision and recall above 0.95, while maintaining an inference time of 0.2 ms per image. These findings highlight the potential of detection-based approaches for AI-assisted seedling selection in horticulture.

Keywords: Cultivar identification, deep learning, durian, leaf image, YOLO

INTRODUCTION

Durian (*Durio zibethinus*) is one of the most economically important fruit crops in Southeast Asia, serving not only as a cultural and culinary icon but also as a high-value commodity in both domestic and international markets. The rapid growth of global demand, particularly from China, has significantly increased durian production in major producing countries such as Thailand, Malaysia, and Indonesia (May et al., 2023). The commercial value of durian is strongly determined by varietal characteristics, including flavour, aroma, texture, and postharvest quality. Therefore, accurate identification of durian cultivars is essential for farmers, nurseries, certification bodies, and other stakeholders to ensure product authenticity and supply chain traceability.

In agricultural systems, varietal identification plays a critical role in farm management, quality assurance, and plant breeders' rights. Misidentification can result in inappropriate agronomic practices, economic losses, and market disputes. Moreover, cultivar-specific traits influence nutrient management, disease resistance, and harvesting time, making accurate identification essential for long-term planning in sustainable durian production (Magandhi et al., 2024).

Conventional and molecular approaches for durian varietal identification have been widely applied. Traditional methods rely primarily on morphological traits, particularly fruit characteristics, which only become observable after several years. Although leaf-based identification during the vegetative stage is possible, it is highly subjective and sensitive to environmental variations. Molecular techniques such as DNA analysis offer high accuracy but are costly, time-consuming, and require specialised laboratory infrastructure, limiting their applicability for routine field use. These limitations highlight a fundamental challenge in early-stage durian varietal identification. As a result, there is currently no rapid, non-destructive, and field-deployable method capable of accurately identifying durian cultivars at the seedling stage under real-world conditions.

Recent advances in Artificial Intelligence (AI), particularly deep learning and computer vision, have significantly improved plant identification using image-based approaches. Convolutional Neural Networks (CNNs) have demonstrated superior performance compared to traditional machine learning methods such as Support Vector Machines (SVMs),

Random Forests (RFs), and K-Nearest Neighbours (KNN), which rely on manually engineered features sensitive to variations in lighting, background, and perspective (Kamilaris & Prenafeta-Boldu, 2018; Wäldchen et al., 2018). CNN-based models can automatically learn hierarchical visual features from raw images, enabling more robust recognition of morphological differences such as leaf shape and venation patterns (LeCun et al., 2015).

Advanced CNN architectures and real-time object detection frameworks have further expanded the applicability of image-based plant analysis. Architectures such as Xception and MobileNetV2 improve classification accuracy while maintaining computational efficiency, making them suitable for real-world applications (Chollet, 2017; Sandler et al., 2018). These models have been successfully applied in crop identification tasks under field conditions (Too et al., 2019; Ullah et al., 2023). In parallel, object detection frameworks such as YOLO (You Only Look Once) enable simultaneous localisation and classification of multiple objects within a single image, which is particularly advantageous in complex agricultural environments. Recent developments in YOLO architectures, including YOLOv11, have improved detection accuracy and efficiency through optimised backbone structures and loss functions (Ahmad et al., 2023; Bochkovskiy et al., 2020). However, most existing studies predominantly formulate plant identification as a single-image classification problem under controlled conditions, where each image contains a single object with minimal background variation. This limitation reduces their applicability in real-world agricultural environments, where multiple leaves may appear simultaneously with occlusion, varying illumination, and complex backgrounds. Recent studies on fruit and leaf-based deep learning, including papaya leaf disease classification, fruit bruise detection, pear leaf disease detection, and MobileNetV2-based fruit classification, further demonstrate the effectiveness of deep learning models in agricultural applications.

Despite these advancements, most existing studies focus on inter-species classification or disease detection, with limited attention to fine-grained intra-species varietal identification under real-world conditions. Several works have applied deep learning to leaf disease detection in fruit crops or fruit-level quality assessment, often using controlled datasets and experimental settings. While these studies demonstrate the effectiveness of deep learning, they provide limited insights into fine-grained intra-species classification, particularly for perennial fruit crops such as durian during the vegetative stage. Furthermore, many prior approaches formulate plant identification solely as an image classification problem, without exploring detection-based strategies that may better capture subtle visual differences under real-world conditions (Mohanty et al., 2016; Singh et al., 2016).

In practice, varietal management in Indonesia and other durian-producing countries remains largely dependent on manual and subjective identification, increasing the risk of mislabelling during seedling distribution and certification processes (Horiuchi et al., 2023). Existing image-based approaches are often developed under controlled conditions,

limiting their robustness in real-world environments characterised by variable lighting, complex backgrounds, and occlusions. Consequently, their applicability in practical durian cultivation systems remains constrained.

Therefore, the core problem addressed in this study is the absence of a robust, accurate, and field-applicable method for early-stage durian cultivar identification under real-world conditions. To address this challenge, this study proposes a deep learning-based framework that integrates lightweight CNN classification models (MobileNetV2 and Xception) with a real-time object detection model (YOLOv11) for intra-species durian cultivar identification using leaf images. The main contributions of this study are threefold: (1) the development of a real-world durian leaf image dataset collected under field conditions; (2) the introduction of a detection-based formulation for fine-grained intra-species cultivar identification, moving beyond conventional classification-based approaches; and (3) a systematic comparison between classification-based and detection-based deep learning methods, demonstrating that detection models can more effectively capture subtle inter-varietal differences for early-stage identification.

MATERIALS AND METHODS

Study Location

Jember Regency, East Java, Indonesia, is a major centre for durian cultivation and development. The region has favourable agroclimatic conditions, including fertile soil, annual rainfall ranging from 1,800-3,000 mm, and temperatures between 24-32 °C, which support the development of durian cultivation and enable the growth of both local cultivars (Bawor and Petruk) and commercial cultivars (Musang King, Monthong, and Kani). This study selected Jember Regency as the study area because it represents an active durian seed production and certification region, providing a realistic context for evaluating a leaf image–based cultivar identification methodology.

Materials and Workflow Overview

This section presents the methodological workflow adopted in this study, encompassing image acquisition, expert-assisted labelling, data preprocessing, model training, and evaluation, as shown in Figure 1. Leaf images were collected under natural lighting conditions, while manual labelling was verified by agricultural experts to ensure annotation reliability. The dataset consisted of leaf images from five durian cultivars (Monthong, Kani, Petruk, Bawor, and Musang King) collected in Jember Regency.

All experiments were conducted on a workstation equipped with an NVIDIA RTX 3090 GPU (24 GB VRAM), an Intel-based CPU (10 cores, 3.7 GHz), 64 GB RAM, and running Ubuntu 20.04 LTS. Model training was performed using Python 3.9 with TensorFlow/Keras and PyTorch frameworks. For inference and deployment-level

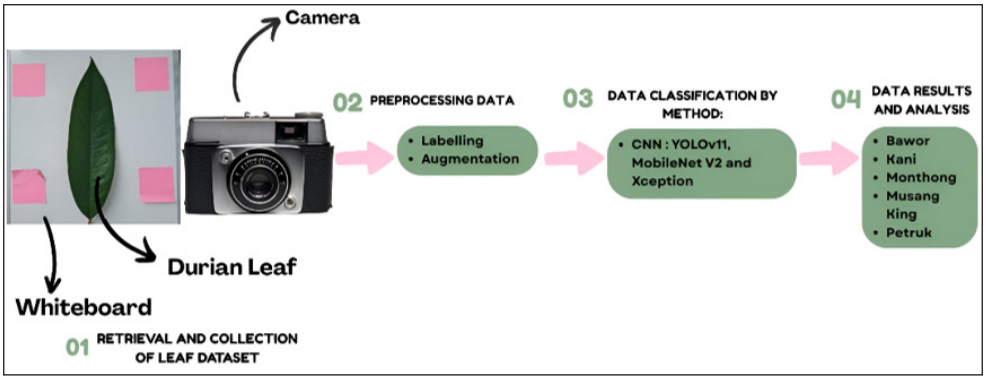


Figure 1. Research stages

evaluation, models were tested under real-time conditions using the Ultralytics YOLOv11 inference pipeline, reflecting practical field deployment scenarios.

Data Collection Procedures

The data collection process began with the identification and validation of five durian cultivars (Bawor, Kani, Monthong, Musang King, and Petruk) conducted in collaboration with local farmers and agricultural experts, as shown in Figure 2. This validation step was included as a methodological control to reduce labelling uncertainty in intra-species classification tasks.

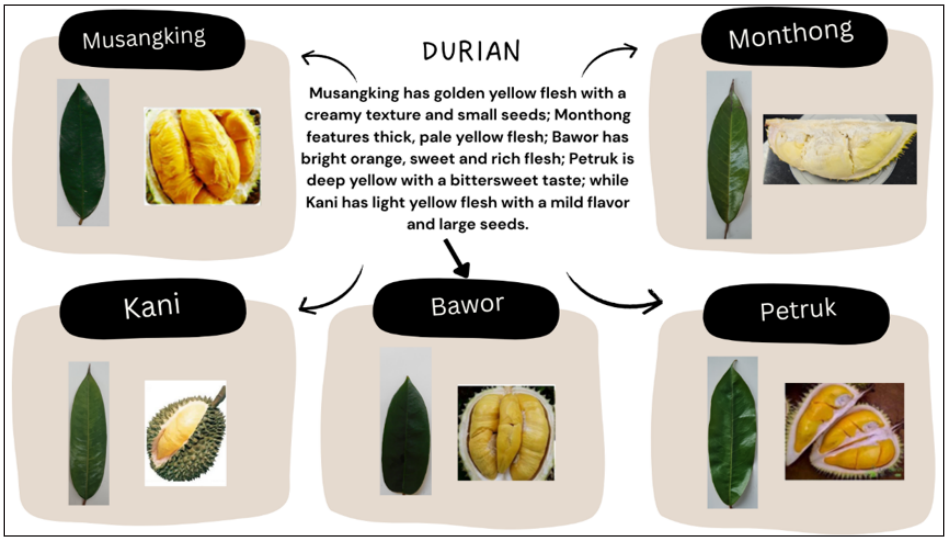


Figure 2. Durian leaf cultivars in Jember Regency (Musang King, Monthong, Kani, Bawor, and Petruk)

A total of 1,788 leaf images were captured using a Nikon Coolpix B500 digital camera (16 MP) at distances ranging from 15 to 50 cm. To minimise background interference and simplify preprocessing, leaves were positioned against a white background with pink markers. All images were collected under natural lighting conditions to reflect practical field acquisition scenarios, rather than laboratory-controlled environments. The dataset distribution across cultivars consisted of Bawor (293), Kani (600), Monthong (443), Musang King (163), and Petruk (289) images.

Data Augmentation Techniques

Data augmentation was employed to enhance model generalisation and simulate real-world variability, as commonly recommended in image-based deep learning studies (Meethongjan et al., 2022; Sengupta et al., 2023). Before augmentation, all images were standardised in terms of resolution and colour space, followed by contrast normalisation to mitigate acquisition-related bias.

Augmentation operations, including rotation, flipping, brightness modification, scaling, cropping, translation, and Gaussian blur, were applied exclusively to the training dataset, with transformation probabilities ranging from 0.3 to 0.5 (Zhang et al., 2024). This strategy was adopted to expose the model to diverse visual conditions rather than to artificially inflate the dataset size.

The dataset was divided into training (80%), validation (10%), and testing (10%) subsets (Table 1). To address class imbalance, the number of augmented samples per class was adjusted to obtain a balanced training set of approximately 600 images per durian cultivar. The validation and testing subsets consisted solely of original (non-augmented) images to prevent data leakage and ensure an unbiased performance evaluation.

Justification for Object Detection-based Approach

While durian varietal identification can be formulated as an image classification problem, practical field scenarios often involve multiple leaves appearing simultaneously within a single frame, partial occlusion, and complex backgrounds.

Table 1
Dataset statistics and class distribution before and after augmentation

Durian Cultivar	Original Images	Augmented Images (Training Only)	Total Training	Validation	Testing
Bawor	293	366	600	29	30
Kani	600	120	600	60	60
Monthong	443	246	600	44	45
Musang King	163	470	600	16	17
Petruk	289	369	600	29	29
Total	1,788	1,571	3,000	178	181

Object detection models, such as YOLOv11, offer the ability to localise and classify multiple leaf instances in real time, which is not achievable using pure image-level classification approaches. This capability is particularly relevant for deployment in plantation-scale monitoring, mobile-based scouting, and automated nursery inspection systems. Therefore, object detection was investigated not as a replacement for classification, but as a scalable and deployment-oriented alternative suitable for real-world agricultural environments.

Deep Learning Architectures

YOLOv11: Object Detection Framework

YOLOv11 is a single-stage object detection framework that formulates object detection as a unified regression problem, simultaneously predicting bounding box coordinates and class probabilities (Redmon et al., 2016). In this study, YOLOv11 was employed primarily to assess its suitability as an object detection module for identifying durian leaves within complex image scenes, rather than to emphasise detection performance outcomes shown in Figure 3.

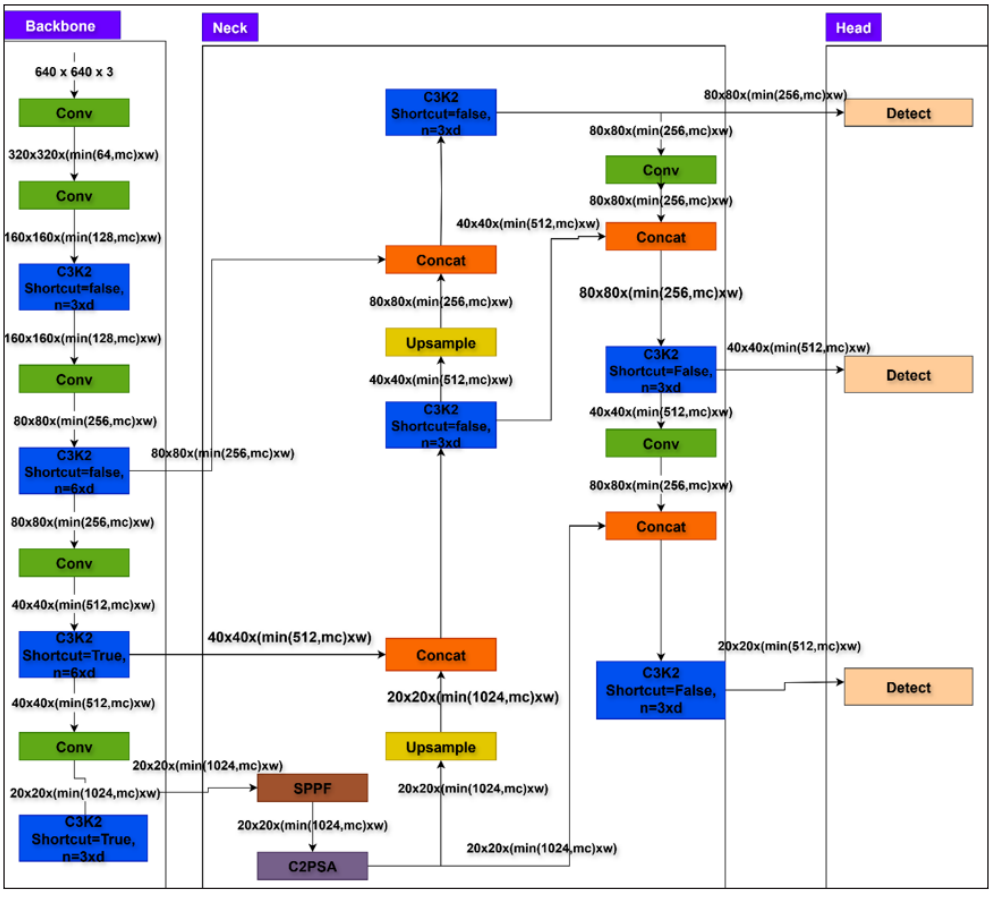


Figure 3. YOLOv11 architecture

This study does not propose architectural modifications but instead focuses on evaluating the applicability of an existing detection framework in real-world agricultural scenarios.

YOLOv11 incorporates a convolutional neural network (CNN) backbone composed of Conv–BatchNorm–SiLU blocks, as defined in the original YOLOv11 architecture, which have been reported to improve training stability and convergence behaviour (Maaroof & Bouhlel, 2024). Spatial pyramid pooling is included as part of the standard YOLOv11 design to preserve spatial information in lower-resolution feature maps, as documented in prior object detection studies. No architectural modifications were introduced to the YOLOv11 backbone in this study.

The model was implemented using the Ultralytics YOLOv11 framework. Training parameters were set to an initial learning rate of 0.001, a batch size of 32, and the Stochastic Gradient Descent (SGD) optimiser (Bottou, 2012). These parameters were kept constant across all experiments to ensure methodological consistency rather than performance optimisation.

MobileNetV2: Lightweight CNN for Resource-constrained Applications

MobileNetV2 is a lightweight convolutional neural network architecture that employs inverted residual blocks and linear bottlenecks to reduce computational complexity while preserving representational capacity (Sandler et al., 2018). In this study, MobileNetV2 was selected to evaluate the feasibility of deploying durian varietal classification models on resource-constrained platforms, such as mobile or edge devices, shown in Figure 4. This architecture has been widely adopted in agricultural image classification tasks, including plant species recognition and fruit quality assessment, due to its favourable trade-off between classification accuracy and computational efficiency (Kazi, 2024; Kusumaningtyas et al., 2024).

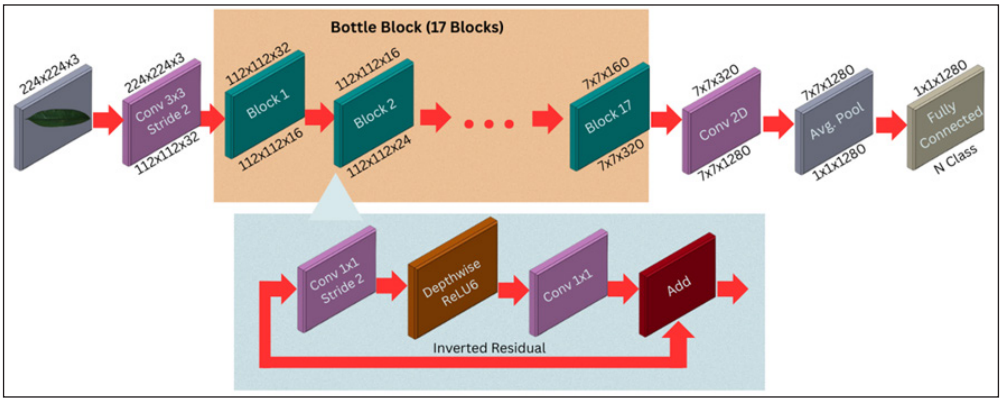


Figure 4. MobileNet V2 architecture

Xception: Deep Feature Extraction Model

Xception is a deep convolutional neural network architecture that replaces standard convolution operations with depthwise separable convolutions, enabling independent learning of spatial and channel-wise features and thereby improving computational efficiency and representational power (Chollet, 2017). In this study, Xception was included as a high-capacity feature extraction model to serve as a comparative benchmark for capturing fine-grained morphological differences among durian leaf cultivars. This architecture has been extensively applied across diverse image analysis domains, including medical imaging, demonstrating its effectiveness in learning complex visual patterns shown in Figure 5 (Maesaroh et al., 2024; Tejada et al., 2024; Zeynali et al., 2024).

Both MobileNetV2 and Xception models were initialised with ImageNet pre-trained weights. Transfer learning was applied by freezing the early feature extraction layers and fine-tuning the task-specific classification layers, following established best practices for small to medium-sized datasets (Yosinski et al., 2014).

Experimental Setup

All leaf images were resized to 640 × 640 pixels to ensure consistent input dimensions across both classification and detection models. Model training was conducted on a workstation equipped with an NVIDIA RTX 3090 GPU, using TensorFlow/Keras and PyTorch frameworks, with training progress monitored via TensorBoard.

Training parameters included an initial learning rate of 0.001, a batch size of 32, and a maximum of 100 epochs. Early stopping was employed as an experimental control strategy, whereby training was terminated if the validation loss failed to improve for 10 consecutive epochs, to mitigate overfitting and enhance reproducibility. Under selected experimental conditions, leaf segmentation was applied as a preprocessing step to reduce background interference, thereby enabling the models to focus on salient leaf morphological features.

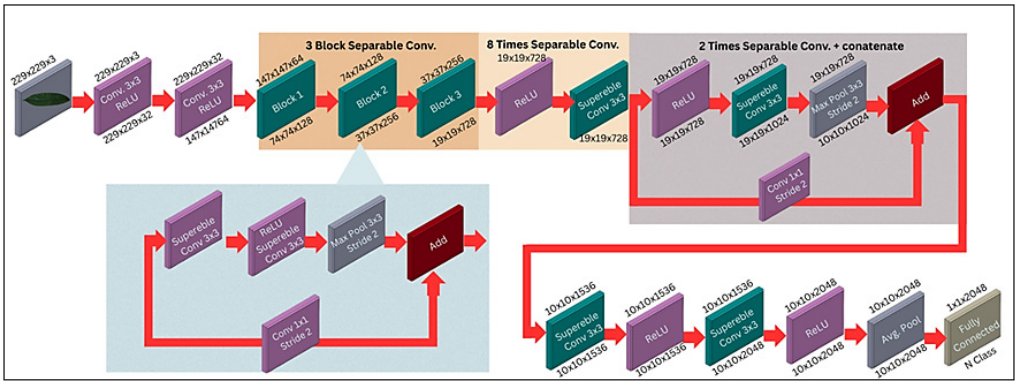


Figure 5. Xception architecture

Evaluation Metrics and Validation Strategy

Model performance was evaluated using a confusion matrix-based approach, from which accuracy, precision, recall, and F1-score were computed (Equations 1 to 4). These metrics were selected to provide a balanced assessment of classification behaviour rather than to emphasise peak performance values.

$$\text{Accuracy} = \frac{TP + TN}{TP+TN+FP+FN} \quad [1]$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad [2]$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad [3]$$

$$\text{F1} - \text{Score} = 2 \times \frac{Precision\ \ x\ Recall}{Precision\ +Recall} \quad [4]$$

Description:

TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

To reduce partitioning bias and improve generalisation reliability, K-Fold Cross Validation (K-Fold CV) was employed, which has been widely adopted in CNN-based classification studies with limited datasets (Gorriz et al., 2024; Hasan et al., 2025; Uzer, 2025). The average accuracy formulation of K-Fold Cross Validation was used to compute the final classification accuracy, Equation 5. For object detection tasks, mean Average Precision (mAP) was used to summarise the precision-recall relationship across all durian cultivars, following standard object detection evaluation practices. The average accuracy formulation of K-Fold Cross Validation:

$$\text{Accuracy_Kfold} = 1/\text{K} + \sum_{n=1}^{K} Accuracy \quad [5]$$

Performance of YOLO-based detectors is typically reported using mean Average Precision (mAP), which summarises the precision-recall trade-off across all classes. Let C be the number of classes. mAP is the mean of the per-class Average Precision (AP), as expressed in Equation 6:

$$\text{mAP} = 1/\text{N} \sum_{i=1}^{N} AccuracyAPi \quad [6]$$

where N is the number of object classes, and APi is the area under the precision-recall curve for class i.

RESULTS AND DISCUSSION

Model Performance Comparison

MobileNetV2 Performance: Accuracy and Training Characteristics

MobileNetV2 is designed to achieve high classification efficiency while maintaining low computational complexity. In this study, MobileNetV2 was employed to classify durian leaf cultivars, achieving classification accuracies ranging from 90% to 95%. The use of depthwise separable convolutions enables a substantial reduction in model parameters and computational cost without compromising classification performance.

Owing to its lightweight architecture, MobileNetV2 is particularly suitable for deployment on resource-constrained devices, such as smartphones and embedded systems. During training, the model effectively learned discriminative patterns from the durian leaf dataset; however, multiple training iterations were required to achieve stable accuracy, consistent with observations reported in previous studies (Huang, Sarabi, & Ragab, 2024; Rahman et al., 2023). Using an input image resolution of 224 × 224 pixels, MobileNetV2 successfully extracted salient leaf features, allowing reliable differentiation among durian cultivars. The achieved loss value of 0.00024 indicates a very low classification error, reflecting strong convergence during training. In addition to its classification performance, the relatively small model size of MobileNetV2 offers advantages in storage efficiency and reduced energy consumption compared to conventional convolutional neural networks.

Nevertheless, optimal performance with MobileNetV2 is highly dependent on the availability of high-quality training data and effective data augmentation strategies. This dependency is consistent with prior work on plant disease detection using MobileNetV2, where enhanced feature extraction through appropriate preprocessing and augmentation yielded classification accuracies of up to 99.71% (Chen et al., 2021). Overall, the results demonstrate that MobileNetV2 provides a reliable and computationally efficient solution for durian leaf classification tasks.

Xception Performance: Accuracy and Training Characteristics

The Xception model achieved high classification performance in durian leaf identification, with accuracy values ranging from 92% to 97%. Its deep network architecture enables the extraction of fine-grained and discriminative visual features, thereby contributing to improved classification accuracy. Similar performance gains have been reported in other domains; for example, Yoon (2025) employed the Xception architecture for brain tumour classification and achieved an accuracy of 99.09%, demonstrating its capability to model complex image patterns.

During training, the classification accuracy of Xception increased steadily with the number of epochs, indicating stable convergence and effective feature learning. The model exhibited strong capability in capturing morphological characteristics of durian leaves, making it well-suited for image classification tasks that require detailed shape and texture recognition.

However, due to its deeper and more complex architecture, Xception requires greater computational resources during both training and inference compared to lightweight models. This trade-off between classification accuracy and computational cost highlights the importance of model selection based on deployment constraints and application requirements.

YOLOv11 Performance: mAP50, mAP50-95, and Inference Speed

YOLOv11 is a single-stage object detection architecture specifically designed for real-time object localisation and classification in complex image scenes. It enables simultaneous object identification and bounding box regression within a unified framework, resulting in high detection accuracy and fast inference performance (Nimma et al., 2025; Sulzbach et al., 2025).

In this study, YOLOv11 was employed to detect durian leaves in natural image environments. The model achieved a mAP50 value approaching 100%, while mAP50–95 values ranged from 97.6% to 99.9%, indicating highly consistent detection performance across varying intersection-over-union thresholds. To the best of our knowledge, the application of the YOLOv11 architecture for durian leaf detection has not been extensively explored in prior studies.

Previous research has demonstrated the effectiveness of earlier YOLO variants in plant-related object detection tasks. For instance, Abid et al. (2024) reported a mean Average Precision (mAP) of 98% using the YOLOv8 architecture for plant leaf disease detection in Bangladesh, highlighting the suitability of YOLO-based frameworks for high-precision agricultural image analysis.

In addition to its detection accuracy, YOLOv11 exhibited a fast inference speed of approximately 0.2 ms per image, underscoring its capability for real-time deployment. The one-stage detection paradigm allows YOLOv11 to identify multiple durian leaf cultivars within a single image without requiring separate processing steps. These characteristics make YOLOv11 particularly well-suited for real-time applications, such as automated durian orchard monitoring and camera-based inspection systems.

Effect of Data Augmentation on Model Performance

Data augmentation, including rotation, horizontal flip and brightness, was used in training the model to increase robustness. Integration of these augmented techniques achieved high

validation accuracy: 99.02% and low validation loss: 0.0341, demonstrating excellent generalisation performance on the unseen dataset. Previous works also find that models trained only with data augmentation are not as generalisable since the diversity of input images is small. The high accuracy of validation obtained in this study indicates that data augmentation is an essential measure to overcome overfitting and be more robust to leaves with variation in orientation and illumination conditions.

Training and Validation Curve Analysis

The accuracy and loss curves of training and validation are shown in Figure 6. The curves show an orderly convergence behaviour in that both training and validation accuracy follow each other closely, with a steady decline in both training and validation loss without clear signs of overfitting.

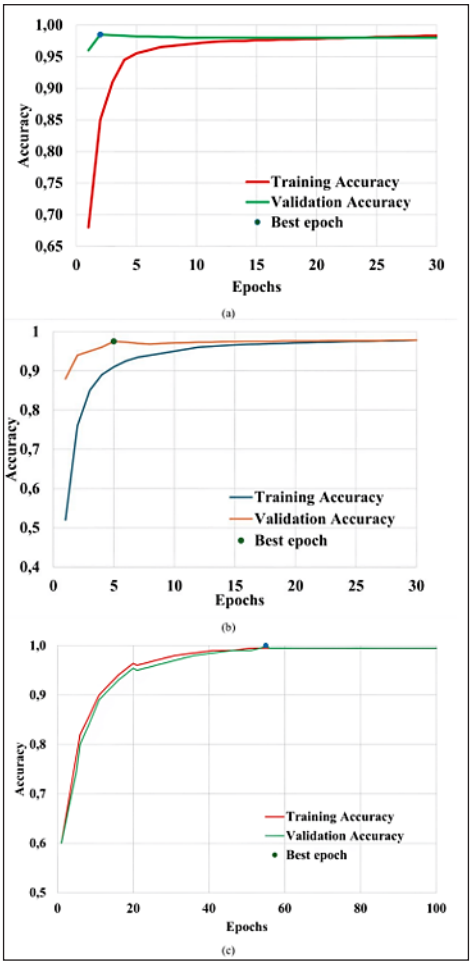


Figure 6. Training and validation accuracy results using the architectures (a) MobileNetV2; (b) Xception; and (c) YOLOv11

Such behaviour is indicative of the model not being overfit. The lack of a wide margin between the training and validation curves suggests that model generalisation is strong, a conclusion supported by the results of K-fold cross-validation analysis.

Comparative Analysis of the Three Architectures in Terms of Accuracy and Speed

The comparative analysis reveals clear performance differences among YOLOv11, MobileNetV2, and Xception for durian leaf detection and classification. Among the evaluated models, YOLOv11 demonstrated the strongest overall performance, achieving a mAP50 value close to 100% across all durian cultivars (Bawor, Kani, Monthong, Musang King, and Petruk), along with consistently high mAP50-95 values ranging from 97.6% to 99.9%. In addition, YOLOv11 exhibited a rapid inference speed of approximately 0.2 ms per image, highlighting its suitability for real-time object detection applications.

In comparison, MobileNetV2 achieved a classification accuracy of approximately 90–95%, which, while competitive, remained lower than that of YOLOv11. The MobileNetV2 architecture is optimised for computational efficiency using depthwise separable convolutions, making it lightweight and well-suited for deployment on resource-constrained platforms such as smartphones and edge devices. However, a key limitation of MobileNetV2 is its relatively slower inference speed compared to YOLOv11, as the model is designed primarily for image classification rather than real-time object detection.

Xception demonstrated higher classification accuracy than MobileNetV2, with performance ranging from 92% to 97%, but at the expense of increased computational complexity. Its deeper architecture and more intricate depthwise separable convolutional operations enhance feature representation, particularly for fine-grained morphological patterns. Nevertheless, this increased complexity results in slower inference compared to both MobileNetV2 and YOLOv11. Consequently, although Xception outperforms MobileNetV2 in classification accuracy, it remains less efficient than YOLOv11 in terms of speed and overall deployment feasibility.

These performance disparities can be primarily attributed to architectural design differences. YOLOv11 is explicitly engineered for object detection, employing a one-stage detection paradigm that performs object classification and bounding box regression simultaneously, thereby enabling high detection accuracy and fast inference. In contrast, MobileNetV2 and Xception are designed for image classification tasks, focusing on categorising a single dominant object per image rather than detecting multiple objects within complex scenes.

The YOLOv11 model used in this study comprises approximately 2.59 million parameters with a computational cost of 6.4 GFLOPs. This relatively low model complexity, compared to deeper convolutional architectures, contributes to its fast inference speed while

maintaining high detection accuracy. These characteristics further underscore the suitability of YOLOv11 for real-time deployment on resource-constrained platforms.

Overall, YOLOv11 emerges as the most effective solution for durian leaf detection due to its superior accuracy and significantly faster inference speed compared to MobileNetV2 and Xception. MobileNetV2 remains a viable alternative when strict hardware constraints are present, while Xception is preferable in scenarios where classification accuracy is prioritised over inference speed. A quantitative summary of the comparative performance of MobileNetV2, Xception, and YOLOv11 is presented in Table 2.

Table 2
Performance comparison of MobileNetV2, Xception, and YOLOv11

Model	Accuracy	Speed	Excellence
MobileNet V2	90-95%	Slower than YOLOv11	Lightweight, suitable for mobile devices
Xception	92-97%	Slower than MobileNet V2 & YOLOv11	High accuracy, but computationally heavier
YOLOv11	97-100% (mAP50-95)	0.2ms per image	Real-time detection, extremely fast and accurate

K-Fold Cross Validation Results

Average Validation Accuracy and Loss Achieved Through K-Fold Cross Validation

The model was assessed through the K-Fold Cross-Validation (CV) method, which can provide more robust estimates of accuracy and mitigate arbitrariness from data partitioning (Refaeilzadeh et al., 2009). The accuracy for the validation results had an average performance of 99.02% with a loss value of 0.0341. This indicates that the model achieves high precision in distinguishing data into five classes, including bawor, kani, monthong, musang king and petruk. Table 3 shows the detailed classification report with precision, recall, and F1-score of each class. The F1-score values for all classes are in the range of 0.95~1.00, suggesting a strong balance between precision and recall (Sokolova & Lapalme, 2009).

Table 3
Precision, recall, and f1-score for each durian cultivar across folds

	Precision	Recall	F1-Score	Support
Bawor	1,00	1,00	1,00	234
Kani	1,00	1,00	1,00	480
Monthong	1,00	0,99	1,00	354
Musang King	0,99	0,92	0,95	131
Petruk	0,95	1,00	0,98	231

The bawor, kani and petruk classes achieved 100% of precision and recall, which means that the classifier identified all samples from these categories properly. On the other hand, there was a significant decrease in recall of the monthong and musang king classes, with both values equal to 0.99 and 0.95, respectively. It is important to mention that for musang king, 11 of the 131 samples were misclassified as petruk by the model. This misclassification indicates that musang king and petruk present the same morphology that could lead to confusion in sample interpretation.

Discussion of the Performance Consistency Across Different Cultivars

The evaluation results indicate that the proposed model can predict most cultivars with a relatively small margin of error. The Bawor, Kani, and Monthong cultivars exhibit near-optimal performance across all evaluation metrics. In contrast, the Musang king cultivar shows a lower recall value compared to the other cultivars, indicating a higher likelihood of misclassification, where Musang king leaves are occasionally predicted as other types.

The Petruk cultivar achieved a recall value of 1.00, demonstrating that the model successfully identified all Petruk samples. However, its precision value was slightly lower, at 0.95, suggesting that a small number of samples from other cultivars were incorrectly classified as Petruk.

Overall, the model demonstrates consistent and robust performance across all folds and cultivars, reflecting a well-balanced dataset and a stable classification model. Musang king was the only cultivar that exhibited a higher tendency for misclassification. This behaviour is likely attributable to similarities in morphological characteristics between Musang king and the Monthong or Petruk cultivars, particularly under certain lighting conditions and image orientations.

Confusion Matrix Analysis

Visual Representation of Prediction Accuracy and Errors

Figure 7 presents the confusion matrix, which clearly illustrates that most of the model's predictions lie along the main diagonal, indicating a high level of classification accuracy. Only a limited number of misclassifications are observed, with most errors occurring in the Musang king class. These off-diagonal entries indicate instances where the predicted labels do not correspond to the ground truth. Such fine-grained classification errors are common in image-based recognition tasks when morphological cues are subtle or visually overlapping (Krause et al., 2015; Wang et al., 2020).

Identification of Specific Misclassification Patterns, Particularly for the Musang King Cultivar

Further analysis of the confusion matrix, as shown in Figure 7, confirms that misclassification errors are predominantly associated with the Musang king class.

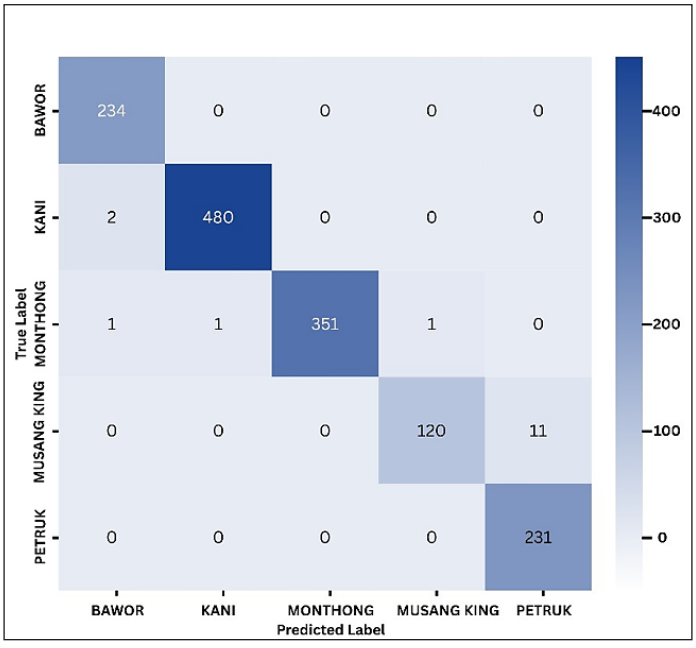


Figure 7. Confusion matrix

As noted by Chicco and Jurman (2020), confusion matrix analysis is essential for identifying class-specific error patterns. The model demonstrates stable and consistent performance for the remaining classes, achieving prediction accuracies exceeding 99%, with no notable cross-class prediction errors. The Musang king cultivar exhibits the highest frequency of misclassification, with some samples being incorrectly classified as Petruk. This confusion is likely attributable to similarities in leaf texture and edge morphology between these two cultivars, particularly when images are captured under comparable growth conditions, lighting environments, and viewing angles. Similar observations have been reported by Zhang and Wu (2012), who noted that cultivars within the same species often share highly similar visual characteristics under analogous imaging conditions.

Interpretation of High Prediction Accuracy for Other Cultivars

The consistently high classification performance observed for the Bawor, Kani, and Petruk cultivars highlights the model's strong capability to distinguish leaf images with clearly differentiated morphological traits. The pronounced intra-class consistency and inter-class separability of these cultivars provide sufficient visual cues, such as vein structures, leaf margins, and blade proportions, enabling the model to extract discriminative features effectively. This reliable differentiation underscores the robustness of the proposed model's feature-learning mechanism, particularly under controlled image acquisition conditions. These findings are consistent with Barré et al. (2017), who demonstrated that high

classification accuracy in plant identification tasks can be achieved when morphological traits are both consistent and visually distinct. Similarly, Too et al. (2019) reported that convolutional neural networks perform exceptionally well in fine-grained classification tasks when trained on datasets with clear inter-class boundaries.

Insights into Visual Similarities Contributing to Misclassifications

The misclassification patterns observed in the Musang king class further emphasise the presence of visual morphological similarities among certain cultivars, a well-documented challenge in plant image classification (Zhang & Wu, 2012). These limitations may be addressed through the incorporation of additional training samples, enhanced data augmentation strategies, or the exploration of more discriminative feature extraction techniques. Despite these challenges, the results indicate that the proposed model architecture exhibits strong generalisation capability and performs effectively across all five durian cultivars. Nevertheless, improving classification performance for the Musang king class remains an important direction for future work.

ROC and AUC Analysis

To further assess the robustness of the classification model, Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) analyses were conducted using a one-vs-rest strategy for each durian cultivar. The model achieved AUC values exceeding 0.98 across all classes, indicating excellent separability among the cultivars. These high AUC scores confirm the model's strong discriminative ability, even in scenarios involving visually similar leaf structures.

Visual Explainability Using Grad-CAM

Grad-CAM visualisation was employed to examine the regions contributing most significantly to the model's classification decisions. The resulting heatmaps indicate that the model primarily focuses on discriminative leaf regions, including vein patterns, margins, and texture characteristics. This behaviour suggests that the model learns meaningful botanical features rather than relying on background artefacts, thereby enhancing both interpretability and reliability of the classification outcomes.

Discussion of Findings and Comparison with Existing Literature

Architectural Factors Behind the Superior Performance of YOLOv11

The superior real-time detection and classification performance of YOLOv11 observed in this study can be primarily attributed to its one-stage detection architecture, which enables simultaneous feature extraction, localisation, and classification within a single forward pass.

Unlike classification-based models that operate on pre-cropped or single-object images, YOLOv11 is inherently designed to handle spatial variability and background noise, conditions that commonly occur in field-acquired agricultural images.

Although MobileNetV2, Xception, and YOLOv11 all achieved high classification accuracy exceeding 99% shown in Figure 6, their training behaviour and deployment characteristics differed substantially. MobileNetV2 converged rapidly, reaching its peak accuracy within only two epochs, which reflects its lightweight design and reduced parameter space optimised for computational efficiency (Sandler et al., 2018). However, this efficiency comes at the cost of limited representational capacity, restricting its ability to capture highly nuanced intra-species variations.

Xception, by contrast, required more training epochs to reach optimal performance, owing to its deeper architecture and extensive use of depthwise separable convolutions, which enhance its ability to model subtle spatial and textural differences among visually similar leaf cultivars (Chollet, 2017). This architectural depth, referring to the greater number of convolutional layers, explains its improved accuracy relative to MobileNetV2, but also accounts for its increased computational burden and reduced inference speed.

YOLOv11 exhibited a longer convergence process compared to MobileNetV2, which is consistent with observations reported in prior object detection research (Xie et al., 2017). Nevertheless, its ability to jointly learn localisation and classification features provides a decisive advantage in real-world scenarios where object scale, orientation, and illumination vary significantly (Bochkovskiy et al., 2020; Redmon et al., 2016). These findings align with previous studies emphasising the inherent trade-offs between model depth, inference speed, and feature discrimination in agricultural image analysis (Kamilaris & Prenafeta-Boldu, 2018; Saleem et al., 2019). Consequently, MobileNetV2 is most suitable for lightweight edge deployments, Xception for applications prioritising feature richness, and YOLOv11 for robust real-time detection and classification in dynamic agricultural environments.

Technical Rationale for the Effectiveness of the Web-Based YOLOv11 System

The successful implementation of a web-based durian leaf classification system using YOLOv11 further demonstrates the practical significance of the proposed approach, as shown in Figure 8. The system's ability to process images uploaded through a browser interface or captured via a live camera is enabled by YOLOv11's low-latency inference pipeline.

The high confidence scores achieved for cultivars such as Monthong (up to 99.9%) are not solely a function of model accuracy but also reflect the consistency between learned feature representations and the morphological characteristics of the input images. The integration of detection and classification within a single model minimises intermediate processing steps, thereby reducing error propagation and improving prediction stability.

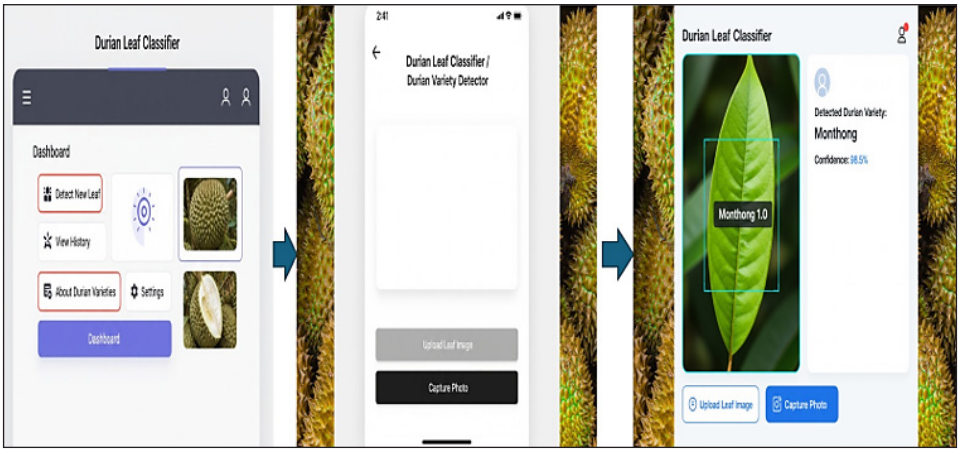


Figure 8. Web-based durian leaf classification

This architectural efficiency explains why YOLOv11 is particularly well-suited for browser-based and real-time agricultural applications, where responsiveness and reliability are critical.

Importance of Real-time Inference Speed in Agricultural Field Applications

Real-time inference speed is a decisive factor in the deployment of computer vision systems in agricultural contexts. The high processing speed demonstrated by YOLOv11 enables immediate feedback during field operations, such as seedling verification, orchard monitoring, and visual quality inspection.

From a practical standpoint, rapid inference reduces dependency on offline laboratory analysis and expert visual assessment, both of which are time-consuming and prone to subjectivity. By enabling near-instantaneous cultivar identification, YOLOv11 facilitates faster decision-making and enhances operational efficiency across multiple stages of the agricultural supply chain. This advantage is particularly relevant in large-scale durian cultivation systems, where timely interventions directly impact productivity and quality control.

Addressing the Challenge of Intra-species Leaf-based Classification

Intra-species classification presents a considerably greater challenge than inter-species classification due to high morphological similarity among cultivars. Durian leaf cultivars often share overlapping features in shape, texture, and venation patterns, which complicates discriminative feature learning.

The high classification performance achieved in this study indicates that deep learning architectures, when combined with effective feature extraction strategies, can capture subtle

intra-class distinctions. This finding extends existing literature that has predominantly focused on interspecific classification and demonstrates the feasibility of applying advanced deep learning models to fine-grained, within-species agricultural identification tasks.

Implications for Durian Seed Certification and Agribusiness Management

The results of this study have important implications for durian seed certification, orchard management, and cultivar control within the broader agribusiness ecosystem. By reducing reliance on manual observation and expert judgment, the proposed approach enhances objectivity and repeatability in cultivar identification.

The integration of real-time detection capabilities with high classification accuracy enables faster verification processes, minimises misidentification risks, and supports scalable deployment across production and distribution stages. As a result, this system has the potential to improve traceability, standardisation, and quality assurance throughout the durian agribusiness value chain.

CONCLUSION

This study addresses the challenge of durian cultivar identification based on leaf morphology, which is inherently difficult due to high intra-species similarity and dependence on expert visual assessment. The results demonstrate that deep learning, particularly the YOLOv11 architecture, effectively overcomes this limitation by learning discriminative spatial features, enabling reliable and non-destructive identification under real-world conditions.

The proposed approach directly addresses the research problem by combining high classification accuracy with real-time inference capability, making it suitable for practical agricultural deployment. To support real-world implementation, a web-based system was developed that allows stakeholders, including nursery operators, orchard managers, and certification authorities, to capture or upload leaf images using standard mobile devices for immediate verification.

Within agricultural workflows, the system can be applied to seedling certification, orchard establishment, and varietal quality control across the durian supply chain. Despite practical constraints such as variable lighting and leaf orientation, the model demonstrates strong robustness and fast inference, enabling efficient field-level decision-making. Overall, this study provides a scalable and application-ready solution that bridges the gap between research and operational needs in durian agribusiness.

ACKNOWLEDGEMENT

We would like to extend our utmost gratitude to the IT Centre of The University of Jember for providing the Supercomputer NVIDIA Station DGX A 100 for this project.

REFERENCES

Abid, S., Jahan, B., Al Mamun, A., Jakir, H., & Hossain Mazumder, S. (2024). Bangladeshi crop leaf disease detection using YOLOv8. *Heliyon, 10*(18), Article e36694. https://doi.org/10.1016/j.heliyon.2024.e36694

Ahmad, A., Saraswat, D., & El Gamal, A. (2023). A survey on using deep learning techniques for plant disease diagnosis and recommendations for the development of appropriate tools. *Smart Agricultural Technology, 3*, Article 100083. https://doi.org/10.1016/j.atech.2022.100083

Barré, P., Stöver, B. C., Müller, K. F., & Steinhage, V. (2017). LeafNet: A computer vision system for automatic plant species identification. *Ecological Informatics, 40*, 50-56. https://doi.org/10.1016/j.ecoinf.2017.05.005

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. *arXiv*. https://doi.org/10.48550/arXiv.2004.10934

Bottou, L. (2012). Stochastic gradient descent tricks. In G. Montavon, G. B. Orr, & K.-R. Müller (Eds.), *Neural networks: Tricks of the trade* (2nd ed., pp. 421-436). Springer. https://doi.org/10.1007/978-3-642-35289-8_25

Chen, J., Zhang, D., Suzauddola, M., & Zeb, A. (2021). Identifying crop diseases using attention embedded MobileNet-V2 model. *Applied Soft Computing, 113*, Article 107901. https://doi.org/10.1016/j.asoc.2021.107901

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(1), Article 6. https://doi.org/10.1186/s12864-019-6413-7

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1251-1258). IEEE. https://doi.org/10.1109/CVPR.2017.195

Gorriz, J. M., Clemente, R. M., Segovia, F., Ramirez, J., Ortiz, A., & Suckling, J. (2024). Is k-fold cross-validation the best model selection method for machine learning? *arXiv*. https://doi.org/10.48550/arXiv.2401.16407

Hasan, M. N., Miah, M. S., Ghose, P., Jannat, T., Bhuyain, M. M. H., Chowdhury, M. S. A., ... & Harun-Ar-Rashid, M. (2025). A Deep Learning Approach for Brain Tumor Diagnosis: Combining an 8 Layer CNN with Rigorous K-Fold Validation. *Mathematical Modelling of Engineering Problems*, *12*(12). https://doi.org/10.18280/mmep.121227

Horiuchi, A., Masuda, K., Shirasawa, K., Onoue, N., Matsuzaki, R., Tao, R., Kubo, Y., Ushijima, K., & Akagi, T. (2023). Genetic basis of lineage-specific evolution of fruit traits in hexaploid persimmon. *DNA Research, 30*(5), Article dsad015. https://doi.org/10.1093/dnares/dsad015

Huang, C., Sarabi, M., & Ragab, A. E. (2024). MobileNet-V2/IFHO model for accurate detection of early-stage diabetic retinopathy. *Heliyon, 10*(17), Article e37293. https://doi.org/10.1016/j.heliyon.2024.e37293

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture, 147*, 70-90. https://doi.org/10.1016/j.compag.2018.02.016

Kazi, A. (2024). Multiclass classification using enhanced MobileNetV2 architecture. *International Journal of Scientific Research in Engineering and Management, 8*(6), 1-5. https://doi.org/10.55041/IJSREM35845

Krause, J., Jin, H., Yang, J., & Fei-Fei, L. (2015). Fine-grained recognition without part annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5546-5555). IEEE. https://doi.org/10.1109/CVPR.2015.7299173

Kusumaningtyas, E. M., Ramadijanti, N., & Rijal, I. H. K. (2024). Convolutional neural network implementation with MobileNetV2 architecture for Indonesian herbal plants classification in mobile app. In *2024 International Electronics Symposium (IES)* (pp. 521-527). IEEE. https://doi.org/10.1109/IES63037.2024.10665862

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444. https://doi.org/10.1038/nature14539

Maaroof, M. K. A., & Bouhlel, M. S. (2025). Real-time object detection using YOLO-8 model: A drone-based approach. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 16*(1), 190-204. https://doi.org/10.58346/JOWUA.2025.I1.011

Maesaroh, I. S., Syaja'ah, K. N., Perkasa, Y. S., Hadianti, S., Riana, D., & Nurmalasari, R. R. (2024). Cervical cancer classification from Pap smear using Xception model. In *2024 10th International Conference on Wireless and Telematics (ICWT)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICWT62080.2024.10674721

Magandhi, M., Sobir, W. Y., & Sudarmono, M. D. (2024). DNA barcoding of endemic durian kura-kura in West Kalimantan, Indonesia. *SABRAO Journal of Breeding and Genetics, 56*(5). https://doi.org/10.54910/sabrao2024.56.5.5

May, O. S., Ismail, W. N. A. B. W., & Jun, C. (2023). The king of fruits dilemma in Malaysia: Discovering durian export challenges to China. *Journal of Engineering Research and Education, 15*, 61-75. https://doi.org/10.58915/jere.v15.2023.652

Meethongjan, K., Hoang, V. T., & Surinwarangkoon, T. (2022). Data augmentation by combining feature selection and colour features for image classification. *International Journal of Electrical and Computer Engineering, 12*(6), 6172-6177. https://doi.org/10.11591/ijece.v12i6.pp6172-6177

Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science, 7*, Article 1419. https://doi.org/10.3389/fpls.2016.01419

Nimma, D., Al-Omari, O., Pradhan, R., Ulmas, Z., Krishna, R. V. V., El-Ebiary, T. Y. A. B., & Rao, V. S. (2025). Object detection in real-time video surveillance using attention-based transformer-YOLOv8 model. *Alexandria Engineering Journal, 118*, 482-495. https://doi.org/10.1016/j.aej.2025.01.032

Rahman, M. M., Basar, M. A., Shinti, T. S., Khan, M. S. I., Babu, H. M. H., & Uddin, K. M. M. (2023). A deep CNN approach to detect and classify local fruits through a web interface. *Smart Agricultural Technology, 5*, Article 100321. https://doi.org/10.1016/j.atech.2023.100321

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 779-788). IEEE. https://doi.org/10.1109/CVPR.2016.91

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of database systems* (pp. 532-538). Springer. https://doi.org/10.1007/978-0-387-39940-9_565

Saleem, M. H., Potgieter, J., & Arif, K. M. (2019). Plant disease detection and classification by deep learning. *Plants, 8*(11), Article 468. https://doi.org/10.3390/plants8110468

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4510-4520). IEEE. https://doi.org/10.1109/CVPR.2018.00474

Sengupta, P., Mehta, A., & Rana, P. S. (2023). Enhancing performance of deep learning models with a novel data augmentation approach. In *2023, the 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-7). IEEE. https://doi.org/10.1109/ICCCNT56998.2023.10308298

Singh, A., Ganapathysubramanian, B., Singh, A. K., & Sarkar, S. (2016). Machine learning for high-throughput stress phenotyping in plants. *Trends in Plant Science, 21*(2), 110-124. https://doi.org/10.1016/j.tplants.2015.10.015

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002

Sulzbach, E., Scheeren, I., Torres Veras, M. S., Tosin, M. C., Ellert Kroth, W. A., Merotto, A., & Markus, C. (2025). Deep learning model optimisation methods and performance evaluation of YOLOv8 for enhanced weed detection in soybeans. *Computers and Electronics in Agriculture, 232*, Article 110117. https://doi.org/10.1016/j.compag.2025.110117

Tejada, C., Espinoza, G., & Subauste, D. (2024). Use of Xception architecture for the classification of skin lesions. In *Proceedings of the International Multi-Conference on Complexity, Informatics and Cybernetics (IMCIC 2024)* (pp. 129-134). https://doi.org/10.54808/IMCIC2024.01.129

Too, E. C., Yujian, L., Njuki, S., & Yingchun, L. (2019). A comparative study of fine-tuning deep learning models for plant disease identification. *Computers and Electronics in Agriculture, 161*, 272-279. https://doi.org/10.1016/j.compag.2018.03.032

Ullah, N., Khan, J. A., Almakdi, S., Alshehri, M. S., Al Qathrady, M., Aldakheel, E. A., & Khafaga, D. S. (2023). A lightweight deep learning-based model for tomato leaf disease classification. *Computers, Materials & Continua, 77*(3), 3969-3992. https://doi.org/10.32604/cmc.2023.041819

Uzer, M. S. (2025). Deep learning-based classification consisting of pre-trained models and the proposed model using k-fold cross-validation for pistachio species. *Applied Sciences, 15*(8), Article 4516. https://doi.org/10.3390/app15084516

Wäldchen, J., Rzanny, M., Seeland, M., & Mäder, P. (2018). Automated plant species identification: Trends and future directions. *PLoS Computational Biology, 14*(4), Article e1005993. https://doi.org/10.1371/journal.pcbi.1005993

Wang, N., Ma, S., Li, J., Zhang, Y., & Zhang, L. (2020). Multistage attention network for image inpainting. *Pattern Recognition, 106*, Article 107448. https://doi.org/10.1016/j.patcog.2020.107448

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1492-1500). IEEE. https://doi.org/10.1109/CVPR.2017.634

Yoon, S. (2025). Brain tumour classification using a hybrid ensemble of Xception and parallel deep CNN models. *Informatics in Medicine Unlocked, 54*, Article 101629. https://doi.org/10.1016/j.imu.2025.101629

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems* (Vol. 27, pp. 3320-3328). https://doi.org/10.48550/arXiv.1411.1792

Zeynali, A., Tinati, M. A., & Tazehkand, B. M. (2024). Hybrid CNN-transformer architecture with Xception-based feature enhancement for accurate breast cancer classification. *IEEE Access, 12*, 189477-189493. https://doi.org/10.1109/ACCESS.2024.3516535

Zhang, Y.-D., & Wu, L. (2012). An MR brain image classifier via principal component analysis and kernel support vector machine. *Progress in Electromagnetics Research, 130*, 369-388. https://doi.org/10.2528/PIER12061410

Zhang, Z., Hou, J., & Lin, Z. (2024). Exploring integration methods for image data augmentation. In *2024 6th International Conference on Internet of Things, Automation and Artificial Intelligence (IoTAAI)* (pp. 707-713). IEEE. https://doi.org/10.1109/IoTAAI62601.2024.10692583

Pertanika J. Sci. & Technol. 34 (3): 1551 - 1577 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Noise-Resilient and Low Complexity LACO-OFDM Using Approximation Log-Likelihood Ratio (approxLLR) Soft-decision Demodulation with Single-FFT Receiver

Norzalina Othman[1, 3] and Mohd Rashidi Che Beson[2]*

[1]*Faculty of Electronic Engineering and Technology, Universiti Malaysia Perlis (UniMAP), Pauh Putra Campus, 02600 Arau, Perlis, Malaysia*
[2]*Faculty of Intelligent Computing, Universiti Malaysia Perlis (UniMAP), Pauh Putra Campus, 02600 Arau, Perlis, Malaysia*
[3]*Electrical, Electronic and Automation Section (EEA), Universiti Kuala Lumpur Malaysian Spanish Institute (UniKL MSI), Kulim Hi-Tech Park, 09000 Kulim, Kedah, Malaysia*

ABSTRACT

Layered Asymmetrically Clipped Optical Orthogonal Frequency Division Multiplexing (LACO-OFDM) is a promising multicarrier modulation scheme applied in Visible Light Communication (VLC) systems. The conventional LACO-OFDM with iterative receiver introduces high computational complexity and low noise resilience, particularly under high-order modulation and multi-layer configurations. A new enhancement of the demodulation technique based on approximate Log-Likelihood Ratio (approxLLR) with a single-FFT receiver was proposed to address this limitation. The approach enhances noise robustness by estimating bit-wise confidence levels via LLR approximation, while significantly reducing complexity by eliminating the need for pairwise FFT/IFFT operations and the layer-by-layer reconstruction of clipping distortion, which are typically required in conventional LACO-OFDM with iterative receiver demodulation schemes. A significant performance improvement at a target bit error rate (BER) of 10^{-4}, effectively demonstrated with a 4 to 5 times enhancement over the conventional scheme under moderate modulation orders. This corresponds to BER reductions of approximately ~0.705, ~0.615, and ~0.591 orders of magnitude at SNR = 10 dB for 16-QAM, 32-QAM, and 64-QAM. For symbol detection on higher layers (layer≥2), 0.7, 0.3, and 0.1 were identified as optimal scale factors for 4-QAM, 16-QAM, and 64-QAM, respectively. The approxLLR SD approach preserves energy efficiency (EE) in the power analysis, despite only slightly increasing power even with a higher layer count configuration.

Keywords: LACO-OFDM, LLR, optical wireless communication, soft-decision demodulation, visible light communication

ARTICLE INFO

Article history:
Received: 25 September 2025
Accepted: 04 January 2026
Published: 19 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.12

E-mail addresses:
norzalina@unikl.edu.my (Norzalina Othman)
rashidibeson@unimap.edu.my (Mohd Rashidi Che Beson)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

INTRODUCTION

Visible Light Communication (VLC) is a subset of optical wireless communication (OWC) that has emerged as a potential substitute for high-speed data transmission utilising modulated light over unregulated bandwidth. VLC, which operates in the visible light spectrum (380-780 nm; 430-790 THz) (Vappangi et al., 2021; Seow et al., 2021), is a complementary standard of radio frequency (RF) based systems by providing a broader bandwidth, unlicensed channels, low power consumption, immunity to electromagnetic interference, and no health hazards.

VLC research has extensively investigated various areas, including experimental hardware designs, modulation techniques, and applications in diverse environments. Different modulation schemes were designed and developed to meet the varied demands of various applications, including high data rates, energy efficiency (EE), spectral efficiency (SE), bit error rate (BER) performance and system complexity. These aspects have been thoroughly studied in prior research on the VLC system. Orthogonal Frequency Division Multiplexing (OFDM) has evolved as a prominent modulation technique due to its multicarrier structure, which enables data to be transmitted simultaneously over multiple orthogonal subcarriers. This scheme exhibits high spectrum and power efficiency, as well as significant tolerance to inter-symbol interference (ISI) and consistent performance over frequency-selective channels, while maintaining the low complexity of a fundamental framework that enables single-tap signal equalisation (Aliaberi et al., 2019; Jativa et al., 2020; Othman et al., 2024; Vappangi et al., 2021; Zhang et al., 2021).

Extensive research on OFDM-based approaches aimed at optimising the performance of OWC, particularly within the field of VLC technology. Othman et al. (2024) classified OFDM-based modulation into two groups: single-component and multi-component (also known as hybrid-component) OFDM schemes. Single-component OFDM includes Direct Current-biased Optical OFDM (DCO-OFDM), Asymmetrically Clipped Optical OFDM (ACO-OFDM), Pulse-Amplitude Modulated Discrete Multitone (PAM-DMT), Unipolar or Flip OFDM, and enhanced U-OFDM. Meanwhile, Asymmetrically Clipped DC-biased Optical OFDM (ADO-OFDM), Hybrid Asymmetrically Clipped Optical OFDM (HACO-OFDM), Asymmetrically Clipped Absolute-Value Optical OFDM (AAO-OFDM), and Layered ACO-OFDM (LACO-OFDM) were classified as multi-component or hybrid-component OFDM variants (Jativa et al., 2020; Vappangi et al., 2021; Zhang et al., 2021). The OFDM signal or data transmitted via the optical channel must be a real and non-negative signal to comply with the intensity modulation and direct detection (IM/DD) principle.

OFDM modulation is widely utilised in the context of its ability to support high data throughput and strong immunity to ISI. Nevertheless, it inherently suffers from a high peak-to-average power ratio (PAPR), which can limit system efficiency.

To overcome this drawback, modified OFDM schemes such as ACO-OFDM and DCO-OFDM have been introduced. The ACO-OFDM improves BER and PAPR by modulating data on only half of the subcarriers, specifically on the odd subcarriers. Furthermore, DCO-OFDM employs the entire spectrum; however, it suffers from inefficient optical power due to the DC bias required to preserve signal non-negativity (Zhang et al., 2025).

LACO-OFDM was an advancement on ACO-OFDM that improves SE in IM/DD-based communication systems. While maintaining the ACO-OFDM principle, multiple layers of ACO-OFDM symbols were divided and superimposed to exploit more subcarriers without requiring DC bias, thereby improving spectral and power efficiency (Nakano et al., 2023; Zhang et al., 2017;). Despite these advantages, this scheme introduced the incrementation of clipping distortion in the higher layers, which degrades the BER performance. An adaptive modulation and advanced receiver architecture have been developed to address these limitations.

The most recent LACO-OFDM research has focused on improving system efficiency while minimising computational complexity. Adaptive LACO-OFDM, introduced in Nakano et al. (2023), modified the modulation order in each layer dynamically according to the noise level. The needs of the process on multiple layers with different modulation orders result in high computational complexity, which in turn diminishes the practicality of the scheme. Previous studies by Liu et al. (2018) and Zhang et al. (2020) focused on improving LACO-OFDM performance by applying layer precoding and optimal power allocation to strengthen receiver sensitivity and system capacity. Additionally, tone-injection-based PAPR-reduction has been employed to suppress peak signals, hence improve BER performance (Zhang et al., 2017).

Diverse low-complexity approaches have been investigated and studied to address the computational demands of conventional LACO-OFDM with iterative receivers caused by repeated Fast Fourier Transform and Inverse Fast Fourier Transform (FFT/IFFT) operations for symbol reconstruction and clipping distortion (Lowery, 2020). This includes low-complexity LACO-OFDM (L-LACO) using half-size FFT/IFFT operations (Bai et al., 2022), single-FFT receivers (Liu et al., 2020), and single-FFT with pairwise maximum likelihood (ML) detection (Li et al., 2021), all of which offer lower complexity while improving BER performance. Furthermore, the adaptive scheme introduced by Yan et al. (2020) and Zhang et al. (2021) has been devised to improve system performance by adjusting the number of layers (layer count) based on signal-to-noise ratio (SNR) conditions. Collectively, these enhancements demonstrate LACO-OFDM's potential as an efficient and scalable solution for VLC and other OWC systems.

Most of the previous studies on LACO-OFDM have employed hard-decision (HD) demodulation for symbol detection. This modulation maps the received symbols directly to the nearest constellation points, resulting in inferior performance under noisy conditions.

Soft-decision (SD) demodulation improves detection reliability by utilising log-likelihood ratios (LLRs) to quantify bit-wise confidence, leading to an increase in BER performance, particularly with high-order modulation schemes. However, the complexity of the full exact LLR computation exponentially increases with high-order modulation, requiring the simpler scheme development.

Olivatto et al. (2017) proposed Simplified LLR demappers for high-order Phase-Shift Keying (PSK) and Amplitude and PSK (APSK) that substantially reduce complexity with no performance loss as compared to the Max-Log-MAP (Maximum a Posteriori) soft-demapper. Meanwhile, Mao et al. (2020) improved a simplified LLR soft demodulation algorithm for 64-QAM, which optimises decision rules to decrease computational complexity while minimising performance loss significantly. However, its effectiveness may be limited when applied to different modulation orders without further modifications. Another study investigated the integration of SD approaches into system-level designs, utilising LLR-assisted Viterbi decoders (Wang et al., 2023), which improve BER performance yet face the constraint length as a constraint.

The implementation of long irregular low-density parity-check (LDPC) codes in SD decoding has demonstrated considerable BER improvements and coding advantages over HD decoding, particularly in high-order QAM applications. This makes SD decoding well-suited to high-capacity optical communication, as reported by Hussein et al. (2021), which exhibits improved system performance while highlighting the importance of implementations that balance performance gains with computational efficiency.

A simplified LLR applied in OFDM with index modulation (OFDM-IM) as presented by Hu et al. (2018), utilises an SNR-dependent radius to identify the number of most likely transmitted symbols. OFDM-IM thereby reducing computational complexity while attaining near-optimal BER performance. The integration of index modulation and bandwidth optimisation in LACO-OFDM demonstrates a significant enhancement in SE and robustness across different channel conditions (Azim et al., 2021). In other related development, Anand et al. (2024) introduced a simplified LLR detector for OFDM systems based on the fractional Fourier transform (FrFT), which notably mitigates intercarrier interference (ICI) typically encountered in conventional intensity modulation OFDM (IM-OFDM) configurations.

Despite LACO-OFDM's promising performance in improving SE, current receiver architectures are computationally intensive. Conventional LACO-OFDM receivers commonly rely on pairwise FFT/IFFT processing for symbol reconstruction and clipping distortion, which incur significant complexity as the number of layers and modulation order increase. Several low-complexity single-FFT receivers have been reported; however, these approaches predominantly adopt HD demodulation that does not exploit bit-wise reliability information. Moreover, existing LLR demodulation techniques are generally developed

for general optical OFDM and do not explicitly account for the layer-dependent clipping noise inherent in LACO-OFDM systems.

To the best of the authors' knowledge, a LACO-OFDM framework that integrates a bit-wise LLR-based SD demodulation with a single-FFT receiver while preserving robustness against clipping distortion has not been previously reported. This work addresses this gap as well as mitigates a drawback encountered by conventional LACO-OFDM with an iterative receiver by proposing a new enhancement approach of LACO-OFDM that incorporates a noise-resilient, low-complexity demodulation scheme as presented in this research article. This enhanced version of LACO-OFDM uses an approximate LLR (approxLLR) SD demodulation with a single-FFT receiver configuration. A layer-by-layer SD framework has been adopted, aimed at enhancing both BER performance and SE. Computational efficiency is achieved by eliminating the need for the pairwise FFT/IFFT operation typically used for symbol reconstruction and clipping distortion suppression as implemented in Lowery (2020). Instead, higher-layer symbols are efficiently detected through direct subtraction of the received symbol from its corresponding modulation reference. Furthermore, the approxLLR preserves bit-level reliability by estimating the likelihoods, mapping probability values to the bit representations, and selecting the most probable symbol for final decoding.

The remaining sections of the article are organised as follows: the proposed modulation scheme is outlined in the Methodology section. Section Results and Discussion present the results and performance evaluation of the LACO-OFDM approxLLR SD demodulation with a single-FFT receiver in comparison to the conventional LACO-OFDM with an iterative receiver scheme. The conclusion section concludes with a summary of the research findings, the implications of the research work and the practical relevance of the proposed approach.

METHODOLOGY

This section presents the proposed LACO-OFDM using approxLLR SD demodulation with a single-FFT receiver applied in a VLC link. It begins with an overview of conventional ACO-OFDM and LACO-OFDM principles, followed by a detailed mathematical model of the transmitter and receiver of the proposed approach. Finally, simulation outcomes of the proposed approach are evaluated and compared with the conventional LACO-OFDM with an iterative receiver scheme using performance analysis formulae as discussed in this section to completely examine the proposed approach's effectiveness and improvements.

ACO-OFDM and LACO-OFDM Principle

ACO-OFDM exploits the signal's anti-symmetry and clipped at zero on the negative values to achieve non-negativity without losing information. The system's design ensures that any distortion caused by clipping only applies to subcarriers that do not carry data,

which occurs on even subcarriers, allowing the receiver to accurately detect symbols on odd subcarriers. This approach is highly beneficial in optical communication to ensure the reliability of transmitted data while fulfilling the requirements of VLC systems.

Conventional ACO-OFDM loads data solely on odd subcarriers, leaving even subcarriers unoccupied. The constellation mapping, Serial to Parallel (SP), Hermitian Symmetry (HS), Inverse Fast Fourier Transform (IFFT), Parallel to Serial (PS), zero clipping and Cyclic Prefix (CP) were imposed in the transmitter, while CP removal, Fast Fourier Transform (FFT) and de-mapping were imposed in the receiver. Despite this method being spectrally efficient, the amount of data it can transmit in each bandwidth is limited. To address this issue, multiple L layers ACO-OFDM modulated symbol on several N Subcarriers were stacked or superimposed to create LACO-OFDM.

Equation 1 defines the number of occupied subcarriers in layered ACO-OFDM N_L can be represented by:

$$N_L = \sum_{l=1}^{L} \frac{N}{2^l} \quad [1]$$

with the maximum number of layers L provided by Equation 2 can be calculated as:

$$L = \log_2 N - 1 \quad [2]$$

However, the commonly applied number of layers is between 4 or 5 layers as reported in Zhang et al. (2019), sufficient to strike the trade-off between throughput and energy consumption.

Each layer's anti-symmetry allows clipping without losing data, and the successive demodulation scheme enables the receiver to independently retrieve data from each layer. Due to the layered approach, which maximises data transmission within a specified bandwidth, LACO-OFDM is adequate for high-speed VLC systems.

In traditional LACO-OFDM with L layers and N Subcarriers, the input data is first mapped to QAM symbols, then HS is applied to the subcarriers to ensure that the signal is real-valued in the time domain. Only the odd subcarriers carry data, resulting in an anti-symmetric time domain after the IFFT, represented by Equation 3 as the frequency domain of the symbol vector X_k :

$$X_k = \left[0, X_1, 0, X_3, \dots, X_{\frac{N}{2}-1}, \dots, X^*_{\frac{N}{2}-1}, \dots, , X^*_3, 0^*, X^*_1\right], k = 0, 1, 2, \dots, \frac{N}{2} - 1 \quad [3]$$

where X_k denotes the M-QAM symbol of k-th subcarrier index and $(.)^*$ denotes complex conjugate after performing HS computation. In the l-th layer, the constellation symbols are assigned to subcarriers with index given by $2_{l(2k+1)}2^{l-1}(2k+1)$ where $k = 0,1,2,\ldots,\frac{N}{2^l}-1$. All other subcarriers are set to zero.

To meet the non-negativity constraint, the negative components of x_n demonstrate by Equation 4 as time domain data symbol, are clipped to zero, resulting in x_{zclip} on the n-th element. The CP computation is then performed for each layer l and is defined as:

$$x_n^{(l)} = \frac{1}{2}\, x_n^{(l)} + x_{zclip}^{(l)}\ , n = 0, 1, 2, \ldots, N-1 \qquad [4]$$

After the CP computational, all symbol in L layer is superimposed and the superimposed symbol represent by $x_{LACO,n}$ before being transmitted to optical channel. The received LACO-OFDM symbols denoted by Equation 5 as:

$$y_n = x_{LACO,n} * h_n + w_n \qquad [5]$$

where:

y_n: received signal

$x_{LACO,n}$: superimposed transmit signal

h_n: channel impulse response

w_n: Additive white Gaussian noise (AWGN)

Note. Shot noise and thermal noise are well modelled as AWGN (Bai et al., 2022)

As in the receiver, after the CP removal and FFT computation, the ACO-OFDM symbol can be detected layer by layer using Equation 6 by employing clipping distortion subtraction to obtain a new received frequency domain of $\hat{R}_k^{(l)}$:

$$\hat{R}_k^{(l)} = \arg min_{X \in S} \left| X - 2\hat{R}_k^{(l-1)} \right|, \qquad \text{k= 1, 3, 5, ….,} \frac{N}{2} - 1 \qquad [6]$$

where:

S : constellation set mapping of X symbol

$\hat{R}_k^{(l-1)}$: received frequency-domain symbol of the previous layer

LACO-OFDM Transmitter

Figure 1 shows the proposed LACO-OFDM transmitter by adding a layer occupied the ACO-OFDM principle. The input data stream is first mapped onto a complex frequency-domain symbol X_{CQAM}, where $CQAM = 1,2,3,\dots,N$ based on the given M-QAM constellation before being split into L parallel layer division for the layer number of l-th as denoted by Equation 7.

$$X^{l}_{2^{l-1}(2k+1)}, \quad k = 0,1,2,\dots,\frac{N}{2^{l}} - 1 \qquad [7]$$

To generate a real-value of time-domain symbol as indicate in Equation 8, HS is imposed on the symbol in the odd subcarrier by setting the first index and $\frac{N}{2}$ index to zero as :

$$X^{l}_{k} = \begin{cases} X^{l}_{k} = X^{*l}_{N-k}, \quad k = 0,1,2,\dots,\frac{N}{2} - 1 \\ X^{l}_{0} = 0 \\ X^{l}_{\frac{N}{2}} = 0 \end{cases} \qquad [8]$$

where (*) denotes the conjugate of complex numbers.

For example, the active symbol in the layer $l = 1$ demonstrate in Equation 9 is represented as follows:

$$X^{1} = \left[0, X_1, 0, X_3, \dots, 0, X_{\frac{N}{2}-1}, 0, X^{*}_{\frac{N}{2}-1}, \dots, , X^{*}_{3}, 0^{*}, X^{*}_{1}, 0\right] \qquad [9]$$

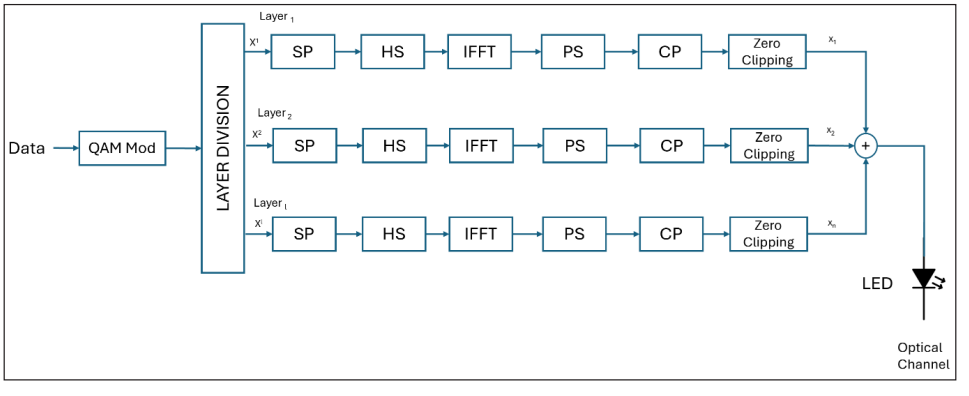


Figure 1. LACO-OFDM approxLLR SD transmitter block diagram

Equation 10 defines the time-domain signal of the layer l-th, x_n^l were obtained after N-point of IFFT, becomes real-valued and anti-symmetric by performing the following equation:

$$x_n^l = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X_k^{(1)} \; exp\left(j \frac{2\pi}{N} nk\right), n = 0,1,2,3, \dots, N \quad [10]$$

which follows a half-wave symmetry provide by Equation 11 as:

$$x_n = x_{n+\frac{N}{2}}, \quad n = 0,1,2,3, \dots, \frac{N}{2} - 1 \quad [11]$$

Hence, the negative time-domain symbol that commonly falls or appears at the even subcarrier can be clipped to zero, $\lfloor x_n^l \rfloor_{zclip}$ as define in Equation 12 without any loss of information.

$$\lfloor x_n^l \rfloor_{zclip} = \begin{cases} x_n^l, & x_n^l \geq 0, \\ 0, & x_n^l < 0 \end{cases} \quad [12]$$

The distortion caused by clipping is compensated for by adding $\frac{1}{4}\ N$ CP to mitigate inter-layer interference (ILI) and ISI.

In the layered ACO-OFDM transmitter, different layers are superimposed in the time domain and written as in Equation 13:

$$x_{LACO,n} = \sum_{l=1}^{L} \lfloor x_n^l \rfloor_{zclip}, \quad n = 0,1,2,3, \dots, N-1 \quad [13]$$

where:

$x_{LACO,n}$ – superimposed time-domain symbol

L – total number of layers of LACO-OFDM

$\lfloor x_n^l \rfloor_{zclip}$ – time-domain zero clipping symbol of the layer l

Lower clipping of the time-domain signal will generate distortion in the frequency domain. Since the effects only fall on even subcarriers in every layer, only the signal of the layer $l \geq 2$ are distorted. The distorted result from the lower layer has been added in the frequency-domain symbol as illustrated in Figure 2.

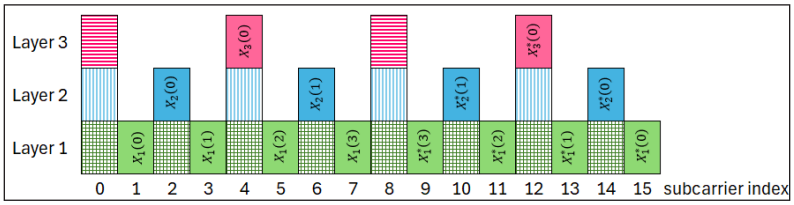


Figure 2. 3-layer ACO-OFDM with 16-point FFT in frequency domain view to illustrate the superimposed layers
Note. (*) It is a symbol of Hermitian symmetry conjugate

Proposed LACO-OFDM Approximate LLR Soft-Decision Algorithm with Single-FFT Receiver

The previous study conducted by Bai et al. (2022) and Cao et al. (2019) imposed an iterative receiver with a subtraction operation of clipping distortion, which is done layer by layer (successive detection). The symbols in the higher layer are recovered after the negative clipping distortion from the corresponding layer has been removed. Even though this approach offered high SE, it faced computational complexity issues.

Figure 3 shows a block diagram of the proposed LACO-OFDM receiver. In this study, an approxLLR is employed in SD demodulation. Then, the approxLLR values of $q = \log_2(M)$ bits of symbol R_k^l are estimated to calculate the bit probabilities of all bits in the symbol. The following paragraph includes a comprehensive explanation regarding the proposed method.

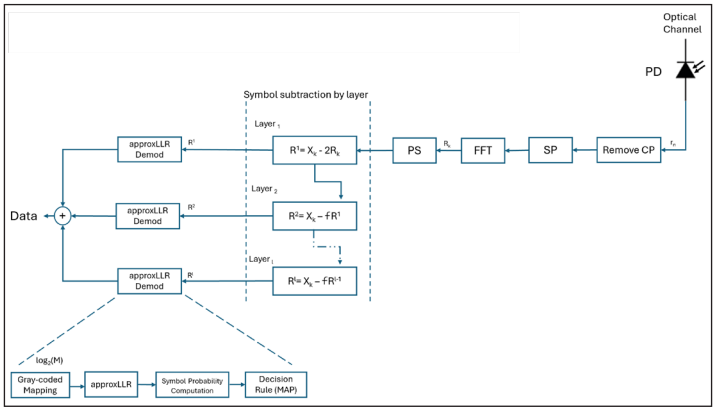


Figure 3. LACO-OFDM approxLLR SD with single-FFT receiver block diagram

For l- In the layer ACO-OFDM, the CP removal is imposed on the received time-domain symbol, r_n and FFT computation was applied to generate the frequency-domain symbol, R_k and represented by Equation 14 as:

$$R_k = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} r_n \, exp\left(j \, \frac{2\pi}{N} nk\right), \quad k = 0,1,2,3, \ldots, \frac{N}{2} - 1 \qquad [14]$$

The first layer $l = 1$ is free from ILI, whereas all subsequent layers are affected by ILI from the preceding layers, ranging from layer $l = 1$ to L. For $l = 1$, the transmitted symbol could be directly detected by using an odd subcarrier of $R_k = R_k^1$ and denoted in Equation 15 by:

$$\hat{R}_k^1 = argmin_{X \in CQAM} \left| X_k - 2R_k^1 \right| \;, k = 1,3,5, \ldots, \frac{N}{2^l} \qquad [15]$$

where $CQAM$ represent the complex number of the constellation set of X symbol and the scale factor of 2 in Equation 15 is because the clipping operation reduces the power of ACO-OFDM symbols in the odd subcarriers by half (Wang et al., 2015).

For the subsequent layer, $l = 2$ to L, the transmitted symbol can be detected using Equation 16 by:

$$\hat{R}_k^l = argmin_{X \in CQAM} \left| X_k - f\hat{R}_k^{(l-1)} \right| \qquad [16]$$

where:

$\hat{R}_k^{(l-1)}$ = received frequency-domain symbol of the previous layer

$k = 1(2^{(l-1)}),3(2^{(l-1)}),5(2^{(l-1)}), \ldots, \frac{N}{2^l}(2^{(l-1)})$

f = scale factor

Figure 4 illustrates the sequential operations executed within the approxLLR demodulation block. Specifically, the diagram presents a step-by-step representation of how the received frequency-domain symbols are decoded, starting with the mapping of the noisy received symbols onto the M-ary QAM constellation, followed by the evaluation of Euclidean distances between the received symbols and candidate constellation points.

These distances are then used to approximate the bit-wise log-likelihood ratios, producing a soft-decision output for each bit position associated with a symbol. A Maximum a Posteriori (MAP) decision rule is applied to select the most probable transmitted symbol, enabling accurate symbol recovery under noise and interference conditions.

Initially, the received frequency-domain symbol $\hat{R}_k^l$ undergoes Grey-coded mapping onto the M-QAM constellation. This process is carried out after subtracting the interference introduced by the preceding layer, as described in the previous step in Equation 16. The use of Grey-coded mapping ensures that each symbol is represented by its corresponding bit sequence in such a way that adjacent constellation points differ by only a single bit, thereby reducing the likelihood of bit errors during demodulation. Each mapped symbol in the constellation can be expressed as a binary vector as described in Equation 17.

Let the constellation size be M, the number of bits per symbol is represented as $q = \log_2(M)$. For the i-th received symbol, the associated bit vector is defined as:

$$b_i = \left[b_{i,1}, b_{i,2}, \ldots, b_{i,q}\right], \; b \in \{0,1\}^q \qquad [17]$$

Figure 5 is the illustration of a 16-QAM constellation with Gray-coded mapping with bit label format $[b_1\ b_2 \mid b_3\ b_4]$ with last-MSB order where:

b_1b_2 = Grey-coded bits for the I-axis (In-phase).

b_3b_4 = Grey-coded bits for the Q-axis (Quadrature).

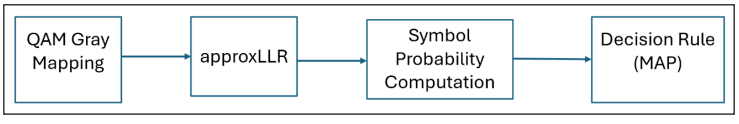


Figure 4. ApproxLLR SD demodulation block diagram

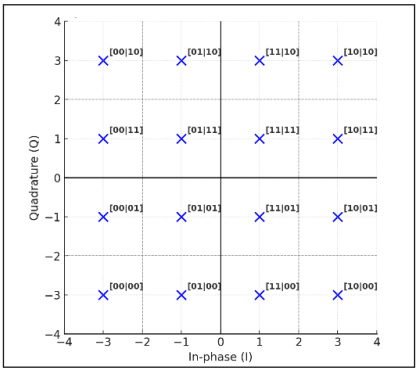


Figure 5. 16-QAM constellation with Gray-coded bit label $[b_1\ b_2 \mid b_3\ b_4]$

To recover these bits at the receiver, each bit is split into 2 subsets, $S_q^{(0)}$ *or* $S_q^{(1)}$ and subsequently the demapper computes approxLLR using Euclidean distance of 1-D nearest squared distance between the binary vector of $b_{i,q}$ and constellation point.

Figure 6 illustrates the 16-QAM constellation diagram, with the in-phase (I) and quadrature (Q) axes defined by the discrete amplitude levels {−3d, −d, +d, +3d}. These levels correspond directly to the 1-D Grey-coded mapping of 16-QAM. The constellation points are explicitly labelled along each axis according to the governing Equation 18:

$$\mathcal{A} = \left\{ \pm(2i-1)d \mid i = 1,2,\ldots,\frac{\sqrt{M}}{2} \right\} \quad [18]$$

where:

$\mathcal{A}$ – Axis level

d - spacing parameter (half of the step between adjacent levels)

i- index (1 to $\frac{\sqrt{M}}{2}$) that generates the inner to outer levels

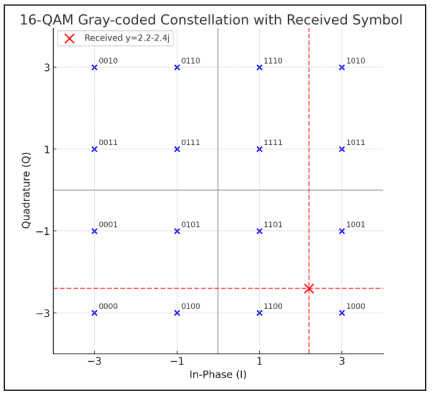


Figure 6. 16-QAM Gray-coded constellation with received bit

Specifically, for the i-th received symbol, the approxLLR corresponding to the q-th bit position in Equation 19 is given by:

$$\mathcal{L}_{i,q} \approx \frac{min_{s \in S_q^{(0)}} |y_i - s|^2 - min_{s \in S_q^{(1)}} |y_i - s|^2}{\sigma^2} \quad [19]$$

where:

$S_q^{(0)}$ – subset of constellation points with bit $b_q = 0$

$S_q^{(1)}$ – subset of constellation points with bit $b_q = 1$

y_i – received symbol after grey-coded mapping demodulation

σ^2 – noise variance of thermal and shot noise (AWGN)

From Equation 19, $\mathcal{L}_{i,q}$ denotes the LLR associated with q-th bit of the i-th received symbol, quantifies the relative likelihood of bit either 1 or 0, which stores the LLR values of all N grey-mapped received symbol across all bit positions in the LLR matrix $\left[\mathcal{L}_{i,q}\right]_{\substack{i=1,\ldots,N \\ q=1,..,\log_2 M}}$. Subsequently, a probability score is assigned for every candidate of a symbol $m \in \{0, 1, \ldots, M-1\}$ by computing the product of the bitwise posterior probability shown in Equation 20, represented by:

$$\tilde{P}_m^{(i)} \propto \prod_{q=1}^{\log_2(M)} P\left(b_q = b_{m,q} | y_i\right) \qquad [20]$$

with each posterior derived directly from the LLR value demonstrate by Equation 21 as:

$$P\left(b_{i,q} = b | y_i\right) = \begin{cases} \frac{e^{\mathcal{L}_{i,q}}}{1+e^{\mathcal{L}_{i,q}}} & , \quad b = 0 \\ \frac{1}{1+e^{\mathcal{L}_{i,q}}} & , \quad b = 1 \end{cases} \qquad \text{where } b \in \{0,1\} \qquad [21]$$

This operation produces an unnormalised likelihood score for every candidate symbol. Finally, the maximum a posteriori (MAP) criterion is applied to identify the most likely transmitted symbol. Equation 22 provides estimated symbol index $\hat{m}^{(i)}$ is obtained by:

$$\hat{m}^{(i)} = \arg max_{m \in \{0,1,\ldots,M-1\}} \tilde{P}_m^{(i)} \qquad [22]$$

By repeating this operation for all received symbols, the final output sequence is $\hat{s} = \left[\hat{m}^{(1)}, \hat{m}^{(2)}, \ldots, \hat{m}^{(N)}\right]$, which constitutes the estimated transmitted symbol stream.

The performance of the proposed approxLLR SD demodulation with a single-FFT approach for the LACO-OFDM system is evaluated through a detailed analysis of the BER and the received power signal. The BER is computed by comparing the detected mismatched bits between the transmitted sequence x_i with the detected received sequence y_i and denoted in Equation 23 Equation 23 for M-QAM over an AWGN channel.

$$BER = \frac{1}{N}\sum_{i=1}^{N}[x_i - y_i] \quad [23]$$

The received power as shown in Equation 24 is estimated after adding AWGN noise and is denoted as:

$$P_{rx} = \frac{1}{N}\sum_{n=1}^{N}|y[n]|^2 = P_{signal} + \sigma^2 \quad [24]$$

where:

P_{rx} : Received power

$y[n]$: received signal after channel noise

N: total number of signal samples

σ^2 : noise variance

The BER performance analysis of both approaches can be calculated by using the order of magnitude difference (OMD) represented by Equation 25:

$$OMD = \log_{10}\frac{BER^{(iter)}}{BER^{(sfft)}} \quad [25]$$

where:

OMD – order of magnitude difference

$BER^{(iter)}$ – BER of conventional iterative- receiver HD at specific SNR

$BER^{(sfft)}$ - BER of single-FFT receiver SD at specific SNR

RESULT AND DISCUSSION

A MATLAB simulation was performed to evaluate the performance of a conventional LACO-OFDM system employing HD demodulation compared to the proposed LACO-OFDM approxLLR with a single-FFT receiver system utilising an SD scheme. Table 1 summarises the parameters employed in the subsequent simulation outcome.

Table 1
Simulation parameter

Description	Parameter
Symbol length (m)	512
No. of subcarriers (N)	1024
Data size	m × N
No of Layer (***L***)	2/3/4
QAM Mod order (M)	4/16/32/64
IFFT size	N-point
FFT size	$\frac{N}{2}$ point
No. of Cyclic Prefix (Ncp) size	$\frac{1}{4} \times N$

Computational Complexity

This section presents a computational complexity analysis of the conventional LACO-OFDM with iterative receiver employing HD demodulation and the proposed approxLLR SD with single-FFT receiver approach, both applied within the LACO-OFDM framework. The complexity evaluation is based on the number of complex-valued multiplications required for the FFT/IFFT operations involved in clipping distortion estimation of the symbol reconstruction-subtraction operations performed at each layer. The complexity of the demapper type approach is also discussed in this section.

The conventional LACO-OFDM with iterative receiver utilising HD demodulation requires pairwise FFT/IFFT with two $\frac{N}{2}$-point FFT and one N-point IFFT operations to regenerate the frequency-domain clipping distortion for each layer of the ACO-OFDM signal. This repeated transformation process significantly contributes to the overall computational complexity. The complexity associated with the N-point IFFT and $\frac{N}{2}$-point FFT operations are denoted in Equation 26 as:

$$IFFT = \mathcal{O}\ (N\ log_2 N\ )$$
$$FFT = \mathcal{O}\left(\frac{N}{2}\ log_2 N\right) \quad [26]$$

The complexity of conventional LACO-OFDM with iterative receiver HD, as denoted in Zhang et al. (2021), can be represented in Equation 27 as:

$$\begin{aligned} \mathcal{O}_{convHD} &= \mathcal{O}(N\ log_2 N) + 2\,\mathcal{O}\left(\frac{N}{2} log_2 N\right) \\ &= \mathcal{O}(N\ log_2 N) + \mathcal{O}(N\ log_2 N) \\ &= 2\mathcal{O}(N\ log_2 N) \end{aligned} \quad [27]$$

For the L – layer of LACO-OFDM, the conventional LACO-OFDM with iterative receiver HD performs layer-by-layer reconstruction and subtraction, iteratively and is presented in Liu et al. (2020) using Equation 28 as:

$$\begin{aligned} \mathcal{O}_{ConvHD} &= L\,.\left(\mathcal{O}(N\ log_2 N) + 2\mathcal{O}\left(\frac{N}{2} log_2 N\right)\right) \\ &= L\,.2\,.\mathcal{O}(N\ log_2 N) \\ &= L2\mathcal{O}(N\ log_2 N) \end{aligned} \quad [28]$$

In HD QAM demodulation, no multiplication or distance computation is required, as it is based on a quantisation decision. Hence, the demapper complexity of HD demodulation per symbol is $\mathcal{O}(1)$. The direct threshold comparison is imposed on every symbol across all layers, with the total number of symbols, N_s as indicate in Equation 29 by:

$$N_s = \frac{N}{2}\left(1 - \frac{1}{2^L}\right) \quad [29]$$

In contrast, the proposed approxLLR SD demodulation utilises a single-FFT receiver, requiring only one N-point FFT operation at the receiver. The associated computational complexity for L – layer LACO-OFDM is represented by Equation 30 as:

$$\mathcal{O}_{Single\ -FFT} = \mathcal{O}(N\ log_2 N) \quad [30]$$

This demonstrates that the approxLLR SD with the single-FFT approach reduces computational complexity by a factor of approximately 2 times compared to the conventional LACO-OFDM with iterative-receiver HD scheme. Such efficiency is particularly beneficial for real-time VLC systems, where low latency and minimal processing time are critical requirements.

The proposed approach is imposed with a bit-wise demodulation operation, and the complexity of approxLLR computation across all layers of L denoted in Equation 31 as:

$$\mathcal{O}_{approxLLR} = \mathcal{O}(N\ log_2 N) + \frac{N}{2}\left(1 - \frac{1}{2^L}\right) \cdot \mathcal{O}\left(q\sqrt{M}\right), \quad [31]$$

where $q = log_2 M$

The demapper complexity of approxLLR $\mathcal{O}\left(q\sqrt{M}\right)$ derive from the arithmetic operation (addition, subtraction, multiplication and comparison) involved in distance calculation, bit-wise minimum selection and scaling computation (refer to Equation 19). By computing a 1-D nearest squared distance on metrics of $t = \sqrt{M}$ amplitude levels in-phase (I) and quadrature (Q) components separately, the computation required $2t$ real subtraction and $2t$ real multiplication for detected subcarriers. The subsequent bitwise LLR computation involves selecting two minima matrices per bit from a subset of size $\frac{t}{2}$ where require $\left(\frac{t}{2} - 1\right)$ comparison. Thus, the required minima comparison per bit is denoted in Equation 32 as:

$$\left(\frac{t}{2} - 1\right) + \left(\frac{t}{2} - 1\right) = t - 2 \quad [32]$$

resulting in $q(t-2)$ comparison across all q bits followed by $1q$ single subtraction and $1q$ single scaling operation per bit.

The final complexity per detected subcarrier of approxLLR demodulation in Equation 33 is summarised by:

$$\begin{aligned} N_{sub} &= 2t + q \quad \text{- subtraction operation of } N \text{ subcarrier} \\ N_{mul} &= 2t + q \quad \text{- multiplication operation of } N \text{ subcarrier} \\ N_{cpr} &= q(t-2) \quad \text{- comparison operation of } N \text{ subcarrier} \end{aligned} \quad [33]$$

The overall approxLLR demodulation complexity $\mathcal{O}\left(q\sqrt{M}\right)$ is derive by $T_{approxLLR}$ summation of operations in Equation 33 makes it the total arithmetic cost per subcarrier denotes in Equation 34 as:

$$\begin{aligned} T_{approxLLR}\ (M) &= N_{sub} + N_{mul} + N_{cpr} \\ &= (2t+q) + (2t+q) + q(t-2) \\ &= t(q+4) \end{aligned} \quad [34]$$

Since multiplying a constant does not change the growth in Big-O $(\mathcal{O})$ analysis and the constant multipliers are omitted, therefore $T_{approxLLR}\ (M) = \mathcal{O}(qt)$ and substitute $t = \sqrt{M}$, then the approxLLR demodulation computation shown in Equation 35 is represented as:

$$T_{approxLLR}\ (M) = \mathcal{O}\left(q\sqrt{M}\right) \quad [35]$$

The most likely transmitted symbol is then obtained with a MAP decision on the LLR symbols, which incurs only $\mathcal{O}(1)$ operation, therefore, the combined demapping and symbol identification complexity remains $\mathcal{O}\left(q\sqrt{M}\right)$.

The approxLLR demappers overhead, combined with the single-FFT receiver architecture, contributes to a favourable complexity–performance trade-off while preserving the benefits of soft-decision demodulation. The complexity remains scalable and efficient, especially when using moderate modulation orders and layers. A comparative summary of conventional LACO-OFDM with iterative-receiver HD versus approxLLR SD with single-FFT receiver complexity is present in Table 2.

Table 2
Comparative summary

Component	Iterative-receiver HD	approxLLR SD with Single-FFT
IFFT/ FFT Complexity	$L2\mathcal{O}(N\ \log_2 N)$	$\mathcal{O}(N\ \log_2 N)$
Demapper Complexity	$\frac{\mathrm{N}}{2}\left(1-\frac{1}{2^{\mathrm{L}}}\right)\cdot\mathcal{O}(1)$	$\frac{\mathrm{N}}{2}\left(1-\frac{1}{2^{\mathrm{L}}}\right)\cdot\mathcal{O}\left(q\sqrt{M}\right)$
Total Complexity	$\mathcal{O}_{ConvHD} = L2\mathcal{O}(N\ \log_2 N) + \frac{\mathrm{N}}{2}\left(1-\frac{1}{2^{\mathrm{L}}}\right)\cdot\mathcal{O}(1)$	$\mathcal{O}_{approxLLR} = \mathcal{O}(N\ \log_2 N) + \frac{\mathrm{N}}{2}\left(1-\frac{1}{2^{\mathrm{L}}}\right)\cdot\mathcal{O}\left(q\sqrt{M}\right)$

Numerical Result

Figure 7 compares the conventional LACO-OFDM with iterative receiver against the proposed LACO-OFDM approxLLR SD with single-FFT receiver across varying QAM orders. At an SNR of 10 dB, the proposed method achieves performance gains of approximately 5×, 4×, and 3.9×, where × represents times, over HD demodulation for 16-QAM, 32-QAM, and 64-QAM, corresponding to order-of-magnitude reductions of 0.705, 0.615, and 0.591, respectively. Conversely, for 4-QAM, the method underperforms, yielding an improvement ratio of just 0.133 (≈ -0.877 order of magnitude). These findings indicate the effectiveness of the proposed approxLLR SD demodulation, which derives bit-wise likelihoods via probabilistic estimation rather than conventional map to the nearest-neighbour decisions performed by HD demodulation, which makes it more susceptible to noise-induced errors. By incorporating bit-wise likelihood information, the SD process can refine the estimation of each transmitted symbol with higher accuracy. The receiver is allowed to differentiate more effectively between closely spaced constellation points, particularly under noisy optical channel conditions. As illustrated in Figure 7, the proposed approach yields a performance gain of approximately ~4dB at a BER of 10^{-4} and about ~3dB at 10^{-5} for 64-QAM. These improvements demonstrate the robustness and efficiency of higher-order QAM formats.

Figure 8 depicts the BER performance of the proposed approach for various M-QAM modulation schemes, 4-QAM, 16-QAM, 32-QAM, and 64-QAM. The results show that 4-QAM has significant noise tolerance, sustaining a continuously low BER under relatively low SNR conditions. The observation demonstrates the practicality of the proposed approach by preserving transmission reliability even in high-noisy conditions. For comparison, 16-QAM and 32-QAM generally achieve a reasonable balance (trade-off) between SE and noise resistance, although they require higher SNR values to maintain acceptable BER performance. While 64-QAM, despite offering the highest SE, exhibits greater vulnerability to noise and demands a substantially higher SNR to ensure reliable data transmission. This sensitivity becomes more evident in LACO-OFDM structures, where asymmetric zero-clipping and superimposed layers introduce additional distortion to the received symbols. The result reinforces the fundamental trade-off in QAM modulation order: lower-order are more reliable in low-SNR, whereas higher-order modulation is suited for delivering high data rates under favourable signal conditions. Thus, selecting the right modulation order is crucial in attaining the optimal balance of SE and error performance.

Figure 9 presents the BER performance of the proposed approach on various QAM orders with different layer counts. The result shows that the 2-layer configuration achieves better BER performance, outperforms the 3-layer result due to less clipping distortion and ILI. Although the 3-layer structure provides higher SE, the additional layer introduces more interference and noise, leading to a noticeable deterioration in BER.

The finding highlights a trade-off in LACO-OFDM systems: increasing the layer count improves SE but adversely affects signal reliability. As a result, the 2-layer configuration works better for applications that require low BER and reliable communication, whereas the 3-layer layout is preferable when increasing SE is the main objective.

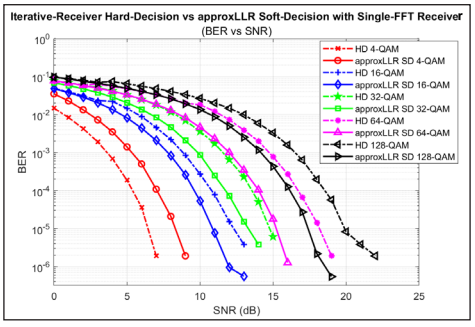


Figure 7. Iterative-receiver HD vs approxLLR SD with single-FFT receiver (different QAM)

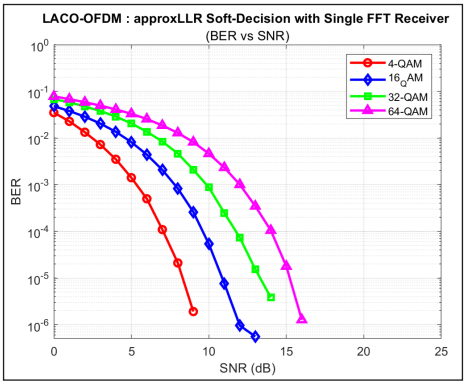


Figure 8. LACO-OFDM approxLLR SD with single-FFT receiver (different QAM)

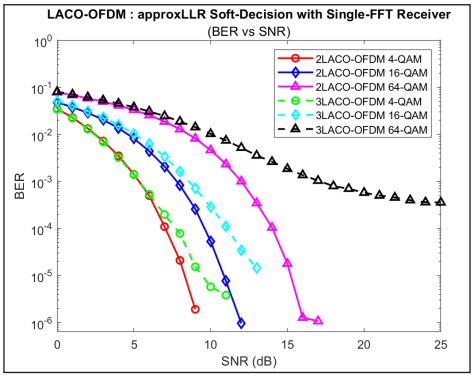


Figure 9. BER performance comparison with different numbers of layers of LACO-OFDM: approxLLR SD with single-FFT receiver

The performance of higher layers in LACO-OFDM is significantly impacted by clipping noise and interference from lower layers. While Layer 1 preserves relatively significant signal power (approximately half of the original due to clipping), higher layers suffer from additional amplitude reduction of superposition, diminishing their contribution to the received signal. Layers 2 and 3 require smaller scaling factors due to weaker amplitudes and higher interference sensitivity, making the conventional 2× scaling factor unsuitable for higher layers.

Under simulation configuration of 10 dB SNR, 30 dBm average received power, and 10% channel attenuation, Figure 10 presents the BER performance for 4-QAM, 16-QAM, and 64-QAM, with optimal scaling factors of 0.7, 0.3, and 0.1 for symbols in higher-layer detection, respectively. The findings show that implementing adaptive scaling factors in higher layers effectively reduces interference and improves BER performance. It contributes a practical trade-off between efficiency and reliability, resulting in a more resilient system.

A comprehensive study was conducted to investigate the relationship between received optical power and energy-per-bit-to-noise density (Eb/No) in a 2-layer LACO-OFDM system employing approxLLR SD demodulation with a single-FFT receiver. The simulation parameters specified in Table 1 were used to analyse the impact of different modulation orders, 4-QAM, 16-QAM, and 32-QAM, on power efficiency and BER performance, with a scaling factor of 0.1 applied to higher layers (layer ≥ 2).

Figure 11 demonstrates a proportional increase in received power with modulation order, indicating reduced noise tolerance in denser constellations. The minimum received powers required for reliable transmission are approximately 0.98 dBm, 7.96 dBm, and 10.98 dBm for 4-QAM, 16-QAM, and 32-QAM at Eb/No = 15dB. Due to the shorter distance between constellation points in higher-order QAM, more received power is required to preserve symbol detection under noisy conditions. The effect becomes more noticeable in OWC systems due to IM/DD constraints (non-negativity signal), which further intensifies the power requirements of higher-order modulation schemes.

Simulation results confirm that the modulation order has a strong influence on both power efficiency and error performance in the proposed approach. While 4-QAM offers excellent noise tolerance and power efficiency, higher-order schemes such as 32-QAM demand significantly more receive power and exhibit higher BER under identical SNR conditions.

Figure 12 demonstrates the received power (in dBm) as a function of Eb/No for a 16-QAM-based LACO-OFDM system under different layer count configurations. As expected, the received power decreases monotonically with increasing Eb/No due to reduced energy demand under improved noise conditions. At Eb/No = 15 dB, the measured received power values are 7.96 dBm (2-layer), 8.05 dBm (3-layer) and 8.08 dBm (4-layer).

These results indicate a slight increase in power consumption as the number of layers increases, attributed to the incremental impact of ILI and clipping noise accumulation. However, the magnitude of difference between the 2-layer and 4-layer systems is limited

to approximately 0.12 dB, indicating that the power overhead introduced by additional layers remains relatively small for moderate modulation orders. This highlights the power efficiency and robustness of the approxLLR SD with a single-FFT receiver scheme in supporting higher-layer configurations without incurring significant energy penalties.

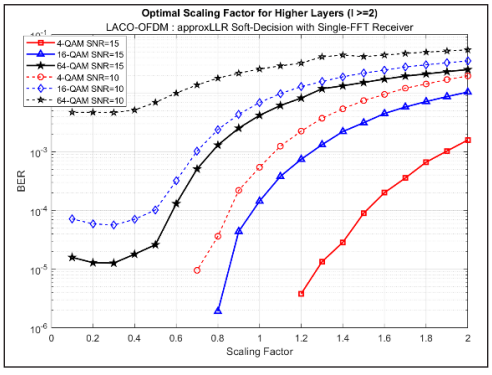


Figure 10. BER versus scale factor on the higher layer of LACO-OFDM approxLLR SD with single-FFT receiver

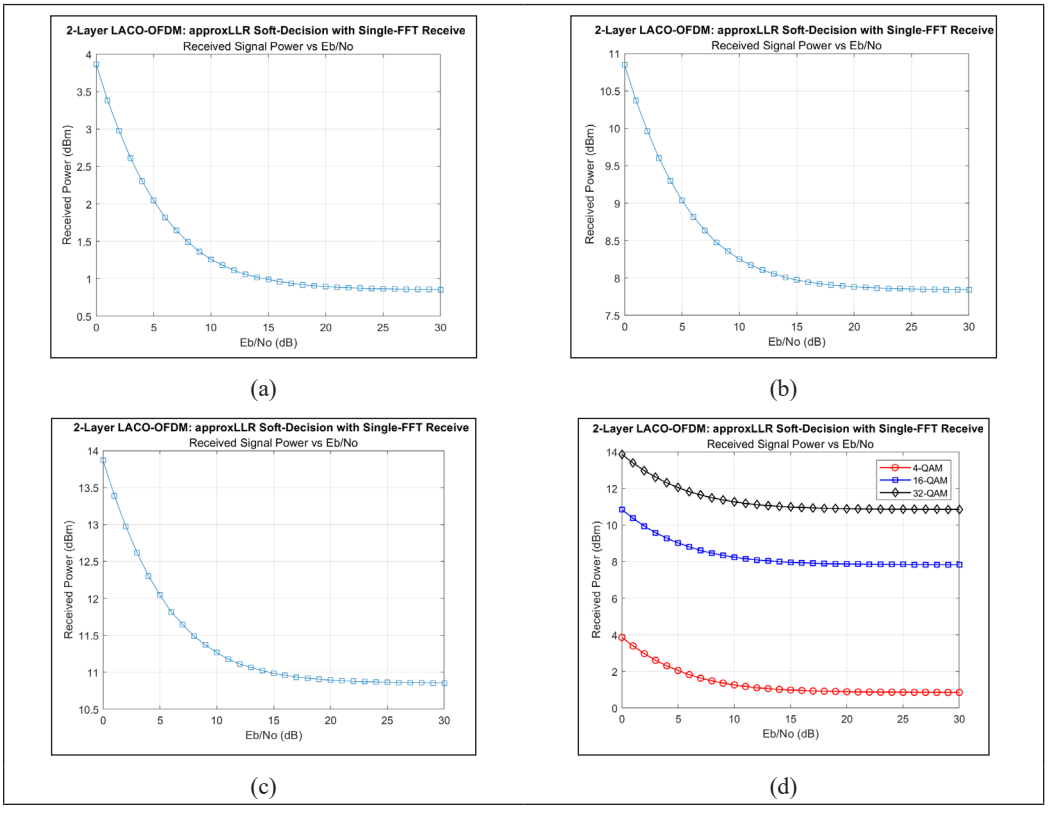


Figure 11. 2-layers of (a) 4-QAM; (b) 16-QAM; (c) 32-QAM; (d) different QAM LACO-OFDM approxLLR SD with single-FFT receiver of BER vs received signal power

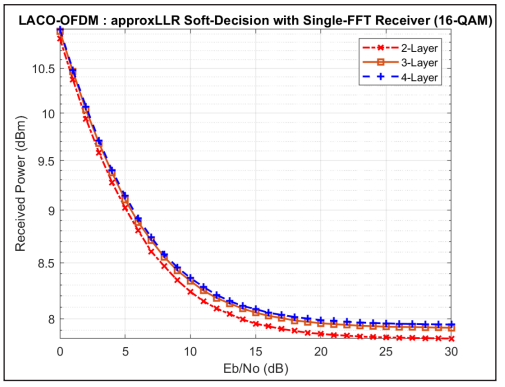


Figure 12. Received power comparison of different numbers of layers, LACO-OFDM approxLLR with 16-QAM

CONCLUSION

The current advancements in various OFDM schemes designed for IM/DD communication have been investigated, and a new theoretical approach of SD demodulation has been derived and validated using MATLAB simulation. The approxLLR SD demodulation has been proposed, integrated with a single-FFT at the receiver module to mitigate the limitation of the conventional LACO-OFDM modulation scheme. The computational bit-wise probabilities for $q = \log_2(M)$ bits, make this SD approach more resilient to channel noise (AWGN) and residual clipping distortion, resulting in a 4 to 5 times performance gain at a target BER of 10^{-4} over the conventional LACO-OFDM iterative-receiver scheme. The incorporation of approxLLR SD demodulation further improves noise-resilience, particularly in higher transmission layers where distortion effects become more pronounced. Enhanced SE is attained through a layered ACO-OFDM structure by enabling layer-by-layer symbol transmission over unused even subcarriers, allowing more effective utilisation of optical bandwidth. With the implementation of single-FFT, which does not require pairwise FFT/IFFT computation to distinguish the frequency-domain of clipping distortion, which makes this proposed approach is relatively simple and the complexity of $\mathcal{O}(N \log_2 N)$, leading to a 2 times reduction of computation can be achieved. Furthermore, the power analysis verifies that the proposed approxLLR SD with single-FFT receiver approach maintains high power and EE, exhibiting only a marginal increase in received power with increasing of layer count.

The performance comparison demonstrates that the proposed LACO-OFDM using approxLLR SD demodulation with a single-FFT receiver consistently outperforms the conventional LACO-OFDM iterative-receiver result while reducing receiver complexity. These results indicate that the proposed approach offers an effective trade-off between performance, SE and implementation simplicity, making it well-suited for VLC system practicality, which requires reliable high data transfer.

ACKNOWLEDGEMENT

The author would like to acknowledge the support from the Fundamental Research Grant Scheme (FRGS) under the grant number of FRGS/1/2022/TK07/UNIMAP/02/82 from the Ministry of Education Malaysia.

DATA AVAILABILITY STATEMENT

The data supporting the findings of this research are accessible upon request from the corresponding author.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest related to this work.

LIST OF ABBREVIATIONS

approxLLR	:	Approximate Log-Likelihood Ratio
AWGN	:	Additive white Gaussian noise
BER	:	Bit error rate
CP	:	Cyclic prefix
EE	:	Energy efficiency
FFT	:	Fast Fourier Transform
HD	:	Hard decision
HS	:	Hermitian symmetry
IFFT	:	Inverse Fast Fourier Transform
ILI	:	Inter-layer interference
IM/DD	:	Intensity modulation and direct detection
ISI	:	Inter-symbol interference
LACO-OFDM	:	Layered Asymmetrically Clipped Optical Orthogonal Frequency Division Multiplexing
LLR	:	Log-Likelihood Ratio
MAP	:	Maximum a posteriori
ML	:	Maximum likelihood
M-QAM	:	M-ary quadrature amplitude modulation
OFDM	:	Orthogonal Frequency Division Multiplexing
OWC	:	Optical wireless communication
PS	:	Parallel to serial
QAM	:	Quadrature amplitude modulation
SD	:	Soft decision
SE	:	Spectral efficiency
SNR	:	Signal-to-noise ratio
SP	:	Serial to parallel
VLC	:	Visible light communication

REFERENCES

Aliaberi, A., Sofotasios, P. C., & Muhaidat, S. (2019). Modulation schemes for visible light communications. In *Proceedings of the 2019 International Conference on Advanced Communication Technologies and Networking* (pp. 1-10). IEEE. https://doi.org/10.1109/COMMNET.2019.8742376

Anand, S., Ashish, & Kumar, P. (2024). A low-complexity LLR detector for FrFT-IM-OFDM in the presence of ICI. In *Proceedings of the 2024 27th International Symposium on Wireless Personal Multimedia Communications* (pp. 1-5). IEEE. https://doi.org/10.1109/WPMC63271.2024.10863056

Azim, A. W., Le Guennec, Y., Chafii, M., & Ros, L. (2021). LACO-OFDM with index modulation for optical wireless systems. *IEEE Wireless Communications Letters, 10*(3), 664-667. https://doi.org/10.1109/LWC.2020.3045119

Bai, R., Hranilovic, S., & Wang, Z. (2022). Low-complexity layered ACO-OFDM for power-efficient visible light communications. *IEEE Transactions on Green Communications and Networking, 6*(3), 1780-1792. https://doi.org/10.1109/TGCN.2022.3147970

Cao, X., Liu, Y., & Hu, D. (2019). Simplified LLR algorithm for M-QAM demodulation. *The Journal of Engineering, 2019*(21), 7370-7375. https://doi.org/10.1049/joe.2019.0634

Hu, Z., Chen, F., Wen, M., Ji, F., & Yu, H. (2018). Low-complexity LLR calculation for OFDM with index modulation. *IEEE Wireless Communications Letters, 7*(4), 618-621. https://doi.org/10.1109/LWC.2018.2802949

Hussein, Y. M., Mutlag, A. H., & Al-Nedawe, B. M. (2021). Comparisons of soft decision decoding algorithms based on an LDPC wireless communication system. *IOP Conference Series: Materials Science and Engineering, 1105*(1), Article 012039. https://doi.org/10.1088/1757-899X/1105/1/012039

Jativa, P. P., Azurdia-Meza, C. A., Canizares, M. R., Zabala-Blanco, D., & Montejo-Sanchez, S. (2020). Performance analysis of OFDM-based VLC schemes in NLOS channels. In *Proceedings of the 2020 South American Colloquium on Visible Light Communications* (pp. 1-6). IEEE. https://doi.org/10.1109/SACVLC50805.2020.9129862

Li, J., Liu, X., Ren, Y., & Huang, Z. (2021). Single-FFT receiver with pairwise maximum likelihood for layered ACO-OFDM. *IEEE Photonics Journal, 13*(4), 1-6. https://doi.org/10.1109/JPHOT.2021.3105812

Liu, G., Peng, M., Liu, C., Chen, M., Zhou, H., & Chen, Q. (2018). Receiver sensitivity improvement of a layered ACO-OFDM system enabled by layered-DFT/OCT precoding technique. In *Proceedings of the 2018 Asia Communications and Photonics Conference* (pp. 1-3). IEEE. https://doi.org/10.1109/ACP.2018.8595736

Liu, X., Li, J., Li, J., & Huang, Z. (2020). Analysis of the single-FFT receiver for layered ACO-OFDM in visible light communications. *Journal of Lightwave Technology, 38*(17), 4757-4764. https://doi.org/10.1109/JLT.2020.2994633

Lowery, A. J. (2020). Spectrally efficient optical orthogonal frequency division multiplexing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 378*(2169), Article 20190180. https://doi.org/10.1098/rsta.2019.0180

Mao, D., Chen, C., & Bai, B. (2020). An improved simplified soft demodulation algorithm for 64QAM signal. In *Proceedings of the 2020 IEEE 3rd International Conference on Electronic Information and Communication Technology* (pp. 157-160). IEEE. https://doi.org/10.1109/ICEICT51264.2020.9334307

Nakano, H., Tamura, K., Ueno, H., Takemoto, R., Cha, J., & Ahn, C. (2023). Each layer adaptation for throughput enhancement of LACO-OFDM. In *Proceedings of the 2023 International Conference on Advanced Technologies for Communications* (pp. 202-206). IEEE. https://doi.org/10.1109/ATC58710.2023.10318938

Olivatto, V. B., Lopes, R. R., & de Lima, E. R. (2017). Simplified method for log-likelihood ratio approximation in high-order modulations based on the Voronoi decomposition. *IEEE Transactions on Broadcasting, 63*(3), 583-589. https://doi.org/10.1109/TBC.2017.2704424

Othman, N., Beson, M. R., Aljunid, S. A., & Endut, R. (2024). Review: Outdoor visible light communication on modulation and receiver denoising scheme. *AIP Conference Proceedings, 3050*(1), Article 030038. https://doi.org/10.1063/5.0195503

Seow, Y. L., Rashidi, C. B., Aljunid, S. A., Ali, N., & Endut, R. (2021). 20 Mb/s experimental demonstration using modulated 460 nm blue LED for underwater wireless optical communications (UOWC). *Journal of Physics: Conference Series, 1878*(1), Article 012069. https://doi.org/10.1088/1742-6596/1878/1/012069

Vappangi, S., Mani, V. V., & Sellathurai, M. (2021). *Visible light communication: A comprehensive theory and applications with MATLAB®*. CRC Press.

Wang, C., Wang, X., Jiang, K., & Zou, J. (2023). Viterbi soft demodulation algorithm for CPFSK systems. In *Proceedings of the 2023 8th International Conference on Communication, Image and Signal Processing* (pp. 522-527). IEEE. https://doi.org/10.1109/CCISP59915.2023.10355684

Wang, Q., Qian, C., Guo, X., Wang, Z., Cunningham, D. G., & White, I. H. (2015). Layered ACO-OFDM for intensity-modulated direct detection optical wireless transmission. *Optics Express, 23*(9), 12382-12393. https://doi.org/10.1364/OE.23.012382

Yan, X., He, J., Shi, J., Zhou, Z., & Tang, Q. (2020). Time-efficient adaptive modulation scheme for LACO-OFDM in VLC systems. *Optics Communications, 474*, Article 126069. https://doi.org/10.1016/j.optcom.2020.126069

Zhang, C., Baharudin, M. A., Guo, T., Wang, J., & Jia, Z. (2025). A survey on the latest developments of OFDM schemes for optimising the capacity and reliability of visible light communication (VLC) systems. *KSII Transactions on Internet and Information Systems, 19*(5), 1625-1647. https://doi.org/10.3837/tiis.2025.05.012

Zhang, T., Sun, L., Zhao, C., Qiao, S., & Ghassemlooy, Z. (2021). Low-complexity receiver for HACO-OFDM in optical wireless communications. *IEEE Wireless Communications Letters, 10*(3), 572-575. https://doi.org/10.1109/LWC.2020.3038181

Zhang, X., Babar, Z., Zhang, R., Chen, S., & Hanzo, L. (2019). Multi-class coded layered asymmetrically clipped optical OFDM. *IEEE Transactions on Communications, 67*(1), 578-589. https://doi.org/10.1109/TCOMM.2018.2869821

Zhang, X., Chen, S., & Hanzo, L. (2020). On the discrete-input continuous-output memoryless channel capacity of layered ACO-OFDM. *Journal of Lightwave Technology, 38*(18), 4955-4968. https://doi.org/10.1109/JLT.2020.2996541

Zhang, X., Wang, Q., Zhang, R., Chen, S., & Hanzo, L. (2017). Performance analysis of layered ACO-OFDM. *IEEE Access, 5*, 18366-18381. https://doi.org/10.1109/ACCESS.2017.2748057

Pertanika J. Sci. & Technol. 34 (3): 1579 - 1598 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Review Article

A Bibliometric Review of Remote Sensing Applications in Water Surge Risk Mapping for Waterfall Areas

Mohammed M. Y. Albarqouni[1], Syarifuddin Misbari[1]*, Nadiatul Adilah Ahmad Abdul Ghani[1], and Mohd Faizal Jamlos[2]

[1]Faculty of Civil Engineering Technology, Universiti Malaysia Pahang, Al-Sultan Abdullah, UMPSA Gambang, Lebuh Persiaran Tun Khalil Yaakob, 26300 Kuantan, Pahang, Malaysia
[2]Faculty of Electrical and Electronics Engineering Technology, Universiti Malaysia Pahang, Al-Sultan Abdullah, Pekan Campus, 26600 Pekan, Pahang, Malaysia

ABSTRACT

Waterfall areas are now in the gold rush era in research in different fields, including ecology and hydrology. There is a critical lack of studies on using remote sensing to detect and map water surges in waterfall areas, particularly in tropical regions, including Malaysia. Thus, this study sheds light on exploring the key research trends on remote sensing applications to detect and map water surges in waterfall areas globally in the past five years, providing insights to guide future research efforts in this critical but underrepresented field. Bibliometric analysis included defining search criteria and keywords, retrieving relevant bibliographic records from the Web of Science database, and conducting a comprehensive analysis of the data among 667 peer-reviewed articles was revealed that there is an immense need to investigate water surge risk in waterfall areas simultaneously as a broader environmental application. Critical visualisation was provided by generated network maps using VOSviewer. The study reveals a global and regional lack of research on using remote sensing to map water surges in waterfall areas. This work emphasises the critical need for targeted research in this niche field, offering actionable insights for advancing Sustainable Development Goals (SDGs) and disaster preparedness in regions prone to water surges.

Keywords: bibliometric analysis, hydrology, remote sensing, water surge risk, waterfall

ARTICLE INFO

Article history:
Received: 18 September 2025
Accepted: 01 May 2026
Published: 19 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.13

E-mail addresses:
mo.albarqouni@gmail.com (Mohammed M. Y. Albarqouni)
syarifuddinm@umpsa.edu.my (Syarifuddin Misbari)
nadiatul@umpsa.edu.my (Nadiatul Adilah Ahmad Abdul Ghani)
faizaljamlos@umpsa.edu.my (Mohd Faizal Jamlos)
* Corresponding author

INTRODUCTION

Freshwater supplies are the most valuable resource for sustaining life on earth; freshwater comprises rivers, snow, waterfalls, underground aquifers and all

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

other bodies of surface water. These sources of freshwater are subject to an increasingly large amount of stress throughout the world because they have both spatial and temporal disparities (Bănăduc et al., 2022). There is also relatively little general knowledge about the availability of freshwater resources in various areas, especially those where there are waterfalls and/or limited amounts of deep groundwater reserves. As a result of a lack of knowledge about the quality and quantity of available freshwater resources during periods of extreme water scarcity and/or abundance may cause significant negative impacts from drought to excessive water surges/flooding. Often, the hydrological extremes that occur will have devastating impacts upon both local and regional communities, resulting in substantial harm to the environmental, economic and social structures of a community (Wang & Xie, 2018). Therefore, it is crucially important to adequately monitor and manage the risk of water surges and droughts. Advances in technology, including remote sensing, drone photography, geographic information systems (GIS) and hydrologic modelling, enable the more accurate monitoring and management of this type of issue.

Flooding as well as water surges in water bodies may lead to structural damage, crop loss, and loss of life (Abdel-Mooty et al., 2021; Nofal & van de Lindt, 2021). Due to the rapid drop in elevation and increase in water speed, water surges in waterfall regions are particularly hazardous. Global statistics show that water surges and flood-related occurrences are having a growing influence around the world. Between 1970 and 2019, floods were responsible for 44% of all disasters caused by weather worldwide, with Asia experiencing the highest proportion, 31% of these events resulting in nearly a million deaths and over $2 trillion in economic losses (World Meteorological Organisation, 2021). In Africa, while floods constituted 60% of weather-related disasters, they caused fewer fatalities compared to droughts. In North and Central America, floods represented 31% of recorded disasters (International Conference on Flood Management, 2022; Climate and Weather-Related Disasters Surge, 2024). Between 2021 and 2023, Malaysia had several devastating water surge disasters that caused the loss of many lives and the destruction of many facilities and infrastructure, emphasising the critical need for better knowledge and control of such phenomena (Table 1). In 2022, incidents included the rescue of 80 visitors at Laman Tiga Budaya waterfall in Melaka on May 16, and the deaths of three individuals, including a pregnant woman, in Baling, Kedah, on July 4 (Figure 1).

Remote sensing plays a pivotal role in water resource mapping by delivering essential data. Satellite-based remotesensing systems could deliver continuous, real-time data with global coverage and exploit their orbital capabilities, also relying on ground inspections for algorithm validation (Wang & Xie, 2018). Additionally, remote sensing technologies have been proven to be highly effective tools in improving many key functions relative to water surge risks, such as mapping accurately affected areas and identifying potential surge incidents before their occurrence (Iqbal et al., 2022; Munawar et al., 2021).

Table 1
Water surge events in Malaysia from 2021 to 2023

Event Name	Date	Location
Gunung Jerai Surge	August 18, 2021	Yan and Tupah, Kedah
Laman Tiga Budaya Waterfall Surge	May 16, 2022	Asahan, Melaka
Baling Surge	July 4, 2022	Baling, Kedah
Lata Hulu Kekabu Surge	November 4, 2022	Setiu, Terengganu
Sungai Pinang waterfall Surge	November 28, 2022	Batang Kali, Selangor
Bukit Tinggi and Janda Baik Surge	February 12, 2023	Pahang
Kampung Asli Pertak Waterfall Surge	June 21, 2023	Kuala Kubu Bharu, Selangor
Jeram Air Putih Surge	July 1, 2023	Kemaman, Terengganu

Note. Adapted from (Selangor Journal, 2024; Sinar Daily, 2023)

Figure 1. Water surge impact on 12 villages in Baling, Malaysia (Malay Mail Newspaper, 2024)

To enhance monitoring and prediction capabilities that are essential in water surge risk management, remote sensing technologies have increasingly been applied in water surge risk mapping of waterfalls. Applications include using satellite imagery to monitor water levels, LiDAR to evaluate terrain and elevation, and vegetation cover analysis to determine erosion susceptibility (Parra, 2022; Trepekli et al., 2022). Further, climate data from satellites and radar are significant to understand precipitation patterns that could potentially signal future surge events (Sahoo & Bhaskaran, 2019), and predictive models can be developed for the forecast of flow rates and identification of high-risk

areas using an integration of this data with both historical and meteorological data, thus creating risk maps used to direct emergency preparedness through GIS system integration (Alaghmand et al., 2010).

Using previous in-situ measurement techniques, which required direct measurements of discharge in waterfall areas, new techniques for estimating discharge have been developed using remote sensing technology and now include the use of synthetic width or stage discharge evaluation curves. The new methodologies use remote sensing from satellites to assess the river's efficient wet width and water surface elevation in order to calculate flow rates. In addition to being able to monitor river discharge at relatively frequent intervals using data which are generally accurate and reliable, these new methodologies can provide this information for large regions that have historically been difficult to reach because of the resolution of the satellite images (e.g. MODIS (Moderate Resolution Imaging Spectroradiometer), ENVISAT (Environmental Satellite), Landsat, etc.) (King et al., 2018; Tourian et al., 2013).

Remote sensing techniques are increasingly being applied in hydrological research due to their growing popularity. The remote sensing technique can provide a high resolution of the spatial-temporal distribution of different types of water resources, as well as assess the operational status of hydraulic infrastructure. These remote sensing methods can also help with monitoring and assessment of droughts, detection of flood water surges, and collection of important information regarding sustainable development and disaster prevention of water resource management.

Water Surge Review Through Bibliometric Analysis

Bibliometric research is the quantitative data collection and analysis used to identify trends in how academia produces, disseminates, and receives published works using statistical/ mathematical methods (Danvila-del-Valle et al., 2019; Diodato, 2013; Donthu et al., 2021). Bibliometric research can be viewed as a study of the production, transmission, and reception of scientific knowledge through the analysis of bibliographic data from each article, including citations, author(s), institution(s), geography, and journal. Examples of how bibliometric research can be applied are to build intellectual networks among researchers, assess the contributions of individual researchers, organisations, or countries, and analyse the research trends across different disciplines (Donthu et al., 2021). The main goal of bibliometric research is to provide a systematic/ objective analysis of the status of a particular field of study.

Researchers can measure the impact of scholarly work on this topic and the relative quality of that work based on how often it is referenced by other peer authors in assessing the effectiveness of bibliometric studies (Danvila-del-Valle et al., 2019; Donthu et al., 2021; van Nunen et al., 2018) and also to determine if these studies have been used by

other researchers in evaluating the quantity of the literature and quality of the literature. This study will use bibliometric studies to find the most influential studies on waterfall flooding in tropical areas like Malaysia and the opportunities for future collaboration to identify current trends in research and provide methods and techniques to monitor and reduce the risk of flooding due to waterfalls because of climate change.

The main benefit of this type of analysis provides researchers, as well as funding agencies, with the ability to make informed decisions on resource allocation based upon the relative strengths and weaknesses in a particular field of research. A second major benefit of the bibliometric analysis process is to assist researchers in understanding the current trends and growth of innovation in a specific field by analysing the trend of publications and citations throughout a given time frame. Bibliometric analysis also offers a way to measure the performance of individual authors and institutions. Through the measurement of citations of an author's works, where those works were published, as well as how geographically widespread those citations were, bibliometrics provide a researcher with a global view of how the scholarly community evolves, and the extent to which international collaboration occurs among academics around the world. The results of a bibliometric study will allow researchers to assess their own scholarly productivity and potentially identify potential future collaborative research opportunities with other scholars in their field of study (Danvila-del-Valle et al., 2019; Diodato, 2013).

VOSviewer (van Eck & Waltman, 2007), BibExcel (Persson et al., 2009), and Bibliometrix (Aria & Cuccurullo, 2017) have become the most widely used applications in the field of bibliometrics by scholars around the world (Garrigos-Simon et al., 2018; Goerlandt et al., 2020). The purpose of this study is to evaluate all existing bibliographic data that pertains to the use of remote sensing technologies for assessing hazards resulting from increased water flow/levels, and for assessing the risk of falling water from waterfalls, due to the importance and urgent need for these assessments. This study will provide an overview of the current state of research in this field and establish a foundation for additional research and knowledge development.

This Bibliometric study accomplished its main objective via a science-mapping analysis, which analysed Co-Citation, Co-Authorship, and Co-Occurrences of Keywords, and also Performance Analysis in the form of Publication Metrics, Citation Statistics, and Identification of Top Publishing Nations using the bibliographic information contained within the WoS database (Alryalat et al., 2019; de Carvalho Pereira et al., 2015). The importance of this study aligns with national goals, global Sustainable Development Goals (SDGs), Environmental, Social, and Governance (ESG) principles and contributes to sustainable resource management and policymaking. By enhancing knowledge of water surge risks in tropical waterfall regions, this research helps fill a critical gap in disaster preparedness measures within the national climate change policy and supports Malaysia's

national efforts to reduce disaster risks. In line with Malaysia's Twelfth Plan (MyGOV, n.d.), this research also supports ecotourism and improves resource management, both of which contribute to sustainable development planning.

The study supports the UN Sustainable Development Goal 6: Clean Water and Sanitation globally by promoting sustainable water resource management and risk mitigation strategies (https://sdgs.un.org/goals/goal6). It also supports the UN Sustainable Development Goal 13: Climate Action by using remote sensing technologies to address climate-related risks (https://sdgs.un.org/goals/goal13), and the UN Sustainable Development Goal 11: Sustainable Cities and Communities by enhancing the resilience of communities near waterfalls (https://sdgs.un.org/goals/goal11). Furthermore, the research supports the principles of environmental governance and good governance (ESG) through data-driven governance in managing water scarcity risks and promoting environmental stewardship through ecosystem conservation. These contributions highlight the study's broader impact on policy innovation and sustainable development.

METHODOLOGY

The principal gaps in using remote-sensing technology for detecting and mapping flood occurrences within watershed regions were identified using a combined analytical method, which consisted of a systematic literature review (SLR) and a bibliometric analysis. The purpose of conducting the SLR first was to completely and impartially evaluate the current body of knowledge to identify and examine all relevant research studies that studied remote-sensing technologies in relation to water-based systems by applying pre-determined inclusion/exclusion criteria. In addition to identifying the thematic content of the articles reviewed, this method allowed us to include in our review those studies that were relevant to our research objectives, thereby allowing us to identify trends and gaps in research that will be useful for the design of future studies. Additionally, we believe that the combination of these two methods will allow us to contribute to the achievement of the Sustainable Development Goals (SDGs) through the identification of regional and global shortcomings in research concerning the use of remote sensing technology to measure the level of water in water catchment regions. Finally, the SLR and Bibliometric Analysis methodologies used in this study will provide a solid foundation for future research in this rapidly developing field, as they will allow us to provide a comprehensive and data-based perspective on the depth of the knowledge base in this subject area.

A bibliometric analysis of the type used in this research employs a consistent three-step procedure (Figure 2), which is as follows: (a) keywords and search queries were defined, (b) searches on databases were conducted and data exported, and (c) the bibliometric analysis was performed. Stage I is presented with more details in the following.

To identify the relevant literature from the database, the first step was to select the appropriate keywords for a search query that would capture the subject area of the study. Therefore, the search query *((“water surge” OR “waterfall areas”) AND (“remote sensing”) AND (“water surge risk mapping”))* was used.

Stage II: The above-mentioned keywords were used to carry out an exhaustive search in a chosen database using the stated keyword query. Many databases contain scientific or technical information; however, the WoS contains a large amount of literature from various areas of science and technology, which provides powerful citation-based measures and tools to study relationships among articles, authors and institutions. Although both WoS and Scopus have generally been found useful for bibliometric studies, they were both the best choices available at the time of writing for the purposes of studying high-quality, peer-reviewed literature that would allow for the determination of the most relevant trends in the area of study.

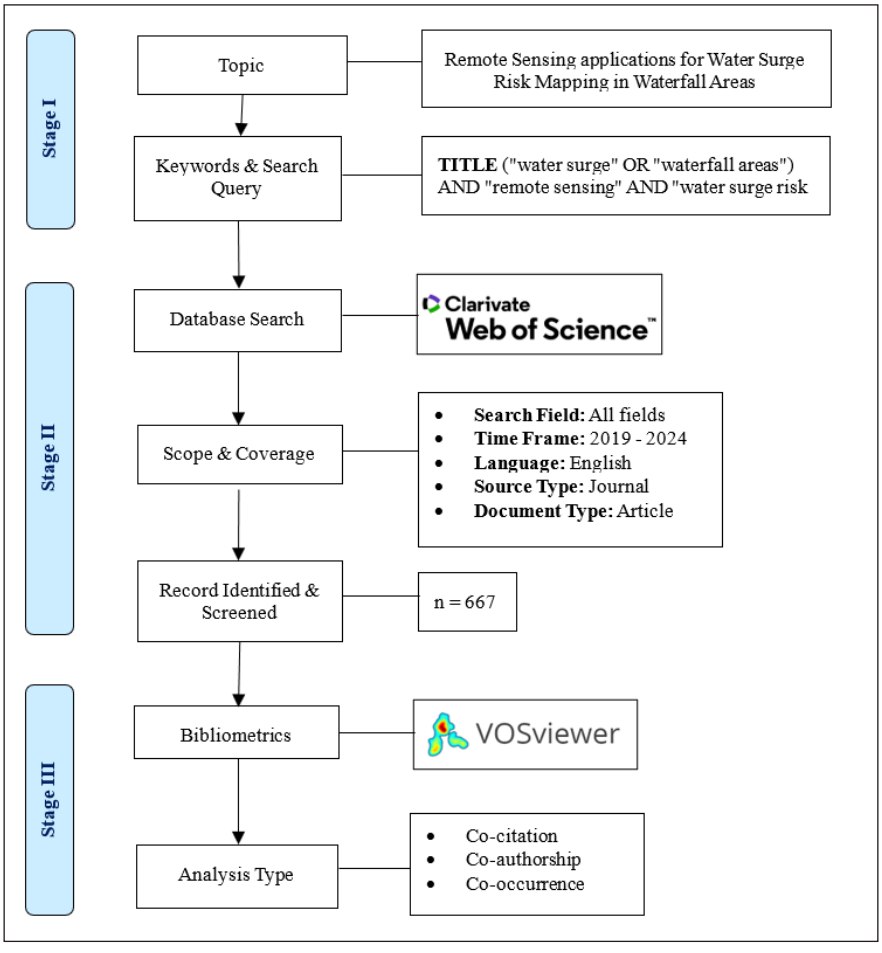


Figure 2. A diagram illustrating the search strategy

Filters were then applied to limit the results of searches made by performing a keyword query to the WoS to the results based on pre-defined criteria. These criteria included categories limited to environmental science and distance education, English as the language of publication, a publication period between 2019 and 2024 (covering the preceding five years), and document formats limited to articles and review articles. The records were then exported for bibliometric analysis.

Stage III: Statistical and graphical analysis was performed on bibliometric records exported from the WoS database. Science maps, which include patterns of co-citation and co-authorship, and performance analysis, which includes metrics such as publication statistics, number of citations, and top publishing countries, were used to present the results. The purpose of these analyses was to identify significant bibliometric patterns in the database. Exported data was used for performance analysis, while Microsoft Excel was used for processing and visualisation. VOSviewer (version 1.6.20), a program well-known in modern bibliometric studies for producing excellent visualisations, can also generate the network maps. The program's default settings were used for the generated maps, with minor adjustments to the threshold settings to improve visual clarity.

RESULTS AND DISCUSSION

Although significant knowledge gaps and trends exist in terms of the risk of water surges from previous research that used a systematic literature review (SLR) and bibliometric analysis, there is a severe lack of previous remote sensing research specifically in tropical waterfall regions such as Malaysia to date, which was also supported by findings from the SLR. Although some of the studies using remote sensing to map water surge in tropical waterfall regions were limited, SLR indicated that many previous research studies had significant shortcomings. Most of the previous research has focused primarily on general hydrology and/or other environmental issues, with most studies utilising traditional ground-based measurement systems instead of advancing with new remote sensing technologies. The study clearly illustrated a pressing need to focus additional research on bridging this technological gap and enhancing community resiliency in vulnerable climates.

Performance Analysis

This is a part of the research on biometrics that contains quantitative tendencies of the bibliography. Trends are identified as follows: yearly number of publications; yearly number of citations; the most important studies; the most important publishing centres; and the most important country contributions.

Publications and Citations

As of November 1, 2024, 667 bibliographic records had been found because of an improved search query made possible by the WoS database. Due to the similar patterns in the Scopus database, there were roughly 2260 publications for the same search criteria. The total number of publications (TP) for remote sensing for water surge risk mapping in waterfall areas increased between 2019 and 2024, reaching a peak of 135 in 2022 before slightly decreasing to 129 in 2024 (Figure 3). The total citations (TC), which peaked at 3142 in 2021, exhibit a different pattern. This discrepancy suggests an expanding corpus of work with waning scholarly influence, perhaps because of a lack of creativity. Notably, studies focusing on water surge risks in waterfall areas are particularly scarce, highlighting an underexplored niche within this domain.

Table 2 lists the top ten most cited papers that discuss environmental research and remote sensing technology to highlight the most important topics and authors. It is also evident that, for at least the last five years, there have been very few scientific studies on the dangers of rushing water in waterfall locations, and none of them has been included in the top ten most referenced publications worldwide. The comparatively low number of citations indicates that the use of remote sensing technology to manage water surge risk, particularly in waterfall areas, is still underexplored. This shows that there is enormous, untapped potential for using these technologies to improve risk assessment, early warning systems, and management techniques in such domains. Expanding research efforts in this area may lead to more inventive solutions and a better knowledge of how remote sensing might reduce the hazards associated with water surges in complex and dynamic environments.

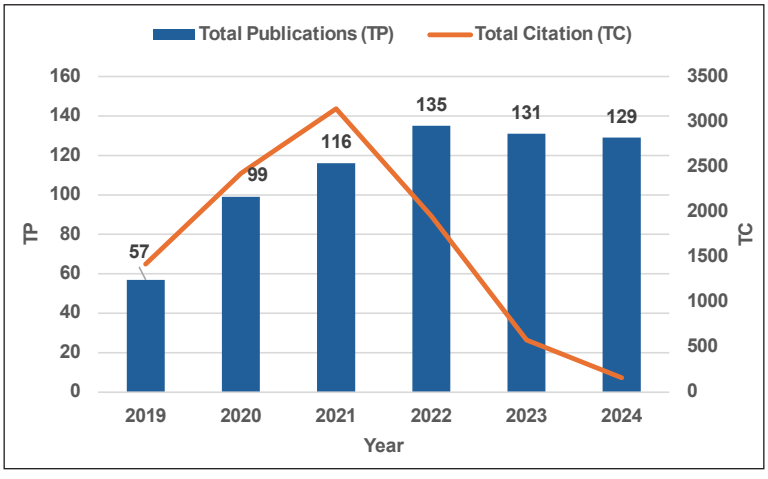


Figure 3. Trends in the number of publications and citations

Table 2
Top 10 publications with the most citations on the use of remote sensing in environmental Sciences

#	Article Title	Authors	No. of Citation
1	Antimicrobials and Antibiotic-Resistant Bacteria: A Risk to the Environment and to Public Health	(Serwecińska, 2020)	383
2	Effect of microplastics in water and aquatic systems	(Issac & Kandasubramanian, 2021)	364
3	Environmental impacts, pollution sources and pathways of spent lithium-ion batteries	(Mrozik et al., 2021)	338
4	Large-scale hydrogen production via water electrolysis: a techno-economic and environmental assessment	(Terlouw et al., 2022)	240
5	Combined Modelling of US Fluvial, Pluvial, and Coastal Flood Hazard Under Current and Future Climates	(Bates et al., 2021)	198
6	Mapping Industrial Symbiosis Development in Europe_ typologies of networks, characteristics, performance and contribution to the Circular Economy	(Domenech et al., 2019)	188
7	Catching a resurgence: Increase in SARS-CoV-2 viral RNA identified in wastewater 48 h before COVID-19 clinical tests and 96 h before hospitalisations	(D'Aoust et al., 2021)	139
8	A High-Resolution Global Dataset of Extreme Sea Levels, Tides, and Storm Surges, Including Future Projections	(Muis et al., 2020)	136
9	Phytoremediation of heavy metals in soil and water: An eco-friendly, sustainable and multidisciplinary approach	(Bhat et al., 2022)	114
10	Assessing Compound Flooding from Landfalling Tropical Cyclones on the North Carolina Coast	(Gori et al., 2020)	104

Top Publishing Countries

Geographical statistics on publications serve as important performance indicators in bibliometric analysis. Table 3 presents a list of countries with the highest number of publications related to the application of remote sensing in environmental sciences. The People's Republic of China tops the list with 197 publications, followed by the United States of America (USA) as the second-largest contributor with 182 publications. While Malaysia ranked 33rd among countries, with 7 scientific consultations in the last 5 years, knowing that these publications also do not discuss the issue of water surge risk mapping in the waterfalls area.

Table 3
The top 10 countries and Malaysia with the most publications

Rank	Country	Number of Publications
1	People's Republic of China	197
2	United States of America (USA)	182
3	England	63
4	Italy	47
5	Germany	42
6	Australia	36
7	Canada	35
8	Spain	33
9	India	32
10	Netherlands	30
33	Malaysia	7

Top Publishing Organisations

Table 4 provides an overview of the top institutions and influential research groups actively contributing to research on remote sensing applications in water surge risk mapping. The Chinese Academy of Sciences and Hohai University are recognised as the leading organisations, with the most publications in this specialist field. Despite their high ranking, the overall number of publications from these universities is still quite low, which reflects the small amount of research that the academic community has produced in this field. up order to improve the research landscape and fill up the current knowledge gaps regarding the application of remote sensing technologies to water surge risk mapping in waterfall areas, this discovery highlights the necessity of expanded efforts and wider collaboration.

Table 4
The ten organisations with the most publications

Rank	Organisation	Number of Publications
1	Chinese Academy of Sciences	40
2	Hohai University	28
3	University of Chinese Academy of Sciences	23
4	Ministry of Natural Resources	14
5	Sun Yat-Sen University	12
6	U.S. Geological Survey (USGS)	12
7	Wuhan University	11
8	Louisiana State University	10
9	Delft University of Technology	9
10	NOAA	9

Science Mapping

The VOSviewer tool is used to create network maps that depict the bibliometric linkages. Network maps, density maps, and overlay maps are just a few of the visuals that this tool provides (van Eck & Waltman, 2007; van Eck & Waltman, 2010). In our investigation, we will concentrate on network maps. Nodes and links are used to create network maps. Each node represents an entity, such as an author, nation, document, institution, or keyword, and the connections show the co-occurrence relationships between these entities. While the thickness of the link reflects the strength of the relationship, the size of the node conveys the importance or prominence of an entity. The degree of association between nodes is reflected in their closeness; stronger relationships are represented by closer nodes. To categorise things into discrete clusters and provide a clear visual depiction of their relationships and grouping patterns, the nodes are also colour coded.

Co-citations

Co-citation analysis is a science mapping method based on the premise that, according to the referred works, publications that are frequently mentioned together have comparable themes. When two publications appear in the reference list of a third publication, they are connected inside a co-citation network. This approach provides insightful information about the conceptual structure and thematic relationships within a subject of study (van Eck & Waltman, 2010). Co-citation-based network maps were used to show the thematic commonalities among texts in the research field. As seen in Figure 4, which displays the co-citation network map for writers, these maps examine the literature by looking at how frequently papers are cited together. The dataset was reduced to 53 writers out of 28,116 by using a minimum citation criterion of 25 for all authors to create this map.

Co-occurrence

The keyword co-occurrence network map, which is shown in Figure 5, depicts the relationships between distinct author-defined terms based on their frequency of appearing together. To create this map, a minimum co-occurrence threshold of 5 was applied, resulting in the inclusion of 51 keywords from a total of 2,634 author-defined terms. The keyword “storm surge” has been referenced in the literature under various terms and emerged as the most frequently occurring keyword, with 85 mentions and a total link strength of 93. Other notable keywords include “sea level rise” (34 occurrences), “coastal flooding” (33 occurrences), and “remote sensing” (18 occurrences).

Moreover, it is important to stress that the current scientific literature does not include or has barely covered terms that are unique to waterfall zones when referring to water surge events.

This finding demonstrates an existing gap within the literature; therefore, it can be inferred that the academic community has not adequately emphasised the aspect of water surge dynamics in its recent publications.

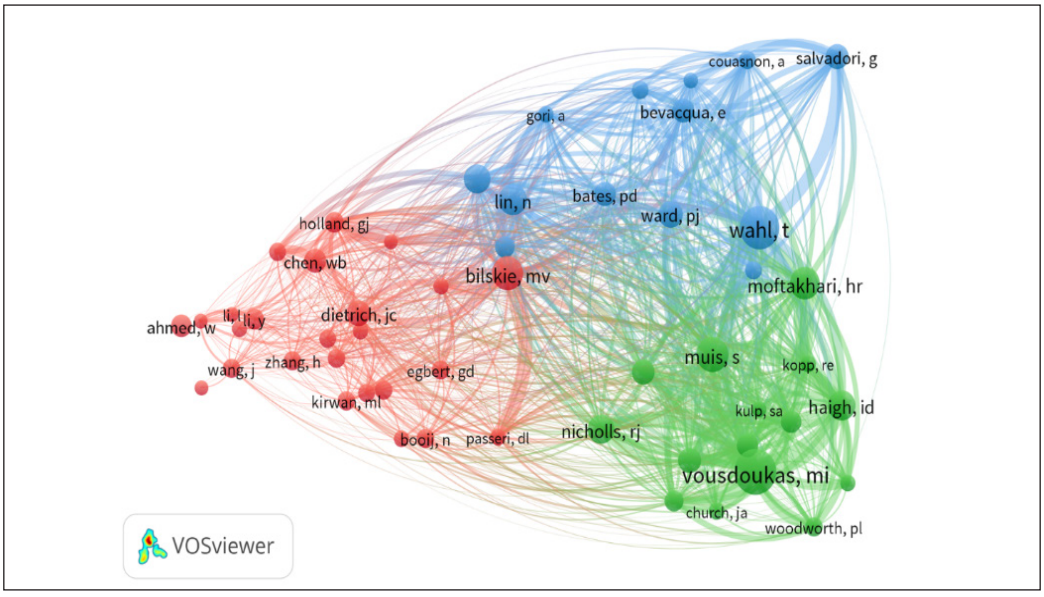


Figure 4. Network map of co-citations among authors

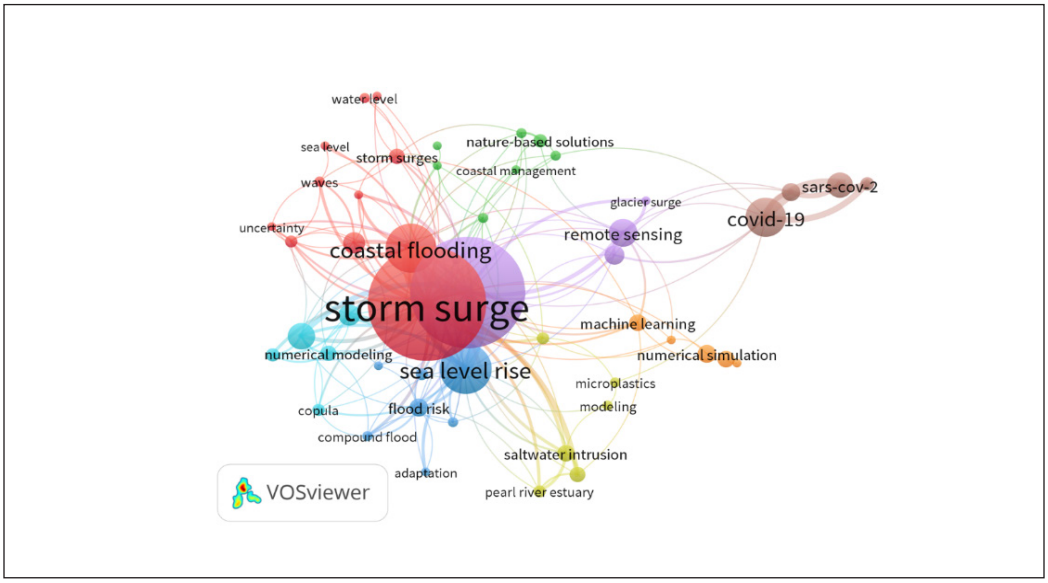


Figure 5. Network map of co-occurrence cited by author keywords

Co-authorship

Co-authorship relationships are visualised in a geographic mapping by nation (as illustrated in Figure 6) that demonstrates the extent of international collaboration between researchers. Those nations that exhibit the greatest number of collaborations among one another, such as China, the United States, England and Australia, have both the highest publication output and are also most actively engaged in international collaborative research. The strength of those alliances is indicated by the width of the connections between nations. Thus, the minimal criterion for map generation was established at 5 publications per country, giving a total of 39 countries from 86. Countries are grouped into distinct clusters, such as the Asia-Pacific-North America region (e.g., China, USA, Australia, Japan), which is observed with the red cluster, Western Europe (e.g., England, Austria, France) with the green cluster, Europe-Asia (e.g., Italy, Poland, Saudi-Arabia, Malaysia) with the blue cluster, Europe-South America (e.g., Spain, Greece, Colombia, Brazil) with the yellow cluster, and Canada, Switzerland and South Africa with purple cluster, demonstrating regional collaboration patterns. Notably, China stands out as the leading contributor with the highest publication output and centrality in the co-authorship network. While the USA and England have the strongest relationship, with a link strength of 30.

The co-authorship map highlights the collaborative relationships among organisations (Figure 7) researching remote sensing applications within environmental science. A minimum threshold of 5 documents per organisation was established, narrowing the analysis to 37 institutions out of a total of 1,267. The Chinese Academy of Sciences and Hohai University are dominating, with the largest nodes denoting their major roles and considerable contributions. Strong collaborative links are seen between Chinese institutions, producing dense networks, especially within the Blue Cluster, which includes Northwest Normal University.

The U.S. Geological Survey, which serves as a focal point in the red group and represents foreign contributions with lesser but still significant collaboration, is another less noteworthy contributor. Regional-focused partnerships can be seen in many of the other groups also, i.e., groups with involvement from the Ministry of Water Resources and European organisations, including Utrecht University and Delft University of Technology. This is due to China being the dominant group on the map, indicating its extensive global presence; however, there are still some examples of foreign collaborations which appear to be less connected than they could be, thus suggesting a possible opportunity to develop more complete, global research networks.

This study provides a multi-dimensional view of the research landscape using both SLR and bibliometric findings. The bibliometric approach was able to quantify these gaps and to place the results in the context of larger trends in research; however, it is the qualitative findings from the SLR that identified methodological limitations and topical gaps.

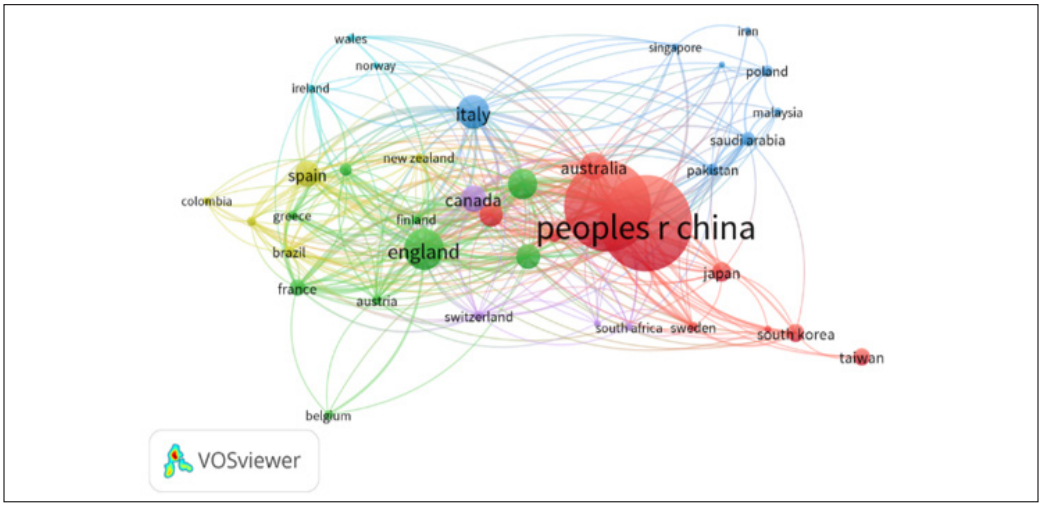


Figure 6. Network map of co-authorship across countries

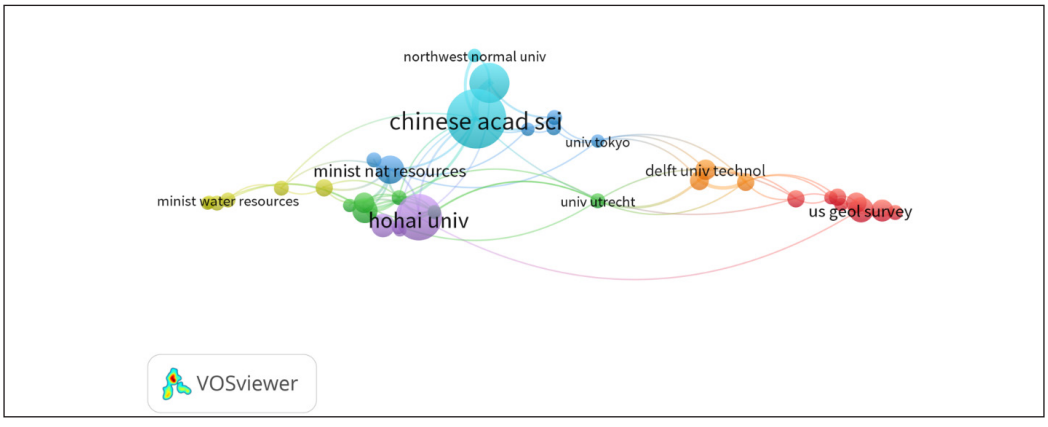


Figure 7. Network map of co-authorship forcited organisations

For example, while the SLR findings related to the lack of regionally specific studies are consistent with the bibliometric analysis's findings related to the lack of studies that focus on tropical regions, the SLR identified areas and topics (e.g., preparedness for disasters in waterfall areas) where collaboration would likely provide the greatest benefit; conversely, the bibliometric approaches quantified the lack of inter-institutional collaboration.

Each of the unique characteristics of the different approaches to data analysis directly supported the goals of the study, as can be seen in Table 5. Citation analysis and Co-Authorship Analysis primarily examined the foundational studies that have defined the field and examined the collaborative relationships among researchers in the field. Performance Analysis identified some of the most recent trends in the production of research in the field.

The subject gaps within the field and the inter-disciplinary connections to other fields were identified using Keyword Co-Occurrence Analysis and Science Mapping, which illustrated the need for additional targeted research in this specific area. The results of the studies were effectively represented through Network Visualisations and provided a clear understanding of the bibliometric connections between authors in the field. Collectively, the studies provided an important addition to the existing body of literature, identified the status of the field and presented practical recommendations for future research projects.

Table 5
Summary of the key findings from each analysis type, the usefulness of each analysis type to achieve the objectives of the study and comparative effectiveness of the analysis types

Analysis Type	Key Insights	Utility for Study	Rank
Performance Analysis	Highlights publication and citation trends, major contributors, and geographic distribution	Useful for identifying key areas in remote sensing research for water surge in waterfalls	1
Network Visualisations	Presents relationships between studies and trends in an intuitive format	Excellent at presenting outcomes efficiently	2
Co-authorship Analysis	Maps collaborative networks among authors and institutions	Identifies top researchers and possible partners in this specialised subject	3
Keyword Co-occurrence	Identifies emerging research topics	Crucial for comprehending the gaps and current areas of interest in remote sensing applications for water surge	4
Co-citation Analysis	Reveals foundational studies	It identifies studies that are widely cited but may ignore recent trends relevant to the past five years	5

CONCLUSIONS

This study has highlighted a very significant research gap regarding the use of remote sensing (RS) methods for evaluating and controlling the risk of water surges in waterfall areas, especially those found in tropical regions, like Malaysia. Even though RS methods have already been used in many different fields of environmental sciences, there are still no studies focused on using RS to monitor and manage the various risks found in waterfall environments, where the environment is constantly changing. This is due to the high level of uncertainty associated with the amount of rainfall experienced in tropical regions, increasing the risk of sudden, unexpected flooding.

The need for a dedicated study regarding RS and Waterfall Environments is evident when examining both the bibliographic study of relevant literature and the bibliometric analysis. The bibliographic study examined the major themes and trends from previous work and included trends that were found, including the number of publications, citation metrics,

and geography of papers that relate to RS and Waterfall Environments. The results from the SLR provided a qualitative component to the bibliographic study's findings to show that most researchers rely on data-based monitoring systems, and further, there was a lack of variety of methodologies used by researchers. The bibliographic study and SLR collectively illustrated that innovative research methods are needed and described a method of potentially advancing this critical area of research. The future studies can then develop and implement real-time monitoring systems and early warning systems that are specific to each location in Malaysia using the unique geology and climate of each waterfall environment by developing multi-disciplinary approaches utilising advanced remote sensing technology (such as satellite imaging, LiDAR, drone technology and GIS-based modelling). These types of programs can contribute towards protecting the ecosystem, disaster preparedness, and sustainable tourism for these natural wonders.

Additionally, studies focused on developing regionally relevant methods for evaluating flood dynamics and hydrologic patterns will lead to more effective and sustainable management of water resources. Finally, the integration of traditional and reference methods into future studies will allow for the identification of unexplored research topics and alignment with the United Nations' Sustainable Development Goals. The convergence of these perspectives will provide researchers with practical insights for enhancing environmental sustainability in highly vulnerable areas.

DECLARATION OF COMPETING INTEREST

There are no conflicts of interest or known personal relationships that may impact the content of this manuscript.

ACKNOWLEDGEMENT

The authors would like to extend our deepest appreciation to the Faculty of Civil Engineering Technology (FCET), University Malaysia Pahang Sultan Abdullah (UMPSA), for providing us with valuable assistance and infrastructure to publish this paper. The authors would also like to acknowledge funding from the UMPSA for the Special Project Research Grant with grant number SPU260101 and Yayasan Pahang Special Grant with grant number RDU240704.

REFERENCES

Abdel-Mooty, M. N., Yosri, A., El-Dakhakhni, W., & Coulibaly, P. (2021). Community flood resilience categorisation framework. *International Journal of Disaster Risk Reduction, 61*, Article 102349. https://doi.org/10.1016/j.ijdrr.2021.102349

Alaghmand, S., Abdullah, R. B., Abustan, I., & Vosoogh, B. (2010). GIS-based river flood hazard mapping in an urban area: A case study in Kayu Ara River Basin, Malaysia. *International Journal of Engineering and Technology, 2*(6), 488-500. https://www.researchgate.net/publication/50422303

Alryalat, S. A. S., Malkawi, L. W., & Momani, S. M. (2019). Comparing bibliometric analysis using PubMed, Scopus, and Web of Science databases. *Journal of Visualised Experiments, 2019*(152), Article e58494. https://doi.org/10.3791/58494

Aria, M., & Cuccurullo, C. (2017). bibliometric: An R tool for comprehensive science mapping analysis. *Journal of Informetrics, 11*(4), 959-975. https://doi.org/10.1016/j.joi.2017.08.007

Bănăduc, D., Simić, V., Cianfaglione, K., Barinova, S., Afanasyev, S., Öktener, A., McCall, G., Simić, S., & Curtean-Bănăduc, A. (2022). Freshwater as a sustainable resource and generator of secondary resources in the 21st century: Stressors, threats, risks, management and protection strategies, and conservation approaches. *International Journal of Environmental Research and Public Health, 19*(24), Article 16570. https://doi.org/10.3390/ijerph192416570

Bates, P. D., Quinn, N., Sampson, C., Smith, A., Wing, O., Sosa, J., Savage, J., Olcese, G., Neal, J., Schumann, G., Giustarini, L., Coxon, G., Porter, J. R., Amodeo, M. F., Chu, Z., Lewis-Gruss, S., Freeman, N. B., Houser, T., Delgado, M., … Krajewski, W. F. (2021). Combined modelling of US fluvial, pluvial, and coastal flood hazard under current and future climates. *Water Resources Research, 57*(2), Article e2020WR028673. https://doi.org/10.1029/2020WR028673

Bhat, S. A., Bashir, O., Ul Haq, S. A., Amin, T., Rafiq, A., Ali, M., Américo-Pinheiro, J. H. P., & Sher, F. (2022). Phytoremediation of heavy metals in soil and water: An eco-friendly, sustainable and multidisciplinary approach. *Chemosphere, 303*, Article 134788. https://doi.org/10.1016/j.chemosphere.2022.134788

D'Aoust, P. M., Graber, T. E., Mercier, E., Montpetit, D., Alexandrov, I., Neault, N., Baig, A. T., Mayne, J., Zhang, X., Alain, T., Langlois, M.-A., Servos, M. R., MacKenzie, M., Figeys, D., Brisson, M., & Delatolla, R. (2021). Catching a resurgence: Increase in SARS-CoV-2 viral RNA identified in wastewater 48 h before COVID-19 clinical tests and 96 h before hospitalisations. *Science of the Total Environment, 770*, Article 145319. https://doi.org/10.1016/j.scitotenv.2021.145319

Danvila-del-Valle, I., Estévez-Mendoza, C., & Lara, F. J. (2019). Human resources training: A bibliometric analysis. *Journal of Business Research, 101*, 627-636. https://doi.org/10.1016/j.jbusres.2019.02.026

de Carvalho Pereira, F., Verocai, H. D., Cordeiro, V. R., Gomes, C. F. S., & Costa, H. G. (2015). Bibliometric analysis of information systems related to innovation. *Procedia Computer Science, 55*, 298-307. https://doi.org/10.1016/j.procs.2015.07.052

Diodato, V. (2013). *Dictionary of bibliometrics*. Routledge. https://doi.org/10.4324/9780203714133

Domenech, T., Bleischwitz, R., Doranova, A., Panayotopoulos, D., & Roman, L. (2019). Mapping industrial symbiosis development in Europe: Typologies of networks, characteristics, performance and contribution to the circular economy. *Resources, Conservation & Recycling, 141*, 76-98. https://doi.org/10.1016/j.resconrec.2018.09.016

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research, 133*, 285-296. https://doi.org/10.1016/j.jbusres.2021.04.070

Garrigos-Simon, F. J., Narangajavana-Kaosiri, Y., & Lengua-Lengua, I. (2018). Tourism and sustainability: A bibliometric and visualisation analysis. *Sustainability, 10*(6), Article 1976. https://doi.org/10.3390/su10061976

Goerlandt, F., Li, J., & Reniers, G. (2020). The landscape of risk communication research: A scientometric analysis. *International Journal of Environmental Research and Public Health, 17*(9), Article 3255. https://doi.org/10.3390/ijerph17093255

Gori, A., Lin, N., & Smith, J. (2020). Assessing compound flooding from landfalling tropical cyclones on the North Carolina coast. *Water Resources Research, 56*(4), Article e2019WR026788. https://doi.org/10.1029/2019WR026788

International Conference on Flood Management. (2022). *A compilation of 2020-2021 global flood events and international experience in flood management*.

Iqbal, U., Barthelemy, J., & Perez, P. (2022). Emerging role of unmanned aerial vehicles (UAVs) for disaster management applications. In *Nanotechnology-based smart remote sensing networks for disaster prevention* (pp. 281-301). Elsevier. https://doi.org/10.1016/B978-0-323-91166-5.00007-0

Issac, M. N., & Kandasubramanian, B. (2021). Effect of microplastics on water and aquatic systems. *Environmental Science and Pollution Research, 28*(16), 19544-19562. https://doi.org/10.1007/s11356-021-13184-2

King, T. V., Neilson, B. T., & Rasmussen, M. T. (2018). Estimating discharge in low-order rivers with high-resolution aerial imagery. *Water Resources Research, 54*(2), 863-878. https://doi.org/10.1002/2017WR021868

Malay Mail. (2024, December 4). *Breaking news, Malaysia, world*. https://www.malaymail.com/

Mrozik, W., Rajaeifar, M. A., Heidrich, O., & Christensen, P. (2021). Environmental impacts, pollution sources and pathways of spent lithium-ion batteries. *Energy & Environmental Science, 14*(12), 6099-6121. https://doi.org/10.1039/D1EE00691F

Muis, S., Apecechea, M. I., Dullaart, J., de Lima Rego, J., Madsen, K. S., Su, J., Yan, K., & Verlaan, M. (2020). A high-resolution global dataset of extreme sea levels, tides, and storm surges, including future projections. *Frontiers in Marine Science, 7*, Article 263. https://doi.org/10.3389/fmars.2020.00263

Munawar, H. S., Hammad, A. W. A., & Waller, S. T. (2021). A review of flood management technologies related to image processing and machine learning. *Automation in Construction, 132*, Article 103916. https://doi.org/10.1016/j.autcon.2021.103916

MyGOV. (n.d.). *Whole of government: Twelfth Malaysia Plan (12MP)*. https://www.malaysia.gov.my/portal/content/31186

Nofal, O. M., & van de Lindt, J. W. (2021). High-resolution flood risk approach to quantify the impact of policy change on flood losses at the community level. *International Journal of Disaster Risk Reduction, 62*, Article 102429. https://doi.org/10.1016/j.ijdrr.2021.102429

Parra, L. (2022). Remote sensing and GIS in environmental monitoring. *Applied Sciences, 12*(16), Article 8045. https://doi.org/10.3390/app12168045

Persson, O., Danell, R., & Wiborg Schneider, J. (2009). *How to use Bibexcel for various types of bibliometric analysis*. International Society for Scientometrics and Informetrics. https://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-25636

Sahoo, B., & Bhaskaran, P. K. (2019). Prediction of storm surge and inundation using climatological datasets for the Indian coast using soft computing techniques. *Soft Computing, 23*(23), 12363-12383. https://doi.org/10.1007/s00500-019-03775-0

Selangor Journal. (2024, November 23). *Water surge incidents in Malaysia*. https://selangorjournal.my

Serwecińska, L. (2020). Antimicrobials and antibiotic-resistant bacteria: A risk to the environment and to public health. *Water, 12*(12), Article 3313. https://doi.org/10.3390/w12123313

Sinar Daily. (2023, July 4). *Water surge incidents in Malaysia since 2021*. https://www.sinardaily.my/article/200361/focus/national/water-surge-incidents-in-malaysia-since-2021

Terlouw, T., Bauer, C., McKenna, R., & Mazzotti, M. (2022). Large-scale hydrogen production via water electrolysis: A techno-economic and environmental assessment. *Energy & Environmental Science, 15*(9), 3583-3602. https://doi.org/10.1039/D2EE01023B

Tourian, M. J., Sneeuw, N., & Bárdossy, A. (2013). A quantile function approach to discharge estimation from satellite altimetry (ENVISAT). *Water Resources Research, 49*(7), 4174-4186. https://doi.org/10.1002/wrcr.20348

Trepekli, K., Balstrøm, T., Friborg, T., Fog, B., Allotey, A. N., Kofie, R. Y., & Møller-Jensen, L. (2022). UAV-borne, LiDAR-based elevation modelling: A method for improving local-scale urban flood risk assessment. *Natural Hazards, 113*(1), 423-451. https://doi.org/10.1007/s11069-022-05308-9

van Eck, N. J., & Waltman, L. (2007). VOS: A new method for visualising similarities between objects. In H.-J. Lenz & R. Decker (Eds.), *Advances in data analysis* (Studies in Classification, Data Analysis, and Knowledge Organisation, pp. 299-306). Springer. https://doi.org/10.1007/978-3-540-70981-7_34

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523-538. https://doi.org/10.1007/s11192-009-0146-3

van Nunen, K., Li, J., Reniers, G., & Ponnet, K. (2018). Bibliometric analysis of safety culture research. *Safety Science, 108*, 248-258. https://doi.org/10.1016/j.ssci.2017.08.011

Wang, X., & Xie, H. (2018). A review on applications of remote sensing and geographic information systems (GIS) in water resources and flood risk management. *Water, 10*(5), Article 608. https://doi.org/10.3390/w10050608

World Meteorological Organisation. (2021, August 31). *Weather-related disasters have increased over the past 50 years, causing more damage but fewer deaths*. https://wmo.int/media/news/weather-related-disasters-increase-over-past-50-years-causing-more-damage-fewer-deaths

Pertanika J. Sci. & Technol. 34 (3): 1599 - 1623 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Carbon Emissions Quantification from Ceramic Tile Production by Using Life-Cycle Assessment (LCA): A Case Study of PT Sinar Karya Duta Abadi, Indonesia

Ransi Pasae[1]*, M. Rijal Idrus[2], Hazairin Zubair[3], and Darhamsyah[4]

[1]Environmental Science Program, Graduate School, Hasanuddin University, Perintis Kemerdekaan Street KM 10, Tamalanrea, 90245 Makassar, South Sulawesi, Indonesia
[2]Hasanuddin University Centre of Excellence for Climate Change Studies, Hasanuddin University, Perintis Kemerdekaan Street KM 10, Tamalanrea, 90245 Makassar, South Sulawesi, Indonesia
[3]Soil Science Study Program, Faculty of Agriculture, Hasanuddin University, Perintis Kemerdekaan Street KM 10, Tamalanrea, 90245 Makassar, South Sulawesi, Indonesia
[4]Environmental Management Study Program, Hasanuddin University, Perintis Kemerdekaan Street KM 10, Tamalanrea, 90245 Makassar, South Sulawesi, Indonesia

ABSTRACT

A major environmental problem in the ceramic tile industry is pollution in the form of carbon emissions from energy-intensive production. The study set out to measure how much carbon is released throughout the ceramic tile production chain and to pinpoint which steps contribute the most. It also explored a few grounded options for cutting those emissions, especially in stages where energy use tends to spike. A cradle-to-gate LCA was carried out assuming a functional unit of 1 m^2 ceramic tile, with primary data extracted from the corporate information for January-December 2024. The system boundary primarily focuses on gate-to-gate processes, while upstream raw material extraction was incorporated through secondary data (emission factors) rather than direct system modelling. According to the LCA results, the carbon footprint of ceramic tile production was estimated at 7.3401 kg $CO_{2\text{-eq}}/m^2$, with fuel combustion and electricity consumption dominating the impact, contributing 67.29% and 30.60% of total emissions, respectively. Sensitivity analysis confirmed that overall emissions are primarily driven by energy-related inputs, particularly kiln fuel use. Based on these findings, two mitigation scenarios were assessed. The installation of a waste heat recovery (WHR) system reduced emissions by 10.09%, while its combination with partial substitution of natural gas by biogas and hydrogen-enriched

ARTICLE INFO

Article history:
Received: 21 October 2025
Accepted: 05 May 2026
Published: 19 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.14

E-mail addresses:
pasaeransi@gmail.com (Ransi Pasae)
rijal.idrus@unhas.ac.id (M. Rijal Idrus)
hazairin-zubair@agri.unhas.ac.id (Hazairin Zubair)
darhamsyah@unhas.ac.id (Darhamsyah)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

fuel increased the reduction potential to 17.16%. Life cycle cost analysis further demonstrated the economic feasibility of both options, yielding negative abatement costs of -82.53 USD/t CO_2 for the WHR scenario and -17.50 USD/t CO_2 for the combined scenario, indicating that meaningful emission reductions can be achieved alongside long-term cost benefits.

Keywords: Ceramic tile production, emissions, environmental impact, Indonesia, life cycle assessment

INTRODUCTION

The world of the ceramic tile industry is a strange one within the construction sector. Tiles are valued because they stand up to traffic, present an intentional and clean appearance in a finished space, and generally don't put materials costs beyond the range of a budget (Chouhan et al., 2024). But even the comfort of familiar surfaces rests on a production chain that links back to hot, mineral-mining intensive kilns and continued dependence on fossil fuels. In many other plants, firing and drying lines operate for almost 365 days a year, using large amounts of electricity and fuel such as natural gas (Ancona et al., 2023). These energy requirements build up the environmental profile of the sector and raise questions more generally about whether it is feasible for a material being widely used to be incorporated into attempts to combat climate change (Xian et al., 2024).

In a world where the industry behaved more sensibly, ceramic processing would run on tightly controlled feeds, recover most of the heat it now wastes and monitor its emissions closely enough to make at least some recommendations about what should come next (Delpech et al., 2017). In reality, things rarely fall into such neat order. Previous research suggests that the global warming potential (GWP) of ceramic tiles can reach 133.60 kg $CO_{2\text{-eq}}/m^2$, with over 90% stemming from CO_2 emissions during natural gas use and raw material processing. The firing stage takes up a disproportionate share of the burden and can easily send the global warming potential of a single square meter of ceramic tile far past the levels commonly documented for many other construction materials (Ibáñez-Forés et al., 2011; Vieira et al., 2023). Temperatures regularly exceeding 1000 °C make meaningful energy savings difficult unless the entire production setup is rethought. Many plants still depend on older machinery that was never built for low-carbon goals, so the gap between what is possible and what actually happens remains quite large (Yuan et al., 2024).

Many researchers have tried to close this gap by digging into process tweaks (comparisons of firing/fuel options) (Wang et al., 2020), running careful evaluations of kiln performance (Ros-Dosdá et al., 2018b), or tracing emissions from the earliest raw-material steps up to the factory gate (Bovea et al., 2007). These attempts have some valuable findings, although they usually concentrate on individual production stages, certain furnace kinds or clusters of plants in Europe and East Asia (Ibáñez-Forés et al., 2011; Ye et al., 2018). As such, the larger picture remains imprecise. Downstream emissions are indirectly

taken into account in many works, or facility-specific aspects affecting results, such as the regional fuel mix and/or the local clay mineralogy, are not considered (Cellura et al., 2011). When that information isn't on offer, there is a real problem. It can also be the case that its determinates a baseline against which emissions reduction needs to be compared, and its consequences will only become apparent when measures are implemented in locations in which production systems appear very differently from those presupposed by global datasets (Fernández-Miguel et al., 2022). And by those places the numbers start to go fuzzy, and there are not, in fact, particularly compelling plans for mitigation where there jolly well should be.

The consequences of this knowledge gap are more sweeping than most people imagine. When emissions data is sparse, companies may overestimate or underestimate what their true carbon liability is. Policymakers thus can create incentives or standards that do not match the real conditions inside factories. Yet not even the architects and developers selecting low-carbon materials are, in practice, able to make fully informed choices (Ibrahim et al., 2024; Suwatno, 2023; Tacconi, 2018). A larger problem is the sparse availability of locally grounded data, which hinders making real progress toward national climate goals. Indonesia illustrates this tension. Despite having a net-zero emission target of 2060, its building sector continues to expand; nevertheless, estimates of the production-stage CO_2 emissions for commonly used construction materials (e.g., ceramic tiles) are surprisingly limited (Bungas et al., 2024; Hermawan & Prabhawati, 2024). Hence, reliable and company-specific LCA investigations are of paramount importance for industrial decarbonization pathways and in reinforcing responsibility for the ceramic tile sector (Ding et al., 2023; Hechelmann et al., 2023; Liu et al., 2023).

A lot of what we know about ceramic tile impacts is still based on studies in places such as Italy, Turkey, Egypt and China, with the firing choices, production approaches and also the clay compositions significantly different from each other (Ali, 2024; Li et al., 2022; Türkmen et al., 2021). Indonesian producers, though, appear only fitfully in the literature, and that absence is significant. Industrial performances vary across Southeast Asia, and many plants in Indonesia use alternative fuels or recycled inputs that are not exactly comparable to global practices (Handaya et al., 2022). With no data from real facilities, any effort to cut emissions will just be based on the uncertain footing of hypotheticals that may have little to do with the realities of production systems in Indonesia. Without data from real facilities, any plan to slash emissions ends up relying on assumptions that may not reflect reality in Indonesia's production systems. In the absence of data from real facilities, any plan to cut emissions winds up resting on assumptions that might not match actual circumstances in Indonesia's production systems.

LCA offers a systematic way of dealing with these uncertainties by following the environmental impacts from the extraction of raw materials to manufacture and distribution.

It applies a procedure that follows ISO 14040 and 14044 (Marson et al., 2024). Currently, LCA is utilised with ceramic tiles to provide industry and policymakers the ability to measure their greenhouse gas (GHG) emissions profiles, as well as pinpoint the production stage in which most of the environmental profile occurs; this allows for discovering possible mitigation strategies that can mitigate carbon footprints (Pini et al., 2014). Used judiciously, LCA also has a habit of illuminating the sources of emissions in reality, like lifting hoods behind which factories rarely peer. By evaluating the actual contributions of all components in production, it closes the gap between efficiencies which engineers draw on paper and the much messier performance that comes out of a factory line (Ferrari et al., 2019; Sappa et al., 2019).

This paper applies a gate-to-gate life cycle assessment (LCA) to PT Sinar Karya Duta Abadi, one of the largest ceramic tile manufacturers in Indonesia. This approach is particularly relevant as it captures emission sources that are directly controllable at the plant level, such as fuel consumption, electricity use, and process-related emissions. By isolating the production phase, the analysis enables a more precise identification of emission hotspots and supports the development of targeted and practically implementable mitigation strategies within industrial operations. Moreover, given the limited availability of facility-specific data in Indonesia, a gate-to-gate perspective provides a robust and context-sensitive basis for evaluating decarbonization opportunities under real operating conditions.

The objective of the study is to quantify CO_2 emissions, identify the most contribute processes within the production system, and develop technically grounded mitigation options under local industrial conditions. In addition to the baseline assessment, the study incorporates sensitivity analysis to evaluate the robustness of the results and to examine the relative influence of key inventory parameters on total emissions. Furthermore, a life cycle cost (LCC) analysis is conducted to assess the economic feasibility of selected emission reduction scenarios, enabling an integrated environmental-economic evaluation. This study specifically examines how the emission profile is shaped by plant-level fuel use, raw material composition, and process configuration, which differ from patterns commonly reported in international studies. The work holds both academic and practical significance, contributing region-specific empirical evidence to LCA research while providing quantitative insights to support industrial decision-making and policy development in the context of increasingly stringent climate and sustainability requirements.

METHODS

Framework

Life cycle assessment based on ISO 14040 and 14044 was carried out to quantify the carbon emissions generated from ceramic tile production in PT Sinar Karya Duta Abadi, Indonesia. The methodological framework is structured with four core phases of

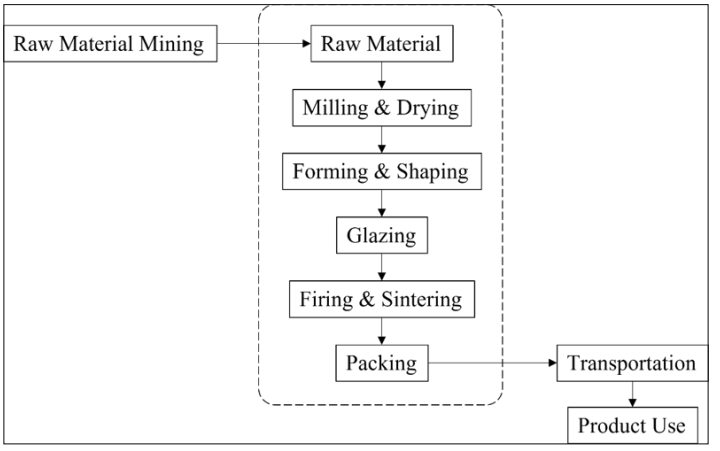


Figure 1. System boundary of ceramic tile production

LCA: goal and scope, life cycle inventory (LCI) analysis, life cycle impact assessment (LCIA) and interpretation, as recommended by standard guidelines for LCA (Vieira et al., 2023). All data collection and modelling were conducted in accordance with gate-to-gate system boundaries to consider emissions associated with on-site production processes, while upstream raw material extraction was incorporated indirectly through secondary emission factors.

Goal and Scope

This study seeks to assess the GWP in $CO_{2\text{-eq}}$ for the production of ceramic tiles in Indonesia from raw material milling and drying through to packing at the factory gate. This study takes a gate-to-gate perspective, as shown in Figure 1. Although the study adopts a gate-to-gate system boundary focusing on on-site manufacturing processes, upstream emissions related to raw material extraction are incorporated indirectly using secondary data (emission factors). Therefore, raw material extraction is not explicitly modelled as part of the foreground system but considered as background data. It is non-comparative and related to the topic of sustainability reporting and carbon-intensive stages. The functional unit is 1 m^2 of ceramic tile (Yuan et al., 2024) and it conforms with the international criterion. That work occurred at PT Sinar Karya Duta Abadi in Indonesia, the type of thing that happens quietly underfoot, for a full year - January through December 2024.

Life Cycle Inventory (LCI)

Our raw data was collected directly on the shop floor of PT Sinar Karya Duta Abadi, where we also performed multiple rounds of walk-throughs and a relatively complete

account of the mass and energy flows. The inventory included the usual technical suspects: how much raw material was used; how much fuel was burning or leaking refrigerant, and the plant's electricity load. For upstream data, in particular, emission factors for the extraction of raw materials and electricity use, the project had to depend on literature values and ecoinvent type databases because regional data sets were not broadly available. All LCI numbers applied in this study are presented in Table 1, Table 2, and Table 3.

Table 1
Emission factors of fuel

Emission Source	Emission Factor (ton $CO_{2\text{-eq}}$/TJ)			References
	CO_2	CH_4	N_2O	
Diesel	74.1000	0.0030	0.0006	(Garg & Pulles, 2006)
Natural gases	56.1000	0.0010	0.0001	(Garg & Pulles, 2006)
Gasoline RON 90	69.3000	0.0030	0.0006	(Garg & Pulles, 2006)

Table 2
Emission factor of raw material, refrigerant, and electricity

Emission Source	Emission Factor	References
Dolomite	0.4773 ton $CO_{2\text{-eq}}$./ton carbonate	(Harnisch & Agyeman-Bonsu, 2006)
Calcium carbonate	0.4397 ton $CO_{2\text{-eq}}$./ton carbonate	(Harnisch & Agyeman-Bonsu, 2006)
Barium carbonate	0.2230 ton $CO_{2\text{-eq}}$./ton carbonate	(Harnisch & Agyeman-Bonsu, 2006)
Calcite	0.4397 ton $CO_{2\text{-eq}}$./ton carbonate	(Harnisch & Agyeman-Bonsu, 2006)
Refrigerant R32	0.6500 ton $CO_{2\text{-eq}}$./ton refrigerant	(Szczęśniak & Stefaniak, 2022)
Refrigerant R410A	1.7250 ton $CO_{2\text{-eq}}$./ton refrigerant	(Szczęśniak & Stefaniak, 2022)
Refrigerant R134	1.0000 ton $CO_{2\text{-eq}}$./ton refrigerant	(Szczęśniak & Stefaniak, 2022)
Refrigerant R407C	1.5255 ton $CO_{2\text{-eq}}$./ton refrigerant	(Szczęśniak & Stefaniak, 2022)
Electricity	0.00087 ton $CO_{2\text{-eq}}$/kWh	(Ministry of Energy and Mineral Resources, 2021)

Table 3
Fuel, material, refrigerant, and electricity consumptions

Emission Source	Consumption
Stationary generator (diesel), L	0.0382
Stationary generator (natural gases), MMBTU	2.0268
Mobile combustion (diesel), L	0.3307
Mobile combustion (gasoline), L	0.0092
Dolomite, kg	0.1863
Calcium carbonate, kg	0.2605
Barium carbonate, kg	0.0233
Calcite, kg	9.3501

Table 3 (*continued*)

Emission Source	Consumption
Refrigerant R32, kg	77.9000
Refrigerant R410A, kg	55.3000
Refrigerant R134, kg	22.0000
Refrigerant R407C, kg	20.5000
Electricity, kWh	83.1961

Table 4
Conversion factors of fuel

Fuel	Conversion Factors
Diesel	0.00003868 TJ/L
Natural gases	0.001055 TJ/MMBTU
Gasoline	0.00003466 TJ/L

Life Cycle Impact Assessment (LCIA)

The quantitative impact assessment was conducted by combining primary and secondary inventory data. The work focused only on the climate change impact category, since the main goal was to estimate carbon emissions expressed as GWP. Other categories (acidification, eutrophication, or resource depletion) were ignored.

The carbon emissions generated from energy consumption can be calculated using Equation 1 (Li et al., 2022), where C_e is the carbon emission generated by energy consumption, M_e is the amount of energy consumed, and F_e is the CO_2 emission factor. Energy consumption is calculated by converting the amount of fuel used into terajoules (TJ) using the conversion factors in Table 4.

$$C_e = M_e \times F_e \quad [1]$$

Carbon released as carbonates break down during high-temperature calcination can be estimated using Equation 2 from (Li et al., 2022). In that equation, C_c represents the emissions tied to the raw materials themselves. M_c is simply the amount of carbonate material that actually gets used, although in practice, this number can vary a bit depending on the batch and how consistently the feedstock is prepared. F_c refers to the emission factor for those carbonate materials, which researchers often treat as a fixed value even though, in reality, different sources of limestone or dolomite can show slight differences in their carbon dioxide output. The emission factors for carbonate decomposition (calcite, dolomite, calcium carbonate, and barium carbonate) were adopted from the 2006 IPCC Guidelines for National Greenhouse Gas Inventories, as reported by Harnisch and Agyeman-Bonsu (2006).

Carbon emissions from refrigerant use are estimated through Equation 3, where C_r represents the emissions produced by the release, M_r is the amount of refrigerant that actually escapes into the air, and F_r is the emission factor that converts that leaked mass into its carbon dioxide equivalent. When Equations 1, 2, and 3 are brought together, they give the overall carbon emissions associated with the ceramic tile life cycle, which is summarised in Equation 4.

$$C_c = M_c \times F_c \quad [2]$$

$$C_r = M_r \times F_r \quad [3]$$

$$C = \sum_{e=1}^{k}(M_e \times F_e) + \sum_{c=1}^{n}(M_c \times F_c) + \sum_{r=1}^{m}(M_r \times F_r) \quad [4]$$

Emission Reduction

The mitigation scenarios were developed based on explicit technical assumptions to ensure transparency and reproducibility. For the WHR scenario, emission reduction was estimated based on the expected thermal recovery efficiency of waste heat recovery systems applied to kiln exhaust gases, which offset a portion of fuel demand. For the combined scenario, partial fuel substitution was modelled using defined blending ratios of biogas and hydrogen-enriched fuel relative to natural gas, along with their respective lower heating values. The detailed parameters and assumptions used in both scenarios are summarised in Table 5.

Table 5
Assumptions for emission reduction scenarios

Scenario	Parameter	Assumed Value	Basic/Justification
WHR	Thermal recovery efficiency	15% effective fuel reduction	Conservative value aligned with the literature range of 8-25% WHR performance in kiln systems
WHR	Heat utilisation application	Preheating of combustion air and raw materials	Standard integration in ceramic roller kiln waste heat recovery systems
WHR	System coverage	Kiln exhaust gases	The main source of thermal loss was identified in the baseline results
Combined scenario	Additional fuel substitution rate	30% reduction of remaining fuel-related emissions	Literature-based conservative assumption for partial substitution in industrial burners
Combined scenario	Biogas fuel type	Renewable biogas (CH_4-rich gas)	Typical substitution fuel in industrial thermal systems
Combined scenario	Biogas lower heating value (LHV)	~20-23 MJ/Nm^3	Standard range for upgraded biogas
Combined scenario	Hydrogen blending ratio	10-20% (volumetric basis)	Conservative blending range for hydrogen-enriched combustion in industrial burners

Table 5 (*continued*)

Scenario	Parameter	Assumed Value	Basic/Justification
Combined scenario	Reference fuel	Natural gas	Baseline kiln fuel in PT Sinar Karya Duta Abadi
Combined scenario	System condition	No change in production output	Consistent with cradle/gate-to-gate LCA assumption in study

Table 6
Data used for life cycle cost assessment

Fuel	Value	Unit
Lifetime	15	Years
Discount rate	8	%
CAPEX – WHR system	1,200,000	USD
CAPEX – Fuel system (biogas + burner retrofit)	600,000	USD
Annual O&M cost – WHR	3	% of CAPEX
Annual O&M cost – fuel system	2	% of CAPEX
Natural gas price	9	USD/GJ
Biogas price	14	USD/GJ
Hydrogen-enriched fuel price	20	USD/GJ
Fuel system price (70% biogas + 30% H2)	6.8	USD/GJ
Emission factor of natural gas	56	kg CO_2/GJ

Life Cycle Cost

The life cycle cost (LCC) analysis was conducted using economic and technical input data summarised in Table 6.

RESULTS AND DISCUSSION

Life Cycle Assessment Result

The life cycle assessment carried out for the ceramic tile production line at PT Sinar Karya Duta Abadi mapped the carbon footprint associated with an annual production output of approximately 14.782 million square meters, representing a large-scale ceramic manufacturing facility operating under Indonesian industrial conditions. Operational data were compiled directly from plant-specific records, including handwritten logs for diesel consumption and refrigerant replenishment, invoice-based records for natural gas and electricity usage, and weighing sheets used in the batching area for raw material inputs. While the level of data uniformity varied across categories, particularly for mobile fuels and refrigerants, site-specific operational data allows the inventory to reflect actual production practices rather than generic or database-derived assumptions, and was therefore considered sufficiently representative of routine plant operations.

Table 7
The carbon emissions of ceramic tile production

Emission Sources	Carbon Emission (kg $CO_{2\text{-eq}}/m^2$)	Total Carbon Emission/year (ton $CO_{2\text{-eq}}$)
Fuel	**4.9394**	**73,019.2563**
Stationary generator (diesel)	0.0101	149.5855
Stationary generator (natural gas)	4.8899	72,286.5541
Mobile combustion (diesel)	0.0386	571.0024
Mobile combustion (gasoline)	0.0008	12.1143
Raw material	**0.1379**	**2,037.7100**
Dolomite	0.0043	64.2821
Calcium carbonate	0.0058	85.1984
Barium carbonate	0.0002	2.3860
Calcite	0.1276	1,885.8461
Refrigerant	**0.0169**	**249.5540**
Refrigerant R32	0.0009	13.3250
Refrigerant R410A	0.0065	95.3925
Refrigerant R134	0.0015	22.0000
Refrigerant R407C	0.0080	118.8365
Electricity	**2.2459**	**33,201.4272**
Total	**7.3401**	**108,507.9500**

Based on the compiled inventory, emission values were calculated for fuel use, electricity consumption, raw materials, and refrigerants. The detailed emission breakdown is presented in Table 7, enabling transparent attribution of emissions to individual input categories. The total carbon footprint of the production phase was estimated at 108,507.95 tons of $CO_{2\text{-eq}}$ per year, providing a quantitative baseline for subsequent scenario evaluation. When normalised to annual production, this corresponds to an emission intensity of 7.3401 kg $CO_{2\text{-eq}}$ per square meter of ceramic tiles produced and the annual production volume of approximately 14.782 million m², confirming the internal consistency of the functional unit scaling across all calculated results.

Carbon emissions were classified into four main sources: fuel consumption, electricity use, raw materials, and refrigerants, allowing clear identification of emission hotspots along the production chain, shown in Figure 2. Fuel consumption emerged as the dominant contributor, accounting for approximately 67.29% of total emissions, equivalent to 73,019.2563 tons of $CO_{2\text{-eq}}$ annually. Within this category, natural gas combustion in stationary kilns accounted for nearly all fuel-related emissions (72,286.5541 tons $CO_{2\text{-eq}}$ per year, or 98.997%), indicating that kiln firing processes remain the primary driver of direct emissions in the assessed production system, while diesel contributed a comparatively minor share.

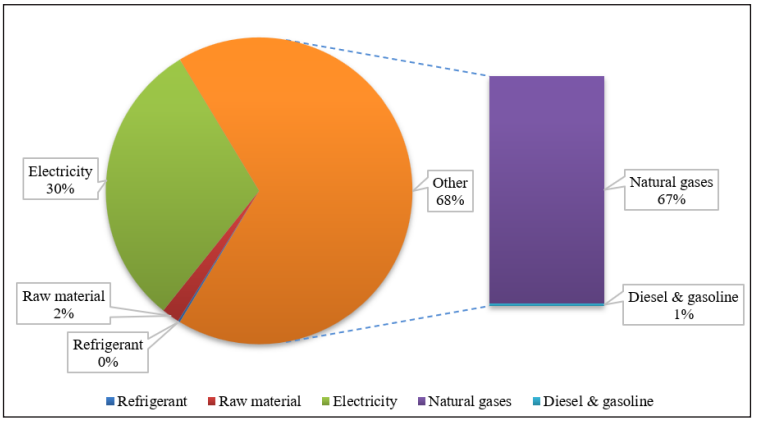


Figure 2. Hierarchical distribution of life cycle carbon emissions from ceramic tile production, highlighting fuel combustion and electricity use as dominant emission hotspots and their detailed contributors

Electricity consumption represented the second-largest emission source, contributing 30.60% of total emissions, or approximately 33,201.4272 tons of $CO_{2\text{-eq}}$ per year. These emissions were associated with purchased electricity used across production processes, including milling, pressing, glazing, and auxiliary operations. The substantial contribution of electricity-related emissions highlights the influence of grid electricity characteristics on the overall carbon footprint of ceramic manufacturing, particularly in regions where fossil-based power generation remains dominant.

Fuel combustion emerged as the dominant contributor primarily due to the high thermal demand of the kiln process, which operates continuously at elevated temperatures and currently does not incorporate heat recovery or preheating systems. In contrast, the emissions associated with purchased electricity are largely influenced by the carbon intensity of the Indonesian electricity grid, which remains dependent on fossil-based generation. Therefore, the higher contribution of fuel reflects both the process characteristics of ceramic firing and the direct dependence on natural gas, whereas electricity-related emissions highlight the indirect effect of the upstream power mix. This distinction suggests that near-term mitigation at the plant level should prioritise kiln efficiency improvements, while deeper reductions in electricity-related emissions will also depend on broader decarbonization of the national power system.

Raw material inputs contributed a smaller fraction of total emissions, amounting to approximately 1.88% (2,037.71 tons $CO_{2\text{-eq}}$ per year). Despite its relatively limited contribution compared to energy-related sources, calcite was identified as the dominant contributor within this category, responsible for 1,885.8461 tons of $CO_{2\text{-eq}}$ annually, accounting for more than 92% of raw-material-related emissions. This finding reflects emissions associated with carbonate decomposition and upstream material processing, which are intrinsic to ceramic body formulation.

Refrigerant-related emissions accounted for 0.23% of the total footprint, corresponding to approximately 249.554 tons of $CO_{2\text{-eq}}$ per year. Although quantitatively small, this category was characterised by high global warming potential values. Among the refrigerants used, R407C and R410A contributed the largest shares, accounting for 47.62% and 38.225% of refrigerant-related emissions, respectively, underscoring the relevance of refrigerant management practices even when total leakage volumes are limited.

Overall, emissions from fuel combustion and electricity consumption (Scope 1 and Scope 2) collectively accounted for approximately 97% of the total carbon footprint. This concentration of emissions within energy-related processes provides a clear and data-driven basis for prioritising mitigation strategies, while also reinforcing the suitability of the assessed system as a case study for evaluating energy-focused emission reduction scenarios in ceramic tile manufacturing.

Studies conducted on Spanish porcelain stoneware reported emission intensities ranging from 130 to 160 kg $CO_{2\text{-eq}}/m^2$ (Ros-Dosdá, Celades, et al. 2018a), while another study reported approximately 133.60 kg $CO_{2\text{-eq}}/m^2$ for glazed stoneware under comparable production conditions (Ibáñez-Forés et al., 2011). Assessments conducted in Turkey reported emission intensities of approximately 14.4-14.5 kg $CO_{2\text{-eq}}/m^2$ (Morfino et al. 2022; Türkmen et al. 2021). The carbon emission intensity of 7.3401 kg $CO_{2\text{-eq}}/m^2$ observed in this study is lower than values reported in several international life cycle assessments of ceramic tile production. This lower value may be attributed to several plant-specific factors, including the kiln configuration, fuel efficiency, operational practices, and the characteristics of the raw material mix used in the studied facility. For example, the plant's production system may exhibit lower thermal losses, more efficient energy utilisation, or a different material formulation compared with facilities reported in previous studies. In addition, variations in system boundary definitions, production scale, and allocation assumptions across studies may also contribute to the observed differences. Therefore, the comparison suggests that emission intensity in ceramic manufacturing is not only influenced by national or regional energy structures, but also by facility-level technical and operational conditions.

In contrast, the emission intensity obtained in this study is closer to values reported for ceramic industries in Brazil and Mexico. Previous study reported emission intensities ranging from 5.07 to 5.31 kg $CO_{2\text{-eq}}/m^2$ for Brazil and 6.00 to 6.33 kg $CO_{2\text{-eq}}/m^2$ for Mexico (Saavedra & Osma, 2025). The proximity of these values suggests that ceramic production systems operating under energy structures dominated by natural gas, with relatively limited electrification of thermal processes, tend to exhibit comparable carbon intensities. These comparisons further indicate that emission intensities in ceramic manufacturing vary widely depending on energy sources, kiln configurations, operational efficiency, and methodological assumptions applied in LCA studies.

Fuel-related emissions dominated the overall footprint, reflecting the inherently energy-intensive nature of high-temperature firing processes in ceramic production. Natural gas-fired kilns, particularly those operating without heat recovery or preheating systems, have consistently been identified as the primary emission hotspot across multiple LCA studies. Similar patterns were reported for Portuguese roof tile production (Quinteiro et al. 2022) and sanitaryware manufacturing (Desole et al. 2024), where kiln firing stages contributed the largest share of environmental burdens. The present findings reinforce the central role of thermal energy demand in shaping the life cycle impacts of ceramic products.

At PT Sinar Karya Duta Abadi, kiln systems are fuelled by natural gas and operate without heat recovery or preheating zones, resulting in substantial thermal losses through exhaust streams. Previous studies, including (Oliveira et al. 2021), have shown that modern roller kilns equipped with waste heat recovery systems can reduce energy consumption by approximately one-fifth, although actual performance is strongly dependent on operational discipline, process control, and maintenance practices. The absence of such systems in the studied facility provides a plausible explanation for the dominance of fuel-related emissions observed in the baseline assessment.

Electricity consumption constituted the second-largest emission source. This trend is consistent with earlier studies, Muthukannan and Chithambar Ganesh (2019) and Wang et al. (2018), which reported elevated electricity-related emissions in production systems involving prolonged drying cycles or high-temperature processes. In the Indonesian context, electricity-related emissions are strongly influenced by the national grid, which remains heavily dependent on coal and natural gas (International Energy Agency, 2024). As a result, improvements in plant-level electrical efficiency alone may have a limited impact unless accompanied by broader decarbonization of the upstream power supply.

Although raw materials contributed a relatively small share of total emissions, calcite dominated this category due to direct CO_2 release during calcination. This finding aligns with previous analyses by a previous study (Mattila et al. 2014), which identified carbonate mineral decomposition as a significant emission source in ceramic manufacturing, even when its associated energy demand is comparatively modest. The result emphasises that process-related emissions, in addition to energy use, should be considered when evaluating mitigation opportunities.

Refrigerant-related emissions accounted for only a minor fraction of the total footprint; however, their high global warming potentials result in disproportionately large climate impacts. R407C and R410A were identified as the main contributors within this category. These findings underscore the importance of improved leak management, regular monitoring, and a gradual transition toward refrigerants with lower climate impacts, particularly as auxiliary systems such as cooling and air conditioning continue to expand in industrial facilities.

Sensitivity Analysis

Sensitivity analysis in LCA is a systematic approach used to examine the extent to which the calculated environmental impacts respond to variations in assumptions, input parameters, and methodological choices. Rather than aiming to modify the numerical results, its primary purpose is to evaluate the robustness and credibility of the conclusions derived from the LCA model. In the present study, sensitivity analysis was performed on the main emission hotspots identified in Figure 2, including fuel consumption (natural gas, diesel, and gasoline), electricity use, raw materials, and refrigerants. These parameters were selected due to their dominant contribution to the overall carbon footprint. The outcomes of the sensitivity analysis are illustrated in Figure 3, providing insight into the relative influence of key inputs on total CO_2-eq emissions and supporting the reliability of the hotspot identification.

The results of the one-at-a-time sensitivity analysis are presented in the tornado chart shown in Figure 3, illustrating the relative influence of variations in key inventory parameters on total annual $CO_{2\text{-eq}}$ emissions. Fuel consumption, including natural gas, diesel, and gasoline, was varied by ±20% and exhibited the highest sensitivity among all evaluated parameters. In particular, changes in natural gas consumption in the kiln system led to the largest deviation from the baseline emissions, resulting in variations of approximately -13.32% to +13.32% in total $CO_{2\text{-eq}}$ emissions. This outcome reflects the dominant contribution of fuel use to the overall carbon footprint of the ceramic tile production system.

Electricity consumption was also varied by ±20% and emerged as the second most influential parameter in the sensitivity analysis. Variations in electricity input resulted in changes of approximately -6.12% to +6.12% relative to the baseline emissions.

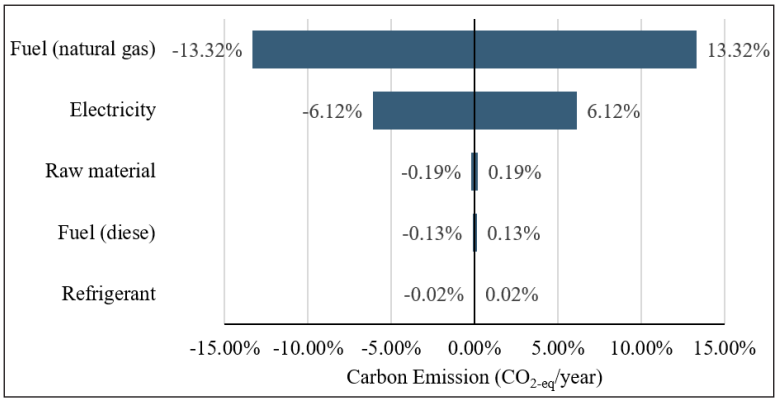


Figure 3. Tornado chart illustrating one-at-a-time sensitivity analysis of total annual carbon emissions from ceramic tile production, showing the relative influence of key input parameters. Positive and negative bars represent ± variation of each parameter around the baseline scenario

Although less pronounced than the effect of fuel consumption, the results indicate that electricity use remains a significant contributor to the variability of total emissions due to its involvement across multiple production stages.

In contrast, raw material inputs were varied by ±10% and produced only minor changes in total $CO_{2\text{-eq}}$ emissions, with deviations remaining below ±0.2% from the baseline. Similarly, refrigerant use, also varied by ±10%, resulted in negligible changes in total emissions. These findings indicate that total carbon emissions from ceramic tile production are primarily sensitive to energy-related parameters, while non-energy inputs exhibit limited influence within the tested variation ranges.

The sensitivity analysis provides important insights into the structural drivers of carbon emissions in ceramic tile production and confirms the robustness of the hotspot identification derived from the baseline LCA results. For fuel inputs (natural gas and diesel), a ±20% variation range was selected, as LCA literature on industrial systems commonly indicates that such a range conservatively represents operational variability and data uncertainty at the plant level. For example, a previous study conducted a sensitivity analysis by increasing and decreasing natural gas consumption by 20%, resulting in an approximately 7% change in global warming potential (GWP) (Hubbard et al. 2020). This finding suggests that a ±20% variation in fuel inputs leads to proportional environmental changes and can therefore be considered a reasonable range for testing model sensitivity. Although fewer studies explicitly address diesel consumption, its role as a major fossil energy carrier justifies the application of a similar ±20% range to account for fluctuations in consumption and uncertainty in diesel-related emission data.

The pronounced sensitivity of total $CO_{2\text{-eq}}$ emissions to variations in fuel consumption, particularly natural gas used in kiln firing, reflects the fundamentally energy-intensive nature of ceramic manufacturing. High-temperature firing processes operate continuously and dominate thermal energy demand; consequently, even conservative variations in fuel input (±20%) lead to substantial shifts in total emissions (approximately -13.32% to +13.32%). This finding is consistent with previous LCA studies on ceramic and other high-temperature industries, which consistently identify kiln systems as the primary leverage point for emission mitigation. Importantly, the magnitude of this sensitivity suggests that uncertainties in fuel consumption data, such as daily production fluctuations, burner efficiency variations, or measurement limitations, can significantly influence absolute emission estimates, while still preserving the relative ranking of emission sources.

A ±20% variation range was also applied to electricity consumption, consistent with practices commonly reported in LCA studies. Previous study, for instance, applied a ±20% variation to electricity input, which resulted in an approximately ±3.8% change in GWP (Hubbard et al. 2020). This indicates that electricity demand at the plant level may realistically vary by up to 20%, due to factors such as process load fluctuations or variations in grid supply conditions, while remaining within plausible operational scenarios.

This range also captures uncertainty associated with electrical energy intensity, including equipment efficiency and utilisation factors, without overstating variability beyond realistic industrial conditions.

For raw materials such as calcite, dolomite, and other fillers, LCA literature rarely provides explicit variation ranges. Raw material composition is generally governed by process formulations and tends to remain relatively stable across production cycles. Consequently, a narrower range of ±10-20% was assumed to conservatively capture potential variability arising from differences in material supply, quality, or inventory data uncertainty. This range also accounts for uncertainties related to material properties, such as chemical composition or moisture content, without exaggerating variability. While some LCA studies in other contexts, such as waste management, have explored much broader composition ranges, a more moderate assumption was adopted here in line with standard industrial LCA practice.

For refrigerants (e.g., R410A and R407C), a ±10% variation range was applied, reflecting their relatively small contribution to total LCA results despite their high global warming potential. As an illustration, previous research reported that leakage of R410A contributed approximately 3% of total GWP in air conditioning systems (Ross & Cheah, 2017). Given the relatively small absolute magnitude of refrigerant-related emissions, a narrower variation range was considered appropriate to represent uncertainty in refrigerant charging and leakage rates at the plant scale. A ±10% range was therefore deemed sufficiently conservative to reflect realistic uncertainty in refrigerant use while ensuring that such variations do not disproportionately influence the overall LCA outcomes.

In contrast, the limited sensitivity of total emissions to variations in raw material inputs and refrigerant use demonstrates that non-energy parameters play a comparatively minor role in shaping the overall LCA outcome within the tested uncertainty ranges. Variations of ±10% in raw material consumption resulted in deviations of less than ±0.2%, reflecting both the relatively small contribution of material-related emissions and the stability of ceramic body formulations in industrial practice. Similarly, refrigerant-related emissions, despite their high global warming potentials, exhibited negligible influence on total results due to their small absolute quantities. These findings indicate that uncertainties associated with material composition, supply variability, or refrigerant leakage rates are unlikely to alter the central conclusions of this study, thereby reinforcing the robustness of the identified energy-related emission hotspots.

From an industrial and policy perspective, the sensitivity results provide clear guidance for prioritising emission reduction strategies in ceramic manufacturing. Measures targeting kiln fuel efficiency, such as waste heat recovery, improved insulation, burner optimisation, and partial substitution with lower-carbon fuels, are likely to deliver the largest and most reliable emission reductions. Electricity-related measures, including process optimisation and procurement of lower-carbon electricity, represent secondary but still meaningful

opportunities. Conversely, efforts focused primarily on modifying raw material inputs or refrigerant management, while relevant for comprehensive environmental management, are unlikely to yield substantial reductions in overall carbon emissions. Integrating these sensitivity insights with life-cycle cost considerations will be critical to translating environmental priorities into economically viable decarbonization strategies for the ceramic industry.

Emission Reduction

Emission reduction scenarios were developed based on the emission distribution identified in the baseline assessment, with particular emphasis on fuel combustion and electricity use as the principal emission sources. The scenarios were derived directly from the dominant contributors identified in Figure 2, ensuring that the proposed mitigation measures target empirically observed emission hotspots rather than hypothetical interventions. A linear modelling approach was applied, assuming that percentage reductions in energy consumption would result in proportional reductions in $CO_{2\text{-eq}}$ emissions. The analysis further assumed constant production output, minimal additional operational costs, and no changes in fuel prices or electricity tariffs, allowing the scenarios to reflect technically realistic improvements under current operating conditions.

Waste heat recovery (WHR) was evaluated as a primary mitigation option, based on the premise that a portion of thermal energy contained in kiln exhaust gases could be recovered for preheating incoming air or raw materials. This option was selected due to its compatibility with existing kiln configurations commonly used in ceramic manufacturing. A conservative 15% reduction in fuel-related emissions was assumed to account for

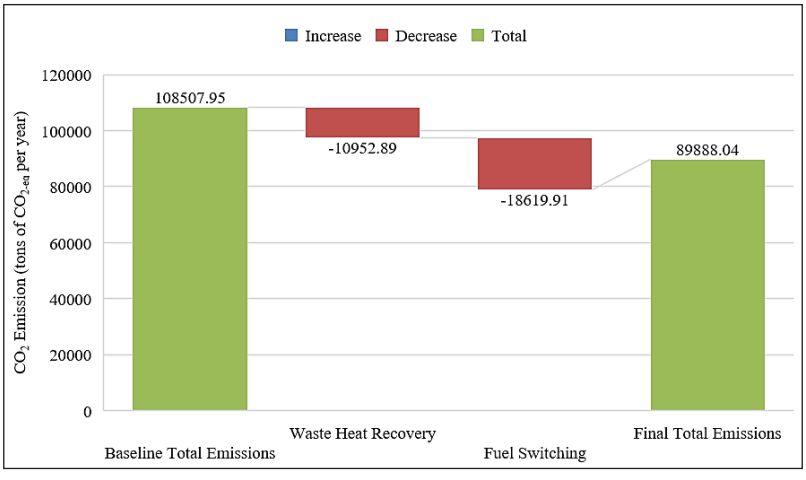


Figure 4. Stepwise emission reduction pathway for ceramic tile production under waste heat recovery (WHR) and combined fuel switching scenarios

variations in kiln design and flue gas temperature. Based on this assumption, emissions were reduced by 10,952.888 tons of $CO_{2\text{-eq}}$ per year, resulting in a revised total emission level of 97,555.062 tons of $CO_{2\text{-eq}}$ per year.

The second scenario integrates Waste Heat Recovery (WHR) with partial substitution of natural gas by biogas and hydrogen-enriched fuel, representing a progressive decarbonization pathway rather than full fuel replacement. In this scenario, an additional 30% reduction was applied specifically to the remaining fuel-related emissions after WHR implementation, consistent with values reported in the literature. As a result, total annual emissions decreased to 89,888.04 tons of $CO_{2\text{-eq}}$ per year, corresponding to an overall reduction of approximately 17.17% relative to the baseline emissions shown in Figure 4.

The emission reduction estimates are based on the technical assumptions summarised in Table 5, ensuring consistency between model inputs and reported results. The emission reduction scenarios evaluated in this study demonstrate that targeted interventions focusing on thermal efficiency and fuel composition can yield meaningful emission reductions in ceramic tile manufacturing. Waste heat recovery represents a particularly attractive option, as ceramic kilns typically lose 20-30% of thermal energy through exhaust gases. By explicitly linking mitigation options to the dominant emission sources identified in the baseline assessment, the analysis provides a coherent pathway from hotspot identification to intervention selection. Previous studies have similarly identified kiln systems as the most effective targets for emission mitigation, reinforcing the relevance of the selected approach (Desole et al. 2024; Quinteiro et al. 2022).

The practical feasibility of these mitigation options is supported by the existing kiln configuration at the studied facility, which relies on natural gas-fired thermal processes and therefore offers clear opportunities for heat recovery and fuel optimisation. Waste heat recovery can be integrated into the current system with relatively limited process disruption, particularly by utilising exhaust heat for preheating incoming air or materials. However, its implementation may require careful engineering design, space availability, and maintenance planning to ensure stable performance under continuous production conditions. In contrast, partial fuel substitution with biogas and hydrogen-enriched fuel may face greater operational constraints, including burner compatibility, fuel supply reliability, and the need for retrofitting of combustion systems. For a facility of this scale, such measures are more realistically implemented in a phased manner, allowing technical adjustments and operational learning to occur gradually while maintaining production continuity. Therefore, the estimated emission reductions represent technically feasible pathways, although their actual deployment would depend on plant-specific engineering and infrastructure readiness.

Reported reductions in energy consumption associated with WHR range from 8% to 25%, depending on kiln design and operating conditions (Yin et al. 2017). The 15%

reduction assumed in this study, therefore, falls within a conservative and technically realistic range, particularly for facilities operating conventional natural gas-fired kilns without integrated heat recovery. Beyond technological interventions, organisational measures such as energy management systems and ISO 50001 certification have been shown to play a critical role in sustaining efficiency gains through continuous monitoring, operator training, and maintenance optimisation (Prasetya et al. 2021), highlighting that operational practices can be as influential as equipment upgrades.

Combining WHR with partial fuel substitution further enhances emission reduction potential. Previous LCA-based models suggest that replacing natural gas with biogas or hydrogen-enriched fuel can reduce CO_2 emissions by approximately 30-50% in ceramic tile production systems (Saavedra & Osma, 2025). Framing fuel switching as a partial and gradual substitution, rather than a complete replacement, reflects realistic constraints related to fuel availability, burner compatibility, and cost. Additional measures, such as solar-assisted preheating, have also been proposed as complementary strategies, particularly in regions with high solar availability (Kumar et al. 2022), offering opportunities to reduce fossil fuel demand without disrupting core thermal processes.

Importantly, the mitigation pathways evaluated in this study do not rely on assumptions of immediate or complete energy system decarbonisation. Instead, they reflect incremental and technically feasible strategies that can be implemented within existing industrial infrastructure and current energy system constraints. While plant-level interventions can deliver significant near-term emission reductions, deeper decarbonisation will ultimately depend on broader transitions in national electricity generation toward cleaner energy sources. In this sense, the scenarios provide a pragmatic bridge between current operational realities and longer-term sustainability objectives.

Table 8
Comparison of emission reduction and life cycle cost results for WHR-only and combined mitigation scenarios

Parameter	Scenario 1: WHR nly	Scenario 2: WHR + Fuel Substitution
Baseline emission (t CO_2/y)	108,507.95	108,507.95
Emission after mitigation (t CO_2/y)	97,555.06	89,888.04
CO_2 avoided (t CO_2/y)	10,952.89	18,619.91
Emission reduction (%)	10.09%	17.16%
CAPEX (USD)	1,200,000.00	1800000
Annual net cost (USD/y)	1,724,285.89	781,291
Total CO_2 avoided (t CO_2)	164,293	279,299
NPV of total cost (USD)	-13,558,988.35	-4887440.301
Cost per ton CO_2 avoided (USD/t)	-82.53	-17.50

The results suggest that integrating life cycle assessment with incremental improvement scenarios offers a useful framework for guiding emission reduction strategies in ceramic industries, particularly in regions where rapid energy system decarbonisation remains challenging.

Life Cycle Cost

The life cycle cost (LCC) results of the evaluated mitigation scenarios are presented in Table 8, which compares the environmental and economic performance of the WHR-only scenario and the combined WHR with fuel substitution scenario. Under Scenario 1, the application of waste heat recovery reduced annual CO_2 emissions from 108,507.95 t CO_2/y to 97,555.06 t CO_2/y, corresponding to an avoided emission of 10,952.89 t CO_2/y or a reduction of 10.09%. Over the assumed technical lifetime of 15 years, the cumulative avoided emissions reached approximately 164,293 t CO_2.

From an economic standpoint, Scenario 1 required a capital investment of USD 1.2 million. Nevertheless, the significant reduction in natural gas consumption generated substantial operational savings, resulting in an annual net cost benefit of USD 1.72 million. When discounted using an 8% discount rate, the net present value (NPV) of the total cost over the system lifetime was estimated at -13.56 million USD. Consequently, the cost per ton of CO_2 avoided reached -82.53 USD/t CO_2, indicating that the WHR system provides both emission mitigation and strong economic returns.

In Scenario 2, which integrates WHR with partial substitution of natural gas by biogas and hydrogen-enriched fuel, annual emissions were further reduced to 89,888.04 t CO_2/y. This corresponds to an avoided emission of 18,619.91 t CO_2/y or a total reduction of 17.16% relative to the baseline. Over the 15-year operational period, the cumulative avoided emissions increased significantly to approximately 279,299 t CO_2.

The combined scenario required a higher capital expenditure of USD 1.8 million due to additional fuel system retrofitting. Although operational fuel costs increased, the scenario still produced a negative NPV of -4.89 million USD. The resulting cost per ton of CO_2 avoided was -17.50 USD/t CO_2, confirming that the combined mitigation pathway remains economically feasible while achieving deeper emission reductions.

The LCC results indicate that waste heat recovery (WHR) represents a highly cost-effective mitigation option for the ceramic industry. The WHR-only scenario yields a strongly negative abatement cost, indicating that the fuel cost savings generated during operation are sufficient to offset both the initial investment and annual operating expenses. This finding is consistent with empirical studies in cement and other high-temperature industrial sectors, which report that WHR systems can simultaneously reduce production costs and greenhouse gas emissions, particularly when bottoming-cycle technologies such as Organic Rankine Cycles (ORC) or heat utilisation for air and raw material preheating are integrated into kiln systems (Marenco-Porto et al. 2023).

In contrast, the combined scenario integrating WHR with partial fuel substitution using biogas and hydrogen-enriched fuel achieves a substantially higher absolute emission reduction but exhibits a lower marginal cost efficiency. This outcome is primarily associated with the price premium of low-carbon fuels and the additional capital and retrofit requirements needed to enable fuel switching. Similar trends have been reported in previous decarbonization pathway analyses for energy-intensive industries, which emphasise that energy efficiency measures (such as WHR) generally represent no-regret options, while fuel substitution pathways often face significant economic and infrastructural barriers unless supported by carbon pricing mechanisms, targeted subsidies, or long-term policy frameworks (Lechtenböhmer et al. 2016).

From a strategic perspective, these results support a hierarchical decarbonization approach. Thermal efficiency improvements, including WHR implementation, should be prioritised as short- to medium-term measures capable of delivering immediate emission reductions together with direct economic benefits. Fuel substitution strategies, such as the gradual introduction of biogas and hydrogen blends, are more suitable as medium- to long-term options, particularly as fuel supply chains mature and supportive policy instruments become available. In this context, the integrated LCA & LCC framework applied in this study provides a robust decision-support tool, enabling industrial stakeholders and policymakers to quantitatively assess trade-offs between emission reduction depth and economic feasibility in the ceramic manufacturing sector.

CONCLUSION

This study applied a gate-to-gate life cycle assessment to a large-scale ceramic tile manufacturing facility in Indonesia, PT Sinar Karya Duta Abadi, revealing a total carbon footprint of 108,507.95 t $CO_{2\text{-eq}}$ per year, equivalent to 7.3401 kg $CO_{2\text{-eq}}/m^2$ of ceramic tiles. Fuel combustion in kiln operations and electricity consumption were identified as the dominant emission hotspots, jointly contributing approximately 97% of total emissions, a finding further confirmed by sensitivity analysis showing strong responsiveness of overall emissions to variations in energy-related parameters. Based on these results, emission reduction scenarios were developed, indicating that waste heat recovery alone could reduce emissions by 10.09%, while its integration with partial fuel substitution decreased total emissions by up to 17.16%. Life cycle cost analysis demonstrated that both mitigation pathways are economically feasible, with negative abatement costs of -82.53 USD/t CO_2 for the WHR scenario and -17.50 USD/t CO_2 for the combined scenario, highlighting that meaningful emission reductions can be achieved alongside long-term economic benefits.

ACKNOWLEDGEMENT

The authors gratefully acknowledge PT Sinar Karya Duta Abadi for the technical support and data provided for the preparation of this manuscript.

REFERENCES

Ali, A. A. M. (2024). Cradle-to-gate life cycle assessment of interior floor material alternatives in Egypt. *Journal of Umm Al-Qura University for Engineering and Architecture, 15*(3), 282-297. https://doi.org/10.1007/s43995-024-00057-1

Ancona, M. A., Branchini, L., Ottaviano, S., Bignozzi, M. C., Ferrari, B., Mazzanti, B., Salvio, M., Toro, C., Martini, F., & Benedetti, M. (2023). Energy and environmental assessment of cogeneration in the ceramic tiles industry. *Energies, 16*(1), Article 182. https://doi.org/10.3390/en16010182

Bovea, M.-D., Saura, Ú., Ferrero, J. L., & Giner, J. (2007). Cradle-to-gate study of red clay for use in the ceramic industry. *The International Journal of Life Cycle Assessment, 12*(6), 439-447. https://doi.org/10.1065/lca2006.06.252

Bungas, K., Adji, F. F., Darung, U., Nyahu, N., Ruthena, Y., & Sosilawaty, S. (2024). Environmental management technologies, environmental policy stringency, energy productivity and pollution emissions: Fresh evidence from Indonesia under the STIRPAT framework. *International Journal of Energy Economics and Policy, 14*(3), 70-78. https://doi.org/10.32479/ijeep.15690

Cellura, M., Longo, S., & Mistretta, M. (2011). Sensitivity analysis to quantify uncertainty in life cycle assessment: The case study of an Italian tile. *Renewable and Sustainable Energy Reviews, 15*(9), 4697-4705. https://doi.org/10.1016/j.rser.2011.07.082

Chouhan, M. K., Meena, N., Agrawal, J., & Gupta, T. (2024). Enhancing the durability of concrete and mortar with ceramic waste: A comprehensive review. *Journal of Scientific Research and Reports, 30*(8), 485-502. https://doi.org/10.9734/jsrr/2024/v30i82272

Delpech, B., Axcell, B., & Jouhara, H. (2017). A review on waste heat recovery from exhaust in the ceramics industry. *E3S Web of Conferences, 22*, Article 00034. https://doi.org/10.1051/e3sconf/20172200034

Desole, M. P., Fedele, L., Gisario, A., & Barletta, M. (2024). Life cycle assessment (LCA) of ceramic sanitaryware: Focus on the production process and scenario analysis. *International Journal of Environmental Science and Technology, 21*(2), 1649-1670. https://doi.org/10.1007/s13762-023-05074-6

Ding, K., Li, A., Lv, J., & Gu, F. (2023). Decarbonising the ceramic industry: Technological routes and cost assessment. *Journal of Cleaner Production, 419*, Article 138278. https://doi.org/10.1016/j.jclepro.2023.138278

Fernández-Miguel, A., Riccardi, M. P., Veglio, V., García-Muiña, F. E., Fernández del Hoyo, A. P., & Settembre-Blundo, D. (2022). Disruption in resource-intensive supply chains: Reshoring and nearshoring as strategies to enable them to become more resilient and sustainable. *Sustainability, 14*(17), Article 10909. https://doi.org/10.3390/su141710909

Ferrari, A. M., Volpi, L., Pini, M., Siligardi, C., García-Muiña, F. E., & Settembre-Blundo, D. (2019). Building a sustainability benchmarking framework of ceramic tiles based on life cycle sustainability assessment (LCSA). *Resources, 8*(1), Article 11. https://doi.org/10.3390/resources8010011

Garg, A., & Pulles, T. (2006). *2006 IPCC guidelines for national greenhouse gas inventories: Volume 2-Energy*. Institute for Global Environmental Strategies. https://www.ipcc-nggip.iges.or.jp/public/2006gl/vol2.html

Handaya, Marimin, Indrawan, D., & Susanto, H. (2022). A comparative life cycle assessment of palm kernel shell in ceramic tile production: Managerial implications for renewable energy usage. *Sustainability, 14*(16), Article 10100. https://doi.org/10.3390/su141610100

Harnisch, J., & Agyeman-Bonsu, W. K. (2006). *2006 IPCC guidelines for national greenhouse gas inventories: Volume 3-Industrial processes and product use*. Institute for Global Environmental Strategies. https://www.ipcc-nggip.iges.or.jp/public/2006gl/vol3.html

Hechelmann, R.-H., Paris, A., Buchenau, N., & Ebersold, F. (2023). Decarbonisation strategies for manufacturing: A technical and economic comparison. *Renewable and Sustainable Energy Reviews, 188*, Article 113797. https://doi.org/10.1016/j.rser.2023.113797

Hermawan, L. T., & Prabhawati, A. (2024). Implementasi *Just Energy Transition Partnership* Indonesia menuju *net zero emissions* tahun 2060 [Implementation of Indonesia's Just Energy Transition Partnership towards net-zero emissions by 2060]. *Jurnal Energi Baru dan Terbarukan, 5*(3), 28-38. https://doi.org/10.14710/jebt.2024.24161

Ibáñez-Forés, V., Bovea, M.-D., & Simó, A. (2011). Life cycle assessment of ceramic tiles: Environmental and statistical analysis. *The International Journal of Life Cycle Assessment, 16*(9), 916-928. https://doi.org/10.1007/s11367-011-0322-6

Ibrahim, K., Jais, M., Md Isa, A. H., & Jaffary, M. A. S. A. (2024). Unveiling the green truth: How transparent are Malaysian manufacturing companies about their carbon footprint? *International Journal of Accounting and Financial Reporting, 14*(4), 20-27. https://doi.org/10.5296/ijafr.v14i4.22327

International Energy Agency. (2024). *Indonesia energy mix*. https://www.iea.org/countries/indonesia/energy-mix

Kumar, L., Ahmed, J., El Haj Assad, M., & Hasanuzzaman, M. (2022). Prospects and challenges of solar thermal for process heating: A comprehensive review. *Energies, 15*(22), Article 8501. https://doi.org/10.3390/en15228501

Lechtenböhmer, S., Nilsson, L. J., Åhman, M., & Schneider, C. (2016). Decarbonising the energy-intensive basic materials industry through electrification: Implications for future EU electricity demand. *Energy, 115*, 1623-1631. https://doi.org/10.1016/j.energy.2016.07.110

Li, B., Li, D., Hu, J., Zhu, X., Wang, H., Jeon, C., Kim, G.-M., & Zeng, Y. (2024). Carbon emissions from municipal solid waste under different classification methods in the context of carbon neutrality: A case study of Yunnan Province, China. *Fuel, 372*, Article 132167. https://doi.org/10.1016/j.fuel.2024.132167

Li, Y., Sharaai, A. H., Ma, S., Wafa, W., He, Z., & Ghani, L. A. (2022). Quantification of carbon emissions and solid waste from pottery production using the life-cycle assessment (LCA) method in Yunnan, China. *Processes, 10*(5), Article 926. https://doi.org/10.3390/pr10050926

Liu, J., Liu, Y., Zheng, Y., & Li, X. (2023). Progress in green and low-carbon technology development in the building ceramics industry. *Journal of Physics: Conference Series, 2639*(1), Article 012008. https://doi.org/10.1088/1742-6596/2639/1/012008

Marenco-Porto, C. A., Fierro, J. J., Nieto-Londoño, C., Lopera, L., Escudero-Atehortua, A., Giraldo, M., & Jouhara, H. (2023). Potential savings in the cement industry using waste heat recovery technologies. *Energy, 279*, Article 127810. https://doi.org/10.1016/j.energy.2023.127810

Marson, A., Zuliani, F., Fedele, A., & Manzardo, A. (2024). Life cycle assessment-based decision making under methodological uncertainty: A framework proposal. *Journal of Cleaner Production, 445*, Article 141288. https://doi.org/10.1016/j.jclepro.2024.141288

Mattila, H. P., Hudd, H., & Zevenhoven, R. (2014). Cradle-to-gate life cycle assessment of precipitated calcium carbonate production from steel converter slag. *Journal of Cleaner Production, 84*, 611-618. https://doi.org/10.1016/j.jclepro.2014.05.064

Ministry of Energy and Mineral Resources. (2021). *Decree of the Minister of Energy and Mineral Resources No. 163.K/HK.02/MEM.S/2021 on the determination of greenhouse gas emission factors for the electricity generation system in Indonesia*. Government of Indonesia. https://jdih.esdm.go.id/

Morfino, A., Gediga, J., Harlow, K., & Mazzanti, B. (2022). Life cycle assessment comparison between zircon and alumina sand applied to ceramic tiles. *Cleaner Engineering and Technology, 6*, Article 100359. https://doi.org/10.1016/j.clet.2021.100359

Muthukannan, M., & Chithambar Ganesh, A. S. (2019). The environmental impact caused by the ceramic industries and assessment methodologies. *International Journal for Quality Research, 13*(2), 315-334. https://doi.org/10.24874/IJQR13.02-05

Oliveira, M. C., Iten, M., & Matos, H. A. (2021). Assessment of energy efficiency improvement in the ceramic industry through waste heat recovery modelling. In M. Türkay & R. Gani (Eds.), *Computer aided chemical engineering* (Vol. 50, pp. 1653-1658). Elsevier. https://doi.org/10.1016/B978-0-323-88506-5.50256-4

Pini, M., Ferrari, A. M., Gamberini, R., Neri, P., & Rimini, B. (2014). Life cycle assessment of a large, thin ceramic tile with advantageous technological properties. *The International Journal of Life Cycle Assessment, 19*(9), 1567-1580. https://doi.org/10.1007/s11367-014-0764-8

Prasetya, B., Wahono, D. R., Dewantoro, A., & Anggundari, W. C. (2021). The role of an energy management system based on ISO 50001 for energy-cost savings and reduction of CO2 emissions: A review of implementation, benefits and challenges. *IOP Conference Series: Earth and Environmental Science, 926*(1), Article 012077. https://doi.org/10.1088/1755-1315/926/1/012077

Quinteiro, P., Almeida, M. I., Serra, J., Arroja, L., & Dias, A. C. (2022). Life cycle assessment of ceramic roof tiles: A temporal perspective. *Journal of Cleaner Production, 363*, Article 132568. https://doi.org/10.1016/j.jclepro.2022.132568

Ros-Dosdá, T., Celades, I., Monfort, E., & Fullana-i-Palmer, P. (2018a). Environmental profile of Spanish porcelain stoneware tiles. *The International Journal of Life Cycle Assessment, 23*(8), 1562-1580. https://doi.org/10.1007/s11367-017-1377-9

Ros-Dosdá, T., Fullana-i-Palmer, P., Mezquita, A., Masoni, P., & Monfort, E. (2018b). How can the European ceramic tile industry meet the EU's low-carbon targets? A life cycle perspective. *Journal of Cleaner Production, 199*, 554-564. https://doi.org/10.1016/j.jclepro.2018.07.176

Ross, S. A., & Cheah, L. (2017). Uncertainty quantification in life cycle assessments: Interindividual variability and sensitivity analysis in LCA of air-conditioning systems. *Journal of Industrial Ecology, 21*(5), 1103-1114. https://doi.org/10.1111/jiec.12505

Saavedra, E. L., & Osma, J. F. (2025). Evaluating decarbonisation technologies and optimising pathways in ceramic tile production through life cycle assessment in Latin America. *Journal of Cleaner Production, 518*, Article 145800. https://doi.org/10.1016/j.jclepro.2025.145800

Sappa, G., Iacurto, S., Ponzi, A., Tatti, F., Torretta, V., & Viotti, P. (2019). The LCA methodology for ceramic tile production through the addition of MSWI bottom ash. *Resources, 8*(2), Article 93. https://doi.org/10.3390/resources8020093

Suwatno, D. S. R. (2023). Ratifikasi terhadap traktat Persetujuan Paris (*Paris Agreement*) sebagai wujud implementasi komitmen Indonesia dalam upaya mitigasi dan adaptasi perubahan iklim [Ratification of the Paris Agreement as a manifestation of Indonesia's commitment to climate change mitigation and adaptation efforts]. *Jurnal Pendidikan Kewarganegaraan Undiksha, 10*(2), 328-340. https://doi.org/10.23887/jpku.v10i2.47076

Szczęśniak, S., & Stefaniak, Ł. (2022). Global warming potential of new gaseous refrigerants used in chillers in HVAC systems. *Energies, 15*(16), Article 5999. https://doi.org/10.3390/en15165999

Tacconi, L. (2018). Indonesia's NDC bodes ill for the Paris Agreement. *Nature Climate Change, 8*(10), 842. https://doi.org/10.1038/s41558-018-0277-8

Türkmen, B. A., Duhbacı, T. B., & Özbilen, Ş. K. (2021). Environmental impact assessment of ceramic tile manufacturing: A case study in Turkey. *Clean Technologies and Environmental Policy, 23*(4), 1295-1310. https://doi.org/10.1007/s10098-021-02035-w

Vieira, A. W., Rosso, L. S., Demarch, A., Pasini, D., Ruzza, S. P., Arcaro, S., Ribeiro, M. J., & Angioletto, E. (2023). Life cycle assessment in the ceramic tile industry: A review. *Journal of Materials Research and Technology, 23*, 3904-3915. https://doi.org/10.1016/j.jmrt.2023.02.023

Wang, Y., Liu, Y., Cui, S., Sun, B., Gong, X., Gao, F., & Wang, Z. (2020). Comparative life cycle assessment of different fuel scenarios and milling technologies for ceramic tile production: A case study in China. *Journal of Cleaner Production, 273*, Article 122846. https://doi.org/10.1016/j.jclepro.2020.122846

Wang, Y., Liu, Y., Gong, X., Wang, Z., & Gao, F. (2018). Life cycle assessment of typical sanitary ceramic products in China. In Y. Han (Ed.), *Advances in energy and environmental materials* (pp. 545-554). Springer. https://doi.org/10.1007/978-981-13-0158-2_57

Xian, Y., Li, N., & Wang, K. (2024). Carbon emissions marginal abatement cost and its influencing factors in the construction and hygienic ceramics manufacturing industries in China. *Environmental Impact Assessment Review, 104*, Article 107352. https://doi.org/10.1016/j.eiar.2023.107352

Ye, L., Hong, J., Ma, X., Qi, C., & Yang, D. (2018). Life cycle environmental and economic assessment of ceramic tile production: A case study in China. *Journal of Cleaner Production, 189*, 432-441. https://doi.org/10.1016/j.jclepro.2018.04.112

Yin, Q., Du, W. J., & Cheng, L. (2017). Optimisation design of heat recovery systems on rotary kilns using genetic algorithms. *Applied Energy, 202*, 153-168. https://doi.org/10.1016/j.apenergy.2017.05.072

Yuan, Q., Zhang, J., Robert, D., Mohajerani, A., Tran, P., Zhang, G., & Pramanik, B. K. (2024). Life cycle assessment of ceramic tiles manufactured using industrial waste fly ash. *Journal of Building Engineering, 97*, Article 110775. https://doi.org/10.1016/j.jobe.2024.110775

Pertanika J. Sci. & Technol. 34 (3): 1625 - 1653 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Field Validation of IES-VE for Daylight Assessment in Tropical High-rise Low-cost Housing: Effects of Building Height, Layout Depth, and Façade Design

Nor Izyan Saleh[1]*, Mohd Firrdhaus Mohd Sahabuddin[1], Norafida Ab Ghafar[1], Zuraina Leily Awaluddin[1], and Bismiazan Abd.Razak[2]

[1]Department of Architecture, Faculty of Built Environment, Universiti Malaya, 50603 Kuala Lumpur, Malaysia
[2]Department of Architecture and Built Environment, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia

ABSTRACT

Daylight is an important component of indoor environmental quality, yet its distribution in tropical high-rise low-cost housing is often uneven due to variations in building height, façade orientation, and unit depth. This study aims to validate the Integrated Environmental Solutions – Virtual Environment (IES-VE) simulation software for assessing daylight performance in tropical high-rise low-cost housing in Kuala Lumpur. A quantitative approach was adopted, combining field measurements and simulation modelling. Field measurements were conducted at different floor levels and room depths to investigate horizontal and vertical daylight distribution within residential units. The Daylight Ratio (DR) was used to represent actual daylight conditions under tropical sky environments, while simulation results were evaluated using the Daylight Factor (DF). The comparison between field measurements and simulation results showed differences ranging from 15% to 20%, which fall within the commonly accepted validation threshold. Statistical validation using Mean Bias Error (MBE) and Root Mean Square Error (RMSE) indicated small deviations between measured and simulated values, while correlation analysis demonstrated a strong agreement with a coefficient of determination (R^2) of 0.9917. The findings confirm that IES-VE can reliably simulate daylight performance in tropical high-rise residential buildings. The validated simulation model provides a reliable platform for future climate-based daylight modelling and visual comfort assessments, including the evaluation of daylight availability, solar exposure, and glare performance. The

ARTICLE INFO

Article history:
Received: 12 October 2025
Accepted: 05 June 2026
Published: 19 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.15

E-mail addresses:
22054446@siswa.um.edu.my (Nor Izyan Saleh)
firrdhaus@um.edu.my (Mohd Firrdhaus Mohd Sahabuddin)
norafida@um.edu.my (Norafida Ab Ghafar)
zurainaleily@um.edu.my (Zuraina Leily Awaluddin)
bismiazan@ukm.edu.my (Bismiazan Abd.Razak)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

outcomes of this study contribute towards the development of evidence-based daylight design recommendations and support future improvements to daylighting guidelines for tropical high-rise residential buildings.

Keywords: Daylight performance, daylight ratio, IES-VE, radiance, residential

INTRODUCTION

Daylight plays an important role in human life and indoor environmental quality. It contributes to energy efficiency by reducing the reliance on artificial lighting while also improving occupants' well-being and mood (Du et al., 2022; Tanveer et al., 2023). Effective daylighting strategies, therefore, enhance indoor environmental quality and support human comfort, particularly visual comfort (DiLaura,2021).

Building regulations, standards, and rating systems are important tools used to evaluate daylight performance in buildings. Various countries adopt daylight assessment methods that correspond to their climatic conditions and technological capabilities. For instance, the static daylight metric known as the Daylight Factor (DF) has been widely applied, particularly in contexts where technical expertise and simulation resources are limited. In contrast, more technologically advanced countries such as the United Kingdom, the United States, and Japan increasingly employ dynamic daylight metrics through Climate-Based Daylight Modelling (CBDM), which allows daylight performance to be simulated using real climatic data.

However, the Daylight Factor (DF) assumes a uniform overcast sky, which may not adequately represent daylight conditions in tropical climates where sky conditions are highly variable (Dahlan et al., 2009). In tropical regions such as Malaysia, sky conditions change rapidly due to dynamic cloud formation. Studies by (Zain-Ahmed et al., 2002) indicate that approximately 85.6% of the sky conditions in Subang are classified as intermediate skies, while (Djamila et al., 2011) reported that around 70-90% of sky conditions in Kota Kinabalu fall within the intermediate sky category. These findings suggest that tropical daylight conditions are dominated by fluctuating cloud patterns rather than stable overcast or clear skies.

In addition, outdoor global illuminance levels in tropical regions are significantly higher than those in temperate climates. Measurements show that global illuminance may exceed 80,000 lux at noon and occasionally reach values above 100,000 lux. During typical daytime working hours (9:00 a.m. to 5:00 p.m.), outdoor illuminance commonly exceeds 20,000 lux (Lim & Heng, 2016). Because of these characteristics, daylight studies in tropical climates often employ measurements derived from actual indoor and outdoor illuminance ratios, commonly referred to as the Daylight Ratio (DR), which provides a more representative indicator of daylight availability under real sky conditions.

In Malaysia, several regulations incorporate daylighting requirements, including the Energy Efficiency and Use of Renewable Energy for Non-Residential Buildings (MS 1525:2019), the Energy Efficiency and Use of Renewable Energy for Residential Buildings (MS 2680:2017), and the Uniform Building By-Laws (UBBL) 1984. These regulatory standards primarily rely on descriptive recommendations such as a minimum opening area of 10% for daylighting and ventilation, window-to-floor ratio (WFR), and the use of static daylight metrics such as the Daylight Factor (DF). However, such approaches may not fully capture daylight performance in tropical climates where solar exposure is intense, and sky conditions are highly dynamic (Mirrahimi et al., 2022). Consequently, discrepancies may occur between simulated daylight performance and actual daylight behaviour in buildings.

Low-cost housing developments often face additional design constraints. Limited technical input during the design stage frequently results in simplified façade configurations, the absence of shading devices, and compact floor plans (Al-Ashwal et al., 2024; Gonzalez-Longo, 2019). These architectural characteristics may lead to uneven daylight distribution, where spaces close to windows receive abundant daylight while deeper interior areas receive significantly less daylight. Furthermore, building height also affects daylight availability. Lower floors may experience reduced daylight penetration due to surrounding obstructions, whereas upper floors may receive excessive direct sunlight that could cause glare and visual discomfort (Kalaimathy et al., 2023; Syed Fadzil et al., 2009) .

In 2015, the Malaysian government introduced a large-scale affordable housing initiative known as the People's Housing Programme (PPR), implemented by the National Housing Department under the Ministry of Urban Wellbeing, Housing and Local Government (Gonzalez-Longo, 2019). According to the UHLG annual report, 169 PPR projects were proposed nationwide, providing approximately 102,896 housing units (Ministry of Urban Wellbeing, 2016). Nearly 47% of these housing units are in Kuala Lumpur and are predominantly constructed as high-rise residential buildings (JPN, 2016; KPKT, 2017).

Given the increasing reliance on simulation tools in building design, accurate simulation models are required to predict daylight performance and guide design decisions. Integrated Environmental Solutions – Virtual Environment (IES-VE) is one widely used platform capable of simulating both static and dynamic daylight conditions. However, before simulation outputs can be reliably applied in tropical housing contexts, the simulation model must be validated against empirical field measurements.

Previous validation studies have reported simulation–measurement deviations of up to 20%, depending on building type and sky condition (Elsiana et al., 2023; Li et al., 2017). In addition to evaluating percentage differences, statistical indicators such as Mean Bias Error (MBE), Root Mean Square Error (RMSE), and correlation analysis are often used to assess the agreement between measured and simulated daylight data.

This paper, therefore, presents a combined fieldwork and simulation study aimed at validating the performance of IES-VE for daylight analysis in tropical high-rise low-cost housing. Field measurements of indoor illuminance were conducted across multiple floor levels and room depths within selected residential units, and the same conditions were replicated in IES-VE using the Radiance simulation engine. The comparison between field measurements and simulation outputs was carried out using static daylight indicators, where the Daylight Ratio (DR) derived from field measurements was compared with the Daylight Factor (DF) obtained from simulation.

The comparison between field measurements and IES-VE simulation results indicates differences ranging between 15% and 20%, which fall within the acceptable validation threshold reported in previous daylight simulation studies. In addition to percentage difference analysis, statistical validation methods, including Mean Bias Error (MBE), Root Mean Square Error (RMSE), and correlation analysis, were applied to evaluate the level of agreement between measured and simulated daylight conditions. Accordingly, this study focuses on the empirical evaluation of daylight performance in tropical high-rise low-cost housing through field measurements and simulation modelling. By comparing measured indoor daylight conditions with simulated outputs, the research validates the predictive reliability of the IES-VE simulation platform under tropical climatic conditions. This integrated approach establishes a methodological framework for evaluating daylight performance in real residential environments, which can support future research employing advanced dynamic daylight metrics such as Useful Daylight Illuminance (UDI), Spatial Daylight Autonomy (sDA), and Annual Sunlight Exposure (ASE).

LITERATURE REVIEW

Daylight in High-rise Housing Buildings

Daylighting is an important element to optimise the indoor environment quality by improving visual comfort and well-being (Sallan et al., 2025). For residential buildings, a good daylight design supports smooth daily activities, harmonises human mood and at the same time fosters productivity, while poorly designed daylight may cause glare and visual discomfort (Al-Ashwal et al., 2024; Al Horr et al., 2016; Hassan & Al-Ashwal, 2015). Artificial lighting can cause stress and eye strain; thus, a good daylight design can avoid this effect and, at the same time, can reduce the reliance on artificial lighting (Knoop et al., 2020; Wirz-Justice et al., 2021).

Previous studies have shown that daylighting was influenced by window characteristics such as size, placement and type of glazing as shown in Table 1. In terms of size, window ratios such as Window-to-Floor (WFR), Window-to-Wall (WWR), and Carpet Area to Window Ratio (CAWR) are important in determining the daylight performance.

Table 1
Previous study on daylighting in residential buildings

Author Location	Daylight Metric	Objective	Method	Key Finding
(Wong et al., 2004) Singapore	Daylight Factor (DF)	To assess the impact of external shading devices on daylighting and glare	Field measurements LIGHTSCAPE simulations	Use shading devices with adjustable angles to balance daylight and glare; avoid direct sky views.
(Syed Fadzil et al., 2009) Penang, Malaysia	Daylight Factor (DF)	To assess the effect of orientation and WWR on daylight levels	Field measurements Simulation	Reduce WWR to 10% or less in tropical climates to avoid glare and overheating.
(Dutta et al., 2017) Kolkata, India	-	To evaluate how window orientation and dynamic external shading affect cooling loads	TRNSYS and ECOTECT simulation	Use dynamic external shading aligned with sun path on south-facing windows.
(Achsani et al., 2018) Bandung, Indonesia	-	To evaluate daylight technologies for high-rise residential buildings in tropical climates	Literature review	Use diffuse light-guiding systems, especially anidolic systems, for deep floor plans in tropical climates
(Syed Husin & Hanur Harith, 2018) Selangor, Malaysia	Daylight Factor (DF)	To analyse the impact of window types and glazing on daylight performance	Field measurements	Use casement windows with obscure glass for privacy and balanced daylight; avoid fixed louvres with clear glass.
(Dabe & Adane, 2019) Nagpur, India	Useful Daylight Illuminance (UDI)	To evaluate daylight performance by floor level and orientation	Field measurements Simulation	Adjust window sizes by floor level and orientation to optimise daylight and thermal comfort.
(Rukiah et al., 2022) Penang, Malaysia	-	To assess air and lighting quality in low-cost high-rise housing	Field observation Survey	Improve corridor ventilation and daylight access in dense residential buildings
(Dhayal et al., 2023) India	-	To review factors influencing indoor visual comfort and assessment techniques	Systematic Literature Review	Promote dynamic lighting design and conduct more studies in residential settings
(Dinapradipta et al., 2024)	Daylight Factor (DF)	To evaluate the performance of adjustable bent louvres for daylighting in tropical apartments	RADIANCE and ECOTECT simulation	Use adjustable two-bent slat louvres for optimal daylight distribution in tropical apartments.

For example, Dabe and Adane (2019) found that optimal CAWR values vary between lower and upper floors to balance UDI and thermal comfort, while Syed Fadzil et al. (2009) showed that excessive WFR can lead to glare. Similarly, glazing properties strongly affect daylight transmittance and visual quality (Kalaimathy et al., 2023; Syed Husin & Hanur Harith, 2018; Wang et al., 2023).

Daylighting performance further improved with daylighting technologies by integrating shading devices, especially in tropical countries. Studies show that horizontal louvres, bent slats, and light shelves can improve daylight distribution, reduce glare, and help deep floor plans daylight penetration (Al-Tamimi & Fadzil, 2011; Dinapradipta et al., 2024; Dutta et al., 2017; Iqbal et al., 2023). Advanced shading technologies, such as the dynamic shading system also proven to achieve fifteen per cent (15%) of energy saving and at the same time can improve the visual comfort (Dutta et al., 2017). These findings show that integrating daylight strategies can optimise the daylight usage as well as reduce the use of artificial lighting in residential buildings. The other benefit of daylighting is toward the occupant's well-being (Tanveer et al., 2023). Recent research shows that indoor visual comfort is influenced by environmental factors, design strategies, and user perceptions (Dhayal & Jha, 2023). In Malaysia, a field study was conducted in low-cost high-rise housing and shows that the residents were experiencing inadequate daylight, particularly at the lower floor level, forcing the residents to depend on artificial lighting, including during the daytime, contributing to dissatisfaction and poor living conditions of the residents. Thus, a well-planned architectural design that addresses the daylight issues is beneficial to the residents, thereby creating a more livable and comfortable residential environment (Rukiah et al., 2022).

Although technical practitioners were aware of the importance of daylight strategies in building design, it is noticeable that most of the common practices still rely on static metrics such as Daylight Factor and simulation-based evaluations, with limited field validation (Nazari & Matusiak, 2024). While dynamic metrics like UDI, sDA, and ASE offer greater reliability, they are rarely applied in low-cost high-rise housing studies (Kalaimathy et al., 2023; Nazari & Matusiak, 2024). Moreover, most fieldwork is confined to single floors, overlooking vertical variation in daylight access. This limits the transferability of findings to real-world contexts. The following subsection outlines the research gap identified from the review and highlights the novelty of the present study.

Research on daylighting in vertical residential buildings has expanded considerably over recent years. However, there are still notable gaps, especially when it comes to affordable housing in tropical climates. Many existing studies tend to depend heavily on computer simulations or theoretical frameworks, with limited validation through actual field measurements. These methods can affect the accuracy and reliability of the findings. Furthermore, in terms of building typology, not much research that focuses on high-rise

residential buildings compared to low-rise residential buildings. Thus, the real challenge and the specific needs in high- rise residential buildings remain unexplored, especially in a tropical country like Malaysia, where extreme climatic factors really affect the daylight performance in buildings. Moreover, existing fieldwork studies for high-rise residential building focusing within single floors, without considering the high-rise setting of the buildings.

Simulation Programs, Integrated Environmental Solutions-Virtual Environment (IESVE)

Over the past fifty years, the development of building performance simulation tools has significantly advanced in the construction industry and has become an important tool in evaluating and improving energy efficiency as well as sustainability in buildings. These tools help technical experts to simulate and analyse many aspects of building services, such as indoor comfort, energy efficiency and building life cycle cost, such as evaluating the thermal conditions of emergency shelters, which was notably utilised by the U.S. government during disaster response efforts (Attia et al., 2009). As these technologies continue to evolve, experts have used these tools not only to analyse energy efficiency but also to predict environmentally friendly design to ensure a better life quality and sustainability for a healthier planet (Crawley et al., 2008).

ECOTECT, EnergyPlus, IES-VE, TRNSYS, and ANSYS are several simulation programs that have been used in the construction industry. For example, research by Ozorhon and Caglayan (2011) focused on assessing the effectiveness of EnergyPlus and IES-VE in designing energy-efficient buildings by comparing these tools to determine their suitability for Building Information Modelling (BIM)-based simulations. The simulations focused on window size and building orientation, and the results showed that IES-VE is more user-friendly, offering easier parameter management and reporting, and supports daylighting analysis.

Thus, this paper presents a combination of fieldwork measurement and simulation measurement that aimed to validate the performance of IES-VE in terms of daylight analysis for high-rise low-cost housing. The fieldwork measurements involved measuring indoor and outdoor illuminance within a depth of typical units located at different floor levels (upper floor level and lower floor level), and the same scenarios were modelled in IES-VE using its Radiance engine. The results were then compared by using static daylight metrics, Daylight Factor (DF) and the introduction of Daylight Ratio (DR), which is more relevant in tropical climates, which remain fundamental to both Malaysian standards and international codes. The study demonstrates that IES-VE predictions differ from field measurements by approximately 15–20%, within the acceptable range reported in prior daylight validation studies. These findings confirm IES-VE's suitability as a predictive tool for assessing daylight in tropical residential housing, while also addressing the challenges in

the field of daylight distribution for high-rise building typology. By addressing these issues, this research can provide further recommendations for using dynamic daylight metrics for more comprehensive results. By doing so, this research will provide comprehensive results tailored to the local context that can help to standardise better building regulations and standards. Ultimately, this study aims to promote the use of dynamic daylight metrics by applying a reliable CBDM software, which can lead to better building design by reducing the use of artificial lighting while optimising the use of daylighting to shape a better and healthier environment for the residents.

METHODOLOGY

This study adopts a quantitative approach involving fieldwork measurements of daylight conditions in selected high-rise low-cost housing developments in Kuala Lumpur. To support model validation, the fieldwork concentrated on the measurement of fundamental luminous parameters, particularly indoor illuminance, which served as the empirical dataset for evaluating simulation accuracy. Mardaljevic et al. (2016) emphasised that confidence in daylight simulation predictions should first be established through comparisons with measured illuminance data from real occupied spaces before more advanced daylight assessments are undertaken. Therefore, the present study adopted illuminance-based static daylight metrics as the validation framework. Once validated, the IES-VE model can be reliably applied in subsequent research stages to assess annual daylight performance and visual comfort using climate-based daylight modelling approaches.

The exact building characteristics were modelled in SketchUp and subsequently exported to the Integrated Environmental Solutions – Virtual Environment (IES-VE) platform to perform daylight simulations using the IES-VE Radiance engine. The accuracy of the simulation model in predicting daylight performance under tropical conditions was assessed using standard static daylight metrics as benchmark indicators. This combined approach aims to provide a clearer understanding of daylight performance in high-rise low-cost housing and to support improved daylight-responsive design practices. By integrating empirical field measurements with simulation modelling, the study enables a direct comparison between measured daylight conditions and simulated outputs to evaluate the predictive reliability of the simulation model under tropical climatic conditions (Quek & Jakubiec, 2021).

Case Study Selection

An initial observation of ten (10) PPR developments in Kuala Lumpur was conducted to identify representative high-rise low-cost housing typologies. From this observation, two projects were selected representing Generation 1 and Generation 2 PPR design. These 2 projects with simple façades and minimal shading elements represent typical Malaysian low-cost housing design (Gonzalez-Longo, 2019). Case study one (1) PPR Intan Baiduri

represents an older generation of housing (built before the year 2000) with a limited basic façade and small floor area; 19.62 m² as shown in Figure 2 and Figure 4, while case study two (2) PPR Seri Aman reflects a newer generation development (post-2019) that incorporates basic façade modifications such as horizontal fins and larger floor area; 20.89 m² as shown in Figure 3 and Figure 5. Case studies 1 and 2 were further categorised into sub-cases (1 and 1a; 2 and 2a) to represent the respective floor levels, as presented in Table 2. Both projects are located within a 1.09 km radius to ensure similar outdoor daylight conditions as shown in Figure 1.

The inclusion of two generations of housing allows for comparative analysis of how daylight conditions vary with design evolution while still being representative of the low-cost housing typology (Gonzalez-Longo, 2019). These developments share common features of Malaysian affordable housing: high-density, narrow floor plan, repetitive unit layouts, single-sided fenestration design, making them suitable testbeds for evaluating daylight distribution and validating simulation outputs. The selection of these case studies also allows comparison between buildings with different façade treatments and orientations, which may influence daylight penetration patterns.

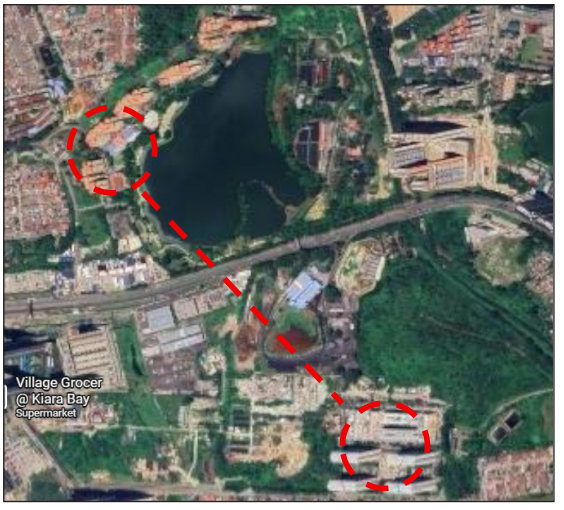


Figure 1. The distance between 2 PPR is 1.09 km

Figure 2. PPR Intan Baiduri

Figure 3. PPR Seri Aman

Table 2
Detailed overview of the on-site low-cost housing units

Case Study	Building Height (m)	Measured Point	Orientation	Floor Area (m²)	Window Area (m²)	WFR %	Shading Devices
1	Gen 1 45m	T01 T02 T03	East	19.62	1.94	9.89	No shading
1a	Gen 1 24m	T01 T02 T03	East	19.62	1.94	9.89	No shading
2	Gen 2 54m	T01 T02 T03	North	20.89	2.16	10.34	With shading
2a	Gen 2 30m	T01 T02 T03	North	20.89	2.16	10.34	With shading

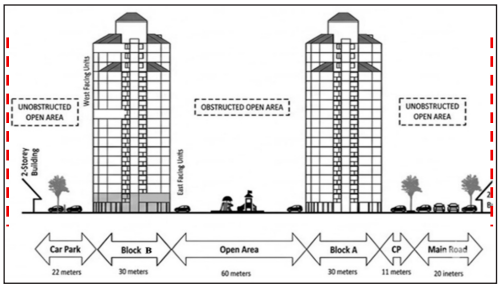


Figure 4. PPR Intan Baiduri Contextual Elevation

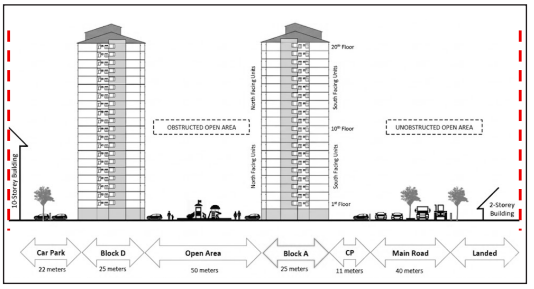


Figure 5. PPR Seri Aman Contextual Elevation

Fieldwork Measurement

Instruments

Indoor illuminance levels were measured using a Lux meter as shown in Figure 8, Onset HOBO U12-012 Temperature/Humidi ty/Light and HOBO Temp/RH/Light/Ext-Analogue MX 1104 with data loggers that had been calibrated to ISO/CIE standards as well as recommended by IESNA 10th Edition as shown in Figure 7, which was checked against a reference source before each measurement session to ensure accuracy as shown in Table 3. Simultaneous outdoor illuminance was recorded using a lux meter (HOBO Temp/RH/ Light/Ext-Analogue MX 1104 with data loggers) positioned in an unobstructed location to provide reference values for calculating Daylight Ratio (DR).

Before field deployment, all instruments were collocated under identical lighting conditions for five days to verify measurement consistency. The readings showed stable agreement within each instrument type, with minor deviations of approximately 0.9 and 0.54.

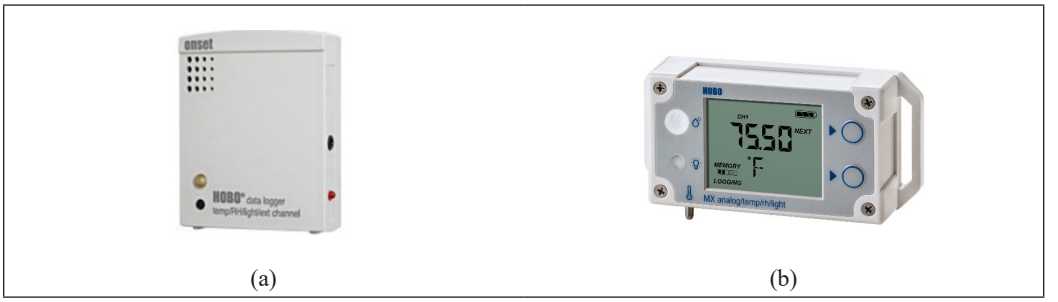
These differences were considered negligible relative to the daylight illuminance range recorded during the fieldwork, as shown in Figure 6.

Figure 6. Verification of measurement consistency

Equipment Type	Purpose	Measurement Uncertainty and Range	Meter Characteristics	Brand Name[a] Examples
Illuminance meter	Establish functional performance of baseline and new lighting equipment	Uncertainty: ±3% Range: ≤0. 1 fc (0.01 lux) to ≥10,000 fc (100,000 lux)	<3% deviation from cosine function for reported single value or ≤10% at incidence angle of 60° for multiple angle reported values. Spectral response within 10% of the CIE spectral luminous efficiency function	Minolta Photo Research Cooke Extech Amprobe Solar Light
Power meter	Establish true RMS power draw of baseline and new lighting equipment	Uncertainty: ±3% Decimal precision of 0.01 amp	Power factor (PF) calculation	Fluke 39/41/41B Extech 4KC20 AEMC 3910
Light on/off data logger	Measure run time of lighting fixtures	Uncertainty: ±1 minute per week; Light threshold adjustment range: 1–100 fc (10–1,000 lux)		Onset Computer Hobo Loggers Dent Instruments SmartLogger Omega OM-53

(a) Brand names listed are examples only. Associated products **may not** meet all of the requirements in this table. Verification of the individual equipment is required.

Figure 7. Example Instrument and Specification (IESNA 10th Edition)



(a) (b)

Figure 8. a) Onset HOBO U12-012 Temperature/Humidity/Light Data Logger; b) HOBO Temp/RH/Light/ Ext-Analogue MX 1104

Table 3
Instruments reference value of accuracy

Equipment	Parameter Measured	Measurement Range	Accuracy	Logging Interval
Onset HOBO U12-012 Temperature / Humidity / Light Data Logger	Illuminance (lux)	0 – 90,000 lux	±5% of reading	1-minute interval
	Air temperature	-20° to 70°C	± 0.35°C from 0° to 50°C	1-minute interval
	Relative humidity	5% to 95% RH	±2.5% RH	1-minute interval
HOBO Temp/RH/Light/ External-Analogue MX1104 Data Logger	Illuminance (lux)	0 to 167,731 lux (15,582 lum/ft2)	±10% typical for direct sunlight	1 second to 18 hours
	Air temperature	-20° to 70°C (-4° to 158°F)	±0.20°C from 0° to 50°C (±0.36°Ffrom 32° to 122°F)	1 second to 18 hours
	Relative humidity	0% to 100% at -20° to 70°C (-4° to158°F)	±2.5% from 10% to 90% (typical)	1 second to 18 hours

Measurement Points

For each project, measurements were conducted at two floor levels representing upper and lower zones, resulting in four case studies. A total of fourteen (14) measurement points (Figure 11) were used, consisting of twelve indoor points (three points per unit) and two outdoor reference points installed outside the highest unit to record external illuminance, as shown in Figure 9. Within each residential unit, measurement points were established across the living room and dining areas, as these spaces represent the primary occupied zones where most daytime activities take place and therefore require adequate daylight availability for visual tasks and general indoor comfort, as shown in Figure 9 and Figure 10. Illuminance measurements were recorded at 0.75 m above the floor level, representing the standard horizontal working plane height commonly adopted in daylight performance assessment. This measurement height was selected to ensure consistency with the recommended procedures outlined by the Illuminating Engineering Society (IES) guidelines (Kruisselbrink et al., 2018) (*Illuminating Engineering Society THE LIGHTING HANDBOOK 10th Edition | Reference and Application*, 2011). The use of this measurement height also follows the methodology adopted in previous daylight field measurement studies conducted in tropical environments (Elsiana et al.,2023; Lim & Heng, 2016), ensuring methodological consistency and comparability of daylight performance evaluation.

There are three (3) measurement points established within each unit, namely T01, T02, and T03, to represent the daylight distribution within the living and dining area based on

the dimensions of the investigated residential units. The combined depth of the living and dining area was approximately 5.6 m in PPR Intan Baiduri and 5.9 m in PPR Seri Aman. Measurement points were positioned at regular 1.5 m intervals extending from the window opening towards the rear of the space to capture the daylight gradient across the occupied zone. This arrangement enabled the assessment of daylight penetration of the area from the façade. According to recommended daylight measurement practices, spaces should be divided into daylighted zones based on distance from the window and the activity within the space, where each zone is represented by a measurement point placed at a critical or representative location of typical illuminance levels. For continuous measurements, one measurement point is sufficient to represent the illuminance condition within each daylighted zone at the working plane level (Kruisselbrink et al., 2018). Therefore, three (3) measurement points were considered sufficient to represent the daylight distribution within the living and dining area, particularly within the active daylight zone extending from the window opening towards the rear of the space, considering the depth of each residential unit.

To capture vertical daylight variation, measurements were conducted at both lower floor levels (approximately 24 m and 30 m above ground level) and upper floor levels (approximately 45 m and 54 m above ground level) in each building. In addition, units with different orientations (south-oriented and east-oriented) were selected to account for variations in solar exposure. This measurement configuration enables the evaluation of daylight distribution within the daylight active zone, both horizontally across the room depth and vertically across different building heights. In total, fourteen (14) measurement points were monitored, consisting of twelve indoor points and two outdoor reference points. Illuminance data were recorded using data loggers at 5-minute intervals over a seven-day monitoring period, allowing continuous measurement of daylight conditions under the same sky conditions for all points.

Figure 9. Placement Point T1, T2 and T3 in PPR Intan Baiduri

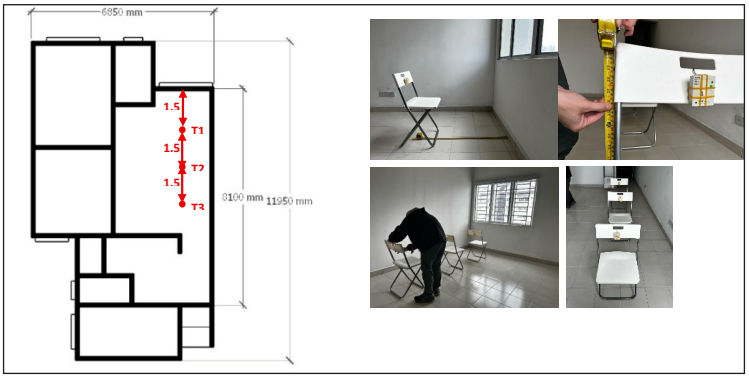


Figure 10. Placement Point T1, T2 and T3 in PPR Seri Aman.

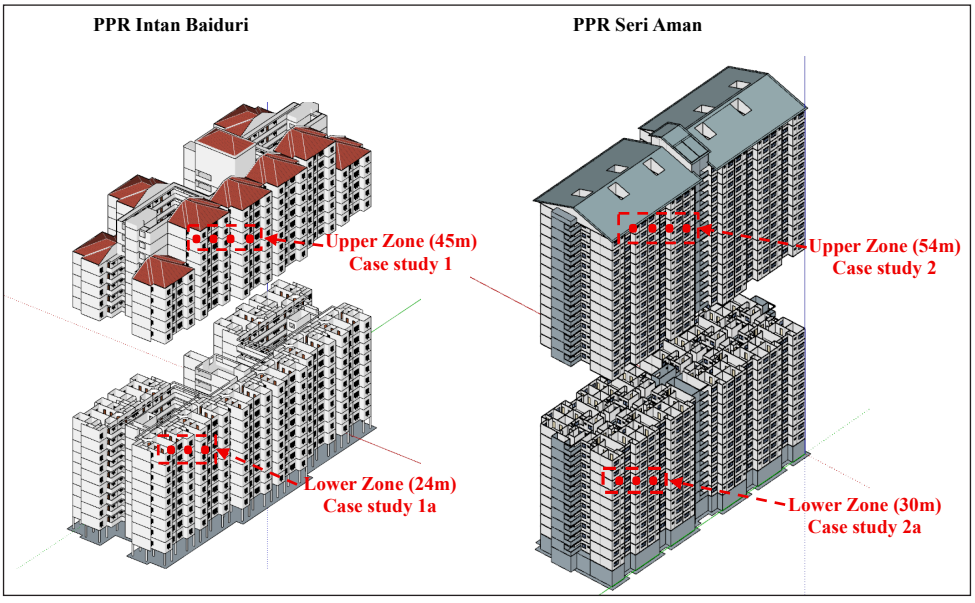


Figure 11. 14 points were set on the same day from 16/07/25 to 22/07/25

Fieldwork Procedure

To ensure accurate assessment of daylighting conditions, all artificial lighting in the selected units was switched off during data collection. Since the units were unoccupied, measurements could be conducted without any interference from internal light sources, allowing the study to focus solely on the impact of natural daylight within the living spaces. Data collection for each unit was conducted on days with stable sky conditions, on the 16th of July 2025 until the 22nd of July 2025 (7 days), simultaneously across different building heights in two low-cost housing schemes. The total number of points involved in the fieldwork is fourteen (14) points, with two (2) points to measure the outdoor

illuminance, and the time used for daylight measurement is between 08:00 and 18:00. Simultaneity between indoor and outdoor measurements was ensured by configuring all data loggers using HOBO logging software before fieldwork. The instruments were synchronised and set to record illuminance at a 5-minute logging interval, allowing indoor and outdoor illuminance data to be captured simultaneously throughout the monitoring period, under intermediate sky conditions, as mentioned in Malaysian Standards. To further minimise the influence of short-term sky variability, the recorded field measurements were verified against corresponding weather station data for each monitoring day to confirm the prevailing sky conditions. This verification procedure was consistently applied throughout the fieldwork period to ensure the reliability of the measured daylight data. The selection of the lowest and highest daylight conditions was based on the average indoor illuminance values, which is a simplified and acceptable method recommended by IES guidelines for determining representative illuminance levels. After the seven-day monitoring period, the average lux values from the three indoor measurement points were compared. The results showed that 18/07/2025 recorded the lowest average illuminance, while 21/07/2025 recorded the highest. This approach allows daylight conditions to be evaluated based on daily illuminance variations rather than annual calculations, using multiple monitoring days, which improves reliability compared to a single-day measurement (Al-Ashwal et al., 2024). This research, therefore, applies appropriate static daylight metrics to evaluate fundamental daylight performance, which are further described in the Criteria of Analysis section.

The selection of the fieldwork period in July was also informed by the daylight availability characteristics described in MS 2680:2017. As illustrated in Figure 12 of the standard, the monthly average hourly illuminance in tropical regions shows that March generally represents the maximum daylight condition, December represents the minimum daylight condition, while June represents the average daylight condition throughout the year. Since July occurs within the mid-year period following June, the outdoor illuminance levels are typically closer to the annual average range rather than the extreme maximum or minimum conditions. Conducting field measurements during this period, therefore, provides relatively stable daylight conditions with moderate variation in outdoor illuminance. This stability is important for daylight validation studies because it reduces extreme fluctuations in sky brightness and allows a more reliable comparison between measured daylight ratios (DR) and simulated daylight factors (DF). Consequently, selecting July as the fieldwork period supports the generation of representative daylight data that can be used to produce more consistent and reliable validation results. Nevertheless, if field measurements were conducted during March or December, when absolute illuminance levels may differ due to variations in solar position and sky conditions, the overall daylight distribution patterns and the validation relationship between measured and simulated results are expected to remain generally consistent.

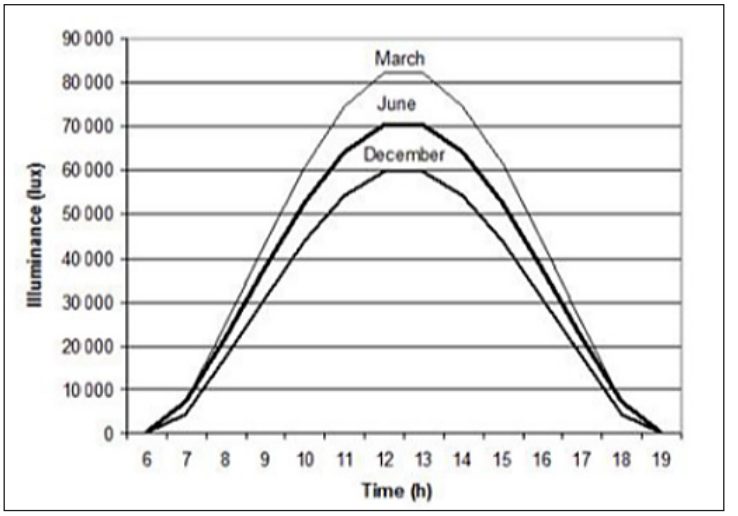
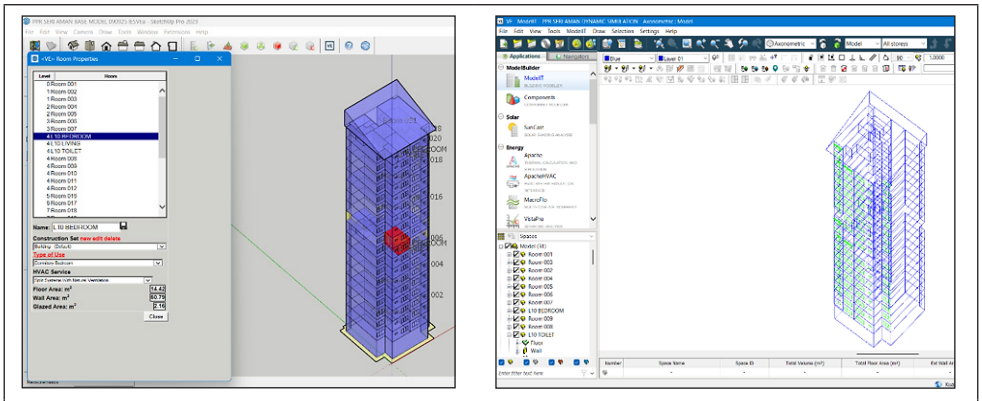


Figure 12. Monthly average hourly illuminance in March (maximum), June (average) and December (maximum)

Simulation Setup Integrated Environmental Solutions – Virtual Environment (IES-VE)

The case study buildings were generated using AutoCAD as shown in Figures 4 and 5, modelled in SketchUp, then exported to the IES-VE platform using the Radiance-IES module. Building geometry, unit layouts, wall finishes, and glazing specifications, as shown in Figure 13, were reconstructed based on architectural drawings and site plan layout. The building geometry was created in SketchUp and exported to IES-VE using the VE plug-in. During the export process, only fully enclosed spaces are recognised as "rooms" by IES-VE, which are automatically highlighted in blue within the SketchUp interface. Any incomplete geometry appears yellow and cannot be exported as a room, ensuring that only properly enclosed spaces are transferred to the simulation model. This process helps maintain geometric fidelity between SketchUp and IES-VE. After export, the room dimensions, wall thickness, and window openings were rechecked using IES-VE tools to confirm that the geometry matched the original SketchUp model before running the daylight simulation, as shown in Figure 13.

External obstructions were not explicitly incorporated into the digital model because the selected case study buildings have minimal surrounding obstruction conditions. The lower measurement level (approximately 24 m above ground, Level 8) is already above typical ground-level elements such as trees or small urban features. In addition, the selected PPR blocks are located with wide building setbacks and without adjacent high-rise buildings directly facing the measured façades, which minimises potential shading effects from surrounding structures. The façade design of the selected units also does not include balconies that could obstruct daylight penetration. Furthermore, the



(a) (b)

Figure 13. a) Building modelling and b) export building to IES-VE

Table 4
Construction material of 3D model in IES-VE

Case Study	Orientation	Floor Area (m²)	Wall / Floor Finish	Wall/Floor Reflectance	Window Area (m²)	Glazing Visible Light Transmittance
1	East-West	19.62	Paint/cement Render	0.7/0.4	1.94	0.7
1a	East-West	19.62	Paint/cement Render	0.7/0.4	1.94	0.7
2	North-South	20.89	Paint/tiles	0.7/0.6	2.16	0.7
2a	North-South	20.89	Paint/tiles	0.7/0.6	2.16	0.7

absence of significant external obstructions was verified through contextual elevation drawings generated using AutoCAD software based on actual site dimensions and building geometry. These technical drawings provided a reliable and accurate representation of the surrounding built environment and confirmed that no substantial neighbouring structures interfered with the daylight access to the selected units. Therefore, external obstruction was considered negligible for the purposes of both the field interpretation and the simulation model. However, the building orientation, window-to-wall ratio (WWR), glazing visible light transmittance (VLT), and external shading devices such as horizontal fins were carefully modelled to replicate the actual conditions of the case study buildings as shown in Table 4. The reflectance values as well as the glazing visible light transmittance used in the model were obtained from product catalogue references and Jabatan Kerja Raya (JKR) specifications and supplemented with default material properties from the IES-VE material library.

Kuala Lumpur, with a latitude of 3.13º and a longitude of 101.55º, is set to represent the tropical climate. The Kuala Lumpur (Subang) IWEC weather file in APlocate IES-VE was used, providing monthly climate data including sun path, direct solar flux, and

sky illuminance distribution Figure 6. In Radiance-IES engine settings, WP Zone Data (Dynamic) is used with 0.75m WP Height and 0.25m WP grid as well as applying the rtrace_dc method, a high-resolution sky model (2305 sky patches), and 20% ground reflectance, while other parameters followed the standard Radiance-IES default configuration. The sky models were refined according to the recommendations of (Elsiana et al., 2023; Lim et al., 2017; Tsang et al., 2022; Wei et al., 2023), which are overcast and intermediate sky to ensure reliable daylight propagation and accurate light distribution within the interior spaces, as shown in Figure 14. They work uses for North-South is overcast sky, while for East-West is intermediate sky condition. The simulation settings, including date and time, were aligned with the corresponding fieldwork measurement periods to replicate the actual conditions and ensure accurate validation of the simulation model, as shown in Table 5.

The simulation generated outputs for illuminance levels (Lux) and Daylight Factor (DF), which were extracted at the same measurement points used during field surveys to enable direct comparison. This point-to-point comparison enables validation of the simulation results against the measured daylight conditions. The simulation framework in this research was designed to allow future extension with dynamic daylight metrics such as Useful Daylight Illuminance (UDI), Spatial Daylight Autonomy (sDA), and Annual Sunlight Exposure (ASE), which could provide a more comprehensive evaluation of daylight performance.

Table 5.
Summary of case study orientation, dates, time intervals, and applied sky conditions

Case Study	Orientation	Date	Time	Sky Condition
1	East-West	18/07/2025 and 21/07/2025	8.00 until 16.00 (1 hour interval)	Intermediate
1a	East-West	18/07/2025 and 21/07/2025	8.00 until 16.00 (1 hour interval)	Intermediate
2	North-South	18/07/2025 and 21/07/2025	8.00 until 16.00 (1 hour interval)	Overcast
2a	North-South	18/07/2025 and 21/07/2025	8.00 until 16.00 (1 hour interval)	Overcast

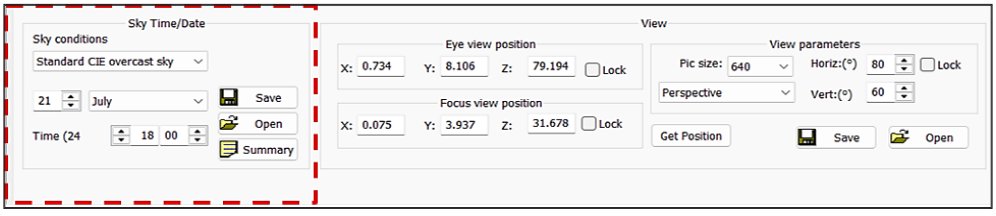


Figure 14. Simulation of dates, time intervals and applied sky conditions in IES-VE

CRITERIA OF ANALYSIS

A traditional static metric known as Daylight Factor (DF), a static daylight metric is used widely to calculate daylight performance in buildings. The calculation of DF uses the ratio

of indoor illuminance over the outdoor illuminance under an overcast sky. However, in tropical regions like Malaysia, the use of DF to measure the daylight performance becomes a big constraint as the Malaysian sky conditions are very unpredictable. As noted by Lim et al. (2010) and further emphasised by Al-Ashwal et al. (2024) DF does not really capture the actual time-varying environmental influences, building orientation, or solar path variations. Thus, this study applied the use of Daylight Ratio (DR), which uses the ratio of actual indoor illuminance compared to actual outdoor illuminance values under actual sky conditions. This method, supported by the work (Al-Ashwal et al., 2024; Kalaimathy et al., 2023; Lim & Heng, 2016), gives a more accurate on-site measurement of daylight availability, which is relevant in tropical environments where the sky condition is variable. Equation 1 for DR is as follows:

$$DR = \frac{Indoor\ Illuminance\ (WPI)}{Outdoor\ Illuminance} \times 100 \quad [1]$$

As for the IES-VE validation process, data gathered from fieldwork measurements were compared with IES-VE simulated data measurements. Both illuminance values and static daylight metrics are considered, with the Daylight Factor (DF) derived from IES-VE simulations and the Daylight Ratio (DR) obtained from fieldwork measurement was used as a basis for comparison. The comparison was undertaken because both DR and DF are dimensionless ratios that normalise indoor illuminance against outdoor illuminance, thereby reducing the influence of fluctuating absolute illuminance levels. Although DF is traditionally associated with the CIE overcast sky assumption, previous studies have demonstrated that DR can serve as a practical field-based equivalent for validating daylight simulation models under real conditions. In the present study, the IES-VE simulation was configured using an intermediate sky condition to replicate the prevailing sky conditions observed during the field measurement period. However, despite the flexibility of the sky model settings, the software continues to report the simulation output using the Daylight Factor (DF) terminology. Therefore, the purpose of the comparison was not to establish absolute equivalence between DR and DF, but rather to evaluate the consistency of daylight distribution patterns and relative daylight performance between measured and simulated conditions across the selected measurement points. Furthermore, (Elsiana et al., 2023b) demonstrated that the correlation between simulation and field measurement results based on DR and DF was strong, significant, and positive, supporting the use of these metrics for evaluating the consistency between measured and simulated daylight performance.

To ensure accuracy, the simulated measurement points are aligned with point of lux meter during fieldwork measurement, with reference height of 0.75 m above the floor level and corresponding room positions.

RESULTS AND DISCUSSION

Within the unit, the results show a reduction from the window point (T01) to the back of the dining room point (T03). The east- facing unit, which is Case Study 1, had reductions of daylight penetration from point T01 to T03 were 63.3% at the upper level and 52.2% for Case Study 1a, while for Case Study 2, which is a north-facing building, the reductions were 52.6% and 54.5% for Case Study 2a. As for the comparison between floor levels, the fieldwork result for Case Study 1 indicates that daylight at the window zone (T01) was 23.3% lower at the lower level, whereas in Case Study 2, the lower level recorded 13.6% lower values than the upper floor. These findings highlight both within-unit daylight loss with depth and indicate within-floor daylight inequity, varying by orientation, façade design, and window-to-wall ratio (WWR).

Next, the field measurement data and the IES-VE simulation outputs are analysed and compared to determine the percentage differences in accuracy and precision. The calculation of these differences follows Equation 2, as recommended by (Ab.Razak, 2023; Lim et al., 2010; Maamari et al., 2006).

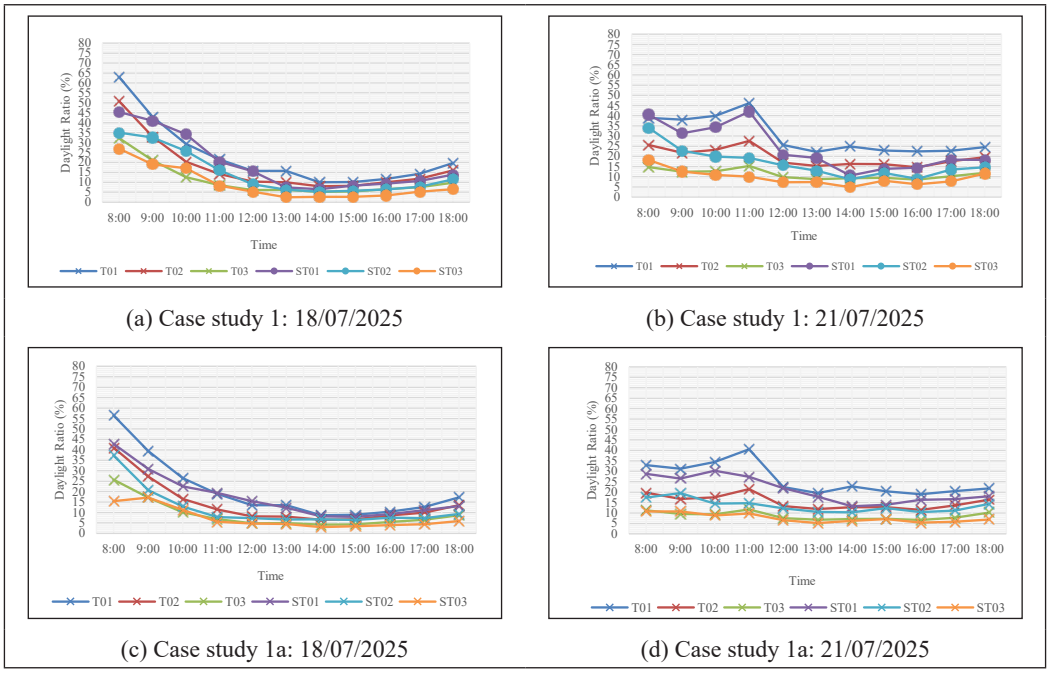
$$PD = \frac{(SM - FM)}{FM} \times 100 \quad [2]$$

Where,
PD. Percentage Difference
SM. Simulation Measurement
FM. Fieldwork Measurement

Maamari et al. (2006) recommended that the acceptable percentage difference between field measurements and simulation results should range between 10% and 20%. This guideline is further supported by the validation of IES-VE studies conducted by Elsiana et al. (2023) who reported differences of 7% and 12% differential of DR in an office building. Heng et al. (2016) Also mention the DR is from 11% - 15% for light pipe performance in an office building. In this study, following the benchmark provided by these scholars, 20% maximum differences were allowed between fieldwork measurements and IES-VE simulations to confirm the reliability and suitability of the software for current and future applications.

Within a two (2) day duration, from 8:00 a.m to 6:00 p.m, the average DR differences between field measurements and IES-VE simulations were compared. In Case Study 1, 18/07/2025 was selected as it showed the lowest lux condition compared to the other days. At points T01-T03 from fieldwork measurement, the average DR values ranged between 11-23%, while the corresponding IES-VE simulation results ranged between 9-19%.

The percentage differences (PD) recorded between fieldwork and simulation values for these points were 16% (T01), 18% (T02), and 20% (T03). Then, on 21/07/2025, where the highest lux condition was recorded, average DR values from field measurements ranged from 11-30%, compared to 10-24% from IES-VE simulations. The percentage differences (PD) were 20% (PD01), 15% (PD02), and 15% (PD03) as shown in Figure 15 and Table 6.



(a) Case study 1: 18/07/2025

(b) Case study 1: 21/07/2025

(c) Case study 1a: 18/07/2025

(d) Case study 1a: 21/07/2025

Figure 15. Daylight Ratio (DR) profiles for case study 1 and 1a, showing field measurements (T01-T03) and IES-VE simulations (ST01-ST03) at hourly intervals from 08:00 a.m. to 6:00 p.m.
Note. Data were collected at one-hour intervals to calculate the average daylight ratio (DR), which was then used for comparison between fieldwork measurements and IES-VE simulations

Table 6
Case study 1 and 1a: Average DR of fieldwork measurement and average DF of simulation in IES-VE, as well as the percentage differences

Date	Time	Fieldwork Measurement (FM) %			Simulated Measurement (SM) %			Percentage Different (PD) %		
Case Study 1		**T01**	**T02**	**T03**	**ST01**	**ST02**	**ST03**	**PD01**	**PD02**	**PD03**
18/07/25	08:00 a.m-06: 00 p.m.	30	20	11	24	17	10	20	15	15
21/07/25	08:00 a.m-06: 00 p.m.	23	17	11	19	15	9	16	18	20
Case Study 1a		**T01**	**T02**	**T03**	**ST01**	**ST02**	**ST03**	**PD01**	**PD02**	**PD03**
18/07/25	08:00 a.m-06: 00 p.m.	21	14	9	18	12	7	15	18	20
21/07/25	08:00 a.m-06: 00 p.m.	26	15	9	21	13	8	19	12	13

For Case Study 1a, on 18/07/2025, the average DR values obtained from field measurements at points T01-T03 ranged between 9-21%, while the corresponding IES-VE simulation results labelled as (ST01-ST03) ranged between 7-18%. The percentage differences recorded between field and simulation values for these points were 15% (T01), 18% (T02), and 20% (T03). On 21/07/2025, the average DR values from field measurements ranged from 9-26%, compared to 8-21% from IES-VE simulations. The corresponding percentage differences were 19% (PD01), 12% (PD02), and 13%(PD03) as shown in Figure 15 and Table 6.

For Case Study 2 and 2a, on 18/07/2025, the average DR values obtained from field measurements at points T01-T03 ranged between 9-19% for case study 2, while the average DR values obtained from field measurements at points T01-T03 ranged between 5-10% for case study 2a. The corresponding IES-VE simulation results ranged between 8-16% for case study 2, while for 2a, the range is between 4-9%. The percentage differences recorded between field and simulation values for these points were 17% (PD01), 9% (PD02), and 20% (PD03) for case study 2, while for case study 2a, the percentage differences were 7% (PD01), 14% (PD02), and 15% (PD03) as shown in Figure 16 and Table 7.

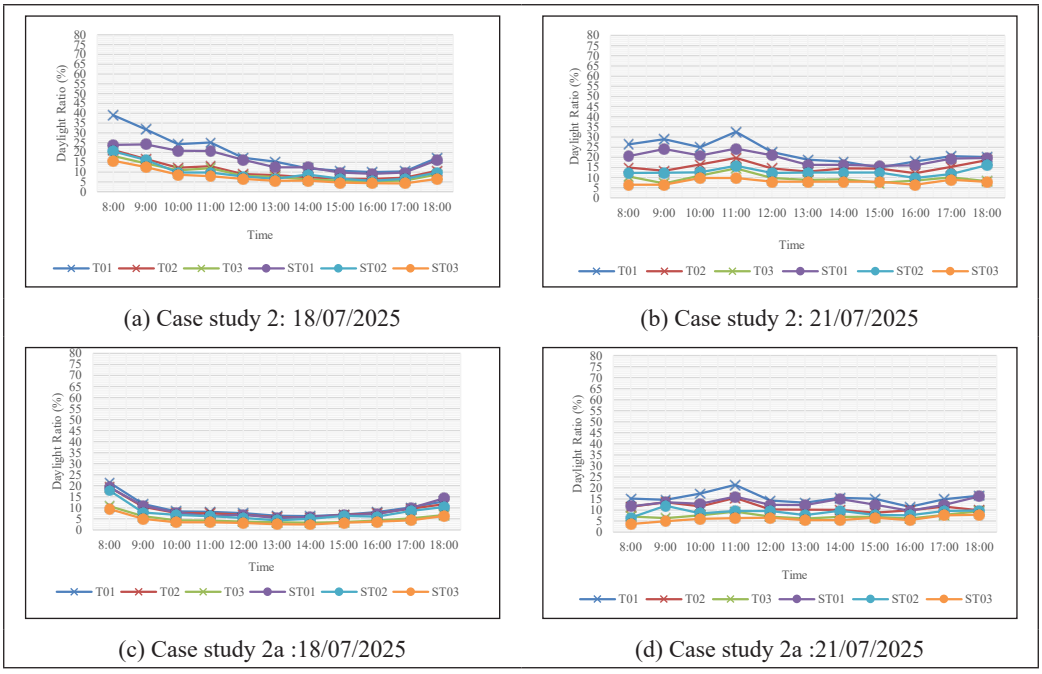


Figure 16. Daylight Ratio (DR) profiles for case study 1 and 1a, showing field measurements (T01-T03) and IES-VE simulations (ST01-ST03) at hourly intervals from 08:00 a.m. to 18:00 p.m.
Note. Data were collected at one-hour intervals to calculate the average daylight ratio (DR), which was then used for comparison between fieldwork measurements and IES-VE

Table 7
Case study 2 and 2a: Average DR of fieldwork measurement and average DF of simulation in IES-VE, as well as the percentage differences

Date	Time	Fieldwork Measurement (FM) %			Simulated Measurement (SM) %			Percentage Different (PD) %		
Case Study 2		**T01**	**T02**	**T03**	**ST01**	**ST02**	**ST03**	**PD01**	**PD02**	**PD03**
18/07/25	08:00 a.m-06: 00p.m.	19	11	9	16	10	8	17	9	20
21/07/25	08:00 a.m-06: 00p.m.	22	15	10	19	13	8	13	16	17
Case Study 2a		**T01**	**T02**	**T03**	**ST01**	**ST02**	**ST03**	**PD01**	**PD02**	**PD03**
18/07/25	08:00 a.m-06: 00p.m.	10	9	5	9	8	4	7	14	15
21/07/25	08:00 a.m-06: 00p.m.	15	11	7	13	9	6	15	20	19

On 21/07/2025, the average DR values obtained from field measurements at points T01-T03 ranged between 10-22% for case study 2, while average DR values obtained from field measurements at points T01-T03 ranged between 7-15% for case study 2a. The corresponding IES-VE simulation results ranged between 8-19% for case study 2, while for 2a, the range is between 6-13%. The percentage differences recorded between field and simulation values for these points were 13% (PD01), 16% (PD02), and 17% (PD03) for case study 2, while for case study 2a, the percentage differences were 15% (PD01), 20% (PD02), and 19% (PD03).

The results indicate that the IES-VE simulations were lower than the fieldwork measurement data, with average differences ranging from 15% to 20% across both test days. These differences show the range of ≤20% as recommended by (Elsiana et al., 2023; Heng et al., 2016; Maamari et al., 2006), therefore confirming the reliability of IES-VE accuracy in imitating daylight performance under tropical high-rise residential conditions. As for the sky condition to be applied in IES-VE simulation, the percentage difference (PD) provides a useful indicator for choosing the right sky condition, weather it is an intermediate or overcast sky. In this context, maintaining PD values below 20% ensures that the selected sky condition reflects actual tropical daylight patterns with acceptable accuracy.

Following the comparison between field measurements and IES-VE simulation results presented in the case study analysis, additional statistical indicators were calculated to further evaluate the agreement between measured and simulated daylight values. In addition to the percentage difference (PD) analysis, Mean Bias Error (MBE) and Root Mean Square Error (RMSE) were applied as statistical validation indicators to quantify the deviation between field measurements and simulation outputs. These indicators are widely used in building performance simulation validation studies to evaluate systematic error and the overall dispersion between measured and simulated datasets (Elsiana et al., 2023). The Mean Bias Error (MBE) represents the average tendency of the simulation model to either

overestimate or underestimate the measured values, while the Root Mean Square Error (RMSE) represents the overall magnitude of deviation between simulated and measured values. Lower values of RMSE and MBE indicate stronger agreement between simulation predictions and field measurement results.

The equations used to calculate the statistical indicators are presented in Equations 3 and 4.

$$\mathrm{MBE} = \frac{1}{\mathrm{n}}\sum(\mathrm{SM} - \mathrm{FM}) \quad [3]$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}}\sum(\mathrm{SM} - \mathrm{FM})^2} \quad [4]$$

Where
SM = Simulation Measurement
FM = Fieldwork Measurement
n = number of observations

The calculated validation indicators for all investigated case studies are presented in Table 7.

The RMSE values range between 1.01% and 3.71%, while the MBE values range between –0.93% and –3.53%. These relatively small deviations indicate a strong agreement between the measured daylight ratio values obtained from field measurements and the daylight factor values generated from the IES-VE simulation. The negative MBE values indicate that the simulation model slightly underestimates the measured daylight ratios, although the magnitude of this deviation remains small. These findings demonstrate that the simulation model can reproduce the daylight distribution observed during field measurements with a high level of accuracy, as shown in Table 8.

Table 8
Statistical validation between field measurement and IES-VE simulation

Case Study	Level	Daylight Condition	MBE (%)	RMSE (%)
PPR IB (1)	L15	Lowest Lux (18/07/25)	-2.83	3.03
PPR IB (1)	L15	Highest Lux (21/07/25)	-3.53	3.71
PPR IB (1a)	L8	Lowest Lux (18/07/25)	-2.43	2.59
PPR IB (1a)	L8	Highest Lux (21/07/25)	-2.67	2.86
PPR SA (2)	L15	Lowest Lux (18/07/25)	-2.07	2.20
PPR SA (2)	L15	Highest Lux (21/07/25)	-2.27	2.41
PPR SA (2a)	L10	Lowest Lux (18/07/25)	-0.93	1.01
PPR SA (2a)	L10	Highest Lux (21/07/25)	-1.97	2.07

Furthermore, the correlation between the measured daylight ratio (DR) and the simulated daylight factor (DF) is illustrated in Figure 17. The scatter plot shows a strong positive linear relationship between simulation and measurement results with a coefficient of determination of $R^2 = 0.9917$, indicating excellent agreement between field measurements and simulation outputs.

CONCLUSION

The results from the fieldwork measurements indicate that daylight distribution within the investigated residential units is uneven across the two housing developments, namely PPR Intan Baiduri (represented by Case Study 1 and Case Study 1a) and PPR Seri Aman (represented by Case Study 2 and Case Study 2a). The findings demonstrate that daylight penetration in high-rise low-cost housing units is significantly influenced by building height, layout depth, and façade configuration. In all investigated units, the window zone (T01) received more than half of the available daylight compared to the rear zone (T03) located deeper within the living space. This pattern confirms that daylight intensity decreases progressively with distance from the façade opening.

When comparing the two residential developments, Case Study 1 and Case Study 1a (PPR Intan Baiduri) recorded higher daylight availability compared with Case Study 2 and Case Study 2a (PPR Seri Aman). This difference indicates that façade design and building orientation play an important role in determining the amount of daylight entering interior spaces in high-rise low-cost housing. In addition, comparisons between floor levels show that upper floors generally receive higher daylight levels than lower floors due to increased sky visibility and reduced external obstruction. Despite these variations, the overall pattern of daylight reduction from the window point (T01) toward the rear of the living room (T03) remains consistent across all case studies.

These findings highlight the significant influence of building height, layout depth, and façade design on daylight penetration and distribution in high-rise residential environments. By analysing these factors, this study provides empirical evidence of the actual daylight performance conditions in high-rise low-cost housing in Kuala Lumpur.

To simulate daylight performance, the Integrated Environmental Solutions – Virtual Environment (IES-VE) software was employed. Therefore, validating the accuracy of the simulation model is essential to ensure its reliability for future applications. In this study, the Daylight Ratio (DR) was adopted as the primary indicator to represent the relationship between indoor and outdoor illuminance under actual sky conditions. The comparison between field measurements and IES-VE simulations revealed relatively small deviations, with percentage differences ranging between 15% and 20%, which remain within the widely accepted validation threshold of $\leq$20% reported in previous daylight simulation studies.

Further statistical validation using Mean Bias Error (MBE) and Root Mean Square Error (RMSE) demonstrated small deviations between the measured and simulated datasets, with RMSE values ranging between 1.01% and 3.71% and MBE values between –0.93% and –3.53%. In addition, correlation analysis between the measured daylight ratio (DR) and the simulated daylight factor (DF) showed a very strong positive relationship with a coefficient of determination of $R^2 = 0.9917$. These results confirm that the IES-VE simulation model provides reliable predictions of daylight performance within the investigated residential units and is suitable for daylighting analysis in tropical high-rise residential buildings.

International guidelines such as the Illuminating Engineering Society of North America (IESNA) and the Malaysian Standard MS 2680:2017 recommend the use of advanced daylight performance metrics, including Useful Daylight Illuminance (UDI), Spatial Daylight Autonomy (sDA), and Annual Sunlight Exposure (ASE). These dynamic metrics are typically evaluated using Climate-Based Daylight Modelling (CBDM) approaches, which analyse daylight availability and glare potential over an entire year using real climate data. However, reliable application of these dynamic metrics requires prior validation of the simulation environment to ensure that model predictions accurately reflect real daylight behaviour.

Once the reliability of the IES-VE simulation model has been established and appropriate sky conditions for tropical contexts have been identified, the software can be applied to evaluate these advanced CBDM-based daylight metrics in future research. The integration of dynamic daylight analysis will allow a more comprehensive assessment of visual comfort, glare potential, and daylight performance in high-rise residential buildings.

Overall, this study contributes to the understanding of daylight performance in tropical high-rise low-cost housing by combining empirical field measurements with validated simulation modelling. The findings provide a reliable methodological framework for evaluating daylight performance in dense urban residential environments and offer useful insights for improving façade design strategies and visual comfort conditions in future high-rise housing developments in Kuala Lumpur.

ACKNOWLEDGEMENT

The authors would like to thank Universiti Malaya for providing the research facilities used in this work. The authors also acknowledge the expertise and technical support provided by the university, which greatly facilitated the data analysis and production of this paper.

REFERENCES

Abd Razak, B. (2023). Verification of IES-VE reading accuracy of naturally ventilated Royal Malaysia Police (RMP) lockup cell thermal comfort performances

Achsani, R. A., Wonorahardjo, S., & Soelami, F. X. N. (2018). A vision of daylight technologies for high-rise residential buildings in the tropics. *IOP Conference Series: Earth and Environmental Science, 152*(1), Article 012013. https://doi.org/10.1088/1755-1315/152/1/012013

Al horr, Y., Arif, M., Katafygiotou, M., Mazroei, A., Kaushik, A., & Elsarrag, E. (2016). Impact of indoor environmental quality on occupant well-being and comfort: A review of the literature. *International Journal of Sustainable Built Environment, 5*(1), 1-11. https://doi.org/10.1016/j.ijsbe.2016.03.006

Al-Ashwal, N. T., Lim, Y. W., Dalumo, D. B., Leng, P. C., Rahamn, N. A., & Mahdzar, S. S. S. (2024). Effect of unit layout and façade design on high-rise apartment indoor daylight conditions in a tropical climate. *International Journal of Sustainable Building Technology and Urban Development, 15*(2), 140-159. https://doi.org/10.22712/susb.20240012

Al-Tamimi, N. A., & Fadzil, S. F. S. (2011). The potential of shading devices for temperature reduction in high-rise residential buildings in the tropics. *Procedia Engineering, 21*, 273-282. https://doi.org/10.1016/j.proeng.2011.11.2015

Attia, S., Beltrán, L., De Herde, A., & Hensen, J. (2009). "Architect-friendly": A comparison of ten different building performance simulation tools. In *Proceedings of the 11th International IBPSA Conference* (pp. 204-211). International Building Performance Simulation Association. https://doi.org/10.26868/25222708.2009.0204-211

Crawley, D. B., Hand, J. W., Kummert, M., & Griffith, B. T. (2008). Contrasting the capabilities of building energy performance simulation programs. *Building and Environment, 43*(4), 661-673. https://doi.org/10.1016/j.buildenv.2006.10.027

Dabe, T. J., & Adane, V. S. (2019). Assessment of daylight into the residential building according to the floor levels for the hot and dry climatic zone. *Journal of Architectural Environment & Structural Engineering Research, 2*(2), 18-29. https://doi.org/10.30564/jaeser.v2i2.376

Dhayal, P., & Jha, B. (2023). Indoor visual comfort: A review of factors and assessments. *ISVS e-Journal, 10*(11), 38-59. https://doi.org/10.61275/ISVSEJ-2023-10-11-03

DiLaura, D. (2021). *The lighting handbook: Reference and application*. Illuminating Engineering Society of North America.

Dinapradipta, A., Sudarma, E., Indrawan, I. A., Krisdianto, J., Arifin, M. F., & Hasya, A. H. (2024). The efficacy of horizontal adjustable bent louver shading for enhancing daylighting performance in walk-up apartment buildings. *IOP Conference Series: Earth and Environmental Science, 1351*(1), Article 012027. https://doi.org/10.1088/1755-1315/1351/1/012027

Du, Y., Li, N., Zhou, L., A, Y., Jiang, Y., & He, Y. (2022). Impact of natural window views on perceptions of indoor environmental quality: An overground experimental study. *Sustainable Cities and Society, 86*, Article 104133. https://doi.org/10.1016/j.scs.2022.104133

Dutta, A., Samanta, A., & Neogi, S. (2017). Influence of orientation and the impact of external window shading on building thermal performance in tropical climate. *Energy and Buildings, 139*, 680-689. https://doi.org/10.1016/j.enbuild.2017.01.018

Elsiana, F., Ekasiwi, S. N. N., & Antaryama, I. G. N. (2023). Validation of IES-VE for assessing daylight performance of building implementing horizontal light pipe and shading systems in the tropics.

International Journal of Application on Sciences, Technology and Engineering, 1(2), 573-582. https://doi.org/10.24912/ijaste.v1.i2.573-582

Syed Fadzil, S. F., Al-Tamimi, N. A., Abdullah, A., & Wan Harun, W. M. (2009). The impact of varied orientation and wall window ratio (WWR) on daylight distribution in residential rooms. In *Proceedings of the CIB W107 International Symposium on Construction in Developing Economies: Commonalities Among Diversities* (pp. 478-485). https://www.irbnet.de/daten/iconda/CIB18158.pdf

Gonzalez-Longo, C., & Mohd Sahabuddin, M. F. (2019). High-rise social housing in hot-humid climates: Towards an "Airhouse" standard for comfort. *Applied Sciences, 9*(23), Article 4985. https://doi.org/10.3390/app9234985

Hassan, A. S., & Al-Ashwal, N. T. (2015). Impact of building envelope modification on energy performance of high-rise apartments in Kuala Lumpur, Malaysia. *International Transaction Journal of Engineering, Management, & Applied Sciences & Technologies, 6*, 91-105. https://doi.org/10.14456/itjemast.2015.11

Heng, C. Y. S., Ossen, D. R., & Lim, Y. W. (2016). Light pipe transporter for high-rise office building in tropical climate. *International Journal of Built Environment and Sustainability, 3*(1), 36-44. https://doi.org/10.11113/ijbes.v3.n1.108

Iqbal, A., Nazir, H., & Awan, M. A. (2023). Design of external shading devices in Mansehra, Pakistan and their role in climate change. *Frontiers in Energy Efficiency, 1*, Article 1244106. https://doi.org/10.3389/fenef.2023.1244106

Kalaimathy, K., Shanthi Priya, R., Rajagopal, P., Pradeepa, C., & Senthil, R. (2023). Daylight performance analysis of a residential building in a tropical climate. *Energy Nexus, 11*, Article 100226. https://doi.org/10.1016/j.nexus.2023.100226

Knoop, M., Stefani, O., Bueno, B., Matusiak, B., Hobday, R., Wirz-Justice, A., Martiny, K., Kantermann, T., Aarts, M. P. J., Zemmouri, N., Appelt, S., & Norton, B. (2020). Daylight: What makes the difference? *Lighting Research & Technology, 52*(3), 423-442. https://doi.org/10.1177/1477153519869758

Li, D. H. W., Lou, S., Ghaffarianhoseini, A., Alshaibani, K. A., & Lam, J. C. (2017). A review of calculating procedures on daylight factor-based metrics under various CIE standard skies and obstructed environments. *Building and Environment, 112*, 29-44. https://doi.org/10.1016/j.buildenv.2016.11.019

Lim, G. H., Hirning, M. B., Keumala, N., & Ghafar, N. A. (2017). Daylight performance and users' visual appraisal for green building offices in Malaysia. *Energy and Buildings, 141*, 175-185. https://doi.org/10.1016/j.enbuild.2017.02.028

Lim, Y. W., Ahmad, M. H., & Ossen, D. R. (2010). Empirical validation of daylight simulation tool with physical model measurement. *American Journal of Applied Sciences, 7*(10), 1426-1431. https://doi.org/10.3844/ajassp.2010.1426.1431

Maamari, F., Fontoynont, M., & Adra, N. (2006). Application of the CIE test cases to assess the accuracy of lighting computer programs. *Energy and Buildings, 38*(7), 869-877. https://doi.org/10.1016/j.enbuild.2006.03.016

Mirrahimi, S., Ibrahim, N. L. N., Nahi, M. H., & Shirazi, A. (2022). Investigation of various window orientations in daylighting performance in the hot-humid climate of Subang, Malaysia. *Civil Engineering and Architecture, 10*(7), 3165-3172. https://doi.org/10.13189/cea.2022.100728

Nazari, M., & Matusiak, B. (2024). Daylighting simulation and visualisation: Navigating challenges in accuracy and validation. *Energy and Buildings, 312*, Article 114188. https://doi.org/10.1016/j.enbuild.2024.114188

Özörhon, B., & Çağlayan, S. (2017). Using building information modelling (BIM)-based tools to support early design decisions. In *Proceedings of the 7th Construction Management Congress* (pp. 619-626).

Rukiah, W. N., Arshard, M., Arminda, W., Anis, T., Raja, Q., & Kadir, A. (2022). Indoor and outdoor air and lighting quality assessment in high-rise low-cost housing in Penang. *Planning Malaysia, 20*(21). https://journals.iium.edu.my/pmr/index.php/pmj/article/view/1107

Sallan, S., Kwong, Q. J., Ling, O. H. L., & Adnan, H. (2025). Post-occupancy evaluation of indoor environmental quality parameters for thermal, visual, and acoustic comfort conditions in a Malaysian green building: A preliminary study. *Jurnal Kejuruteraan, 37*(5), 2441-2451. https://doi.org/10.17576/jkukm-2025-37(5)-32

Syed Husin, S. N. F., & Hanur Hrith, Z. Y. (2018). The performance of daylight through various windows for residential buildings. *Asian Journal of Environment-Behaviour Studies, 3*(6), 169-178. https://doi.org/10.21834/aje-bs.v3i6.247

Tanveer, S., Qureshi, S., & Rasool, A. (2023). The effect of views through windows on apartment dwellers in an urban setting: A case study of Hyderabad apartments. *Jurnal Kejuruteraan, 35*(6), 1289-1297. https://doi.org/10.17576/jkukm-2023-35(6)-03

Tsang, E. K. W., Li, D. H. W., & Li, S. Y. (2022). Predicting daylight illuminance for 15 CIE standard skies using a simple software tool. *Frontiers in Sustainable Cities, 4*, Article 792997. https://doi.org/10.3389/frsc.2022.792997

Wang, X., Zhang, L., Su, X., & Yang, H. (2023). Daylighting and energy performance of the window with transparent insulation slats in the humid subtropical climate zone. *Energy and Buildings, 300*, Article 113685. https://doi.org/10.1016/j.enbuild.2023.113685

Wei, G., Sarman, A. M., Gungat, L., Li, M., Shen, L., & Fang, G. S. (2023). Application of building information modelling (BIM) six dimension (6D) in green buildings: A case study of sunshine and daylight analysis. *Jurnal Kejuruteraan, 35*(2), 511-521. https://doi.org/10.17576/jkukm-2023-35(2)-23

Wirz-Justice, A., Skene, D. J., & Münch, M. (2021). The relevance of daylight for humans. *Biochemical Pharmacology, 191*, Article 114304. https://doi.org/10.1016/j.bcp.2020.114304

Wong, N. H., Istiadji, A. D., & Ng, E. (2004). Effect of external shading devices on daylighting penetration in residential buildings. *Lighting Research & Technology, 36*(4), 317-333. https://doi.org/10.1191/1365782804li126oa

Pertanika J. Sci. & Technol. 34 (3): 1655 - 1683 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Gamification for Extended Reality Informative Space (XRIS) Visualisation in the Construction Industry

Changsaar Chai, Anne Ann Yii Chua*, and Felix Kai Han Yong

School of Architecture, Building and Design, Faculty of Innovation and Technology, Taylor's University, 47500 Subang Jaya, Selangor, Malaysia

ABSTRACT

Intuitive visualisation tools are essential for the effective coordination of construction projects to convey complex spatial and technical information to diverse stakeholders. Traditional design documentation using 2D drawings and static 3D models cannot accurately depict complex spatial relationships and construction sequences. This results in design misinterpretations, inefficient constructability reviews and rework costs. Most of the current extended reality (XR) deployments in construction are essentially used as visualisation tools. Gamification usage in the industry is mainly oriented to the educational workflow context. This study develops an Extended Reality Information Space (XRIS), a HoloLens-based system that spatially anchors construction data onto physical environments. The system operates through gaze-triggered information tags and embedded game-based mechanics. Three functional layers covering immersive visualisation, information management, and interaction and gamification were comprised in this system. Semi-structured interviews were conducted with eight construction professionals and two XR specialists to identify the benefits of XRIS implementation and the impact of gamification on XRIS. Data were then analysed using thematic analysis. The findings suggest that XRIS enhanced spatial understanding, user accessibility, team communication, and motivation. Gamification integrated in the visualisation layer maintained the users' engagement through the behaviour of spatial exploration and task completion. Both mechanisms supported deeper retention and more precise articulation of design logic than passive documentation review. This study contributes a professional-oriented, empirically evaluated XRIS configuration advancing immersive tool design for construction practice.

Keywords: Construction industry, extended reality, Extended Reality Informative Space (XRIS), gamification, visualisation

ARTICLE INFO

Article history:
Received: 01 December 2025
Accepted: 31 May 2026
Published: 19 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.16

E-mail addresses:
cs.chai@taylors.edu.my (ChangSaar Chai)
annechuaannyii@sd.taylors.edu.my (Anne Ann Yii Chua)
0369865@sd.taylors.edu.my (Felix Kai Han Yong)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

INTRODUCTION

Design coordination, stakeholder communication, and spatial information delivery are still significant challenges in the construction sector. Costly rework and schedule delays are a result of design misinterpretation, information asymmetry, and inefficient constructability reviews (Le et al., 2024). A survey found that 48% of all reworks arise from poor project data and miscommunication, leading to an estimated USD 31.3 billion in avoidable rework costs (FMI & PlanGrid, 2018). Traditional two-dimensional (2D) drawings and static three-dimensional (3D) models cannot represent complex spatial relationships and construction sequencing (Abouelkhier et al., 2025). These emphasise the need for more intuitive, interactive approaches that support collaborative decision-making.

Extended Reality (XR), encompassing Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR), has emerged as a key digital development in architecture, engineering, and construction (AEC). Integrations between XR and Building Information Modelling (BIM) provide spatial awareness, clash detection and constructability analysis (Safikhani et al., 2022). On-site progress monitoring and information overlay are also enabled through AR applications. However, most XR deployments remain as standalone visualisation tools, and limited evidence can be found of professional benefits or continued use in the daily work of projects (Trindade et al., 2023).

Gamification has shown potential to address engagement and motivation in professional settings. In construction, it has demonstrated benefits for hazard recognition, safety behaviour, and sustainable design involvement (Oke et al., 2024). Its broader application, though, is constrained by technical complexity, usability barriers, and poor integration with live project data (Davila Delgado et al., 2020). While the individual merits of XR and gamification in construction are documented, empirical evaluations of integrated, gamified XR systems targeting professional end-users remain scarce. Limited study has examined the perceived benefits and adoption performance of a gamified XR platform that spatially anchors construction data within an immersive environment for professional use.

To address these gaps, this study presents the Extended Reality Informative Space (XRIS), a gamified XR platform that integrates spatially anchored construction data with interactive game-based mechanisms in an immersive environment. This study includes two objectives: (1) to identify the benefits of XRIS implementation in the construction industry, and (2) to evaluate the impact of gamification on XRIS use. It delivers empirical evidence on the integration of gamification and XR for professional construction workflows, an area currently underrepresented in the literature. These findings guide the design of gamified XR tools supporting spatial understanding, stakeholder engagement, and collaborative decision-making.

LITERATURE REVIEW

There has been a growing amount of research interest on the use of gamification in construction practice. A keyword co-occurrence analysis of the current research landscape was performed using VOSviewer software based on a Scopus database search of documents with the keywords "construction" and "gamification", resulting in 375 documents. The VOSviewer keyword network in Figure 1 shows that "gamification" is strongly linked to educational terms such as "students", "teaching", and "learning systems". A separate cluster, however, consisting of professional construction-related terms like "construction projects", "construction management", and "project management", can be perceived but is relatively less dense.

These clusters are structurally separated, suggesting that gamification studies in the construction domain have focused on educational gamification. Its use as an operational mechanism within professional construction workflows, particularly in combination with immersive XR environments, remains largely unexamined. This gap is not merely topical but methodological. Existing studies have not empirically assessed the perceptions of construction professionals toward gamified, immersive visualisation tools designed for project-level tasks such as design review, spatial coordination, and information retrieval.

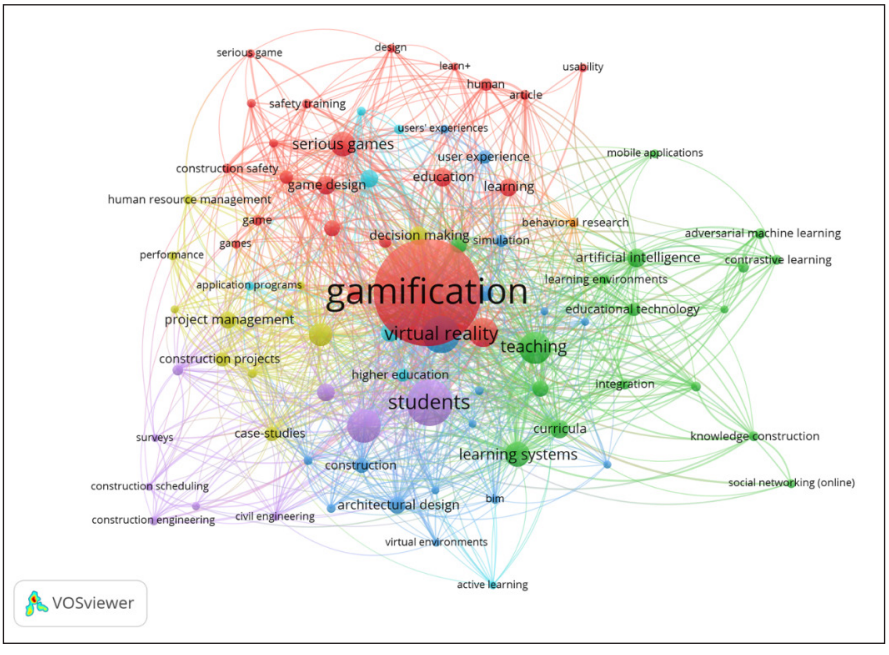


Figure 1. Keyword co-occurrence analysis

Gamification in Digital Technologies

Gamification applies game design elements such as points, challenges and leaderboards to non-game contexts (Koivisto et al., 2014). Reward systems, feedback, and progression mechanisms are organised to influence user behaviour. Reward systems are points or badges for completing tasks. Feedback loops provide immediate feedback and help users take certain actions. Progression mechanics allow for a progression of levels, milestones, or unlockable content to give the player a sense of progression. These elements transform routine activities into structured, goal-oriented experiences. These principles have been applied in education, corporate training and health and wellness. The evolution of gamification further reflects a shift from entertainment-oriented video games to productivity-focused tools. This is gradually integrating with new technologies such as AR and VR (Lee et al., 2022).

In construction, gamified applications have emerged to support stakeholder engagement and project monitoring (Yu et al., 2024). Guimarães et al. (2024) developed a game-like BIM application for on-demand monitoring of construction projects. Further research has confirm the integration of gamification techniques with BIM-based VR has enabled public engagement in architectural design (Yalçın, 2024). A BIM-VR system with gamified simulations was also developed for health risk assessment in urban infrastructure (Chen et al., 2024). Collectively, these developments illustrate the growing integration of gamification digital construction tools across diverse project functions, highlighting the need for more immersive and interactive approaches that sustain user engagement with construction information.

Extended Reality (XR) in Construction

Computer graphics simulation in the manufacturing and defence sectors has led to the development of Extended Reality (XR). It has since evolved into a transformative technology within the AEC industry. XR refers to a set of technologies that have the capacity to integrate digital content with physical environments to varying degrees. It is applied in the construction process from the design to the operation and maintenance (Li et al., 2024). The early applications were on architectural walkthrough systems (Brooks, 1987). These showed the potential of immersive visualisation to reduce design misinterpretation. XR was then extended to 4D construction simulation and scheduling visualisation (Dawood et al., 2003). As computing power matured, XR advanced into safety training and hazard simulation. Immersive environments enabled workers to practise high-risk tasks without physical exposure (Sacks et al., 2013). The release of head-mounted devices later broadened access across project teams. This development also accelerated research output within the AEC domain. More recently, XR has supported on-site assembly guidance, inspection, progress monitoring, and remote collaboration (Gornall et al., 2025).

Visualisation remains one of XR's strongest assets in construction. It enhances spatial understanding of building structures and construction sequencing across project stages (Safikhani et al., 2022).

Benefits of Gamification

Gamification boosts user engagement by fulfilling psychological needs for competence, autonomy, and relatedness (Bitrián et al., 2021). Repetitive tasks are less boring with interactive activities and rewards. Well-designed game mechanics tap into intrinsic motivation, supporting goal attainment and challenge-oriented behaviours (Manzano-León et al., 2021). Beyond engagement, gamified challenges, levels, and real-time feedback promote active learning, deepen understanding, and enhance long-term recall (Liu et al., 2023; Ortiz-Rojas et al., 2025). Badges, milestones, and progress indicators make effort visible and motivate users to complete tasks more efficiently and consistently (Encarnação et al., 2021; Bizzi, 2023; Oke et al., 2023). At the group level, team-based quests and shared rewards foster collaboration and collective problem-solving, stimulating communication and trust among respondents (Patricio et al., 2022). Table 1 summarises the benefits of gamification identified in recent studies that are relevant to this research.

Table 1
Benefits of gamification

Benefits	References
Enhancing User Engagement	(Bitrián et al.,2021; Chang et al., 2022; Manzano-León et al., 2021)
Improved Learning and Retention	(Jato-Espino et al., 2024; Liu et al., 2023; Ortiz-Rojas et al., 2025)
Motivation and Productivity	(Bizzi, 2023; Encarnação et al., 2021; Oke et al., 2023)
Collaboration and Teamwork	(Bizzi, 2023; Encarnação et al., 2021; Patricio et al., 2022)

Challenges of Gamification

The interrelated challenges of gamification adoption in construction impede its progress from controlled environments to real projects. Usability represents the most immediate barrier, as gamified systems misaligned with complex, time-sensitive construction workflows risk disrupting site activities and undermining operational efficiency (Guimarães et al., 2024; Oke et al., 2023). User fatigue occurs even in well-designed systems, where prolonged exposure erodes the novelty effect that initially drives engagement (Almeida et al., 2023; Priyantini et al., 2023; Yang et al., 2024). Attributing performance gains solely to gamification remains difficult, as its effects are rarely separable from other factors (Klein, 2021; Monteiro et al., 2021; Oke et al., 2024). Competitive mechanics such as leaderboards may reinforce hierarchical tensions and discourage equitable participation across diverse

cultural and demographic groups (Gomez et al., 2024; Henry et al., 2024; Othman et al., 2023). These challenges, summarised in Table 2, suggest that successful gamification requires careful contextualisation.

Table 2
Challenges of gamification

Challenges	References
Usability issues	(Guimarães et al., 2024; Oke et al., 2023; Priyantini et al., 2023)
User fatigue	(Almeida et al., 2023; Guimarães et al., 2024; Yang et al., 2024)
Outcome attribution difficulty Cultural and demographic constraints	(Gomez et al., 2024; Henry et al., 2024; Klein, 2021; Monteiro et al., 2021; Oke et al., 2024; Othman et al., 2023)

Impacts of Gamification

There is evidence that gamification has an impact, but this relates to conditions where gains are shown to exist or to vanish. Reported outcomes hinge on task characteristics, organisational context, and design quality. Similar interventions yield different results across settings (Hammedi et al., 2024; Jacobides et al., 2024). A major constant constraint is the lack of motivation. Feedback, rewards and structured tasks influence work practices in the short term (Patricio et al., 2022), but behavioural change rarely persists beyond the mechanics that engender it (Manzano-León et al., 2021; Mohanty & Christopher, 2024). Collaboration follows a similar pattern. Cooperation can be facilitated by shared goals, although primarily when working relationships already exist (Bizzi, 2023; Cheng & Chau, 2022; Leite et al., 2022). While efficiency improvements have been documented (Leite et al., 2023; Rahiman et al., 2023), it is challenging to clearly separate the impact of gamification from organisational and task-related factors (Jacobides et al., 2024). Table 3 summarises these impacts.

Table 3
Impacts of gamification

Impact	References
Context-dependent engagement	(Hammedi et al., 2024; Jacobides et al., 2024; Sailer & Homner, 2020)
Extrinsic motivation dependency	(Manzano-León et al., 2021; Mohanty & Christopher, 2024; Patricio et al., 2022)
Conditional collaboration Attribution difficulty in performance gains	(Bizzi, 2023; Cheng & Chau, 2022; Jacobides et al., 2024; Leite et al., 2022; Leite et al., 2023; Rahiman et al., 2023)

Benefits of XR

XR introduces capabilities that address longstanding limitations in construction project workflows. Enhanced visualisation is the most frequently cited benefit. XR enables stakeholders to interact with 3D representations before construction begins, supporting clash detection, design review, and client communication more effectively than 2D drawings (Arslan et al. 2024; Alhady et al. 2024; Pham et al. 2025). Real-time feedback further strengthens spatial immersion, as XR platforms allow users to manipulate models and receive immediate system responses, reducing reliance on sequential review processes (Bhatarai et al., 2025; Hajji et al., 2022; Yu et al., 2024). XR also supports training and simulation in controlled settings, allowing workers to rehearse high-risk tasks without physical exposure (Alzarrad et al., 2024; Liu et al., 2022; Speiser & Teizer, 2024). Additionally, immersive environments support decision-making during design and constructability assessments by enabling project teams to evaluate alternatives spatially rather than abstractly (Chen et al., 2024; Delgado et al., 2020; Wang & Tung, 2023). Table 4 summarises these capabilities.

Table 4
Benefits of XR

Impact	References
Enhanced Visualisation	(Arslan et al.,2024; Alhady et al., 2024; Pham et al.,2025)
Real-Time feedback and interaction	(Bhatarai et al., 2025; Hajji et al., 2022; Yu et al., 2024)
Training and simulation	(Alzarrad et al., 2024; Liu et al., 2022; Speiser & Teizer, 2024)
Decision-Making	(Chen et al., 2024; Delgado et al., 2020; Wang & Tung, 2023)

Challenges of XR

XR adoption in construction remains constrained by interrelated resource, technical, and organisational barriers. High hardware prices and content development expenses restrict access, particularly for small and medium-sized enterprises with limited budgets (Ali et al., 2024; Davila Delgado et al., 2020; Pavlou et al., 2021). These pressures are compounded by technological limitations, as processing constraints and limited device performance reduce the reliability of real-time XR environments (Gornall et al., 2025; Heo et al., 2023; Struye et al., 2024). Even where financial and technical conditions are met, organisational factors introduce further friction. Cultural resistance and reluctance to adopt unfamiliar workflows slow uptake (Jalo & Pirkkalainen, 2024; Schwarz et al., 2024). Poor compatibility between XR platforms and legacy software infrastructures further weakens integration into existing project processes (Fernández-Caramés & Fraga-Lamas, 2024; Vona et al., 2025). These barriers diminish XR's reliability as a decision-support tool during critical construction phases. Most evaluations focus on technical feasibility, leaving user perceptions of usability

and engagement within professional workflows underexamined. These challenges are outlined in Table 5.

Table 5
Challenges of XR

Challenges	References
High implementation costs	(Ali et al., 2024; Davila Delgado et al., 2020; Pavlou et al., 2021)
Technological limitations Resistance to change	(Alam et al., 2025; Heo et al., 2023; Irani et al., 2023; Jalo & Pirkkalainen, 2024; Schwarz et al., 2024; Struye et al., 2024)
Integration with legacy systems	(Fernández-Caramés & Fraga-Lamas, 2024; Irani et al., 2023; Vona et al., 2025)

Impacts of XR

There is emerging evidence on the impact of XR in professional settings. Improved learning outcomes are the most consistently reported impact. Immersive task rehearsal produces stronger knowledge retention and faster skill acquisition than conventional instruction (Margheritti et al., 2025; Pavlou et al., 2021). These gains feed into a second outcome of reduced errors and rework. Early clash detection and verification on design phase helps project teams resolve discrepancies before construction (Alhady et al., 2024; Alshboul and Shehadeh, 2025). Enhanced stakeholder decision-making has also been noted with the aid of immersive visualisation to help stakeholders evaluate design alternatives and construction sequences more effectively (Mi & Li, 2024; Oral et al., 2023). On an industry level, the adoption of XR can drive innovation across the industry, fostering the use of digital workflows in project delivery and organisational practice (Nyqvist et al., 2025; Radanovic et al., 2023). Table 6 summarises these impacts. Overall, the available evidence is positive but limited to context-specific and sometimes small-scale trials. The outcomes will be based on the readiness of the organisation and its digital skills, as well as its integration into the already existing platforms. Issues of pilot experience transferability to real-world site conditions, usability and motion sickness indicate that impacts of XR are mediated but not guaranteed.

Table 6
Impacts of XR

Impact	References
Improved learning outcomes	(Margheritti et al., 2025; Pavlou et al., 2021)
Reduced errors and rework	(Alhady et al., 2024; Alshboul and Shehadeh, 2025)
Enhanced stakeholder decision-making	(Mi & Li, 2024; Oral et al., 2023)
Industry Innovation	(Nyqvist et al., 2025; Radanovic et al., 2023)

Extended Reality Information Space (XRIS)

XRIS is an extension to the traditional XR visualisation that incorporates structured information management in immersive environments. XR-BIM platforms are mainly used for the rendering of 3D models for visual inspection. XRIS, on the other hand, adds a spatial information layer that binds data to the actual or virtual object when the object is gazed. This distinction matters. XR-BIM is concerned with the visualisation of the model, while XRIS is about the immersive environment as an active space of information. Users access, manipulate and react to project information in real time.

XRIS comprises three integrated layers as shown in Figure 2. The visualisation layer is an immersive visualisation that leverages XR technologies to represent construction environments. The information management layer bridges the gap between BIM, GIS, and IoT platforms to specific spatial locations (Morris et al., 2021). The interaction and gamification layer controls access to the user via spatial control, including gaze activation, proximity detection and gesture selection. It also includes elements such as points, challenges, and rewards to keep users engaged in the game.

The system boundary of XRIS is defined by spatially anchored data within a single immersive session. Users interact with a bounded set of construction components, each linked to embedded information tags. The system does not extend to live project management functions such as scheduling, procurement, or cost tracking. It focuses instead on information retrieval, spatial comprehension, and design communication within an immersive setting. Though not yet widely adopted, XRIS holds potential to improve spatial cognition, real-time decision-making, and digital collaboration among project stakeholders.

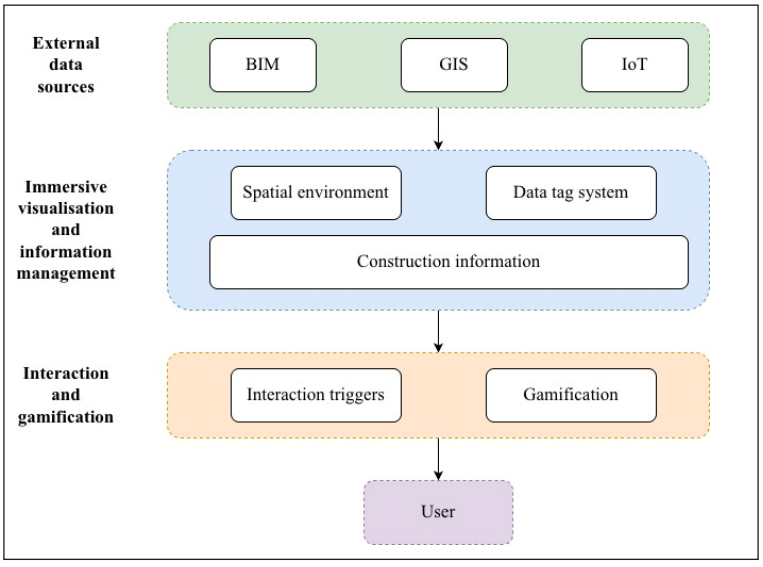


Figure 2. Functional architecture diagram of XRIS

Table 7 shows a comparison of the findings of the current study in relation to previous works. Platforms without gamification demonstrated visualisation and coordination benefits but did not address sustained engagement during information retrieval. Prior gamified platforms recorded incidental engagement from task completion and measured retention narrowly as hazard recall or procedural knowledge transfer. Neither construct was linked to gamification as a deliberate design mechanism. Unlike prior platforms where gamification functioned as an external layer, XRIS embedded it within the visualisation layer as part of the information retrieval mechanism. Gaze-triggered mechanics and spatial discovery structured sustained engagement, while active interaction supported respondent-reported conceptual retention of construction information. This gap in the existing literature informed the empirical evaluation presented in this study.

Table 7
Comparison of the present study with previous studies

Author	Technology	Engagement	Retention	Gamification	User Context
Current study	XR	Sustained engagement confirmed	Conceptual retention reported	Embedded gaze-triggered mechanics	XR specialist and Construction professionals
(Speiser & Teizer, 2024)	VR + BIM	Not measured	Hazard recall supported	None	Construction professionals
(Alhady et al., 2024)	XR + BIM	Not measured	Not measured	None	Infrastructure professionals
(Chen et al., 2024)	BIM-VR + gamification	Scenario engagement reported	Not measured	Scenario-based	Mixed (students/ professionals)
(Yalçın,2024)	BIM-VR	Participation reported	Not measured	Game mechanics	Public stakeholders
(Margheritti et al., 2025)	XR + serious games	Enhanced engagement reported	Knowledge transfer supported	Serious games	Workplace safety trainees

METHODS

A qualitative research design was adopted to identify the benefits of XRIS implementation in the construction industry and evaluate the impact of gamification on XRIS use. A qualitative method was appropriate to generate in-depth narrative data, as the research centred on subjective experiences and perceptual responses that cannot be adequately captured through numerical measurement (Creswell & Creswell, 2017).

A purposive sampling strategy recruited respondents with direct experience in XR technologies or construction project workflows. The sampling frame required at least five years of industry experience and active professional involvement. Determining an appropriate sample size was guided by the concept of data saturation, which occurs when no new information or themes emerge from additional data collection. Guest et al. (2006) demonstrated that data saturation becomes evident after six in-depth interviews. A larger sample size can hinder the depth of analysis in qualitative inquiry (Boddy, 2016).

Respondents engaged in a controlled session with a gamified XRIS prototype on a HoloLens at the Vortex XR Lab, Taylor's University, as shown in Figure 3. Each respondent explored the prototype independently within a standardised timeframe. This setup provided a consistent experiential baseline for the interviews. Each respondent then took part in a semi-structured interview. The protocol covered four thematic areas: spatial understanding, system accessibility, gamification experience, and collaborative potential.

Figure 3. XRIS prototype session at the Vortex XR Lab

Research Framework

Figure 4 presents the integrated research framework designed to guide the methodological process of this study. The framework starts with XR sessions, whereby respondents engaged with an XRIS prototype with a HoloLens. Two forms of data were collected from each session: interaction logs documenting user navigation patterns and spatial engagement, and feedback transcripts capturing verbal responses during and after the experience. These dimensions provided the basis for connecting XRIS experiences and gamification design features to user responses across the thematic categories identified in the study.

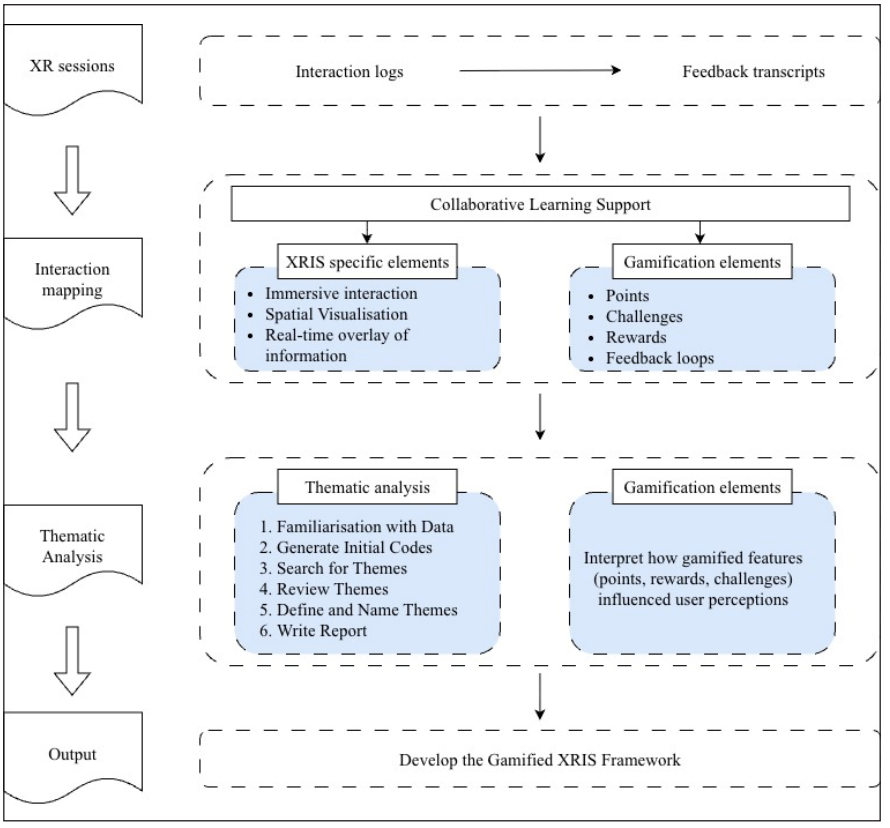


Figure 4. Research framework

Thematic Analysis

A thematic analysis was conducted following Braun and Clarke's (2006) six-phase framework. After familiarisation with the transcripts, descriptive and interpretive codes were generated and iteratively reviewed to refine emerging patterns into coherent themes. The findings were consolidated into two primary categories: (A) Benefits of XRIS in construction and (B) Impact of Gamification on XRIS Use.

The coding structure followed a two-tier sub-theme system within each thematic group. Two sub-themes were constructed for each main theme. For example, Improved Visualisation comprised easier spatial relationship comprehension (A1a-i) and a more immersive experience than 2D drawings (A1a-ii), while User Engagement was split into encouragement of extended use (B1a-i) and stimulation of curiosity and task completion (B1a-ii). This allowed for more in-depth analysis and for different perceptions to be captured. Table 8 presents the complete coding structure along with the themes, sub-themes and distribution of respondents.

Table 8
Summary of responses on findings

Finding / Coding Group	Theme	Sub-theme	Code	Respondent									
				R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
Finding 1 – Benefits of XRIS in Construction													
Improved Visualisation	Spatial understanding	Easier spatial relationship comprehension	A1a-i		✓	✓	✓	✓	✓	✓	✓	✓	✓
		More immersive than 2D drawings	A1a-ii	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
User Accessibility	Information access	Information is easy to locate	A2a-i		✓	✓	✓	✓	✓	✓	✓	✓	✓
		Intuitive interaction with the system	A2a-ii	✓	✓	✓		✓	✓	✓			✓
Team Communication	Collaboration	Easier to explain design decisions	A3a-i		✓	✓		✓	✓		✓	✓	✓
		Supports through shared view	A3a-ii	✓	✓	✓	✓			✓		✓	✓
User Motivation	Engagement	More interesting than traditional tools	A4a-i		✓	✓	✓	✓	✓	✓	✓	✓	✓
		Encourages learning through experience	A4a-ii	✓	✓	✓			✓		✓		
Finding 2 – Impact of Gamification on XRIS Use													
User Engagement	Motivation	Encourages extended use	**B1a-i**		✓	✓			✓	✓	✓	✓	
		Stimulates curiosity and task completion	**B1a-ii**	✓	✓	✓		✓	✓			✓	
Game Features	Challenges and Rewards	Makes learning fun	**B2a-i**	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		Increases system exploration	**B2a-ii**	✓	✓	✓			✓		✓		
Learning Experience	Knowledge Retention	Enhances understanding via interaction	**B3a-i**	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		Supports better memory of construction concepts	**B3a-ii**	✓	✓	✓			✓		✓	✓	✓

RESULTS AND DISCUSSION

Respondents' Profile

The various professional positions from the respondents covered the entire project lifecycle. From design advisory and feasibility assessment (consultant) to coordination and construction sequencing on site (contractor), and evaluation from XR specialists and academics. Ten respondents came from four different occupational groups, providing cross-disciplinary coverage in the evaluation. The XR specialists had experience in XR facility management, hardware deployment, and immersive content development, with firsthand experience in applications using HoloLens. This background enabled them to assess the technical fidelity and interaction design of the prototype from a development standpoint. Table 9 provides succinct details about the interviewees' backgrounds. By the eighth interview, no new themes emerged, confirming that the sample was sufficient for thematic saturation (Lim, 2025). The final two interviews confirmed existing patterns without adding new codes.

Table 9
Demographic profile of respondents

Respondent ID	Discipline	Years of Industry Experience	XR Experience Level
R1	XR Specialist	> 10	Advanced
R2	XR Integration Designer	> 5	Advanced
R3	Consultant	> 5	Basic
R4	Consultant	> 10	Basic
R5	Consultant	> 15	Basic
R6	Contractor	> 20	Basic
R7	Contractor	> 10	Basic
R8	Contractor	> 5	Basic
R9	Academia	> 10	Basic
R10	Academia	> 15	Basic

Benefits of XRIS in Construction

Improved Visualisation

The empirical findings demonstrated that XRIS significantly enhanced visualisation as it allowed users to view the spatial relationship and get access to information more intuitively compared to the traditional two-dimensional way of representation. The respondents stressed that the direct overlay of the data on virtual objects fixed the gap between the abstract data and the physical layout. The system eliminated the need of cross-referencing external manuals or fixed slides as it spatially anchored the information related to the component.

The page made it easier to comprehend the spatial relationship between components. -R6

The spatial placement allowed me to link each element with its function without confusion. -R8

The way the info popped up next to each component made it easier for me to understand what I was looking at. It's clearer than trying to read a manual. -R3

The info tag comes out beside the item, so I can understand straight away what it is for. -R4

It helps to give more of your space differently, give a new meaning to what spatial spaces mean. -R1

These results support the findings obtained by Arslan et al. (2024), who found that immersive 3D environments can promote spatial knowledge and reduce misunderstanding. The ability of the users to find their way through interrelations and absorb data conforms to the propositions of Azmi et al. (2017), who argued that XRIS facilitates a deeper understanding of intricate designs and spatial relationships in the design planning processes. In addition, this finding aligns with more generalised evidence that suggests interactive visualisation helps to promote a better understanding of choices compared to non-interactive media (Mi & Li, 2024; Oral et al., 2023). Hence, this allows the users to manipulate complicated data without the frustration that is often caused by conventional documentation.

User Accessibility

The empirical results indicated that the XRIS environment was concomitant with high user access, as indicated by an extraordinarily short learning curve. These are the results compared to the current fears of digital complexity in the context of construction. Users claimed an intuitively easy experience that requires minimal training, which highlights the instantaneous usability of the system.

Honestly, it didn’t take more than a minute to figure out. Everything felt intuitive. -R3

Didn’t need any guide. I just wear and explore already. -R7

Very easy to use. Just try one time, already can use. -R5

These outcomes contradict the usability issues described by Priyantini et al. (2023), suggesting that the problematic usability can be alleviated by an effective interface design. The ability of the respondents to communicate with each other using simple gestures was consistent with the long-term interaction goal expressed by Bitriany et al. (2021). Moreover, the users noted that the system is responsive, especially in the efficiency of gaze-based triggers, which were used to achieve smooth and hands-free access to information.

> Information access was seamless. It appeared automatically as I focused on a component. -R6

> Basically, you just look at wherever you want, and then basically the information is there. -R1

> It auto comes out when I look at it. Very fast, very simple. -R10

The automaticity can be linked to the benefits of real-time feedback and interaction described by Hajji et al. (2022) and Yu et al. (2024). The system also made it easier to navigate and interactively manipulate datasets in immersive settings, which supports the propositions of Morris et al. (2021).

Team Communication

According to the respondents, the XRIS system greatly helped in the presentation of design logic and construction concepts to the team members. The respondents noted that the direct interaction with the spatial model allowed them to come up with eloquent explanations. This further helped them to properly coordinate their individual understanding to the group observation.

> After I tried XRIS, I had a clearer idea how to explain my design logic in the group. – R3

> When I talked to my friends, I just walked them through what I saw. That made my explanation clearer. – R9

> It can support my ability to explain construction ideas if I communicate with teams in the future. – R6

These results support the existing body of literature about team dynamics in which shared visual tools are linked to better shared objectives and more transparency (Cheng and

Chau, 2022; Bizzi, 2023). The internalisation of spatial data allows stakeholders to exchange knowledge more efficiently, which justifies the findings by Encarnação et al. (2021) on improved communication and trust in a team-based environment. Additionally, the respondents identified how the use of spatial referencing can be useful to minimise ambiguity in discussions. The ability to recognise and highlight visual items allowed users to produce more accurate messages.

> You want to point out to a specific location or specific object... it makes it easier to send the message to your team member. -R1

> When our group discuss the assignment, we can just share what we see in the VR. -R10

This is in line with the advantages reported by Hajji et al. (2022) and Yu et al. (2024), who established that XR will help in remote coordination and immediate reactions to design alterations. The fact that it is possible to anchor the discussion on specific visual objects confirms the possibility of XRIS to facilitate group problem-solving (Patricio et al., 2022).

User Motivation

Active interaction in the XRIS system significantly contributes to the element of user motivation as it transforms passive observation into an active exploration. The respondents also noted that they always found the experience interesting, and they rated the experience highly when compared to traditional industry practices, such as didactic presentations or when using manuals. The immersive features embedded in the system made the users engage in the project data understanding voluntarily and give extra time and effort.

> Besides just looking at a website, [users] can walk around... it makes the [process] much more fun. -R2

> Better engaging way... unsuspectingly spending more time, engaging more time with the application. -R1

> I think it is more engaging instead of looking at the [static] slide. -R6

These findings are in line with the research on gamification in the workplace, which argues that interactive exercises are attention-grabbing and lead to long-term engagement (Bitrián et al., 2021; Chang et al., 2022). The unintentional time expenditure of the

respondents is consistent with the results of Encarnação et al. (2021) and Bizzi (2023), who found that motivational cues strengthen effort and prompt users to perform tasks more effectively. Respondents indicated that physical movement and exploration resulted in better memory retention compared to passive briefing.

> Since I was exploring actively, I think I retained more than usual. -R3

> I retained the concepts better because I was actively involved. -R6

> It makes me feel like I want to learn more after the testing to XRIS. -R8

These results confirm the hypotheses of Ortiz-Rojas et al. (2025) who argue that gamified tasks provide active learning experiences, enhance a better understanding, and enhance long-term memory of information. The shift in the consumption to the active enquiry and the inquiry shows how XRIS can support the adoption and continuation of preferred work practices, as Kim and Castelli (2021) propose.

Impact of Gamification on XRIS Use

User Engagement

The qualitative data revealed that the gamified mechanisms within the XRIS system significantly prolonged user engagement, fostering a state of sustained interaction, whereby stakeholders voluntarily invested more time in exploration than anticipated. Respondents described a compelling attraction to the system, whereby the drive to complete tasks overrode the perception of time passing. Curiosity-driven engagement was observed in respondents who voluntarily extended their session time due to an intrinsic interest in the immersive experience. These users described an unplanned desire to continue exploring without external instruction or prompting.

> That would encourage you to stay more without realising it because you are trying to achieve a particular gamified task within the application. -R1

> I spent more time than expected. I didn’t want to stop halfway. -R3

> I engaged with the content longer because the feature grabs my [attention]. -R6

The results were congruent with the literature on gamification mechanics, which claimed that these activities were successful in attracting attention, maintaining engagement and avoiding disengagement by satisfying psychological needs for competence and autonomy (Bitrián et al., 2021; Chang et al., 2022). Gamification also produced an excitement that enhanced the level of attentional investment. The treasure-hunt framing of XRIS drew respondents in before task-driven engagement became a conscious goal. This distinction matters analytically, as sustained interaction describes persistence while excitement-driven attention capture prevents disengagement. Self-Determination Theory further illuminates this, as the spatial challenge structure satisfied the need for competence by triggering intrinsic curiosity rather than extrinsic reward-seeking (Ryan & Deci, 2000). This effect was magnified in the immersive environment, which focused attention on the gamified content (Smith & Mulligan, 2021).

> The layout encouraged full exploration of all elements. -R6

> I want to stay longer just to make sure I got through all the parts. -R9

> Besides just looking at a website, [users] can walk around, doing treasure hunting, it makes the [process] much more fun. -R2

This behaviour validated Encarnação et al. (2021) and Bizzi (2023), who noted that progress indicators reinforce effort and task completion. The immersive aspect of XRIS transformed routine information retrieval into an active task, reducing boredom caused by repetitive technical tasks (Almeida et al., 2023). High engagement translates into readiness to adopt XRIS professionally, supporting well-designed gamified systems, maintaining desired work practices (Kim & Castelli, 2021).

> For me, I think I will continue to use it in the future if it's developed in the construction industry. -R8

> I would say it encouraged more, for example, how can I apply VR and XR in construction architecture? -R2

Game Features

XRIS effectively developed a mission-oriented attitude, transforming traditional information retrieval into a systematic challenge. The respondents repeatedly compared the experience

to a treasure hunt, a mini mission. They also noted that the implicit gamification encouraged them to search proactively and find certain data points in space. This task structure provided an intrinsic compass, ensuring that all the contents were covered even where no specific guidelines were given.

> It gamified the learning without needing traditional scores or points. -R8

> The learning is like a game, because we need to walk around the room and search for the info... the system guided me naturally. -R6

> Similar to game elements, like having to find where the spot is located, making it challenging and fun. -R10

> It felt like I was completing a task, kind of gamified. -R3

These results correspond to the existing literature on gamified challenges, which are the ones that have been found to facilitate active learning and further comprehension (Liu et al., 2023; Ortiz-Rojas et al., 2025). The fact that the respondents have specifically mentioned the aspect of natural guidance. On the other hand, facilitated the discussion that effective gamified interfaces can enhance usability and lessen the cognitive burden of intricate navigation. Completionist behaviour was one of the results of this mechanic that was notable. There was a psychological desire to check every node reported by its users to make sure that they did not miss any important information, which is not so inherent in the traditional document reviews.

> I explored every tag just to make sure I didn't miss anything. -R3

> I moved through the entire layout because it felt like a complete learning sequence. -R8

> User can walk around, doing treasure hunting, it makes the [process] much more fun. - R2

This confirmed the claim of Encarnação et al. (2021) and Bizzi (2023) that gamified features, including progress indicators or the design of a quest, strengthen the effort and motivate users to perform tasks more regularly. The system should make the users eager to explore the content more carefully by positioning the construction data inspection as a discovery process, thus increasing the likelihood of retaining the information (Ortiz-Rojas et al., 2025).

Learning Experience

The findings indicated that XRIS was interactive, which led to a better understanding of elements of construction in comparison to the passive modes of learning. The respondents highlighted that they could perceive spatial elements and labels directly, which contributed to the increased conceptual clarity. Through the interaction with the information in a spatial environment, they indicated having greatly improved their capacity to understand the logic behind the construction assemblies.

> It helps me understand more about the elements of the construction. -R2

> Interactive learning enhanced my conceptual clarity. -R6

> By interacting with each label, I think it makes my understanding better. -R8

These observations were also supported by the knowledge about XR-based training, which implied that spatial knowledge is better learned in immersive environments and is less prone to misinterpretation (Arslan et al., 2024). The claim by the respondents that the system enhances conceptual understanding was in line with Azmi et al. (2017), who argued that XRIS visualisation enables users to have a better understanding of complex designs and spatial relationships.

Out of direct understanding, the data indicated that the practicality of the experience was a great enhancer of the memory. Respondents observed that the process of physically navigating and interacting with the data produced a greater cognitive impression as compared to traditional reading, and, therefore, they were able to remember the information much better.

> Interacting with the system helped me understand the content more deeply... I can still recall what I saw because I was hands-on. -R3

> The experience is unique, and I can memorise things better through the visualisation. -R9

> The key takeaways remain with me because of the immersive experience. -R6

This supports the research of Ortiz-Rojas et al. (2025), who assumed that gamified interactive challenges facilitated active learning and increased long-term information retention. Moreover, it proves the conclusion made by Pavlou et al. (2021) that knowledge retention and practical preparedness can be improved through XR-based skills training.

Lastly, the findings indicated improved efficiency in learning. Respondents also said that the interactive model made them quicker at processing information, partly because the device was hands-free and therefore, they could do multiple activities at the same time and access information instantly. This was in line with the literature that reported enhanced task performance and skill learning as the main advantage of XR integration (Liu et al., 2022; Speiser and Teizer, 2024).

> When I interact with myself, I understand faster. -R4

> Can understand the elements faster besides just searching the info on the internet. -R7

> You free your hands when you are using it to get data... basically you can do other stuff like hold other things. -R1

CONCLUSION

This study presents a gamified Extended Reality Informative Space (XRIS), a HoloLens-based platform that spatially anchors construction data within an immersive environment through gaze-triggered information retrieval and embedded game-based mechanics. These advances address persistent challenges in the construction industry, such as design misinterpretation, information asymmetry, and inefficient constructability reviews that arise from conventional 2D documentation and static 3D models. Through semi-structured interviews with industry stakeholders and XR specialists, this study identified the benefits of XRIS implementation and the impact of gamification on XRIS use.

The findings indicated that XRIS supported spatial cognition and improved access to construction information relative to traditional methods. Respondents described the system as intuitive, reporting gains in perception, retention, and articulation of design logic. Gamified discovery cues were associated with sustained user engagement throughout the controlled prototype sessions. Respondents distinguished curiosity-driven exploration from task-driven persistence prompted by completion goals and progress cues.

This study goes beyond existing XR-BIM and gamification research by integrating spatially anchored information retrieval, immersive interaction triggers, and embedded gamification within a single professional workflow configuration. In contrast to the generic XR-BIM, XRIS is a system that integrates motivational elements into the visualisation layer as opposed to being an external layer. This configuration was evaluated with practising construction stakeholders, a population underrepresented in gamified XR research. This positions the study as a meaningful contribution to the design of immersive tools for professional construction practice.

Several limitations bound these findings. The primary purpose of this study was to focus on the development of an XRIS prototype that can be applied practically in the industry and establishing proof of concept in controlled conditions before its deployment in the field. The controlled laboratory environment was hence not entirely representative of environmental conditions and pressures on working construction sites. Respondents also voiced concerns over long-term use of the devices and the learning curve for new users. The constraints of ergonomic and usability may affect the feasibility of adoption.

Future research should consider the performance of XRIS with alternative head-mounted devices over longer periods. Comparative experimental designs would enhance causal claims by benchmarking XRIS with non-gamified XR and conventional review to assess whether there is discomfort reduction or onboarding time reduction. The collaborative potential suggested by respondents requires multi-user deployment studies to be conducted. A mixed-method assessment, utilising efficiency, error, and decision quality measures, along with qualitative feedback, would evaluate the impact of workflow.

ACKNOWLEDGEMENT

The authors wish to pay gratitude to Taylor's University Malaysia for providing a conducive research environment.

REFERENCES

Abouelkhier, N., Rauf, A., Luyten, L., & Shafiq, M. T. (2025). Mitigating early-phase miscommunication between architects and structural engineers: A systematic review and framework linking BIM/VR strategies to communication barriers. *Architectural Engineering and Design Management*, 1-23. https://doi.org/10.1080/17452007.2025.2586665

Alam, N., Saha, N., Gadow, V., Harik, R., & Ryu, J. (2025). Role of extended reality (XR) technologies in maintenance operations: Trends, challenges, and integration in Industry 4.0. *Manufacturing Letters, 44*, 1545-1557. https://doi.org/10.1016/j.mfglet.2025.06.174

Alhady, A., Alanany, M., Khodair, Y., Salem, S., & El Maghraby, Y. (2024). Integrating building information modelling (BIM) and extended reality (XR) in the transportation infrastructure industry. *Advances in Bridge Engineering, 5*(1), Article 22. https://doi.org/10.1186/s43251-024-00132-6

Ali, O., Murray, P., Al-Ahmad, A., & Tahat, L. (2024). An integrated framework for addressing the challenges and strategies of technology adoption: A systematic review. *Emerging Science Journal, 8*(3), 1215-1242. https://doi.org/10.28991/ESJ-2024-08-03-025

Almeida, C., Kalinowski, M., Uchôa, A., & Feijó, B. (2023). Negative effects of gamification in education software: Systematic mapping and practitioner perceptions. *Information and Software Technology, 156*, Article 107142. https://doi.org/10.1016/j.infsof.2022.107142

Alshboul, O., & Shehadeh, A. (2025). Adaptive integration of BIM and Navisworks for real-time clash detection using the XGBoost algorithm. *Journal of Construction Engineering and Management, 152*(1), Article 04025232. https://doi.org/10.1061/JCEMD4.COENG-16001

Alzarrad, A., Miller, M., Durham, L., & Chowdhury, S. (2024). Revolutionising construction safety: Introducing a cutting-edge virtual reality interactive system for training US construction workers to mitigate fall hazards. *Frontiers in Built Environment, 10*, Article 1320175. https://doi.org/10.3389/fbuil.2024.1320175

Arslan, O., Akdoğan, A., & Dogan, M. D. (2025). TinkerXR: In-situ, reality-aware CAD and 3D printing interface for novices. In *Proceedings of the ACM Symposium on Computational Fabrication* (pp. 1-19). Association for Computing Machinery. https://doi.org/10.1145/3745778.3766651

Azmi, N. F., Chai, C. S., & Chin, L. W. (2017). Building information modelling (BIM) in the architecture, engineering, and construction (AEC) industry: A case study in Malaysia. In *Proceedings of the 21st International Symposium on Advancement of Construction Management and Real Estate* (pp. 401-412). Springer. https://doi.org/10.1007/978-981-10-6190-5_36

Bhatarai, R., Banihashemi, S., Shakouri, M., & Antwi-Afari, M. (2025). Integration of augmented reality with building information modelling: Design optimisation and construction rework reduction perspective. *Archives of Computational Methods in Engineering, 32*(4), 2233-2254. https://doi.org/10.1007/s11831-024-10211-6

Bitrián, P., Buil, I., & Catalán, S. (2021). Enhancing user engagement: The role of gamification in mobile apps. *Journal of Business Research, 132*, 170-185. https://doi.org/10.1016/j.jbusres.2021.04.028

Bizzi, L. (2023). Why to gamify performance management? Consequences of user engagement in gamification. *Information & Management, 60*(3), Article 103762. https://doi.org/10.1016/j.im.2023.103762

Boddy, C. R. (2016). Sample size for qualitative research. *Qualitative Market Research: An International Journal, 19*(4), 426-432. https://doi.org/10.1108/QMR-06-2016-0053

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Brooks, F. P., Jr. (1987). Walkthrough-A dynamic graphics system for simulating virtual buildings. In *Proceedings of the 1986 Workshop on Interactive 3D Graphics* (pp. 9-21). Association for Computing Machinery. https://doi.org/10.1145/319120.319122

Chang, D., Xiang, Y., Zhao, J., Qian, Y., & Li, F. (2022). Exploration of brain–computer interaction for supporting children's attention training: A multimodal design based on attention network and gamification design. *International Journal of Environmental Research and Public Health, 19*(22), Article 15046. https://doi.org/10.3390/ijerph192215046

Chen, Q., Li, C., Xu, X., Mao, P., & Xiong, L. (2024). Intelligent systems integrating BIM and VR for urban subway microenvironmental health risks management. *Buildings, 14*(7), Article 1912. https://doi.org/10.3390/buildings14071912

Cheng, C., & Chau, C. L. (2022). Gamification-based intervention for enhancing team effectiveness and coping flexibility: Randomised controlled trial. *Frontiers in Psychiatry, 13*, Article 941252. https://doi.org/10.3389/fpsyt.2022.941252

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. SAGE Publications.

Davila Delgado, J. M., Oyedele, L., Beach, T., & Demian, P. (2020). Augmented and virtual reality in construction: Drivers and limitations for industry adoption. *Journal of Construction Engineering and Management, 146*(7), Article 04020079. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001844

Dawood, N., Sriprasert, E., Mallasi, Z., & Hobbs, B. (2003). Development of an integrated information resource base for 4D/VR construction process simulation. *Automation in Construction, 12*(2), 123-131. https://doi.org/10.1016/S0926-5805(02)00045-6

Encarnação, R., Reuter, J., Ferreira Dias, M., & Amorim, M. (2021). Gamification as a driver of motivation in organisations: A bibliometric literature review. In *Proceedings of the Ninth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'21)* (pp. 167-172). Association for Computing Machinery. https://doi.org/10.1145/3486011.3486440

Fernández-Caramés, T. M., & Fraga-Lamas, P. (2024). Forging the industrial metaverse towards Industry 5.0: Where augmented/mixed reality, IIoT, opportunistic edge computing, and digital twins meet. *IEEE Access, 12*. https://doi.org/10.1109/ACCESS.2024.3422109

FMI, & PlanGrid. (2018). *Construction disconnected: The high cost of poor data and miscommunication*. Autodesk. https://www.autodesk.com/blogs/construction/construction-disconnected-fmi-report/

Guimarães, A., Cavalcanti, M., & Vasconcelos, B. M. (2024). Gamification using technologies for occupational safety training in the civil construction sector. *Work, 77*(2), 477-485. https://doi.org/10.3233/WOR-220685

Gomez, M., Pahor, A., Carrillo, A., Seitz, A., & Jaeggi, S. (2024). Promoting executive functions in older adults: The role of gamification and individual differences. *Innovation in Aging, 8*(Suppl. 1), 1152. https://doi.org/10.1093/geroni/igae098.3695

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods, 18*(1), 59–82. https://doi.org/10.1177/1525822X05279903

Gornall, J., Peña, A., Pinto, H., Rojas, J., Correa, F., & García, J. (2026). Extended reality in Construction 4.0: A systematic review of applications, implementation barriers, and research trends. *Applied Sciences, 16*(1), Article 9. https://doi.org/10.3390/app16010009

Hajji, R., Kharroubi, A., Ben Brahim, Y., Bahhane, Z., & El Ghazouani, A. (2022). Integration of BIM and mobile augmented reality in the AECO domain. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLVI-4/W3-2021*, 131-138. https://doi.org/10.5194 isprs-archives-XLVI-4-W3-2021-131-2022

Hammedi, W., Leclercq, T., & Steils, N. (2024). Gamification myopia: Satiation effects in gamified activities. *Journal of Service Research, 27*(2), 213-230. https://doi.org/10.1177/10946705231190873

Henry, J., Li, F., & Arnab, S. (2024). On the pre-perception of gamification and game-based learning in higher education students: A systematic mapping study. *Simulation & Gaming, 55*(6), 985-1010. https://doi.org/10.1177/10468781241271082

Heo, J., Bhardwaj, K., & Gavrilovska, A. (2023). FleXR: A system enabling flexibly distributed extended reality. In *Proceedings of the 14th Conference on ACM Multimedia Systems* (pp. 1-13). Association for Computing Machinery. https://doi.org/10.1145/3587819.3590966

Irani, Z., Abril, R. M., Weerakkody, V., Omar, A., & Sivarajah, U. (2023). The impact of legacy systems on digital transformation in European public administration: Lessons learned from a multiple-case analysis. *Government Information Quarterly, 40*(1), Article 101784. https://doi.org/10.1016/j.giq.2022.101784

Jacobides, M. G., Ma, M. D., Trantopoulos, K., & Vassalos, V. (2024). The business value of gamification. *California Management Review, 66*(2), 91-107. https://doi.org/10.1177/00081256231218469

Jalo, H., & Pirkkalainen, H. (2024). Effect of user resistance on the organisational adoption of extended reality technologies: A mixed-methods study. *International Journal of Information Management, 75*, Article 102731. https://doi.org/10.1016/j.ijinfomgt.2023.102731

Jato-Espino, D., Vila-Cortavitarte, M., Rodriguez-Hernandez, J., & Castro-Fresno, D. (2024). Gamification as a tool to improve educational and training outcomes in civil engineering. *Journal of Civil Engineering Education, 150*(4), Article 03124001. https://doi.org/10.1061/JCEECD.EIENG-1962

Kim, J., & Castelli, D. M. (2021). Effects of gamification on behavioural change in education: A meta-analysis. *International Journal of Environmental Research and Public Health, 18*(7), Article 3550. https://doi.org/10.3390/ijerph18073550

Klein, G. (2021). Viewing gamification design limitations and weaknesses through a pandemic lens. *Societies, 11*(4), Article 137. https://doi.org/10.3390/soc11040137

Koivisto, J., & Hamari, J. (2014). Demographic differences in perceived benefits from gamification. *Computers in Human Behaviour, 35*, 179-188. https://doi.org/10.1016/j.chb.2014.03.007

Le, N., Tran, D., & Sturgill, R. (2024). Content analysis of three-dimensional model technologies and construction applications: Current trends and future directions. *Sensors, 24*(12), Article 3838. https://doi.org/10.3390/s24123838

Lee, M., Chai, C., Xiong, Y., & Gui, H. (2022). Technology acceptance model for building information modelling-based virtual reality (BIM-VR) in cost estimation. *Journal of Information Technology in Construction, 27*, 904-925. https://doi.org/10.36680/j.itcon.2022.044

Leite, R. M. C., Winkler, I., & Alves, L. R. G. (2022). Visual management and gamification: An innovation for disseminating information about production to construction professionals. *Applied Sciences, 12*(11), Article 5682. https://doi.org/10.3390/app12115682

Leite, R. M. C., Alves, L. R. G., Cardoso, L. D. S. P., & Neto, H. M. M. (2023). How has gamification in the production sector been developed in manufacturing and construction workplaces? *Buildings, 13*(10), Article 2614. https://doi.org/10.3390/buildings13102614

Li, T., Hu, H., Ma, H., Ma, J., & Li, Q. (2025). Using virtual reality to enhance learning performance and address educational resource disparities in architectural history courses. *Sustainability, 17*(3), Article 866. https://doi.org/10.3390/su17030866

Lim, W. M. (2025). Sample size in qualitative research: Moving from data saturation to theoretical saturation. *Journal of Global Marketing, 38*(5), 449-459. https://doi.org/10.1080/08911762.2025.2590757

Liu, P., Xing, J., Xiong, R., & Tang, P. (2022). *Sharing construction safety inspection experiences and site-specific knowledge through XR-augmented visual assistance* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2205.15833

Liu, S., Ma, G., Tewogbola, P., Gu, X., Gao, P., Dong, B., … Wu, Y. (2023). Game principle: Enhancing learner engagement with gamification to improve learning outcomes. *Journal of Workplace Learning, 35*(5), 450-462. https://doi.org/10.1108/JWL-11-2022-0160

Manzano-León, A., Camacho-Lazarraga, P., Guerrero, M. A., Guerrero-Puerta, L., Aguilar-Parra, J. M., Trigueros, R., & Alias, A. (2021). Between level up and game over: A systematic literature review of gamification in education. *Sustainability, 13*(4), Article 2247. https://doi.org/10.3390/su13042247

Margheritti, S., Marcucci, S., & Miglioretti, M. (2025). Bridging the gaps: Examining the impact of technology-based active learning in workplace safety training through a systematic literature review. *Safety, 11*(1), Article 5. https://doi.org/10.3390/safety11010005

Mohanty, S., & Christopher, B. P. (2024). The role of gamification research in human resource management: A PRISMA analysis and future research direction. *SAGE Open, 14*(2), Article 21582440241243154. https://doi.org/10.1177/21582440241243154

Monteiro, R. H. B., de Almeida Souza, M. R., Oliveira, S. R. B., dos Santos Portela, C., & de Cristo Lobato, C. E. (2021). The diversity of gamification evaluation in software engineering education and industry: Trends, comparisons, and gaps. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET)* (pp. 154-164). IEEE. https://doi.org/10.1109/ICSE-SEET52601.2021.00025

Morris, A., Guan, J., & Azhar, A. (2021). An XRI mixed-reality Internet of Things architectural framework towards immersive and adaptive smart environments. In *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)* (pp. 68-74). IEEE. https://doi.org/10.1109/ISMAR-Adjunct54149.2021.00026

Nyqvist, R., Peltokorpi, A., Lavikka, R., & Ainamo, A. (2025). Building the digital age: Management of digital transformation in the construction industry. *Construction Management and Economics, 43*(4), 262-283. https://doi.org/10.1080/01446193.2024.2416033

Oke, A. E., Aliu, J., Tunji-Olayeni, P., & Abayomi, T. (2024). Application of gamification for sustainable construction: An evaluation of the challenges. *Construction Innovation, 24*(4), 1066-1084. https://doi.org/10.1108/CI-09-2022-0247

Oke, A. E., Aliu, J., Mwanaumo, E. M. U. M., Abayomi, T., & Kahanji, C. (2023). Leveraging gamification to enhance productivity and employee engagement in the Nigerian construction industry. *Built Environment Project and Asset Management, 13*(6), 813-829. https://doi.org/10.1108/BEPAM-04-2023-0064

Oral, E., Chawla, R., Wijkstra, M., Mahyar, N., & Dimara, E. (2023). From information to choice: A critical inquiry into visualisation tools for decision-making. *IEEE Transactions on Visualization and Computer Graphics, 30*(1), 359-369. https://doi.org/10.1109/TVCG.2023.3326593

Ortiz-Rojas, M., Chiluiza, K., Valcke, M., & Bolanos-Mendoza, C. (2025). How gamification boosts learning in STEM higher education: A mixed-methods study. *International Journal of STEM Education, 12*(1), Article 1. https://doi.org/10.1186/s40594-024-00521-3

Othman, N. A. F., Zainoddin, A. I., Jaini, A., Ismail, M., & Radzi, S. F. M. (2023). Management of gamified lesson: Understanding aesthetics and acceptance among learners. *Information Management and Business Review, 15*(4), 155-162. https://doi.org/10.22610/imbr.v15i4(SI)I.3588

Patricio, R., Moreira, A. C., & Zurlo, F. (2022). Gamification in innovation teams. *International Journal of Innovation Studies, 6*(3), 156-168. https://doi.org/10.1016/j.ijis.2022.05.003

Pavlou, M., Laskos, D., Zacharaki, E. I., Risvas, K., & Moustakas, K. (2021). XRSISE: An XR training system for interactive simulation and ergonomics assessment. *Frontiers in Virtual Reality, 2*, Article 646415. https://doi.org/10.3389/frvir.2021.646415

Pham, V. B., Wong, S. P. P., & Abbasnejad, B. (2025). *A systematic review of criteria influencing the integration of BIM and immersive technology in building projects*. https://hdl.handle.net/10779/rmit.28570676

Priyantini, D., Amalin, A. M., & Amalia, N. (2023). Designing user interface and user experience by using user-centred design on the gamified platform "Ezedu." *Jurnal Teknologi Pendidikan: Jurnal Penelitian dan Pengembangan Pembelajaran, 8*(4), 834–844. https://doi.org/10.33394/jtp.v8i4.8944

Radanovic, M., Khoshelham, K., Fraser, C. S., & Acharya, D. (2023). Continuous BIM alignment for mixed reality visualisation. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, X-1/W1-2023*, 279-286. https://doi.org/10.5194/isprs-annals-X-1-W1-2023-279-2023

Rahiman, H. U., Kodikal, R., & Suresh, S. (2023). Game on: Can gamification enhance productivity? *F1000Research, 12*, Article 818. https://doi.org/10.12688/f1000research.131579.2

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68-78. https://doi.org/10.1037/0003-066X.55.1.68

Sacks, R., Perlman, A., & Barak, R. (2013). Construction safety training using immersive virtual reality. *Construction Management and Economics, 31*(9), 1005-1017. https://doi.org/10.1080/01446193.2013.828844

Safikhani, S., Keller, S., Schweiger, G., & Pirker, J. (2022). Immersive virtual reality for extending the potential of building information modelling in the architecture, engineering, and construction sector: A systematic review. *International Journal of Digital Earth, 15*(1), 503-526. https://doi.org/10.1080/17538947.2022.2038291

Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. *Educational Psychology Review, 32*(1), 77-112. https://doi.org/10.1007/s10648-019-09498-w

Schwarz, A., Verkooijen, K., de Vet, E., & Simons, M. (2024). Perceived barriers and facilitators regarding the implementation of gamification to promote physical activity in the neighbourhood: Interview study among intermediaries. *JMIR Serious Games, 12*, Article e52991. https://doi.org/10.2196/52991

Smith, S. A., & Mulligan, N. W. (2021). Immersion, presence, and episodic memory in virtual reality environments. *Memory, 29*(8), 983-1005. https://doi.org/10.1080/09658211.2021.1953535

Speiser, K., & Teizer, J. (2024). Formalising virtual construction safety training: A schematic data framework enabling real-world hazard simulations using BIM and location tracking. *Journal of Information Technology in Construction, 29*, 980-1004. https://doi.org/10.36680/j.itcon.2024.043

Struye, J., Van Damme, S., Bhat, N. N., Troch, A., Van Liempd, B., Assasa, H., Lemic, F., Famaey, J., & Torres Vega, M. (2024). Toward interactive multi-user extended reality using millimetre-wave networking. *IEEE Communications Magazine, 62*(8), 54-60. https://doi.org/10.1109/MCOM.001.2300804

Trindade, N. V., Ferreira, A., Pereira, J. M., & Oliveira, S. (2023). Extended reality in AEC. *Automation in Construction, 154*, Article 105018. https://doi.org/10.1016/j.autcon.2023.105018

Vona, F., Warsinke, M., Kojić, T., Voit-Antons, J.-N., & Möller, S. (2025). User-centric evaluation methods for digital twin applications in extended reality. In *2025 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)* (pp. 142-146). IEEE. https://doi.org/10.1109/AIxVR63409.2025.00028

Yalçın, Z. Ö. (2024). Integrating gamification with BIM for enhancing participatory design. *Journal of Computational Design, 5*(2), 317-344. https://doi.org/10.53710/jcode.1505309

Yang, H., Wang, L., Hu, Z., & Li, D. (2024). Understanding the failing of social gamification: A perspective of user fatigue. *Electronic Commerce Research and Applications, 64*, Article 101369. https://doi.org/10.1016/j.elerap.2024.101369

Yu, D., Chen, W., Meng, T., Yu, H., & Mahamadu, A. M. (2024). Applications of immersive technologies in construction computing: A systematic literature review. In *Proceedings of the 2024 European Conference on Computing in Construction* (Vol. 5). European Council on Computing in Construction. https://doi.org/10.35490/EC3.2024.230

Yu, X., Lee, B., & Sedlmair, M. (2024). Design space of visual feedforward and corrective feedback in XR-based motion guidance systems. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1-15). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642143

Pertanika J. Sci. & Technol. 34 (3): 1685 - 1702 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Studying the Effects of Knowledge Distillation and the Fragility of Lightweight BERT-based Models

Allistair Nallie Konsil[1]*, Stephanie Chua[1], Jacey Lynn Minoi[1], Md Mizanur Rahman[2], Gerraint Gillan[2], Rafazila Ramli[2], Rasitasam Safii[3], Ahmad Sofian bin Shminan[4], and Lee Jun Choi[4]

[1]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[2]Department of Community Medicine and Public Health, Faculty of Medicine and Health Sciences, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[3]Kulliyyah of Medicine, International Islamic University Malaysia, Jalan Sultan Ahmad Shah, Bandar Indera Mahkota, 25200 Kuantan, Pahang, Malaysia
[4]Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia

ABSTRACT

Large language models (LLMs) such as BERT (Bidirectional Encoder Representations from Transformers) have caused strides of progress in the field of natural language processing (NLP), with lightweight variants such as DistilBERT, MobileBERT and TinyBERT being developed to lower the resource requirements to deploy the models in real-world settings. However, while past research has investigated improving non-distilled models on a variety of benchmarks by changing their architecture, there are limited studies that explore how the lightweight variants may perform in the same conditions. This study applies modifications and ensemble techniques on lightweight BERT models in extractive question answering (QA) to address this gap. The experiments were conducted using three datasets: SQuAD, AdversarialQA, and a newly curated Sexual and Reproductive Health QA (SRHQA) dataset consisting of 1000 samples. From the results, it was shown that applying the same modifications that would enhance the base BERT models to the lightweight variants generally caused a 0.14 - 5.20% F1 decrease in performance, with marginal exceptions observed in specific dataset-model combinations.

ARTICLE INFO

Article history:
Received: 08 December 2025
Accepted: 06 June 2026
Published: 19 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.17

E-mail addresses:
allistair99@gmail.com (Allistair Nallie Konsil)
chlstephanie@unimas.my (Stephanie Chua)
jacey@unimas.my (Jacey Lynn Minoi)
rmmizanur@unimas.my (Md Mizanur Rahman)
22010275@siswa.unimas.my (Gerraint Gillan)
22010339@siswa.unimas.my (Rafazila Ramli)
razitasham@iium.edu.my (Rasitasam Safii)
sasofian@unimas.my (Ahmad Sofian bin Shminan)
cljun@unimas.my (Lee Jun Choi)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

Ensembling, on the other hand, showed improvements across all the datasets, ranging from 2.19 - 19.46% F1 over the BERT baseline. The results of the study highlight the sensitivity of the lightweight models, their trade-offs with efficiency, and that an ensemble is a valid approach to utilising the lightweight models without architectural modifications.

Keywords: BERT, ensemble, lightweight models, question answering, robustness

INTRODUCTION

Transformer-based large language models such as BERT have been shown to perform quite well compared to previous architectures in several NLP tasks, with key advantages being their ability to process more tokens in parallel and being able to produce high-quality embeddings. However, one of the drawbacks of the base BERT model is that it is computationally intensive, making it difficult to deploy where resources are limited, such as edge devices. This led to the development of lightweight variants like DistilBERT, MobileBERT and TinyBERT, which apply several techniques to reduce their overall parameter count.

While these compressed models have their benefits of being faster and having a lower barrier for use, there has been some research that shows they are also more fragile. Their performance varies when they are trained in architectural or when retrained in more specialised domains. This makes them less reliable and brings into question their ability for use in more high-stakes domains such as sexual and reproductive health (SRH), where misinformation can have dire consequences.

Ensembling shows promise in enhancing the stability of LLMs by aggregating multiple predictions, which in turn reduces variance and the effects of outliers. However, this technique is often utilised for its ability to edge out extra performance on benchmarks using large non-distilled models, with limited usage for lightweight models in more niche areas like extractive question answering on a domain. With these gaps in mind, this study investigates how lightweight BERT-based models perform on a mix of general and domain-specific datasets for extractive question answering after applying architectural modifications, while also looking into how ensembling can be utilised to improve the performance using the modified models.

Related Works

Transformer Model

A large part of research done in NLP before 2017 utilised neural networks such as recurrent neural networks (RNNs) or long short-term memory networks (LSTMs), which are sequential in nature, with a major restriction being that they would become exponentially more expensive as the sequence length grew.

The introduction of the Transformer architecture (Vaswani et al., 2017) was a paradigm shift because it showed that neural networks can be trained to process tokens simultaneously instead of just sequentially. They utilised self-attention to create high-quality embeddings that translated into improved performance on benchmarks. The design choices also allowed for faster training and iteration on longer texts; other than that, they used positional encoding, which helped preserve the word order, while residual connections reduced the problem of vanishing gradients and improved stability. All these design choices contributed to the Transformer architecture achieving state-of-the-art results on a variety of NLP tasks, such as machine translation and sentiment analysis, surpassing previous architectures at the time.

The Transformer architecture, which was originally used for machine translation, consisted of an encoder and decoder component. However, researchers realised you can utilise just the encoder or decoder for specialised tasks, for example, a decoder-only architecture such as GPT (generative pretrained transformer) has been shown to excel at auto-regressive sequence modelling and is effective at open-ended text generation. Other than that, BERT (Devlin et al., 2019), which is an encoder-only model pretrained on two pre-training tasks, masked language modelling and next sentence prediction, has been shown to produce high-quality embeddings that significantly improve the performance on downstream tasks. However, due to their large size, these models are computationally expensive.

Lightweight BERT Models

With the success of large pretrained transformer models and the demand for a more practically deployable solution that can be utilised in resource-constrained conditions, this led to research branching into increasing efficiency. DistilBERT Sanh et al. (2019) was one of the earliest studies in reducing the parameter count of the base BERT model. This was done by reducing the number of encoder layers and using knowledge distillation to make the remaining layers mimic the teacher model; they were able to retain 95% of BERT's performance with a reduced parameter count. DistilBERT showcased that transferring knowledge from a large teacher model to a smaller student model is a viable strategy for reducing the total parameter count and the computational cost of inference.

MobileBERT Sun et al. (2020) was a study aiming to develop a model that could be used on edge devices. They utilised architectural modifications such as bottleneck structures and thin layers to compress the model, retaining performance. MobileBERT also applies knowledge distillation to distil the intermediate representations and predictions from the teacher BERT.

TinyBERT Jiao et al. (2020) focus on knowledge distillation, where they employ a two-stage distillation process, first by doing a general distillation, then a task-specific

distillation. For both general and task-specific distillation, they investigated passing the generated embeddings, attention and hidden layers from the teacher to the student while still retaining performance even with drastically reducing the number of parameters. The three models utilised knowledge distillation to reduce the parameter count while retaining their performance, allowing them to be used in contexts where a full-sized BERT model may be too computationally taxing. However, there is still the question of whether these models can really be utilised in areas where accuracy is critical, although they may be efficient and could be deployable with weak robustness and sensitivity to modifications, the trade-off may not be worth it.

Modifications to BERT

After the introduction of BERT, extensive usage showed that the generated embeddings were able to perform quite well in downstream tasks. This motivated researchers to try to create variants of the architecture that may edge out more performance. One of the most common and easily implemented modifications is the addition of layers utilising other neural network architectures, such as BiLSTM or convoluted neural networks, with the aim of adding better capture of sequential dependencies. For example, Dhar and Chauhan (2023) added BiLSTM layers into DistilBERT and showed that it could improve the performance of classification on web content datasets. Similarly, Fang et al. (2025) applied a BERT-BiLSTM model for detecting malicious comments and noted that they were able to get improved performance compared to the baseline BERT model.

Although these studies show that modifications as simple as adding another layer can improve performance, some studies have reported that for distilled models, these modifications may cause worsening performance. Diddee et al. (2022) and Lorenzoni et al. (2024) reported in their papers that some modifications to architecture or hyperparameters can disturb the optimised weights of lightweight models, which points to a trade-off between model performance and robustness in distilled models. For this study, fragility is defined as the tendency for performance to degrade when the models are subjected to changes in hyperparameters, architecture or subject domain.

Ensemble Learning

Ensemble learning is a technique that integrates the predictions of multiple models to improve the accuracy and robustness. It has been used extensively in NLP due to the technique being able to reduce variance and make up for the weakness of individual models. Traditional ensemble strategies such as bagging Breiman (1996) and boosting Schapire (1990) demonstrate how training either multiple independently sampled models or sequentially refined learners can yield stronger generalisation, while stacking Wolpert (1992) introduces a meta-learning framework that combines model outputs

through an additional predictive layer. Averaging is the selected ensemble method due to several reasons. Unlike the previously mentioned methods, averaging does not require ensemble-specific training, such as training a meta learner; this helps isolate the effects of modifications from improvements introduced during training.

Ensembles in Question Answering

The ensemble technique is effective in a variety of NLP tasks, including extractive question answering. For instance, Lee et al. (2019) combined BioBERT with BiLSTM-Attention layers for biomedical natural language inference, using majority voting to aggregate predictions from multiple models. Zhou et al. (2019) used weighted averaging with BERT and BiDAF models to achieve top results on SQuAD 2.0. Pranesh et al. (2020) introduced QuesBELM, a stacking ensemble of BERT-based models for Google's Natural Questions dataset, demonstrating the benefits of combining models of varying architectures.

In social media health text classification, Dang et al. (2020) applied ensembles of BERT-Large, Bio+Clinical BERT, and BERT-Base, achieving higher F1 scores through averaging and cross-validation. Pearce et al. (2021) added BiLSTM layers into BERT embeddings, forming a layered ensemble that improved extractive QA performance across NewsQA and CovidQA. These studies showed the potential and effectiveness of using ensembles for improving the performance and stability on a variety of tasks and datasets.

Despite extensive work on ensemble learning with large BERT models, there is limited research focusing on using lightweight BERT in ensembles. The robustness of distilled models when applying architectural modifications also remains underexplored, especially in extractive QA tasks. This study addresses these gaps by examining the sensitivity of lightweight BERT models to modifications and evaluating whether averaging ensembles can enhance robustness. The findings aim to highlight the effects of applying naïve modifications to optimised distilled models without further knowledge distillation.

MATERIALS AND METHODS

Data Sources and Collection

The study uses three datasets, all of which are in English, representing both general-purpose and domain-specific contexts for extractive question answering tasks. The first dataset is the Stanford Question Answering Dataset (SQuAD). SQuAD is a widely used benchmark dataset for extractive QA. It contains over 100,000 question-answer pairs based on Wikipedia articles with extractive QA requiring the model to determine the answer span of a question within a given context. The large sample size and comprehensive range of general topics make it a suitable dataset for setting the baseline model performance to compare with existing studies. The second dataset is AdversarialQA, which extends the QA challenge by introducing adversarial examples designed to test the robustness of QA systems.

During the development of the dataset, a human-in-the-loop approach was used, where a comparison of the F1 score from human and model predictions was done. Only questions where the base model achieves an F1 score below a threshold of 40% are retained. Adding AdversarialQA to the study allows us to evaluate how lightweight BERT variants handle complex contexts. Lastly, the third dataset is the Sexual and reproductive health question answering (SRHQA) dataset, which was custom-curated for this research. It focuses on the topic of sexual and reproductive health, an area where accuracy is critical due to the consequences of misinformation. Sources were selected through a systematic review; paragraphs of text were extracted from the sources and were then proofread and verified by professionals from two universities in Malaysia. The dataset contains approximately 1,000 annotated samples, which consist of context, question and answer pairs. The SRHQA dataset is used to evaluate how well models can perform under domain shift.

The SRHQA dataset was constructed using publicly available educational and medical texts sourced from verified educational and medical publications. No human subjects were recruited or involved in data collection; all source material was drawn from existing published documents. Accordingly, formal IRB or ethics board approval was not required under standard institutional research guidelines. The dataset has been made publicly available on HuggingFace under a CC-BY 4.0 license to support reproducibility and future research in domain-specific QA.

To further characterise the dataset, Table 1 provides aggregate corpus statistics. The dataset comprises 64 passages and 1,010 QA pairs. Average context length is 379.6 words per passage (range: 100-507 words), and average answer span length is 7.3 words (range: 1-57 words). Domain coverage spans sexual health and STIs, anatomy and puberty, relationships and emotions, menstruation and hygiene, mental and emotional wellbeing, reproduction and pregnancy, abuse, consent and safety, cancer and medical conditions, and gender and sexual identity.

Data Preprocessing

The domain-specific dataset was formatted to follow the structure of SQuAD to keep it consistent with the other two datasets. The annotation tool Haystack was utilised for labelling and formatting the dataset for the extractive question answering task.

Table 1
SRHQA dataset aggregate statistics

Statistic	Value
Total passages	64
Total QA pairs	1,010
Average Context Length	379.6 (range: 100-507)
Average Answer Span Length	7.3 (range: 1-57)

Additional steps were taken to clean the dataset, such as removing unnecessary characters and adjusting the token length of the text to fit within the model context window.

The dataset was then partitioned into two subsets, one for training and another for validation, with a ratio of 8 to 2. The training subset was used during the model and ensemble training process, enabling the models to learn the relationships between questions and their corresponding answers, optimise performance, and assess generalisation capabilities. The validation subset, on the other hand, was reserved for evaluating the model's performance on held-out validation data.

Model Configuration

In this study, four pre-trained models were selected for extractive question answering. These include BERT (csarron/bert-base-uncased-squad-v1), DistilBERT (distilbert/distilbert-base-uncased-distilled-squad), MobileBERT (csarron/mobilebert-uncased-squad-v1), and TinyBERT (deepset/tinybert-6l-768d-squad2), all of which were obtained from the Hugging Face model repository. Each selected model was pretrained on the SQuAD dataset since it provides a large sample for training, allowing the models to get familiar with the extractive QA task while also setting a baseline. Input tokenisation was performed using the respective tokeniser for each model. TinyBERT, however, utilises the BERT base tokeniser due to its architectural compatibility. The maximum input sequence length was set to 512 tokens for all models to ensure a consistent and sufficient context window.

It should be noted that the TinyBERT checkpoint used (deepset/tinybert-6l-768d-squad2) was pre-trained on SQuAD v2, while the remaining models were pre-trained on SQuAD v1. This versioning difference was retained as it reflects the best publicly available TinyBERT checkpoint for extractive QA at the time of experimentation. SQuAD v2 introduces unanswerable questions in addition to the answerable questions present in SQuAD v1; however, since evaluation in this study focuses exclusively on answerable span prediction across all datasets, the practical impact of this checkpoint difference is expected to be minimal. The relative performance patterns of TinyBERT remain consistent with the other lightweight models across AdversarialQA and SRHQA, suggesting that the versioning difference does not materially affect the study's conclusions.

To evaluate the effects of architectural modification, a bidirectional LSTM (BiLSTM) layer was added to each model variant. The BiLSTM layer was placed between the transformer encoder and the final classification layer. To make sure that the BiLSTM layer is compatible with the encoder output, the hidden size is set to match the dimensions. The modified models were fine-tuned using the same hyperparameters as the base model to isolate the impact of the architectural change.

During training, the models were trained for three epochs using a learning rate of 3e-5 and the Adam optimiser with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 1e\text{-}8$. The learning rate followed a linear scheduling strategy. The training batch size was set to 6, while evaluation

was performed with a batch size of 60. A fixed random seed of 42 was used throughout all experiments. This was a deliberate methodological choice to ensure full deterministic reproducibility. The hyperparameters of the training and evaluation setup are shown in Table 3. All experiments were conducted on Google Colab using a single T4 GPU with approximately 16 GB of VRAM, an Intel Xeon CPU and 13 GB of RAM. The versions used for the software are listed below in Table 2.

Transfer Learning Setup

Each of the pretrained models was fine-tuned on the AdversarialQA and SRHQA datasets to get the performance before and after being adapted to each dataset. The pretrained weights were loaded from the Hugging Face repository during initialisation and then fine-tuned on the dataset through single-stage training. During fine-tuning, variants of each base model were created, which included the addition of a BiLSTM layer. This was done to investigate the sensitivity of lightweight models to architectural changes. Apart from this modification, all models followed the same training pipeline and hyperparameter settings that were specified in the model configuration section. After fine-tuning, model performance was assessed, with evaluation done on the held-out validation split of each respective dataset.

Table 2
Software environment

Component	Version
Python	3.10
PyTorch	2.0.1+cu118
HuggingFace Transformers	4.35.0
CUDA	11.8
datasets (HF)	2.14.0

Table 3
Hyperparameters

Parameter	Value
Batch size (Training)	6
Batch size (Evaluation)	60
Epochs	3
Random seed	42
Learning rate	3e-5
Optimiser	Adam
Adam β_1	0.9
Adam β_2	0.999
Adam ε	1e-8
Maximum Context Length	512 tokens

Ensemble Implementation

The averaging ensemble is generally composed of fine-tuned variants of the base lightweight models, DistilBERT, MobileBERT and TinyBERT. These models were chosen to evaluate the robustness and whether using an ensemble can mitigate any performance drops. Variants of these models, fine-tuned on specific datasets, were grouped accordingly. For instance, to evaluate the ensemble on the AdversarialQA dataset, the corresponding fine-tuned variants of the three lightweight models were selected. Each base model was trained independently to ensure that any performance improvements are from the ensemble instead of shared training procedures, while also accommodating the hardware setup.

Predictions from the three models were aggregated using a probability averaging strategy. This was done by adapting the question answering pipeline from Hugging Face to accept multiple models simultaneously. During inference, each model produces a probability distribution over the potential answer tokens. Z-score normalisation is then applied to these distributions to make the magnitude of predictions consistent across models to prevent any one model from dominating the final prediction. The final ensemble prediction was obtained by averaging the normalised probabilities token-wise. This averaged probability distribution was then passed to the existing post-processing function, which maps the probabilities to the final answer span. Model diversity was primarily derived from differences in distillation levels and architectural variations among the lightweight models.

Evaluation Metrics

Model performance was evaluated using metrics used in the SQuAD paper for the extractive question answering task, which are F1 and EM, which focus on partial and exact correctness of predictions. F1 is used to measure the overlap between the prediction and the ground truth; it acts as a measure for soft correctness and is calculated by the harmonic mean of precision and recall, as shown in Equation 1. The Exact Match (EM) is a binary metric that evaluates whether the model can predict the same answer as the ground truth. Together, these metrics provide a comprehensive assessment of both approximate and exact model performance.

$$F_1\ Score = \frac{2\ \times Precision \times Recall}{Precision + Recall} \quad [1]$$

RESULTS AND DISCUSSION

Individual Model Performance

The baseline evaluation of lightweight BERT models revealed that all three variants, DistilBERT, MobileBERT, and TinyBERT, performed well on g eneral-purpose datasets but exhibited varying degrees of sensitivity when the addition of a BiLSTM layer was applied, as shown in Table 4. On the SQuAD dataset, DistilBERT achieved an F1 score of 86.56% and an Exact Match (EM) of 78.94%, while MobileBERT and TinyBERT recorded F1 scores of 89.94% and 88.91%, and EM scores of 82.5% and 81.49%, respectively. These results align with previous studies reporting competitive performance for lightweight variants in standard QA tasks. However, with the additional layer, a consistent decrease in performance is shown, ranging from 0.13 to 0.31% for all lightweight models. These decreases aren't seen in the non-distilled model, which has an increase of 0.16% F1 and 0.41% EM. The ensemble also shows an increase of 2.9% F1 and 3.8 EM.

The AdversarialQA dataset was designed to confuse models by including adversarial samples that scored a low F1 score during the curation of the dataset. Due to this, a lower overall performance is to be expected, as seen by the F1 and EM scores of all models in Table 5. Following fine-tuning on the AdversarialQA training set, DistilBERT achieved 23.2% EM and 32.79% F1, while MobileBERT and TinyBERT scored 29.6% EM / 39.53% F1 and 27.5% EM / 37.85% F1 for the base models, respectively. These results demonstrate the model's sensitivity to context perturbations and the limitations of modified distilled architectures in handling adversarial samples. However, the use of an ensemble was able to help in reducing the weakness of individual models, resulting in the highest F1 and EM out of the evaluated models.

In the SRHQA dataset, which contains domain-specific terminology, baseline models displayed a further decrease in performance, as shown in Table 6. MobileBERT reached 73.94% F1 and 57.23% EM, and TinyBERT recorded 62.57% F1 and 42.08% EM.

Table 4
SQuAD dataset

Model	F_1	EM	F1 Difference	EM Difference
DistilBERT	86.56	78.94	-	-
DistilBERT + BiLSTM	86.64	78.81	0.08	-0.13
MobileBERT	89.94	82.5	-	-
MobileBERT + BiLSTM	89.63	82.19	-0.31	-0.31
TinyBERT	88.91	81.49	-	-
TinyBERT + BiLSTM	88.77	81.28	-0.14	-0.21
BERT	88.11	80.48	-	-
BERT + BiLSTM	88.27	80.89	0.16	0.41
Ensemble	91.01	84.28	2.9	3.8

Table 5
AdversarialQA dataset

Model	F_1	EM	F1 Difference	EM Difference
DistilBERT	32.79	23.20	-	-
DistilBERT + BiLSTM	32.49	22.20	-0.30	-1.00
MobileBERT	39.53	29.60	-	-
MobileBERT + BiLSTM	39.84	28.90	0.31	-0.70
TinyBERT	37.85	27.50	-	-
TinyBERT + BiLSTM	37.64	27.20	-0.21	-0.30
BERT	38.08	27.00	-	-
BERT + BiLSTM	38.33	27.10	0.25	0.10
Ensemble	40.27	30.00	2.19	3.00

Table 6
SRHQA dataset

Model	F_1	EM	F1 Difference	EM Difference
DistilBERT	59.25	40.2	-	-
DistilBERT + BiLSTM	60.08	40.4	0.83	0.2
MobileBERT	79.14	64.26	-	-
MobileBERT + BiLSTM	73.94	57.23	-5.2	-7.03
TinyBERT	66.23	44.26	-	-
TinyBERT + BiLSTM	62.57	42.08	-3.66	-2.18
BERT	65.86	42.48	-	-
BERT + BiLSTM	68.61	44.26	2.75	1.78
Ensemble	85.32	67.23	19.46	24.75

DistilBERT, when modified, on the other hand, achieved 60.08% F1 and 40.4% EM, an increase from the baseline performance. This result is considered an outlier instead of a pattern since the same performance gains are not seen on other datasets. The reduced performance highlights the challenge of transferring lightweight models to specialised domains and the importance of assessing model robustness beyond standard benchmarks.

Impact of Architectural Modifications

To evaluate how architectural modifications affect the performance of lightweight transformer models, a BiLSTM layer was added between the BERT encoder and the final output head of each base model. The non-distilled model showed the expected outcome of having a slight performance boost, which is consistent with the results seen from other studies. However, for the lightweight models, a consistent decrease in performance is seen across all three datasets.

This suggests that the internal representations acquired during distillation are highly optimised, and additional structural modifications disrupt the alignment learned from the teacher models. In contrast, the base BERT model generally showed stable or improved performance under similar modifications, which further supports that lightweight models are less robust compared to the base model.

From the results, there is a trade-off in efficiency and robustness seen in how the lightweight models adapt to the addition of a BiLSTM layer in their architecture and in how the base BERT model performs. Generally, we see a decrease in performance. This can be attributed to the smaller parameter count of the lightweight models and the learned representations from the teacher already making use of the limited parameters. With the addition of the BiLSTM layer, the model needs to relearn representations that can utilise the characteristics of the new layer. For lightweight models, this is difficult; however, for the non-distilled BERT model, we can see that it can adapt by inferring from the increased performance across the three datasets. Practically, when working with compact models, architectural modifications need to be carefully considered and planned.

These results also show lightweight models having limited ability for generalising across different datasets and domains. For example, when evaluating on SQuAD, a general domain dataset and SRHQA, a domain-specific dataset, both of which follow the same formatting, they perform at different levels of accuracy. This is reflected in lower F1 and EM scores on SRHQA relative to SQuAD. The limited ability to generalise can be attributed to the training process. Since these models are obtained through knowledge distillation, they inherently have a narrower representation space than their teacher models. This reduces flexibility when exposed to domains with distinct linguistic structures. The AdversarialQA dataset was excluded from this generalisation analysis, as its intentionally difficult questions yielded consistently low F1 scores (<40%), making direct comparison inappropriate.

The general pattern of weakening performance in lightweight models following the addition of a BiLSTM layer can be attributed to the limited internal representations produced by knowledge distillation. Because knowledge distillation compresses the teacher's learned representations into a smaller parameter space, the student models carry less internal redundancy, making them more sensitive to downstream changes (Bai et al., 2024). When a BiLSTM layer is added after the transformer encoder, it introduces additional sequential processing that requires the model to project its compressed representations into a new feature space. For full-sized BERT, which retains broader representational capacity and higher internal redundancy, such structural additions are feasible without disturbing the underlying representations. The contrast between BERT's stable or improved performance and the lightweight models' weakening performance supports the hypothesis that the fragility of distilled models is specifically a consequence of the limitations introduced during the distillation process itself (Diddee et al., 2022; Lorenzoni et al., 2024).

An exception to this general trend is observed with DistilBERT on the SRHQA dataset, where the addition of a BiLSTM layer produced a marginal improvement (F1: 59.25% → 60.08%, EM: 40.20% → 40.40%). This improvement does not replicate across other datasets. DistilBERT degrades under BiLSTM modification on AdversarialQA (-0.30 F1) and shows no meaningful improvement on SQuAD (+0.08 F1). The non-replicating nature of this result identifies it as context-specific rather than structural. This study employed three datasets spanning general-purpose, adversarial, and domain-specific contexts precisely to test whether the fragility pattern holds across conditions. The directional consistency of degradation across datasets, 7 of 9 lightweight model-dataset combinations show degradation or no meaningful improvement, distinguishes the general finding from random variation. An isolated positive deviation that does not replicate across the same model on other datasets cannot be attributed to architectural advantage and does not undermine the overall pattern.

The BERT + BiLSTM configuration on AdversarialQA yields F1 of 38.33% and EM of 27.10%, representing an improvement of +0.25 F1 and +0.10 EM over the BERT baseline (F1 38.08%, EM 27.00%). All models in Table 5 were fine-tuned on the AdversarialQA training set before evaluation. Among lightweight models, DistilBERT and TinyBERT show the expected degradation under BiLSTM modification (-0.30 and -0.21 F1, respectively), while MobileBERT shows a marginal positive deviation (+0.31 F1). Consistent with the DistilBERT SRHQA case, this deviation does not replicate across datasets — MobileBERT degrades under BiLSTM modification on both SQuAD (-0.31 F1) and SRHQA (-5.20 F1). The cross-dataset consistency of MobileBERT's degradation on two of three datasets, and the absence of replication of the positive deviation, confirms that the AdversarialQA result is context-specific and does not constitute evidence of a structural architectural benefit. This behaviour aligns with Tran and Kretchmar (2024), who showed that models trained on SQuAD-style datasets do not inherently generalise to adversarial inputs and often fail under misleading or contradictory question formulations.

Ensemble Performance

Table 7 presents the comparison between the performance of the BERT baseline and the averaging ensemble. For the SQuAD dataset, the ensemble was able to score an F1 of 91.01 and EM of 84.28, improving by 2.9 and 3.8 points from the BERT baseline, respectively. A similar trend is observed on the AdversarialQA dataset, where the ensemble obtains an F1 score of 40.27 and an EM score of 30.00, which is a substantial improvement of 2.19 in F1 and 3.00 in EM over the baseline model. The results on SRHQA show the largest improvements with the ensemble scoring an F1 of 85.32 and EM of 67.23, which improve on the results of the base BERT model by 19.46 and 24.75, respectively.

Table 7
Ensemble performance

Dataset	Model	F_1	EM	F1 Difference	EM Difference
SQuAD	BERT	88.11	80.48	-	-
	Ensemble	91.01	84.28	2.9	3.8
AdversarialQA	BERT	38.08	27.00	-	-
	Ensemble	40.27	30	2.19	3.00
SRHQA	BERT	65.86	42.48	-	-
	Ensemble	85.32	67.23	19.46	24.75

In general, a trend appears when comparing the results across all three datasets, which is that the use of an ensemble to aggregate the predictions often resulted in better performance than the base model. This is especially evident when the sub model performs well, as seen by MobileBERT's results on the SRHQA dataset, which in turn contributes largely to the overall ensemble performance. These improvements show that a simple method, such as averaging, can stabilise and mitigate weaknesses of individual models, providing a practical approach for improving lightweight models without resorting to architectural-level changes.

Across all experiments, the results indicate that lightweight models struggle to accommodate architectural changes. Adjustments such as adding extra layers frequently led to declines in performance when compared to their unmodified counterparts. While some isolated cases showed gains in F1 or EM, these improvements were inconsistent across datasets and typically involved trade-offs elsewhere. In contrast, the base BERT model demonstrated stable and consistent gains when the same BiLSTM layer was introduced, suggesting that larger, non-distilled architectures are more resilient to structural modifications. Meanwhile, the averaging ensemble approach yielded reliable improvements across all datasets without requiring any architectural changes, reinforcing ensembling as an effective and robust strategy for enhancing lightweight model performance.

CONCLUSION

The results of the experiments highlight that the lightweight transformer models are sensitive to naïve modifications, as seen through the addition of a BiLSTM layer, which consistently resulted in degraded performance and through the model's inability to retain performance when undergoing domain shift. Using averaging, an improvement in F1 and EM was seen across all datasets, with the highest increase from the SRHQA dataset. Overall, the results indicate that ensembles can stabilise the predictions of individual models and that the method of distillation provides enough diversity to make averaging viable.

Theoretical and Practical Contributions

The study provides empirical evidence on the fragility of lightweight models when architectural changes are applied. Previous studies have explored applying ensembling

and modifications to BERT; however, there are limited studies that look into applying them to lightweight models. By recording the results of how these techniques affect the distilled models, this research contributes to the knowledge of trade-offs between efficiency, robustness and model complexity.

The results of this study provide practical considerations when deploying lightweight models on edge devices, such as whether transfer learning is appropriate, whether redistilling is required and whether modifications are a viable idea. Although the lightweight models are more appropriate for edge or mobile devices because of their lower requirements, usage requires careful deliberation. Ensembling is a reliable way of using multiple models to improve overall performance without increasing the architectural complexity of each model. Domain shift requires careful consideration and ample preparation, especially for the data to sufficiently adapt the general-purpose model to a specialised domain.

Limitations and Future Directions

The results of the experiments highlight the fragility of distilled BERT models when architectural modifications are applied. To address the fragility, future research should look into utilising the existing distilled weights instead of trying to retrain the model. For example, techniques like Low-Rank Adaptation (LoRA) could be an area to pursue since it minimally disrupts the original weights.

Dataset size was also observed to be a factor in the model performance on the respective validation set, with models trained on the larger dataset, SQuAD, often showing better generalisation than those trained on the much smaller SRHQA dataset. Future research could investigate methods for enhancing the generalisation of these lightweight models in low-resource settings.

The use of an ensemble showed a general improvement in performance over the base BERT model; the strategy that was used was simple averaging. From the base results, it was shown that MobileBERT often outperformed the other lightweight models. A weighted averaging configuration with MobileBERT having a higher weight may be more effective. Other than this, research could investigate how to specifically address the weakened robustness of lightweight post-distillation, possibly through a secondary distillation phase like in TinyBERT or an improved knowledge distillation pipeline. Future work should also consider input-level robustness evaluation through controlled perturbations such as typo injection, synonym replacement, and context shuffling, to quantify model sensitivity beyond a single adversarial dataset.

Overall, these directions emphasise the need to balance model efficiency with robustness, highlighting that performance alone is insufficient when lightweight models are highly sensitive to architectural changes. Continued research along these lines could enhance the practical utility and adaptability of distilled models for real-world applications.

In conclusion, while lightweight transformer models offer better efficiency, they also show notable sensitivity when modifications are added to their architecture and domain shifted. Using averaging ensembles, the predictions of multiple models can be stabilised and the performance enhanced, especially in domain-specific contexts. These findings contribute to both theoretical insights and practical guidance for the development and deployment of reliable, efficient NLP systems. By highlighting the trade-offs between computational efficiency, robustness, and accuracy, this research informs best practices for leveraging distilled BERT models in real-world extractive question answering applications.

ACKNOWLEDGEMENT

This work was supported by the Vice-Chancellor High Impact Research Grant (UNI/F08/VC-HIRG/85491/P05-02/2022). The authors would like to acknowledge Universiti Malaysia Sarawak (UNIMAS) for the support provided towards this publication.

REFERENCES

Bai, G., Chai, Z., Ling, C., Wang, S., Lu, J., Zhang, N., Shi, T., Yu, Z., Zhu, M., Zhang, Y., Song, X., Yang, C., Cheng, Y., & Zhao, L. (2024). Beyond efficiency: A systematic survey of resource-efficient large language models. *arXiv*. https://doi.org/10.48550/arXiv.2401.00625

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123-140. https://doi.org/10.1007/BF00058655

Dang, H., Lee, K., Henry, S., & Uzuner, Ö. (2020). Ensemble BERT for classifying medication-mentioning tweets. In G. Gonzalez-Hernandez, A. Z. Klein, I. Flores, D. Weissenbacher, A. Magge, K. O'Connor, A. Sarker, A.-L. Minard, E. Tutubalina, Z. Miftahutdinov, & I. Alimova (Eds.), *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task* (pp. 37-41). Association for Computational Linguistics. https://aclanthology.org/2020.smm4h-1.5/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dhar, A., & Chauhan, I. (2023). A multi-category content classification approach using DistilBERT transformer and BiLSTM. In *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)* (pp. 727-732). IEEE. https://doi.org/10.1109/ICCCIS60361.2023.10425205

Diddee, H., Dandapat, S., Choudhury, M., Ganu, T., & Bali, K. (2022). Too brittle to touch: Comparing the stability of quantisation and distillation towards developing low-resource MT models. In P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A.

Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi, & M. Zampieri (Eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)* (pp. 870-885). Association for Computational Linguistics. https://aclanthology.org/2022.wmt-1.80/

Fang, Z., Zhang, H., He, J., Qi, Z., & Zheng, H. (2025). Semantic and contextual modelling for malicious comment detection with BERT-BiLSTM. In *2025 4th International Symposium on Computer Applications and Information Technology (ISCAIT)* (pp. 1867-1871). IEEE. https://doi.org/10.1109/ISCAIT64916.2025.11010257

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., & Liu, Q. (2020). TinyBERT: Distilling BERT for natural language understanding. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4163-4174). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.372

Lee, L.-H., Lu, Y., Chen, P.-H., Lee, P.-L., & Shyu, K.-K. (2019). NCUEE at MEDIQA 2019: Medical text inference using an ensemble BERT-BiLSTM-attention model. In D. Demner-Fushman, K. B. Cohen, S. Ananiadou, & J. Tsujii (Eds.), *Proceedings of the 18th BioNLP Workshop and Shared Task* (pp. 528-532). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5058

Lorenzoni, G., Portugal, I., Alencar, P., & Cowan, D. (2024). Exploring variability in fine-tuned models for text classification with DistilBERT. *arXiv*. https://doi.org/10.48550/arXiv.2501.00241

Pearce, K., Zhan, T., Komanduri, A., & Zhan, J. (2021). A comparative study of transformer-based language models on extractive question answering. *arXiv*. https://doi.org/10.48550/arXiv.2110.03142

Pranesh, R. R., Shekhar, A., & Pallavi, S. (2020). QuesBELM: A BERT-based ensemble language model for natural questions. In *2020 5th International Conference on Computing, Communication and Security (ICCCS)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICCCS49678.2020.9277176

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv*. https://doi.org/10.48550/arXiv.1910.01108

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning, 5*(2), 197-227. https://doi.org/10.1007/BF00116037

Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., & Zhou, D. (2020). MobileBERT: A compact task-agnostic BERT for resource-limited devices. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2158-2170). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.195

Tran, S. Q., & Kretchmar, M. (2024). Towards robust extractive question answering models: Rethinking the training methodology. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 2222-2236). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.129

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *arXiv*. https://doi.org/10.48550/arXiv.1706.03762

Wolpert, D. H. (1992). Stacked generalisation. *Neural Networks, 5*(2), 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1

SUPPLEMENTARY DATA

Code and Data Availability

Training and evaluation code can be found at https://github.com/AllistairNK/SRHQAExperiments. The SRHQA dataset can be found at https://huggingface.co/datasets/allistair99/SRH1K. The dataset is released under the CC-BY 4.0 license.

Pertanika J. Sci. & Technol. 34 (3): 1703 - 1725 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Enhanced Deepfake Image Detection and Classification Using Deep Learning

Debasish Samal[1], Dimple Nagpal[2], Prateek Agrawal[2,3,4]*, Vishu Madaan[3], Chhavi Sharma[5], and Wou Onn Choo[6]

[1]Department of Computer Applications, School of Computer Applications, Lovely Professional University, 144411 Phagwara, Punjab, India
[2]Department of Computer Science and Engineering, School of Computer Science and Engineering, Lovely Professional University, 144411 Phagwara, Punjab, India
[3]Department of Computer Science and Engineering, Rungta College of Engineering and Technology, Rungta International Skills University, 490024 Bhilai, Chhattisgarh, India
[4]Department of Computer Science and Information Technology, Faculty of Data Science and Information Technology, INTI International University, 71800 Nilai, Negeri Sembilan, Malaysia
[5]Department of Computer Science and Engineering, School of Engineering and Technology, SGT University, 122505 Gurugram, Haryana, India
[6]Department of Computer Science and Information Technology, Faculty of Data Science and Information Technology, INTI International University, 71800 Nilai, Negeri Sembilan, Malaysia

ABSTRACT

The development of Generative Adversarial Networks (GANs) for generating realistic deepfake content through artificial intelligence brings a complex task to authenticate deepfake images. The spread of deepfakes leads to widespread distrust among people while simultaneously hurting both personal and community-based reputations through deceptive information distribution. This paper aims to develop an optimised deep learning model based on the EfficientNetV2-B1 architecture designed specifically for binary image classification of distinguishing real or fake. The proposed method has been implemented on the extensive 140k Real and Fake faces dataset from Kaggle. As other existing detection and classification approaches rely heavily on pre-trained models and a limited dataset, our model delivers compelling performance through a customised training methodology. As a result, the model was able to achieve 99.91% training and 98.76% testing accuracy coupled with the precision, recall and F1-score at 99.28%, 98.43%, 98% respectively. Furthermore, the model's performance is compared to the current techniques to show its reliability. The model's

ARTICLE INFO

Article history:
Received: 04 October 2025
Accepted: 25 May 2026
Published: 19 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.18

E-mail addresses:
debasishsamal01@gmail.com (Debasish Samal)
dimplenagpal009@gmail.com (Dimple Nagpal)
dr.agrawal.prateek@gmail.com (Prateek Agrawal)
dr.vishumadaan@gmail.com (Vishu Madaan)
chhavibhardwaj12@gmail.com (Chhavi Sharma)
wouonn.choo@newinti.edu.my (Wou Onn Choo)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

predictions are also interpreted using XAI visualisations, providing explainable insights into the areas of an image that contribute to its classification as either real or fake.

Keywords: AI security, computer vision, deepfake detection, efficientNetV2-B1, image forensics, peaceful society, social safety, social security

INTRODUCTION

Deepfake technology has quickly advanced to generate realistic, high-resolution media displaying individuals making false statements and acting in ways that do not truly exist. This is quite threatening to the privacy and safety of a human being. A sampling of a movie star could be produced to make it appear as if they are endorsing a brand or making offensive comments (Kim & Cho, 2021). Their lives can be disrupted because of how damaging this could be to their reputation, career, and overall credibility. Beyond the safety threat, deepfakes can be weaponised against individuals to disclose sensitive and private information, or worse, persuade someone into malicious acts. The rapid and relentless advancement of deepfake technology compels individuals and organisations to be hyper-vigilant and implement layers of security for any potential concerns associated with deepfake-related issues (Kosarkar et al., 2023). One of the key components in effective detection and classification practices related to this issue requires advanced deep learning methods (Ein et al., 2022) to counter generative AI methods implemented by GANs (Goodfellow et al., 2014), which acknowledges why this is a critical issue to address. Accurate classifications are among the most important considerations for deepfake detection and classification (Krueger et al., 2023), whether that's accurately classifying potential malignant cells in neuroimaging data or identifying directly fraudulent financial transactions.

These advanced methods anticipate and neutralise potential risks, while at the same time improving overall security measures. Without advanced detection and classification tools, subjects would be more vulnerable to scams, hacking, and possibly misdiagnosed conditions in healthcare settings. Investing in the latest technologies and consistently upgrading detection systems allows individuals to be one step ahead of risks and ultimately make it a safer environment for everyone (Kumar et al., 2023).

Rapid releases of deepfakes are abusing people’s right to privacy and causing irreparable harm to their reputations. As shown in Figure 1, the face of German-born seven-time Mr Olympia & Hollywood actor "Arnold Alois Schwarzenegger" (Right) is overlaid on the previous actor "Clint Eastwood" (Left) from the Dirty Harry movie series, all created by using a DeepFaceLab tool. This makes it hard to decipher for most audience members which photo was the original and which image was the deepfake variant.

Deepfakes are such a trend now that they are being created in rapid succession and making headlines, even today, constantly abusing people’s privacy and causing damage to

their reputation. In 2025, just before the Mahakumbh festival (the world's largest cultural and spiritual gathering that only takes place every 144 years in India), pictures of World Wrestling Entertainment (WWE) wrestlers with fake backgrounds have been disseminated, e.g., in Figure 2.

False scenarios of the Indian actor and comedian, Brahmanandam, and popular wrestlers like John Cena, Brock Lesnar, and Roman Reigns being created and depicted in AI-generated photos and videos on social media are leading to misinformation and bewilderment among fans as well as the public. The distribution of these AI-generated deepfakes reminds us of the moral quandaries that come with very advanced technologies. These image forgery cases are a glaring example of the need to be digitally literate and cautious when dealing with online content.

Figure 1. Deepfake example: Original image (left) and fake image (right)
Note. Adapted from YouTube (https://www.youtube.com/watch?v=i42HGtt5fao)

Figure 2. Viral AI-generated photos of well-known celebrities from DFRAC: John Cena and Brock Lesnar (left), Brahmanandam and Roman Reigns (right)
Note. Adapted from DFRAC (https://www.dfrac.org)

Motivation

Deepfakes hold potential terrorist threats to the truthfulness of information and could potentially sway public opinion to influence political vectors or even defame persons by way of fake materials. Keeping both society as a whole and individuals safe from the negative consequences of deepfake technology is indeed a tough mission.

The CNN model, which has been proposed in this study, proves reliable for image-related binary classification problems. It keeps its best accuracy even in practical computer vision applications like image categorisation, no matter how small its footprint is. Consequently, it can be the first step in tackling the problem of deepfake identification from various sources. As the deepfake creation process keeps on improving, the deepfake detection system that needs to be robust will also have to undergo the adaptation process.

Main Contributions

The research paper's main contributions are as follows:

- Architectural Innovation: The development of a customised and computationally efficient deep learning architecture for deepfake image detection by integrating custom convolutional blocks into the EfficientNetV2-B1 backbone. The research is utilised to replace the standard EfficientNet classification top with a custom head specifically designed to improve the model's ability to capture subtle features of a face image.
- Demonstrated real or fake Image Classification through XAI: The research presented in this study involves Explainable AI (XAI) using Gradient-weighted Class Activation Mapping (Grad-CAM). It allows the visual identification of regions within the face image, which greatly impacts the prediction of the model.
- This research follows a novel training approach for the proposed model. The core function of this approach is the synergistic employment of the Adamax optimiser with a custom 'lr_ask' callback function to adjust the learning rate dynamically. The regularisation strategy used in the study combined Global Max Pooling with specific L1/L2 regularisers and a dropout rate of 0.4 to improve model performance. This new approach results in a demonstrable and optimal balance between high accuracy and computational efficiency capable of running on moderate GPUs.
- The comparative efficiency analysis highlights key metrics performance such as Precision, Recall, and F1-Score. It also compares the proposed model with related research works to show the model's suitability for deployment on resource-limited devices, which is a critical factor for practical applications.

The rest of the research paper's framework includes a detailed examination of existing literature in the 'Literature Review' section, which provides an overview of various deepfake detection techniques and relevant studies. The paper provides detailed information

about the proposed model's structural details with the dataset used in the section 'Materials and Methods'. Section 'Results and Discussion' presents the performance results of the proposed EfficientNetV2B1 model. The study summarises its main results in the conclusion and proposes potential avenues for additional research in future studies.

LITERATURE REVIEW

The identification of manipulated media that includes synthetic images and videos created by AI calls for the detection and classification of deepfakes to be done accurately. Conventional methods allowed the processing of manually created features from photos or video movements. The fake video shows that there are detailed inconsistencies in the movement, the textures, and the facial features (Abu-Ein, 2022; Li et al., 2020).

Image-Based CNNs, for instance, are largely used for the purposes of deepfake detection and image classification, besides being engaged to solve the irregularities in movements (Zhang et al., 2022), textures (Sharma et al., 2022), and facial landmarks (Wen & Xu, 2019). At the same time, the CNN model can detect the signs of an anomaly, for instance, the inconsistent expressions or the unusual textures that are a result of the deepfake process. It is possible to reconfigure this model for video deepfake, where it looks for temporal differences between several frames (Dhar et al. 2021).

Tolosana et al. (2020) have noted the success of pre-trained autoencoders in the task of copying a group of images and comparing the copied data to a reconstructed image made by a trained model. If the reconstructed image differs from the original, it may be a modified one. The approach, by and large, is the identification of low-level artefacts in deepfakes.

The methods created by Goodfellow et al. (2020) leveraged the technology of GANs to fabricate and, at the same time, detect deepfakes. To improve the generalisation for detecting adversarially created fakes, these methods make use of a trained discriminator, referred to as 'D' in the GAN architecture and a generator 'G' to create images for distinguishing between artificially generated and authentic ones. Besides that, adversarial training is also used to amplify the performance, which means that in the training process of the model, a photo or video with substantially altered input designed to trick the model into giving incorrect results is included. This, therefore, increases the model's capability of detecting deepfakes and makes it tougher for different types of manipulation of the model.

One of the main products of this development was the creation of the Vision Transformer (ViT) model by Indigo's Enseign Meetier Lesche Infantilien lab (2024), which offers different ViT models for the detection of state-forensic crimes. Vision Transformers can identify even very small changes in the image or failures since they can represent an image's total information. Along with transfer learning, the models rely on pre-trained models on the deepfake datasets, a feature that has proven to be very useful in the domain of deepfake images.

Liang et al. (2023) decided to employ ensemble methods to improve the accuracy of detection. Thus, they merged the predictions of several models, including CNNs, RNNs, and ViTs, to get one final prediction. Ensemble methods combine a multitude of techniques (Suratkar et al., 2023) that blend the different strengths of each and, as a result, get a more effective measure against deepfakes, particularly when a wide range of manipulations is used. The effectiveness of these techniques is not constant but rather changes depending on the type of deepfake and the context. The main success is generally when the combinations of multiple methods are used, which allows for more robust protection in the face of the targeted characteristics that keep on evolving (Fatoni et al., 2025). The literature review in Table 1 gives a detailed overview of the different approaches that have been implemented, along with the description of their methodology, advantages, and drawbacks.

Table 1
Summary of deepfake detection methods

Methodology	Approach	Strengths	Limitations
Physical Face Feature Extraction	Recognises anomalies in movement, textures, and facial features (Zhang et al., 2022).	Effective for specific visual inconsistencies in manual manipulations.	Controlled scalability; not effective against sophisticated AI-created deepfakes.
Facial Appearance Attribute-based CNNs	Operates convolutional layers to spot artefacts such as inconsistent expressions or unusual textures (Dhar et al., 2021).	Robust performance in image classification identifies spatial anomalies.	Does not verify temporal information; requires extensive datasets.
Autoencoder-based Detection	Assesses original and restructured face images to identify discrepancies (Tolosana et al., 2020).	Identifies low-level anomalies; focuses on reconstruction inaccuracies.	Performance depends heavily on model size and the quality of the dataset.
GAN Discriminator-based Detection	Uses GAN discriminators to differentiate between real and artificial content (Goodfellow et al., 2020).	Generalises well to adversarial content; resilient against wide manipulation.	Computationally intensive; requires hyperparameter tuning & a less-used detection approach.
Vision Transformers (ViTs)	Encodes global image information and draws on transfer learning on pre-trained deepfake datasets (Wang et al., 2024).	Effective at identifying minor artefacts; excels at global representation of face features.	Computationally demanding; needs adjustment for specific datasets.
Ensemble Techniques	Combines CNNs, RNNs, and ViTs for improved precision (Liang et al., 2023; Suratkar et al., 2023).	Aggregates powers of different models; resistant to varied manipulations.	Increased complexity and computational needs require thorough integration.

The application of deep learning over image data has become a foundational element of modern cyber forensic investigations, providing robust frameworks for identifying digital anomalies (Awasthi et al., 2023). As the complexity of manipulations scales from static images to continuous video streams, recent 2026 breakthroughs emphasise spatiotemporal deep learning architectures such as 3DCNNs, 3DResNets, and Variational Autoencoders (VAEs) to facilitate real-time video-based deepfake detection (Agrawal et al., 2026). Concurrently, across broader computer vision domains, there is a pronounced shift towards highly optimised, lightweight image recognition models. Techniques such as architectural pruning and attention mechanisms have been successfully deployed in lightweight models like ILN-YOLOv8 to drastically reduce computational overhead while maintaining high precision (Zhou et al., 2025). These parallel advancements in cyber forensics and lightweight architectural design strongly reinforce the necessity of parameter-efficient models capable of rapid, accurate inference in resource-constrained environments.

While existing methods like ensemble models and Vision Transformers achieve high detection accuracy, they suffer from significant limitations. Large backbones like VGG19 and ResNet50 have millions of parameters and require substantial GFLOPs, making them unsuitable for mobile or real-time deployment, resulting in high computational cost. Other limitation includes models overfitting to specific artefacts in the training set.

This research addresses these gaps by proposing a lightweight EfficientNetV2-B1 architecture. Unlike existing heavy models, the proposed approach prioritises a low parameter count (~8.2M) and computational efficiency (<1.0 GFLOPs) without sacrificing accuracy, specifically targeting resource-constrained environments.

MATERIALS AND METHODS

The aimed methodology for the study is illustrated in the flow diagram shown in Figure 3. The pipeline begins with an initial pre-processing phase, where the input facial images undergo resizing and augmentation to ensure consistency. Following this, the processed images are fed into the EfficientNet-based model architecture for feature extraction and analysis. Finally, the network outputs a binary image classification, determining whether the given input is a real or fake image.

Figure 4 illustrates the input layer, which accepts 150x150 pixel face images and undertakes image normalisation to proceed with feature extraction. The proposed methodology imposes an EfficientNetV2-B1-based architecture to detect deepfake images, taking advantage of its efficient convolutional layers and inverted residual blocks to extract detailed spatial features from input images.

The Fully Connected Layer is accountable for computing the final classification scores. This layer generates the scores that inform decision-making. The Output Layer utilises these scores to produce the ultimate prediction, categorising the input image as either authentic or a forgery, with the fake one being termed as such.

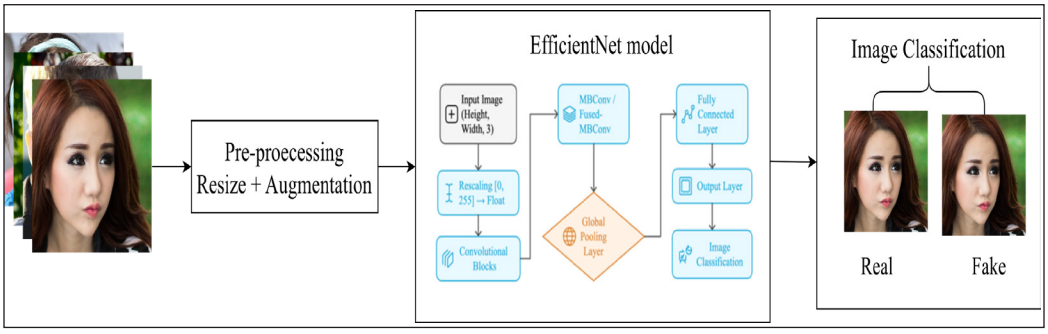


Figure 3. Proposed methodology

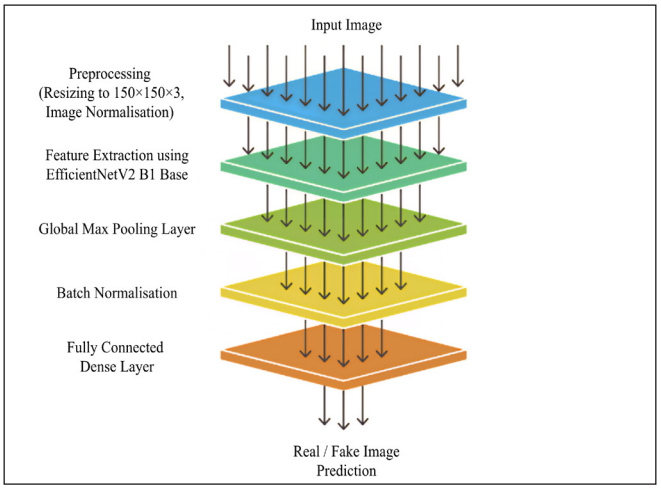


Figure 4. Stages of model implementation

Rationale for Using EfficientNet in Deepfake Detection

Both EfficientDet (Tan et al., 2020) and EfficientNet (Tan et al., 2019a) are strong neural network architectures; nevertheless, they have been designed to handle different issues. EfficientNet is a deep learning model that is mostly used in image classification tasks. It achieves this by effectively scaling model dimensions – depth, width, and resolution, using compound scaling and MBConv blocks, thereby bringing about higher accuracy and efficiency of image-level prediction tasks.

The main difference with EfficientDet is that the latter is a combination of EfficientNet as a backbone architecture with a Bi-directional Feature Pyramid Network (BiFPN) and uses this fusion for multi-scale features merging. This merges the features at different scales to allow accurate localisation and detection of objects; thus, it is a network designed for object detection & segmentation tasks. The training of EfficientNet from scratch on a big,

diverse dataset is easy, and the model converges quickly. Besides, pre-trained EfficientNet models on large datasets, such as ImageNet, provide a lever in transfer learning; thus, there is a faster convergence rate and an increased generalisation capability. The design of the model's balanced complexity is such that it avoids the problem of overfitting while it is doing so effectively on a dataset that has both real and fake images. EfficientNet is top among the choices for the deepfake detection task, and the rationale of this is that it is very accurate, it brings about efficient feature extraction, and the image classification task requirements are met through it.

EfficientNetV2-B1 Base Architecture

The EfficientNetV2 B1 model (Tan et al., 2021), one of the EfficientNetV2 family of architectures, is the basic network utilised for the work of feature extraction. The design is essentially made up of several layers of convolutions and inverted residual blocks, which allow it to capture the hierarchies of spatial and even the finest features in the input images. It is the combination of convolutional and pooling layers that has been shown in the standard architecture structure in Figure 5(a). Compared to the more complicated designs, this base model is a light, advanced deep learning architecture with 8.2 million parameters, which is aimed at lessening the complexity on relatively large datasets and is still computationally feasible for low-resource environments, such as a device with moderate GPU capabilities. Table 2 gives a summary of all the parameters and their values used in the proposed model categorically.

The design features multiple layers of convolutions, which are supplemented by such operations as Fused-MBConv (Tan et al., 2019b) for quicker execution and lower memory usage, that basically 'pull' complex spatial and hierarchical features from the input images. The fundamental measures include a first convolutional layer that is highly efficient in processing the input image. In these layers, the breakdown of standard convolution into

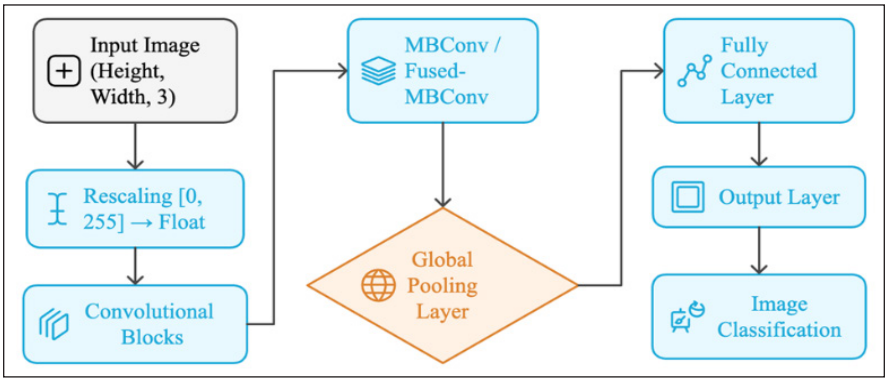


Figure 5(a). Standard structure of EfficientNetV2-B1 base model

depthwise and pointwise convolution results in substantial efficiency loss while the complexity of features is still maintained.

The suggested technique takes full advantage of EfficientNetV2-B1 features to build an accurate and computationally efficient deepfake detection model. Once the model accepts the hyperparameters, it achieves the highest accuracy of distinguishing real and fake images by employing a mixture of multi-scale features and advanced convolution blocks for capturing subtle patterns. The model is equipped with 'categorical_crossentropy' loss, optimised with 'Adamax' optimiser, further upgraded with 'lr_ask' callback function and data augmentation for better generalisation and higher accuracy.

In Figure 5(b), the standard EfficientNetV2-B1 backbone extracts features, which are then processed by the custom classification head utilising GAP, L1/L2 regularisation, and aggressive Dropout (0.4) to enhance deepfake image detection.

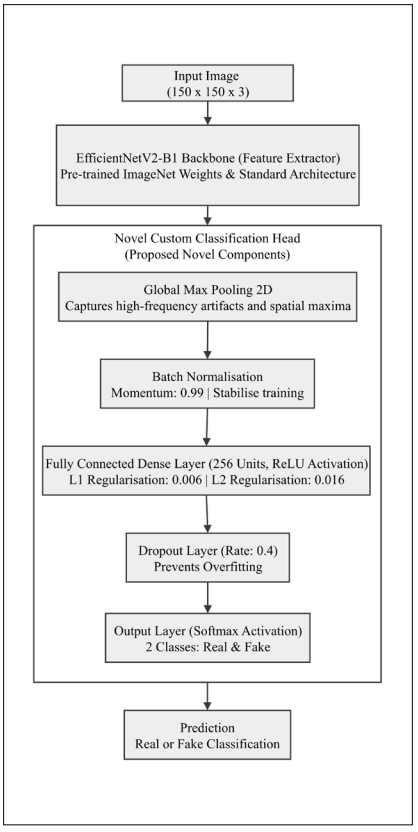


Figure 5(b). Detailed architecture of the proposed model

Table 2
Summary of hyperparameters and values

Category	Parameters	Value
Input and Architecture Specifications	Image Size	150 × 150
	Base Model	EfficientNetV2-B1
	Pooling Strategy	Global Max Pooling
	Trainable Layers	All layers (base_model.trainable=True)
	Batch Normalisation	Momentum: 0.99 Epsilon: 0.001
	Custom Dense Layer	256 Units
	Dense Layer Activation	ReLU
	Output Layer Activation	Softmax
Regularisation Techniques	Kernel Regulariser	L2 (l = 0.016)
	Activity Regulariser	L1 (l = 0.006)
	Bias Regulariser	L1 (l = 0.006)
	Dropout Rate	0.4 (Seed: 123)

Table 2 (*continued*)

Category	Parameters	Value
Training and Optimisation Parameters	Batch Size	20
	Number of Epochs	15
	Optimiser	Adamax
	Learning Rate (lr)	User-defined in a function parameter
	Loss Function	Categorical Cross-Entropy
	Metrics	Accuracy, Precision, Recall, F1- Score
	Callback: LR_ASK	Dwell: True, Factor: 0.4

Experimental Setup

The model was trained, tested, and evaluated using TensorFlow and Keras in Python on a PC equipped with an Intel Core i7 processor. The research utilised an NVIDIA GeForce RTX 3060 graphics processing unit equipped with 16 gigabytes of random-access memory.

Dataset Information

The 140k Real and Fake Faces Kaggle dataset by Tunguz (2019), which consists of 140,000 images, has been utilised in the study. Of these, 70,000 real expressions were created by NVIDIA's Flickr-Faces-HQ (FFHQ) by Rougetet (2019), and their corresponding fake expressions were created by StyleGAN from 1 million FAKE faces by Tunguz (2020). Various machine and deep learning models can be effectively trained on the look of real human faces using photos from the dataset, as depicted in Figure 6.

This dataset was chosen for the study because it includes a variety of subsets and three well-balanced core components: test, validation, and training sets. This makes it easier to construct and manage deep learning models in an organised manner, which can then be improved with new raw data. Researchers can use the dataset's CSV files, which include related metadata and annotations, for further study.

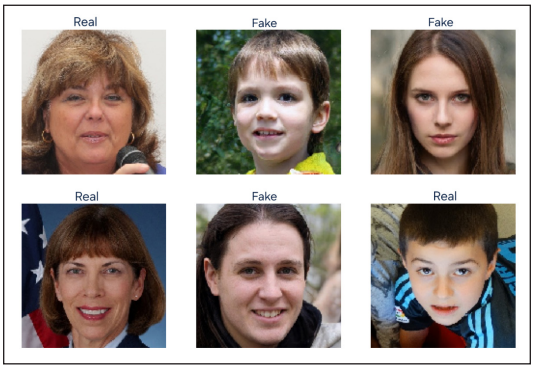


Figure 6. Real and fake image samples from the chosen dataset

In Figure 7, the bar chart illustrates the allocation of images by label in the training dataset, utilising 10,000 images for both 'fake' and 'real' categories, each in equal quantities.

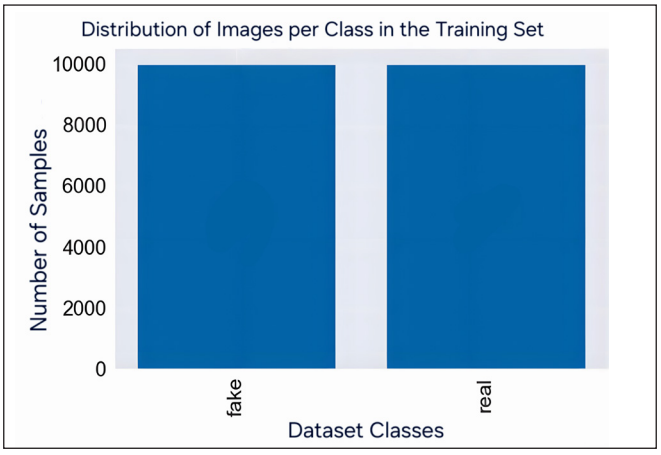


Figure 7. Balanced distribution of images in Training Set

Evaluation Metrics

The study measures model performance across various evaluation metrics, namely accuracy, recall, precision, F1 score, true positive rate (TPR), and false positive rate (FPR), after model selection and tuning.

Accuracy is defined as the number of correctly predicted instances divided by the total number of instances, as mathematically expressed in Equation 1 (Ahmed et al., 2020; Baldi et al., 2000).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad [1]$$

where TP means the number of true positives, TN is the number of true negatives, FP represents the number of false positives, and FN shows the number of false negatives.

Precision is described as the proportion of correct positive predictions to the total number of actual positive and negative cases, as mathematically expressed in Equation 2 (Iheanacho et al., 2021).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad [2]$$

Recall is defined as the proportion of actual positives that are correctly identified, calculated by dividing true positives by the total number of positives Equation 3. High recall is particularly crucial in deepfake detection, as failing to identify a synthetic image can lead to the unchecked spread of misinformation (Martin & Stevens, 2023).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad [3]$$

F1 Score: The F1 score is referred to as the balanced F score due to its properties that show it as the harmonic mean of precision and recall, hence suitable for an imbalanced class Equation 4, Davis & Goadrich, 2006).

$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad [4]$$

RESULTS AND DISCUSSION

Model Loss, Accuracy and F1-score at Various Training Stages

Throughout all epochs, the model showed outstanding training accuracy with minimal loss, as shown in Figure 8(a), 8(b), and 8(c). The model achieved a benchmark validation accuracy persistently exceeding 98.12%. The maximum validation accuracy reached was 98.76% at Epoch 11. Slight variations in validation loss and accuracy seen in later epoch stages indicate that the model was nearing its maximum optimal performance threshold.

Training and Validation Results

Table 3, summarises the training and validation results from Epoch 1 to Epoch 15. These results are based on the training process of the model for the deepfake detection task. For each epoch, metrics such as Training Loss, Training Accuracy, Validation Loss, Validation Accuracy, Validation Loss Improvement, Learning Rate, Next Learning Rate, and Duration in Seconds are given. By the basics of training, the model exhibited very good results in accuracy and loss decrement. Training for the first epoch, the model had a very high training loss of 3.0211 and achieved a training accuracy of 83.65%. Besides, the validation loss was 1.0067, which yielded a validation accuracy of 93.84%. A combined L1 (Lasso) and L2 (Ridge) regularisation strategy (0.006/0.016) has been employed specifically to prevent the model from overfitting to high-frequency noise artefacts common in GAN-generated images.

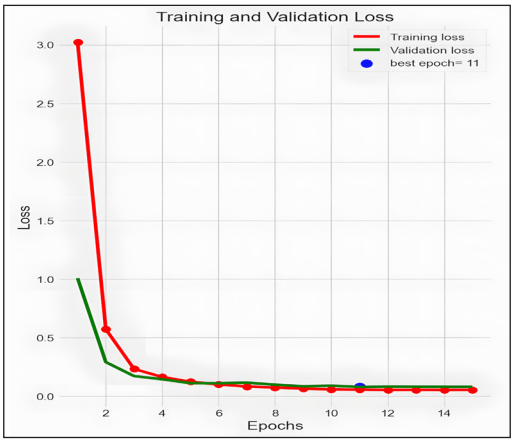


Figure 8(a). Proposed model training and validation Loss over Epochs

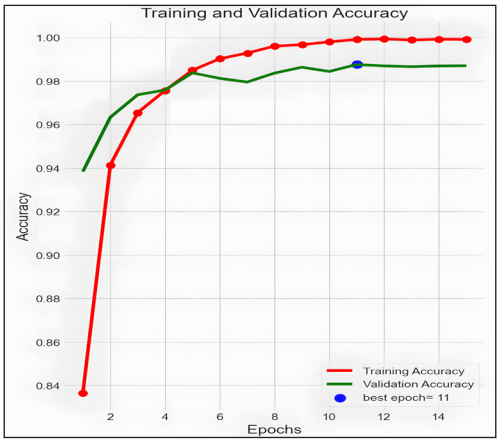


Figure 8(b). Proposed model training and validation Accuracy over Epochs

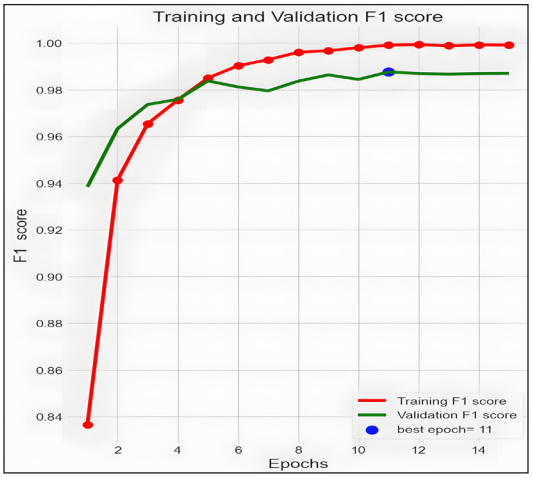


Figure 8(c). Proposed model training and validation F1-score over Epochs

The learning rate was kept at 0.001 for the first six epochs, which led to a gradual improvement of the model's performance. This trend was mainly due to the model's gradual improvement in performance until the 6th epoch, whereby it gained a training accuracy of 99.02% and a validation accuracy of 98.12%. With the start of Epoch 7, the changes in the validation loss became more pronounced and thus the loss curve started fluctuating. This sign indicated that the model was close to the point of convergence. To make the training process more stable, the training rate was lowered to 0.0004. The training accuracy rates of the model were extremely high from Epoch 11 onwards, no less than 99.88%, while only small variations were noticed in validation loss and accuracy values. The validation loss seemed to be stabilising at around 98.66% to 98.76%,

suggesting that the model was not significantly overfitting within the 15 training epochs and was thus effectively generalising.

By employing pre-trained weights from ImageNet and fine-tuning all its layers, the model manages to reveal the subtle features of fake images that are typically missed by less intricate neural networks.

Table 3
Summary of EfficientNetV2-B1 model training and validation results

Epoch No.	Train Loss	Train Accuracy	Valid Loss	Valid Accuracy	V_Loss % Improvement
1	3.0211	83.65	1.0067	93.84	0.0
2	0.5706	94.13	0.2898	96.33	71.21
3	0.2314	96.54	0.1705	97.36	41.17
4	0.1625	97.56	0.1431	97.58	16.08
5	0.1223	98.5	0.1095	98.37	23.5
6	0.0981	99.02	0.1093	98.12	0.14
7	0.0816	99.29	0.1142	97.95	-4.44
8	0.0723	99.61	0.0969	98.37	11.31
9	0.0635	99.67	0.0831	98.63	14.25
10	0.0565	99.8	0.0883	98.44	-6.23
11	**0.0546**	**99.91**	**0.078**	**98.76**	**6.13**
12	0.0526	99.93	0.0804	98.69	-3.07
13	0.0533	99.88	0.0798	98.66	-2.3
14	0.0523	99.92	0.0793	98.69	-1.57
15	0.0529	99.91	0.0792	98.7	-1.45

Note. The text in bold signifies the best results

Confusion Matrix Analysis

The efficiency of the model is thoroughly assessed using a detailed confusion matrix analysis, as shown in Figure 9, which offers insights into its precision and recall for both classes.

The machine successfully identified 9843 images as true and 9929 as false, so it is safe to say that it performed well. 71 counterfeit images were identified as true ones, while 157 authentic images were mislabeled as false. The model represents a high level of accuracy, with the largest number of true positives and true negatives supporting this. Besides, the study found 228 errors in 20000 tests, leading to an accuracy of 98.76 and an F1 score of 98.86 as depicted in Figure 10.

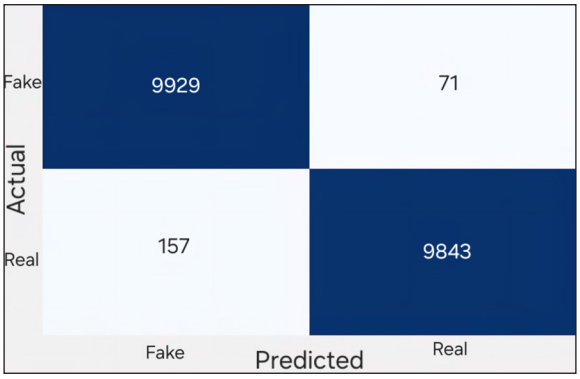


Figure 9. Confusion matrix of the EfficientNetV2-B1 model

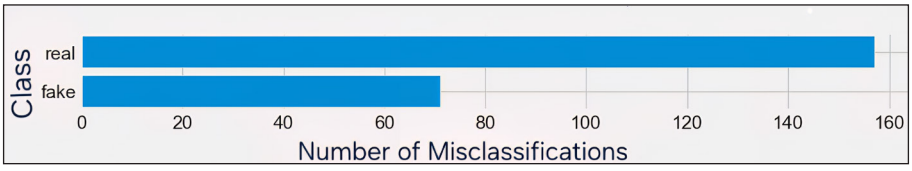


Figure 10. Classification errors occurred on Test Set by Class

Studies of the misclassification of the test dataset revealed that misclassification of real and fake samples is the main source of differences in the error rate of classification. The model tended to mistake genuine images for counterfeit the most, and about 160 were mislabelled, while only around 70 were imposter cases. The low false positives and false negatives rates pinpoint that the model is effective in reducing misclassification errors.

The confusion matrix values were used for determining the evaluation metrics of the classification model - precision, recall, and F1-score. The model was characterised by a precision of 99.28%, which means that the greatest share of the images labelled as real actually belong to the real class. The model obtained a recall of 98.43%, thus it was able to find true-positive examples among all real-world instances. The F1-score, computed as the harmonic mean of precision and recall, was 98.86%, which indicated a fair compromise between precision and recall. The combined use of these metrics gives us the conclusion that the model is a very effective instrument for separating real from artificial, with a minimal rate of misclassification.

The exceptional accuracy (98.76%) and F1-score (98.86%) achieved by the proposed model can be directly attributed to the architectural synergy between the EfficientNetV2-B1 backbone and the customised classification head. The high precision and recall are driven by the backbone's compound scaling, which efficiently captures both low-level textural anomalies and higher-level structural inconsistencies typical of GAN-generated faces. Furthermore, the robust performance is heavily influenced by the aggressive regularisation

strategy implemented. By applying Global Max Pooling in conjunction with combined L1/L2 regularisers and a steep Dropout rate (0.4), the network was explicitly penalised for memorising dataset-specific biases. This forced the model to learn generalised, high-frequency forgery artefacts rather than overfitting to the balanced properties of the 140k dataset, resulting in highly balanced predictive capabilities across both classes.

Comparative Study with Existing Models

Unlike previous research, which often relies on limited feature analysis and model tuning, the research offers a thorough analysis of how certain characteristics make predictions due to the efficient training with customised hyperparameter tuning. This study compares the proposed model's specifications along with the obtained test accuracy to previous research, demonstrating that the custom CNN model achieves higher performance in detecting deepfakes than other models.

The paper compares its approach to three existing deep learning-based methods, as described in Table 4. Sobowale et al. (2024) used VGG19 and ResNet50 models and achieved accuracy rates of 91.59% and 96.61%, respectively. Vaishnavi et al. (2024) achieved a 94% accuracy when utilising DenseNet121. Sharma et al. (2022) achieved test accuracy rates of 95.85%, 93.98%, and 86.63% using their custom CNN, ResNet50, and VGG16 models.

Table 4
Comparison of proposed model with previous deepfake detection models

Author	Model	Dataset	Accuracy (in %)	Parameter Count (M)	FLOPS (G)
(Sobowale et al., 2024)	VGG 19	140K	91.59	25.6	4.1
(Sobowale et al., 2024)	ResNet50	140K	96.61	25.58	4.09
(Vaishnavi et al., 2024)	DenseNet121	140K	94	9.4	3
(Sharma et al., 2022)	VGG 16	140K	86.63	138	15.3
(Sharma et al., 2022)	ResNet50	140k	93.98	23	4.09
(Sharma et al., 2022)	CNN	140k	95.85	11.99	2
(Tan et al., 2019a)	EfficientNetB0	RVF10K	80.30	~5.3	~0.39
(Dosovitskiy, 2020)	ViT-B/16	RVF10K	88.70	~86	~17.5
(Kim, 2025)	Swin Transformer	RVF10K	91.20	~28	~4.5
(Samal et al., 2025)	ResNet50V2	RVF10K	91.95	25.66	~4.1
(Kumar, 2024)	PViT	140K	91.92	~86	~17.5
Proposed Model	**EfficientNetV2-B1**	**140K**	**98.76**	**~8.2**	**<1.0**

Note. The dashes indicate metrics that were not uniformly reported in the original cited studies. Parameter counts and FLOPs for standard external baselines are based on approximate standard architectural values for 224×224 and 150×150 input resolutions

The major benefit of the proposed model, which is based on EfficientNetV2-B1, lies in its enhanced performance, innovative lightweight architectural structure, and detailed training procedure when compared to current deepfake detection methods. In contrast to conventional models such as VGG, ResNet, DenseNet, and basic CNNs used in previous research, EfficientNetV2-B1 offers a scalable and parameter-efficient architecture that enables deeper and accelerated learning. Larger models such as VGG16 and others may theoretically offer advantages in capturing long-range dependencies, but they are more computationally intensive, making them unsuitable for deployment on devices with limited resources. Kumar (2024) utilised a Parallel Vision Transformer (PViT) on the same 140k dataset, achieving 91.92% accuracy. The new approach achieved an accuracy of 98.76%, demonstrating that a well-tuned, lightweight CNN can outperform heavy transformer architectures on such a binary classification task while maintaining a lower computational footprint. The proposed model surpasses the performance of all previously recorded models, and it also displays high precision, recall, and F1-score, making it a highly practical and state-of-the-art solution for identifying deepfakes.

To further contextualise the performance of the proposed EfficientNetV2-B1 approach, an extended cross-dataset architectural comparison was conducted against recent state-of-the-art models, including Vision Transformers (ViTs) and the DeFakeNet architecture proposed by Samal et al. (2025). While direct 1:1 benchmarking requires identical datasets, cross-dataset evaluation provides critical insights into architectural scalability, parameter efficiency, and data dependency. For instance, heavy architectures like ResNet50V2 (25.66M parameters) and ViT-B/16 (~86M parameters) achieved 91.95% and 88.70% test accuracies, respectively, on a highly curated 10K split of the RVF10K dataset (Kunichetty, 2023). Furthermore, an older generation EfficientNetB0 baseline yielded only 80.30% on the same data. In contrast, the proposed EfficientNetV2-B1 model limits the computational burden to approximately 8.2M parameters yet successfully scales to the extensive 140k dataset to achieve 98.76% accuracy. This comparison underscores that the proposed lightweight architecture, paired with a specialised training methodology, possesses superior feature extraction capabilities when deployed across larger, more diverse datasets.

GRAD-CAM Visualisation for Deepfake Detection

This study incorporates Explainable AI (XAI) via Gradient-weighted Class Activation Mapping (Grad-CAM) to improve the interpretability and transparency of classification decisions. This entails resizing the image to a target dimension of 150x150 pixels, scaling pixel values to the range [0, 1], and appending a batch dimension to ensure compatibility with the model's input format.

As shown in Figure 11, a grad_model is generated, accompanied by a side-by-side comparison of the original image (left) and its corresponding Grad-CAM overlay.

This XAI approach facilitates not only the visual verification of the model's key areas during prediction but also enhances the reliability of the deep learning system in real-world applications (Li et al., 2020).

Figure 11. Test image prediction: Original image (Left) - Actual image label - Fake, *Grad-CAM* visualisation (Right) - Prediction - Fake (63.17%)

Ethical Considerations and Societal Impact

As deepfake technology advances, ethical concerns relating to dataset privacy, potential misuse of the technology, and algorithmic bias have become paramount. The dataset by Tunguz (2019) primarily consists of public figures or consented human faces from the web. While commonly used for research, bias in such datasets is observed, which makes the trained model difficult to generalise to unseen manipulated images. Future work must prioritise diverse, consented datasets to ensure equitable detection performance across all racial and gender groups. Evidence indicates the proposed EfficientNet model shows high potential in securing social safety by detecting forgeries, but there is also a dual-use risk where understanding detection boundaries could help adversaries generate better deepfakes. To mitigate this, the research advocates for responsible AI practices, ensuring that such tools are deployed by verified entities (e.g., social platforms, govt. cyber security bodies) rather than being open-sourced without guardrails.

CONCLUSION

In conclusion, the primary contribution of this research is the development of a highly optimised, lightweight deep learning architecture for deepfake image classification, achieved by integrating a customised classification head onto an EfficientNetV2-B1 backbone. The major findings validate the efficacy of this approach, with the model achieving an exceptional test accuracy of 98.76% and a precision of 99.28% on the extensive 140k

benchmark dataset, demonstrating robust discriminative capabilities between real human faces and AI-generated fake images. The critical practical significance of this work lies in its architectural efficiency; utilising approximately 8.2M parameters and requiring fewer than 1.0 GFLOPs, the proposed model drastically reduces computational overhead compared to massive deep residual networks and heavy transformer architectures. This structural efficiency makes the model exceptionally well-suited for real-time deployment in resource-constrained environments, mobile edge computing, and live social media filtering pipelines. While the current methodology establishes a strong baseline on a highly curated dataset, future research trajectories will prioritise extensive cross-dataset validation on uncurated, in-the-wild video frames, such as the Celeb-DF dataset. Furthermore, the integration of lightweight spatiotemporal attention mechanisms will be explored to continuously fortify digital identity authentication against increasingly sophisticated generative manipulations.

ACKNOWLEDGEMENT

This research is self-funded, and no financial support has been given for the study.

REFERENCES

Abu-Ein, A. (2022). Analysis of the current state of deepfake techniques creation and detection methods. *Indonesian Journal of Electrical Engineering and Computer Science, 28*(3), 1659–1668. https://doi.org/10.11591/ijeecs.v28.i3.pp1659-1668

Agrawal, P., Pathak, D., Madaan, V., Verma, P. K., & Choo, W. O. (2026). Spatiotemporal deep learning for real-time video-based deepfake detection using 3DCNN, 3DResNet, TCN, and VAE. *Scientific Reports, 16*, Article 18200. https://doi.org/10.1038/s41598-026-49090-1

Ahmed, J., & Ahmed, M. (2020). Ontological-based approach of integrating big data: Issues and prospects. In A. Kumar, M. Paprzycki, & V. Gunjan (Eds.), *ICDSMLA 2019* (Lecture Notes in Electrical Engineering, Vol. 601, pp. 365-378). Springer. https://doi.org/10.1007/978-981-15-1420-3_38

Awasthi, A., Das, P., Gupta, R., Varma, R., Sharma, S., Gupta, A., & Khan, H. (2023). CNN-based deep learning approach over image data for cyber forensic investigation. In *Handbook of research on thrust technologies' effect on image processing* (pp. 174-192). IGI Global. https://doi.org/10.4018/978-1-6684-8618-4.ch011

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics, 16*(5), 412-424. https://doi.org/10.1093/bioinformatics/16.5.412

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233-240). Association for Computing Machinery. https://doi.org/10.1145/1143844.1143874

Dhar, A., Acharjee, P., Biswas, L., Ahmed, S., & Sultana, A. (2021). *Detecting deepfake images using a deep convolutional neural network* [Undergraduate thesis, BRAC University]. BRAC University Institutional Repository. https://dspace.bracu.ac.bd/xmlui/handle/10361/15933

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv*. https://doi.org/10.48550/arXiv.2010.11929

Fatoni, F., Kurniawan, T. B., Dewi, D. A., Zakaria, M. Z., & Muhayeddin, A. M. M. (2025). Fake vs. real image detection using a deep learning algorithm. *Journal of Applied Data Sciences, 6*(1), 366-376. https://doi.org/10.47738/jads.v6i1.490

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (Vol. 27, pp. 2672-2680).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139-144. https://doi.org/10.1145/3422622

Iheanacho, C. O., Osoba, D. O., & Eze, U. I. (2021). Evaluation of predominant risk factors for type 2 diabetes mellitus among outpatients in two Nigerian secondary health facilities. *African Health Sciences, 21*(2), 693-701. https://doi.org/10.4314/ahs.v21i2.33

Kim, C. (2025). Distinguishing AI-generated and real images using Swin Transformer. In *2025 IEEE Integrated STEM Education Conference (ISEC)* (pp. 1-6). IEEE. https://doi.org/10.1109/ISEC64801.2025.11147328

Kim, E., & Cho, S. (2021). Exposing fake faces through deep neural networks combining content. *IEEE Access, 9*, 123493-123503. https://doi.org/10.1109/ACCESS.2021.3110859

Kosarkar, U., Sarkarkar, G., & Gedam, S. (2023). Revealing and classification of deepfake video images using a customised convolutional neural network model. *Procedia Computer Science, 218*, 2636-2652. https://doi.org/10.1016/j.procs.2023.01.237

Krueger, N., Vanamala, M., & Dave, R. (2023). Recent advancements in the field of deepfake detection. *arXiv*. https://doi.org/10.48550/arXiv.2308.05563

Kumar, B. C. (2024). Deepfake detection using parallel vision transformers. In the *NeurIPS Workshop on Generative AI and Creativity*. https://neurips.cc/virtual/2024/98369

Kumar, L., Yadav, R. K., & Yuvaraj, S. (2023). Detecting fake faces in smart cities' security surveillance using an image. *International Journal of Scientific Methods in Engineering and Management, 1*(4), 39-48.

Kunichetty, S. (2023). *Real vs. fake faces-10k (RVF10K)* [Data set]. Kaggle. https://www.kaggle.com/datasets/sachchitkunichetty/rvf10k

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020). Face X-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5001-5010). IEEE. https://openaccess.thecvf.com/content_CVPR_2020/html/Li_Face_X-Ray_for_More_General_Face_Forgery_Detection_CVPR_2020_paper.html

Liang, B., Wang, Z., Huang, B., Zou, Q., Wang, Q., & Liang, J. (2023). Depth map guided triplet network for deepfake face detection. *Neural Networks, 159*, 34-42. https://doi.org/10.1016/j.neunet.2022.11.027

Martin, D., & Stevens, E. (2023). Statistical validation of machine learning models for diabetes prediction. In *Machine learning in healthcare conference proceedings* (pp. 110-118).

Rougetet, A. (2019). *Flickr-Faces-HQ dataset (FFHQ)* [Data set]. Kaggle. https://www.kaggle.com/datasets/arnaud58/flickrfaceshq-dataset-ffhq

Samal, D., Nagpal, D., Agrawal, P., Madaan, V., & Choo, W. O. (2025). DeFakeNet: A ResNet50V2-based deep learning model for deepfake detection and classification. *Journal of Innovative Image Processing, 7*(4), 1356-1373. https://doi.org/10.36548/jiip.2025.4.015

Sharma, J., Sharma, S., Kumar, V., Hussein, H. S., & Alshazly, H. (2022). Deepfakes classification of faces using convolutional neural networks. *Traitement du Signal, 39*(3), 1027-1037. https://doi.org/10.18280/ts.390330

Sobowale, A. A., Adetona, B. J., Soladoye, A. A., & Omodunbi, B. A. (2024). Deepfake face recognition through modified and improved deep transfer learning. *Uniosun Journal of Engineering and Environmental Sciences, 6*(1). https://doi.org/10.36108/ujees/4202.60.0111

Suratkar, S., & Kazi, F. (2023). Deepfake video detection using transfer learning approach. *Arabian Journal for Science and Engineering, 48*, 9727-9737. https://doi.org/10.1007/s13369-022-07321-3

Tan, M., & Le, Q. V. (2019a). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 6105-6114). PMLR. http://proceedings.mlr.press/v97/tan19a.html

Tan, M., & Le, Q. V. (2019b). MixConv: Mixed depthwise convolutional kernels. In *Proceedings of the British Machine Vision Conference*. https://doi.org/10.48550/arXiv.1907.09595

Tan, M., Pang, R., & Le, Q. V. (2020). EfficientDet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10781-10790). IEEE. https://doi.org/10.1109/CVPR42600.2020.01049

Tan, M., & Le, Q. V. (2021). EfficientNetV2: Smaller models and faster training. *arXiv*. https://doi.org/10.48550/arXiv.2104.00298

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion, 64*, 131-148. https://doi.org/10.1016/j.inffus.2020.06.014

Tunguz, B. (2019). *140k real and fake faces* [Data set]. Kaggle. https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces

Tunguz, B. (2020). *1 million fake faces* [Data set]. Kaggle. https://www.kaggle.com/datasets/tunguz/1-million-fake-faces

Vaishnavi, K. D., Bindu, L. H., Sathvika, M., Lakshmi, K. U., Harini, M., & Ashok, N. (2024). Deep learning approaches for robust deepfake detection. *World Journal of Advanced Research and Reviews, 21*(3), 1629-1636. https://doi.org/10.30574/wjarr.2024.21.3.0889

Wang, Z., Cheng, Z., Xiong, J., Xu, X., Li, T., Veeravalli, B., & Yang, X. (2024). A timely survey on vision transformer for deepfake detection. *arXiv*. https://doi.org/10.48550/arXiv.2405.08463

Wen, L., & Xu, D. (2019). Face image manipulation detection. *IOP Conference Series: Materials Science and Engineering, 533*(1), Article 012054. https://doi.org/10.1088/1757-899X/533/1/012054

Zhang, M., Wang, H., He, P., Malik, A., & Liu, H. (2022). Improving GAN-generated image detection generalisation using unsupervised domain adaptation. In *2022 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICME52920.2022.9859763

Zhou, X., Kan, J., Rosely, N. F. L. M., Duan, X., Cai, J., & Zhou, Z. (2025). ILN-YOLOv8: A lightweight image recognition model for crimped wire connectors. *IEEE Access, 13*, 5193-5202. https://doi.org/10.1109/ACCESS.2025.3525564

Pertanika J. Sci. & Technol. 34 (3): 1727 - 1744 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Occupational Health Implications of High-frequency Whole-body Vibration Exposure in Mine Workers

S. Joel Vinoth Singh[1,2], I. Siva[3]*, JT. Winowlin Jappes[2], and Chithirai Pon Selvan[4]

[1]Architectural, Engineering and Process Team, Barry-Wehmiller Design Group, Prodapt IT Park, Guindy, 600032 Chennai, Tamil Nadu, India
[2]Department of Mechanical Engineering, Kalasalingam Academy of Research and Education, Anand Nagar, 626126 Krishnankoil, Tamil Nadu, India
[3]Department of Mechanical Engineering, Mahendra Engineering College, Namakkal, 637503 Tamil Nadu, India
[4]School of Science and Engineering, Curtin University Dubai, Dubai International Academic City, P.O. Box 345031 Dubai, United Arab Emirates

ABSTRACT

Whole-body vibration (WBV) exposure is a significant occupational hazard in mining and heavy machinery operations, which can lead to long-term musculoskeletal and physiological disorders. This study investigates the magnitude of vibration exposure experienced by mine equipment operators under actual field conditions. Vibration measurements were conducted using a tri-axial accelerometer integrated with a customised data acquisition system. The signals were processed using appropriate filtering and frequency weighting techniques according to ISO 2631-1 standards. Key parameters such as root mean square (RMS) acceleration, daily exposure value A(8), and vibration dose value (VDV) were evaluated and compared with the Health Guidance Caution Zone (HGCZ) limits. The results show that the operators of dumpers and wheel loaders are exposed to vibration levels exceeding recommended threshold limits, particularly along the vertical (Z) axis, thereby increasing the risk of musculoskeletal disorders. The findings highlight the need for effective vibration mitigation strategies, including ergonomic interventions, improved seat suspension systems, and optimised work schedules. This research offers useful guidance on how to evaluate and control whole-body vibration (WBV) exposure in mining settings.

Keywords: ISO 2631-1, mining equipment, occupational health, RMS acceleration, whole-body vibration

ARTICLE INFO

Article history:
Received: 07 February 2026
Accepted: 01 June 2026
Published: 25 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.19

E-mail addresses:
cliffrajjoel@gmail.com (Joel Vinoth Singh)
drisivamech@gmail.com (Siva)
winowlin@klu.ac.in (Winowlin Jappes)
pon.selvan@curtindubai.ac.ae (Chithirai Pon Selvan)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

INTRODUCTION

Whole-body vibration (WBV) is a significant occupational hazard in mining and heavy machinery operations for workers who are exposed to prolonged vibration during equipment operation. Continuous exposure to WBV has been associated with a wide range of health issues like musculoskeletal disorders (MSDs), lower back pain, cardiovascular disturbances, gastrointestinal complications, and neurological impairments (Chaudhary et al., 2022). Out of these issues, lower back disorders are the most frequently reported, especially among dumpers and heavy vehicle operators, due to sustained exposure to vertical vibration components during long working hours (Upadhyay et al., 2022a, 2022b). These health conditions affect worker well-being and lead to reduced productivity with increased occupational health burden in mining industries (Upadhyay et al., 2024).

The vibration frequencies in the range of 4-8 Hz coincide with the natural frequency of the human spine, leading to resonance effects that amplify biomechanical stress. Prolonged exposure to such frequencies can result in significant physiological changes like degeneration of spinal discs, impaired nerve function, and reduced peripheral blood circulation. Vibration-induced reduction in oxygen supply can adversely affect vital organs such as the brain, heart, kidneys, and muscles, leading to long-term health complications (Zhukova et al., 2022). These findings highlight the importance of understanding vibration exposure from a mechanical and biomedical view (Singh et al., 2018; Singh et al., 2019).

Many field-based and analytical studies have investigated WBV exposure in mining environments. For instance, case-control studies conducted on dumper operators have demonstrated a higher prevalence of musculoskeletal disorders compared to non-exposed groups, emphasising the direct relationship between WBV exposure and occupational health risks (Kumar et al., 2022). Similarly, large-scale investigations in Indian mining sectors have utilised questionnaire-based approaches such as the Nordic Musculoskeletal Questionnaire (NMQ), along with vibration measurements, to record discomfort and health impacts among operators (Mandal et al., 2013; Upadhyay et al., 2022a). These studies underline the necessity for regular monitoring, early diagnosis, and preventive strategies to mitigate the adverse effects of WBV exposure.

Advanced analytical techniques have been employed to enhance the understanding of WBV-related risks. Bayesian network models have been used to identify critical risk factors and predict the chance of developing musculoskeletal disorders among dumper operators (Atal et al., 2022). Experimental studies have focused on evaluating the transmissibility characteristics of vehicle seats using Seat Effective Amplitude Transmissibility (SEAT) analysis, which plays a key role in reducing vibration transmission to the human body (Sharma & Mandal, 2021). These investigations demonstrate that machine design and operational conditions significantly influence vibration exposure levels.

Studies in the agricultural and transport sectors provide valuable insights into vibration behaviour under various operating conditions. Experimental investigations using techniques such as Fast Fourier Transform (FFT) analysis have identified dominant vibration frequencies typically in the range of 0.8-4 Hz, which overlap with the natural frequencies of different parts of the human body (Singh et al., 2022). Parameters such as vehicle speed, load, terrain condition, and operator posture have been found to significantly affect vibration magnitude and exposure duration. These findings point out the importance of both time-domain and frequency-domain analysis in assessing WBV exposure (Harsha et al., 2014; Hildebrand et al., 2008; Liang & Chiang, 2008; Tiemessen et al., 2008).

Occupational safety initiatives have focused on incorporating Industry 4.0 technologies and machine learning for effective management of Whole-Body Vibration (WBV). For instance, Sakinala et al. (2024) successfully utilised machine learning techniques to predict WBV risk for heavy equipment operators in underground mining, while Vikram et al. (2024) demonstrated how varying operational parameters and rock types significantly influence exposure levels. The contemporary frameworks now leverage digital monitoring and decision-support systems to transition from reactive compliance to proactive ergonomic risk mitigation (Misita et al., 2024; Ullah et al., 2025). Cross-sectional evidence also confirms a high correlation between prolonged exposure to heavy machinery and chronic musculoskeletal disorders, which calls for these updated monitoring strategies (Mpendu et al., 2026).

Despite having extensive literature, most studies are conducted either under controlled laboratory environments or involve large datasets with complex modelling approaches (Basri & Griffin, 2013; Beard & Griffin, 2016; Bovenzi, 2015). There is a lack of field-based investigations that capture real-time vibration exposure under actual mining conditions using compact and customised data acquisition systems. Limited attention has been given to directly correlating measured vibration parameters with ISO 2631-1 Health Guidance Caution Zone (HGCZ) limits for practical and immediate risk assessment. This gap is particularly relevant in developing regions, where site-specific conditions and operational practices may significantly influence vibration exposure levels (Bandyopadhyay et al., 2012; Charles et al., 2018; Combs et al., 2018; Du et al., 2018).

Whole-body vibration (WBV) research (Nyarubeli et al. 2026) has focused on improving measurement accuracy and real-time exposure assessment using wearable sensor technologies and smart data acquisition systems. Studies published after 2022 have highlighted the use of wireless sensing devices and integrated monitoring frameworks for capturing field-level vibration data with higher reliability. Data-driven approaches, including machine learning-based risk classification and predictive modelling, are being used to assess occupational exposure and associated health risks (Darabont et al. 2025). Studies also emphasise the importance of combining experimental measurements with

ergonomic assessment to develop mitigation strategies in mining and heavy machinery environments (Kumar et al., 2001; Sorainen et al., 1998). These developments indicate a more precise, real-time, and application-oriented WBV assessment methodology, particularly under practical field conditions.

The present study aims to experimentally evaluate whole-body vibration exposure among mine equipment operators under real working conditions. The specific objectives of this study are to (i) measure vibration levels using a tri-axial accelerometer integrated with a customized data acquisition system, (ii) analyse vibration exposure in terms of root mean square (RMS) acceleration, daily exposure value A(8), and vibration dose value (VDV), and (iii) assess the associated health risks by comparing the measured values with ISO 2631-1 standards. The study provides practical insights into vibration exposure characteristics across different types of mining equipment and highlights the need for mitigation strategies, ergonomic improvements and exposure management practices.

EXPERIMENTAL METHODOLOGY

The experimental methodology adopted to quantify whole-body vibration (WBV) exposure in mine equipment operators under real working conditions. The methodology covers the overall measurement plan (selection of operators and machines, sensor locations, and measurement duration), the instrumentation and calibration procedures, and the data acquisition settings used to capture tri-axial acceleration signals. The recorded vibration signals were then processed using ISO-recommended frequency-weighting and filtering to obtain exposure metrics such as RMS acceleration, daily exposure value A(8), vibration dose value (VDV), and crest factor. These parameters were compared with relevant ISO guideline limits to enable a consistent assessment of occupational health risk across different equipment categories.

Measurement Strategy

Whole-body vibration (WBV) measurements were conducted under actual mining operating conditions to evaluate the vibration exposure experienced by equipment operators. Physical measurements were obtained from three critical locations: two seating interfaces (seat–operator contact on the back and buttocks) and the arm positions. These locations were selected to capture both whole-body and hand-arm vibration components during machine operation.

Instrumentation and Data Acquisition

The vibration measurements were carried out using a tri-axial accelerometer integrated with a human vibration measurement system (Larson Davis HVM200). The device

is compliant with ISO 8041 standards and can measure whole-body and hand-arm vibration in accordance with ISO 2631 and ISO 5349 guidelines. The accelerometer has a measurement range suitable for low-frequency vibration (typically 0.5–80 Hz) and a sensitivity appropriate for human vibration analysis. The system supports three-channel simultaneous data acquisition along orthogonal axes (X, Y, and Z). The data acquisition was performed, ensuring adequate resolution for capturing both low-frequency WBV and transient vibration peaks. The acquired signals were stored and processed using the manufacturer-supported software interface.

Calibration Procedure

Before data collection, the vibration measurement system was calibrated using standard calibration procedures to ensure measurement accuracy and reliability. The calibration was performed in accordance with manufacturer guidelines, and sensor alignment was verified to match the human body coordinate system specified in ISO 2631-1. This ensures that the measured vibration components accurately represent the directional exposure experienced by the operator.

Data Collection Procedure

Vibration data were recorded under normal working conditions of mining equipment. The measurements were conducted continuously for a specific duration while the operators performed their routine tasks. Appropriate care was taken to ensure that sensor placement did not interfere with the operator's activity or machine operation. Each measurement was conducted for a duration of 20 minutes, and the data were recorded continuously with an averaging interval of 5 seconds to capture variations in vibration levels during operation. The recorded signals represent acceleration as a function of time for each axis. These time-domain signals were further processed to evaluate vibration exposure parameters.

Signal Processing and Analysis

The acquired vibration signals were processed using standard procedures recommended in ISO 2631-1. The raw acceleration signals were filtered and frequency-weighted to account for human sensitivity to different vibration frequencies. The processed signals were then used to compute key vibration exposure parameters. A band-limited frequency weighting filter was applied to the raw acceleration signals to account for human sensitivity to vibration across different frequency ranges. The processed signals were used to compute RMS acceleration, A(8) and VDV values.

Evaluation Parameters

Root Mean Square (RMS) Acceleration

The values of acceleration as a function of time are obtained from the vibration measurements conducted in the field. Subsequently, the computations are executed utilising the subsequent formulas, Equation 1. RMS value of the acceleration can be calculated by:

$$rms_{ws} = [\frac{1}{T}\int_0^T a_w{}^2\,(t)dt]^{1/2} \quad [1]$$

Daily Exposure Value A(8)

In general, A (8) Equation 2 is the daily exposure value for 8 hours of work. The Daily Exposure Action Value (EAV) is a measure used to evaluate the magnitude of whole-body vibration (WBV) exposure that a person undergoes throughout a typical workday, usually lasting 8 hours. It acts as a limit that determines when special measures should be implemented to reduce the impact of vibration. EAV is quantified using the root mean square (RMS) acceleration, often denoted in units of metres per second squared (m/s²).

$$A(8)_S = k_S\, rms_{ws}\sqrt{\frac{T_{exp}}{T_0}} \quad [2]$$

where 's' represents the axis along which the measurement is carried out, rms_{ws}-weighted RMS acceleration, T - signal duration, a_w - weighted acceleration, $A(8)_S$ - Daily exposure value, k_S - weighted factor, T_{exp} = Time of exposure, T_0 = 8 hours. And can be represented by another relationship as mentioned in Equation 3,

$$A(8)_S = a_w\sqrt{\frac{T_{exp}}{T_{ref}}} \quad [3]$$

Where, a_w is the RMS acceleration value (in m/s²) of the vibration. T_{exp} is the total exposure duration during the workday. T_{ref} is the reference duration (8 hours or 28,800 seconds).

Vibration Dose Value (VDV)

The vibration dose value can be calculated as per Equation 4.

$$VDV_{ws} = [\int_0^T a_w{}^4(t)dt]^{1/4} \quad [4]$$

where 's' represents the axis along which the measurement is carried out, $VDVws$ – weighted vibrational dose value, T – signal duration, a_w – weighted acceleration.

$$VDV_s = k_s \, VDV_{ws} \sqrt{\frac{T_{exp}}{T_{meas}}} \quad [5]$$

(or)

$$VDV_s = k_s \, rms_{ws} \, T^{0.25} \quad [6]$$

where 's' represents the axis along which the measurement is carried out, $rmsws$ – weighted RMS acceleration, T – signal duration, VDV_s - Vibration dose value as per Equations 5 and 6, k_S - weighted factor, T_{exp} - daily duration of vibration exposure, T_{meas} - duration of measurement taken.

Crest Factor

Crest factor can be calculated as per Equation 7,

$$\text{Crest factor} = \frac{\text{Peak value}}{\text{RMS value}} \quad [7]$$

If the crest factor > 9, then the VDV value must also be considered along with the A(8) to conclude vibration exposure.

Assessment Based on ISO Standards

The calculated vibration parameters were evaluated in accordance with ISO 2631-1 guidelines. The obtained values of A(8) and VDV were compared with the Health Guidance Caution Zone (HGCZ) limits to determine the level of risk associated with vibration exposure.

Equipment Configuration and Sensors Placement

Whole-body vibration (WBV) measurements were carried out using a seat pad sensor integrated with a tri-axial accelerometer. The sensor was mounted on a rubber adapter and placed at the seat–operator interface to capture vibration transmitted to the human body during machine operation. The operator was seated directly on the sensor pad to ensure accurate transmission of vibration signals.

In addition to WBV measurements, hand-arm vibration was recorded using accelerometers mounted on specialised adapters positioned on the operator's hand interface

(handle or grip surface). Care was taken to ensure that sensor placement did not interfere with normal machine operation while maintaining firm contact for reliable data acquisition.

The sensors were aligned according to the standard human body coordinate system, where the vertical direction (Z-axis) corresponds to the spinal axis of the operator. Proper alignment ensures that vibration components are accurately captured in all three orthogonal directions (X, Y, and Z), as illustrated in Figure 1.

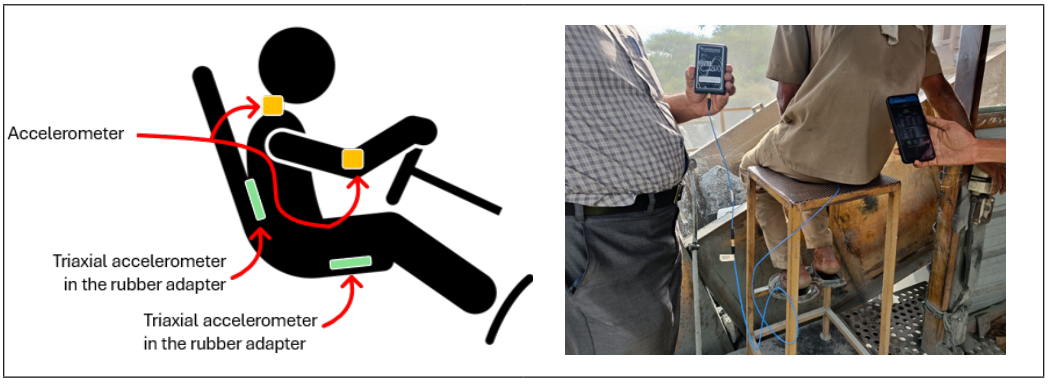


Figure 1. Freebody diagram of the measuring system and placement of sensors, (b) WBV measurement setup for the crusher operator

Instrument Description

The vibration measurements were performed using the Larson Davis HVM200 Human Vibration Meter, a portable and robust device designed for evaluating human exposure to vibration. The instrument is a three-channel system compliant with ISO 8041:2005 standards and supports measurements in accordance with ISO 2631 (whole-body vibration) and ISO 5349 (hand-arm vibration).

The device is equipped with integrated Wi-Fi capability for real-time data transmission and monitoring. It incorporates built-in frequency weighting filters and signal processing capabilities required for accurate assessment of human vibration exposure. The system is also compatible with standard occupational health guidelines, including the American Conference of Governmental Industrial Hygienists (ACGIH) Threshold Limit Values (TLVs) and European Directive 2002/44/EC.

Applicable Standards

The evaluation of vibration exposure in this study is based on the guidelines provided by ISO 2631-1, which defines methods for quantifying and assessing whole-body vibration in occupational environments. The standard considers key parameters such as vibration magnitude, frequency, and exposure duration to establish acceptable exposure limits.

According to ISO 2631-1, the human body is particularly sensitive to vertical vibration in the frequency range of 4-8 Hz, which coincides with the natural frequency of the spine. Prolonged exposure within this range can lead to resonance effects and increased health risks.

The ISO standards also provide frequency weighting factors for different axes of vibration, enabling a more accurate assessment of human response to vibration. These weighting factors are used in the calculation of RMS acceleration, A(8), and VDV values, as presented in Table 1.

Human Body Coordinate System

The measurement of vibration exposure requires proper alignment of sensors according to standardised coordinate systems. For whole-body vibration, the ISO 2631 coordinate system is used, where the X-axis represents the fore-and-aft direction, the Y-axis represents the lateral direction, and the Z-axis represents the vertical direction aligned with the spine.

Sensors for WBV measurements are typically placed at the seat interface for seated operators, while additional placements such as the backrest or foot platform may be used depending on the application. For hand-arm vibration measurements, ISO 5349 guidelines are followed, where sensors are positioned on the hand, wrist, or tool interface to capture transmitted vibration.

Proper sensor placement is essential to ensure consistency, accuracy, and comparability of vibration measurements across different studies. The coordinate system used in this study is illustrated in Figure 2.

Standard Operating Procedure for Vibration Measurement

Along with accurate measurement of vibration magnitude, precise estimation of exposure duration is essential for reliable assessment of human vibration exposure. In this study, a structured standard operating procedure (SOP) was adopted to ensure consistency in data acquisition and analysis.

Measurement Approaches

Two standard approaches are generally followed for evaluating vibration exposure:

Task-Based Measurement Approach

Vibration measurements are recorded only during periods when the operator is actively exposed to vibration. Each measurement corresponds to a specific machine operation or working condition. The total exposure duration is calculated by summing only the active vibration periods. This approach provides an accurate characterisation of vibration levels for individual tasks; however, it requires detailed observation and time tracking of work activities.

Table 1
ISO standard guidance on values of the Caution Zone for a human

Whole Body Vibration (WBV)	**ISO 2631 – 1**	
	Daily Exposure Value (m/s^2)	**Vibration Dose Value ($m/s^{1.75}$)**
Exposure Action Value	0.43	8.5
Exposure Limit Value	0.86	17
Hand-arm vibration (HAV)	**ISO 5349 – 1**	
Exposure Action Value	2.5	0.4
Exposure Limit Value	5	0.8

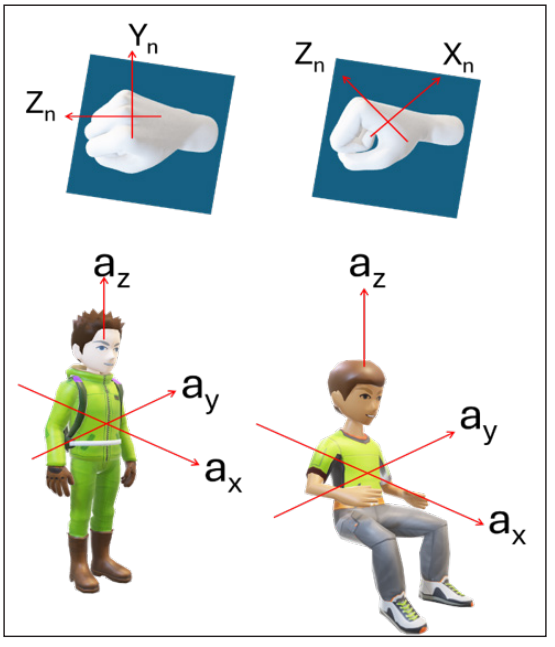


Figure 2. Coordinate system of the hand and whole body according to ISO standards for the placement of the sensor

Full-shift Measurement Approach

A continuous measurement is carried out over an extended duration covering multiple operating conditions, including machine operation, idle periods, and breaks. The resulting data represents an average exposure over the entire work cycle. While this method simplifies data collection, it may not capture variations associated with specific tasks or operating conditions.

In the present study, an appropriate combination of these approaches was adopted to ensure both accuracy and practical feasibility of measurements.

Measurement and Signal Processing Procedure

The vibration magnitude was primarily quantified in terms of acceleration, as it is the most widely accepted parameter for evaluating human vibration exposure. The accelerometer provides a signal proportional to acceleration, which is further processed to obtain meaningful exposure parameters.

The acquired signals were subjected to digital filtering to remove noise and isolate the frequency range relevant to human vibration exposure. Frequency weighting was applied separately for each axis (X, Y, and Z) to account for the directional sensitivity of the human body to vibration, as specified in ISO standards.

The processed signals were then used to compute vibration parameters such as RMS acceleration, A(8), and VDV, which provide measures of continuous and transient vibration exposure.

Data Acquisition and Software Tools

Data acquisition and processing were carried out using the Larson Davis HVM200 system along with its dedicated software tools. The G4 LD Utility software was used for post-processing and analysis of recorded vibration data in a desktop environment. The LD Atlas mobile application was used for real-time monitoring and data acquisition during field measurements.

These tools enabled efficient data collection, storage, and analysis, ensuring accuracy and reliability of the vibration exposure assessment, which supports the research workflow illustrated in Figure 3.

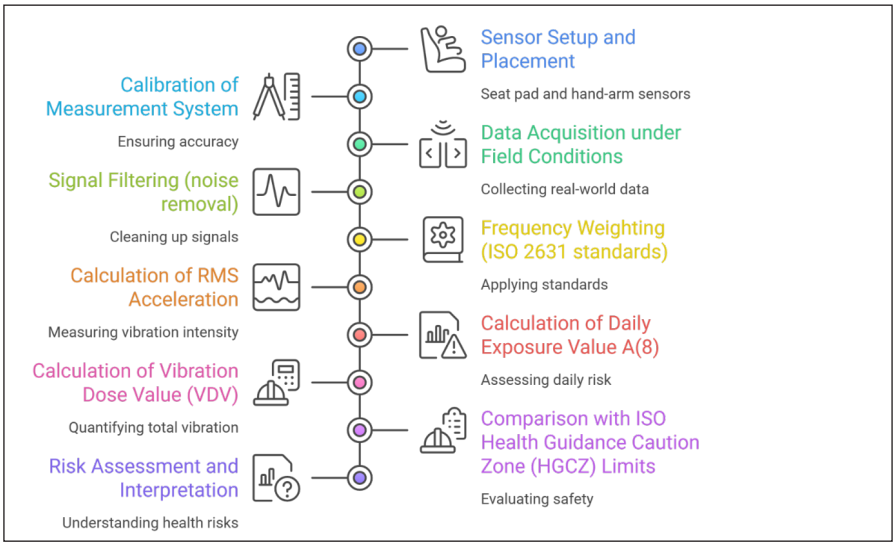


Figure 3. Workflow

RESULTS AND DISCUSSION

Measurement Overview and Data Organisation

Whole-body vibration (WBV) measurements were carried out on four operators handling different categories of mining equipment, namely crushers, dumpers, and wheel loaders. The data acquisition system was configured with a storage interval of 1 second and an averaging interval of 5 seconds to capture both steady-state and transient vibration behaviour. Each measurement was carried out for a duration of 20 minutes under normal working conditions to ensure consistency across all observations.

A structured questionnaire was used to collect operator-specific details, including age, height, weight, body mass index (BMI), years of experience, and daily working hours. These parameters were considered important in understanding the influence of individual characteristics on vibration exposure and associated health risks. The collected data were systematically organised and presented in tabular form, categorised based on the type of equipment operated and the operator profile. Table 2 presents the details of the operators, while Table 3 summarises the measured vibration parameters used for further analysis.

The experiment involved four operators: Operator 1 and Operator 2 (crusher operators), Operator 3 (dumper operator), and Operator 4 (wheel loader operator). This categorisation enabled comparison of vibration exposure across different types of mining equipment and operational conditions.

Owing to the exploratory nature of the study and the limited number of operators available under actual field conditions, the analysis was primarily descriptive. Comparative evaluation was performed using RMS acceleration, A(8), VDV, and crest factor values across equipment categories. The observed differences were interpreted with reference to ISO 2631-1 Health Guidance Caution Zone limits.

Vibration Magnitude Analysis

The measured vibration data across the three orthogonal axes (X, Y, and Z) provide important insights into the exposure levels experienced by different operators. It is observed that the vertical (Z-axis) component of vibration is dominant in most cases, which aligns with the typical transmission of vibration through the seat operator interface during machine operation.

Operators 1 and 2, associated with crusher operations, exhibit relatively low vibration levels across all axes. The root mean square (RMS) acceleration values are below 0.1 m/s^2, and the corresponding A(8) values remain within safe limits. Similarly, the vibration dose values (VDV) for these operators are below 0.5 m/s$^{1.75}$, indicating minimal exposure and low health risk under current operating conditions. These findings can be attributed to the relatively stationary nature of crusher operations and shorter exposure durations.

Table 2
Physical and technical information about the studied human

Personal Factor	Operator 1	Operator 2	Operator 3	Operator 4
Machine	Crusher	Crusher	Dumper	Wheel loader
Age (year)	53	55	31	29
Weight (kg)	55	65	60	75
Height (m)	1.8	1.65	1.7	1.65
Experience (year)	25	32	8	12
Body Mass Index (kg/m^2)	16.97	23.87	20.76	27.55
Working hours/day (hour)	4	4	6	8

Table 3
Various data were measured for different operators in Mines

Indicators	X- axis				Y- axis				Z- axis			
Operator	1	2	3	4	1	2	3	4	1	2	3	4
a_{rms} (m/s^2)	0.0623	0.0406	0.7092	0.1241	0.0806	0.0615	1.0035	0.1461	0.7884	0.6642	1.5146	1.2773
a_{peak} (m/s^2)	0.4445	0.2536	3.6708	0.7437	0.5829	0.3697	5.7151	0.8960	6.7202	5.7254	8.5772	8.3764
Crest factor	7.1347	6.2473	5.1760	5.9928	7.2325	6.0123	5.6952	6.1326	8.5239	8.6200	5.6630	6.6057
A(8) (m/s^2)	0.0617	0.0402	0.8599	0.1737	0.0798	0.0609	1.2167	0.2045	0.5575	0.4697	1.3317	1.2773
VDV(m/s$^{1.75}$)	0.2593	0.1689	5.7288	1.5653	0.3271	0.2496	8.1371	1.8380	2.2376	1.8852	8.8005	11.211

Operators 3 and 4, associated with dumper and wheel loader operations, exhibit significantly higher vibration exposure. Operator 3 recorded the highest RMS acceleration along the Z-axis (1.5146 m/s^2), with an A(8) value of 1.3317 m/s^2 and a VDV of 8.8005 m/s$^{1.75}$. Similarly, Operator 4 exhibited an A(8) value of 1.2773 m/s^2 and a VDV of 11.2108 m/s$^{1.75}$, indicating severe vibration exposure. These elevated values are primarily due to continuous machine movement, uneven terrain, and longer operating durations.

Comparison with ISO Standards

The measured vibration parameters were evaluated based on the ISO 2631-1 Health Guidance Caution Zone (HGCZ). As summarised in Table 1, the whole-body vibration exposure action value is 0.43 m/s^2 for an 8-hour working period; exceeding this threshold necessitates the implementation of control measures to reduce vibration exposure.

The results indicate that crusher operators (Operators 1 and 2) operate well below the EAV, suggesting a low risk of immediate health effects. However, dumper and wheel loader operators (Operators 3 and 4) significantly exceed this threshold, placing them within or above the upper limit of the HGCZ. This indicates a high probability of developing vibration-related health disorders if exposure continues over extended periods.

Crest Factor and Transient Effects

The crest factor values for all operators were found to be greater than 6, indicating the presence of transient shocks or peak accelerations during machine operation. Although the crest factor values are below the critical limit of 9, the repeated occurrence of such transient loads can contribute to cumulative stress on the musculoskeletal system.

In such scenarios, reliance solely on RMS-based metrics may underestimate the severity of exposure. Therefore, the inclusion of VDV provides a more comprehensive evaluation of vibration exposure, particularly for conditions involving intermittent shocks and uneven operating surfaces.

Health Risk Assessment

The quantitative analysis of vibration data highlights significant differences in occupational exposure and associated health risks among the operators. Operators 3 and 4 are subjected to high levels of vibration, particularly along the vertical axis, which is closely associated with spinal loading. Prolonged exposure to A(8) values exceeding 1.2 m/s^2 can lead to serious health issues such as lower back pain, spinal disc degeneration, and circulatory disturbances in the lumbar and pelvic regions.

The elevated VDV values further indicate cumulative exposure to vibration energy, increasing the likelihood of soft tissue damage and long-term musculoskeletal disorders.

Factors such as operator posture, machine design, terrain conditions, and individual characteristics such as BMI and age may amplify these effects.

Although the vibration levels for Operators 1 and 2 are within acceptable limits, long-term exposure, even at moderate levels, may contribute to cumulative health risks. Therefore, continuous monitoring and preventive measures are recommended for all operators regardless of current exposure levels.

Implications and Mitigation Strategies

The findings of this study emphasise the need for effective vibration mitigation strategies in mining environments. Engineering controls, such as improved seat suspension systems, vibration isolation mechanisms, and optimised vehicle design, can significantly reduce vibration transmission to the operator.

Administrative measures, including job rotation, reduced exposure duration, and scheduled rest periods, can further minimise the risk of vibration-related health issues. Periodic medical examinations and ergonomic training programs are essential for early detection and prevention of musculoskeletal disorders.

The integration of advanced monitoring systems, wearable health monitoring technologies, including real-time vibration tracking and data-driven exposure analysis, can enhance occupational safety and support informed decision-making. Without timely intervention, prolonged exposure to high vibration levels may lead to chronic health conditions, reduced productivity, and increased occupational health burden.

CONCLUSION

This study investigated the significance of whole-body vibration (WBV) exposure on the health of mine workers operating different categories of heavy machinery. Experimental measurements conducted under real field conditions provided valuable insights into the variation of vibration exposure across different types of equipment and operators. The analysis revealed clear differences in vibration magnitude and associated risk levels depending on both operational conditions and individual characteristics.

The results indicate that crusher operators (Operators 1 and 2) experience vibration levels within the acceptable limits specified by ISO 2631-1, suggesting low immediate health risks. However, prolonged exposure over several years may still contribute to cumulative musculoskeletal effects, highlighting the need for continuous monitoring and preventive health assessment. In contrast, dumper and wheel loader operators (Operators 3 and 4) were exposed to significantly higher vibration levels, particularly along the vertical (Z) axis. The recorded A(8) and VDV values exceed the Health Guidance Caution Zone (HGCZ), indicating a high probability of developing vibration-induced disorders such as lower back pain, spinal degeneration, and associated physiological complications.

These findings emphasise the urgent need for implementing vibration mitigation strategies in mining environments. Engineering interventions such as improved seat suspension systems, vibration isolation mechanisms, and optimised vehicle design can significantly reduce vibration transmission. Administrative measures, including reduced exposure duration, job rotation, and structured work-rest cycles, are essential to minimise occupational risk. Periodic medical examinations and ergonomic awareness programs should also be incorporated to identify early symptoms and prevent long-term health deterioration.

The integration of advanced technologies such as real-time vibration monitoring, artificial intelligence-based exposure tracking, and predictive analytics can enhance early risk detection and support data-driven decision-making for occupational safety. Overall, this study highlights that without appropriate preventive measures, operators exposed to high-magnitude vibrations are at a significant risk of developing chronic musculoskeletal and neurological disorders. Therefore, long-term vibration control strategies and health-focused interventions are critical for ensuring sustainable and safe mining operations.

A limitation of the present study is the relatively small number of operators and equipment categories investigated. Future studies involving larger operator populations and additional mining equipment are recommended to improve the generalisability of the findings.

ACKNOWLEDGEMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors would like to express their sincere gratitude to the Centre for Composite Materials, Kalasalingam Academy of Research and Education, for providing the facilities and support necessary to carry out this research.

REFERENCES

Atal, M. K., Palei, S. K., Chaudhary, D. K., Kumar, V., & Karmakar, N. C. (2022). Occupational exposure of dumper operators to whole-body vibration in opencast coal mines: An approach for risk assessment using a Bayesian network. *International Journal of Occupational Safety and Ergonomics, 28*(2), 758-765. https://doi.org/10.1080/10803548.2020.1828551

Bandyopadhyay, A., Dev, S., & Gangopadhyay, S. (2012). A study on the prevalence of musculoskeletal disorders among the coalminers of Eastern Coalfields of India. *International Journal of Occupational Safety and Health, 2*(2), 34-37. https://doi.org/10.3126/ijosh.v2i2.6596

Basri, B., & Griffin, M. J. (2013). Predicting discomfort from whole-body vertical vibration when sitting with an inclined backrest. *Applied Ergonomics, 44*(3), 423-434. https://doi.org/10.1016/j.apergo.2012.10.006

Beard, G. F., & Griffin, M. J. (2016). Discomfort of seated persons exposed to low-frequency lateral and roll oscillation: Effect of backrest height. *Applied Ergonomics, 54*, 51-61. https://doi.org/10.1016/j.apergo.2015.11.010

Bovenzi, M. (2015). A prospective cohort study of neck and shoulder pain in professional drivers. *Ergonomics, 58*(7), 1103-1116. https://doi.org/10.1080/00140139.2014.935487

Charles, L. E., Ma, C. C., Burchfiel, C. M., & Dong, R. G. (2018). Vibration and ergonomic exposures associated with musculoskeletal disorders of the shoulder and neck. *Safety and Health at Work, 9*(2), 125-132. https://doi.org/10.1016/j.shaw.2017.10.003

Chaudhary, D. K., Palei, S. K., Kumar, V., & Karmakar, N. C. (2022). Whole-body vibration exposure of heavy earthmoving machinery operators in surface coal mines: A comparative assessment of transport and non-transport earthmoving equipment operators. *International Journal of Occupational Safety and Ergonomics, 28*(1), 174-183. https://doi.org/10.1080/10803548.2020.1785154

Combs, B., Heaton, K., Raju, D., Vance, D. E., & Sieber, W. K. (2018). A descriptive study of musculoskeletal injuries in long-haul truck drivers: A NIOSH national survey. *Workplace Health & Safety, 66*(10), 475-481. https://doi.org/10.1177/2165079917750935

Darabont, D. C., Cioca, L. I., & Badea, D. O. (2025). Digital approaches for mitigating occupational vibration exposure in the context of Industry 4.0. *Safety, 11*(2), Article 31. https://doi.org/10.3390/safety11020031

Du, B. B., Bigelow, P. L., Wells, R. P., Davies, H. W., Hall, P., & Johnson, P. W. (2018). The impact of different seats and whole-body vibration exposures on truck driver vigilance and discomfort. *Ergonomics, 61*(4), 528-537. https://doi.org/10.1080/00140139.2017.1372638

Harsha, S. P., Desta, M., Prashanth, A. S., & Saran, V. H. (2014). Measurement and biodynamic model development of seated human subjects exposed to a low-frequency vibration environment. *International Journal of Vehicle Noise and Vibration, 10*(1-2), 1-24. https://doi.org/10.1504/IJVNV.2014.059627

Hildebrand, R., Keskinen, E., & Navarrete, J. A. R. (2008). Vehicle vibrating on a soft compacting soil half-space: Ground vibrations, terrain damage, and vehicle vibrations. *Journal of Terramechanics, 45*(4), 121-136. https://doi.org/10.1016/j.jterra.2008.09.003

Kumar, A., Mahajan, P., Mohan, D., & Varghese, M. (2001). Tractor vibration severity and driver health: A study from rural India. *Journal of Agricultural Engineering Research, 80*(4), 313-328. https://doi.org/10.1006/jaer.2001.0755

Kumar, V., Palei, S. K., Karmakar, N. C., & Chaudhary, D. K. (2022). Whole-body vibration exposure vis-à-vis musculoskeletal health risk of dumper operators compared with a control group in coal mines. *Safety and Health at Work, 13*(1), 73-77. https://doi.org/10.1016/j.shaw.2021.10.007

Liang, C. C., & Chiang, C. F. (2008). Modelling of a seated human body exposed to vertical vibrations in various automotive postures. *Industrial Health, 46*(2), 125-137. https://doi.org/10.2486/indhealth.46.125

Mandal, B. B., Pal, A. K., & Sishodiya, P. K. (2013). Vibration characteristics of mining equipment used in Indian mines and their vibration hazard potential. *International Journal of Environmental Health Engineering, 2*(1), Article 45. https://doi.org/10.4103/2277-9183.122440

Misita, M., Spasojević Brkić, V., Đurić, R., & Perišić, M. (2024). Decision support system for mining machinery risk mitigation driven by ergonomics and contextual theory. *Applied Sciences, 14*(15), Article 6413. https://doi.org/10.3390/app14156413

Nyarubeli, I., Sakwari, G., Chombo, S., Reuben, S., Msangi, N., Duguza, R., Francis, B., Lwaho, S., Omar, A., & Mduma, J. (2026). Whole-body vibration exposure and chronic low back pain among heavy machine operators and long-haul truck drivers in Tanzania: A cross-sectional study. *Frontiers in Public Health, 14*, Article 1735367. https://doi.org/10.3389/fpubh.2026.1735367

Sakinala, V., Paul, P. S., & Moparthi, J. R. (2024). Assessment of HEMM operators' risk exposure due to whole-body vibration in underground metalliferous mines using machine learning techniques. *Mining, Metallurgy & Exploration, 41*(4), 1515-1530. https://doi.org/10.1007/s42461-024-01009-y

Sharma, A., & Mandal, B. B. (2021). Attenuation of mechanical vibration during transmission to the human body through mining vehicle seats. *Mining, Metallurgy & Exploration, 38*(3), 1449-1461. https://doi.org/10.1007/s42461-021-00406-x

Singh, A., Samuel, S., Singh, H., Singh, J., Prakash, C., & Dhabi, Y. K. (2022). Whole-body vibration exposure among the tractor operator during soil tillage operation: An evaluation using ISO 2631-5 standard. *Shock and Vibration, 2022*, Article 6412120. https://doi.org/10.1155/2022/6412120

Singh, A., Singh, L. P., Singh, S., Singh, H., Chhuneja, N. K., & Singh, M. (2019). Evaluation and analysis of occupational ride comfort in rotary soil tillage operation. *Measurement, 131*, 19-27. https://doi.org/10.1016/j.measurement.2018.08.046

Singh, A., Singh, L. P., Singh, S., Singh, H., & Prakash, C. (2018). Investigation of occupational whole-body vibration exposure among Indian tractor drivers. *International Journal of Human Factors and Ergonomics, 5*(2), 151-165. https://doi.org/10.1504/IJHFE.2018.092240

Sorainen, E., Penttinen, J., Kallio, M., Rytkönen, E., & Taattola, K. (1998). Whole-body vibration of tractor drivers during harrowing. *American Industrial Hygiene Association Journal, 59*(9), 642-644. https://doi.org/10.1080/15428119891010820

Tiemessen, J. H., Hulshof, C. T. J., & Frings-Dresen, M. H. W. (2008). Low back pain in drivers exposed to whole-body vibration: Analysis of a dose–response pattern. *Occupational and Environmental Medicine, 65*(10), 667-675. https://doi.org/10.1136/oem.2007.035147

Upadhyay, R., Bhattacherjee, A., Patra, A. K., & Chau, N. (2022a). Association between whole-body vibration exposure and musculoskeletal disorders among dumper operators: A case-control study in Indian iron ore mines. *Work, 71*(1), 235-247. https://doi.org/10.3233/WOR-205140

Upadhyay, R., Jaiswal, V., Bhattacherjee, A., & Patra, A. K. (2022b). Role of whole-body vibration exposure and posture of dumper operators in musculoskeletal disorders: A case study in metalliferous mines. *International Journal of Occupational Safety and Ergonomics, 28*(3), 1711-1721. https://doi.org/10.1080/10803548.2021.1932111

Upadhyay, R., Senapati, A., Chau, K., Bhattacherjee, A., Patra, A. K., & Chau, N. (2024). Development of an intervention programme to reduce whole-body vibration exposure based on occupational and individual determinants among dumper operators. *International Journal of Occupational Safety and Ergonomics, 30*(1), 41-55. https://doi.org/10.1080/10803548.2022.2159176

Vikram, P., Aruna, M., & Arif, S. (2024). Evaluating whole-body vibration exposure in rock-breaker operators: Influence of rock type and operational parameters. *Journal of Mines, Metals and Fuels, 72*(11), 1163-1177. https://doi.org/10.18311/jmmf/2024/44527

Zhukova, A. G., Kizichenko, N. V., Gorokhova, L. G., & Kazitskaya, A. S. (2022). Experimental models of vibration disease: A literature review. *Hygiene and Sanitation, 101*(7), 776-782. https://doi.org/10.47470/0016-9900-2022-101-7-776-782

Pertanika J. Sci. & Technol. 34 (3): 1745 - 1767 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

A LiDAR and IoT Integrated Approach for Intelligent Monitoring of Lubricant Liquid Levels: A Lab Scale Study on Petronas Products

Nur Shaebah Mat Hussain[1]*, Mohd Khairol Anuar Mohd Ariffin[1], Siti Azfanizam Ahmad[1], and Zulkifli Mohamed[2]

[1]Department of Mechanical and Manufacturing Engineering, Faculty of Engineering, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[2]Department of Mechanical Engineering, Faculty of Mechanical Engineering, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

ABSTRACT

This study investigated the development of an Internet of Things (IoT) based lubricant level monitoring system. Monitoring lubricant levels is essential for maintaining the reliability, efficiency, and longevity of industrial machinery. Conventional lubricant level inspection methods are often manual, time-consuming, and unsuitable for continuous monitoring. To address this limitation, this study develops the system integrates a Light Detection and Ranging (LiDAR) LiDAR sensors and IoT technology using an ESP32 microcontroller for real-time lubricant level measurement and remote data monitoring. The proposed system continuously measures lubricant levels and transmits the collected data through an IoT platform, enabling users to access monitoring information remotely. Laboratory tests were carried out using two Petronas products. They are base oil, known as ETRO 6+ and product, termed as Syntium 800 10W-40. Five measurement tests were performed for each lubricant to evaluate the stability and consistency of the monitoring system. The collected data were analysed using statistical parameters, including mean and standard deviation. The test results showed stable performance with standard deviations of 10.6 mm and 9.27 mm, respectively. The proposed system offers a practical solution for predictive maintenance applications. Future work should focus on evaluating the system under actual industrial environments involving vibration, temperature fluctuations, and other operational disturbances to further validate its industrial applicability and long term reliability. These findings indicate that the developed LiDAR-IoT monitoring system is capable of providing accurate and continuous lubricant level monitoring under controlled laboratory conditions.

Keywords: Industrial automation, IoT, LiDAR sensor, lubricant level monitoring, real-time data

ARTICLE INFO

Article history:
Received: 12 April 2026
Accepted: 06 June 2026
Published: 25 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.20

E-mail addresses:
shaebahhussain@gmail.com (Nur Shaebah Mat Hussain)
khairol@upm.edu.my (Mohd Khairol Anuar Mohd Ariffin)
s_azfanizam@upm.edu.my (Siti Azfanizam Ahmad)
zulkifli127@uitm.edu.my (Zulkifli Mohamed)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

INTRODUCTION

Advances in Industrial Revolution 4.0 technology have created technology such as sensors, from the ordinary to the best, such as the integration of light detection sensors (LiDAR). A LiDAR is often used in navigation mapping because of its precise, contactless nature, which is suitable for measuring lubricating oil levels (Epelle & Gerogiorgis, 2020; Manish et al., 2022). LiDAR sensor intelligence has been integrated with a robust IoT framework to create a transformative approach towards real-time measurement of the residual lubricant level in a tank (Tanati et al., 2023). In the case of this study, the LiDAR sensor system is to measure the fluid level in the tank and send the data through a microcontroller (ESP32) and upload it to a cloud platform via Wi-Fi or MQTT.

The setup is enabled with real-time monitoring and alert systems to prevent machine failure. Continuous data collection, remote or close observation and prompt decision making allow for the concentration of sensors in high resolution with any proper digital connection (Xin et al., 2024), which improves operational performance and reduces reliance on unscheduled maintenance. The lubricating systems have self-adjusting control and optimisation based on real-time operating conditions, which have been significantly enhanced by recent developments in sensor attenuation, wireless communication protocols, and artificial intelligence (AI) (Liu et al., 2024).

Traditional monitoring methods usually rely on manual inspections or simple sensors such as float-based systems that are often exposed to inaccuracies, delayed responses, and human error (Sparham et al., 2014). These traditional methods lack the swiftness in detecting any abnormal situations early, which leads to system production shutdown and any unnecessary problems. Most industries still practice manual or basic monitoring methods due to limitations in awareness and the use of technology. This method will usually lead to issues such as unexpected shortage of lubricant stock, resulting in shutdowns in the production line (Zhang et al., 2025). With the advent of industrial technology, the industry is beginning to realise the need to introduce smart technology to reduce risks, as well as increase efficiency and optimise resource utilisation (Benziane et al., 2025).

As a result, improvement with the use of intelligent monitoring is a critical step towards better predictive maintenance and operational efficiency. Failure to maintain lubrication levels is often associated with a lack of routine maintenance, especially in machinery and motor equipment (Chen et al., 2020). Therefore, continuous monitoring of lubricant levels is a must to ensure operational reliability and extend the service life of equipment (Hanaysha, 2016; Quinchia et al., 2014).

The transformation of smart monitoring technology has revolutionised the way industrial systems manage operations in production, in addition to maintenance and sustainability (Varalakshmi et al., 2025). This is because one of the branches of this transformation is the Industry Internet of Things (IIoT), where industry players are

shifting from a reactive maintenance approach to predictive maintenance and automation methods.

In this study, the LiDAR and IoT framework is a good system that can be used in the oil and gas industries (demonstrates significant potential for application in the oil and gas industry). Radar, Lidar, and Photocells-based method systems typically involve higher initial costs due to the sophisticated technology employed, but they offer real-time and accurate data, justifying their investment (Lo Castro et al., 2024). The study focuses on the development and validation of LiDAR-IoT integrated monitoring systems for lubricant level measurement (the study focuses on the development and validation of an integrated LiDAR-IoT monitoring system). An intelligent system with a LiDAR sensor for level monitoring, which is contactless, can significantly improve the performance of the lubricating system, especially in ensuring smooth and more efficient mechanical operation (The intelligent contactless monitoring system performance by ensuring smoother and more efficient mechanical operations). The study makes direct comparisons with other sensing technologies (Direct comparisons with other sensing technologies), such as ultrasonic, capacitive, and float-based systems (float-based systems), which are beyond the scope of the study.

The literature review highlights several key factors that influence its design and effectiveness. These factors include operational efficiency, security and reliability, cost reduction, scalability, and sustainability. Hence, this study establishes a performance standard for next-generation smart lubrication management systems while addressing the gaps in the current systems (Wilińska et al., 2025). A comprehensive benchmarking analysis of conventional sensing technologies is recommended as future research to further evaluate the relative performance, accuracy and practicality of the proposed system.

The research gap and novelty have been strengthened by highlighting those existing studies on liquid level monitoring that focus on conventional sensing technologies, whereas the integration of LiDAR with IoT for lubricant level monitoring remains underexplored. Furthermore, limited studies have investigated the feasibility of LiDAR-based monitoring specifically for Petronas lubricant products, incorporating real-time cloud-based visualisation and remote monitoring capabilities. Internet of Things (IoT) technology has emerged as a promising solution for addressing these challenges (Ahlawat & Rana, 2021; Boddu et al., 2023).

Efficiency in operation is defined as the ability of the system to convert the input into a desired output with minimal loss. In lubrication management, the efficiency of the system depends on the accuracy of the system configuration, timely feedback and continuous monitoring to avoid mechanical failure. In traditional inspection methods that rely heavily on manual reading, which lack consistency and are prone to human error, this often leads to delayed maintenance and the incurrence of huge losses due to a stop in production.

Therefore, the novelty of this work lies not only in overcoming the limitations associated with conventional methods but also in the development and validation of an integrated LiDAR via an IoT framework for intelligent lubricant level monitoring applications, which remains insufficiently explored in existing literature.

It mirrors the real-world counterpart and allows simulation, monitoring, and analysis. Combining machine vision with IoT sensor data, cloud storage, and machine learning algorithms can proactively enhance the accuracy of measurement prediction, machining responses and schedule maintenance (Kaur et al., 2020; Liu et al., 2022).

LiDAR sensors provide superior spatial stability, response time and non-contact measurement capabilities. The LiDAR is resistant to surface pollution or turbulence, making it a good tool for continuous and stable monitoring. However, many existing studies focus on environmental or hydrological applications, such as flood monitoring and terrain mapping, rather than lubrication systems that differ significantly in properties such as viscosity and optics (Mast et al., 2024). Therefore, a more reliable monitoring system is the LiDAR and IoT that help to improve efficiency in production operations and the overall supply chain.

Safety and reliability are two important factors of a monitoring system used in the industry. Failure to provide lubricant stock will cause many to be affected because the decline in stock level and supply will cause the machine to shift and line shutdown due to overheating, and lead to catastrophic mechanical failure (Shi et al., 2021). IoT-based monitoring technology provides a preventive and early warning system that reduces risk and ensures safe operations. LiDAR, with its non-contact measuring ability and optical properties, enhances the system's predictability by eliminating issues related to corrosion, electrical interference and mechanical wear associated with sensors such as in non-contactless sensor systems

On the other hand, when integrated with IoT platforms, LiDAR data can trigger automatic alerts and predictive maintenance responses, reducing shutdown time in unplanned production (Hatta Antah et al., 2021). Despite these advantages, the environmental factors such as oil mist, reflective surfaces, and particulate matter can affect the quality of the signal, underscoring the need for advanced calibration and adaptive algorithms to maintain accuracy in real-world industrial settings (Schneider et al., 2021).

In addition, cost saving is another factor that is the main driver for using this type of tape monitoring system (Muneeswaran et al., 2023). Intelligent maintenance powered by IoT can reduce maintenance costs and minimise maintenance losses due to line and unplanned shutdowns. Although LiDAR was once considered a more expensive technology than ultrasonic sensors, advances in solid-state LiDAR have significantly reduced costs. When considering its long-term durability, minimal maintenance requirements and high reliability, smart LiDAR systems offer an excellent return on investment. However,

existing studies rarely provide comprehensive cost-benefit analyses that compare LiDAR with conventional sensors in smart monitoring of the lubricant level, suggesting research opportunities to evaluate technoeconomic to offset performance, cost, and long-term value (Castro et al., 2024).

Scalability is another important aspect, in addition to features that are suitable for modern industrial systems. Scalable systems can evolve and adapt to meet growing operational needs without the need for a complete redesign (Rashidifar et al., 2022). For IoT Frame protocol support, communication protocols such as MQTT are indeed the preferred choice as they allow multiple sensors to work more efficiently and seamlessly within a unified framework. LiDAR sensors are one of the best sensors that have been designed to be compatible with common microcontrollers such as Arduino, ESP32 and Blynk Apps that are suitable and easily integrated into larger systems. But it will become more difficult in terms of finding compatibility in terms of latency, power management and data transmission as sensors increase.

Machine vision systems have demonstrated the capability to perform non-contact and real-time tool condition monitoring through both direct and indirect measurement approaches. These systems can effectively detect tool wear and machining conditions, reducing dependence on manual inspection while improving monitoring efficiency in manufacturing operations (Pimenov et al., 2024). Meanwhile, the various techniques, including image enhancement, feature extraction, machine learning, and deep learning algorithms, for evaluating tool wear, surface quality, and defect detection reported that these approaches enable automated monitoring and quality control, contributing to improved measurement accuracy and manufacturing reliability (Ercetin et al.,2024). Several algorithms for flank wear assessment have been reported to have high measurement accuracy and repeatability. The study found that the effectiveness of machine vision systems for automated, non-contact, and real-time monitoring supports predictive maintenance and productivity improvement in modern manufacturing environments (Makhesana et al., 2025).

These studies demonstrate the growing adoption of intelligent sensing, machine vision, and automated monitoring technologies in Industry 4.0 applications. Although previous research primarily focused on tool condition monitoring and machining performance assessment, the findings highlight the advantages of non-contact sensing for real-time industrial monitoring. The advantages of these developments, the present study proposes a LiDAR sensor integrated with IoT technology for continuous lubricant level monitoring. The proposed system provides real-time data acquisition and remote monitoring capabilities, offering a practical solution for predictive maintenance and smart manufacturing environments.

METHODOLOGY

A prototype simulation of a laboratory-scale level monitoring system has been developed, made of smart liquids using the integration of LiDAR and IoT, specifically for smart lubricant tanks. This prototype is a replica of the existing tank in the industry but has been reduced in size to allow conditions to mimic the actual lubricant storage operation. The test setup aims to evaluate the ability of LiDAR sensors to accurately detect fluid levels and transmit data through IoT networks using microcontrollers.

Based on the research flow chart shown in Figure 1, this test was conducted in two different phases, namely hardware installation and software implementation phases. The hardware installation phase begins with a study on the design and fabrication of laboratory-scale prototypes. Several suitable tank designs were identified to be used as a prototype, and the PUGH matrix was employed to determine the most suitable design alternative

Based on the matrix, we chose the cylinder-shaped tank design. The prototype was developed, taking into consideration the material as well as measurement to ensure its performance behaviour closely resembles real-world operating conditions. The conversion from a full-scale industrial system to a compact laboratory setup allows experiments to be conducted under controlled conditions. In doing so, it minimises external interference temperature and environmental changes. This allows for a reliable testing and validated data collection process. In the software installation phase, ESP32 and Blynk were used to create a physical dashboard monitoring system. This system is then connected with the LiDAR sensors, enabling real-time collection, transmission, and display of the lubricant

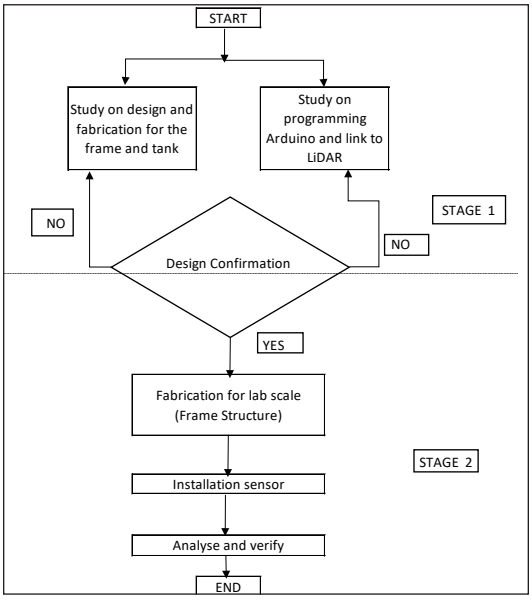


Figure 1. Research flow chart

level. The integration of the monitoring system with an IoT connection allows for two-way communication between physical hardware and a cloud-based monitoring interface. This allows users to observe, record, and analyse system performance remotely through mobile or web applications. Automation of data synchronisation helps to provide timely updates and reduce latency in sensor readings.

After which, a trial of the integrated monitoring system was carried out to validate its mechanical system's performance, accuracy of the sensors installed, as well as the IoT-based monitoring capabilities for potential customisation in an industrial lubrication management application.

During the test period, the usability of the LiDAR sensor was evaluated. Environmental parameters, such as temperature and laboratory conditions, were not considered; they remained constant under controlled experimental conditions as the prototype was built to gauge the readings of the lubricant level in the tank, as shown in Figure 2.

To facilitate the test, two different PETRONAS lubricant samples were used: the raw lubricant, identified as PETRONAS ETRO6+ base oil, and the final product, derived from the same base oil, PETRONAS Syntium 800 10W-40. These samples were tested to observe the standard deviation between the two samples, focusing on the expected variations arising from the inherent characteristics of LiDAR sensors, which differ significantly from conventional ultrasonic sensors commonly used in lubricant level monitoring. The prototype system shown in Figure 3 was tested under controlled laboratory conditions. The following experimental parameters, namely laboratory temperature, surrounding environment, fluid velocity, and viscosity, were kept constant.

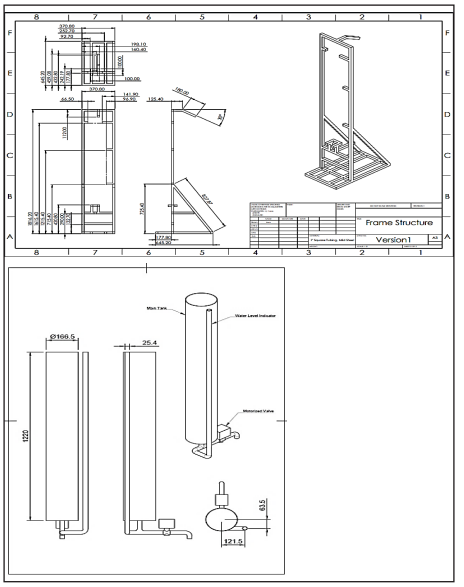


Figure 2. 2D Drawing of the prototype: top: frame; bottom: tank

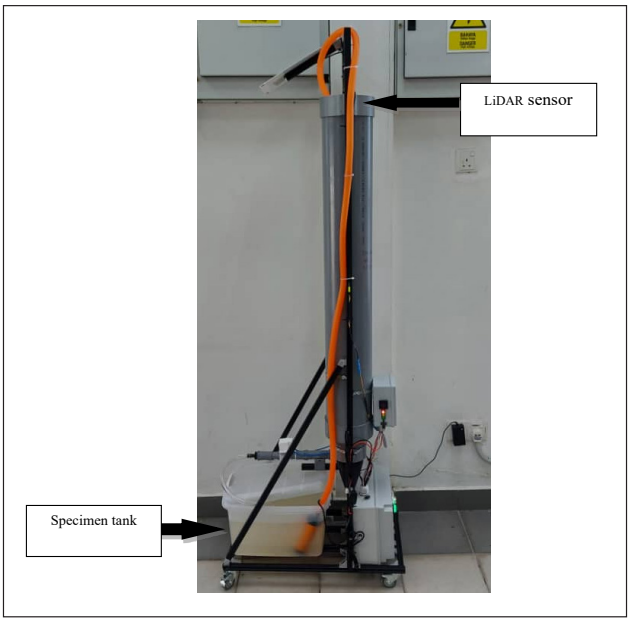


Figure 3. Prototype system

The system showed reliable performance in monitoring lubricant levels. However, further testing is required to evaluate its performance in real industrial conditions. Two PETRONAS lubricant samples: base oil PETRONAS ETRO 6+ and the finished lubricant product PETRONAS Syntium 800 10W-40 were used to assess accuracy as well as the response rate of the LiDAR sensor. The system is designed to reduce reliance on manual maintenance and prevent lubricant run-out that could compromise machine operation. The experiment was repeated five times to ensure the reliability and validity of the data collected. The repetition principle is important in scientific research because the observational indices are random variables, which require a certain number of samples to reveal their changing regularity. The repetition principle helps stabilise the mean and the standard deviation, thereby enabling sample statistics to more accurately represent population parameters. Thus, the statistical inference will be reliable. This article discussed the repetition principle from the perspective of common sense and speciality with examples. Repeated measurements reduce the impact of random errors, thereby improving data robustness and reliability for subsequent analysis (Hu et al. 2011).

Two different laboratories at two different locations were used for the purpose of this study. The fabrication and development stage was carried out at the Sports Engineering and Artificial Intelligence Centre (SEA-IC) Laboratory, UiTM Shah Alam, while the system calibration testing was carried out at the Fluid Laboratory, UPM. The objective of this study is to find out the standard deviation value of two lubricating oils from Petronas ETRO6+ base oil and compare it with the blended lubricating oil to get the standard deviation, which

provides information on the consistency and variation of measurement between the two lubricants. As illustrated in the system circuit diagram shown in Figure 4, the electrical configuration operates under controlled electrical current conditions, and its configuration was carefully developed to ensure both functional performance and operational safety. As the system circuit diagram demonstrates, the electrical configuration operates through controlled electrical current, and its configuration has been carefully developed to ensure not only functionality but also operational safety.

Every single component of the electrical hardware is orderly arranged, systematically connected, to minimise the risk of electric degradation, without compromising on the performance. The integration of electrical circuits with power supplies, switches, resistors, sensors and other essential elements and components is chosen for its distinctive function in ensuring the effectiveness, stability and reliability of the system. The circuit is designed to support the required electrical load while also providing adequate safety features in safety protection in case of an emergency; therefore, it also ensures the safety and longevity of the system. The system technique has been used to build a circuit so that the prototype works properly. The diagrams that have been created provide a clear picture and function as a reference for any changes or future changes. The electrical components are carefully connected to the mechanical system to guarantee they function properly. This organised design allows the prototype to perform as planned and provides a dependable and sturdy solution for its use.

The prototype was designed systematically. The circuit schematics were prepared to have a better understanding and serve as a reference for future modifications or any improvements. The integration of electrical components with the mechanical system was carefully done, so that the components and system complement each other. Hence, creating a robust and reliable system that can be used as intended.

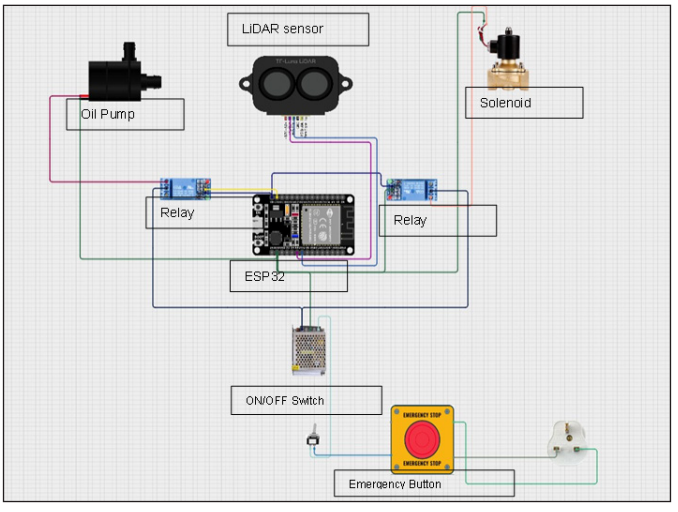


Figure 4. System circuit

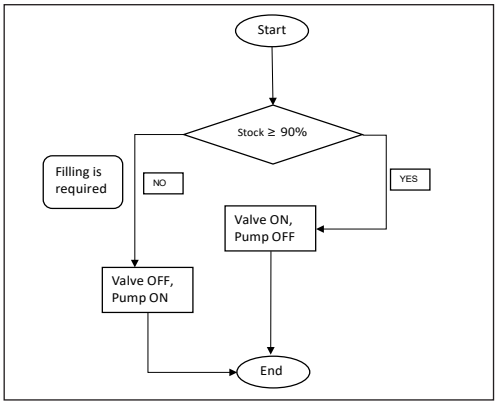


Figure 5. Programming process flow chart

Figure 5 shows the programming flowchart, which explains the logical sequence of the automatic control process. The system starts by continuously monitoring the lubricant level in the tank. It then checks whether the lubricant level is sufficient, based on the tank's full capacity. The flowchart serves to visually explain the logical sequence of the automatic control process.

The monitoring of levels of lubricant in the tank begins with the 'Start' button, where the result notes are used to check if the lubricant levels in the tank are at the optimum level, which is 100 per cent of the tank's capacity. If the reading is at 90 per cent or more, the system will automatically identify it as sufficiently filled. After which, the valve is turned on to release the excess fluid, and the pump is turned off to stop the lubricant from flowing into the tank. On the other hand, if the reading falls below 90 per cent, the pump will be turned on, for lubricants to flow into the tanks, while the valve is turned off, until it reaches its optimal level.

The program for the sensor was developed using the TTGO ESP32 (NodeMCU) microcontroller. It serves as the main controller for the IoT-based lubricant level monitoring system. It is also responsible for sensor startup, real-time data collection, basic signal filtering, and data transmission to the Blynk IoT platform for remote monitoring and visualisation. Also, the program used a modular coding structure to ensure clarity, scalability, and help with future maintenance. The main reason for choosing NodeMCU is its features. The NodeMCU has built-in WiFi, low power consumption and easy sensor integration. This open-source platform is cost-effective and more efficient when compared to alternatives, such as Raspberry Pi and Arduino Mega. The NodeMCU can activate the pump and valve accordingly when the lubricant level reaches a certain threshold. Through IoT monitoring, automatic control guarantees timely fluid management, can prevent overflow or leakage, and enables predictive maintenance.

The following pseudocode shows the sequence and structure of the program's logic. It provides an overview of the functions and operation of the prototype system:

```
BEGIN
  //Initialise Serial Communication
  Initialise Serial
    //Initialise LIDAR sensor with pins (RX, TX)
  Initialise LIDAR communication with pins (RX, TX)
  Initialise the TFMPlus library with the LIDAR device
    //Initialise TFT display
  Initialise TFT display (invert display, set rotation, set text size)
    // Set pin modes for LEDs, relays, and buzzer
  Set pin modes for:
    LED Yellow, LED Red, LED Green
    Relay 1, Relay 2, Relay 3, Relay 4
    Buzzer Pin
      //Initialise Blynk connection for remote monitoring
  Initialise Blynk connection with authentication token
    // Start system operation
  PRINT "System initialisation complete."
    // Main loop
  WHILE true:
      // Read distance, strength, and temperature from LIDAR
    IF LIDAR data is available:
      Read distance, strength, and temperature from LIDAR
        // Display data on TFT screen
    DISPLAY "Distance: " + distance + " cm" on TFT screen
    DISPLAY "Strength: " + strength on TFT screen
    DISPLAY "Temperature: " + temperature + " degC" on TFT screen
        // Send distance value to Blynk app
    SEND distance data to Virtual Pin V1 on Blynk
        // Control LEDs based on distance
    IF distance > 90:
      Turn ON red LED
      Turn OFF the yellow LED
      Turn OFF green LED
      Trigger Buzzer for 3 short bursts (tone)
    ELSE IF distance > 70:
      Turn OFF red LED
      Turn ON yellow LED
      Turn OFF green LED
      Turn OFF Buzzer
    ELSE:
      Turn OFF red LED
      Turn OFF yellow LED
      Turn ON green LED
      Turn OFF Buzzer
    // Control relays based on virtual pin values from Blynk app
    IF Virtual Pin V0 has value:
      Control relay connected to "oren"
    IF Virtual Pin V2 has value:
      Control relay connected to "hitam"
    // Execute Blynk functions for app updates and communications
    RUN Blynk app loop (Blynk Edgentrun)
  END WHILEEND
```

RESULT AND DISCUSSION

During the initial stages, the frame and tank design were developed using both 2D and 3D computer-aided design (CAD) models. These models figured the system's geometry, layout of the component, and structural configuration. These models provide a comprehensive representation of the lubricant monitoring system, ensuring that all mechanical and electronic components can be smoothly integrated into housing and mounting structures.

After this, there is the validation of the prototype, which used Finite Element Analysis (FEA). It measures the mechanical performance, stress distribution and deformation of the prototype when use. The FEA evaluates the maximum, average and minimum stress concentration that causes deformation when the prototype is subjected to mechanical loads and vibrations. FEA is specially chosen as it can provide pertinent insights towards modifications, in terms of the selection of materials, as well as the thickness of the prototype. This is purposefully carried out to ensure structural rigidity, durability, and stability of the prototype. The simulation was carried out using a 3D CAD model prototype. Fixed constraints were placed at the base and distribution loads along the support surface to simulate the actual condition when in use.

As presented in Table 1, it is crucial to evaluate the prototype performance through simulation analysis to ensure compliance with engineering best practices, as well as the principles of precision design and cost-effective prototyping.

Table 1
Finite Element Analysis (FEA) summary table

Analysis Type	Purpose	Key Focus	Findings / Objective
Static FEA (No External Loads)	Assess frame behaviour under self-weight and gravitational force	Structural deformations, stress distribution, stiffness, displacement	Identify areas of potential weakness or stress concentration, ensuring robustness under normal conditions
Static Load Analysis	Evaluate frame response to specific external loads	Local deformation, stress concentrations, and potential failure modes	Determine the frame's ability to resist bending, shearing, or torsional stresses under operational loads
Loaded FEA (With External Loads)	Simulate frame's behaviour under complex loading scenarios (e.g., dynamic, thermal)	Impact of dynamic loads, thermal effects, and vibrations on frame performance	Identify worst-case scenarios, assess performance under transient forces, and consider potential design optimisation
Comprehensive FEA Assessment	Combine static and loaded analyses to evaluate overall structural integrity	Load distribution, material properties, stress analysis across scenarios	Inform design adjustments, including material selection, geometry, and reinforcement for enhanced stability

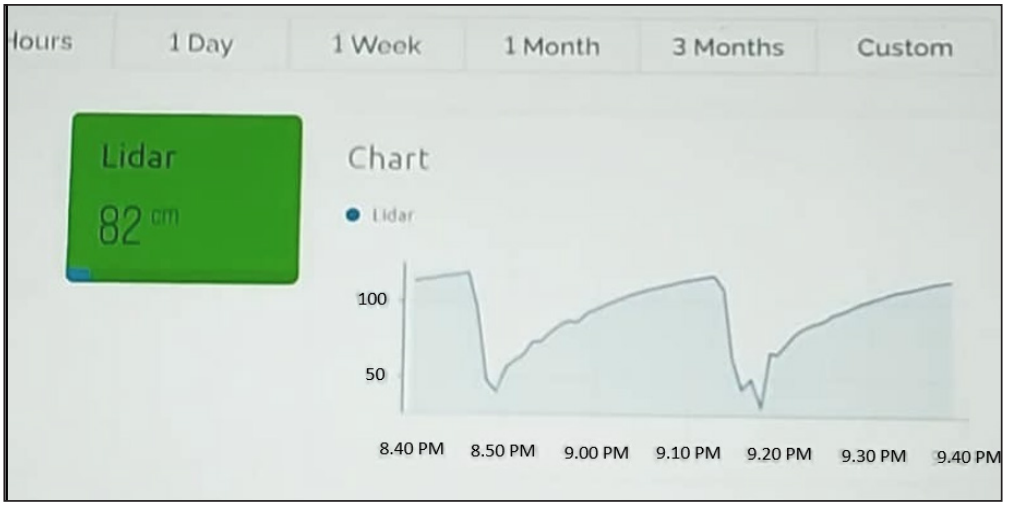


Figure 6. Real-time visualisation of lubricant level data measured by the LiDAR sensor on the IoT dashboard

The system and prototype were put to the test and repeated five times, to make sure that the results obtained are valid and reliable. From these repeated measurements, the mean and standard deviation were calculated. The comparison of standard deviation values between the base oil PETRONAS ETRO 6+ and the finished lubricant PETRONAS Syntium 800 10W-40 helps to evaluate the sensor's consistency and sensitivity. It revealed how the difference in fluid viscosity and composition influences LiDAR measurement stability in controlled experimental conditions.

Figure 6 illustrates that the graph and the real data appeared as digital numbers. Calibration is critical, as it helps to make sure that the system is functioning well. Before the experiment, the sensors were calibrated to ensure measurement stability and accuracy during data collection. At the same time, readings were also tested and displayed through the Blynk application.

The calibration process is designed to align the sensor output voltage and distance reading with the actual fluid level in the tank, minimising systematic failure or avoiding any error and ensuring that subsequent measurements reflect actual physical variations rather than sensor interference. At the start of the calibration process, the LiDAR sensor is placed at a fixed height above the tank surface. Distance readings were then recorded as the fluid level changed. Each reading was compared with the actual fluid height measured in the tank.

The test was repeated five times, under the same conditions and prototype. This is done to ensure that the data collected is reliable and valid. Repetition, in taking the measurement, helps in identifying random variation due to sensor sensitivity or acute testing fluctuation. The collected data were then analysed. They were analysed using the cumulative standard

Table 2
Results of loading lube and unloading lube by using two different specimens from two products of PETRONAS: Raw material and end product

No.	Material	Colour	Loading Lube by Using a Pump	Unloading Lube by Using a Solenoid
1	Base Oil, PETRONAS Etro 6+		Low viscosity enables easy pumping	Poor LiDAR detectability due to optical clarity
2	PETRONAS Syntium 800, 10W-40		Higher viscosity may require increased pump pressure	Good LiDAR detectability due to high opacity

deviation (σ), a statistical parameter that measures the difference in reading when compared to the mean value. degree of measurement spread relative to the mean value.

The result of the data analysis from the test carried out, as Table 2, has confirmed the reliability of the LiDAR-based liquid level monitoring system, even when using lubricants with different properties. These findings reinforce the use of LiDAR technology for monitoring of lubricant level at a larger scale, like industrial lubrication. This has happened because of their repeatability, operational efficiency, accuracy and adaptability, which are critical for predictive maintenance and stock levels.

The result of the data analysis from the test carried out, as Table 2, has confirmed the reliability of the LiDAR-based liquid level monitoring system, even when using lubricants with different properties.

These findings reinforce the use of LiDAR technology for monitoring of lubricant level at a larger scale, like industrial lubrication. This has happened because of their repeatability, operational efficiency, accuracy and adaptability, which are critical for predictive maintenance and stock levels.

As shows as Table 3 and 4, by using Equation 1, the Cumulative Standard Deviation (σ) for the PETRONAS ETRO 6+ base oil and product of PETRONAS Syntium 800 10W-40 was derived from five sets of measured data for each specimen. The graphical representations indicate a degree of uniformity and stability in the readings of the sensors, where lower standard deviation values indicate higher value accuracy and consistent sensor performance. A lower Cumulative Standard Deviation (σ) value corre sponds to a more consistent

Table 3
PETRONAS ETRO 6+: Data collection to get sigma value

PETRONAS ETRO 6+					
Sequence, n	1	2	3	4	5
Time taken, (s)	244	260	230	248	257
xi-μ	-3.8	12.2	-18	0.2	9.2
$(xi - \mu)^2$	144	149	317	0	84.6
Cumulative Distribution	0.36	0.87	0.05	0.51	0.81

Table 4
PETRONAS Syntium 800, 10W-40: Data collection to get sigma value

PETRONAS Syntium 800, 10W-40					
Sequence number, Ո	1	2	3	4	5
Time, (second)	195	210	185	205	190
xi-μ	-2	13	-12	8	-7
$(xi - \mu)^2$	4	169	144	64	49
Cumulative Distribution	0.41	0.92	0.1	0.81	0.23

reading, confirming that the system is capable of high repeatability and low measurement uncertainty.

To quantify the static variance in sensor readings, the standard deviation (σ) of the lubricant level measurements was calculated by using the formula in Equation 1.

The formula calculation shall be:

$$\sigma = J\overline{\frac{1}{n-1}\sum_{i=1}^{n}[x_i - \pi]^2} \quad [1]$$

Sigma, σ = standard deviation
n = total number of observations
x_i = each value from the data
μ = the mean of the data

Then the mean deviation of the value from the mean is determined as: $\sum_{i=1}^{n}[x_i - \pi]^2$

To calculate x_i-μ Every number has to be minus the mean value.
244-247.8 = -3.80, and so on; using the same formulae.

Using the numbers acquired, the mean deviation can be calculated.
Let's use a value as an example: $= \sum_(i=1)\text{^}n \ [\![(x_i-\mu)^2]\!] \ = (-2)^2 = 4$
By adding all the values from $(x_1-197)^2$ to $(x_n-197)^2$,
The final total of the Sum of Differences is: 430
and the variance of the PETRONAS ETRO 6+ is: 430/5
Thus, the Standard Deviation of PETRONAS ETRO 6+, $\sigma = \sqrt{86} = 9.27$

The graph of the cumulative and sequence for PETRONAS Etro6+ is plotted in Figure 7.

The static evaluation allows for the identification of measurement dispersion and serves as a performance indicator for accuracy, precision and stability of LiDAR sensors. When the accuracy of the system is benchmarked against conventional level monitoring methods such as ultrasonic and capacitive sensors, it demonstrates performance and reliability that is consistent with various environmental conditions, including temperature, viscosity and optical interference. These findings underscore the potential of integrated LiDAR sensors and IoT systems to deliver monitoring accuracy, fast response times, and significant maintenance cost reductions. The sensor calibration in measurement, static verification and comparative tests proves the effectiveness of this system in the management of real-world industrial lubrication.

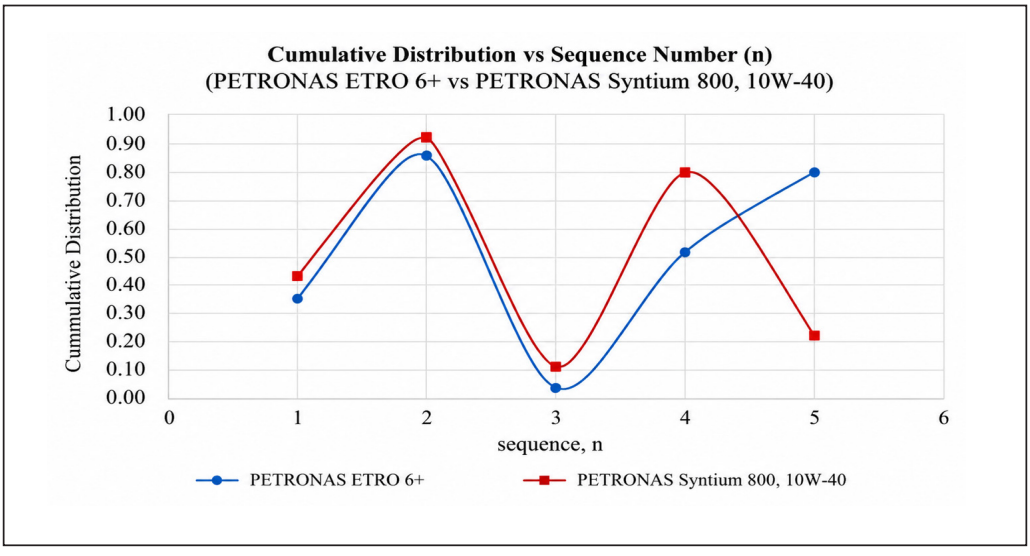


Figure 7. Combined cumulative distribution profiles of PETRONAS ETRO 6+ and PETRONAS Syntium 800 (10W-40)

Table 5
Comparison between PETRONAS 6+ and Syntium 800, 10W-40

Type	PETRONAS ETRO 6+	PETRONAS Syntium 800, 10W-40
Data quantity	5	5
Sum	1239	985
Mean, μ	247.8	197
Subtract from the Mean	564.8	430
Variance, σ^2	112.96	86
Standard Deviation	10.63	9.27

The statistical results summarised in Table 5, which compares PETRONAS ETRO 6+ and PETRONAS Syntium 800 (10W-40). The observed differences in mean, variance, and standard deviation indicate variations in lubricant level measurements, reflecting the distinct monitoring characteristics of the two lubricant samples.

The calculation of both data showed that the difference of standard deviation is 1.36, and by comparing the data of standard deviation, it shows how consistent and reliable the results are in different datasets. The data showed a lower standard deviation, which means that the values are closer to the mean and are less affected by random factors. A higher standard deviation means that the test scores are farther from the mean and thus more influenced by random factors.

Figure 7 shows the combined cumulative distribution profiles of PETRONAS ETRO 6+ and PETRONAS Syntium 800 (10W-40). The graph PETRONAS ETRO 6+ represents in a blue line, the normalised cumulative distribution function derived from the measured lubricant level, providing an overview of the trend variation and stability of the measurements from various experiments. A product of lubricant, PETRONAS Syntium 800 (10W-40), represented in the red line, has properties of higher viscosity and greater optical opacity compared to the PETRONAS ETRO 6+ base oil. The plot captured that the sensor was responding consistently over five measurement sequences by providing a perspective on how fluid characteristics influence LiDAR signal stability and measurement precision.

The inconsistency of the fluctuation curve in the red line indicates of PETRONAS ETRO 6+ that the LiDAR readings in the measurement, this is because the variation reflects the sensor's characteristic capabilities, high transparency and low reflectivity, which can reduce the sensor’s ability to capture consistent reflected light signals from the fluid surface. As LiDAR technology relies on the time of flight (ToF) of laser pulses

to determine distance, any reduction in the returned signal strength may lead to minor measurement deviations or signal.

For red lines for PETRONAS ETRO 6+, the fluctuation of the graph shows the cumulative distance values that the sensor captured, a temporary reduction in detection accuracy, such as signal loss caused by the lubricant's optical clarity. On the other hand, the smoother portions of the curve indicate a higher measurement consistency, where the LiDAR readings are more stable and repeatable. The standard deviation (σ) calculated from these readings quantifies the extent of dispersion in the measured lubricant levels. A moderate standard deviation (σ) value indicates that the sensing remains to maintain a level of accuracy even when the degree of measurement variability occurs due to the optical properties of the base oil. This finding is in line with the transparent liquid properties that produce less reliable LiDAR reflections compared to more pigmented lubricants such as PETRONAS Syntium 800 10W-40.

For as PETRONAS Syntium 800 10W-40, the pattern demonstrates two prominent peaks with a relatively smooth transition between sequences, indicating repetitive yet stable measurement behaviour. Even though there are minor fluctuations observed, the overall data trend suggests improved reflectivity and signal reliability, due to lubricant properties and the reduced light transmittance. These optical properties enhance the return intensity of the laser beam, allowing the LiDAR sensor to provide more consistent distance readings.

The comparison between the PETRONAS ETRO 6+ base oil, which exhibited measurement irregularities, was due to its transparency. The PETRONAS Syntium 800 10W-40 data showed narrower dispersion around the mean, resulting in a lower standard deviation (σ) value. This lower standard deviation (σ) indicates greater precision and repeatability, validating that the LiDAR system performs more effectively with opaque or semi-opaque lubricants that minimise beam scattering. The PETRONAS Syntium 800 10W-40 results confirm that higher viscosity and optically opaque lubricants enhance LiDAR detection reliability by strengthening signal reflection. This finding supports the conclusion that fluid optical density plays a key role in determining LiDAR measurement precision. As a result, the use of LiDAR sensors in industrial lubrication monitoring is suggested for the product of lubricants, where consistent level detection is essential for predictive maintenance, efficiency control, and system automation.

The system performance was validated through a series of statistical data collection from experiments by two specimens from PETRONAS ETRO 6+ and PETRONAS Syntium 800 10W-40, which show standard deviations of both are 10.6 and 9.27. These values indicate that both lubricants demonstrated acceptable measurement stability, with

PETRONAS Syntium 800 10W-40 showing slightly higher precision due to its greater optical pigmentation and viscosity.

The validation process encompassed the observation of real-time variations in lubricant levels, pump activation durations, and the total volume of lubricant dispensed. The data has been recorded in the cloud. To ensure that the system is reliable, these measurements have been benchmarked against the data obtained through manual monitoring techniques. Comparative observations confirm that the LiDAR integrated monitoring system performs better in all aspects, such as achieving maximum accuracy, consistency, and operational efficiency, in addition to lowering the probability of human error and response delays typically associated with manual observation. The findings validate that the system has the potential for real-time, contactless monitoring applications, particularly in industrial lubrication management, where measurement precision and automation are critical for predictive maintenance.

The cumulative distribution results confirm that the level monitoring system by using LiDAR based liquid level monitoring system can detect the level of lubricant in the base oil properties. However, signal calibration and optimisation may be required to increase detection stability when measuring optically clear or low viscosity liquids.

CONCLUSION

This study successfully designed, built and tested a prototype experiment on a laboratory scale for a lubricant level monitoring system through IoT, providing information on measurement accuracy, system reliability and operational efficiency for applications, especially in the oil and gas industry. The proposed system has been designed to provide real-time and continuous monitoring of lubricant stock levels in the tank, while reducing the need for manual inspections and minimising human error. The system also delivers data consistently under different environmental conditions and fluid properties. Thanks to the combination of modern sensing technology and IoT connection protocols, there are effective means to facilitate.

Validation of the experiment conducted through data analysis showed that the system's capacity can maintain high accuracy, durability and adaptability to various lubricant parameters. The results also show that the proposed system is not only good at improving monitoring performance but can also help promote predictive maintenance techniques by allowing accurate decisions to be made in a timely and informed manner. This feature is essential for maximising resource utilisation, improving equipment availability, and reducing downtime in industrial environments. Furthermore, analysis

from real industry provides a complete picture of the main elements that drive system performance, laying the groundwork for future improvements and scalability.

As this research was limited to a laboratory-scale study, the long-term durability and continuous operational testing were beyond the scope of the present study. In real industrial applications, the system may be exposed to various operating conditions, including mechanical vibrations, temperature fluctuations, dust accumulation, oil mist contamination, and electromagnetic interference, all of which may potentially affect the sensor performance and measurement stability. Nevertheless, the non-contact measurement principle of the LiDAR sensor minimises mechanical wear and reduces direct exposure to the monitored lubricant that may contribute to improved operational longevity. For future research recommendation can focus on extended-duration testing under actual industrial operating conditions to evaluate sensor degradation, calibration drift, system reliability, maintenance requirements, and long-term proposed measurement consistency.

ACKNOWLEDGEMENT

This research was funded by the Putra Graduate Initiative Grant (GP-IPS/2021/9705900) from Universiti Putra Malaysia (UPM). Therefore, the researchers would like to thank the Sports Engineering and Artificial Intelligence Centre (SEA-IC) of the Faculty of Mechanical Engineering, Universiti Teknologi MARA, Shah Alam, for providing the research facilities and expert support, as well as the Fluid Laboratory of Universiti Putra Malaysia for providing the laboratory space and facilities necessary to conduct experimental testing and data collection activities. Finally, the research team would also like to thank everyone who contributed to this research directly and indirectly from start to completion.

LIST OF ABBREVIATIONS

IoT	:	Internet of Things
LiDAR	:	Light detection sensor
Wi-Fi	:	Wireless fidelity
MQTT	:	Message Queuing Telemetry Transport
IIoT	:	Industry Internet of Things
TTGO	:	refers to a popular series of ESP32-based development boards
ESP32	:	Espressif 32-bit Microcontroller
NodeMCU	:	Node MicroController Unit

REFERENCES

Ahlawat, P., & Rana, C. (2021). An era of recommendation technologies in IoT: Categorisation by techniques, challenges and future scope. *Pertanika Journal of Science & Technology, 29*(4), 2355-2381. https://doi.org/10.47836/pjst.29.4.07

Benziane, S. (2025). Exploring the synergy of intelligent systems: Cloud computing, industrial Internet of Things (IIoT), big data analytics, and neural networks. In *Applied neural networks in the AI era: From theory to real-world impact* (pp. 241-268). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3373-4571-0.ch009

Boddu, R. D., Ragam, P., Pendhota, S. P., Goni, M., Indrala, S., & Badavath, U. R. (2023). IoT-based smart agricultural monitoring system. In K. S. Raju, A. Govardhan, B. T. Rao, S. N. Mohanty, & S. B. Sriram (Eds.), *Proceedings of the fourth international conference on computer and communication technologies: IC3T 2022* (pp. 377-385). Springer. https://doi.org/10.1007/978-981-19-8563-8_36

Chen, J., Patra, J., Pradel, M., Xiong, Y., Zhang, H., Hao, D., & Zhang, L. (2020). A survey of compiler testing. *ACM Computing Surveys, 53*(1), Article 4. https://doi.org/10.1145/3363562

Epelle, E. I., & Gerogiorgis, D. I. (2020). A review of technological advances and open challenges for oil and gas drilling systems engineering. *AIChE Journal, 66*(4), Article e16842. https://doi.org/10.1002/aic.16842

Ercetin, A., Der, O., Akkoyun, F., Gowdru Chandrashekarappa, M. P., Şener, R., Çalışan, M., Olgun, N., Chate, G., & Bharath, K. N. (2024). Review of image processing methods for surface and tool condition assessments in machining. *Journal of Manufacturing and Materials Processing, 8*(6), Article 244. https://doi.org/10.3390/jmmp8060244

Hanaysha, J. (2016). Testing the effect of service quality on brand equity of the automotive industry: Empirical insights from Malaysia. *Global Business Review, 17*(5), 1060-1072. https://doi.org/10.1177/0972150916656656

Hatta Antah, F., Khoiry, M. A., Abdul Maulud, K. N., & Abdullah, A. (2021). Perceived usefulness of airborne LiDAR technology in road design and management: A review. *Sustainability, 13*(21), Article 11773. https://doi.org/10.3390/su132111773

Hu, L. P., Bao, X. L., & Wang, Q. (2011). The repetition principle in scientific research. *Journal of Chinese Integrative Medicine, 9*(9), 937-940. https://doi.org/10.3736/jcim20110903

Kaur, M. J., Mishra, V. P., & Maheshwari, P. (2019). The convergence of digital twin, IoT, and machine learning: Transforming data into action. In M. Farsi, A. Daneshkhah, A. Hosseinian-Far, & H. Jahankhani (Eds.), *Digital twin technologies and smart cities* (pp. 3-17). Springer. https://doi.org/10.1007/978-3-030-18732-3_1

Liu, C., Zhu, H., Tang, D., Nie, Q., Zhou, T., Wang, L., & Song, Y. (2022). Probing an intelligent predictive maintenance approach with deep learning and augmented reality for machine tools in IoT-enabled manufacturing. *Robotics and Computer-Integrated Manufacturing, 77*, Article 102357. https://doi.org/10.1016/j.rcim.2022.102357

Liu, Z., Fan, Q., Liu, J., Zhou, L., & Zhang, Z. (2024). Robust intelligent monitoring and measurement system for downhole dynamic liquid level. *Sensors, 24*(11), Article 3607. https://doi.org/10.3390/s24113607

Lo Castro, F., Ariza Alvarez, M. L., De Luca, M., & Iarossi, S. (2024). A review of vehicle speed measurement methods for the statistical pass-by noise testing. *INTER-NOISE and NOISE-CON Congress and Conference Proceedings, 270*(3), 8386-8397. https://doi.org/10.3397/IN_2024_4086

Makhesana, M. A., Bagga, P. J., Patel, K. M., Patel, H. D., Balu, A., & Khanna, N. (2025). Comparative analysis of different machine vision algorithms for tool wear measurement during machining. *Journal of Intelligent Manufacturing, 36*(7), 4567-4591. https://doi.org/10.1007/s10845-024-02467-3

Manish, R., Hasheminasab, S. M., Liu, J., Koshan, Y., Mahlberg, J. A., Lin, Y. C., & Habib, A. (2022). Image-aided LiDAR mapping platform and data processing strategy for stockpile volume estimation. *Remote Sensing, 14*(1), Article 231. https://doi.org/10.3390/rs14010231

Mast, T., Neighborgall, C., Taheri, M., Holton, C., & Ahmadian, M. (2020). Considerations for sensor selection for detecting top-of-rail (TOR) lubrication. In *ASME/IEEE Joint Rail Conference* (Paper No. V001T13A001). American Society of Mechanical Engineers. https://doi.org/10.1115/JRC2020-8028

Muneeswaran, V., Nagaraj, P., Akhila, A., Sudeepthi, L., Venkateswararao, B., & Advytha, M. S. (2023). Design of IoT-based heterogeneity and computing system. In *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1594-1597). IEEE. https://doi.org/10.1109/ICICCS56967.2023.10142365

Pimenov, D. Y., da Silva, L. R., Ercetin, A., Der, O., Mikolajczyk, T., & Giasin, K. (2024). State-of-the-art review of applications of image processing techniques for tool condition monitoring on conventional machining processes. *The International Journal of Advanced Manufacturing Technology, 130*(1), 57-85. https://doi.org/10.1007/s00170-023-12679-1

Quinchia, L. A., Delgado, M. A., Reddyhoff, T., Gallegos, C., & Spikes, H. A. (2014). Tribological studies of potential vegetable oil-based lubricants containing environmentally friendly viscosity modifiers. *Tribology International, 69*, 110–117. https://doi.org/10.1016/j.triboint.2013.08.016

Rashidifar, R., Bouzary, H., & Chen, F. F. (2022). Resource scheduling in cloud-based manufacturing system: A comprehensive survey. *The International Journal of Advanced Manufacturing Technology, 122*(11), 4201-4219. https://doi.org/10.1007/s00170-022-09873-y

Schneider, D., Shrotri, A., Flatt, H., Stübbe, O., Wolf, A., Lachmayer, R., & Bunge, C. A. (2021). Impact of industrial environments on visible light communication. *Optics Express, 29*(11), 16087-16104. https://doi.org/10.1364/OE.421757

Shi, X., He, S., Xie, X., & Sun, Y. (2021). Review on feature extraction of lubrication and wear fault diagnosis in tribology systems. *Tribology, 43*(3), 241-255. https://doi.org/10.16078/j.tribology.2021066

Sparham, M., Sarhan, A. A., Mardi, N. A., & Hamdi, M. (2014). Designing and manufacturing an automated lubrication control system in CNC machine tool guideways for more precise machining and less oil consumption. *The International Journal of Advanced Manufacturing Technology, 70*(5-8), 1081-1090. https://doi.org/10.1007/s00170-013-5325-y

Tanati, M. K., & Ponnusamy, M. (2023). Real-time investigation of LoRaWAN architecture by the LoRa communication module and the ultrasonic sensor. In *2023 Innovations in Power and Advanced Computing Technologies (i-PACT)* (pp. 1-5). IEEE. https://doi.org/10.1109/I-PACT58649.2023.10434398

Varalakshmi, S., Sajitha, M., Pandi, V. S., Ravi, S., Giri, P., & Samhan, A. A. A. (2025). Improving industrial IoT systems with AI to support smart factory operations, predictive maintenance, and real-time monitoring. In the *2025 Global Conference in Emerging Technology (GINOTECH)* (pp. 1-6). IEEE. https://doi.org/10.1109/GINOTECH63460.2025.11077046

Wilińska, I., & Wilkanowicz, S. (2025). Advancements in environmentally friendly lubricant technologies: Towards sustainable performance and efficiency. *Energies, 18*(15), Article 4006. https://doi.org/10.3390/en18154006

Xin, C., Xu, Y., Zhang, Z., & Li, M. (2024). Micro-opto-electro-mechanical systems for high-precision displacement sensing: A review. *Micromachines, 15*(8), Article 1011. https://doi.org/10.3390/mi15081011

Zhang, X., Zhen, D., Feng, G., Zhou, Z., Liang, X., Zhang, H., & Gu, F. (2025). Investigation on acoustic emission characteristics of lubricating grease in a hydrodynamic lubrication regime. *Tribology Letters, 73*(2), Article 59. https://doi.org/10.1007/s11249-025-01995-0

Pertanika J. Sci. & Technol. 34 (3): 1769 - 1800 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Human-in-the-Loop Soft Prompt Tuning for Adverse Drug Event Extraction from Clinical Notes

Salisu Modi[1], Khairul Azhar Kasmiran[2]*, Nurfadhlina Mohd Sharef[2], and Mohd Yunus Sharum[2]

[1]Department of Computer Science, Faculty of Computing, Sokoto State University, 840101 Sokoto, Sokoto State, Nigeria

[2]Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

ABSTRACT

Adverse drug events (ADEs) are significant healthcare issues, increasing costs and medication durations. Extracting ADE information is crucial for enhancing healthcare delivery and drug safety. Transformer-based models' performance has recently improved through soft prompt tuning in the ADE task. Current ADE extraction models require large, annotated datasets and offer limited interpretability for clinical use. In addition, the complexity and black-box nature of transformer-based models limiting incorporating human preference and make interpretation difficult, which are critical in healthcare, hindering its full adoption. To mitigate this problem, this research proposes a human-in-the-loop learning method to tune transformers with domain expert input to complement limited data and incorporate expert preference. The method evolves in two phases: initially soft prompt tuning the model for multi-task learning of dual sequence labelling, ADE extraction with an additional soft prompt that guides the model, followed by iterative prediction, validation, and feedback refinement to optimise the model. Visualisation techniques display model predictions to users, enhancing understanding. The self-attention weights aid in diagnosing and explaining the model using saliency maps and attention flow diagrams. A graphical user interface allows experts to provide corrective labels for misclassified samples, thus refining the model. The result of the final model evaluated on testing sets from TAC 2017 and N2C2 2018 datasets, achieving state-of-the-art performance of 0.9404 and 0.9132 for N2C2 concept and relation extraction and 0.8723 and 0.5506 for TAC 2017 concept and relation extraction. In conclusion, this research

ARTICLE INFO

Article history:
Received: 25 November 2025
Accepted: 07 June 2026
Published: 25 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.21

E-mail addresses:
gs63125@student.upm.edu.my (Salisu Modi)
k_azhar@upm.edu.my (Khairul Azhar Kasmiran)
nurfadhlina@upm.edu.my (Nurfadhlina Mohd Sharef)
m_yunus@upm.edu.my (Mohd Yunus Sharum)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

demonstrates the effectiveness of incorporating human feedback to improve model performance in complex scenarios, demonstrating its effectiveness in improving drug safety surveillance and drug monitoring.

Keywords: Adverse drug event, drug safety surveillance, human-in-the-loop learning, optimisation, soft prompt tuning

INTRODUCTION

Extracting adverse drug events (ADEs) from narrative documents has been a topic of interest in recent years. The goal is to identify the drug entities and their related attributes and then determine the causal relationships between the drug and the adverse effects caused by the drug. This task is significant in improving healthcare delivery and drug safety surveillance. The increasing rate of ADE cases reported worldwide is becoming worrisome as it increases medication costs and causes damage to public health (Xia, 2022). Most ADE cases are preventable when appropriate measures are taken (Ong et al., 2024). The early stage of monitoring and investigating the side effects of a drug is through clinical trials with volunteer patients before the approval and marketing of new drug products. This approach is ineffective due to its shorter duration and lack of volunteer patients, especially with drugs with longer latency (Xia, 2022). Consequently, spontaneous reporting becomes the next option to complement the earlier approach at the post-marketing stage of drug usage. The spontaneous reporting system (SRS) relies on the clinician's assumption during the patient's medication period or the report from the affected patient. This approach is characterised by underreported cases and incompleteness (Kim et al., 2023).

With improved performance, large language models (LLMs) based on transformer neural networks have been widely adopted for various ADE extraction tasks (Gu et al., 2023; Jamil et al., 2026; Ren & Wang, 2023; Sahoo et al., 2024). This can be attributed to their transfer learning capabilities, which enhance downstream-specific performance on downstream tasks. Adapting pre-trained language models (PLMs) through fine-tuning leaves an objective gap between pre-training and downstream task objectives and also provides a memory-efficient approach. The soft prompting mitigates this gap by integrating task-specific trainable prompts to guide the model on the downstream task (Tian et al., 2024). However, this requires a huge amount of annotated data from the downstream task. Healthcare-related data are primarily scarce and restricted due to privacy constraints and a lack of domain expertise, consistent annotations.

Moreover, transformer-based models are black-box and complex. They are composed of complex architectures with billions or trillions of trainable parameters. The models are trained on large volumes of data, making interpreting and diagnosing their decision-making processes challenging. Adverse drug event extraction involves establishing causality

relationships between the drug entities and the adverse effects caused by their usage during patient treatments. This type of connection could sometimes involve complex, ambiguous and polysemous entities that would otherwise be difficult for a machine-learning model to handle accurately. Incorporating human preferences through domain experts into the modelling process could improve model performance and reduce time-consuming upfront data annotations. However, most current approaches treat humans as passive participants (i.e. sending input and receiving output processed by the model) during model development (Gómez-Carmona et al., 2024). On the one hand, an interactive human-in-the-loop learning approach is a form of learning where models are developed with human involvement throughout the process. In this approach, end-users actively contribute to model creation by continuously providing training parameters, reviewing model outputs, and offering feedback on intermediate results to improve performance (Wondimu et al., 2022).

The black-box nature of the LLM models has also been a significant concern for an effective clinical natural language processing (NLP) model for ADE extraction. Incorporating human factors, model interpretation, and explanations in model development is crucial for improving model reliability and capability in handling complex situations (Zitu et al., 2025). Interactive human-in-the-loop learning (IHITLL) enables the incorporation of human judgment and feedback into PLM's model adaptation for ADE extraction, thereby mitigating the risk of opaque decision-making in the model (Wondimu et al., 2022).

This work proposes a novel approach that combines transformer-based model soft prompt tuning with interactive human-in-the-loop interfaces to optimise the model's performance tailored for a multi-task sequence labelling of ADE concept and relation extraction.

This paper introduced the human-in-the-loop process to this task to enable humans to understand, interpret and modify the model that extracts ADE cases from unstructured clinical documents. The work designed interfaces to support model interpretation, diagnosis and refinement in model development. The approach first tunes the model using a soft-prompt tuning approach on the training set. Further, it optimises the model with feedback based on the prediction outcome from the sample validation data. The method leverages visualisation techniques to display the model prediction result and utilises the self-attention weight generated by the model to interpret and diagnose the model decision markings. The proposed approach was experimented on several transformer-based models, such as BERT and SCIBERT, as a baseline. All the experimented models have shown an improvement compared to traditional fine-tuning, indicating the generality of the proposed hybrid approach.

In summary, the objectives of this research are as follows:

1. To propose a novel, interactive human-in-the-loop soft prompt tuning of the transformer model method with three levels of interaction: model understanding, model diagnosis

and model refinement, for improving model interpretation and explanations.

2. To experiment with the proposed approach to multi-task dual sequence labelling of adverse drug event extraction tasks of concepts and relations on two public datasets, N2C2 2018 and TAC 2017.
3. To improve the ADE concept and relation extraction F1-score using human feedback-based soft prompt tuning.

The architecture of PLMs based on transformers is inherently complex. It features multiple nested attention layers and processing blocks that perform numerous linear and non-linear transformations on input sequences to generate output (Vaswani et al., 2017). This intricate structure makes the internal workings of these models challenging to interpret, resulting in opaque decision-making processes (Budd et al., 2021). Integrating human factors into the decision-making process during model development and deployment is crucial, especially for healthcare-related tasks such as ADE extraction.

Interpreting neural network-based models has been a significant concern for effectively developing practical systems for NLP tasks (Accenture, 2019). Enhancing model understanding and mitigating the risks associated with the black-box nature of PLM is an ongoing area of research. Various methods have been developed to improve the explainability of neural network-based models, for instance Kim & Kim, (2022) proposed Shapley Additive exPlanations (SHAP). This facilitates incorporating human factors into the human-in-the-loop process of developing intelligent systems for NLP tasks. Techniques such as Active Learning (AL) (Riesener et al., 2024; Shelmanov et al., 2021), Reinforcement Learning with human-in-the-loop learning (RLHILL) (Bai et al., 2022; Ouyang et al., 2022), Explainable AI (XAI), and IHITLL (Zhao & Liu, 2021) have been applied to various NLP tasks.

Moreover, human feedback can optimise LLM's performance on different downstream tasks. An example is the success of OpenAI's InstructGPT (Ouyang et al., 2022), which incorporated user feedback and intent into its GPT-3.5 model using reinforcement learning from human feedback (RLHF). The initial model had 1.3 billion parameters and was fine-tuned using a feedback corpus to include helpfulness, honesty, and harmlessness. Based on comparisons by human expert evaluators, InstructGPT was preferred over GPT-3.5, which had 175 billion parameters.

The approach has also been explored by Thoppilan et al., (2022), that proposed LaMDA, which fine-tuned LLMs to be interesting, helpful, and factually accurate using supervised learning techniques that combine generative and discriminative approaches. Similarly, (Bai et al., 2022) applied preference modelling and RLHF to fine-tune LLMs, aligning the models to be helpful, harmless, and honest. This approach involves iteratively updating the model with feedback from online crowd workers.

Recently, many parameter-efficient fine-tuning approaches have been promising to make language model adaptation more efficient, and they each have distinct characteristics. Adapter-based fine-tuning (Houlsby et al., 2019) introduces inference latency, as the adapter layers are executed sequentially, unlike the original pre-trained transformer layers executed in parallel. LoRA (Hu et al., 2021) and its variants approximate the pre-trained weights with two matrices, requiring alterations to the structure of the pre-trained model. The soft prompting (Liu et al., 2022; Li & Liang, 2021) It is a lightweight and straightforward approach that requires updating only a few prompt parameters without altering the core parameters of the pre-trained model. The study by Peng et al., (2024) Driven further developments in ADE extraction tasks by adapting a deep prompt-tuning approach to compare four learning strategies: fine-tuning, hard prompting, soft prompting with a frozen model, and soft prompting with an unfrozen model. This system utilised GatorTron clinical LLMs to identify ADE concepts and perform end-to-end RE from clinical documents.

Despite the success of these approaches in optimising LLM performance for various NLP tasks, most existing methods are based on generative AI models using decoder-only transformer architectures for Natural Language Generation (NLG) tasks. The Natural Language Understanding (NLU) tasks, such as ADE extraction, have not fully utilised these promising approaches to improve LLM adaptation. In the active learning model, uncertain samples are to be annotated by an oracle or human expert (Quennelle et al., 2025). In RLHF, human feedback shapes reward signals for reinforcement learning, and human judgment guides the model optimisation (Bai et al., 2022). The interactive Machine learning involved a continuous loop of training with user input, while the human iterative refine output (Teso et al., 2023). The proposed IHITLL framework involves iteratively collecting domain expert feedback based on model predictions, then utilising the feedback to refine the model weight on few iterations.

Additionally, while RLHF is the most prominent approach for incorporating human preferences into model fine-tuning, IHITLL has not been extensively explored. Furthermore, current approaches have placed less emphasis on interpreting models to enhance human understanding of their opaque decision-making process. The effectiveness of IHITLL in the domain of ADE extraction is still uncertain. Therefore, it is essential to explore this approach using PLM. Incorporating human feedback into model training can reduce the need for manual labelling and address complex data situations in the ADE concept and relation extraction subtasks through real-time human feedback. To the knowledge of the author, no study explores IHITLL for multi-task ADE extraction tasks, which this study aimed to investigate through soft prompt tuning of PLMs. This paper presents an interactive human-in-the-loop learning method designed to optimise model performance by incorporating human feedback in the model tuning on a token-level sequence modelling task for ADE extraction.

METHODOLOGY

This section details the proposed hybrid method of interactive human-in-the-loop multi-task learning tuning of a transformer-based model to optimise its performance on multi-task sequence labelling ADE extraction tasks. The key novelty of the proposed method is introducing interactive learning with three levels of interaction to tune a transformer-based model. Three levels of user interaction with the model and data are provided to enhance user understanding of model predictions, diagnosis and refinement, with a user interface for model prediction understanding, a user interface for model interpretation, and an interface for interactive model refinement. The overall flow of the procedure is depicted in Figure 1, which comprises two phases. The initialisation phase is where the transformer model is initially tuned using the soft prompt tuning approach. The interactive model optimisation is where the initiated model is tested and iteratively improved from sample data by the domain expert.

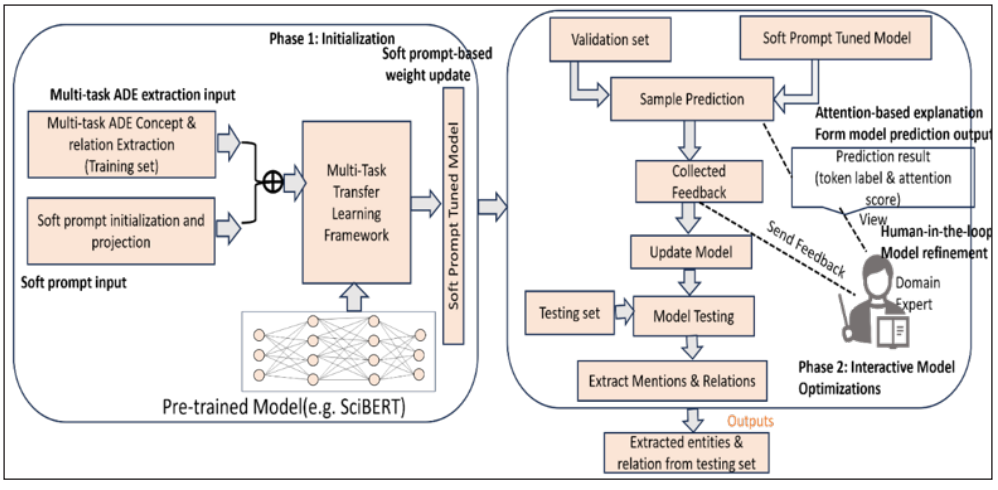


Figure 1. IHITLL-based fine-tuning framework

For more effective design and implementation of interactive human-in-the-loop methods, two human-in-the-loop concerns are as follows:

- Who is the human-in-the-loop, and what role do they play? This research utilises datasets provided by TAC 2017 and N2C2 2018 ADE extraction challenges. The task was handled as sequence labelling, where the model is trained to predict a corresponding label to each token within the sequence. Therefore, the human-in-the-loop can be a clinical expert or data scientist developing and training the model based on the pre-annotated dataset provided by the organisers. The role of the human is to provide corrective feedback based on model prediction according to the gold standard annotation through an interactive interface.

- How to mitigate model and human bias during the interactive model optimisation: The dataset was bifurcated into training and testing sets. For the interactive model tuning, 10% of the training set is used to optimise the model through predictive and refined cycles. These datasets have been carefully annotated based on the annotation guidelines provided by the challenges organisers, detailed in (Demner-Fushman et al., 2018; Henry et al., 2020). During the model optimisation process, corrective feedback is provided only from the gold-standard annotations, thereby avoiding incorporating unverified human factors into the model refinement. The rationale for selecting 10% of the training data for interactive refinement is to allow the model to be initially fine-tuned on a larger portion of the dataset, thereby acquiring domain knowledge, while simultaneously reducing the annotation burden on the human-in-the-loop expert.

Problem Formulation

The task of ADE extraction involves two sub-tasks: concept and relation extraction. The process starts with identifying named entities within a sequence of tokens named concepts, denoted as $S = \{t_1, t_2, \ldots, t_n\}$, where n represents the sequence length. The goal is to extract a set of positive entities $E = \{e_1, e_2, \ldots, e_i\}$ where each entity $e_i \in E$. has one or more relationships with other entities or entities in E, and a corresponding set of relations $R = \{r_1, r_2, \ldots, r_j\}$.

Recently, many parameter-efficient fine-tuning approaches have been promising to make language model adaptation more efficient, and they each have distinct characteristics. Adapter-based fine-tuning (Quentin, 2019) introduces inference latency, as the adapter layers are executed sequentially, unlike the original pre-trained transformer layers executed in parallel, LoRA (Wallis, 2021) and its variants approximate the pre-trained weights with two matrices, requiring alterations to the structure of the pre-trained model. In contrast, soft prompting is a lightweight and straightforward approach that requires updating only a few prompt parameters without altering the core parameters of the pre-trained model (Chen et al., 2024).

The soft prompt tuning method prepends task-specific instructions as a learnable prompt, instructing the pre-trained model on downstream task objectives. The aim is to minimise the multi-task loss together with the additional learnable prompt, as in Equation 1. Here, the model parameters are updated, as shown in Figure 2b.

$$L_{\text{MTL-ADE}}\left(\theta_{\text{pre-trained}}, \theta_{\text{tasks}}, \theta_{\text{prompt}}\right) = \sum_{t=1}^{T} \delta t \left(-\sum_{n=1}^{N_t} \sum_{r=1}^{R_t} y_{nc}^{(t)} \log P_{nc}^{(t)}\left(\theta_{\text{pre-trained}}, \theta_{\text{tasks}}, \theta_{\text{prompt}}\right)\right) \quad [1]$$

For parameter-efficient soft prompt tuning (frozen model), instead of updating the pre-trained parameters $\theta_{\text{pre-trained}}$, only the task-specific parameters θ_{tasks} and the additional θ_{prompt} trainable vectors generated from the embedding of the prompt token of the input sequence

are updated. The overall multi-task loss in Equation 2 is minimised by updating only the θ_{tasks} and θ_{prompt} parameters while leveraging the knowledge of the pre-trained model during the fine-tuning process, as shown in Figure 2a.

$$L_{MTL\text{-}ADE}(\theta_{tasks}, \theta_{prompt}) = \sum_{t=1}^{T} \delta t \left(- \sum_{n=1}^{N_t} \sum_{r=1}^{R_t} y_{nc}^{(t)} \log P_{nc}^{(t)}(\theta_{pre\text{-}trained}, \theta_{tasks}, \theta_{prompt}\right) \quad [2]$$

Where $\theta_{tasks}=[\theta_{ner}; \theta_{re}]$ and $\theta_{prompt} = [\theta_{ner\text{-}prompt}; \theta_{re\text{-}prompt}]$ and N_t and R_t are the numbers of tokens for the ADE-concept and ADE-relation input sequences. $y_{nc}^{(t)}$ is the true label for the n-th token ADE-concept and ADE-relation, and $P_{nc}^{(t)}$ is the predicted probability of the n-th token ADE-concept and ADE-relation. Here, all the parameter sets ($\theta_{pre\text{-}trained}$, θ_{tasks}, θ_{prompt}) are learnable and have been adjusted to minimise the multi-task loss.

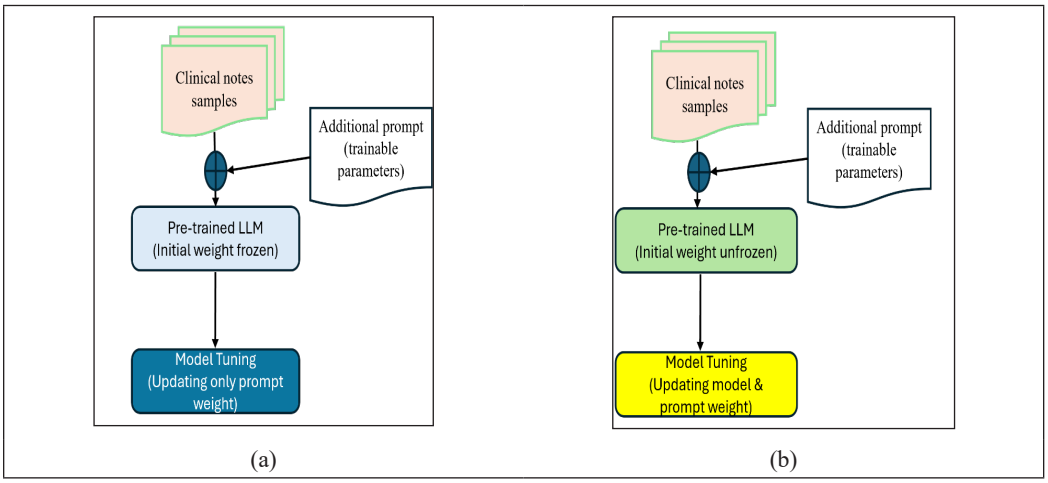


Figure 2. Soft prompt model tuning strategies. (a) Frozen model tuning and (b) unfrozen model tuning

Datasets

The N2C2 2018 dataset (Henry et al., 2020), derived from clinical narratives, was used for the ADE extraction challenge. This dataset contains annotations for nine entities: drug, strength, form, dosage, frequency, route, duration, reason, and ADE entities, all linked to a drug entity as their source. There are eight possible relations between these entities. The proposed model and methods were trained and evaluated using the official dataset splits, which consist of 303 training records and 202 testing records.

The TAC 2017 dataset (Belousov et al., 2019) consists of 200 drug labels in XML format, divided into a training set of 101 labels and a test set of 99. The dataset contains five attributes related to Adverse Drug Reactions (ADR): Animal, Drug Class, Factor,

Negation, and Severity. Moreover, the dataset has three types of relationships: Effect (linking severity to ADR), Hypothetical (linking animal, drug class, or factor mentions to ADR), and Negated (linking negation or factor mentions to ADR)

Prompt Initialisation and Projection

To ena ble the adaptation of LLM models on the downstream token-level ADE extraction tasks with task-specific objectives in the tuning process, a learnable prompt is prepended to the input embeddings generated by the embedding layer of the model, as in Equation 3. The procedure initially used a textual prompt and then converted it to a vector as in Equation 4.

$$E=\text{Embed}(X)\in R^{Bz*L_X*d_{model}} \quad [3]$$

Where X is the input sequence, Lx is the length of the input sequence, and Embed is the transformer embedding layer.

$$P \in R^{(1,0)*L_p*d_{model}} \quad [4]$$

Where P is the prompt vector, L_p is the prompt length and d_{model} The model dimension is 768 for the BERT-based models. The prompt is then projected through the linear transformation, as in Equation 5.

$$P= W_p P + b_p \quad [5]$$

Where W_p is the projection weight, and b_p is the bias term. The projected prompt is broadcast to the batch size (B_z) as in Equation 6.

$$P'=\text{broadcast}(P) \in R^{Bz*L_P*d_{model}} \quad [6]$$

One major modification to the original model setting is expanding the maximum sequence length. The experimental PLMs use a fixed sequence length (512 for the BERT-based models). To enable the model to process the broadcast soft prompt embeddings as in Equation 6, the method extended the model's maximum input sequence length to accommodate the combined input sequence as in Equation 7. Consequently, the attention mask in Equation 8, token type IDs in Equation 9, and sequence labels were also padded to match the new sequence length as in Equation 10. The procedure for creating the extended input is outlined in Algorithm 1, presented in Figure 3. Only the prompt projection parameters (Wp, bp) gradient are updated while the model parameters remain frozen.

$$C_{input}=[P';E] \in R^{Bz*(Lp+Lx)} \quad [7]$$

$$M = \left[1^{Bz*Lp}; Mx\right] \in \{0,1\}^{Bz*(Lp+Lx)} \quad [8]$$

$$T = \left[0^{Bz*Lp}; Tx\right] \in \{0,1\}^{Bz*(Lp+Lx)} \quad [9]$$

$$Y = \left[-1^{Bz*Lp}; Yx\right] \in Z^{Bz*(Lp+Lx)} \quad [10]$$

Algorithm 1: Procedure for Soft Prompting Input and Projection

Input: Task Input data, Top-k Prompt Tokens feature select, List of Prompt Template, Pre-trained Model, Number of Task

Steps:

For task in Number of Task **Do:**

1. *Pmt1* ← ***SelectPrompt****(Pmt, c_task_id)* #Get the prompt for the task from the list
2. *Pmt_tokenized* ← ***Tokenizer****(Pmt1)*
3. *Pmt1_emb* ← ***Embedding***(Pmt_*tokenized*) #embed the tokenized prompt using pre-trained model
4. *tokenized_input* ← ***Tokenizer(task_*Input data***) # tokenization
5. *X__input* ←***Embedding***(*tokenized_input*) #Get the embedding of the input data
6. *X_con_input* ←***[Pmt1_emb ;X__input]*** #Concatenate the input and soft prompt
7. *P_input*← ***EmbeddingSpace***(*X_con_input*) #Projection Layer
8. *Mask* ← ***Expand*** *(mask, 1's, top-k*) #Expand the attention mask of batch input to the length of top-k prompt fill with 1's
9. *Pmt_label* ←***fill****(-1, len(top-k))*
10. *Task_ids* ← ***Expand*** *(task_ids, 0's, top-k*) #Expand the tasks ids of batch input to the length of top-k prompt fill with 0's
11. *Input_labels* ← *[X_input[label], pmt_lable]* #Expand the labels of batch input to the length of top-k prompt fill with -1's
12. *Extended_input* ← *[P_input, Mask, Task_ids, Input_label]* *# combine*

end for

Output: Extended_input

Figure 3. Procedure for soft prompt initialisation and projection

Multi-task Model Soft Prompt Tuning on ADE Extraction

The task is presented as a dual sequence labelling problem for the Adverse Drug Event (ADE) concept and relation extraction at the token level. Each token in the input sequence is assigned a corresponding label. An extended BIO tagging scheme, inspired by (El-Allaly et al., 2021) is used to assign labels to tokens in the sequence for each task. In the first task, all Adverse Drug Reactions (ADR) mentioned in drug labels and drug names in the clinical notes are initially identified to create a set of positive entities with at least one relation with mention attributes. In the second task, attributes of the initially identified positive entities are determined, followed by linking the relations between the entities and their attributes. To guide the Language Model (LM) on the downstream task, a soft

prompt tuning approach is employed using multiple prompts, one for each task, as detailed in the previous sections. A multi-task learning framework (Liu et al., 2019) is utilised to model the two tasks simultaneously. The tasks are input at the input layer, and the PLM generates a shared representation, which is then passed to the fully connected layer for final token-level classification. This research investigates the capabilities of small-scale LLMs; as such, the models' parameters are left unfrozen to leverage the transfer learning abilities of the model for improved performance with minimal resources.

Human-in-the-loop Interactive Model Optimisation

During the interactive optimisation process, the initialised best model obtained from soft prompt tuning (unfrozen) is subjected to sample data from the validation set, one task at a time. The model makes an initial prediction on the sample data to predict a label for each token in the sequence. The model output is displayed to the user via an interactive interface using an alignment plot diagram, which shows the token, the label predicted by the model for the token, and the gold standard label of the token side by side for the human-in-the-loop. A typical example is shown in Figure 6. Additionally, the model's attention weight assigned to each feature, which are the tokens in the sequence, indicates the attention given by the transformer self-attention mechanism to each token that influences the model's decision-making. These attention weights are displayed using attention saliency heatmaps and other visualisation methods to improve the human-in-the-loop's understanding of model prediction results and decision-making processes. The human-in-the-loop provides feedback based on model prediction and diagnosis. An interactive interface allows the human-in-the-loop to accept model predictions and proceed to the next instance or provide corrective feedback to the misclassified token via JSON file upload or through manual labelling from the drop-down list of gold standard annotation labels. This process continues iteratively until all validation instances are executed.

Fine-tuning transformer-based models involves updating the initial pre-trained model parameters to adapt the pre-trained knowledge to the new task. This task requires saving the updated parameters for inference. To handle this constraint and avoid continued perturbation of model parameters for each instance, a list is created for each task that stores all the corrective feedback during the iterations. When all the validation instances are exhausted, the model is retrained to incorporate human feedback for a few epochs, saving model checkpoints at each epoch to mitigate catastrophic forgetting by reverting to the previous checkpoint whenever forgetting occurs. Further training for a few epochs prevents the model from overfitting the small sample from human-in-the-loop feedback. In this research, an optimal performance was achieved with 2 epochs. The final refined model is saved for inference.

The saved model is loaded for testing on the test set during inference. The model makes predictions on the test sample. The predicted ADE concepts and relations are then extracted using the extraction module. An official evaluation script provided by the challenge organisers is used to evaluate the performance of the final model. The overall procedure is given in Algorithm 2 in Figure 4**.**

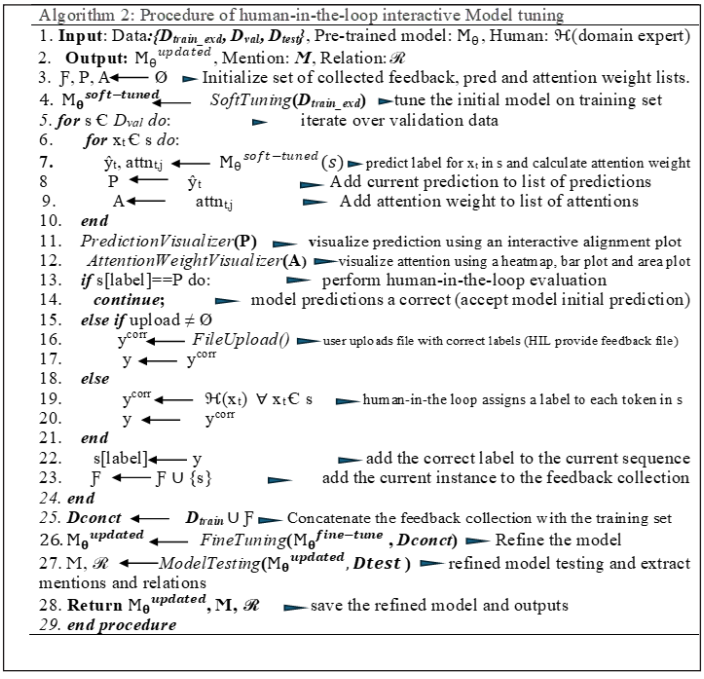


Figure 4. Procedure for human-in-the-loop optimisation

Interactive Interfaces

Interactive machine learning (IML) has recently become a popular approach to achieving human-machine collaboration, a concern of Industry 5.0 (Kumar et al., 2024). However, the two key questions are: How will the human-in-the-loop of model development understand the prediction results provided by the model and how to diagnose the model decision to comprehend the cause of misclassification? To address this concern, the proposed method encompasses two primary avenues for effective interaction between the user and the model development process. On the one hand, the human-in-the-loop component provides continuous feedback with correctional labels to the misclassified samples throughout the model refinement process. On the other hand, the interaction visualisation component allows for human-in-the-loop and model exchange. To this end, three visual interfaces are designed for model explanations and refinement. Firstly, to improve user understanding of the model prediction result, the method utilises an interactive alignment plot to effectively

display each token with its corresponding prediction label, with a hover over the actual label. This enables even non-technical experts to understand the result of the model prediction. Secondly, the method utilises the transformer model self-attention weight that indicates the importance of each token to the overall model decision-making to predict output for interpreting and diagnosing model decision-making. Lastly, the work designed an interface to enable users to interact and provide corrective feedback for model refinement. Details of each interface are given in the subsequent sections.

The diagram in Figure 5 shows a typical interaction for concept extraction subtasks, showcasing the sequence of activities to accomplish model predictions and refinement. Figure 6 shows the main user interface for the system interaction with a human-in-the-loop.

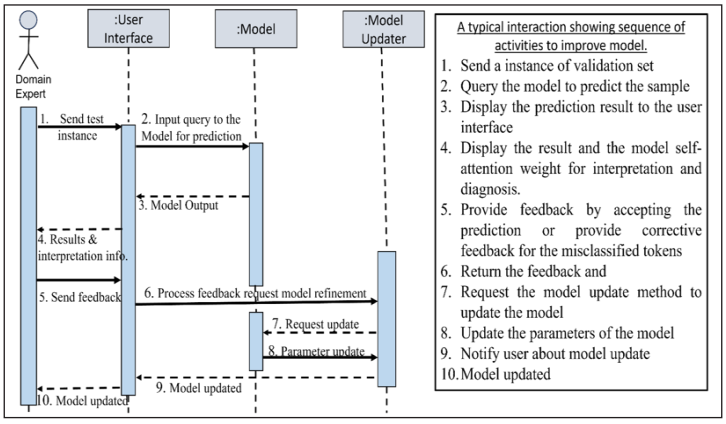


Figure 5. Sample UML interaction diagram showcasing the task of single instance concept extraction

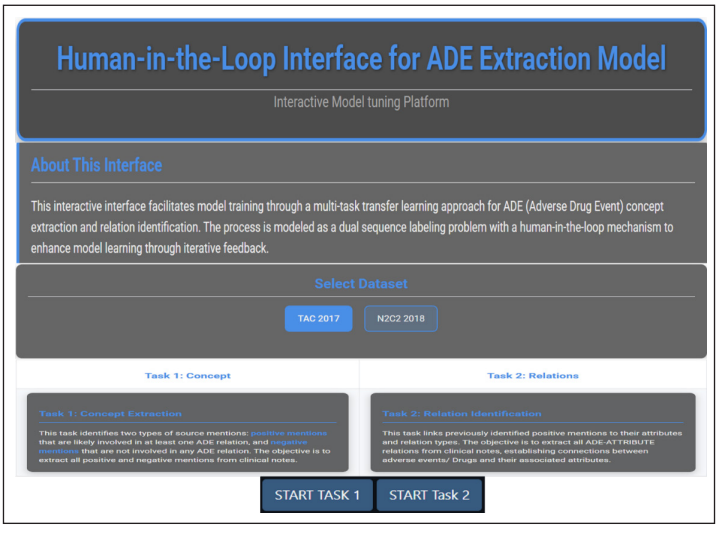


Figure 6. Main page of the interactive system

Model Prediction Understanding Interface

The success of explanations in an IHITLL-based model depends mainly on how well humans can comprehend and interpret their predictions. Inadequate explanations may confuse users and introduce unintended bias during the model refinement (Kumar et al., 2024). Sequence labelling involves training the model to predict a corresponding label for each token within the input sequence, as in Equation 11. Each token is displayed with its corresponding predicted label, as in Equation 12.

$$X= (x_1,x_2……..x_n) \quad [11]$$

Where x_1 is the i^{th} token in the sequence, and n is the number of tokens in the sequence.

$$\hat{y}_n=\text{argmax } (Py_n|x) \quad [12]$$

where $P(y_n|x)$ is the probability distribution over labels for the token x_n given the entire sequence x, using a prediction function defined as $\hat{y}_n=\boldsymbol{f}(x_t;\theta)$ where $\boldsymbol{f}(\cdot;\theta)$ is the model function parameterised by weights θ.

To effectively increase the understanding of model results, especially for non-medical experts, the method uses a Plotly Python library to design an interactive alignment plot diagram sample shown in Figure 7 displaying each token aligned to its corresponding predicted and actual label side-by-side. This gives a granular insight into the model's prediction results, allowing a fine-grained understanding of the prediction. The work employed a colour-coding scheme to differentiate between the correctly predicted tokens (designated by green) and the misclassified tokens (designated by red) as defined in Equation 13.

$$\text{Colour}_t = \begin{cases} \text{green, if } \hat{y}_n = y_n \\ \text{red, if } \hat{y}_n \neq y_n \end{cases} \quad [13]$$

When, $\hat{y}_n$ is the predicted label, and y_n is the true label

Hover text and colour coding present interpretable feedback to the user, enabling even the non-technical expert to comprehend model predictions intuitively. Hovering over predicted and actual labels for each token allows the user to evaluate the model's performance for a given instance quickly. Figure 7 depicts an example outcome for the sequence input “intrathecal methotrexate hists”.

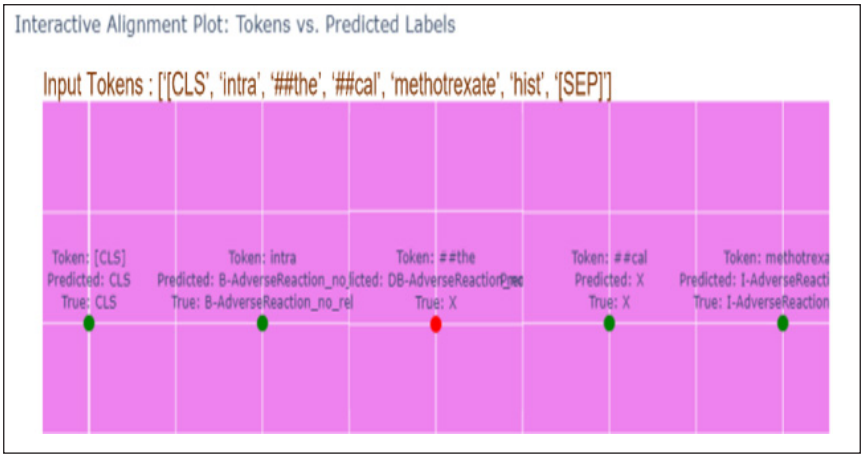


Figure 7. Sample Interactive Alignment Plot diagram for model prediction understanding

Transformer Self-attention Weight-based Model Diagnosis Interface

In sequence labelling tasks, each word or sub-word is a feature for the model's decision-making process. During the contextual representation of the input sequence, the transformer model utilises self-attention mechanisms to determine the importance of each feature within the input sequence. The self-attention mechanisms calculate the score for each token, showing how much attention a particular input should pay to other elements in the given sequence, irrespective of their relative distance. In essence, the transformer self-attention compares all input sequence members with each other and assigns importance, as in Equations 14 and 15.

$$\alpha_{t,i} = \frac{\exp(\text{score}_{t,i})}{\sum_{j=1}^{T} \exp(\text{score}_{t,j})} \quad [14]$$

where $\alpha_{t,i}$ represents the attention weight assigned to token x_i when predicting the label for token x_t, and $\text{score}_{t,i}$ is the attention score, calculated as:

$$\text{Score}_{t,i} = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad [15]$$

where Q, K and V are obtained from the linear transformation of input(X) embeddings.

Expert feedback is considered a feature that the model is backpropagated by the update methods, as in Algorithm 2. The soft prompt weights are appended to the feature representations so that they can be backpropagated jointly. At the classification layer, the

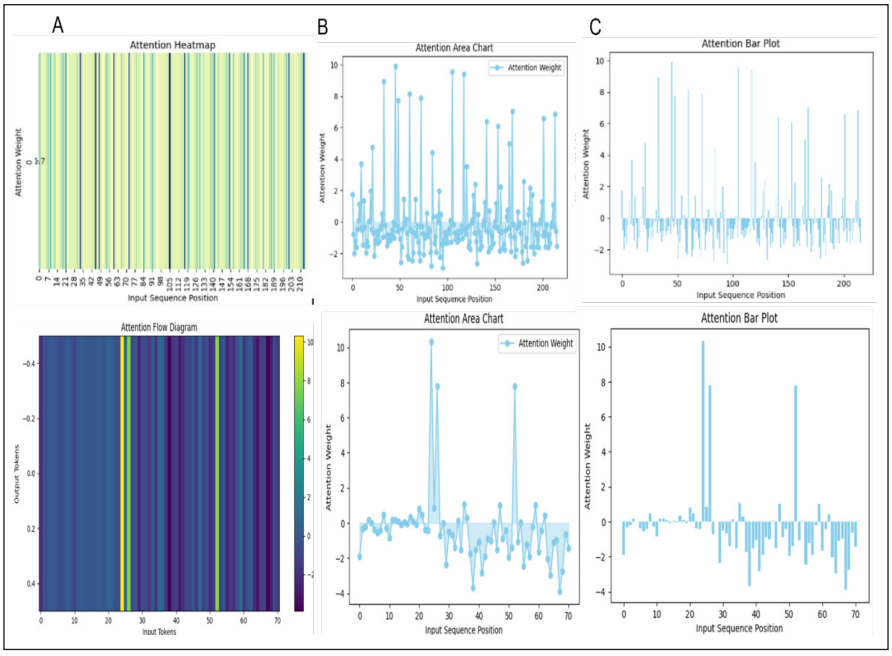


Figure 8. Sample attention interpretation for some input sequence. (a) Attention heatmap/flow (b) Attention area plot (c) Attention bar plot

soft prompt tokens are excluded from the softmax operation, as they are assigned negative label values. The weight update strategy modifies the trained model parameters for a limited number of iterations to incorporate human feedback. To bring the model transparency and interpretation to the human-in-the-loop, the work leverages the attention weight assigned to each token by the transformer self-attention mechanism to visualise its feature importance, thereby explaining the model diagnosis. Visualising each token's importance score helps the user understand which part of the input contributes significantly to the final model prediction, thereby enabling the user to ascertain whether a correct emphasis is given to each token or requires further adjustment of the model decision-making.

Although SHAP provides feature importance scores, it is computationally expensive to combine with transformer-based models and is less intuitive for textual data. In contrast, LIME relies on random perturbations, which may lead to unstable or unreliable explanations (Qian et al., 2023). Both methods are more suitable for post-hoc model interpretation, whereas attention-based approaches offer real-time explanations that can be directly leveraged by human experts.

To this end, the method utilises various alternative visualisation tools, including an attention saliency heatmap, attention flow heatmap, attention bar plot and attention area plot to display the attention weight for the input sequence -- typical examples are shown in Figure 8. The attention heatmap (a) uses a colour-coding scheme with a score to indicate

the relevant token: the darker the colour, the higher the importance score, and the relevance assigned to the input position, the lighter the colour. The bar plot (c) and area plot (b) showcase the relevant percentage of each input position in the sequence.

Visualisation of model judgements increases the chances of a practical model refinement process and the reliability of the end users.

Model Refinement Through Annotation Feedback

The interactive alignment plot and colour coding techniques to increase model prediction understanding allow for immediate visibility of misclassification and areas that require indispensable improvement. This visualisation presents an actionable insight into areas where the model should be refined, suggesting further model tuning. In addition, highlighting the prediction of the results can help spot misclassification, thereby facilitating error analysis of the model. This gives room for understanding why particular predictions go wrong and the pattern of how the model works. In addition, visualising model diagnosis gives clear insight into the model interpretation and explains its internal operation and decision-making to the human-in-the-loop.

To incorporate human feedback into the model's further optimisation process, the work developed an interactive interface to collect user input, as shown in Figure 8. When the model predicts a given instance, the human-in-the-loop can accept if all tokens are correctly labelled (Figure 9a). Otherwise, an expert provides corrective feedback for the misclassified tokens. Considering the longer sequences in clinical documents, the method developed an interface that allows users to upload a JSON file containing the corresponding labels for each token in the sequence (Figure 9b). This reduces the tedious burden of manually labelling each sequence, especially the ones with many misclassified tokens. Alternatively, iterative

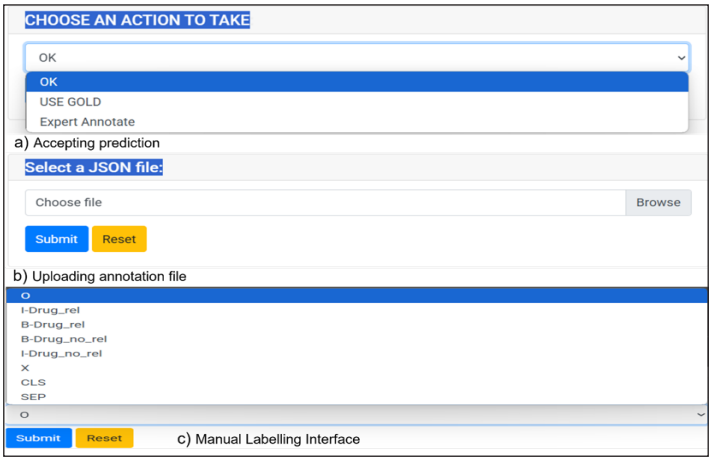


Figure 9. Designed interfaces. (a) Accepting model prediction (b) Uploading feedback (c) Manual labelling interface

labelling for each token in the sequence interface was equally provided to increase the system's usability, flexibility, and robustness for the end users (Figure 9c). A gold standard annotation was used as the reference lookup for the human-in-the-loop to make corrections to ensure unbiased feedback.

To incorporate human feedback, corrected labels are replaced with the model prediction for the misclassified tokens, and the model parameters are updated with new concatenated data during retraining, as shown in Equation 16.

$$\theta updated = \theta - \eta \quad _{\theta} L(D_{new}; \theta) \qquad [16]$$

Where θ are the model parameters, $\mathscr{L}$(Dnew;θ) is the loss function over the updated dataset, and η is the learning rate.

RESULTS AND DISCUSSION

This section summarises the experimentation conducted with the chosen model to optimise the soft-tuned model performance. This paper showcases how sample validation sets are utilised to further optimise the model based on the prediction result provided by the model and the corrective feedback provided by the human-in-the-loop model development. Table 1 shows sample experiments of model predictions on some data instances with user feedback.

Base Models

In this research, transformer-based models are utilised that have a stronger capability of contextually handling longer and more diverse sequences (Elbiach et al., 2025). Based on the soft prompting tuning performance of the base models on the training set, utilising 10% for validation, the work selected the best-performing model among the base models for the interactive human-in-the-loop optimisation experiment. The BERT model (Devlin et al., 2019) It is a PLM based on Masked Language Modelling for NLU tasks; the research utilised BERT-base. The SciBERT model (Beltagy et al., 2019) builds upon the BERT architecture and is pre-trained on a large corpus of 1.14 million full-text papers from Semantic Scholar. There are two available versions of Sci-BERT: sci-vocab and base-vocab. The work utilised the sci-vocab model. The models are tuned using the soft prompt tuning unfrozen (STU) strategy on multi-task ADE extraction.

Experimental Setup

The experiment was conducted on a single Tesla V100 GPU server running CUDA version 11.7 and a 16-core CPU. A maximum sequence length of 512 for fine-tuning and a batch size of 2 were used. The learning rate was set to 5e-5, with the cross-entropy loss function and the Adamax optimiser. A weight decay of 0.05 and a dropout rate of 0.1 to prevent overfitting

Table 1
Sample experiment of human-in-the-loop interaction for model tuning (TAC 2017 & N2C2 2018 datasets)

Input Sequence	Dataset	Task	Correctly Classified Tokens	Misclassified Tokens	Expert Feedback
"Heparin-induced thrombocytopenia. Secondary 1. End"	TAC 2017	ADE-concept	10	5	Provide a correct label to the misclassified tokens
"Seven-day course of ciprofloxacin and remained asymptomatic. 13. HTN Pt was normotensive with an episode of hypotension as described above. BP medications were held, and BP was monitored throughout his stay. It is recommended that his medications be started as an outpatient after monitoring his BP for hypotension and reevaluating his hypertension. 14. H o alcohol abuse Pt was monitored on CIWA, with prn Ativan. 15. FEN cardiac diabetic diet. 16. Prophylaxis Initially, SC heparin, pneumoboots when off heparin. Ambulatory towards the end of his stay. Bowel regimen, PPI. 17. Access PICC"	N2C2 2018	ADE-concept	111	38	Provide a correct label to the misclassified tokens
"Sedation with Versed 3.5, Fentanyl 75, and Phenergen 25 pt was noted to be apneic with an O2 sat of 77%. Bag ventilation was initiated with an increase in his Sats to 100%. He was given Narcan 400 mcg IM and Flumazenil 200 mg IV."	N2C2 2018	ADE-relation	50	12	Provide a correct label to the misclassified tokens
"Intrathecal methotrexate Hist"	TAC 2017	ADE-concept	6	1	Provide a correct label to the misclassified tokens
ASA 325, Amlodipine 10, Carvedilol 25 Hospital 1, Clobetasol q AM, Furosemide 40, Glipizide 5, Moexipril 30, Rosuvastatin 5, OxyContin 20. Discharge Medication"	N2C2 2018	ADE-relation	30	32	Provide a correct label to the misclassified tokens
"Contrast material opacifying"	N2C2 2018	ADE-concept	3	0	Accept model predictions
"Other serious infections: increased risk of bacterial, viral, fungal, and protozoal infections, including opportunistic infections and tuberculosis."	TAC 2017	ADE-relation	24	24	Accept model predictions
"Potential worsening of infections (e.g., existing tuberculosis; fungal, bacterial, viral, or parasitic infections; ocular herpes simplex)."	TAC 2017	ADE-relation	20	5	Provide a correct label to the misclassified tokens

were applied. The models were retrained for two epochs on the TAC 2017 dataset and ten on the N2C2 2018 dataset. Additionally, 10% of the training data was used for validation to optimise the model using an IHITLL process, and the test set was used for the final evaluation.

Experimental Results

This section presents the result of an interactive human-in-the-loop model fine-tuning of the transformer-based model for multi-task ADE extraction (IHITLFT-MADE) for concept and relation from N2C2 2018 and TAC 2017 datasets.

Result on N2C2 2018 Dataset

Figure 10 shows the experimental result for concept extraction. For the overall F1 score, the IHITLFT-MADE obtained 0.9404, which is ahead of BERT-STU with 0.9244 by 1.6% and SciBERT-STU with 0.9254 by 1.5%. Similarly, for RE, for the overall F1 score, the IHITLFT-MADE obtained 0.9132, which is ahead of the BERT-STU with 0.8916 by 2.16% and SciBERT-STU with 0.8925 by 2.07%.

A paired Student's T-test was employed to evaluate the performance of IHITLFT-MADE against the baseline model's performance and determine its significance. Table 2 shows that the results are statistically significant (p-values of 0.0010 and 0.0100) for concept extraction and (p-values of 0.0050 and 0.0010) for relation extraction.

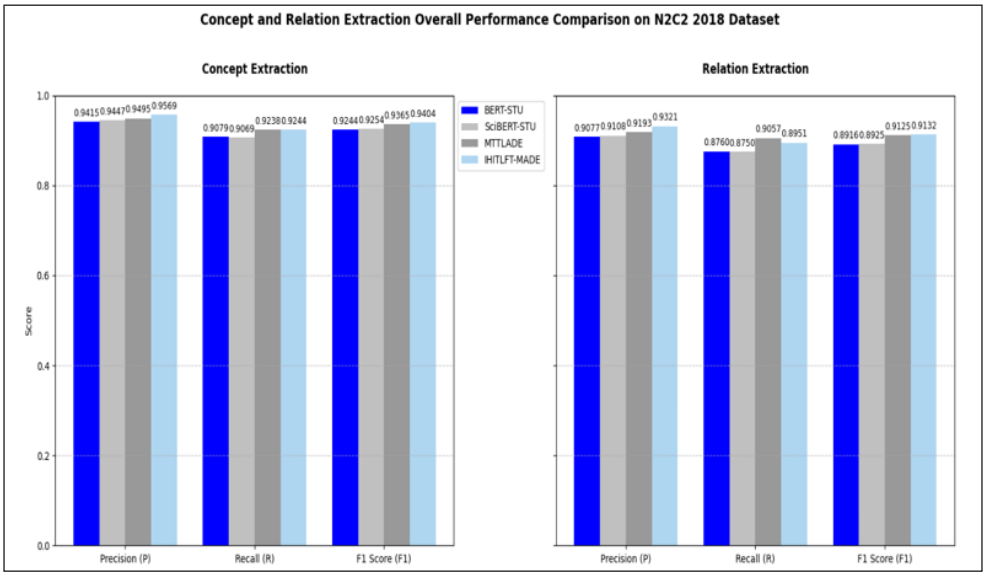


Figure 10. IHITLF-MADE result compared with baseline models: BERT-STU and SciBERT-STU for concepts and relations extraction on the N2C2 2018 dataset

Table 2
Significance test analysis of IHITLFT-MADE against baseline models on N2C2 2018

Measurement	System1	System 2	Statistics	Conclusion
		Concept		
Precision, Recall and F1-score (N=3)	IHITLFT-MADE Mean = 0.9405 SD =0.0162	BERT-STU Mean = 0.9246 SD = 0.0168	t = 50.2130 p = 0.0010	Significant
		SciBERT-STU Mean = 0.9256 SD = 0.0189	t = 9.7340 p = 0.0100	Significant
		Relation		
Precision, Recall and F1-score (N=3)	IHITLFT-MADE Mean = 0.9134 SD = 0.0185	BERT-STU Mean = 0.8917 SD = 0.0158	t = 14.1760 p = 0.0050	Significant
		SciBERT-STU Mean = 0.8927 SD = 0.0179	t = 59.7560 p = 0.0010	Significant

Note. Mean is the mean of precision, recall and F1 score, SD is the standard deviation of the score, N is the number of measurements, t is the statistical t-score, and p is the statistical p-value. $P < 0.05\%$ (95% confidence interval) indicates the performance is significant between the two systems. STU: soft prompt tuning unfrozen

Result on TAC 2017 Dataset

Figure 11 shows the experimental result for concept extraction; for the overall F1-score, the IHITLFT-MADE obtained 0.8723, which is ahead of the BERT-STU with 0.8532 by 1.91% and SciBERT-STU with 0.8708 by 0.15%. Similarly, for the relation extraction, for the overall F1-score, the IHITLFT-MADE obtained 0.5506, which is ahead of the BERT-STU with 0.5115 by 3.91% and SciBERT-STU with 0.5333 by 1.73%.

Table 3 shows that IHITLFT-MADE is statistically significant (p-values of 0.0050) over BERT-STU for concept extraction and (p-values of 0.0140) over SciBERT-STU for relation extraction.

Ablation Study

To evaluate the contribution of each component within the proposed framework, this section presents the results of experiments conducted with three configurations: baseline model fine-tuning, soft prompt only, and the human feedback mechanism. The three interaction interfaces. The first two interaction components, understanding and diagnosis, primarily utilise the model's output to enhance interpretability. In contrast, the refinement component directly contributes to model performance by incorporating human feedback. The outcomes of the ablation study are summarised in Table 4.

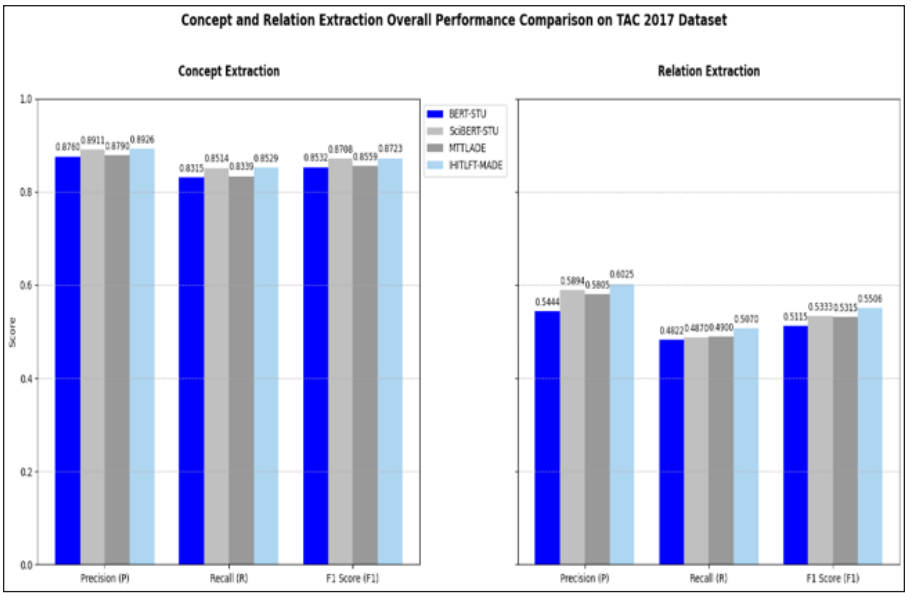


Figure 11. IHITLF-MADE result compared with baseline models: BERT and SciBERT soft prompt tuning (unfrozen) for concepts and relations extraction on the TAC 2017 dataset

Table 3
Significance test analysis of IHITLFT-MADE against baseline models on TAC 2017

Measurement	System 1	System 2	Statistics	Conclusion
		Concept		
Precision, Recall and F1-score (N=3)	IHITLFT-MADE Mean =0.8726 SD =0.0198	BERT-STU Mean = 0.8535 SD = 0.0222	t = 13.7320 p = 0.005	Significant
		SciBERT-STU Mean = 0.8711 SD = 0.0198	t = - p = - SRD = 0	Not Significant
		Relation		
Precision, Recall and F1-score (N=3)	IHITLFT-MADE Mean =0.5533 SD =0.0478	BERT-STU Mean = 0.5127 SD = 0.0311	t = 4.2160 p = 0.052	Not significant
		SciBERT-STU Mean = 0.5365 SD = 0.0512	t = 8.3690 p = 0.0140	Significant

Note. Mean is the mean of precision, recall and F1 score, SD is the standard deviation of the score, N is the number of measurements, t is the statistical t-score and p is the statistical p-value. $P < 0.05\%$ (95% confidence interval) indicates the performance is significant between the two systems. STU: soft prompt tuning unfrozen

Table 4
Ablation study

Metric	Baseline Model Fine-tuning				Baseline Model Soft Tuning				IHITLF-MADE	
	N2C2 2018 Dataset									
	BERT		SciBERT		BERT		SciBERT			
	concept	relation	concept	relation	concept	relation	concept	relation	concept	relation
P	0.9383	0.8885	0.9433	0.9121	0.9415	0.9077	0.9447	0.9108	**0.9569**	**0.9321**
R	0.9026	0.8257	0.9051	0.8715	0.9079	0.8760	0.9069	0.8750	**0.9244**	**0.8951**
F1	0.9201	0.8560	**0.9238**	**0.8913**	0.9244	0.8916	**0.9254**	**0.8925**	**0.9404**	**0.9132**
	TAC 2017 Dataset									
P	0.8614	0.5333	0.8919	0.5805	0.8760	0.5444	0.8911	0.5894	0.8926	0.6025
R	0.8164	0.4724	0.8464	0.4900	0.8315	0.4822	0.8514	0.4870	0.8529	0.5070
F1	0.8383	0.5010	**0.8685**	**0.5315**	0.8532	0.5115	**0.8708**	**0.5333**	0.8723	0.5506

ADE Extraction Outcomes

This section presents sample outcomes of the ADE extraction for concept and relation tasks. Tables 5 and 6 show the extracted concepts and relations for the N2C2 2018 and TAC 2017 datasets, respectively.

The models' predictions for each task extract the entities mentioned within the clinical text. An extraction module was created to extract all the mentioned segments. For instance, each mention has a segment, text, section start, and length for the TAC drug label. Based on the extracted positive mentions, their related attributes and relation pairs are extracted and written to the annotation files for comparison with the gold annotation files.

Table 5
Sample extraction outcome from testing set in N2C2 2018

Document	Concept Extraction	Relation Extraction
Clinical notes-100130	Drugs: Decadron, Oxacillin, Sulphate, Lipitor, Vanco ADE: eye discharge Dosage: 2 Form: drops Reason: seizure Prophylaxis, ophthalmic Involvement Duration: one week, 3 days Frequency: q.i.d Route: Intravenous Strength: 10 mg	Duration-Drug: Arg1: Decadron Arg2: one week Frequency-Drug: Arg1: Decadron, Arg2: q.i.d Reason-Drug: Arg1: ophthalmic Involvement, Arg2: Oxacillin Route-Drug: Arg1: Intravenous, Arg2: Oxacillin ADE-Drug: Arg1: eye discharge, Arg2: Dilantin Form-Drug: Arg1: drops, Arg2: Sulphate Strength-Drug: Arg1: 10 mg, Arg2: Lipitor Duration-Drug: Arg1: for 3 days Arg2: Vanco Dosage-Drug: Arg1: Lipitor, Arg2: 2

Table 5 (*continued*)

Document	Concept Extraction	Relation Extraction
Clinical-10059	Drugs: Vancomycin, Chemotherapy, Fluid, Metoprolol, Furosemide, Carbamazepine, Chlorhexidine Gluconate ADE: Tumour lysis Syndrome, beta-blocker toxicity. Dosage: 15 ml, 1 Duration: for 8 days Frequency: ONCE Form: Oral Rinse Route: PO Reason: VAP	Duration-Drug: Arg1: for 8 days, Arg2: Vancomycin Frequency-Drug: Arg1: ONCE, Arg2: Furosemide Reason-Drug: Arg1: VAP, Arg2: Vancomycin Route-Drug: Arg1: PO, Arg2: Carbamazepine ADE-Drug: Arg1: Tumour lysis Syndrome, Arg2: Chemotherapy Form-Drug: Arg1: Oral Rinse, Arg2: Chlorhexidine Gluconate Strength-Drug: Arg1: 400 mg, Arg2: Carbamazepine Dosage-Drug: Arg1: boluses, Arg2: Fluid Dosage-Drug: Arg1: 15 ml, Arg2: Chlorhexidine Gluconate

Table 6
Sample extraction outcome from testing set in TAC 2017

Drug Label	Concept	Relation
ACTEMRA	ADRs: infection, decrease in neutrophil counts, decrease in platelet count, Stevens-Johnson Syndrome, viral reactivation Severity: serious, below 1000 per mm 3 Negation: without, no Factor: Risk, may Drug class: immunosuppressive biologic therapies	ADR-Severity (Effect): Arg1: Decrease in neutrophil counts Arg2: below 1000 per mm 3 ADR-Severity (effect): Arg1: Infection Arg2: Serious ADR-Negation (Negated): Arg1: Decrease in neutrophil counts Arg2: without ADR-Factor (Hypothetical): Arg1: Stevens-Johnson Syndrome Arg2: Risk ADR-Drug class (Hypothetical): Arg1: Viral Reactivation, Arg2: immunosuppressive biologic therapies
ASCLERA	ADRs: Decrease visual acuity, Increase in bilirubin, ALT elevation, embryotoxic Severity: Grade 3-4, Grade 1-2 Negation: No Factor: Can, Intestinal perforation Animal: Rats	ADR-Factor (Hypothetical): Arg1: Decrease visual acuity Arg2: Can ADR-Negation (Negated): Arg1: Increase in bilirubin Arg2: No ADR-Severity (Effect): Arg1: ALT elevation Arg2: Grade 3-4 ADR-Animal (Hypothetical): Arg1: embryotoxic, Arg2: Rats

Complexity Analysis

Although the performance gain and improved extraction outcome, the model still struggles with certain complex entities and relations. The common one is the case of ambiguous entities within the dataset. For example, in the TAC 2017 dataset, the text input "no clear relationship between decreases in neutrophils". The term "no" here is ambiguous, as it can stand as a normal term or as a negation term. Similarly, for N2C2 2018 dataset, for inputs "patient had significant delay in recovery of mental status, initially attributed to build up of benzodiazepines used for sedation" and "His extubating was initially limited both by agitation requiring sedation and by requirement for high PEEP to maintain oxygenation.". the sedation here is ambiguous standing as reason for drug in the first input and as a drug in the second input.

Another form of complexity experienced in this implementation is the expansion of the standard sequence length of the BERT model to accommodate the appended soft prompt. This expansion could increase the computational resource usage. However, as a trade-off, it led to improved model performance.

DISCUSSION

The results presented for the proposed IHILTLFT-MADE for N2C2 2018 (Figure 10) and TAC 2017 (Figure 11) for concept and relation, respectively. The proposed method employed an interactive human-in-the-loop to optimise the adaptation of transformer-based models with human feedback. As shown in the tables, the results of the best baseline models before the incorporation of the human feedback and benchmarks are included for comparison. The IHITLTFT-MADE consistently outperforms all the interim models across all tasks and datasets. For instance, the challenging entities in N2C2 are ADE and reason entities due to their polysemous nature. For example, the dataset labelled the gold annotation "headache" as ADE and the reason. Similarly, relation types connecting the entities to drugs have been challenging types. The proposed method achieved 0.6717 and 0.7879 for ADE and reason entities and 0.6454 and 0.7385 for ADE-drug and reason-drug relations, ahead of the best baseline model SciBERT-STU with 0.5966 and 0.7330 for the entities and 0.5536 and 0.6747 for the relations. The consistent improvement of the proposed optimisation method over the interim methods demonstrated the effectiveness of the proposed hybrid framework for ADE extraction. This signifies the impact of incorporating human-in-the-loop into fine-tuning to refine the model prediction with correct labels for the misclassified instances. This study focuses on improving ADE extraction from clinical textual documentation using LLM-based techniques.

The statistical analysis test confirmed the significant difference between the IHITLF-MADE and all the baselines across all tasks and experimental datasets. This implies that the final model has a significantly greater prediction capability in detecting ADE cases than the interim models.

Erroneous extraction of clinical concepts such as drug names, adverse events, or indications for drug administration can introduce misinformation into patient records, thereby compromising treatment decisions and patient safety. Understanding the characteristics of low-confidence entities and the complex relations that models fail to capture is essential for designing effective post-processing checks and for incorporating real-world pharmacovigilance experts as human-in-the-loop participants during model redevelopment. The performance gains achieved by the proposed methods provide evidence that integrating human-in-the-loop learning with pretrained language model adaptation can bridge the performance gap left by traditional approaches. This research has demonstrated the potential to improve ADE extraction, drug safety surveillance, and clinical research in general. The proposed model is a foundation in addition to its immediate practical applications. The model achieves state-of-the-art performance on benchmark datasets, demonstrating its feasibility in supporting large-scale pharmacovigilance and clinical text mining. The graphical user interface and interaction interfaces provided enable direct involvement by domain experts during system development and practical use.

Comparison with Benchmark

Figures 10 and 11 summarise the comparison of the results between IHITLFT-MADE and MTTLADE for both N2C2 2018 and TAC 2017. The MTTLADE system has been a state-of-the-art study that has recently been shown to outperform all the systems that participated in the challenge organised by the two benchmark datasets (El-Allaly et al., 2021). For the N2C2 dataset, IHITLFT-MADE achieved an overall F1 score of 0.9404, ahead of MTTLADE with 0.9365 by 0.39% for the concept extraction task, and IHITLFT-MADE achieved an F1 score of 0.9132, ahead of MTTLADE with 0.9125 by 0.07% for the RE task. Similarly, for the TAC dataset, IHITLFT-MADE achieved an overall F1 score of 0.8723, ahead of MTTLADE with 0.8559 by 1.34% for the concept extraction task, and IHITLFT-MADE achieved an F1 score of 0.5506, ahead of MTTLADE with 0.5315 by 1.91% for the RE task.

The statistically significant analysis shown by Table 7 between IHILFT-MADE and the benchmark MTTLADE using the paired student test for TAC (p-value of 0.0110 for concept and 0.0060 for relation) and approximate randomisation test with randomisation value of 10,000 (p-value of 1.9996×10^{-5}) using the official script provided by the N2C2 challenge organiser shows that IHILFT-MADE's performance in predicting ADE cases is statistically significant compared to MTTLADE's performance.

Error Analysis

Although the proposed hybrid system has demonstrated state-of-the-art performance, the system cannot handle some complex entities and relations within the experimented datasets.

Table 7
IHTLFT-MADE against benchmark MTTLADE statistical significance analysis on TAC 2017 result

Measurement	Task	IHITLFT-MADE	MTTLADE	Statistics	Conclusion
Precision, Recall and F1-score (N=3)	Concept	Mean=0.8726 SD = 0.0198	Mean=0.8552 SD = 0.0226	t = 9.2680 p= 0.0110	Significant
	Relation	Mean= 0.5533 SD = 0.0478	Mean=0.5340 SD = 0.0453	t=13.3160 p = 0.0060	Significant

Note. Mean is the mean of precision, recall and F1 score, SD is the standard deviation of the score, N is the number of measurements, t is the statistical t-score, and p is the statistical p-value. $P < 0.05\%$ (95% confidence interval) indicates the performance is significant between the two systems

The most notable error cases in ADE concept extraction are on concepts with a small number of instances and abbreviations in the training set. For instance, in the TAC 2017 dataset, the text input "TMA was reported in 3 of 2258 patients", the ADR "TMA" only appeared a few times in the training set, so the system failed to extract a similar entity in the test set. Similarly, on N2C2 2018, rare instances of like "QID" drug frequency and "p.o." route of taking a drug are rarely mentioned within the training set. These rare instances with abbreviations are some of the specific cases in which the model failed to extract correctly. Other cases in concept extraction include ambiguous and polysemous entities. For the relation task, failures are related to multiple relations in a single sentence. For example, in the TAC 2017 dataset "The incidence of Grade 3 and 4 depressive disorders", the "depressive disorders" here is related to both "Grade 3" and "Grade 4" severity of ADR with an effect type of relation. The model identifies only one relationship. Similarly, in the N2C2 dataset, identifying the relation between the challenging entities of ADE and Reason entities to the drug was among the failure cases observed in the system.

CONCLUSION

This work presented a proposed interactive human-in-the-loop-based method to further optimise the model tuning through human feedback during the fine-tuning process. Current ADE extraction models require large, annotated datasets and offer limited interpretability for clinical use. The proposed method adopts the best-performing baseline and applies interactive learning techniques to optimise its performance on multi-task ADE extraction. Visualisation techniques are employed to improve human-in-the-loop understanding of model prediction and diagnosis. A transformer-based attention score assigned to features is utilised to interpret model decision-making. To refine the model, interactive interfaces are provided to incorporate human feedback with corrective labels for the misclassified samples.

The proposed method outperformed all the baseline models and the benchmark system in the two experimental datasets, demonstrating its capabilities in extracting more intricate entities and relations from clinical documents. The statistically significant analysis shows that the performance of the proposed method is statistically significant compared to the baseline and benchmark methods. The final evaluation of the proposed framework using an official evaluation script demonstrates its capability, showing a comparable extraction performance to the gold standard annotation provided by the challenge organisers. This implies that the proposed system can be employed for real-time, accurate and reliable extraction of ADE cases from narrative documents.

In future research, the work plan is to employ real-time clinical data and engage a real-time pharmacovigilance expert to interact with the system for identifying ADE instances within the narrative documents and to assess its robustness in handling imperfect or conflicting annotations. In addition, large language models with larger parameters will be investigated using a frozen model tuning strategy for parameter-efficient model adaptation and resource limitations.

ACKNOWLEDGEMENT

This research is supported by Universiti Putra Malaysia and the Ministry of Higher Education, Malaysia, under the Fundamental Research Grant Scheme (FRGS/1/2023/ICT02/UPM/02/3).

SUPPLEMENTAL MATERIAL/DATA

All the datasets utilised for this research were obtained based on data usage agreements. Part of the agreement is that the datasets should not be shared with any third party without the organisation's prior consent. Consent and access to data can be obtained from https://N2C2.dbmi.hms.harvard.edu/data-sets. Codes are available at https://github.com/salisushagari/IHILT-MADE.

REFERENCES

Accenture. (2019). *Explainable AI: The next stage of human–machine collaboration*. https://www.accenture.com/gb-en/insights/technology/explainable-ai-human-machine

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., . . . Kaplan, J. (2022). *Training a helpful and harmless assistant with reinforcement learning from human feedback* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2204.05862

Belousov, M., Milosevic, N., & Dixon, W. (2019). *Extracting adverse drug reactions and their context using sequence labelling ensembles in TAC2017* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1905.11716

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3615-3620). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1371

Budd, S., Robinson, E. C., & Kainz, B. (2021). A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis, 71*, Article 102062. https://doi.org/10.1016/j.media.2021.102062

Chen, K., Pang, Y., & Yang, Z. (2024). *Parameter-efficient fine-tuning with adapters* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2405.05493

Demner-Fushman, D., Fung, K. W., Do, P., Boyce, R. D., & Goodwin, T. R. (2018). Overview of the TAC 2018 drug–drug interaction extraction from drug labels track. In *Proceedings of the 2018 Text Analysis Conference (TAC 2018)* (pp. 1-10). National Institute of Standards and Technology. https://tac.nist.gov/publications/2018/additional.papers/TAC2018.DDI.overview.pdf

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

El-Allaly, E. D., Sarrouti, M., En-Nahnahi, N., & Ouatik El Alaoui, S. (2021). MTTLADE: A multi-task transfer learning-based method for adverse drug events extraction. *Information Processing & Management, 58*(3), Article 102473. https://doi.org/10.1016/j.ipm.2020.102473

Elbiach, O., Grissette, H., & Habib, E. (2025). Benchmarking large language models for adverse drug reaction extraction in social media and clinical texts. *Results in Engineering, 28*, Article 107362. https://doi.org/10.1016/j.rineng.2025.107362

Gómez-Carmona, O., Casado-Mansilla, D., López-de-Ipiña, D., & García-Zubia, J. (2024). Human-in-the-loop machine learning: Reconceptualising the role of the user in interactive approaches. *Internet of Things, 25*, Article 101048. https://doi.org/10.1016/j.iot.2023.101048

Gu, Y., Zhang, S., Usuyama, N., Woldesenbet, Y., Wong, C., Sanapathi, P., Wei, M., Valluri, N., Strandberg, E., Naumann, T., & Poon, H. (2023). *Distilling large language models for biomedical knowledge extraction: A case study on adverse drug events* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2307.06439

Henry, S., Buchan, K., Filannino, M., Stubbs, A., & Uzuner, O. (2020). 2018 N2C2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association, 27*(1), 3-12. https://doi.org/10.1093/jamia/ocz166

Jamil, S., Saha, S., & Misra, R. (2026). Enhancing adverse drug event extraction and summarisation for cancer drugs through large language models. *Journal of Biomedical Informatics, 178*, Article 105009. https://doi.org/10.1016/j.jbi.2026.105009

Kim, S., Kang, T., Chung, T. K., Choi, Y., Hong, Y. S., Jung, K., & Lee, H. (2023). Automatic extraction of comprehensive drug safety information from adverse drug event narratives in the Korea Adverse Event Reporting System using natural language processing techniques. *Drug Safety, 46*(8), 781-795. https://

doi.org/10.1007/s40264-023-01323-2

Kim, Y., & Kim, Y. (2022). Explainable heat-related mortality with random forest and SHapley Additive exPlanations (SHAP) models. *Sustainable Cities and Society, 79*, Article 103677. https://doi.org/10.1016/j.scs.2022.103677

Kumar, S., Datta, S., Singh, V., Datta, D., Singh, S. K., & Sharma, R. (2024). Applications, challenges, and future directions of human-in-the-loop learning. *IEEE Access, 12*, 75735-75760. https://doi.org/10.1109/ACCESS.2024.3401547

Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimising continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 4582-4597). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.353

Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4487-4496). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1441

Liu, X., Ji, K., Fu, Y., Tam, W. L., Du, Z., Yang, Z., & Tang, J. (2022). P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 61-68). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-short.8

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730-27744.

Peng, C., Yang, X., Smith, K. E., Yu, Z., Chen, A., Bian, J., & Wu, Y. (2024). Model tuning or prompt tuning? A study of large language models for clinical concept and relation extraction. *Journal of Biomedical Informatics, 153*, Article 104630. https://doi.org/10.1016/j.jbi.2024.104630

Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys, 55*(9), Article 194, 1-33. https://doi.org/10.1145/3561048

Quennelle, S., Malekzadeh-Milani, S., Garcelon, N., Faour, H., Burgun, A., Faviez, C., Tsopra, R., Bonnet, D., & Neuraz, A. (2025). Active learning for extracting rare adverse events from electronic health records: A study in paediatric cardiology. *International Journal of Medical Informatics, 195*, Article 105761. https://doi.org/10.1016/j.ijmedinf.2024.105761

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). *Parameter-efficient transfer learning for NLP* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1902.00751

Ren, Y., & Wang, Z. (2023). A tree-structured neural network model for joint extraction of adverse drug events. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 4938-4941). IEEE. https://doi.org/10.1109/BIBM58861.2023.10386041

Riesener, M., Kuhn, M., Schümmelfeder, S., Xiao, D., Norheim, J., Rebentisch, E., & Schuh, G. (2024). Active learning with pre-trained language models for named entity recognition in requirements engineering. *Procedia CIRP, 128*, 339-344. https://doi.org/10.1016/j.procir.2024.06.027

Sahoo, P., Singh, A., Saha, S., Chadha, A., & Mondal, S. (2024). Enhancing adverse drug event detection with multimodal dataset: Corpus creation and model development. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 11214-11226). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.667

Shelmanov, A., Puzyrev, D., Kupriyanova, L., Belyakov, D., Larionov, D., Khromov, N., Kozlova, O., Artemova, E., Dylov, D. V., & Panchenko, A. (2021). Active learning for sequence tagging with deep pre-trained models and Bayesian uncertainty estimates. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 1698-1712). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.145

Teso, S., Alkan, Ö., Stammer, W., & Daly, E. (2023). Leveraging explanations in interactive machine learning: An overview. *Frontiers in Artificial Intelligence, 6*, Article 1066049. https://doi.org/10.3389/frai.2023.1066049

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., . . . Le, Q. V. (2022). *LaMDA: Language models for dialogue applications* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2201.08239

Tian, Z., Zhang, H., & Wang, Y. (2024). Personalised soft prompt tuning in pre-trained language models: Bridging multitask transfer learning and crowdsourcing learning. *Knowledge-Based Systems, 305*, Article 112646. https://doi.org/10.1016/j.knosys.2024.112646

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998-6008). Curran Associates, Inc.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-rank adaptation of large language models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2106.09685

Ong, J. C. L., Chen, M., Ng, N., Elangovan, K., Tan, N. Y. T., Jin, L., Xie, Q., Ting, D. S. W., Rodriguez-Monguio, R., Bates, D. W., & Liu, N. (2024). *Generative AI and large language models in reducing medication-related harm and adverse drug events: A scoping review* [Preprint]. medRxiv. https://doi.org/10.1101/2024.09.13.24313606

Wondimu, N. A., Buche, C., & Visser, U. (2022). *Interactive machine learning: A state-of-the-art review* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2207.06196

Xia, L. (2022). Historical profile will tell? A deep learning-based multi-level embedding framework for adverse drug event detection and extraction. *Decision Support Systems, 160*, Article 113832. https://doi.org/10.1016/j.dss.2022.113832

Zhao, Y., & Liu, J. (2021). Human-in-the-loop based named entity recognition. In *2021 International Conference on Big Data Engineering and Education (BDEE)* (pp. 170-176). IEEE. https://doi.org/10.1109/BDEE52938.2021.00037

Zitu, M. M., Owen, D., Manne, A., Wei, P., & Li, L. (2025). Large language models for adverse drug events: A clinical perspective. *Journal of Clinical Medicine, 14*(15), Article 5490. https://doi.org/10.3390/jcm14155490

Pertanika J. Sci. & Technol. 34 (3): 1801 - 1827 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Empirical Analysis of Factors in User Control Model for Cloud Data Migration

Ishak Iskandar[1]*, Imbaji Injuwe Danga[2], Hamidah Ibrahim[1], and Fatimah Sidi[1]

[1]Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[2]Department of Information and Communication Technology, Taraba State University, 660213 Jalingo, Taraba State, Nigeria

ABSTRACT

This study aims to develop, with empirical validation, a model of the factors influencing user control during on-premises-to-cloud data migration in Software-as-a-Service (SaaS) environments. The model is grounded in the technology-organisation-environment framework and control theory. The research examines how security, cost, legal compliance, and personnel knowledge affect user control outcomes through standards and performance as control metrics. A quantitative research approach was employed, using survey data collected from 55 cloud computing professionals selected through purposive sampling. The data were analysed using descriptive statistics in SPSS, and their structural relationships were evaluated through Partial Least Squares Structural Equation Modelling (PLS-SEM) in SmartPLS. The results indicate that cost significantly influences both standards and performance, while security significantly affects standards but not performance. Legal compliance shows a significant relationship with performance but not standards, whereas personnel knowledge does not exhibit a significant effect on either standards or performance. Additionally, standards were found to have a significant impact on performance, confirming their role as a critical control mechanism. The measurement model demonstrated strong reliability and validity, with Cronbach's alpha values ranging from 0.753 to 0.955 and factor loadings for all indicators exceeding 0.7, confirming validity. The study contributes to the cloud computing domain by providing a proposed model for assessing user control during migration execution, extending beyond traditional cloud adoption frameworks. Practically, the findings offer guidance to organisations in prioritising cost management, legal compliance, and structured standards in managing cloud data migration processes.

ARTICLE INFO

Article history:
Received: 01 August 2025
Accepted: 07 April 2026
Published: 25 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.22

E-mail addresses:
iskandar_i@upm.edu.my (Ishak Iskandar)
Imbaji.danga@tsuniversity.edu.ng (Imbaji Injuwe Danga)
hamidah.ibrahim@upm.edu.my (Hamidah Ibrahim)
fatimah@upm.edu.my (Fatimah Sidi)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

The model can be applied as a diagnostic and decision-support tool for improving transparency, accountability, and performance in cloud data migration projects.

Keywords: Analysis, hypothetical model, on-premise, SaaS, structural equation modelling (SEM), user control

INTRODUCTION

The continuous adoption of cloud computing by organisations has necessitated efficient and controlled migration processes to maintain data integrity, ensure security, and optimise operational performance. This study analyses the research constructs of the hypothetical model of the user control framework for on-premises to cloud data migration, focusing on the identified user control constructs: standards, performance, security, legal, cost, and personnel knowledge. The research further investigates its findings by means of Structural Equation Modelling (SEM), enquiring whether there is an association between the identified user control constructs and their affecting factors, as asserted by the hypothetical model.

While studies like Pahl and Xiong (2013) have explored the technical aspects of cloud data migration, leaving a gap in addressing the user visibility and control during the migration process. The trend in cloud migration frameworks has been on the Cloud Service Provider's (CSP) offerings, scalability, performance, optimisation, and cost-related issues, without a clear-cut on user involvement (Abou_el_ela Abdou Hussein, 2021; Islam et al., 2017; Khan & Khan, 2019). This has led to the clamouring concerns about user visibility and control of the data migration process, as limited user involvement, reduced visibility, and control in the process persist. It is based on the premise of the challenges that this study intends to address the problem through a critical analysis of factors affecting user control in the process of cloud data migration. By focusing on user control affecting factors and the control metric, the proposed framework should guide users on how to monitor key factors that affect the cloud data migration process, as it intends to enhance visibility and control.

Factors that are seen in the data migration process affecting user control can be explained in terms of technology, organisation, and environmental standpoint. This aligns with the Technology-Organisation-Environment (TOE) framework, hence, investigating the factors within the TOE framework. Additionally, the measurement of the outcome of the user control factors is grounded in the control theory, thereby providing the basis for evaluating the extent to which users have control during the data migration process. The two models together (TOE and control theory) provide a theoretical foundation for analysing the factors in two different contexts. Particularly, the TOE framework serves as a model that projects the understanding of how technology adoption in organisations moving to cloud are implementation, vis-a-vis how innovations are shaped by factors within the technological perspective, and the organisational, and environmental (Pochyla, 2019).

It is therefore suitable for identifying factors affecting user control in this study due to their popular usage in computer science research, particularly in cloud migration and cloud adoption case studies (Kamal et al., 2020; Murtezani & Metin, 2016; Zboril & Svata, 2022). While Control theory is known for its basis in designing and implementing effective control measures that are inherent to organisational processes. It also lay the emphasises on the importance of the establishment of standards, procedures, and performance metrics. These are vital components of the management control process (Dědečková,2020).

This approach tends to ensure that operational activities in control processes are aligned with strategic objectives. Thereby, experience the efficiency and effectiveness of the control process. Standards as a vital metric of control process represent the available predefined criteria for performance evaluation and ensuring organisational goals are consistently evolving. This attribute of standards makes it a reference point for measuring actual performance outcomes; furthermore, standards provide the basic control measures needed for organisations to refer to when performance is required.

Contextually, the use of standards in the cloud data migration process includes compliance with industry requirements, strategy, and best practices, by ensuring that migration processes align with standard procedures in achieving organisational objectives. For example, adherence to security standards like ISO/IEC 27001 guarantees data integrity throughout the migration process. Resulting in user control and trust issues due to the service reliability experienced (Laurent de la Vaissière, 2011). Performance metric, on the other hand, measures how standards in the management control process can provide quantitative measures to evaluate and monitor user control in the data migration process. This emphasises the application of performance metrics in domains like: database systems, real-time systems, and cloud computing (Abdelzaher et al. 2008). Making it easy to identify deviations from expected outcomes.

The usage of the TOE framework to examine factors that influence organisations' preparedness to adopt and implement technological innovations has been widely used. For example, studies like Yang et al. (2015) and Oladoyinbo et al. (2024) have investigated factors affecting cloud computing adoption or cloud migration. To make it noteworthy, some studies refer to data migration to the cloud as cloud adoption, whereas most studies maintain the data migration in terms of transferring or moving data to the cloud environment. Besides cloud adoption advantages, there is a combination of organisational and technical factors that affects user control in the process of migrating data to the cloud, for users in Software as a Service (SaaS). The affecting factors are not limited to personnel skills, legal, and security requirements that become issues of concern looking at the organisational perspective. Similar studies have also, in an effort to integrated organisational and technical factors in SaaS, highlighted the challenges of data migration and found that the technical expertise of the staff involved in the process plays an important role in determining user control (Iqbal & Colomo-Palacios, 2019).

Conclusively, the TOE framework offers the foundation for identifying the user control affecting factors for a successful cloud data migration process, by addressing key factors that influence the process. When combined with control theory and focusing on user control, this approach conceptualises the factors in user control in the cloud data migration process.

Related Research Works

This subsection examines related works and provides insight into the research constructs Grob & Schill (2012) proposed a user-centric governance and control framework. Their work significantly contributes to understanding how governance can shape data migration practices. However, a major limitation is the neglect of user experience as part of regulatory compliance considerations. Perra (2015) explored integrated user control mechanisms, such as predefined permissions and automated decision-making during data migration. The study shows that while these mechanisms enhance user control in security and privacy matters, they simultaneously reduce user involvement in certain functionalities. The multi-step process of data migration, as outlined in Sarmah (2018), shares similarities with a user control framework for cloud data migration in SaaS environments. It involves analysing legacy data, performing quality assurance testing, and executing phased transitions to minimise risks and ensure data integrity.

Patrick & Satyanarayana (2020) examined the optimisation of service level agreements (SLAs) by balancing cloud provider offerings with user experience. Their findings reveal that users typically lack control over their data once stored in the cloud, and SLA issues are only addressed after migration, often through penalties. The absence of pre-migration SLA considerations underscores the need to empower users to define service preferences using quality of service indicators embedded within migration frameworks. The migration to cloud-based services, as explored in Ali et al. (2021), sheds light on various factors influencing the migration to cloud environments. While it identifies critical organisational and technical challenges, the focus remains on adoption rather than on user control mechanisms during migration. Sadoughi et al. (2020), in a review of studies, highlighted cloud migration factors categorised into technological, organisational, environmental, and individual domains. These factors closely align with the factors that affect the implementation of user control in cloud data migration in SaaS. Their work is useful in understanding what drives adoption, but it lacks emphasis on factors affecting user control during data migration. Amin et al. (2021) provide insights into the challenges associated with cloud data migration, particularly focusing on the migration of locally stored data to public cloud platforms.

Al Mudawi et al. (2022) developed the Adoption of Cloud Computing in Government (ACCE-GOV) framework to analyse the critical factors affecting the adoption of cloud computing in Saudi government systems. The study identifies factors such as security,

trust, technology readiness, top management support, and regulatory issues as significant in cloud adoption. While valuable in analysing cloud adoption factors, it does not explore user-specific control elements within cloud data migration contexts. This opens the door for enhancing such frameworks by integrating user-driven control factors to improve transparency and engagement during migration. Vadlamani et al. (2023) explored strategies for cross-platform data migration specifically designed for enterprise data warehouses (EDWs). While these works share common challenges with this work, it leaves a gap for extending such strategies to include user-centric controls for on-premises to cloud migration processes, as shown in Table 1.

Table 1
Summary of related research works

Author	Focus/ Methodology	Findings/ Limitations	Identified Gap	Proposed Contribution
(Grob & Schill, 2012)	User-centric governance and control in data migration	Addresses the cloud provider trust issue, regulatory compliance, and user-controlled migration.	It neglects user experience as part of regulatory compliance.	Ensure user experience as part of the control concern.
(Perra, 2015)	Integrated user control mechanisms like predefined permissions and automated decision-making.	Ensures user control in security and privacy; conversely, limits user involvement in some functionalities.	Absence of visibility into data flows, compliance enforcement, and integrity maintenance during migration.	Extends user control by incorporating legal and security oversight through the process.
(Sarmah,2018)	Multi-step data migration process with risk management and rollback strategies.	Focuses on minimising risks and testing but does not involve user-driven feedback or control mechanisms.	The framework is too general for data migration.	The development of a specific purpose framework for data migration is timely.
(Patrick & Satyanarayana, 2020)	Optimisation of service level agreement	Define a balance between cloud provider offerings and cloud user experience. Users have no control over their data stored in the cloud.	Does not address service level agreement issues before migration, hence uses a penalty as a solution.	Allow the user to determine their service preference through quality service indicators in the framework.

Table 1 (*continued*)

Author	Focus/ Methodology	Findings/ Limitations	Identified Gap	Proposed Contribution
(Ali et al., 2021)	Analysis of factors influencing cloud adoption, including compatibility, complexity, cost, and security.	Identifies key organisational and technical challenges but focuses on adoption rather than user-centric control mechanisms for migration.	No emphasis on user control factors over cloud data migration.	Consider user control factors.
(Sadoughi et al., 2020)	Factors influencing cloud adoption are categorised into technological, organisational, and individual domains.	Highlights adoption factors but lacks focus on user-centric control factors during the migration process.	Limited focus on engaging users to manage cloud data migration challenges.	Investigate more factors affecting user-centric control migration workflows.
(Amin et al., 2021)	Insights into cloud data migration challenges, focusing on the migration of locally stored data to the public cloud.	Highlights the importance of planning, impact analysis, and visibility into security, but lacks mechanisms for real-time user control in SaaS environments.	Lack of user control and transparency during migration processes in SaaS platforms.	Introduces user-centric tools like dashboards to enhance visibility and control in cloud migration workflows.
(Al Mudawi et al., 2022)	ACCE-GOV framework addressing adoption factors like security, trust, and technology readiness.	Analyses adoption factors but does not explore factors that are user specific.	Lack of integration of user control factors in cloud data migration contexts.	Enhances the adoption framework by integrating user-driven control factors.
(Vadlamani et al. ,2023)	Strategies for cross-platform data migration for enterprise data warehouses (EDWs).	Addresses data compatibility and integration complexities. No user control mechanism.	The study did not cover on-premises to cloud migration.	Consider on-premises to cloud migration.

THEORETICAL BACKGROUND OF CONSTRUCTS

To investigate measures of user control in cloud data migration, this study integrates the TOE framework and control theory to develop a theoretical model of the research constructs. Traditionally, cloud migration models focus on CSP prospects, technical efficiency and cost-effectiveness, often neglecting user control measures that enhance transparency, compliance, and security. This study aims to establish a structured model that effectively

evaluates measures of user control in cloud migration processes by leveraging the TOE framework and control theory

Technology-Organisation-Environment (TOE) Framework

The TOE framework, presented by Tornatzky (1990) Classify the factors influencing technology adoption into three dimensions, namely: technological factors; security, performance, and adherence to standards influence the technical feasibility of cloud migration. Organisational factors, such as cost, personnel knowledge, and operational readiness, affect the organisation's ability to manage migration effectively. And environmental factors: legal and regulatory compliance, market pressures, and industry standards impact the decision-making process.

Control Theory

Control theory provides a foundation for measuring control outcomes within the cloud migration processes. It establishes the user control metrics (standards and performance) as integral components of the management control process (Dedeckova, 2020). Control, as an element of managerial process, requires setting up practical standards for organisational activities, evaluating actual performance based on the set standards, and identifying possible deviations by comparing actual outcomes against the predetermined standards (Manyuchi & Sukdeo, 2021). Standards are inherent to certain specifications, such as ISO/IEC 27001, where compliance with security and data integrity protocols is essential, thereby strengthening user trust in cloud migration (Laurent de la Vaissière, 2011). The use of performance metrics to assess projects of cloud data migration ensures that organisations monitor and assess how users can control the migration process.

Integrating TOE and Control Theory in the Cloud Data Migration Process

This study makes three primary theoretical contributions to cloud computing and user control in cloud data migration. First, while prior studies are grounded in the TOE framework and primarily examine determinants of cloud adoption readiness, intention, or migration feasibility, this study shifts focus from the adoption decision to data migration with user control in focus. It conceptualises user control not as an adoption outcome, but as a measurable construct during the transition phase of on-premises to cloud data migration. Secondly, this study integrates TOE determinants with principles from control theory to develop a structured evaluative model. Unlike prior models that treat technological, organisational, and environmental factors as independent factors of cloud adoption, this framework connects these factors to control outcome metrics, standards and performance. Thirdly, the study further positions user control as a measurable multidimensional construct

validated empirically using PLS-SEM. Existing literature discusses visibility, trust, governance, and compliance in conceptual terms, but rarely provides validated indicators for measuring user control during migration execution. This study addresses that gap by offering validated constructs that can serve as a foundation for future empirical work in cloud governance research

The factors of TOE are integrated with metrics of the control theory to allow for a comprehensive fusion and analysis of user control measures in cloud data migration. This is done by aligning the technological, organisational, and environmental factors with control measures. The aim is to identify and measure key indicators of user control. Using the proposed model through empirical testing with SEM, the study offers a theoretical foundation for gaining an understanding of how user control factors are measured in the process of migrating on-premise data to the cloud.

The Study Objectives

The study aims to identify the factors that affect user control in cloud data migration, analyse the relationships among the identified factors, and propose a model for user control in the cloud data migration process.

The Proposed Model of the Research

The Research model for user control affecting factors in cloud data migration is developed based on the technology organisation environment framework and the control theory. The model defines key factors influencing user control and hypothesises the relationships among them.

The integration of technological, organisational, and environmental factors, as outlined in the TOE framework, establishes a structured approach for understanding the dynamics of factors of user control in cloud migration. Additionally, control theory provides a foundation for measuring performance and standards metrics as outcomes for user control in the data migration process. The proposed model was tested for reliability, and a pilot study was carried out before the data collection stage. This is to ensure that the measurement instruments are aligned with the objectives of the empirical study. The research model put forward the relationships among the constructs, including security, cost, legal compliance, personnel knowledge, standards, and performance, in determining the extent to which users are involved in the control of the migration project.

The proposed model will be tested using SEM to examine the relationships between the identified factors and validate the model's effectiveness in improving user control mechanisms. Figure 1 illustrates the hypothesis model of the user control framework.

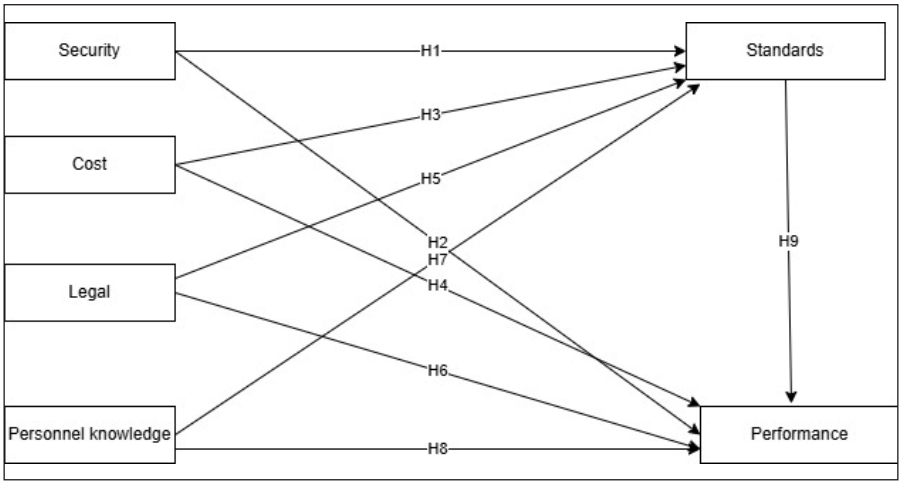


Figure 1. Hypothesis model of the user control framework

Hypothesis Formulation

The research basis for this empirical investigation proposes nine hypotheses to examine the relationships between the key factors influencing user control in migrating on-premises data to the cloud. The hypotheses explored the impact of the relationship among the user control measures, security, cost, legal, and personnel knowledge on standards and performance.

Impact of Security on Standards and Performance

Security is a crucial issue in the data migration process. Considering measures like: data integrity, ensuring that data is complete, accurate, and consistent as it moves from source to target (Yang et al., 2015); confidentiality, referring to the protection of sensitive information from unauthorised access, disclosure, or theft (Fisher et al., 2023) data doss, the permanent or temporary unavailability of data during the transition process (Dziadosz et al., 2021) and privacy, monitoring of privacy in cloud data migration is essential to ensure security and compliance (Maniah et al., 2021). Hence, users need to adhere to security measures to ensure that data migration standards are met, which in turn affects the entire performance of the data migration process. Thus, the following hypotheses are inferred:

H1: Security has a significant effect on standards in the user control of the data migration process.

H2: Security significantly influences performance in the user control of the data migration process.

Impact of Cost on Standards and Performance

Cost considerations are essential in determining the feasibility and sustainability of user control in the cloud data migration process. Managing data migration costs could contribute to standards and migration performance in measures like application and data cost, storage cost, connectivity cost, and consultancy cost (Bhatt & Sehgal, 2024; Islam et al., 2023; Mansouri et al., 2019; Varma, 2024). Based on the assertion, the following hypotheses are considered:

H3: Cost has a significant effect on standards in the user control of the data migration process.

H4: Cost has a significant effect on performance in the user control of the data migration process.

Impact of Legal Factors on Standards and Performance

User control is affected by certain legal factors that users must consider to ensure compliance and mitigate risks (Chowdhury, 2024). Legal considerations are important to the successful implementation and operation of cloud services; they comprise policy adherence, service level agreements (SLAs), compliance with regulations, and governance frameworks (Rahulamathavan et al., 2014)**.** This has become necessary to ensure that data migration processes align with established standards, procedures and performance measures. The following hypothesis statements reflect the significance of legal factors on standards and performance:

H5: Legal compliance has a significant effect on standards in the user control of the data migration process.

H6: Legal compliance has a significant effect on performance in the user control of the data migration process.

Impact of Personnel Knowledge on Standards and Performance

The expertise and knowledge of personnel involved in cloud data migration directly influence the implementation of control measures. Trained personnel acquire skills that could enhance adherence to certain standards and improve performance. Hence, diverse skills, including: technical, technical requirements for monitoring the systems (Malouche et al., 2017) communication, coordination among teams and departments, making everyone involved to understands the goal, timeline, and responsibility (Rao et al., 2021) and business skills, strategic business acumen can align the data migration process objectives, with the business goals (Alkhalil et al., 2013) to ensure a seamless migration project. Consequently, the following hypotheses were formulated:

H7: Personnel knowledge has a significant effect on standards in the user control of the data migration process.

H8: Personnel knowledge has a significant effect on performance in the control of the data migration process.

Impact of Standards on Performance in Data Migration

Standards usually, is the benchmark for evaluating the success of cloud data migration by ensuring that user control factors adhere to established standards. This helps users to maintain consistency throughout the migration process by exploring various user control measures related to how standards impact performance. This tests the significance of standards on performance using the hypothesis statement below:

H9: The standards of the data migration process have a significant effect on its performance.

The hypotheses H1, H2, H3, H4, H5, H6, H7, H8, and H9 tend to outline the theoretical foundation of the empirical investigation, so that the proposed research model could be validated through SEM.

MATERIALS AND METHODS

Research Design

This study employs a largely the use of quantitative research approach and the qualitative research approach in limited cases for the investigation of factors that influence user control in data transitioning to the cloud. It analyses the key control measures such as security, cost, legal compliance, personnel knowledge, standards, and performance from the integrated TOE framework and control theory. A survey was carried out to collect data from respondents who are computer scientists, particularly cloud computing professionals. The study further allows for the use of suitable statistical techniques and analytical software applications to test the relationships among variables.

To achieve the objective of the study, the research methodology is structured into five phases to ensure a comprehensive and systematic approach to evaluating the measures of user control in the cloud data migration process. The five phases of the research methodology include:

1. Reviewing existing literature to identify user control metrics, the phases of cloud data migration, and factors that affect user control.
2. The development of a hypothetical model that integrates user control factors from the TOE context and their accompanying outcome in control theory.
3. The development and validation of the research instrument and conducting a pilot study.

4. The data collection and analysis phase.
5. The model validation phase.

Data Collection Instrument

Primarily, the data collection instrument was developed using constructs from existing literature, which were reviewed by an expert before distribution. This is to ensure that ambiguities are removed and items are refined to enhance clarity. Before administering the questionnaire, it was divided into two Sections: Section A has general demographic information (e.g., gender, years of experience, educational background), Section B has indicators of key constructs that measure user control.

First, a pilot study was conducted to ensure content validity and reliability of constructs. Then, a purposive sampling technique was used in selecting the respondents. Purposive sampling is more appropriate for in-depth research where expertise or specific characteristics are critical (Akpan & Piate, 2023). This is to ensure that only experienced professionals participate in the survey. A total of 55 valid responses were collected for the main survey. The adequacy of this sample size is supported by evidence from purposive sampling literature and SEM studies, which indicate that meaningful saturation and reliable parameter estimation can be achieved with similar sample sizes when constructs are well defined (Etikan et al., 2016; Lakens, 2022). The method of sampling and determining the sample size are of great importance in research problems that rely on surveys. Determining the sample size is not an issue in purposive sampling. The survey instruments were administered using online Google Forms; in some cases, direct email contacts were used to reach out to the respondents by targeting cloud professionals globally through their LinkedIn profiles. Responses were then collected and analysed via SPSS and smartPLS.

User Control Measures

The study identified six key measures relevant to user control in cloud data migration:

1. Security: involving issues of confidentiality, integrity, data loss, and privacy.
2. Cost: covers connectivity, storage, application, data, and consultancy expenses.
3. Legal: examines adherence to regulations compliance, service-level agreements (SLAs), governance, and policies.
4. Personnel Knowledge assesses the technical, business, and communication skills of professionals involved in migration.
5. Standards: includes best practices, authentication, strategy, and industry.
6. Performance: evaluates quality of service, reliability, throughput, transfer speed, downtime, and user experience.

Each measure was incorporated into the survey instrument to assess its influence on SaaS users. The distribution of items within the constructs is presented in Table 2.

Table 2
The measure of user control in cloud data migration process

SN	Constructs	Measurement Items
1	Security	SECU1, SECU2, SECU3. SECU4
2	Cost	COST1, COST2, COST3, COST4
3	Legal	LGAL1, LGAL2, LGAL3, LGAL4
4	Personnel knowledge	PKNW1, PKNW2, PKNW3
5	Standards	STND1, STND2, STND3, STND4, STND5, STND6, STND7, STND8, STND9, STND10, STND11, STND12
6	Performance	PERF1, PERF2, PERF3, PERF4, PERF5, PERF6, PERF7, PERF8, PERF9, PERF10, PERF11, PERF12

Respondents

The pilot study targeted professionals involved in the cloud computing field, including cloud users, Cloud Service Provider (CSP), cloud auditors, cloud academics/researchers, and other IT professionals, to assess the reliability of the survey instrument. Shrestha (2021) supposedly pointed out that, reliability measures the consistency of responses within a construct, critical to pilot studies, and is commonly evaluated using Cronbach's alpha.

Questionnaires were distributed via Google Forms, with an expected response range of 15 to 50 respondents, aligning with the minimum requirement for pilot studies (Bujang et al., 2024). The data collection process was closely monitored, and responses were checked at 33 points. At this number, the acceptable threshold for reliability assessment is met. The data collected from respondents were exported to excel sheet for cleaning and coding, followed by further analysis in SPSS. Findings from the analysis evaluated the reliability of the constructs, and the results are presented in Table 3.

Table 3
Reliability value of the constructs

SN	Constructs	Abbreviation in SPSS	Reliability
1	Security	SECU	0.882
2	Cost	COST	0.817
3	Legal	LGAL	0.856
4	Personnel knowledge	PKNW	0.799
5	Standards	STND	0.972
6	Performance	PERF	0.967

The reliability value shows that security has 0.882, cost has 0.817, legal has 0.856, personnel knowledge has 0.799, standards have 0.972, and performance has 0.967. This means that all the constructs have their reliability values within the acceptable threshold.

According to Edeh et al. (2023) Composite reliability or Cronbach's alpha values ranging from 0.60 to 0.70 are acceptable. This implies that the survey instrument demonstrates a high level of reliability. Hence, justifying its usage for the main survey.

Instrument Validation and Data Analysis

Before the field administration of the research instrument, several preparatory steps were taken. These included content validity by experts, identification of the target respondents to be used for the pilot study, selection of an appropriate sample size, administration of the research instrument, and conducting the reliability assessment as reported in the prior section.

Demographic Characteristics

Table 4 presents the demographic distribution of respondents who participated in the study. The data collected provides insight into the professional background, experience, educational qualifications, and familiarity with the on-premises to cloud data migration process.

Table 4
Respondents demographic information

Demographic Variable	Frequency	Percentage
Cloud Profession		
Cloud user	30	54.4
Cloud service provider	2	3.7
Cloud researcher/academic	8	14.6
Cloud auditor	2	3.7
Other IT professions	13	23.6
Total	55	100
Range of Years of Working Experience		
0-5	19	34.5
6-10	17	30.8
11-15	8	14.6
16-20	7	12.7
21-25	2	3.7
Above 26	2	3.7
Total	55	100

Table 4 (*continued*)

Demographic Variable	Frequency	Percentage
Highest Academic Qualification		
Bachelor	17	30.8
Master	29	52.7
PhD.	8	14.6
Others	1	1.9
Total	55	100
Familiarity with the On-premises to Cloud Data Migration Process		
No	17	30.8
Yes	38	69.2
Total	55	100
Level of Understanding of the Cloud Data Migration Process		
Basic	25	45.5
Intermediate	19	34.5
Expert	11	20.0
Total	55	100

Cloud Profession

The study identified most respondents (54.4%) to be cloud users, next to cloud users are other IT professionals (23.6%), this category belongs to computer science, but they are not professionals in cloud computing. 14.6% are either cloud researchers or academics. Whereas, cloud service providers and cloud auditors made the least proportion of the respondents with 3.7% each. The distribution of respondents indicates that the study engaged a diverse group of cloud computing professionals and other IT experts.

Years of Working Experience

The respondents had varying lengths of years of professional working experience. The highest range is 0-5 years (34.5%), followed by 6-10 years (30.8%), 11-15 years (14.6%), 16-20 years (12.7%), while 21-25 years, and above 26 years made the least range of working experience among respondents, with 3.7% each. This clearly indicates that a mixture of early career and seasoned professionals was involved in the study.

Highest Academic Qualification

52.7% of the respondents had a Master's degree, while 30.8% had a Bachelor's degree. 14.6% had PhD degrees, and 1.9% hold other qualifications, other than conventional degree qualifications. The academic composition reflects how well educated the respondents are; this suggests the credibility of their responses.

Familiarity with On-premise to Cloud Data Migration

About 69.2% of the respondents indicated their familiarity with the on-premise to cloud data migration process, while 30.8% indicated a trace of knowledge in on-premise to cloud, but are without practical or prior experience of the process. The percentage highlights that most respondents had relevant knowledge of the research context, which makes their feedback valuable.

Level of Understanding of Cloud Data Migration

In terms of hierarchy of respondents' level of understanding; 45.5% had a basic understanding, 34.5% were rated intermediate in their understanding. While 20% of the respondents are experts. This distribution shows how the varying levels of expertise can provide a balance in assessing the response from different perspective which could reduce bias. The insight from the demographic data confirms the appropriateness of the selected respondents for the investigation and their suitability in all the perspectives of the study objectives.

RESULTS AND DISCUSSION

Summary of Statistics

The outcome of the pilot study provided insights into the reliability of constructs and the distribution of responses among respondents. It also guided the main investigation through the collection of data and analysis for descriptive analysis. The descriptive statistics were employed to summarise demographic variables, including respondents' professional background, years of experience, educational qualifications, familiarity with the migration process in cloud data, and level of expertise.

The sample size consists of 55 respondents. The summary of the percentages of respondents' distribution includes cloud users (54.5%), IT professionals (23.7%), cloud researchers/academics (14.6%), CSP and cloud auditors had 3.6% each. The respondents' working experience ranges from 0-5 years to over 26 years, indicating the inclusion of both high and low professional levels in the study. The highest proportion of academic qualification among the respondents was a Master's degree (52.7%), followed by a bachelor's degree (30.8%). 14.5% have a PhD degree, and the least proportion was other IT professionals with 1.9%.

On the respondents' familiarity with cloud data migration, 69.1% were reported having prior experience in cloud data migration, while 30.9% were only familiar but had no direct practical involvement. Additionally, the respondents' level of understanding about cloud data migration had varying proportions; Those with basic knowledge are 45.5%, intermediate (34.5%), and expert (20.0%).

After the descriptive analysis, the research constructs were tested for reliability, and all the constructs had their Cronbach's alpha value above the acceptable threshold of 0.7.

Particularly, the range falls between 0.799 and 0.972, demonstrating a strong reliability of the constructs. This finding has proven the reliability of the research instrument to be suitable for the main study.

Evaluating the Measurement Model

The measurement model was assessed for reliability and validity of the constructs and their indicator. Having its basis in the hypothetical model of the measures in the user control framework for cloud data migration, as illustrated in Figure 1, the theoretical foundation for understanding the factors influencing the migration process and their corresponding control metrics was established through the framework. This led to the model that highlights the relationships among the constructs and their impact on each other. The framework also integrates standards and performance metrics as key control measures to ensure an efficient and secure migration process.

The hypothesis testing result as illustrated in the study measurement model shown in Figure 2, demonstrates that 4 out of the 9 hypotheses were not supported. The unsupported hypotheses are found in the relationships between security and performance, legal and standards, and the relationship that personnel knowledge has with both standards and performance. The supported hypotheses alongside their corresponding indicators have their relationship in the proposed framework mapped with a straight line, whereas the unsupported hypotheses are mapped with dotted lines.

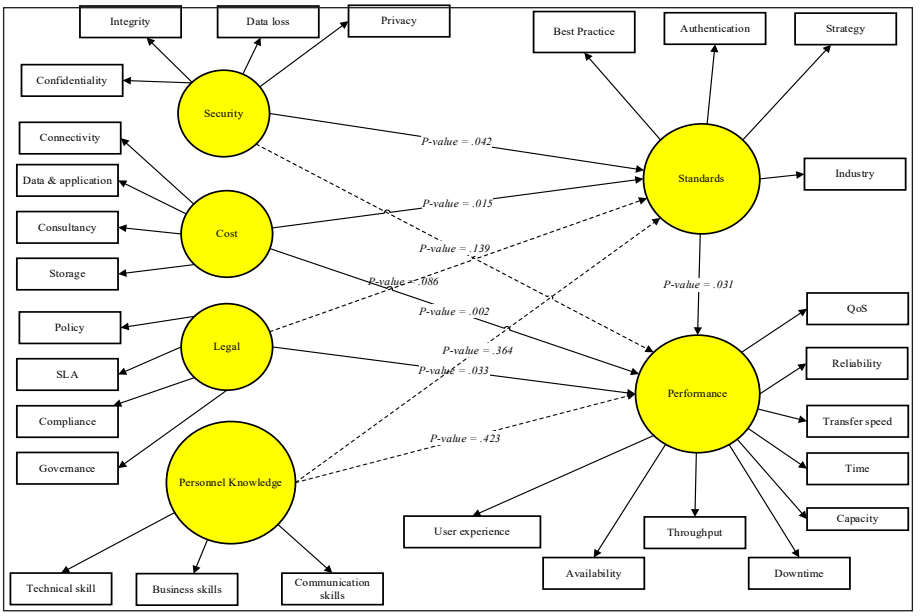


Figure 2. The study measurement model (a structural diagram representing relationships between security, cost, legal compliance, personnel knowledge, and migration success)

Further to this, Table 5 presents the relationships identified among the constructs, and the measurement properties of the measurement model.

Table 5
Reliability and validity of the constructs in the measurement model

Variable	Indicators	Factor Loading	Cronbach's Alpha	Composite Reliability	AVE
Cost	COST1	0.855	0.830	0.887	0.663
	COST2	0.853			
	COST3	0.745			
	COST4	0.800			
Legal	LGAL1	0.730	0.818	0.875	0.637
	LGAL2	0.828			
	LGAL3	0.846			
	LGAL4	0.811			
Performance	PERF1	0.788	0.951	0.957	0.648
	PERF2	0.780			
	PERF3	0.807			
	PERF4	0.818			
	PERF5	0.778			
	PERF6	0.855			
	PERF7	0.799			
	PERF8	0.782			
	PERF9	0.803			
	PERF10	0.825			
	PERF11	0.815			
	PERF12	0.808			
Personnel knowledge	PKNW1	0.751	0.753	0.852	0.650
	PKNW2	0.896			
	PKNW3	0.806			
Security	SECU1	0.793	0.854	0.901	0.696
	SECU2	0.835			
	SECU3	0.877			
	SECU4	0.829			

Table 5 (*continued*)

Variable	Indicators	Factor Loading	Cronbach's Alpha	Composite Reliability	AVE
Standards	STND1	0.839	0.955	0.961	0.674
	STND2	0.773			
	STND3	0.843			
	STND4	0.857			
	STND5	0.704			
	STND6	0.884			
	STND7	0.866			
	STND8	0.910			
	STND9	0.805			
	STND10	0.747			
	STND11	0.765			
	STND12	0.834			

Reliability Analysis of Constructs

Sequel to the Confirmatory Factor Analysis (CFA) results, the convergent validity, discriminant validity, and reliability of all the multi-variate constructs were analysed, based on the upheld guidelines from existing literature**:** (Ab Hamid et al., 2017). The Composite Reliability (CR) was assessed in terms of the degree to which construct indicators are free from random error and yield consistency at a range from 0.753 to 0.955, which justifies the degree to which the indicators are free from random error. Therefore, consistency in the constructs is upheld.

Using another technique of the measurement model, the discriminant validity, also known as divergence validity, was assessed through the Fornell and Larcker criterion cross-loadings, and the Heterotrait-monotrait ratio (HTMT) as the primary criterion for discriminant validity assessment. The HTMT paired matrix values for all the constructs are presented in Table 6. The validity results show that the HTMT values are below 0.9; as such, discriminant validity has been established. This is also reported so in Roemer et al. (2021) and Yusoff et al. (2020) where the HTMT criterion analysis was used for the validity of constructs.

Table 6
Discriminant validity - the HTMT paired matrix values of constructs

	COST	LGAL	PERF	PKNW	SECU	STND
COST						
LGAL	0.682					

Table 6 (*continued*)

	COST	LGAL	PERF	PKNW	SECU	STND
PERF	0.947	0.543				
PKNW	0.773	0.762	0.637			
SECU	0.564	0.713	0.623	0.663		
STND	0.740	0.656	0.799	0.632	0.682	

On the disposition of the Fornell-Larcker criterion, the square root of the average variance extracted (AVE) presents each latent construct and suggests that they ought to exceed the correlations of all the other latent constructs. Implying to note that the AVE for each construct should exceed its highest correlation with other constructs in the model. It is based on this principle that studies like Henseler et al. (2015) and Hilkenmeier et al. (2020) leveraged on the Fornell-Larcker criterion to establish their claims.

Hypothesis Testing

The result of the hypotheses tested in a relational model using Structural Equation Modelling (SEM), as presented in Table 7 and Figure 2, showed that security has a significant influence on standards, with a p-value of 0.042, at a 0.05 significance level. Therefore, H1 was supported. This finding highlights the importance of implementing security measures by addressing confidentiality, data loss, data integrity, and privacy concerns in accordance with cloud standards to enhance user control in the data migration process. For H2, which suggested that security has a significant impact on performance in the control of the data migration process, with a p-value of 0.139, indicating that there was no support for the hypothesis, hence security has no impact on the performance of users in the control of the migration project. H3, has it that cost is effectively high on standards in the control of the data migration process, with a p-value of 0.015; the hypothesis was supported. Similarly, H4 cost shows a significant effect on performance in the control of the data migration process, with a p-value of 0.002; hence, the hypothesis was supported. Whereas H5 postulated that legal factors have a significant effect on standards in the control of the data migration process. The hypothesis was not supported due to a p-value of 0.086. Legal factor, however, shows a significant effect on performance, hence H6 is supported with a p-value of 0.033. The hypotheses related to personnel knowledge (H7 and H8) are not supported, with p-values of 0.364 and 0.423, respectively. Suggesting that personnel knowledge has no significant influence on either standards or performance. This implies that personnel knowledge may

be a baseline competency in cloud environments; thus, variance may be insufficient to predict control outcomes. Alternatively, structured standards may mitigate dependence on individual expertise. Finally, the impact of standards on performance (H9) is supported, with a p-value of 0.031, indicating that standards significantly contribute to performance outcomes in the model.

Table 7
Path coefficient analysis and hypotheses testing

Hypothesis	P-value	Decision
H1: SECU → STND	0.042	Supported
H2: SECU → PERF	0.139	Not supported
H3: COST → STND	0.015	Supported
H4: COST → PERF	0.002	Supported
H5: LGAL → STND	0.086	Not supported
H6: LGAL → PERF	0.033	Supported
H7: PKNW → STND	0.364	Not supported
H8: PKNW → PERF	0.423	Not supported
H9: STND → PERF	0.031	Supported

Effect Size of Variables

Table 8 presents the effect size (F-square values) of independent variables, showing their impact on Performance (PERF) and Standards (STND) in the model.

Table 8
Effect size of variables

Variable	F^2	Effect Size
COST → PERF	0.977	Large
COST → STND	0.213	Moderate
LGAL → PERF	0.066	Small
LGAL → STND	0.021	Small
PKNW → PERF	0.002	No effect
PKNW → STND	0.001	No effect
SECU → PERF	0.055	Small
SECU → STND	0.141	Small
STND → PERF	0.288	Moderate

The F^2 from the Table based on Cohen (1988) guidelines, an effect size is considered small if F^2 is ≥ 0.02 and < 0.15, moderate if ≥ 0.15 and < 0.35, while F^2 is considered large if it is 0.35 or above. While an effect size of < 0.02 means no effect. Thus, the path

from COST to PERF shows a large effect size ($F^2 = 0.977$), suggesting that cost has a substantial and dominant influence on performance within the model. The paths COST to STND, SECU to STND, and STND to PERF all demonstrate small and moderate effect sizes ($F^2 = 0.213$, 0.141, and 0.288, respectively), indicating that these relationships have meaningful, though not dominant, impacts on the associated dependent variables. On the other hand, LGAL to PERF ($F^2 = 0.066$), LGAL to STND ($F^2 = 0.021$), and SECU to PERF ($F^2 = 0.055$) display small effect sizes, showing that legal considerations and security contribute modestly to performance and standards. Notably, the paths from PKNW (Personnel Knowledge) to PERF ($F^2 = 0.002$) and PKNW to STND ($F^2 = 0.001$) exhibit negligible effect sizes, implying that personnel knowledge has no impact on performance or standards in this context.

Practical Application of the Proposed User Control Framework

The proposed user control framework provides a structured and practical approach for assessing and managing user control throughout the cloud data migration lifecycle. It aligns with key migration phases: pre-migration planning, risk assessment, data preparation, migration execution, and post-migration validation and optimisation. while embedding a dedicated user control module that ensures continuous evaluation and intervention. This module projects user control by integrating measurable constructs, including standards (e.g., best practices, authentication, and strategy), performance (e.g., reliability, transfer speed, and availability), and influencing factors such as security, cost, legal compliance, and personnel knowledge.

A key practical feature of the framework is the computation of a User Control Level (UCL), which provides a quantitative basis for decision-making. The UCL is evaluated against a defined threshold range to determine whether migration conditions are adequate. When the control level meets the acceptable threshold, migration activities can proceed with confidence. However, if the control level falls below the required standard, the framework triggers a feedback mechanism, prompting users to reassess and improve relevant control measures before continuing.

This iterative evaluation process makes the framework directly applicable as a decision-support and diagnostic tool for organisations. It enables practitioners to monitor readiness, identify weaknesses in control mechanisms, and implement corrective actions in real time. By translating abstract control concepts into measurable indicators and actionable steps, the model enhances transparency, accountability, and effectiveness in cloud data migration processes.

CONCLUSION

This study examines the factors that influence user control in the cloud data migration process and analyses the relationships between the constructs and their indicators. The reliability values of the constructs of the model show internal consistency with a Cronbach's alpha value greater than 0.80; this value exceeds the recommended threshold of 0.70, indicating consistency (Sarstedt et al., 2019). The constructs of this study were also validated through the measurement model. Additionally, the convergent and discriminant validity of all constructs/indicators were established through AVE and CR. The final model was tested in SEM to validate the hypotheses and examine the effect of each construct on the data migration process. Results from further analysis highlight constructs that are supported by the hypothesis statement, showing the significance of the relationship among constructs of user control affecting factors and the control metric that measures their outcome. This provides users with key constructs to consider in the cloud data migration process. The proposed model could provide a comprehensive framework for user control measurement in cloud data migration. The implications of this study also extend to industry practitioners, policymakers, and researchers, based on the multidisciplinary context that was integrated to support the model development. This suggests that future research could translate the framework into actionable guidance.

Although cloud service provider size was not examined in this study, it represents an important contextual variable that may moderate the relationship between TOE factors and user control outcomes. Future research should investigate cloud service providers' size as a moderating variable, to better understand how providers size can affect customers' support and user experience. Potentially through multivariate analysis comparing large and small cloud service providers.

ACKNOWLEDGEMENT

No grant for this study. The authors warmly acknowledge the support provided by the academic experts and professionals who contributed their valuable insights and guidance throughout the study. Their support and constructive feedback were instrumental in refining the research framework and enhancing the overall quality of this work.

REFERENCES

Ab Hamid, M. R., Sami, W., & Mohmad Sidek, M. H. (2017). Discriminant validity assessment: Use of Fornell and Larcker criterion versus HTMT criterion. *Journal of Physics: Conference Series, 890*(1), Article 012163. https://doi.org/10.1088/1742-6596/890/1/012163

Abdelzaher, T., Diao, Y., Hellerstein, J. L., Lu, C., & Zhu, X. (2008). Introduction to control theory and its application to computing systems. In Z. Liu & C. H. Xia (Eds.), *Performance modelling and engineering* (pp. 185-215). Springer. https://doi.org/10.1007/978-0-387-79361-0_7

Abou_el_ela Abdou Hussein, A. A. (2021). Data migration need, strategy, challenges, methodology, categories, risks, uses with cloud computing, and improvements in its using with cloud using suggested in the proposed model (DMig 1). *Journal of Information Security, 12*(1), 79-103. https://doi.org/10.4236/jis.2021.121004

Akpan, E. E., & Piate, R. C. (2023). Assessment of different methods of sampling techniques: The strengths and weaknesses. *Shared Seasoned International Journal of Topical Issues, 9*(1), 64-83.

Ali, O., Shrestha, A., Osmanaj, V., & Muhammed, S. (2021). Cloud computing technology adoption: An evaluation of key factors in local governments. *Information Technology & People, 34*(2), 666-703. https://doi.org/10.1108/ITP-03-2019-0119

Alkhalil, A., Sahandi, R., & John, D. (2013). Migration to cloud computing: The impact on IT management and security-An exploration of senior technical and security professional views. In *Proceedings of the 1st International Workshop on Cloud Computing and Information Security* (pp. 196-200). Atlantis Press. https://www.atlantis-press.com/article/9904.pdf

Al Mudawi, N., Beloff, N., & White, M. (2022). Developing a framework of critical factors affecting the adoption of cloud computing in government systems (ACCE-GOV). In K. Arai (Ed.), *Intelligent computing* (Lecture Notes in Networks and Systems, Vol. 283, pp. 520-538). Springer. https://doi.org/10.1007/978-3-030-80119-9_32

Amin, R., Vadlamudi, S., & Rahaman, M. M. (2021). Opportunities and challenges of data migration in cloud. *Engineering International, 9*(1), 41-50. https://doi.org/10.18034/ei.v9i1.529

Bhatt, P. C. P., & Sehgal, N. K. (2024). *Project management in cloud applications*. Springer. https://doi.org/10.1007/978-3-031-53890-2

Bujang, M. A., Omar, E. D., Foo, D. H. P., & Hon, Y. K. (2024). Sample size determination for conducting a pilot study to assess the reliability of a questionnaire. *Restorative Dentistry & Endodontics, 49*(1), Article e3. https://doi.org/10.5395/rde.2024.49.e3

Chowdhury, R. A. (2024). *Key factors and challenges in cloud adoption: A case study of Kartverket* [Master's thesis, University of South-Eastern Norway].

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203771587

Dědečková, N. (2020). Control, controlling and its objectives in the organisation. *SHS Web of Conferences, 83*, Article 01009. https://doi.org/10.1051/shsconf/20208301009

Dziadosz, P., Granosik, E., Hieronimus, S., Marciniak, T., Mier, J., Novak, J., Pastusiak, B., Purta, M., & Sokoliński, O. (2021). *Cloud 2030: Capturing Poland's potential for accelerated digital growth*. McKinsey & Company. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/cloud-2030-capturing-polands-potential-for-accelerated-digital-growth

Edeh, E., Lo, W.-J., & Khojasteh, J. (2023). Review of *Partial least squares structural equation modeling (PLS-SEM) using R: A workbook*. *Structural Equation Modelling: A Multidisciplinary Journal, 30*(1), 165-167. https://doi.org/10.1080/10705511.2022.2108813

Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics, 5*(1), 1-4. https://doi.org/10.11648/j.ajtas.20160501.11

Fisher, W., Craft, R., Ekstrom, M., Sexton, J., & Sweetnam, J. (2023). *Data confidentiality: Identifying and protecting assets against data breaches* (NIST Special Publication 1800-28). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.1800-28

Grob, S., & Schill, A. (2012). Towards user-centric data governance and control in the cloud. In J. Camenisch & D. Kesdogan (Eds.), *Open problems in network security: iNetSec 2011* (pp. 132-144). Springer. https://doi.org/10.1007/978-3-642-27585-2_11

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modelling. *Journal of the Academy of Marketing Science, 43*(1), 115-135. https://doi.org/10.1007/s11747-014-0403-8

Hilkenmeier, F., Bohndick, C., Bohndick, T., & Hilkenmeier, J. (2020). Assessing distinctiveness in multidimensional instruments without access to raw data-A manifest Fornell-Larcker criterion. *Frontiers in Psychology, 11*, Article 223. https://doi.org/10.3389/fpsyg.2020.00223

Iqbal, A., & Colomo-Palacios, R. (2019). Key opportunities and challenges of data migration in cloud: Results from a multivocal literature review. *Procedia Computer Science, 164*, 48-55. https://doi.org/10.1016/j.procs.2019.12.153

Islam, R., Patamsetti, V., Gadhi, A., Gondu, R. M., Bandaru, C. M., Kesani, S. C., & Abiona, O. (2023). The future of cloud computing: Benefits and challenges. *International Journal of Communications, Network and System Sciences, 16*(4), 53-65. https://doi.org/10.3390/jrfm10020010

Islam, S., Fenz, S., Weippl, E., & Mouratidis, H. (2017). A risk management framework for cloud migration decision support. *Journal of Risk and Financial Management, 10*(2), Article 10. https://doi.org/10.3390/jrfm10020010

Kamal, A. A., Shaipullah, N. M., Truna, L., Sabri, M., & Junaini, S. N. (2020). Transitioning to online learning during the COVID-19 pandemic: Case study of a pre-university centre in Malaysia. *International Journal of Advanced Computer Science and Applications, 11*(6), 217-223. https://doi.org/10.14569/IJACSA.2020.0110628

Khan, R. A., & Khan, D. A. (2019). Cloud migration: Standards and regulatory issues with their possible solutions. *International Journal of Advanced Networking and Applications, 10*(6), 4113-4119. https://doi.org/10.35444/ijana.2019.100613

Lakens, D. (2022). Sample size justification. *Collabra: Psychology, 8*(1), Article 33267. https://doi.org/10.1525/collabra.33267

Laurent de la Vaissière, S. H. (2011). *How to ensure control and security when moving to SaaS/cloud applications*. Deloitte. https://www.deloitte.com/assets/Dcom-Global/Local%20Assets/Documents/TMT/dttl_tmt_CloudAdoptionSurveyDeloitte_CIONet.pdf

Malouche, H., Ben Halima, Y., & Ben Ghézala, H. (2017). Enterprise preparation for cloud migration: Assessment phase. In *2017, IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)* (pp. 652-659). IEEE. https://doi.org/10.1109/AICCSA.2017.23

Maniah, Soewito, B., Lumban Gaol, F., & Abdurachman, E. (2022). A systematic literature review: Risk analysis in cloud migration. *Journal of King Saud University-Computer and Information Sciences, 34*(6), 3111-3120. https://doi.org/10.1016/j.jksuci.2021.01.008

Mansouri, Y., Nadjaran Toosi, A., & Buyya, R. (2019). Cost optimisation for dynamic replication and migration of data in cloud data centres. *IEEE Transactions on Cloud Computing, 7*(3), 705-718. https://doi.org/10.1109/TCC.2017.2659728

Manyuchi, M. M., & Sukdeo, N. (2021). Importance of the control function in effective operational planning and rational decision-making in mining organisations. In *Proceedings of the International Conference on Industrial Engineering and Operations Management* (pp. 407-411). IEOM Society International. https://doi.org/10.46254/SA02.20210215

Murtezani, D. R., & Metin, B. (2016). Assessing cloud computing readiness and adoption. In *2016, IEEE 14th International Symposium on Intelligent Systems and Informatics (SISY)* (pp. 263-266). IEEE. https://doi.org/10.1109/SISY.2016.7601509

Oladoyinbo, T. O., Adeyemo, I. A., & Oladoyinbo, O. B. (2024). The TOE theory and cloud computing: Exploring factors affecting the adoption of cloud computing. *International Journal on Cybernetics & Informatics, 13*(4), 33-45. https://doi.org/10.5121/ijci.2024.130404

Pahl, C., & Xiong, H. (2013). Migration to PaaS clouds-Migration process and architectural concerns. In *2013, IEEE 7th International Symposium on the Maintenance and Evolution of Service-Oriented and Cloud-Based Systems* (pp. 86-91). IEEE. https://doi.org/10.1109/MESOCA.2013.6632740

Patrick, Z. P. G., & Satyanarayana, K. V. V. (2020). Optimisation of service level agreements (SLAs) within SaaS cloud IT infrastructure. *Journal of Critical Reviews, 7*(1), 414-420. https://doi.org/10.31838/jcr.07.01.82

Perra, C. (2015). A framework for user control over media data based on a trusted point. In *2015 IEEE International Conference on Consumer Electronics (ICCE)* (pp. 1-2). IEEE. https://doi.org/10.1109/ICCE.2015.7066294

Pochyla, M. (2019). Smart contracts and Internet of Things future. In *Proceedings of the 12th International Conference on Strategic Management and Its Support by Information Systems* (pp. 461-470). https://doi.org/10.1016/j.procs.2019.09.460

Rahulamathavan, Y., Pawar, P. S., Burnap, P., Rajarajan, M., Rana, O. F., & Spanoudakis, G. (2014). Analysing security requirements in cloud-based service level agreements. In *Proceedings of the 7th International Conference on Security of Information and Networks* (pp. 73-76). Association for Computing Machinery. https://doi.org/10.1145/2659651.2659735

Rao, G. M., Srinivas, K., Samee, S., Venkatesh, K., Dadheech, P., Raja, L., & Yagnik, G. (2021). A secure and efficient data migration over cloud computing. *IOP Conference Series: Materials Science and Engineering, 1099*(1), Article 012082. https://doi.org/10.1088/1757-899X/1099/1/012082

Roemer, E., Schuberth, F., & Henseler, J. (2021). HTMT2-An improved criterion for assessing discriminant validity in structural equation modelling. *Industrial Management & Data Systems, 121*(12), 2637-2650. https://doi.org/10.1108/IMDS-02-2021-0082

Sadoughi, F., Ali, O., & Erfannia, L. (2020). Evaluating the factors that influence cloud technology adoption: Comparative case analysis of health and non-health sectors: A systematic review. *Health Informatics Journal, 26*(2), 1363-1391. https://doi.org/10.1177/1460458219879340

Sarmah, S. S. (2018). Development of a general data migration framework in a case organisation. *Science and Technology, 8*(1), 1-10.

Sarstedt, M., Hair, J. F., Cheah, J. H., Becker, J. M., & Ringle, C. M. (2019). How to specify, estimate, and validate higher-order constructs in PLS-SEM. *Australasian Marketing Journal, 27*(3), 197-211. https://doi.org/10.1016/j.ausmj.2019.05.003

Shrestha, N. (2021). Factor analysis as a tool for survey analysis. *American Journal of Applied Mathematics and Statistics, 9*(1), 4-11. https://doi.org/10.12691/ajams-9-1-2

Tornatzky, L. G., Fleischer, M., & Chakrabarti, A. K. (1990). *The processes of technological innovation*. Lexington Books.

Vadlamani, S., Agarwal, N., Chintha, V. R., Shrivastav, E. A., Jain, S., & Goel, O. (2023). Cross-platform data migration strategies for enterprise data warehouses. *International Research Journal of Modernization in Engineering, Technology and Science, 5*(11), 1-10. https://doi.org/10.56726/IRJMETS46858

Varma, Y. (2024). Performance optimisation in cloud-based ML training: Lessons from large-scale migration. *American Journal of Data Science and Artificial Intelligence Innovations, 4*, 109-126.

Yang, Z., Sun, J., Zhang, Y., & Wang, Y. (2015). Understanding SaaS adoption from the perspective of organisational users: A tripod readiness model. *Computers in Human Behavior, 45*, 254-264. https://doi.org/10.1016/j.chb.2014.12.022

Yusoff, A. S. M., Peng, F. S., Razak, F. Z. A., & Mustafa, W. A. (2020). Discriminant validity assessment of religious teacher acceptance: The use of HTMT criterion. *Journal of Physics: Conference Series, 1529*(4), Article 042045. https://doi.org/10.1088/1742-6596/1529/4/042045

Zboril, M., & Svata, V. (2022). Cloud adoption framework. *Procedia Computer Science, 207*, 483-493. https://doi.org/10.1016/j.procs.2022.09.103

Pertanika J. Sci. & Technol. 34 (3): 1829 - 1859 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Towards Inherently Safer Dry Reforming for Carbon Capture: Trends, Gaps, and Insights

Fatin Aliya[1,2], Mardhati Zainal Abidin[3]*, and Azizul Buang[1]

[1]Department of Chemical Engineering, Faculty of Engineering, Universiti Teknologi PETRONAS, 32610 Bandar Seri Iskandar, Perak Darul Ridzuan, Malaysia

[2]Faculty of Ocean Engineering Technology, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia

[3]Department of Chemical Engineering, College of Engineering and Technology, University of Doha for Science and Technology, P.O. Box 24449, Doha, Qatar

ABSTRACT

Dry reforming of methane (DRM) is considered a promising technology to produce syngas since it offers possibilities for carbon dioxide (CO_2) conversion, sustainable generation of hydrogen, and environmentally friendly production. However, the efficiency of the current process, its sustainability, and its scalability remain critical considerations despite the numerous scientific developments in the field. Process safety needs special attention while addressing the problem of carbon deposition, catalyst deactivation and other hazards associated with DRM at the industrial level and operational risks, to ensure the practicality and reliability of large-scale applications. This study aims to evaluate global research trends at the intersection of DRM and process safety through a bibliometric analysis. In total, 1,036 research papers collected from the Scopus database within the timeframe of 1997-2024 were used in this study via a multi-phase process of material collection and analysed using bibliometric techniques, including publication trend analysis, country and author productivity assessment, collaboration network analysis, and keyword co-occurrence mapping. The results show a significant increase in research activity after 2017. Among the productive countries and authors, China and the USA are considered the key drivers and contributors to the topic. Four major safety-related themes have been defined as follows: safety in syngas production, carbon formation and catalyst deactivation, catalyst stability and anti-poisoning capabilities, and process intensification with innovative reactor design. These results indicate that safety issues are becoming more integrated into the DRM research field, especially through studies relating to operational stability, performance of reactors, and mitigating hazards.

ARTICLE INFO

Article history:
Received: 06 September 2025
Accepted: 07 June 2026
Published: 25 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.23

E-mail addresses:
fnaliya@umt.edu.my (Fatin Aliya)
Mardhati.Abidin@udst.edu.qa (Mardhati Zainal Abidin)
azizul.buang@utp.edu.my (Azizul Buang)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

This paper attempts to provide a systematic analysis of emerging safety themes, identify current research gaps, and propose future directions involving inherently safer design, predictive safety monitoring, and pilot-scale validation. The outcomes may support researchers and industry practitioners in developing safer and more reliable DRM technologies for large-scale implementation.

Keywords: Bibliometric analysis, carbon capture and utilisation (CCU), catalyst, dry reforming, sustainable syngas safety

INTRODUCTION

The increased demand for energy has altered the landscape due to industrialisation for years (Tanksale et al., 2010). Additionally, energy demand is expected to continue increasing despite the current low oil prices. Methane (CH_4) and Carbon Dioxide (CO_2) are the most abundant greenhouse gases and are believed to be the major contributors to climate change challenges (Ampomah et al., 2024). The emission of greenhouse gases leads to a surge in global temperatures, which is the leading factor in most ecological and environmental challenges. Therefore, it is paramount that greenhouse gases are reduced and utilised to maintain the balance required for the sustenance of the earth through various energy conversions. The primary components of greenhouse gases are CO_2 and CH_4. The production of CO_2 will come from the combustion of fossil fuels. At the same time, methane emissions can arise from various sources, including coal mining, biogas fermentation ecosystems, paddy fields, and manganese-rich constructed wetlands (Liu et al., 2023; Tucker et al., 2021).

According to the United States Environmental Protection Agency, in 2011, about one-third of all CH_4 emissions in the United States came from CH_4 production at landfills. Landfill gas is approximately 40-45% CH_4 and 55-60% CO_2 by volume (Abdullah et al., 2017). Methane dry reforming (DRM) typically occurs at high temperatures, as it is an endothermic process. Methane is the major component of natural gas. Most natural gas reservoirs are located far from industrial zones and are typically produced offshore. Accordingly, technical and financial considerations regarding the transportation of this valuable commodity from offshore locations to prospective consumers have resulted in massive flaring of natural gas worldwide (Lunsford, 2000). In response to the challenges posed by global climate change and the need for sustainable energy sources, Carbon dioxide Capture and Storage (CCS) has been implemented globally to reduce carbon dioxide emissions. DRM has received considerable attention due to its efficiency in converting CH_4 and CO_2 into syngas, which consists of hydrogen (H_2) and carbon monoxide (CO). The produced syngas from DRM has an optimal H_2/CO ratio of approximately 1, which is considered ideal for Fischer-Tropsch (F-T) synthesis of methanol, ammonia, and other chemical products (Bachmann et al., 2023; He et al., 2020).

DRM has attracted considerable interest in recent years due to its environmental and economic benefits, including CO_2 utilisation and syngas production. Its application in industry has grown rapidly, mainly due to improvements in catalyst design and reactor technology (Shi et al., 2020). The risk factors related to high temperatures, irregular pressure, flammable vapours, and potential catalyst deactivation are strongly manifested in the DRM. Therefore, with the advancement and widespread implementation of DRM technologies worldwide, safety requirements and design guidelines have been gradually developed for the systems (Chen et al., 2022; Gobbo et al., 2018). These requirements and guidelines support the safe operation and reliability

The additional hazards encountered during the operation and maintenance of the DRM process are summarised in Table 1. Therefore, to prevent a few accidents from occurring, safety performance in the design and operational phases of the DRM system should be comprehensively analysed (Amelse & Behrens, 2021). International safety standards and guidelines, including those developed by the International Organisation for Standardisation (ISO) and other regulatory bodies, outline safety requirements for high-temperature chemical reactors, which include DRM processes. Recently, DRM has gained significant attention due to its environmental and economic merits, including CO_2 utilisation and syngas production (Oyama et al., 2012). Additionally, the industrial utilisation of DRM has been developing rapidly, primarily through improvements in catalyst design and reactor engineering. A significant amount of work related to the safety aspects associated with DRM catalysts, reactor configuration, operating conditions, and process integration strategies has emerged in the literature (Soleimani & Lehner, 2022). Although several excellent reviews exist on safety-related DRM (Noor et al., 2013). A macroscopic overview of the knowledge map in this area is still lacking.

The research questions were developed after a preliminary review of the DRM literature, which revealed that most studies mainly focused on catalyst activity, syngas production efficiency and CO_2 conversion performance, while systematic discussion on process safety and inherent safety remains limited (Hussien et al., 2022; Sandoval-Diaz et al., 2022; Usman et al., 2015). Existing bibliographic studies also focus on catalytic thermodynamic aspects, without explicitly linking research developments to industrial safety considerations and safety engineering frameworks. As a result, this study aims to identify the main research themes in DRM and evaluate how these topics contribute to understanding safety in the design, operation and integration of catalysts, reactors and processes.

To address the gaps in the current understanding of dry reforming of methane, this study is conducting the following research questions: (1) What are the main research groups and trends in DRM, particularly in terms of catalyst design, reactor performance and process safety? (2) How can the insights of the literature inform the considerations

of process safety, including risks such as carbon deposition, catalyst deactivation, hotspot formation, and thermal runaway? Based on these questions, the study explores whether safety-related research themes can be systematically identified and interpreted through bibliometric mapping. Bibliometric analysis using VOSviewer allows for the identification of research clusters that can then be linked to specific safety-related aspects. For example, catalyst design groups inform strategies for preventing coking and sintering, while reactor design groups highlight the importance of temperature and pressure management. Process integration clusters provide insight into the adoption of inherent safety principles and standards. This approach strengthens the conceptual framework of research by linking knowledge trends to practical safety frameworks, highlighting areas for future research and industrial applications.

Although a quantitative risk ranking was beyond the scope of this bibliometric study, the identified hazards can be qualitatively prioritised based on their frequency of occurrence in the literature and potential impact on DRM operation. The most critical hazards, such as carbon monoxide release, hotspot formation, reactor overpressure, and carbon deposition, are considered due to their association with toxic exposure, thermal instability, equipment damage, and process disruption (Pawar et al., 2017). Catalyst deactivation and sulphur poisoning are generally regarded as secondary hazards.

Table 1
Common hazards during the expected lifetime of dry reforming

No.	Hazard Type	Hazard Situation
1	Mechanical Hazard	Reactor vessel failure due to prolonged exposure to high pressure and material fatigue. Catalyst bed collapse caused by uneven gas distribution and mechanical stress. Blockage in reactor tubes due to excessive carbon deposition, leading to increased pressure drop.
2	High-Temperature Hazard	Localised overheating (hotspots) caused by uneven heat distribution in the catalyst bed. Thermal stress cracking of reactor walls during rapid heating or cooling cycles. Thermal runaway reaction due to uncontrolled methane feed rate.
3	Explosion and Flammability Hazard	Formation of an explosive methane-air mixture due to leaks in the reactor system. Accumulation of unburned methane in the reactor during startup.
4	Toxic Gas Hazard	Accidental release of CO due to a leak in the reactor seal. Incomplete methane conversion, causing high CO emissions. Operator exposure to toxic gases during catalyst replacement.
5	Corrosion and Erosion Hazard	Reactor wall thinning caused by prolonged exposure to high-temperature CO_2. Catalyst support degradation reduces structural integrity. Corrosion-induced leakage in heat exchangers or reactor tubes.

Note. Adapted from (Chen et al., 2022; Lunsford, 2000; Miao et al., 2021; Oyama et al., 2012)

However, prolonged catalyst degradation can contribute indirectly to process safety concerns via process instability and increased carbon accumulation. Corrosion and material degradation remain important long-term integrity issues, particularly under high-temperature operating conditions (Durán et al., 2023).

Bibliometrics is the cross-cutting science of quantitative analysis of all knowledge carriers utilising mathematics and statistics. A comprehensive body of knowledge can be generated using bibliometrics that integrates mathematics, statistics, and bibliography and focuses on quantification (Merigó et al., 2019). In recent years, bibliometrics has been extensively employed to analyse and visually demonstrate the state, structure, hotspots, and research trends in specific fields (Osarogiagbon et al., 2021). A few researchers have conducted bibliometric studies in the field of dry reforming. Alhassan et al. (2022) performed a bibliometric analysis of dry reforming of methane for hydrogen production based on Web of Science (WoS) data. They set up the effect of certain parameters on the DRM reaction. A bibliometric study on biogas reforming for the synthesis of sustainable aviation fuel, providing a statistical overview of the literature on nickel catalyst reforming and the thermodynamic influences of temperature, pressure, and the H_2O/CH_4 ratio (Duarte et al., 2024; Li et al., 2020).

Although previous systematic research has contributed to valuable insights into catalyst development, hydrogen production, thermodynamic optimisation and sustainable fuel synthesis in DRM, it generally lacks a clear discussion of process safety and inherent safety considerations (Ávila-Neto et al., 2009). Unlike previous bibliographic analyses, this study focuses specifically on safety-related aspects of DRM, including catalyst deactivation, carbon deposition, thermal stability of reactors, toxic gas formation and operational risks. Furthermore, this study expands previous bibliometric approaches by linking identified research groups to established safety engineering frameworks and inherent safety principles, which provides practical implications for the safe implementation of industrial DRM. To the best of the authors' knowledge, based on searches conducted in the Scopus database using combinations of the keywords 'dry reforming of methane', 'bibliometric analysis', 'scientometric analysis', 'safety' and 'process safety', no previous bibliometric study examining DRM from a process safety perspective was identified.

METHODOLOGY

Bibliometric analyses were carried out to identify research trends, collaborative networks, and thematic clusters in the field of dry reforming of methane. VOSviewer was used to construct and visualise co-authorship, keyword co-occurrence, and citation networks. The tool is suitable for analysing large-scale datasets from databases such as Scopus and Web of Science, as it generates graphical maps showing research clusters, linking strengths, and emerging topics (van Eck & Waltman, 2010). Analysis followed standard library

procedures, including data extraction using relevant keywords, data cleaning to unify author and institution names, network construction based on co-occurrence and reference relationships, cluster identification using the VOSviewer cluster algorithm, and visualisation of networks to highlight dominant research themes, emerging trends, and knowledge gaps. Bibliometric mapping using VOSviewer is widely recognised as a rigorous method for exploring scientific fields and supporting empirical literature reviews

The research identified the search terms for retrieving articles through a systematic screening process. The research began with gathering articles from the Scopus database through focused online searches. The query string was refined by incorporating a broad range of safety-related terms associated with dry reforming of methane. To improve search coverage, the query string incorporated a wide range of process safety and operational risk terms in addition to “dry reforming”. These included hazard identification, risk assessment, process control, emergency shutdown systems, thermal runaway, hotspot formation, gas leakage, toxic emissions, catalyst deactivation, sulphur poisoning, and inherently safer design concepts. This strategy was intended to capture publications addressing safety-related DRM issues even when the term “safety” was not explicitly stated. The complete search string employed in this study is presented in Table 2.

1,506 articles were identified using the search string "dry reforming" and safety terms in the title and abstract fields across various document types, including journal articles, conference articles, and book chapters, from 1996 to 2024. Afterwards, the exclusion criteria were used to filter out duplicate papers, non-English language papers, and papers outside the defined scope, and hence, the number of papers was reduced to 1,036 journal articles, as shown in Table 3. Lastly, the filtering process included specific themes related to dry reforming safety, such as syngas production, carbon management, catalytic stability, poisoning resistance, and process intensification.

Table 2
Search string

TITLE-ABS ("Process safety" OR "Hazard identification" OR "HAZID" OR "Risk assessment" OR "Emergency shutdown system*" OR "Process control" OR "Inherent safe*" OR "Inherent* safe*" OR "inherent* safet* design*" OR "chemical route selection" OR "process design" OR "synthesis route" OR "ISD" OR “Carbon deposit*” OR "Material degradation" OR "High* temperature corrosion*" OR "Reactor design*" OR "Catalyst poison*" OR "Heat distribution" OR "Thermal runaway*" OR "Hotspot formation" OR "Temperature control*" OR "Thermal stress*" OR "Gas leak*" OR "CO exposure" OR "Toxic* emission*" OR "Flammable gas*" OR "catalyst deactivation" OR "sulfur poison*") AND ("dry reforming*" OR "dry reforming of methane" OR "dry reforming methane") AND PUBYEAR > 1995 AND PUBYEAR < 2025 AND (LIMIT-TO (DOCTYPE, "ar")) AND (LIMIT-TO (SRCTYPE, "j")) AND (LIMIT-TO (LANGUAGE, "English"))

Note. Search query performed in Scopus (accessed December 2024). Filters applied: journal articles, English language, publication years 1996 to 2024

Table 3
Refinement process

Stage	Description	Records
Initial Retrieval	Scopus search using DRM and safety-related keywords	1,506
Database Filtering	Publication year (1997–2024), English language, journal articles	1,094
Screening	Removal of duplicate and non-relevant publications	1,036
Final Dataset	Records included in the bibliometric analysis	1,036

Therefore, the three-stage filtering process ensures a systematic and comprehensive analysis of bibliometrics on dry reforming safety-related topics, as illustrated in Figure 1. Hence, a total of 1036 articles were identified through the final search string. The refinement process was used in the bibliometric analysis. Consequently, a total of 1036 articles were sourced from 81 countries/regions, 2,302 organisations, 239 sources, and 4,671 authors. Scopus was selected as the preferred database for this study because it provides broad coverage of journals in engineering, energy, catalysis, and process safety. This database also provides standardised bibliographic data suitable for use with VOSviewer software. Scopus has been widely used in bibliometrics due to its broad database coverage and citation indexing (Lam et al., 2023; Pranckutė, 2021). Nevertheless, it is acknowledged that the inclusion of other databases, such as Web of Science, Engineering Village, or Dimensions, may result in variations in publication counts, citation metrics, and collaboration networks due to differences in database coverage and indexing practices.

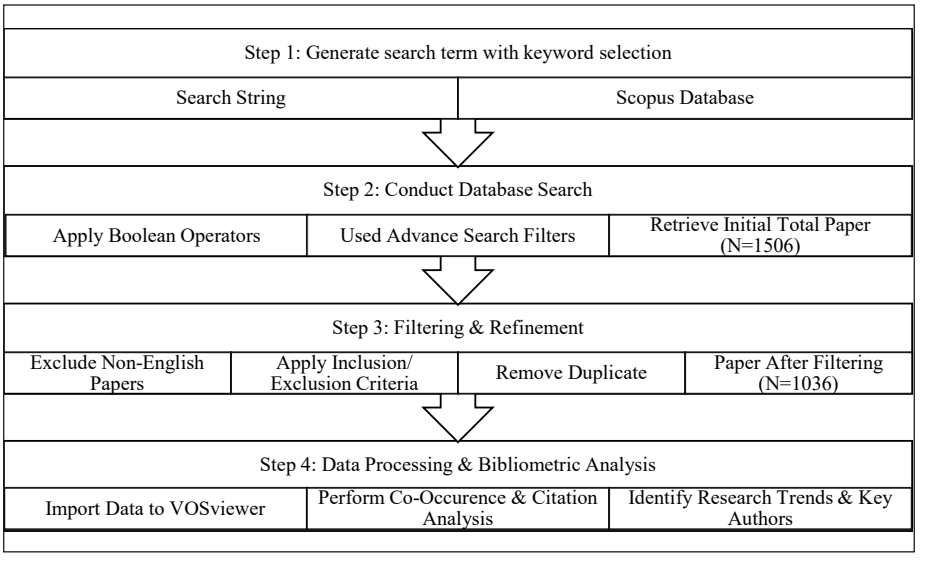


Figure 1. Search strategy and data refinement process for bibliometric analysis of safety-related DRM literature

RESULT AND DISCUSSION

Annual Publication Trend

The rapid growth of science and technology has significantly increased the volume of scientific knowledge. Scientific publication is crucial for disseminating knowledge. Hence, its evolution can indicate scientific improvement in many fields (Collivignarelli et al., 2020). The first published document in the reforming field was issued by Mleczko et al. (1997) from Germany, who proposed the reformer-combustor for CO_2 reforming of methane to achieve high yield and energy efficiency in synthesis gas production in 1997. In 2024, the total number of publications concerning dry reforming safety-related reached a historical peak of 170. Annual trends, cumulative data, and trends in global and major country publications are shown in Figure 2.

Between 1998 and 2021, 25 years, research publications on dry reforming safety-related have increased noticeably. The central active countries are China, the United States, Spain, the United Kingdom, and South Korea. Before 2010, research in this field was evenly distributed among all countries. After 2017, the number of Chinese publications on dry reforming safety-related issues increased sharply compared to other countries. The significant increase in dry reforming publications from China since 2017, as shown in Figure 2, is closely linked to the country's increased emphasis on utilising CO_2 and implementing low-carbon industrial transformation, as outlined in its national energy and climate strategies (Central Committee of the Communist Party of China, 2016). Dry reforming aligns well with these priorities by enabling the simultaneous conversion of syngas and CO_2, supported by increased government funding and strong institutional incentives for high-impact research results. Beginning in 2019, China has dominated publication output. There is a high prevalence of research institutions in China, which is related to the advancement of Chinese researchers and the volume of their publication output (Zhou & Leydesdorff, 2006; Zeng et al., 2020). Yet the increase in publications is mainly driven by the growth of research and by increasing interest in the field. It does not mean, however, that the impact of safety can be measured directly through publications.

Regarding the total number of publications per year, 2018 and 2022 saw the most significant increases compared to the previous year, with increases of 31% and 53%, respectively. Safety-related DRM research has not progressed at the same pace as catalyst and reactor development for industrial applications. Active and durable catalysts have been recently developed, and newly designed reactors have accelerated the commercialisation of dry reforming technologies. Recent advances in syngas production have been widely recognised as significant contributions to sustainable process development. Consequently, the number of publications related to safety in syngas production technologies has increased significantly since 2020.

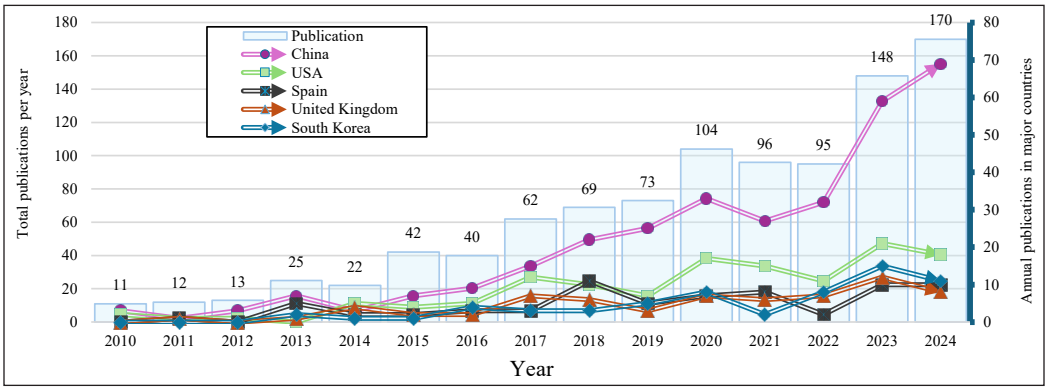


Figure 2. Annual trends in global and major country safety-related DRM publications

Source Journal Analysis

An academic journal is a peer-reviewed publication. Academic journals are the means of disseminating information from intellectual research. Most cases comprise research papers, review articles, short communications, and book reviews 49. A search on Scopus yields 1,036 publications published in 239 journals. Table 4 summarises the top 15 source journals, ranking by the number of publications in dry reforming safety-related. International Journal of Hydrogen Energy (IJHE) leads by a mile with 110 article publications. Chemical Engineering Journal (CEJ) is the most recent in the above list, with an average publication year of 2021.32. The highest average citation is Applied Catalysis B: Environmental (ACBE) at 81.09. According to the result, in terms of the CiteScore 2024, there were 8 journals with an average CiteScore above 10. Although placed fifth in the list, ACBE secured the highest CiteScore at 37.3 compared to other journals. The journal with the lowest CiteScore was Industrial and Engineering Chemistry Research (6.5). This may be attributed to its specialised focus, which restricts its audience, or the presence of other journals in the field that garner greater attention and citations. The journal's scope may not consistently correspond with emerging high-impact research areas, potentially leading to a lower citation rate relative to more widely influential publications. The CJE has not been indexed in Scopus since 1993. Some researchers utilise CiteScore as a metric to assess the annual readership of a journal. CiteScore is vital to authors in their decision of where to publish. CiteScore is a substitute for the Clarivate Analytics Impact Factor, which measures a journal's impact from citation data within the Scopus database. While CiteScore is useful, it should not be used to decide on a journal. Considering all factors aside from the CiteScore, what matters is the journal's contribution towards reaching the right audience and advancing the discipline (Shaifudin et al., 2022).

Table 4
The top 15 most productive journals on various research with their most cited articles from 1998 to 2024

Rank	Journal	NP	TP (%)	TC	CiteScore (2024)	APY	AC	Publisher
1	International Journal of Hydrogen Energy	110	10.62	3030	12.8	2019.13	27.55	Elsevier
2	Chemical Engineering Journal	50	4.83	1139	19.6	2021.32	22.78	Elsevier
3	Fuel	42	4.05	1124	13.8	2020.52	26.76	Elsevier
4	Industrial and Engineering Chemistry Research	34	3.28	1031	6.5	2015.53	30.32	American Chemical Society
5	Applied Catalysis B: Environmental	33	3.19	2676	37.3	2017.24	81.09	Elsevier
6	Applied Catalysis A: General	31	2.99	1484	8.3	2015.52	47.87	Elsevier
7	Catalysis Today	30	2.90	1141	10.9	2019.47	38.03	Elsevier
8	Catalysts	29	2.80	2189	7.4	2020.48	11.03	Multidisciplinary Digital Publishing Institute (MDPI)
9	Journal of CO2 Utilization	25	2.41	894	14.7	2020.28	35.76	Elsevier
10	Catalysis Science and Technology	21	2.03	540	8.0	2018.81	25.71	Royal Society of Chemistry
11	ACS Catalysis	19	1.83	616	19.0	2020.75	38.50	American Chemical Society
12	Chemical Engineering Science	18	1.74	527	7.6	2017.06	29.28	Scopus
13	Fuel Processing Technology	17	1.64	430	16.6	2018.35	25.29	Elsevier
14	Energy and Fuels	16	1.54	321	9.3	2019.44	20.06	American Chemical Society
15	Energy Conversion and Management	14	1.35	532	19.0	2020.64	39.43	Elsevier

Note. NP: Number of publications; TP: Total publications; TC: Total citations; APY: Average publication year; AC: Average citations per document

Quantitative Analysis of Productive Countries/Regions

A quantitative analysis of productive countries/regions can help researchers and institutions identify those with significant recent achievements in a particular research field. The geographical information of the corresponding authors in the 1036 publications shows that this safety-related DRM research is distributed in 81 countries/regions.

Table 5 lists the top 10 productive countries/regions, including the USA from North America, China, South Korea, India, Japan, and the United Arab Emirates from Asia, as well as the UK, Germany, Italy, and Spain from Europe. China produced a total of 322 publications, far more than any other country. This is credited to factors related to job titles and publications, government incentives, and the greater number of researchers. Spain has the largest AC of 47.12 and the most advanced APY of 2018.39, indicating the deepest research cumulation in dry reforming safety-related. This implies that the measures of publication productivity and research influence should be viewed independently. Even though China leads all other countries in the publication numbers concerning safety-related DRM research, using citation-based measures, there exist several countries that have relatively more research influence than their publications would imply. Consequently, publication productivity as well as citation-based measures need to be taken into consideration.

Table 5
The top 10 countries and academic institutions that publish the most various publications

Rank	Country	Continent	Publications	SCP (%)	AC	APY	TLS	TPI	The Most Productive Academic Institution
1	China	Asia	322	71.74	26.56	2020.39	88	40	University of Chinese Academy of Sciences
2	United States	North America	139	46.04	31.02	2018.96	74	14	Texas A&M University
3	Spain	Europe	74	45.95	47.12	2018.39	40	14	Universidad De Zaragoza
4	United Kingdom	Europe	71	19.72	35.18	2018.61	57	10	Imperial College London
5	South Korea	Asia	67	19.40	14.79	2019.93	25	7	Hanyang University

Table 5 (*continued*)

Rank	Country	Continent	Publications	SCP (%)	AC	APY	TLS	TPI	The Most Productive Academic Institution
6	Germany	Europe	59	37.29	30.97	2018.34	28	6	Ruhr-University Bochum
7	United Arab Emirates	Asia	54	16.67	20.63	2018.63	31	15	King Saud University
8	Italy	Europe	48	47.92	35.27	2018.63	20	7	Sapienza Università di Roma
9	Japan	Asia	48	54.17	21.23	2018.79	22	6	Kyushu University
10	India	Asia	44	63.64	25.41	2020.95	16	6	Birla Institute of Technology and Science

Note. SCP: Single country publication; AC: Average citation; APY: Average publication year; TLS: Total link strength; TPI: Total publication academic institution

The USA and Germany, supported by world-class research centres and secure funding initiatives, are the major drivers of the industry, demonstrating their passion for safety research and green energy. South Korea and Japan, leaders in innovation in fuel processing technology, have had consistent production capabilities, showcasing their passion for catalyst manufacturing and process safety improvements. On the other hand, new hotspots, such as India and the United Arab Emirates, have displayed increased activity by highlighting their focus on advanced research in sustainable hydrogen production and CO_2 utilisation. While China has the largest number of publications, Spain's citation-weighted rate indicates an increased scholarly contribution, potentially due to its strong connection with top European universities and industry leaders. These findings indicate that while China leads in research volume, countries such as Spain and Germany contribute significantly to the foundational and high-impact studies in this domain. The geographical distribution of research efforts underscores the need for fostering international collaboration to advance safety frameworks in dry reforming. Strengthening cross-border partnerships can enhance knowledge exchange, promote technology transfer, and ensure a more comprehensive understanding of safety challenges. Future research should focus on bridging the gap between high-output countries and high-impact research institutions to establish globally recognised safety standards for industrial-scale dry reforming applications.

Authors and Their Cooperation Network

Knowledge maps based on the author-co-author network illustrate international co-authorship patterns, major research groups, and leading contributors to dry reform in the safety area. Based on the Scopus dataset, a total of 4,671 authors from 1,036 publications were identified and formed 108 distinct collaboration groups. These clusters are generated using VOSviewer, based on co-authorship links, where the size of nodes represents the output of the publication, and the strength of the links reflects the intensity of the collaboration. The significance of the cluster was evaluated using a combination of quantitative and structural criteria, including cluster size (in terms of author number), total link strength, and the presence of highly productive and well-connected primary authors. Although many clusters were identified, most of them were small, weakly linked groups, indicating fragmentary and topical research efforts.

As a result, two clusters were classified as important, characterised by substantially greater membership, stronger internal collaboration, and sustained research productivity. These two dominant clusters, referred to as clusters 1 and 2, led by Bilbao Javier and Muhler Martin as shown in Figure 3, represent the most established and influential research groups in dry safety reform. The large number of independent clusters also indicates the absence of a globally integrated research network in this area, which can be attributed to the diversity of the technology reform and the wide range of safety-related challenges encountered by different research teams.

The team in cluster 1, led by Bilbao Javier, primarily focuses on the experimental development and simulation-based optimisation of advanced catalysts for hydrogen and syngas production through various reforming processes, including DRM and steam reforming of bio-oil and biomass volatiles (Santamaria et al., 2020a; Santamaria et al., 2020b). Nickel (Ni) and Rhodium (Rh) are used for synthesising catalysts that are doped with various promoters, such as Cerium Oxide (CeO_2), Magnesium Oxide (MgO), Calcium (Ca), and Lanthanum Oxide (La_2O_3), and supported on Aluminium oxide (Al_2O_3), Silicon Dioxide (SiO_2), and FeCrAl-fibre materials. X-ray diffraction (XRD), scanning electron microscopy (SEM/TEM), BET surface area, and transverse phase polarisation (TPR/TPO) are employed as the sophisticated characterisation methods for the evaluation of the structural and chemical properties of the catalysts in single cluster 1 research (Dang et al., 2021; Ferrandon et al., 2022; Valecillos et al., 2024). It applies reaction kinetics, density functional theory, and computational fluid dynamics for the reaction process, heat and mass transfer, and process optimisation (Martinez-Hernandez et al., 2024). The models can estimate the performance impact by varying catalyst composition and reactor conditions, allowing for stable and efficient system design. Under controlled high-temperature environments, performance testing is conducted to quantify stability, catalytic activity, and coke resistance. Gas chromatography measures product gases, including water, carbon monoxide, and methane (Liu et al., 2023; Vicente et al., 2014).

Moreover, numerous research works focus on different methods involving biogas synthesis to diesel by Fischer-Tropsch catalysts, methanol synthesis from steel mill off-gases, and partial methane oxidation (Mhadeshwar et al., 2007; Jiang et al., 2023). This study utilises industrial exhaust, alternative fuels, and extended chemical production targets beyond syngas and hydrogen. Reactor level performance evaluation, analysing yields of products like syngas, diesel, or methanol, and catalytic lifetime at challenging feedstocks like biogas or steel mill gases. These publications strongly emphasise the integration of innovative, safer catalyst design with broad mechanistic investigations that address the challenge of sustainability and enhance energy and chemical production conversion processes (He et al., 2020).

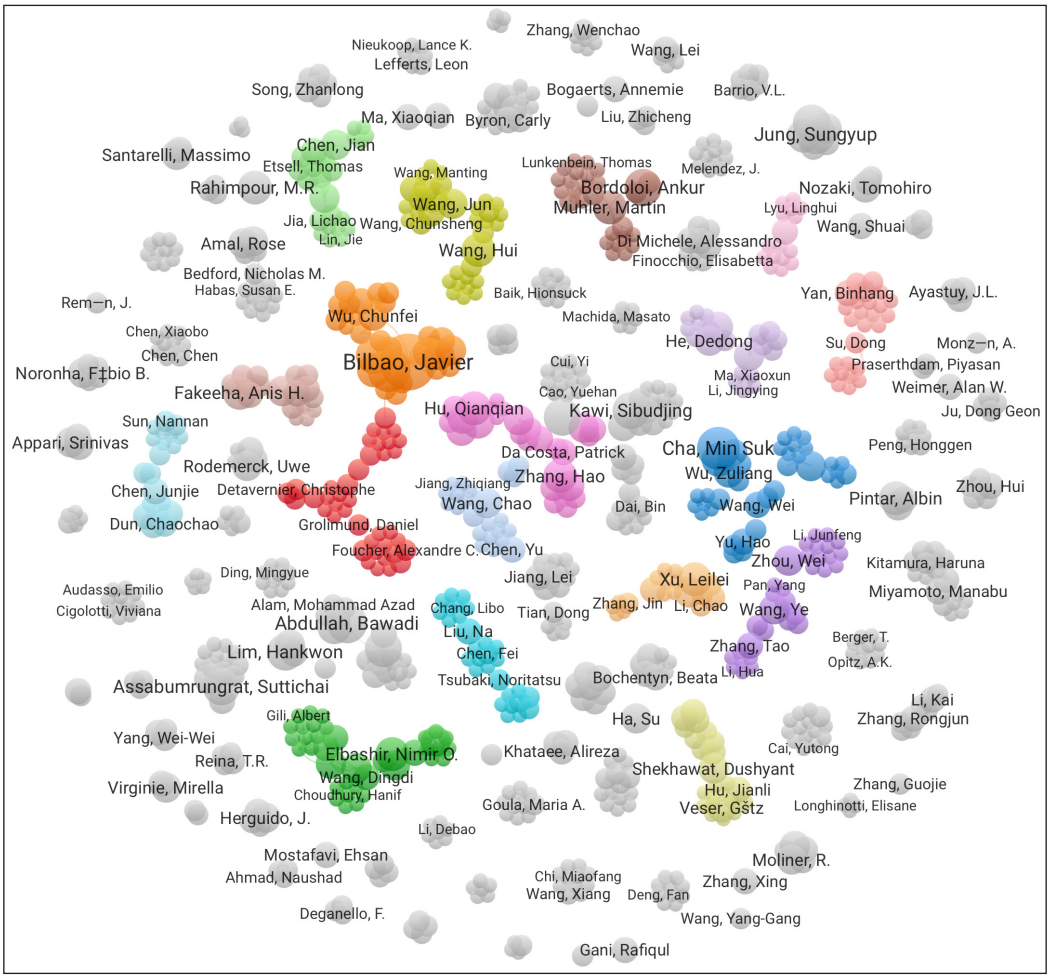


Figure 3. Overall author cooperation network of research on safety-related dry reforming

The second cluster, directed by Muhler Martin, exhibits notable similarities in its emphasis on developing new catalyst systems and comprehending their performance in high-temperature, energy-intensive processes. A key focus in these investigations is the development of high-performance catalysts that exhibit improved stability, activity, and resistance to deactivation. Catalytically active materials, such as Gadolinium-Ruthenium (Gd-Ru) nanoparticles on Zirconium-Cerium (Zr-Ce) nanorods, Nickel-based systems, and solution combustion-synthesised nanoalloys, are assessed for their performance under various operating conditions, including high temperatures and the presence of contaminants (Mette et al., 2016). Material design and nanostructure incorporation into support systems in alloys and modifications are vital aspects for maximising catalytic performance and stability at industrially feasible conditions (Das et al., 2021; Gazi et al., 2024). A second area is the application of in-situ and advanced characterisation techniques to investigate the activity of the catalyst in actual operating conditions. Particle dynamics and deactivation mechanisms, such as coke deposition, metal sintering, and poisoning by impurities, like steel mill off gases, are investigated using X-ray microscopy and spectroscopic techniques. (Beheshti Askari et al., 2020). In addition to researching cooperation patterns, clusters of authors identified have contributed to several important developments in safety-related DRM and scale-up, including enhancing catalyst stability, reducing carbon deposits, improving sulphur poisoning resistance, and optimising the design of reactors. These developments are important to reduce process instability, mitigate carbon-related operational problems, reduce hotspot formation, and improve the long-term reliability of industrial-scale DRM systems.

Table 6 shows the information on the top 10 authors with more than six publications. This table shows that three authors are from Spain, and one author is from the United Arab Emirates, South Korea, Malaysia, Singapore, Thailand, the United Kingdom, and Vietnam. While China leads in the total number of publications on the safety of dry reforming, it is remarkable that no Chinese authors feature among the top 10 most productive or highly cited authors. This may be due to several factors, not least of which is the popularity of long-list collaborations, in which Chinese researchers often take middle positions that limit individual recognition. This will lower the general citation impact of any researcher compared to the focus on specialised fields (Demeter et al., 2025). Citation metrics and journal visibility further extend this deficit, as many Chinese-authored articles were published in regional or lower-impact journals, which may lower their international reach and citation potential. These findings highlight a shortcoming of existing bibliometric studies, which often overlook the fine points of an author's contribution and the interaction among collaborators.

Table 6
List of the most prolific authors related to the study

Rank	Author	Country	Total Publication	TC	AC	APY	TLS
1	Bilbao, Javier	Spain	14	975	69.6	2018.8	54
2	Gayubo, Ana G.	Spain	8	646	80.8	2017.9	30
3	Cha, Min Suk	United Arab Emirates	8	215	26.9	2018.9	6
4	Jung, Sungyup	South Korea	7	82	11.7	2022.4	20
5	Abdullah, Bawadi	Malaysia	7	155	22.1	2021.4	15
6	Kawi, Sibudjing	Singapore	6	376	62.7	2020.7	22
7	Remiro, Aingeru	Spain	6	191	31.8	2020.2	22
8	Assabumrungrat, Suttichai	Thailand	6	146	24.3	2020.3	20
9	Tu, Xin	United Kingdom	6	407	67.8	2020.2	20
10	Vo, Dai-Viet N.	Vietnam	6	135	22.5	2021.2	20

Note. TC: Total citation; APY: Average publication year; AC: Average citation per document

Bibliometric analysis of leading authors in dry reforming and safety reveals that most highly cited works originate from Spain, with Javier Bilbao and Ana G. Gayubo as the leading contributors. First among them is Javier Bilbao, who has amassed an overall citation count of 975, emerging as a forceful researcher. He has a high average annual citation rate of 69.6 and is highly relevant and cited in the academic world. Thus, he has an academic influence spanning several years. Due to his consistently high citation rates, Bilbao is also recognised as a primary specialist in dry reforming research related to safety. Ana G. Gayubo from Spain ranks second, with 646 citations from 8 publications and excellent academic visibility. The research by Gayubo is from 2017 and represented one steady contribution to the field, presenting a critical part of the continuous investigation and mentioning the safety of dry reforming.

The data also reveal a significant representation from several countries. For instance, the UAE's Min Suk Cha and South Korea's Sungyup Jung, although having few publications, have contributed much to the field. Cha, on the other hand, with one publication and eight citations, exhibits a lesser impact in citation count yet shows promise for future research output. Sungyup Jung has made significant contributions over the last three years, as evidenced by seven publications and 82 citations, with an average publication year of 2022. His influence in the field is on the rise, as evidenced by his recent publications and citation data. International growth is enjoyed by Abdullah Bawadi of Malaysia, with seven publications and 155 citations, and Sibudjing Kawi of Singapore, with 6 publications and 376 citations.

The high citation counts for Kawi, combined with an average of 62.7 citations per year, underscore that his work is gaining increased momentum in the literature, especially in recent years, 2020 and 2021. Therefore, LS is part of this bibliometric research, which quantifies the network of co-authors and collaborations. High values of LS indicate that an author who is actively involved in the partnership for the diffusion and impact of their research can have a more significant impact. Ana G. Gayubo shows an LS of 30, which is very high, with a notable degree of multidisciplinary in several research projects she has participated in. Kawi has 22 TLS, which would disclose a collaboration strategy that extends the scope of his work. Other authors, such as Aingeru Remiro from Spain with 22 LS, Assabumrungrat from Thailand with 20 LS, and Xin Tu from the UK with 20 LS, are highly collaborative, which speaks in many instances to the prestige among scholars and the integration of their contributions within a broader academic community. These researchers, although somewhat fewer overall in citations or less in several publications, supplement this developing network of scholars in dry reformation and safety.

This bibliometric analysis highlights the leading role played by Spanish scholars Bilbao and Gayubo, as well as the increasing international contributions from primarily Southeast Asia and the Middle East. The steady rise in APY values for many authors, combined with their increasing total citations and LS, reflects an increased emphasis on dry reforming and safety in recent years, as young authors begin to be recognised. Safety in dry reforming represents an active domain of investigation. At the same time, it is undergoing continuous development and is characterised by robust international cooperation and a pronounced focus on the growing importance of this research in the years to come.

Identification of Research Themes

Compared to previous library studies that focused primarily on catalyst development, hydrogen production and thermodynamic optimisation, this study emphasises the integration of process safety and inherent safety considerations into DRM research trends. This allows to interpret the identified research clusters from both a technological point of view and from the point of view of industrial safety and operational reliability. Keywords represent a basic summary of publications, and high-frequency keyword analysis is a common approach to identifying hot topics and themes in research areas (Lan et al., 2022). This paper analysed all keywords, including author keywords. Before analysis, singular and plural forms of words are reduced to the singular. Hence, "catalyst" and "catalysts" are reduced to "catalyst", and "kinetics" is reduced to "kinetic." Keywords with related terms are combined into one phrase each, as shown, for example, the words "DRM," "dry reforming," and "methane reforming" are combined into "dry reforming of methane."

The threshold value of keyword occurrence frequency was set to 5, and 134 keywords were identified. The sensitivity assessment was carried out using keyword occurrence thresholds of 3, 5, 7, 8 and 10. Lower thresholds create a larger number of low-frequency keywords, creating fragments and fewer interpreted groups. In contrast, higher thresholds reduced thematic coverage by excluding many relevant keywords associated with safety-related DRM research. The threshold of five occurrences was therefore chosen because it ensured the most balanced representation of the research landscape while maintaining the clarity and interpretability of clusters. The resulting network identified four clusters. Figure 4 illustrates the four identified clusters, labelled as Cluster 1 (red), Cluster 2 (green), Cluster 3 (blue), and Cluster 4 (orange). These four clusters contain 40, 27, 51, and 16 keywords, respectively, whose central keywords and the relevant LS are sorted in Table 7.

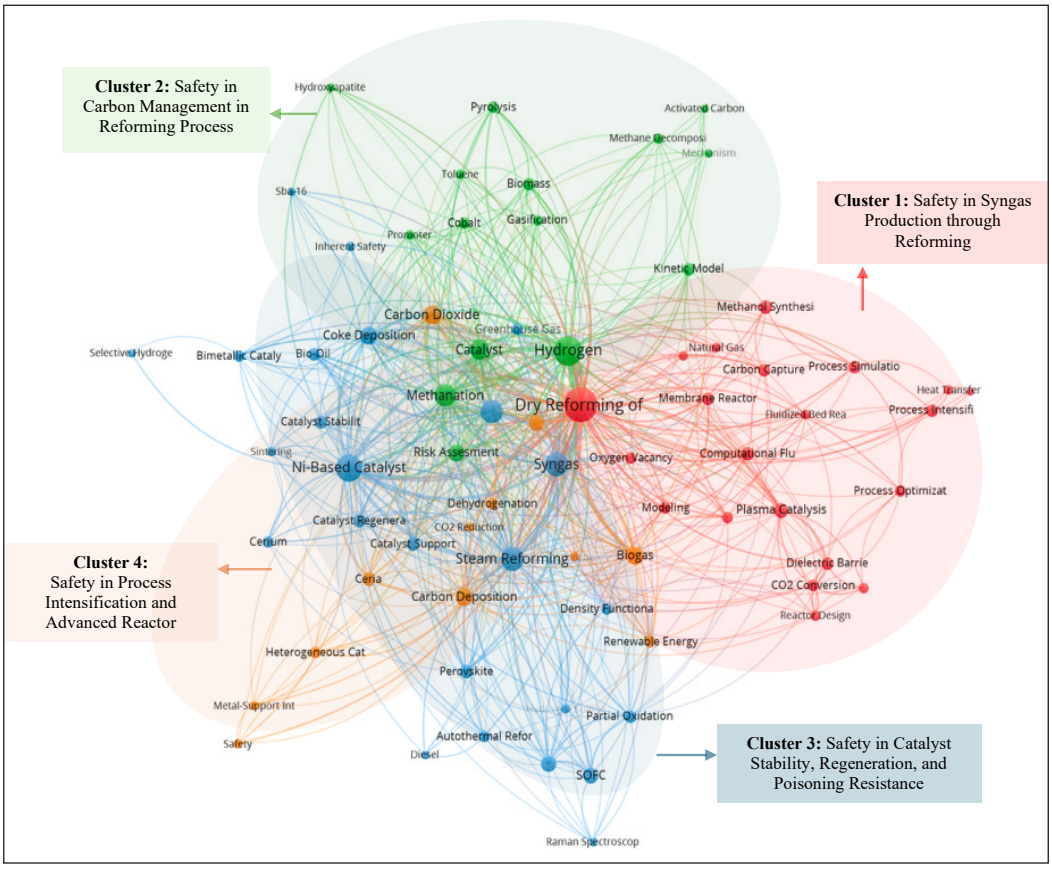


Figure 4. Distribution of co-occurrence network clusters for safety-related dry reforming publications keywords

Table 7
Top 10 keywords in each co-occurrence network cluster according to total link strength in safety-related dry reforming publications

Cluster 1: Safety in Syngas Production Through Reforming		Cluster 2: Safety in Carbon Management in Reforming Process		Cluster 3: Safety in Catalyst Stability, Regeneration, and Poisoning Resistance		Cluster 4: Safety in Process Intensification and Advanced Reactor Design	
Dry Reforming of Methane	467	Hydrogen	298	Ni-based Catalyst	269	Carbon Dioxide	111
Plasma Catalysis	39	Methanation	152	Syngas	192	Biogas	99
Dielectric Barrier Discharge	31	Catalyst	103	Steam Reforming	191	Carbon Deposition	64
Computational Fluid Dynamics	26	Risk Assessment	61	Catalyst Deactivation	158	Coke Resistance	48
Modelling	26	Cobalt	37	Coke Deposition	96	Ceria	39
Chemical Looping	24	Biomass	30	Sulphur Poisoning	50	Renewable Energy	20
Membrane Reactor	24	Promoter	25	Catalyst Regeneration	49	Dehydrogenation	18
CO2 Conversion	20	Pyrolysis	25	Bio-Oil	39	Heterogeneous Catalysis	16
Carbon Capture	18	Gasification	22	Catalyst Support	39	Safety	11
Fluidised Bed Reactor	18	Kinetic Model	22	Perovskite	38	Ethylene	10

The bibliometric analysis resulted in several distinct research clusters in dry reforming of methane (DRM) that can be systematically mapped to safety-relevant aspects, thereby clarifying their practical implications for inherently safer process design. These clusters extend beyond descriptive research trends and demonstrate direct relevance to industrial safety standards, reactor design, and catalyst management strategies. In particular, the catalyst-related clusters are closely associated with risks of catalyst deactivation, sintering, and carbon deposition. Reactor-oriented clusters emphasise temperature management, heat removal, and pressure control, while clusters related to process integration and operation align with established safety engineering frameworks such as ISO 45001, API 521, and conventional HAZID/HAZOP methodologies. Table 8 summarises the explicit mapping between bibliometric clusters, corresponding safety principles, and their operational implications, illustrating how bibliometric insights can be translated into actionable guidance for safer DRM systems.

Table 8
Mapping of bibliometric clusters to safety themes, dominant hazards, and relevant safety engineering frameworks

Cluster	Theme	Dominant Research Topics	Safety Issues Identified	Representative Safety Framework
Cluster 1	Safety in Syngas Production Through Reforming	Reactor performance, catalyst activity, syngas generation	Hotspots, thermal instability, and CO handling	HAZOP/ Process Safety Management
Cluster 2	Safety of Carbon Management in Reforming	Carbon deposition, catalyst deactivation	Reactor blockage, pressure rise, loss of flow distribution	HAZID
Cluster 3	Safety of Catalyst Stability and Regeneration	Coking, sulphur poisoning, catalyst regeneration	Process instability, regeneration hazards	HAZOP
Cluster 4	Safety in Process Intensification and Advanced Reactor Design	Advanced reactors, heat transfer enhancement	Overpressure, thermal stress, and equipment integrity	API 521

Note. Adapted from (Hendershot, 2012; Omran et al., 2020; Usman et al., 2015)

The four clusters are categorised into four themes according to the publications where these keywords are found and related to the safety of dry reforming. Cluster 1 theme is "Safety in Syngas Production through Reforming", and it addresses the optimisation of the reforming process, catalysis, and reactor design for safe and efficient syngas production (Cherif et al., 2022). It also addresses several safety aspects, including reactor stability and catalyst performance, as well as environmental concerns (Zhao et al., 2024; Li et al., 2019). Cluster 1 emphasises CO_2 recycling and the ecological aspects of methanol production. Safety and environmental analyses are incorporated into syngas production processes through modelling, simulation, and optimisation, which are essential for safe and efficient catalytic reactions in reforming (Gyanwali et al., 2023; Saqline et al., 2021).

Cluster 2 is on the Safety of Carbon Management in the Reforming Process. This leads to a focus on carbon deposition and catalyst deactivation. Various studies have identified the stability of the catalyst and prevention of carbon deposition as key safety issues in the catalytic reforming process (Kim et al., 2024). Many studies have focused on methane and hydrogen production through various catalytic techniques that optimise reaction protection and process efficiency (Durán et al., 2023; Hondo et al., 2023; Valle et al., 2013)). Multiple reviews are available that assess the environmental and economic aspects of catalytic processes, often including safety analyses of the manufacturing systems. (Liu et al., 2023). The research in this cluster also focuses on catalyst deactivation and carbon deposition, which are two of the most persistent challenges in DRM systems (Boscherini et al., 2023).

In addition to reducing catalytic performance, excessive carbon accumulation may lead to reactor blockages, increased pressure dropped and formation of local hotspots, which may affect equipment integrity and process stability (Dannesboe et al., 2020). As a result, strategies aimed at mitigated formation of coke and catalyst deactivation not only improve process efficiency but also contribute to safer and more reliable DRM operations.

Cluster 3 is on the Safety of Catalyst Stability, Regeneration, and Poisoning Resistance. The stability, deactivation mechanism, and performance optimisation of various catalysts in several reforming and hydrogenation processes have been investigated. This work presents and discusses the following challenges, such as coking problems, sulphur poisoning, and catalyst regeneration, which are essential challenges that must be considered while developing long-term stability and safety for catalytic reforming processes. Some articles dwell on deactivation mechanisms, such as sulphur poisoning and coke formation, and restoration methods or ways of preventing losses in catalysts' activity (Naidu et al., 2022; Li et al., 2021). Featured is the stability of the catalyst in reforming and hydrogenation processes for effectiveness and safety during operations, which is the subject of many papers. (Li et al., 2021; Shao et al., 2023). Various works investigate ways to decrease coke formation, a critical problem for maintaining catalyst activity in dry reforming processes. (Manavi & Liu, 2023). In addition to causing problems related to catalysts, sulphur poisoning and catalyst degradation could also lead to process instability and deviation from normal operating conditions (Yuan et al., 2023; Li et al., 2024). Effective catalyst regeneration and poisoning resistance, therefore, play an important role not only in maintaining catalyst activity and longevity but also in supporting the safe and stable operation of DRM systems (Pham Minh et al., 2021).

Cluster 4 focuses on Safety in Process Intensification and Advanced Reactor Design. Process intensification, advanced reactor design, and optimisation of catalysts are repeatedly mentioned throughout the studies to improve operational safety and minimise the risk. Research on core-shell nanostructured catalysts, perovskite-modified catalysts, and strong metal-support interactions suggests the development of advanced reactors and catalysts that can perform effectively and safely at high temperatures and pressures (Ahmad et al., 2022, Duma et al., 2024; Yang et al., 2022). A trickle bed reactor and porous crystalline catalysts ensure better heat and mass transfer and reduce potential hazards. A few quantitative risk assessment literature reviews focus on process safety and the strategy for amine-based CO_2 capture, related to mitigating risks in industrial-scale operations (Aliyu, 2024). Some findings from chemical kinetic modelling in related studies suggest its role in forecasting and mitigating probable hazardous reactions to enhance reactor design safety (Kanayama et al., 2024). Compared to other technologies, this theme is more advanced. These four themes were categorised into the circle of the co-occurrence matrix based on safety-related DRM research that could give clarity among researchers unfamiliar with the area and external investors about the research framework.

Some identified research clusters also demonstrate the relevance of designing DRM systems using an intrinsically safer design principle (ISD). Catalyst optimisation studies help to develop substitution and moderation strategies by allowing lower temperature operation and reducing carbon deposition trends. Process intensification and advanced reactor configurations support the principle of simplification and minimisation, reducing equipment inventory, improving heat transfer efficiency and reducing operational instability. In addition, hazards classification methods, DRM process risk mapping and inherent safety indicators can provide a systematic approach to assessing operational risks related to reactor temperature, pressure changes, toxic gas formation and catalytic degradation. These approaches are still relatively little explored in the DRM literature and are important directions for future research aimed at safety as shown in Table 8.

Identified bibliographic clusters can be interpreted in a broader context of process safety management and inherently safer design (ISD). In this study, research themes were mapped to safety frameworks based on the dominant keywords identified in Table 6 and the dominant research themes discussed in representative publications associated with each cluster. For example, keywords such as carbon deposition, catalyst deactivation, sulphur poisoning and risk assessment were used to determine the link between specific clusters and related safety considerations. The principles of inherently safer design refer to strategies to reduce risks at the source through minimisation, substitution, moderation and simplification. Support frameworks such as HAZID and HAZOP facilitate the systematic identification and assessment of hazards during the operation and expansion of DRM, while API 521 provides guidelines for thermal protection systems (Jafarian et al., 2023). Therefore, these frameworks are discussed as a relevant safety perspective for interpreting DRM research trends, rather than the methodology explicitly applied to each analysed publication.

The application of inherently safer design (ISD) principles in DRM systems can be achieved through several practical approaches. Minimisation may be achieved by reducing combustible gas inventory in the process, while substitution can involve using coke-resistant catalysts to reduce carbon deposition and catalyst degradation (Athar et al., 2022). Moderation may be achieved through improved temperature-control strategies to limit hotspot formation and thermal stress, whereas simplification can be promoted through reactor designs with improved heat distribution and reduced operational complexity (Hendershot, 2012). Collectively, these measures contribute to safer and more reliable DRM operation while supporting process efficiency and equipment integrity.

Evolution of Keywords

Various elements can reflect the development of a field, and keywords are key elements that indicate the evolution of research hotspots and scale within a field over time (Tan et al., 2021). This evolution shows that “catalyst” has the highest frequency in all the years in this

safety-related DRM research from 2013 to 2024, with the highest frequencies in 2023 and 2024, at 75 and 113, respectively. The keywords “hydrogen” and “methanation” rank second over the past 10 years, with the highest number, 74, recorded in 2024. “Ni-Based Catalyst” occurs most frequently in 2023 and 2024, with 30 and 47 occurrences, respectively. This is because most of the publications in 2023 (Dong et al., 2023; Wajahat ul Hasnain et al., 2023) and 2024 (Li et al., 2024; Lin et al., 2024) focus on the safety of catalyst stability, regeneration, and poisoning resistance in the dry reforming process. Safety is an immediate concern for the development of dry reforming processes. The terms "catalyst deactivation" and "coke deposition" track over time, along with the "risk assessment", reflect much of the present work being directed at resolving issues in catalyst stability and longevity concerns, as well as possible process safety. This tendency is characterised by a growing awareness of the risk assessment in dry reforming, particularly its operational reliability, environmental impact, and compensation options. Nevertheless, the integration of systems thinking into assessment approaches remains insignificant, illustrating the need for increased scientific study that examines issues of safety in conjunction with catalytic functionality. Furthermore, despite a greater focus on catalyst stability, there is a critical dearth of studies that comprehensively examine the risk factors affecting the overall dry reforming process, such as feedstock variability and reactor performance during transient conditions. Extensive system-wide studies of issues such as reactor heat removal, pressure control, and emergency shutdown systems should be included in process safety investigations alongside the catalysts' lifetime. Although DRM is an endothermic reaction, it operates at very high temperatures, and uneven heat supply can lead to local hot spots or cold spots, which can accelerate undesirable side reactions, carbon deposits, or catalyst sintering. These thermal gradients also impose mechanical constraints on the reactor, increasing the risk of secondary escape reactions and making heat removal and distribution a critical safety consideration. No work has been done to investigate the use of safety design to minimise unwanted side reactions and runaway reactions. Real-time monitoring and predictive analysis are likely to minimise major safety hazards, but most research employs traditional, conventional risk assessment methods.

Despite being commonly regarded as an eco-friendly process for CO_2 utilisation, dry reforming yields undesirable byproducts, such as carbon monoxide and higher hydrocarbons. Those generate extra concern about environmental issues. If improperly managed, these byproducts may lead to equipment fouling, increased maintenance needs, and potentially harmful emissions. Future studies can investigate new process control technologies, including AI-driven safety monitoring systems, machine learning-based failure prediction, and adaptive reactor systems that can dynamically adapt to various operating scenarios (Osarogiagbon et al., 2021). Among the essential concerns is the limited number of research works on large-scale industrial applications of dry reforming under realistic operation conditions.

Many works emphasise the importance of small-scale experiments, which can be inadequate in addressing the safety issues of industrial applications. Scaling up dry reforming from small to industrial scale requires intensive hazard evaluation, lifecycle safety estimation, and pilot-scale demonstration for verifying the efficacy of safety measures under actual operating conditions (Bachmann et al., 2023). Resolving these deficiencies will make dry reforming a long-term, sustainable, and industrially feasible process. Future research should prioritise catalyst performance and place significant emphasis on comprehensive safety evaluations, encompassing fundamental safety concepts, advanced risk assessment techniques, and real-time monitoring technologies, to maximise both process safety and efficiency. By applying a multidisciplinary framework that incorporates chemical engineering, safety science, and data technologies, researchers and industry stakeholders can advocate for a more secure and robust application of dry reforming in industrial hydrogen and syngas production.

CONCLUSION

This study provides a comprehensive bibliometric overview of safety-related themes within the dry reforming of methane (DRM) research landscape. In general, it demonstrates that active challenges to industrialisation concern the catalysts' deactivation, carbon deposition, and process stability. The primary safety concerns remain related to operations at high temperatures, the integrity of the reactors, and the hazards associated with carbon formation. Researchers focus on catalyst design, reactor optimisation, and process control strategies as the most important domains for improving performance and ensuring safety. This knowledge gap needs to be addressed to pave the way for the safer industrial application of dry reforming. Future studies should develop risk assessment models for dry reforming processes that systematically outline all possible hazards. Principles of safety in catalyst and reactor design should focus on minimising the risks associated with high-temperature operation and material degradation. Therefore, future work should investigate critical operational risks such as catalyst deactivation, coke formation, and thermal runaway to develop a more robust process with high stability for long-term performance. Safety-driven approaches can be combined to enhance the reliability and sustainability of dry-reforming technologies. Such efforts should include measurable process indicators relevant for DRM operation, including reactor temperature gradients for hotspot detection, pressure drop as an indicator of carbon deposits and blockages, carbon monoxide concentrations for toxic gas monitoring, loss of catalyst activity and coke growth trends as indicators of catalyst degradation and long-term operating stability. The integration of these variables with predictive analysis can improve early warning of hazards and support safer industrial DRM operations. Although the current study is based mainly on Scopus-indexed peer-reviewed academic publications, future safety assessments of DRM systems may benefit from the

inclusion of complementary sources such as pilot studies, industrial reports, patents, and process safety incident databases. These sources can provide additional insights into operational challenges, equipment reliability, scale-up considerations and safety incidents in the real world that are not always recorded in academic literature. Integration of scientific and industrial evidence can therefore strengthen future safety-related DRM assessments and support safer commercialisation pathways.

Within the coming five years, research related to the safety of DRMs needs to go beyond just being conducted using bibliometrics to the extent of being verified at a pilot scale and ultimately applied industrially. Some aspects to which research attention needs to be devoted include establishing DRM hazard databases that are standardised, validating safety indicators for pilot-scale operation, installing predictive safety monitoring systems and incorporating inherently safer design concepts in reactor scaling up. Stronger collaboration between academia and industry will also be essential to translate research findings into reliable and commercially viable DRM technologies.

ACKNOWLEDGEMENT

This research was supported by the MTJA Research Cess Fund under the grant "Direct Utilisation of CO_2-NG for Simultaneous Production of Syngas and Energy" (015ME0-377). The authors sincerely appreciate this financial support.

REFERENCES

Abdullah, B., Abd Ghani, N. A., & Vo, D.-V. N. (2017). Recent advances in dry reforming of methane over Ni-based catalysts. *Journal of Cleaner Production, 162*, 170-185. https://doi.org/10.1016/j.jclepro.2017.05.176

Ahmad, N., Wahab, R., Manoharadas, S., Alrayes, B. F., Alam, M., & Alharthi, F. A. (2022). The role of strontium in $CeNiO_3$ nano-crystalline perovskites for greenhouse gas mitigation to produce syngas. *Molecules, 27*(2), Article 356. https://doi.org/10.3390/molecules27020356

Ahmad, Y. H., Mohamed, A. T., El-Sayed, H. A., Kumar, A., & Al-Qaradawi, S. Y. (2022). Design of Ni/La_2O_3 catalysts for dry reforming of methane: Understanding the impact of synthesis methods. *International Journal of Hydrogen Energy, 47*(97), 41294-41309. https://doi.org/10.1016/j.ijhydene.2021.10.190

Alhassan, M., Jalil, A. A., Nabgan, W., Hamid, M. Y. S., Bahari, M. B., & Ikram, M. (2022). Bibliometric studies and impediments to valorisation of dry reforming of methane for hydrogen production. *Fuel, 328*, Article 125240. https://doi.org/10.1016/j.fuel.2022.125240

Aliyu, A. A. (2024). *IEAGHG workshop on comparative techno-economic assessment of commercially available CO_2 conditioning technologies* (Technical Review 2024-TR02). IEAGHG. https://doi.org/10.62849/2024-TR02

Amelse, J. A., & Behrens, P. K. (2021). *Sequestering biomass for natural, efficient, and low-cost direct air capture of carbon dioxide* (Version 4) [Preprint]. Preprints.org. https://doi.org/10.20944/preprints202106.0212.v4

Ampomah, W., Morgan, A., Koranteng, D. O., & Nyamekye, W. I. (2024). CCUS perspectives: Assessing historical contexts, current realities, and prospects. *Energies, 17*(17), Article 4248. https://doi.org/10.3390/en17174248

Athar, M., Shariff, A. M., Buang, A., Umer, A., & Zaini, D. (2022). Inherently safer process route ranking index (ISPRRI) for sustainable process design. *Journal of Loss Prevention in the Process Industries, 80*, Article 104909. https://doi.org/10.1016/j.jlp.2022.104909

Ávila-Neto, C. N., Dantas, S. C., Silva, F. A., Franco, T. V., Romanielo, L. L., Hori, C. E., & Assis, A. J. (2009). Hydrogen production from methane reforming: Thermodynamic assessment and autothermal reactor design. *Journal of Natural Gas Science and Engineering, 1*(6), 205-215. https://doi.org/10.1016/j.jngse.2009.12.003

Bachmann, M., Völker, S., Kleinekorte, J., & Bardow, A. (2023). Syngas from what? Comparative life-cycle assessment for syngas production from biomass, CO_2, and steel mill off-gases. *ACS Sustainable Chemistry & Engineering, 11*(14), 5356-5366. https://doi.org/10.1021/acssuschemeng.2c05390

Beheshti Askari, A., Al Samarai, M., Morana, B., Tillmann, L., Pfänder, N., Wandzilak, A., Watts, B., Belkhou, R., Muhler, M., & DeBeer, S. (2020). In situ X-ray microscopy reveals particle dynamics in a NiCo dry methane reforming catalyst under operating conditions. *ACS Catalysis, 10*(11), 6223-6230. https://doi.org/10.1021/acscatal.9b05517

Boscherini, M., Storione, A., Minelli, M., Miccio, F., & Doghieri, F. (2023). New perspectives on catalytic hydrogen production by the reforming, partial oxidation and decomposition of methane and biogas. *Energies, 16*(17), Article 6375. https://doi.org/10.3390/en16176375

Central Committee of the Communist Party of China. (2016). *The 13th Five-Year Plan for economic and social development of the People's Republic of China (2016-2020).* National Development and Reform Commission. https://en.ndrc.gov.cn/policies/202105/P020210527785800103339.pdf

Chen, Q., Wang, J., Gao, M., Liu, L., & Tao, J. (2022). Review on loss prevention of chemical reaction thermal runaway: Principles and application. *Emergency Management Science and Technology, 2*(1), 1-8. https://doi.org/10.48130/EMST-2022-0010

Cherif, A., Nebbali, R., Sen, F., Sheffield, J. W., Doner, N., & Nasseri, L. (2022). Modelling and simulation of steam methane reforming and methane combustion over continuous and segmented catalyst beds in an autothermal reactor. *International Journal of Hydrogen Energy, 47*(20), 9127-9138. https://doi.org/10.1016/j.ijhydene.2021.12.250

Collivignarelli, M. C., Collivignarelli, C., Carnevale Miino, M., Abbà, A., Pedrazzani, R., & Bertanza, G. (2020). SARS-CoV-2 in sewer systems and connected facilities. *Process Safety and Environmental Protection, 143*, 196-203. https://doi.org/10.1016/j.psep.2020.06.049

Dang, C., Luo, J., Yang, W., Li, H., & Cai, W. (2021). Low-temperature catalytic dry reforming of methane over Pd-promoted Ni–CaO–$Ca_{12}Al_{14}O_{33}$ multifunctional catalyst. *Industrial & Engineering Chemistry Research, 60*(51), 18361-18372. https://doi.org/10.1021/acs.iecr.1c04010

Dannesboe, C., Hansen, J. B., & Johannsen, I. (2020). Catalytic methanation of CO_2 in biogas: Experimental results from a reactor at full scale. *Reaction Chemistry & Engineering, 5*(1), 183-189. https://doi.org/10.1039/c9re00351g

Das, S., Sengupta, M., Bag, A., Saini, A., Muhler, M., & Bordoloi, A. (2021). Gd–Ru nanoparticles supported on $Zr_{0.5}Ce_{0.5}O_2$ nanorods for dry methane reforming. *ACS Applied Nano Materials, 4*(3), 2547-2557. https://doi.org/10.1021/acsanm.0c03140

Demeter, M., Goyanes, M., Háló, G., & Xu, X. (2025). The internationalisation of Chinese social sciences research: Publication, collaboration, and citation patterns in economics, education, and political science. *Policy Reviews in Higher Education, 9*(1), 81-107. https://doi.org/10.1080/23322969.2024.2438240

Dong, S., Pu, Y., Niu, Y., Zhang, L., Wang, Y., & Zhang, B. (2023). Interstitial carbon in Ni enables high-efficiency hydrogenation of 1,3-butadiene. *Acta Physico-Chimica Sinica, 39*(11), Article 2301012. https://doi.org/10.3866/PKU.WHXB202301012

Duarte, R. B., Pimenta, J. L. C. W., & de Matos Jorge, L. M. (2024). An overview on biogas reforming for synthesis of sustainable aviation fuel. *International Journal of Hydrogen Energy, 85*, 210-227. https://doi.org/10.1016/j.ijhydene.2024.08.330

Duma, Z. G., Swartbooi, A., & Musyoka, N. M. (2024). Thermocatalytic decomposition of methane to low-carbon hydrogen using $LaNi_{1-x}Cu_xO_3$ perovskite catalysts. *Applied Catalysis A: General, 677*, Article 119703. https://doi.org/10.1016/j.apcata.2024.119703

Durán, I., Dietrich, B., Hofberger, C., Stoppel, L., Uhlenbruck, N., & Wetzel, T. (2023). CO_2 impact on methane pyrolysis as a key issue of using biogas as an educt: A theoretical study. *International Journal of Energy Research, 2023*, Article 3684046. https://doi.org/10.1155/2023/3684046

Ferrandon, M. S., Byron, C., Celik, G., Zhang, Y., Ni, C., Sloppy, J., McCormick, R. A., Booksh, K., Teplyakov, A. V., & Delferro, M. (2022). Grafted nickel-promoter catalysts for dry reforming of methane identified through high-throughput experimentation. *Applied Catalysis A: General, 629*, Article 118379. https://doi.org/10.1016/j.apcata.2021.118379

Gazi, M. J., Khurana, D., Kaishyop, J., Khan, T. S., Bhandari, S., & Bordoloi, A. (2024). Solution combustion-derived nanoalloys: Robust and efficient catalyst systems for partial oxidation of methane. *International Journal of Hydrogen Energy, 51*, 562-579. https://doi.org/10.1016/j.ijhydene.2023.08.261

Gobbo, J. A., Busso, C. M., Gobbo, S. C. O., & Carreão, H. (2018). Making the links among environmental protection, process safety, and Industry 4.0. *Process Safety and Environmental Protection, 117*, 372-382. https://doi.org/10.1016/j.psep.2018.05.017

Gyanwali, K., Karki, S., Adhikari, P., Devkota, S., & Aryal, P. (2023). Techno-economic assessment of green urea production utilising municipal solid waste and hydropower in Nepal. *Journal of Cleaner Production, 419*, Article 138320. https://doi.org/10.1016/j.jclepro.2023.138320

He, J., Laudenschleger, D., Schittkowski, J., Machoke, A., Song, H., Muhler, M., Schlögl, R., & Ruland, H. (2020). Influence of contaminants in steel mill exhaust gases on $Cu/ZnO/Al_2O_3$ catalysts applied in methanol synthesis. *Chemie Ingenieur Technik, 92*(10), 1525-1532. https://doi.org/10.1002/cite.202000045

Hendershot, D. C. (2012). Inherently safer design: The fundamentals. *Chemical Engineering Progress, 108*(1), 40-42. https://doi.org/10.1016/j.jchas.2012.09.006

Hondo, E., Gani, T. Z. H., Kosari, M., Xi, S., Bella, N., Ashok, J., Tan, J. Y., Wang, T., Bian, H., Lim, K. H., Chen, L., Chang, J., Borgna, A., & Kawi, S. (2023). Unveiling the roles of precursor structure and controlled sintering on Ni-phyllosilicate-derived catalysts for low-temperature methane decomposition. *ACS Sustainable Chemistry & Engineering, 11*(24), 8786-8799. https://doi.org/10.1021/acssuschemeng.3c00120

Hussien, A. G. S., & Polychronopoulou, K. (2022). A review on the different aspects and challenges of the dry reforming of methane (DRM) reaction. *Nanomaterials, 12*(19), Article 3400. https://doi.org/10.3390/nano12193400

Jafarian, M., Haseli, P., Saxena, S., & Dally, B. (2023). Emerging technologies for catalytic gasification of petroleum residue-derived fuels for sustainable and cleaner fuel production: An overview. *Energy Reports, 9*, 2198-2226. https://doi.org/10.1016/j.egyr.2023.01.116

Jiang, L., Li, D., Deng, G., Lu, C., Huang, L., Li, Z., Xu, H., Zhu, X., Wang, H., & Li, K. (2023). Design of hybrid $La_{1-x}Ce_xCoO_{3-\delta}$ catalysts for lean methane combustion via creating active Co and Ce species. *Chemical Engineering Journal, 456*, Article 141054. https://doi.org/10.1016/j.cej.2022.141054

Kanayama, K., Grégoire, C. M., Cooper, S. P., Almarzooq, Y., Petersen, E. L., Mathieu, O., Maruta, K., & Nakamura, H. (2024). Experimental and chemical kinetic modelling study of ethylene carbonate oxidation: A lithium-ion battery electrolyte surrogate model. *Combustion and Flame, 262*, Article 113333. https://doi.org/10.1016/j.combustflame.2024.113333

Kim, Y. J., Kim, M. J., Kim, D. H., Mnoyan, A., & Lee, K. (2024). Enhanced methane dry reforming with Ni/SiO_2 catalysts featuring hierarchical external nanostructures. *Catalysts, 14*(4), Article 265. https://doi.org/10.3390/catal14040265

Lam, W. S., Lam, W. H., & Lee, P. F. (2023). The studies on chitosan for sustainable development: A bibliometric analysis. *Materials, 16*(7), Article 2857. https://doi.org/10.3390/ma16072857

Lan, J., Wei, R., Huang, S., Li, D., Zhao, C., Yin, L., & Wang, J. (2022). In-depth bibliometric analysis of research trends in fault diagnosis of lithium-ion batteries. *Journal of Energy Storage, 54*, Article 105275. https://doi.org/10.1016/j.est.2022.105275

Li, L., Chen, J., Zhang, Q., Yang, Z., Sun, Y., & Zou, G. (2020). Methane dry reforming over activated carbon-supported Ni catalysts prepared by solid-phase synthesis. *Journal of Cleaner Production, 274*, Article 122256. https://doi.org/10.1016/j.jclepro.2020.122256

Li, L., Zhang, H., Li, X., Kong, X., Xu, R., Tay, K., & Tu, X. (2019). Plasma-assisted CO_2 conversion in a gliding arc discharge: Improving performance by optimising the reactor design. *Journal of CO_2 Utilisation, 29*, 296-303. https://doi.org/10.1016/j.jcou.2018.12.019

Li, R., Xu, W., Deng, J., & Zhou, J. (2021). Coke-resistant Ni–Co/ZrO_2–CaO-based microwave catalyst for highly effective dry reforming of methane by microwave catalysis. *Industrial & Engineering Chemistry Research, 60*, 17458-17468. https://doi.org/10.1021/acs.iecr.1c03164

Li, Y., Li, Y., Na, L., Xiao, T., Cui, Y., Li, P., & Lü, Z. (2024). Sulfur poisoning and O^{2-}-pumping regeneration of $La_{0.75}Sr_{0.25}Cr_{0.5}Mn_{0.5}O_{3-\delta}$–Ni–$CeO_2$ anode. *Journal of Power Sources, 614*, Article 235061. https://doi.org/10.1016/j.jpowsour.2024.235061

Lin, J., Wu, S., Tang, C., Chen, X., & Zheng, Y. (2024). Roles of different Ni–Si interactions in methane combustion under oscillating temperature conditions. *Journal of Colloid and Interface Science, 668*, 512–524. https://doi.org/10.1016/j.jcis.2024.04.184

Liu, G., Sun, S., Sun, H., Zhang, Y., Lv, J., Wang, Y., Zeng, J., Yan, Z., & Wu, C. (2023). Integrated CO_2 capture and utilisation: A promising step contributing to carbon neutrality. *Carbon Capture Science & Technology, 7*, Article 100116. https://doi.org/10.1016/j.ccst.2023.100116

Lunsford, J. H. (2000). Catalytic conversion of methane to more useful chemicals and fuels: A challenge for the 21st century. *Catalysis Today, 63*(2-4), 165-174. https://doi.org/10.1016/S0920-5861(00)00456-9

Manavi, N., & Liu, B. (2023). Mitigating coke formation for dry reforming of methane on dual-site catalysts: A microkinetic modelling study. *The Journal of Physical Chemistry C, 127*(5), 2274-2284. https://doi.org/10.1021/acs.jpcc.2c06788

Martinez-Hernandez, A. (2024). Dry reforming of methane with a Ni-based catalyst: A kinetic and thermodynamic analysis. *Reaction Kinetics, Mechanisms and Catalysis, 137*(5), 2617-2639. https://doi.org/10.1007/s11144-024-02658-2

Merigó, J. M., Miranda, J., Modak, N. M., Boustras, G., & de la Sotta, C. (2019). Forty years of *Safety Science*: A bibliometric overview. *Safety Science, 115*, 66-88. https://doi.org/10.1016/j.ssci.2019.01.029

Mette, K., Kühl, S., Tarasov, A., Willinger, M. G., Kröhnert, J., Wrabetz, S., Trunschke, A., Scherzer, M., Girgsdies, F., Düdder, H., Kähler, K., Ortega, K. F., Muhler, M., Schlögl, R., Behrens, M., & Lunkenbein, T. (2016). High-temperature-stable Ni nanoparticles for the dry reforming of methane. *ACS Catalysis, 6*(10), 7238-7248. https://doi.org/10.1021/acscatal.6b01683

Mhadeshwar, A. B., & Vlachos, D. G. (2007). A catalytic reaction mechanism for methane partial oxidation at short contact times, reforming, and combustion, and for oxygenate decomposition and oxidation on platinum. *Industrial & Engineering Chemistry Research, 46*(16), 5310-5324. https://doi.org/10.1021/ie070322c

Miao, C., Cai, L., Wang, Y., Xu, X., Yang, J., He, Y., Li, D., & Feng, J. (2021). Array-modified moulded alumina-supported PdAg catalyst for selective acetylene hydrogenation: Intrinsic kinetics enhancement and thermal effect optimisation. *Industrial & Engineering Chemistry Research, 60*, 8362–8374. https://doi.org/10.1021/acs.iecr.1c00423

Mleczko, L., Malcus, S., & Wurzel, T. (1997). Catalytic reformer–combustor: A novel reactor concept for synthesis gas production. *Industrial & Engineering Chemistry Research, 36*, 4459-4465. https://doi.org/10.1021/ie970245t

Naidu, B. N., Kumar, K. D. P. L., Saini, H., Kumar, M., Kumar, T. N., & Prasad, V. V. D. N. (2022). Coke deposition over Ni-based catalysts for dry reforming of methane: Effects of MgO-Al_2O_3 support and ceria and lanthana promoters. *Journal of Environmental Chemical Engineering, 10*(1), Article 106980. https://doi.org/10.1016/j.jece.2021.106980

Noor, Z. Z., Yusuf, R. O., Abba, A. H., Abu Hassan, M. A., & Mohd Din, M. F. (2013). An overview of energy recovery from municipal solid wastes (MSW) in the Malaysian scenario. *Renewable and Sustainable Energy Reviews, 20*, 378-384. https://doi.org/10.1016/j.rser.2012.11.050

Omran, A., Yoon, S. H., Khan, M., Ghouri, M., Chatla, A., & Elbashir, N. (2020). Mechanistic insights for dry reforming of methane on Cu/Ni bimetallic catalysts: DFT-assisted microkinetic analysis for coke resistance. *Catalysts, 10*(9), Article 1043. https://doi.org/10.3390/catal10091043

Osarogiagbon, A. U., Khan, F., Venkatesan, R., & Gillard, P. (2021). Review and analysis of supervised machine learning algorithms for hazardous events in drilling operations. *Process Safety and Environmental Protection, 147*, 367-384. https://doi.org/10.1016/j.psep.2020.09.038

Oyama, S. T., Hacarlioglu, P., Gu, Y., & Lee, D. (2012). Dry reforming of methane has no future for hydrogen production: Comparison with steam reforming at high pressure in standard and membrane reactors. *International Journal of Hydrogen Energy, 37*(13), 10444-10450. https://doi.org/10.1016/j.ijhydene.2011.09.149

Pawar, V., Appari, S., Monder, D. S., & Janardhanan, V. M. (2017). Study of the combined deactivation due to sulfur poisoning and carbon deposition during biogas dry reforming on supported Ni catalyst. *Industrial & Engineering Chemistry Research, 56*(30), 8448-8455. https://doi.org/10.1021/acs.iecr.7b01662

Pham Minh, D., Pham, X. H., Siang, T. J., & Vo, D.-V. N. (2021). Review on the catalytic tri-reforming of methane-Part I: Impact of operating conditions, catalyst deactivation and regeneration. *Applied Catalysis A: General, 621*, Article 118202. https://doi.org/10.1016/j.apcata.2021.118202

Pranckutė, R. (2021). Web of Science (WoS) and Scopus: The titans of bibliographic information in today's academic world. *Publications, 9*(1), Article 12. https://doi.org/10.3390/publications9010012

Sandoval-Diaz, L. E., Schlögl, R., & Lunkenbein, T. (2022). Quo vadis dry reforming of methane? - A review of its chemical, environmental, and industrial prospects. *Catalysts, 12*(5), Article 465. https://doi.org/10.3390/catal12050465

Santamaria, L., Arregi, A., Lopez, G., Artetxe, M., Amutio, M., Bilbao, J., & Olazar, M. (2020a). Effect of La_2O_3 promotion on a Ni/Al_2O_3 catalyst for H_2 production in the in-line biomass pyrolysis reforming. *Fuel, 262*, Article 116593. https://doi.org/10.1016/j.fuel.2019.116593

Santamaria, L., Artetxe, M., Lopez, G., Cortazar, M., Amutio, M., Bilbao, J., & Olazar, M. (2020b). Effect of CeO_2 and MgO promoters on the performance of a Ni/Al_2O_3 catalyst in the steam reforming of biomass pyrolysis volatiles. *Fuel Processing Technology, 198*, Article 106223. https://doi.org/10.1016/j.fuproc.2019.106223

Saqline, S., Chua, Z. Y., & Liu, W. (2021). Coupling chemical looping combustion of solid fuels with advanced steam cycles for CO_2 capture: A process modelling study. *Energy Conversion and Management, 244*, Article 114455. https://doi.org/10.1016/j.enconman.2021.114455

Shaifudin, M. S., Kamaruzzaman, W. M. I. W. M., Badruddin, M. A., Suhaimi, A. M. A. A. M., Nasir, N. A. M., Hamidi, N. A. S. M., Abdullah, W. R. W., Lee, O. J., & Mohd Ghazali, M. S. (2022). Exploring the global publications on varistors using the Scopus database through a bibliometric analysis. *Journal of Asian Ceramic Societies, 10*(2), 438-452. https://doi.org/10.1080/21870764.2022.2068748

Shao, S., Zhang, P., Li, X., & Yu, Y. (2023). Steam reforming of the simulated aqueous fraction of bio-oil based on pre-reforming with dolomite. *Fuel, 344*, Article 128116. https://doi.org/10.1016/j.fuel.2023.128116

Shi, C., Labbaf, B., Mostafavi, E., & Mahinpey, N. (2020). Methanol production from water electrolysis and tri-reforming: Process design and techno-economic analysis. *Journal of CO_2 Utilisation, 38*, 241-251. https://doi.org/10.1016/j.jcou.2019.12.022

Soleimani, S., & Lehner, M. (2022). Tri-reforming of methane: Thermodynamics, operating conditions, reactor technology and efficiency evaluation-A review. *Energies, 15*(19), Article 7159. https://doi.org/10.3390/en15197159

Tan, H., Li, J., He, M., Li, J., Zhi, D., Qin, F., & Zhang, C. (2021). Global evolution of research on green energy and environmental technologies: A bibliometric study. *Journal of Environmental Management, 297*, Article 113382. https://doi.org/10.1016/j.jenvman.2021.113382

Tanksale, A., Beltramini, J. N., & Lu, G. Q. M. (2010). A review of catalytic hydrogen production processes from biomass. *Renewable and Sustainable Energy Reviews, 14*(1), 166–182. https://doi.org/10.1016/j.rser.2009.08.010

Tucker, C. L., Bordoloi, A., & van Steen, E. (2021). Novel single-pass biogas-to-diesel process using a Fischer-Tropsch catalyst designed for high conversion. *Sustainable Energy & Fuels, 5*(22), 5717–5732. https://doi.org/10.1039/d1se01299a

Usman, M., Wan Daud, W. M. A., & Abbas, H. F. (2015). Dry reforming of methane: Influence of process parameters-A review. *Renewable and Sustainable Energy Reviews, 45*, 710–744. https://doi.org/10.1016/j.rser.2015.02.026

Valecillos, J., Landa, L., Elordi, G., Remiro, A., Bilbao, J., & Gayubo, A. G. (2024). Are Rh catalysts a suitable choice for bio-oil reforming? The case of a commercial Rh catalyst in the combined H_2O and CO_2 reforming of bio-oil. *Catalysts, 14*(9), Article 571. https://doi.org/10.3390/catal14090571

Valle, B., Remiro, A., Aguayo, A. T., Bilbao, J., & Gayubo, A. G. (2013). Catalysts of Ni/α-Al_2O_3 and Ni/La_2O_3–α-Al_2O_3 for hydrogen production by steam reforming of the bio-oil aqueous fraction with pyrolytic lignin retention. *International Journal of Hydrogen Energy, 38*(3), 1307-01318. https://doi.org/10.1016/j.ijhydene.2012.11.014

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3

Vicente, J., Ereña, J., Montero, C., Azkoiti, M. J., Bilbao, J., & Gayubo, A. G. (2014). Reaction pathway for ethanol steam reforming on a Ni/SiO_2 catalyst, including coke formation. *International Journal of Hydrogen Energy, 39*(33), 18820-18834. https://doi.org/10.1016/j.ijhydene.2014.09.073

Wajahat ul Hasnain, S. M., Salam Farooqi, A., Singh, O., Hidayah Ayuni, N., Victor Ayodele, B., & Abdullah, B. (2023). Response surface optimisation of hydrogen-rich syngas production by the catalytic valorisation of greenhouse gases (CH_4 and CO_2) over Sr-promoted Ni/SBA-15 catalyst. *Energy Conversion and Management: X, 20*, Article 100451. https://doi.org/10.1016/j.ecmx.2023.100451

Yang, J., Wang, J., Zhao, J., Bai, Y., Du, H., Wang, Q., Jiang, B., & Li, H. (2022). CO_2 conversion via dry reforming of methane on a core-shell Ru@SiO_2 catalyst. *Journal of CO_2 Utilisation, 57*, Article 101893. https://doi.org/10.1016/j.jcou.2022.101893

Yuan, B., Zhu, T., Han, Y., Zhang, X., Wang, M., & Li, C. (2023). Deactivation mechanism and anti-deactivation measures of metal catalysts in the dry reforming of methane: A review. *Atmosphere, 14*(5), Article 770. https://doi.org/10.3390/atmos14050770

Zeng, J., Xu, R., Sun, R., Niu, L., Liu, Y., Zhou, Y., Zeng, W., & Yue, Z. (2020). Evaluation of methane emission flux from a typical biogas fermentation ecosystem in China. *Journal of Cleaner Production, 257*, Article 120441. https://doi.org/10.1016/j.jclepro.2020.120441

Zhao, X. G., Zhao, Y. X., Liu, Q. Y., & He, S. G. (2024). Dry reforming of methane to syngas mediated by rhodium-cobalt oxide cluster anions Rh_2CoO^-. *The Journal of Physical Chemistry Letters, 15*, 9167-9174. https://doi.org/10.1021/acs.jpclett.4c01961

Zhou, P., & Leydesdorff, L. (2006). The emergence of China as a leading nation in science. *Research Policy, 35*(1), 83-104. https://doi.org/10.1016/j.respol.2005.08.006

Pertanika J. Sci. & Technol. 34 (3): 1861 - 1880 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

An IoT-enabled Decision Support System for Real-time Line Balancing in Semiconductor Manufacturing

Nur Ain Qistina Muhammad Shafee[1], Effendi Mohamad[1]*, Mohd Soufhwee Abd Rahman[1], Arfauz A. Rahman[2], and Teruaki Ito[3]

[1]Department of Manufacturing Engineering, Faculty of Industrial and Manufacturing Technology and Engineering, Universiti Teknikal Malaysia Melaka (UTeM), Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
[2]Department of Mechanical and Aerospace Engineering, School of Mechanical and Aerospace Engineering, Queen's University Belfast, Ashby Building, Stranmillis Road, BT9 5AH Belfast, Northern Ireland, United Kingdom
[3]Department of Business Administration, Faculty of Business Administration, Matsuyama University, 4-2 Bunkyo-cho, 790-8578 Matsuyama-shi, Ehime, Japan

ABSTRACT

This study presents iDSS-ProLean, an IoT-enabled Decision Support System (DSS) developed to enhance line balancing in semiconductor production. Traditional methods lack real-time adaptability, while iDSS-ProLean integrates sensor-based data acquisition, Firebase cloud analytics, and a mobile app to support responsive, data-driven decisions. The system architecture includes ESP32 modules, multiple sensors, cloud processing, and an Android interface for the operation feedback. A feasibility study was conducted using 60 production runs, where Line Balancing Efficiency (LBE) served as the main performance metric. Paired t-tests revealed p-values above 0.05 and t-values near zero, indicating consistent data transmission and system stability. These results affirm the reliability of the mobile DSS and its real-time performance under varying production conditions. The study demonstrates how integrating Lean Manufacturing (LM) tools with IoT technologies enables dynamic line optimisation. It's proven that LBE is highly correlated with NoM (0.96), TPT (0.97), and CT (0.93), confirming that machine count, processing time, and cycle time strongly influence LBE. Overall, the IDSS-ProLean framework proves to be an effective and adaptive decision support tool for LM applications, showcasing the value of IoT-enabled models in replacing static, time-consuming methods with real-time, intelligent systems.

Keywords: Decision support system, Internet of Things, line balancing, manufacturing system, semiconductor sector

ARTICLE INFO

Article history:
Received: 28 November 2025
Accepted: 03 June 2026
Published: 25 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.24

E-mail addresses:
qistinashafee@gmail.com (Nur Ain Qistina Muhammad Shafee)
effendi@utem.edu.my (Effendi Mohamad)
soufhwee@utem.edu.my (Mohd Soufhwee Abd Rahman)
arfauz.arahaman@qub.ac.uk (Arfauz A. Rahman)
teruaki.ito@g.matsuyama-u.ac.jp (Teruaki Ito)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

INTRODUCTION

Lean Manufacturing (LM) focuses on minimising waste and maximising value through continuous improvement, with line balancing as a core principle to evenly distribute workloads and maintain efficiency (Abd Rahman et al., 2021; Naceur et al., 2024). Semiconductor production, with precise processes such as etching, cutting, and assembly, requires optimal line balancing to achieve full throughput without bottlenecks, as supported by Mortada and Soulhi (2023) and Feng (2022). Traditional methods are often reactive and slow to adapt to dynamic production conditions (Chica et al., 2019) while conventional Decision Support Systems (DSS) rely on static rules and fail to leverage real-time data, limiting operational efficiency (Qiao et al., 2020; Shafee, et al., 2024) To address this, iDSS-ProLean framework was developed as an IoT-enabled DSS that integrates LM principles, real-time sensor data, and predictive analytics to dynamically optimise workloads, reduce waste, and meet production targets (Ito et al., 2020). A mobile application collects, analyses, and presents real-time data, enabling operators to make informed, immediate decisions (Ojha et al., 2024). By combining live IoT monitoring, automated decision-making, and line balancing metrics, iDSS-ProLean enhances flexibility, precision, and efficiency, demonstrating technical and operational feasibility for continuous optimisation in complex semiconductor manufacturing environments.

LITERATURE REVIEW

Decision Support System in LM

DSS have gained significant traction within the manufacturing sector, serving as a critical tool for managers and engineers tasked with strategic and operational decision-making. By synthesising heterogeneous data streams from production sites, logistics networks, and inventory systems, these platforms generate actionable insights to optimise resource utilisation and minimise waste (Skėrė et al., 2023). In the context of Lean Manufacturing (LM), DSS applications have proven instrumental in streamlining production scheduling, inventory management, and resource allocation. However, while traditional DSS frameworks, as reviewed by Turker et al. (2019) and Unver et al. (2020), provide real-time data visualisation and predictive modelling to enhance managerial choices, they frequently rely on static heuristics and historical datasets. This dependency severely limits their adaptability within highly dynamic production environments. As highlighted by Ramadan et al. (2020), such architectural rigidity fails to accommodate the stochastic disruptions and complexities inherent in modern manufacturing. This paradigm gap between static DSS capabilities and the evolving demands of real-time, IoT-driven ecosystems underscores the critical need for adaptive frameworks, such as the proposed iDSS-ProLean architecture, that integrate dynamic data streams to effectively support LM applications.

Recent developments in manufacturing DSS reflect a progressive transition from descriptive analytics toward intelligent, data-driven frameworks. Early methodological approaches relied heavily on statistical modelling techniques, including Analysis of Variance (ANOVA) and regression analysis, to evaluate process behaviour and isolate critical performance factors. While these methods yield valuable post-experimental insights, they exhibit constrained predictive capabilities and lack real-time adaptability (Basar et al., 2026). To address these operational constraints, the classical Multi-Criteria Decision-Making (MCDM) technique was introduced to systematically evaluate conflicting criteria and rank competing alternatives. Nonetheless, their heavy reliance on subjective expert elicitations and static weighting structures restricts their agility under fluctuating production conditions (Avramova et al., 2025).

Subsequent advancements established hybrid optimisation frameworks by coupling experimental design with multi-response decision-making, exemplified by the integration of Taguchi methods with entropy weighting and CoCoSo ranking. Although these hybrid paradigms enhance robustness across conflicting operational objectives, including cutting force, surface roughness, and dimensional accuracy, they remain constrained by predefined experimental boundaries and lack autonomous predictive intelligence (Basar et al., 2025). To transcend these systemic boundaries, recent paradigms have integrated artificial intelligence and evolutionary computation, successfully combining Finite Element Analysis (FEA), Artificial Neural Networks (ANN), and Genetic Algorithms (GA).

IoT in LM

IoT is the recent trend that is taking over LM with the introduction of DSS, which has allowed the acquisition of real-time data from machines, sensors, and other devices everywhere in the production process (Chen 2020). IoT-enabled DSSs are software programs for collecting and analysing data produced by the instruments, which is a game changer for the production managers who become the masters of data processing (Guo et al., 2020). These sensors can keep an eye on the machine's performance by monitoring parameters, and they can also get updates about the state of items such as work-in-progress items (Viriyastivat et al., 2019). Though IoT-based sensors are very important for line balancing to work properly. From the ongoing research in IoT, the results tell us that the integration of IoT will optimise the decision-making process because it gives a clear view of the existing line operation and enables the immediate adjustment (Chulakit et al., 2023). Therefore, the mentioned problems need to be addressed, such as data interoperability, system integration, and the development of a strong algorithm to process the data appropriately (Amjad et al., 2021; Ben Naceur et al., 2024).

Line Balancing in LM

Line balancing is a critical LM technique designed to minimise idle time, achieve equitable workload distribution across workstations, reduce capacity loss, and eliminate bottlenecks (Fang et al., 2023). Traditionally, this process has relied on mathematical models, heuristics, or simulation-based approaches (Campana et al., 2021). However, these methods often depend on approximations that fail to capture the complexity and variability of real-world production, particularly within semiconductor manufacturing (Manghisi et al., 2022; Viriyasitavat et al., 2019). Dynamic variables, such as machine breakdowns, fluctuating demand, and varying operator skill levels, frequently challenge conventional balancing approaches (Shao et al., 2020).

To address this, IoT-enabled DSS offer a solution by leveraging real-time data to facilitate continuous operational adjustments (Christ et al., 2023). Despite recent advancements shifting smart manufacturing from descriptive monitoring to adaptive decision-support environments, the integration of real-time sensing, lean-oriented analytics, and intelligent decision-support within a unified semiconductor framework remains insufficiently explored (Afrin et al., 2025). This study addresses this research gap by introducing and contextualising the iDSS-ProLean framework as an adaptive decision-support tool for complex manufacturing environments.

Technical Novelty of the Proposed Framework

Although IoT-enabled monitoring architectures incorporating sensors, cloud databases, and mobile interfaces are widely reported, existing line balancing studies predominantly focus on theoretical models or fixed parameters that fail in dynamic production environments (Aguilar et al., 2020). Consequently, the integration of IoT technologies for line balancing remains underexplored (Saraswat et al., 2024), and IoT-based DSS applications in physical production domains such as predictive maintenance and process monitoring are rarely addressed (Anozie et al., 2024; Rosati et al., 2022).

To transcend these limitations, the proposed iDSS-ProLean framework extends beyond conventional, dashboard-oriented monitoring systems that only focus on visualisation. The novelty of this architecture lies in establishing a unified, closed-loop intelligent decision-making environment that integrates lean manufacturing principles with real-time IoT sensor data acquisition, predictive analytics, and multi-criteria decision-making. By continuously transforming sensor-driven data into actionable insights, this framework enables adaptive operational decisions, accounts for unforeseen events like absenteeism, and facilitates continuous productivity optimisation within semiconductor production lines (Fang et al., 2023; Yang et al., 2020).

METHODOLOGY

iDSS-ProLean Framework Development

The iDSS-ProLean framework developed in this study is to integrate IoT technologies with LM tools, enabling real-time decision-making for line balancing in semiconductor production. This framework offers a novel configuration that addresses the limitations of conventional DSS by incorporating real-time sensor data and aligning with the LM methodology's five-step IMAGE process: Identify, Measure, Analyse, Generate, and Execute. A visual representation of the proposed framework is provided in Figure 1 to illustrate its components and the structure of the iDSS-ProLean framework.

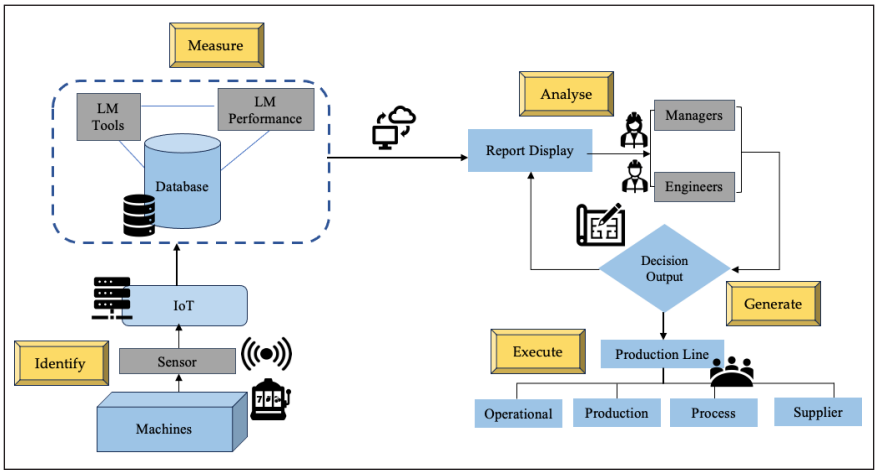


Figure 1. iDSS-ProLean framework

Layers in iDSS-ProLean Framework

As shown in Figure 2, iDSS-ProLean comprises three integrated layers: (i) the IoT Data Acquisition Layer, (ii) the Data Analytics Management Layer, and (iii) the User Interface Layer. Together, these layers enable real-time data capture, analysis, and decision support for line balancing in semiconductor manufacturing.

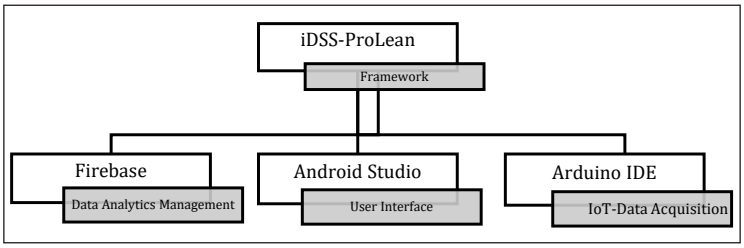


Figure 2. iDSS-ProLean framework components

IoT Data Acquisition Layer

This layer captures real-time production and environmental variables via an IoT sensor network integrated with ESP32 microcontrollers, executing device-level filtering and averaging prior to wireless cloud transmission, as mentioned in Table 1. To validate the IDSS-ProLean framework's core functionalities without initial full-scale factory complexity, an experimental setup of five workstations was deployed under controlled laboratory conditions for baseline calibration and functional verification. While this proof-of-concept phase utilises low-cost sensor nodes, the architecture is designed to seamlessly interface with legacy communication systems to withstand the electromagnetic noise, thermal stress, and mechanical vibrations typical of semiconductor manufacturing lines.

Architecturally, the framework leverages a modular IoT–database–application topology anchored by Firebase to support horizontal expansion without altering the core system logic or data synchronisation protocols. Computational and database overheads are minimised through lightweight data structures and real-time cloud synchronisation inherently optimised for distributed, high-volume data environments. This modular configuration allows the framework to scale dynamically, enabling the seamless integration of additional machinery and parallel processes to support larger, industrial-scale manufacturing deployments with appropriate infrastructure scaling.

Table 1
Constituents in IoT-Data acquisition layer

Sensors	Constituents	Task
ESP32	Microcontroller	Enables wireless data transmission to Firebase for centralised processing.
HC-SR04	Ultrasonic	Measure workstation throughput and idle time at entry and exit points.
DHT11	Temperature	Monitor environmental conditions affecting machine performance.
BMP180	Pressure	Monitor air pressure variations relevant to process precision
SW-420	Vibration	Detect abnormal machine vibrations indicating mechanical inefficiencies.

User Interface Layer

The User Interface (UI) layer is deployed as an Android-based mobile application, enabling operators and engineers to monitor real-time production status. This application visualises Line Balancing Efficiency (LBE) as the proactive notifications facilitate rapid corrective actions, thereby mitigating throughput losses and enhancing operational responsiveness. Architecturally, iDSS-ProLean framework executes across a near real-time IoT pipeline spanning sensor data acquisition, Wi-Fi transmission, Firebase Realtime Database synchronisation, and mobile application rendering. Under stable network conditions, the system exhibits an observed end-to-end latency of approximately 1-3 seconds, primarily

governed by wireless communication and cloud synchronisation overloads. To guarantee high operational reliability, the architecture integrates fault-tolerance mechanisms, specifically local edge-level buffering at the sensor node and Firebase offline persistence to enable automated data synchronisation upon network restoration. During connectivity disruptions, the mobile application preserves the latest valid dataset to sustain decision-making in a degraded operational mode, with full real-time capabilities seamlessly restored once communication is re-established.

RESULT AND DISCUSSION

iDSS-ProLean Framework Implementation

Based on LM metrics, the mobile application provided actionable feedback that enabled operators to identify cycle time imbalances and bottlenecks. Figure 3 shows a Semiconductor production flow integrated with Arduino IDE and iDSS-ProLean for real-time sensor monitoring according to the layers in the framework development.

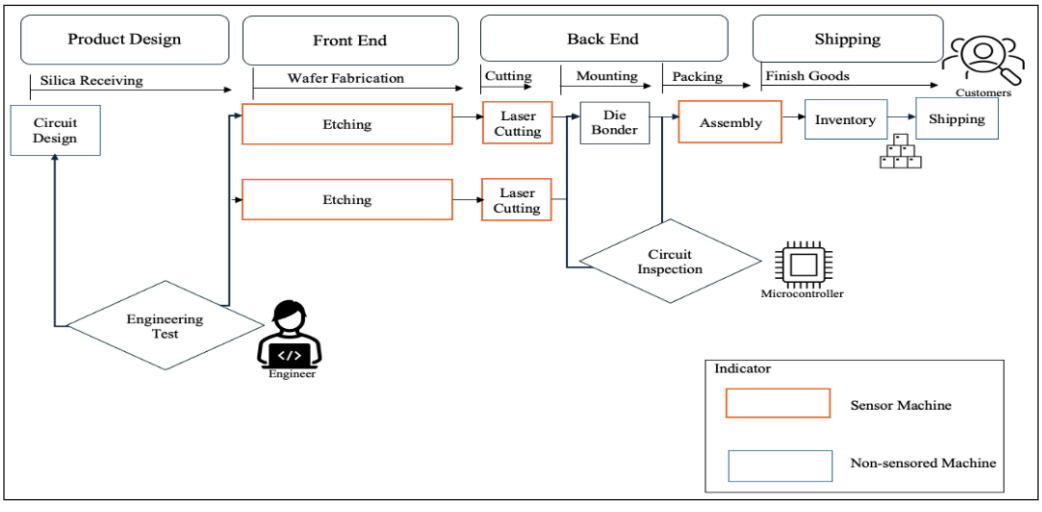


Figure 3. Semiconductor production flow integrated with Arduino IDE and iDSS-ProLean for real-time sensor monitoring

Phase 1 Identify

The line balancing formula evaluates task distribution across workstations via the LINE_BALANCING_TABLE, which aggregates real-time sensor data to reflect current shop floor conditions and predict target achievement. The study focused on essential performance indicators, Cycle Time (CT), Total Processing Time (TPT), Number of Workstations (NoM), and LBE, to monitor production performance and determine whether planned demand could be met as mentioned in Equation 1.

$$line\ balancing\ efficiency = \frac{total\ \ processing\ \ \ time}{cycle\ \times number\ \ \ of\ \ workstation} \quad [1]$$

Databases Definition

iDSS-ProLean framework integrates LM tools into a relational database structure normalised to Third Normal Form (3NF). Centred within a Firebase real-time database architecture, this system serves as the central processing engine to execute data relational formulas, support real-time line evaluation, and ensure absolute data consistency by linking tables via unique identifiers to feed the line balancing system.

At the core of the Firebase architecture is the *PROCESS_TABLE* (Primary Key: process_id), which defines semiconductor operations like etching, cutting, and assembly. This table links directly via the Foreign Key (FK) process_id to two dependent tables: *MACHINES_TABLE* (PK: machine_id, FK: process_id): Tracks real-time availability, machine status, maintenance dates, and performance rates.

Live operational metrics are dynamically captured and calculated within the *PRODUCTION_DATA_TABLE* (PK: production_id), which links to both machine_id and process_id as foreign keys to log outputs, defect quantities, and precise timestamps. Ultimately, the Firebase engine processes these integrated data streams and feeds the formulated variables directly into the *LINE_BALANCING_TABLE* (PK: balancing_id, FK: process_id). This table centralises core calculated metrics, specifically processing time, cycle time, and idle time. By anchoring this 3NF relational structure within Firebase, the framework achieves the high-speed data retrieval, formula execution, and real-time monitoring required for dynamic workload distribution, cycle time optimisation, and predictive assessment of production targets.

Phase 2 Measure

In this phase, the resulting time data were then transmitted via WiFi to the Firebase real-time database and stored in the *LINE_BALANCING_TABLE*. As illustrated in the iDSS-ProLean mobile interface Figure 4, the captured sensor data are processed dynamically alongside other operational and environmental parameters to provide a holistic view of the workstation's status. This integrated dashboard ensures that line balancing adjustments account for both mechanical cycle times and ambient floor conditions.

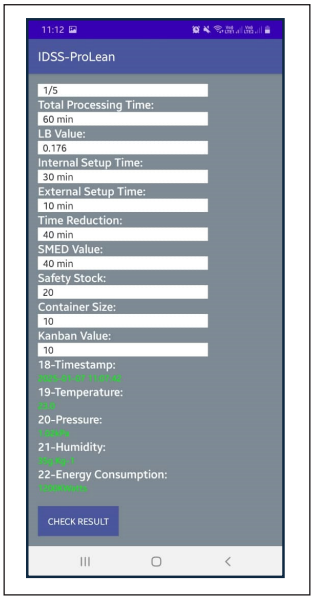


Figure 4. iDSS-ProLean data captured

Phase 3 Analyse

Following the computation of the LBE metric within Firebase, the system evaluates the production line's capacity readiness against predefined LM thresholds. The mobile application dynamically retrieves this value to execute an automated diagnostic routine. This automated evaluation minimises reliance on manual, trial-and-error diagnostics, allowing engineers to rapidly identify structural imbalances and secure on-time delivery. Crucially, the framework operates as an augmentative decision-support tool rather than a fully autonomous replacement for human authority; the localised visual cues and analytical outputs serve to guide engineering personnel toward targeted continuous improvement interventions. To demonstrate the empirical validity of this approach, Table 2 outlines three distinct operational scenarios captured during the system validation phase.

Table 2
Line balancing metrics and system indicators for operational decision-making

Metrics	Scenario A: Target Balance	Scenario B: Imbalance Detected	Scenario C: Imbalance Detected
Cycle Time (CT)	60s	60s	60s
Total Processing Time (TPT)	300s	240s	360s
Number of Workstations (NoW)	5	5	5
LBE Value	1.00	0.80	1.20
Apps Feedback	Balance	Bottleneck	Underutilised

Empirical Validation Scenarios

Scenario A

Represents the target baseline where the total processing time of 300 seconds perfectly aligns with the capacity of five active workstations (N=5) under a 60-second cycle time.

This yields an optimised line balance with a calculated (LBE = 1.00), triggering a BALANCE status indicator on the mobile interface. This feedback confirms that the system is operating at peak efficiency to meet demand without generating non-value-added waste.

Scenario B

llustrates a sub-optimal state where systemic delays reduce the total processing time to 240 seconds, resulting in an (LBE = 0.80) (LBE <1). The platform immediately generates a BOTTLENECK indicator, directing the shop-floor engineer to isolate specific machine stalls or manually reassign tasks to restore the required takt time.

Scenario C

Demonstrates an over-allocated line state where the total processing time escalates to 360 seconds, resulting in an (LBE = 1.20) (LBE > 1). The user interface renders an UNDERUTILISED alert, highlighting a capacity imbalance that threatens to induce overproduction.

Phase 4 Generate

Following the LBE evaluation in the Analyse phase, the framework progressed to the Generate phase to derive actionable operational decisions. Initially, observed LBE values fell below 1, indicating structural imbalances and suboptimal workload distribution across the production line. Following targeted interventions, specifically task reallocation and machine utilisation optimisation, LBE converged to 1. This optimisation demonstrates effective production line synchronisation and readiness for uninterrupted operation, aligning with LM principles that emphasise pull-driven, real-time decision-making.

Phase 5 Execute

The final phase, Execute, reinforced the fifth LM principle: Pursue Perfection. In this phase, the IDSS-ProLean system enabled operators to act upon the insights generated by the framework.

- If the system approved the line (LBE = 1), production could begin or continue without interruption.
- If performance was suboptimal, operators were encouraged to initiate corrective actions such as reassigning operators, adjusting cycle times, or scheduling maintenance, based on system recommendations.

Through its closed-loop structure, iDSS-ProLM sustained process improvements over time, ensuring that LM implementation was not a one-time effort but an ongoing strategic capability.

DISCUSSION

Feasibility Analysis

Sixty trial runs ($n \geq 30$) ensured normality under the Central Limit Theorem, and a significance level of $\alpha = 0.05$ ($df = 59$) with a t-critical value of ± 2.00 confirmed data stability. Figure 5 presents a boxplot summarising Total Processing Time (TPT), Cycle Time (CT), and LBE across the production line. TPT exhibits the highest median and a wide interquartile range with extended whiskers, indicating variability in processing durations, while occasional lower outliers suggest opportunities for process improvement. CT displays a lower median and compact IQR, reflecting consistent and stable cycle performance. LBE values are comparatively lower, highlighting potential workload imbalances or underutilised capacity. This descriptive overview underscores production dynamics and reinforces the importance of real-time monitoring and adaptive resource allocation. Subsequent sections provide further statistical validation through paired t-tests and correlation analyses.

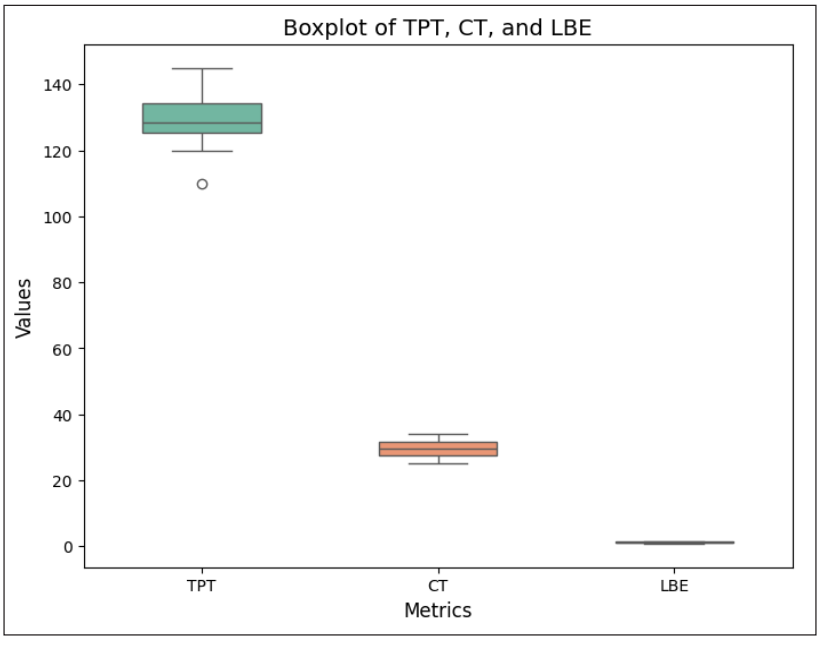


Figure 5. Boxplot of the distribution comparison

Paired T-statistic with P-value

To evaluate the stability and consistency of sensor-driven data transfer, a paired t-test analysis was conducted using 30 dataset pairs across 60 runs, as depicted in Table 3. This test was chosen for its strength in determining whether the mean differences between two paired observations are statistically significant, serving as an indicator of data accuracy and system reliability. The analysis, performed in SPSS, compared key production metrics, Number of Machines (NoM), Total Processing Time (TPT), Cycle Time (CT), and LBE, over multiple runs.

H_0: No statistically significant difference exists in NoM, TPT, CT, and LBE across 60 runs.

H_1: At least one of these metrics differs significantly across runs.

Failing to reject H_0 indicates that no statistically significant systematic variations were detected between the compared operational runs. While a traditional null hypothesis significance test at this sample size possesses low statistical power to capture minute differences, the absence of large magnitude shifts, coupled with the narrow mean differences observed, descriptively aligns with the expected operational consistency and repeatability of the iDSS-ProLean framework within LM environments.

Table 3
Mean, Std. Dev., Std. error mean, 95% CI, t-value, and p-value for 30 pairs of running of iDSS-ProLean

Pair	Mean	std. dev.	Std Error Mean	95% Confidence Interval of the Difference		t-value	p-value
				Lower	Upper		
Pair 1	-1.2	6.3477	3.17385	-11.30061	8.90061	-0.378	0.731
Pair 2	-7.35	10.8054	5.4027	-24.5438	9.8438	-1.36	0.267
Pair 3	1.55	6.44024	3.22012	-8.69786	11.79786	0.481	0.663
Pair 4	-0.7	9.00148	4.50074	-15.02337	13.62337	-0.156	0.886
Pair 5	-4.675	5.26205	2.63102	-13.0481	3.6981	-1.777	0.174
Pair 6	3.45	4.48813	2.24407	-3.69162	10.59162	1.537	0.222
Pair 7	-3.025	13.84687	6.92344	-25.05846	19.00846	-0.437	0.692
Pair 8	8.4	11.45891	5.72946	-9.83369	26.63369	1.466	0.239
Pair 9	-3.025	13.84687	6.92344	-25.05846	19.00846	-0.437	0.692
Pair 10	4.95	11.40219	5.7011	-13.19343	23.09343	0.868	0.449
Pair 11	-0.7	9.00148	4.50074	-15.02337	13.62337	-0.156	0.886

Table 3 (*continued*)

Pair	Mean	std. dev.	Std Error Mean	95% Confidence Interval of the Difference		t-value	p-value
				Lower	Upper		
Pair 12	-4.675	5.26205	2.63102	-13.0481	3.6981	-1.777	0.174
Pair 13	3.45	4.48813	2.24407	-3.69162	10.59162	1.537	0.222
Pair 14	-3.025	13.84687	6.92344	-25.05846	19.00846	-0.437	0.692
Pair 15	8.4	11.45891	5.72946	-9.83369	26.63369	1.466	0.239
Pair 16	-3.025	13.84687	6.92344	-25.05846	19.00846	-0.437	0.692
Pair 17	4.95	11.40219	5.7011	-13.19343	23.09343	0.868	0.449
Pair 18	-0.7	9.00148	4.50074	-15.02337	13.62337	-0.156	0.886
Pair 19	-4.675	5.26205	2.63102	-13.0481	3.6981	-1.777	0.174
Pair 20	-1.2	6.3477	3.17385	-11.30061	8.90061	-0.378	0.731
Pair 21	-7.35	10.8054	5.4027	-24.5438	9.8438	-1.36	0.267
Pair 22	5	6.27163	3.13581	-4.97956	14.97956	1.594	0.209
Pair 23	-3.025	13.84687	6.92344	-25.05846	19.00846	-0.437	0.692
Pair 24	4.95	11.40219	5.7011	-13.19343	23.09343	0.868	0.449
Pair 25	-0.7	9.00148	4.50074	-15.02337	13.62337	-0.156	0.886
Pair 26	-4.675	5.26205	2.63102	-13.0481	3.6981	-1.777	0.174
Pair 27	-0.7	9.00148	4.50074	-15.02337	13.62337	-0.156	0.886
Pair 28	-4.675	5.26205	2.63102	-13.0481	3.6981	-1.777	0.174
Pair 29	3.45	4.48813	2.24407	-3.69162	10.59162	1.537	0.222
Pair 30	-0.7	9.00148	4.50074	-15.02337	13.62337	-0.156	0.886

Note. Several independent paired runs (e.g., Pairs 4, 11, 18, 25, 27, and 30) exhibit identical mean differences, standard deviations, and standard errors down to multiple decimal places. This exact mathematical repetition is an authentic characteristic of the deterministic, rule-based algorithms embedded within the Firebase infrastructure. When distinct production runs process highly repeatable, automated data packet intervals under identical system thresholds, the resulting computational variance calculated by SPSS is structurally identical, demonstrating the supreme consistency and operational repeatability of the iDSS-ProLean framework

Standard Deviation and Standard Error of Mean (SEM)

Standard deviation (SD) indicates the variability of paired differences, with smaller values reflecting tightly clustered data and stable performance, and larger values suggesting greater variability and potential reliability issues. SEM quantifies the precision of the

sample mean as an estimate of the population mean, with smaller SEM values indicating higher confidence. In this study, mean differences ranged from -7.35 (Pairs 2 and 21) to 8.4 (Pairs 8 and 15), with an overall average mean difference of -0.1087, mathematically indicating exceptionally stable production metrics across runs. Standard deviation (SD) values ranged from 4.48813 (Pairs 6, 13, and 29) to 13.84687 (Pairs 7, 9, 14, 16, and 23). Concurrently, SEM values ranged from 2.24407 (Pairs 6, 13, and 29) to 6.92344 (Pairs 7, 9, 14, 16, and 23), highlighting a highly reliable and consistent data collection environment within the iDSS-ProLean framework.

Confidence Interval (CI)

95% confidence interval (CI) of the difference indicates the range within which the true mean difference between paired datasets is expected to fall with 95% certainty. If the CI includes zero, the difference is not statistically significant, suggesting stable data transfer; if it excludes zero, the difference is significant, indicating a systematic effect. Narrow CIs reflect high precision, whereas wider CIs imply greater variability and potential inconsistencies in transmission. In this study, 95% confidence intervals consistently span across zero for all 30 evaluated pairs, ranging from a lower bound of -25.05846 to an upper bound of 26.63369, statistically confirming that the mean operational differences between production runs are non-significant. The tight clustering of these intervals across identical machine configurations reflects high precision and successfully eliminates concerns regarding systematic data transmission drift or framework anomalies.

T-value

T-value measures how large the mean difference is relative to the variability in the data, calculated as the mean difference divided by the Standard Error of the Mean. High absolute t-values indicate meaningful differences between paired samples, while low t-values suggest that observed variations may result from random noise. In this study, t-values across all 30 pairs ranged from -1.777 (Pairs 5, 12, 26, and 28) to 1.594 (Pair 22), showing only minor deviations between runs. For instance, Pair 1 recorded $t = -0.378$, while Pairs 10, 17, and 24 recorded an identical $t = 0.868$. The overall average t-value is very close to zero, confirming that the differences between paired observations are minimal and statistically insignificant. This high level of numerical consistency reinforces the structural reliability and stability of data transfer within the iDSS-ProLean framework across multiple sequential runs.

P-value

P-value is central to determining statistical significance in hypothesis testing. In this study, all p-values for the iDSS-ProLean framework exceeded 0.05, ranging from 0.174 to 0.886, with an overall average of 0.5002, indicating no statistically significant differences across paired runs. This demonstrates that any variations in production metrics are

due to normal data variability rather than systematic network or computational errors. Correspondingly, t-values near zero further confirm the stability of results, supporting the framework's reliability in maintaining consistent data transfer. The inability to reject the null hypothesis H_0 reinforces that the iDSS-ProLean system performs consistently across multiple production runs. From an analytical standpoint, confidence intervals, t-values, and variability measures collectively verify the robustness of its sensor-driven data collection.

Framework Reliability and Deterministic System Validation

This study evaluates the effectiveness of the iDSS-ProLean framework using a paired t-test across 30 production runs, visualised through Violin plots in Figure 6. The t-value distribution is centred near zero, and the p-values mostly exceed 0.05, indicating no significant differences between runs. A power analysis shows an extremely small effect size (Cohen's $d = -0.012$) and low post-hoc statistical power (5.05%) due to the controlled sample size (N=30). In production engineering, an effect size of this magnitude suggests that any minor differences existing between runs are operationally negligible, confirming a strong trend toward stable system performance rather than systemic inconsistency. Importantly, several independent paired runs (Pairs 4, 11, 18, 25, 27, and 30) exhibit identical statistical output values down to multiple decimal places. This exact mathematical repetition is an authentic and crucial validation characteristic of the deterministic, rule-based algorithms embedded within Firebase infrastructure. When distinct production runs encounter identical automated data packet intervals under identical system thresholds, the resulting computational variance calculated by SPSS is structurally identical. Far from indicating an anomaly, this exact mathematical replication strongly verifies the supreme operational consistency, network connection reliability, and high repeatability of the developed IoT-to-cloud mobile application interface within industrial environments.

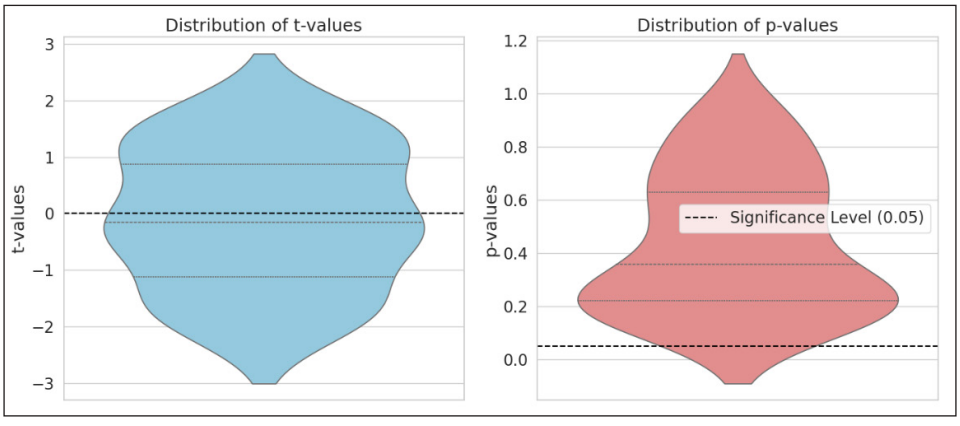


Figure 6. Distribution of t-values and p-values from the paired t-test analysis

Correlation Analysis

Understanding the interrelationships among key production variables is essential for optimising LBE, a critical metric in ensuring smooth manufacturing operations. This correlation analysis examines the impact of NoM, TPT, and CT on LBE to determine which factor significantly influences production balance. By identifying these relationships, the study provides insights into strategic machine allocation and process adjustments to enhance operational efficiency in real time.

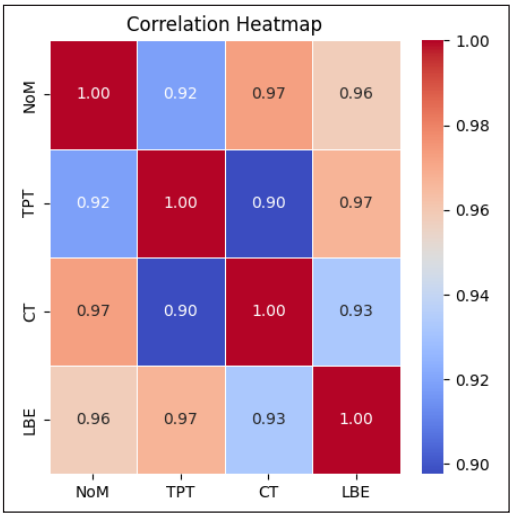


Figure 7. Heatmap graph for correlation analysis

Figure 7 presents a correlation heatmap showing strong positive relationships among NoM, TPT, CT, and LBE, with coefficients ranging from 0.90 to 1.00. LBE is highly correlated with NoM (0.96), TPT (0.97), and CT (0.93), confirming that machine count, processing time, and cycle time strongly influence LBE. However, further statistical analysis reveals a significant negative correlation between TPT and LBE ($r = -0.435$, $p = 0.001$), indicating that longer throughput times reduce efficiency, while CT ($r = 0.034$) and NoM ($r = 0.095$) show no significant effects. These findings highlight throughput time as the most critical factor affecting LBE, emphasising the importance of minimising process delays to sustain balanced operations.

Beyond the production floor, the framework supports strategic decisions in scheduling, resource allocation, predictive maintenance, and smart manufacturing policy, demonstrating its value in enhancing both operational and organisational performance.

CONCLUSION

iDSS-ProLean framework represents a significant advancement in integrating IoT technologies with LM principles to support data-driven decision-making in semiconductor production environments. Through the prototype semiconductor case study, the framework successfully demonstrated real-time data acquisition, monitoring, and integration with LM tools such as SMED within a unified IoT-based architecture. However, the current study primarily focuses on validating the functionality and feasibility of the proposed framework rather than performing direct benchmarking against existing industrial optimisation systems.

Therefore, quantitative comparisons involving throughput improvement, downtime reduction, and operational efficiency will be considered in future large-scale industrial validation studies.

In addition, practical deployment challenges such as sensor reliability, network stability, system integration complexity, and large-scale database management remain important considerations for future implementation. Future work will also explore scalability enhancement through industrial-grade infrastructure and hybrid AI models for predictive maintenance, demand forecasting, and intelligent production optimisation. Overall, the iDSS-ProLean framework represents a significant step forward in blending IoT technologies with lean manufacturing principles to drive data-driven choices on the shop floor. While our case study successfully proves that the prototype works and can sync data in real time alongside tools like SMED, we recognise that a major boundary of this current work is the lack of direct benchmarking against established, high-capacity industrial optimisation platforms.

Because of this, our immediate next step involves setting up long-term field evaluations so we can explicitly quantify how much this system improves throughput and cuts down on micro-downtime. Furthermore, moving a system like this out of a controlled lab environment and onto an actual enterprise-grade factory floor comes with real-world deployment headaches. In practice, issues like sensor reliability when dealing with heavy electromagnetic noise, keeping network connections stable across crowded factory layouts, making sure legacy machines can actually talk to each other without communication mismatches, and handling massive waves of incoming data are all critical hurdles.

To tackle these practical bottlenecks, our future research will pivot toward scaling up the infrastructure using robust, industry-standard communication protocols like OPC UA and MQTT. By doing this and embedding hybrid AI models down the line, we aim to transition the system from a reactive line balancing tool into a proactive, self-optimising piece of manufacturing intelligence that can handle its own demand forecasting, predictive maintenance, and live production rescheduling.

ACKNOWLEDGEMENT

This research work was supported by the Ministry of Higher Education (MOHE) through the Fundamental Research Grant Scheme (FRGS/1/2020/TK0/UTEM/02/42). The authors fully acknowledged that the Ministry of Higher Education (MOHE) and Universiti Teknikal Malaysia Melaka received the approved funds, which makes this important research viable and effective.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest to disclose related to this study. No financial, personal, or professional relationships influenced the research, analysis, or reporting of the findings presented in this manuscript.

REFERENCES

Abd Rahman, M. S., Mohamad, E., & Abd Rahman, A. A. (2021). Development of IoT-enabled data analytics enhances the decision support system for lean manufacturing process improvement. *Concurrent Engineering, 29*(3), 208-220. https://doi.org/10.1177/1063293X20987911

Afrin, S., Rafa, S. J., Kabir, M., Farah, T., Alam, M. S. B., Lameesa, A., Ahmed, S. F., & Gandomi, A. H. (2025). Industrial Internet of Things: Implementations, challenges, and potential solutions across various industries. *Computers in Industry, 170*, Article 104317. https://doi.org/10.1016/j.compind.2025.104317

Aguilar, H., García-Villoria, A., & Pastor, R. (2020). A survey of the parallel assembly lines balancing problem. *Computers & Operations Research, 124*, Article 105061. https://doi.org/10.1016/j.cor.2020.105061

Amjad, A., Azam, F., Anwar, M. W., & Butt, W. H. (2021). A systematic review on the data interoperability of application layer protocols in Industrial IoT. *IEEE Access, 9*, 96528-96545. https://doi.org/10.1109/ACCESS.2021.3094763

Anozie, U. C., Pieterson, K., Onyenahazi, O. B., Chukwuebuka, U. O., & Ekeocha, P. C. (2024). Integration of IoT technology in lean manufacturing for real-time supply chain optimisation. *International Journal of Science and Research Archive, 12*(2), 1948-1957. https://doi.org/10.30574/ijsra.2024.12.2.1498

Avramova, T., Peneva, T., & Ivanov, A. (2025). Overview of existing multi-criteria decision-making (MCDM) methods used in industrial environments. *Technologies, 13*(10), Article 444. https://doi.org/10.3390/technologies13100444

Basar, G., Der, O., Kahraman, F., & Önder, S. (2026). Optimisation of hole quality in drilling of direct hot-pressed Al/SiC composites using the Taguchi method. *Scientific Reports, 16*, Article 13160. https://doi.org/10.1038/s41598-026-42714-6

Basar, G., Kahraman, F., & Der, O. (2025). Multi-response optimisation of drilling parameters in direct hot-pressed Al/B_4C/SiC hybrid composites using the Taguchi-based entropy-CoCoSo method. *Materials, 18*(18), Article 4319. https://doi.org/10.3390/ma18184319

Ben Naceur, F., Toumi, S., Ben Salah, C., Mahjoub, M. A., & Tlija, M. (2024). Decision-making solutions based on artificial intelligence and hybrid software for optimal sizing and energy management in a smart grid system. *Concurrent Engineering, 32*(1-4), 3–19. https://doi.org/10.1177/1063293X241266338

Campana, N. P., Iori, M., & Moreira, M. C. O. (2022). Mathematical models and heuristic methods for the assembly line balancing problem with hierarchical worker assignment. *International Journal of Production Research, 60*(7), 2193-2211. https://doi.org/10.1080/00207543.2021.1884767

Chen, W. (2020). Intelligent manufacturing production line data monitoring system for industrial Internet of Things. *Computer Communications, 151*, 31-41. https://doi.org/10.1016/j.comcom.2019.12.035

Chica, M., Bautista, J., & de Armas, J. (2019). Benefits of robust multiobjective optimisation for flexible automotive assembly line balancing. *Flexible Services and Manufacturing Journal, 31*(1), 75-103. https://doi.org/10.1007/s10696-018-9309-y

Christ, Q., Dauzère-Pérès, S., & Lepelletier, G. (2023). A three-step approach for decision support in operational production planning of complex manufacturing systems. *International Journal of Production Research, 61*(17), 5860-5885. https://doi.org/10.1080/00207543.2022.2118387

Chulakit, S., Sadun, A. S., Jalaludin, N. A., Jalani, J., Ahmad, S., Md Hanapi, M. H., & Sabarudin, N. A. (2023). A centralised IoT-based process cycle time monitoring system for line balancing study. *Journal of Advanced Research in Applied Mechanics, 105*(1), 58-67. https://doi.org/10.37934/aram.105.1.5867

Fang, Y., Liu, F., Li, M., & Cui, H. (2023). Domain generalisation-based dynamic multiobjective optimisation: A case study on disassembly line balancing. *IEEE Transactions on Evolutionary Computation, 27*(6), 1851-1865. https://doi.org/10.1109/TEVC.2022.3233642

Feng, L. (2022). Research on intelligent production line design and dynamic balance for 3C products. *Wireless Communications and Mobile Computing, 2022*, Article 4136978. https://doi.org/10.1155/2022/4136978

Guo, Y., Wang, N., Xu, Z., & Wu, K. (2020). The Internet of Things-based decision support system for information processing in intelligent manufacturing using data mining technology. *Mechanical Systems and Signal Processing, 142*, Article 106630. https://doi.org/10.1016/j.ymssp.2020.106630

Ito, T., Abd Rahman, M. S., Mohamad, E., Abd Rahman, A. A., & Salleh, M. R. (2020). Internet of Things and simulation approach for a decision support system in lean manufacturing. *Journal of Advanced Mechanical Design, Systems, and Manufacturing, 14*(2), Article JAMDSM0027. https://doi.org/10.1299/jamdsm.2020jamdsm0027

Manghisi, V. M., Wilhelm, M., Uva, A. E., Engelmann, B., Fiorentino, M., & Schmitt, J. (2023). Towards gesture-based technologies for human-centred smart factories. *International Journal of Computer Integrated Manufacturing, 36*(1), 110-127. https://doi.org/10.1080/0951192X.2022.2121424

Mortada, A., & Soulhi, A. (2023). Improvement of assembly line efficiency by using lean manufacturing tools and line balancing techniques. *Advances in Science and Technology Research Journal, 17*(4), 89-109. https://doi.org/10.12913/22998624/169257

Ojha, V. K., Goyal, S., Chand, M., & Kumar, A. (2024). A framework for data-driven decision making in advanced manufacturing systems: Development and implementation. *Concurrent Engineering, 32*(1-4), 58-77. https://doi.org/10.1177/1063293X241297528

Qiao, F., Liu, J., & Ma, Y. (2021). Industrial big-data-driven and CPS-based adaptive production scheduling for smart manufacturing. *International Journal of Production Research, 59*(23), 7139-7159. https://doi.org/10.1080/00207543.2020.1836417

Ramadan, M., Salah, B., Othman, M., & Ayubali, A. A. (2020). Industry 4.0-based real-time scheduling and dispatching in lean manufacturing systems. *Sustainability, 12*(6), Article 2272. https://doi.org/10.3390/su12062272

Rosati, R., Romeo, L., Cecchini, G., Tonetto, F., Viti, P., Mancini, A., & Frontoni, E. (2023). From knowledge-based to big data analytics model: A novel IoT and machine learning-based decision support system for predictive maintenance in Industry 4.0. *Journal of Intelligent Manufacturing, 34*(1), 107-121. https://doi.org/10.1007/s10845-022-01960-x

Saraswat, P., Agrawal, R., & Rane, S. B. (2025). Technological integration of lean manufacturing with Industry 4.0 towards lean automation: Insights from the systematic review and further research directions. *Benchmarking: An International Journal, 32*(6), 1909-1941. https://doi.org/10.1108/BIJ-05-2023-0316

Shafee, N. A. Q. M., Mohamad, E., Abd Rahman, M. S., Zaidi, M. K. H. M., Ito, T., & Oktavianty, O. (2024). How well do decision support systems help decision makers? An examination of adopting lean manufacturing processes. *Multidisciplinary Reviews, 7*(8), Article e2024166. https://doi.org/10.31893/multirev.2024166

Shao, H., Moroni, G., Li, A., Liu, X., & Xu, L. (2020). Simultaneously solving the transfer line balancing and buffer allocation problems with a multi-objective approach. *Journal of Manufacturing Systems, 57*, 254-273. https://doi.org/10.1016/j.jmsy.2020.09.009

Skėrė, S., Žvironienė, A., Juzenas, K., & Petraitienė, S. (2023). Optimisation experiment of production processes using a dynamic decision support method: A solution to complex problems in industrial manufacturing for small- and medium-sized enterprises. *Sensors, 23*(9), Article 4498. https://doi.org/10.3390/s23094498

Turker, A. K., Aktepe, A., Inal, A. F., Ersoz, O. O., Das, G. S., & Birgoren, B. (2019). A decision support system for dynamic job-shop scheduling using real-time data with simulation. *Mathematics, 7*(3), Article 278. https://doi.org/10.3390/math7030278

Unver, B., Kabak, Ö., Topcu, Y. I., Altinisik, A., & Cavusoglu, O. (2020). A decision support system for proactive failure prevention: A case in a leading automotive company. *Journal of Enterprise Information Management, 33*(5), 845-880. https://doi.org/10.1108/JEIM-09-2019-0264

Viriyasitavat, W., Xu, L. D., Bi, Z., & Sapsomboon, A. (2019). New blockchain-based architecture for service interoperations in the Internet of Things. *IEEE Transactions on Computational Social Systems, 6*(4), 739-748. https://doi.org/10.1109/TCSS.2019.2924442

Yang, W., Li, W., Cao, Y., Luo, Y., & He, L. (2020). Real-time production and logistics self-adaptation scheduling based on information entropy theory. *Sensors, 20*(16), Article 4507. https://doi.org/10.3390/s20164507

Pertanika J. Sci. & Technol. 34 (3): 1881 - 1904 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Review Article

Optical Sensing Techniques for Microplastic Identification: Comparative Evaluation, Limitations, and Future Recommendations

Muhammad Nurullah Waliyullah Mohamed Nazli[1], Irneza Ismail[1]*, Fatin Hamimi Mustafa[2], Wan Zakiah Wan Ismail[1], Juliza Jamaludin[1], and Latiffah Karim[3]

[1]Department of Electrical and Electronic Engineering, Faculty of Engineering and Built Environment, Universiti Sains Islam Malaysia, 71800 Nilai, Negeri Sembilan, Malaysia
[2]Department of Electronic and Computer Engineering, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johor Bahru, Johor, Malaysia
[3]Food Biotechnology Programme, Faculty of Science and Technology, Universiti Sains Islam Malaysia, 71800 Nilai, Negeri Sembilan, Malaysia

ABSTRACT

Microplastics (MP) have actively polluted various water sources, including oceans, rivers and lakes. MPs are tiny plastic particles, typically less than 5 millimetres in size. They originate from degraded larger plastics or microbeads in products. It can enter human bodies through ingestion or inhalation. Their small size allows them to penetrate tissues and organs, potentially causing inflammation. In the environment, MP accumulates in soil, water, and air. They can disrupt ecosystems by entering the food chain, affecting aquatic life. The objective of this paper is to provide a comprehensive review of established MP identification methods and to recommend new technologies for future environmental monitoring. The methodology employed a comprehensive literature review and comparative analysis of current MP detection techniques. The non-optical analytical tools, such as Pyrolysis-Gas Chromatography-Mass Spectrometry (Py-GC-MS), Nuclear Magnetic Resonance (NMR), X-ray Diffraction (XRD), Thermogravimetric Analysis (TGA), and Scanning Electron Microscopy (SEM), alongside prominent optical methods, include standard Fourier Transform Infrared (FTIR), Attenuated Total Reflection Fourier-Transform Infrared Spectroscopy (ATR-FTIR), Raman spectroscopy, Nile Red fluorescence, and UV-Visible (UV-Vis) spectroscopy. The results obtained conclude that no single method is

ARTICLE INFO

Article history:
Received: 18 November 2025
Accepted: 05 April 2026
Published: 25 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.25

E-mail addresses:
nurullah.waliyullah95@raudah.usim.edu.my (Muhammad Nurullah Waliyullah Mohamed Nazli)
dr.irneza@usim.edu.my (Irneza Ismail)
fatinhamimi@utm.my (Fatin Hamimi Mustafa)
drwanzakiah@usim.edu.my (Wan Zakiah Wan Ismail)
juliza@usim.edu.my (Juliza Jamaludin)
latiffah16@usim.edu.my (Latiffah Karim)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

perfect, but Micro-Fourier Transform Infrared (μ-FTIR) spectroscopy is highly recommended as the primary optical sensing approach for standardised monitoring due to its optimal balance of speed, reliability, and accessibility. The study also highlights that the integration of Artificial Intelligence (AI) and Machine Learning (ML) is the future of optical MP detection, capable of automating complex spectral matching and noise reduction. Potential applications of the research proposed multi-modal framework can be directly applied to large-scale, cost-effective environmental monitoring, enabling broader geographical coverage and aiding policymaking, particularly in developing nations.

Keywords: Analytical method, artificial intelligence, environmental pollution, microplastics, optical sensing techniques, polymer identification

INTRODUCTION

Plastic is highly valued by both manufacturers and customers due to its versatility, lightweight nature, durability, and affordability. As a result, plastic production has greatly increased (Jung et al., 2023). At present, packaging makes up around 40% of all plastic produced, and most packaging is made to be used just once before being thrown away. Tragically, this increase in plastic use has resulted in a huge rise in the number of plastic products that are accidentally or purposefully lost and end up in the environment. It is estimated that nearly 60% of all produced polymers have gathered in junkyards or the environment (Geyer et al., 2017; Zhao et al., 2023). As plastic can be made in various sizes and shapes, the tiniest sizes (less than 5 mm) of manufactured plastic products are referred to as "microplastics".

MPs are classified into two types, primary MP and secondary MP (Pham et al., 2023). The differences between these classes need to be thoroughly examined to properly tackle the problem of MP pollution (Bermúdez & Swarzenski, 2021). Primary MP are produced in small sizes to serve a functional purpose. They include microbeads that are being added to cosmetic and personal care formulations, plastic pellets that are used in manufacturing processes in various industries, and MP that are added to cleaning products in households (Singh & Mishra, 2023). On the contrary, secondary MPs are created when larger plastic particles break down and disintegrate (Contreras-Llin & Díaz-Cruz, 2024). Plastic bottles, bags and containers that are exposed to environmental stressors such as UV radiation, temperature variations, and physical abrasion eventually diminishes them into smaller fragments over time.

One of the most concerning environmental problems in Malaysia is the high level of MP contamination in the country's aquatic environment, affecting the freshwater and marine ecosystems of the country and has the potential to affect human beings (Landrigan, 2020). The issue has its roots in the intensive use of plastics in different industries and in everyday life, thus contributing to the discharge of MP into rivers, lakes, and coastal zones.

These tiny, barely noticeable plastic pieces may linger in the environment for a long time, endangering aquatic life and making it simple for them to make it into the food chain (Zambrano-Pinto et al., 2024). Environmental exposure causes significant surface degradation and cracking on plastic polymers. Ageing alters spectral signatures, which can complicate accurate optical detection. These structural changes also increase the potential for toxic chemical absorption. Degradation creates microscopic cracks and pores on the plastic surface. Degradation also alters the physicochemical properties of the MP. This expanded surface area allows more pollutants to attach to particles (Ge et al., 2023).

There is empirical data of MP contamination that can be recorded along with the populations of brachyuran crabs that live within the Setiu Wetlands, Malaysia. Environmental effects like UV radiation and tidal forces break macroplastics into fragments. Biological activities and wave action further degrade these plastics into micro-sized particles. Land-derived materials frequently move between locations through daily tides and water currents (Jansen et al., 2024). These natural processes facilitate the heavy accumulation of MP within mangrove areas. A species of crab that is prevalent in the area is exposed to MP via the environment it lives in, as the wetlands are increasingly polluted with plastic trash (Abd Rahim et al., 2023). Brachyuran crabs are filter feeders; therefore, consuming tiny particles, such as MP, as food sources. These ingestions may lead to several harmful outcomes on their health and survival, such as poor feeding efficiency, poor nutrient uptake, and physical damage to their digestive tracts. It is recommended that future research pairs optical detection with controlled laboratory feeding trials to distinguish between incidental ingestion and selective feeding behaviours in Brachyuran crabs. In addition, the build-up of MP in the bodies of crabs may pose a potential threat to further trophic levels, hence influencing the overall ecological integrity of the Setiu Wetlands. The Department of Environment monitors river data to evaluate regional pollution controls. The outcome of the estuary of the Terengganu River is utilised to give guidance on municipal limits on the industrial discharge. Such results affect the ideas of Integrated River Basin Management (IRBM) plans that state governments implement. Certain River Basin Management plans are also required by the European Union Water Framework Directive (WFD). Both models provide systematic avenues where the local communities can mitigate plastic pollution. Localised coordinated control ensures that all the nearby ecosystems have cleaner water. The compliance with such international standards improves the accuracy of environmental protection operations (Lim et al., 2023).

The identification of MP is often based on the various methods of analysis, covered by microscopy, stereoscopy, and Liquid Chromatography-Mass Spectrometry (LCMS). Microscopy, e.g. electron microscopy, allows observation and localisation of MP according to size, shape and colour. The use of stereoscopic microscopy increases the depth perception, which enhances the precise detection and measurement of MPs in the specimens of the environment (Feld et al., 2021; Lee et al., 2023).

LCMS is an effective analytical method that can distinguish and quantify the various forms of MP according to their chemical composition. These techniques are inevitable in the determination of the degree of MP pollution, as well as the development of effective mitigation measures, which will consequently lower the environmental pollution and protect the health of people against possible damage.

Various optical sensing methods are also used in MP analysis, such as UV-Visible (UV-Vis) spectroscopy, Fourier Transform Infrared Spectroscopy (FTIR), Raman spectroscopy, and Nile Red staining. UV-Vis spectroscopy quantifies the absorbance of light by the sample molecules, and thus it is possible to determine the concentration of MP based on their specific absorption properties (Jiao et al., 2022). FTIR spectroscopy determines the chemical composition of MP through the analysis of molecular vibrations of the sample, therefore, providing information about the type of plastic which is present. Like the FTIR, Raman spectroscopy complements chemical composition information, but it functions using different principles, thus increasing the strength of the analysis. Nile red staining is a fluorescence technique that specifically stains MP, making them easily observed under a fluorescent microscope. Figure 1 illustrates the annual publication expenses of research on MP between 2020 and 2024.

The most dominant source in all the years is journal articles. In the year 2020, there were around 1,200 reports of journal publications. This number steadily grew to a high of 2,500 in the year 2023. There is a minor decrease in 2024, and approximately 900 journal publications. There was also an increase in generic publications, which, however, showed a significant growth in 2023. Sections of books and conference proceedings corresponded to similar tendencies, only in much smaller amounts. There were stable outputs in each category, indicating the same level of scholarly interest. The least contribution was given by the book category every year. The highest in 2023 could be the result of the increased global anxiety or additional investments. The decline in 2024 shows the effect of COVID-19 towards academic research. The lockdown in 2020 and 2021 halted the research and publications in 2024 because most organisations and schools closed to encourage everyone to stay home. Only essential emergency services like medical and food supplies remained operational (Rume & Islam, 2020). In addition, complex identification methods also create significant labour bottlenecks for modern researchers. Labour-intensive sampling procedures naturally limit the total volume of published studies. Strict analytical requirements have increased the time needed to complete research (Debraj & Mulky, 2025). Although the increase in publications in this field is declining in 2024, it has considerable scientific and practical relevance. In addition, the current research aims to fill the research gap that has occurred recently and spur further research in the field. This tendency is a sign of good academic activity and peer-reviewed work. These include generic works, probably policy documents or institutional reports. The statistics show growing awareness of MP pollution in the world. The further increase in the number confirms the necessity of environmental urgency of the problem.

Figure 2 illustrates the number of studies utilising optical methods for MPs identification from 2020 to 2024. Primary databases were searched with Figure 2 in the search of Web of Science, Scopus and Google Scholar with an aim of building a strong base for this review. Besides, as inclusion criteria, the studies were included when they presented primary experimental information on the characterisation of MP or when they implied a new analytical workflow. In the event literature was not written in English, the preliminary screening was done on titles and abstracts, then a comprehensive eligibility check of the full text was done.

Research activity gradually increased from 2020 to 2023. In 2020, 17 studies were recorded on this topic. The number rose to 21 in 2021 and 24 in 2022. A peak of 25 studies occurred in 2023. This growth indicates rising interest in optical-based techniques. Optical methods are valued for their non-destructive nature. In 2024, the count dropped significantly to only 10 studies. Although there is a decreasing trend, the importance of MP identification in securing the healthiness of biodiversity (Tan & Mohd Zanuri, 2023). It also suggests the difficulties and complexities of handling, preparing and identification of MP samples (Cutroneo et al., 2020).

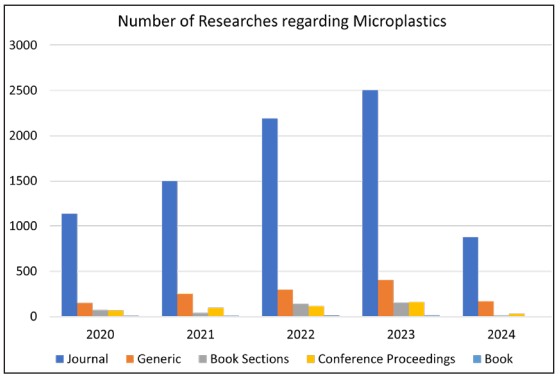


Figure 1. Number of research papers regarding MPs

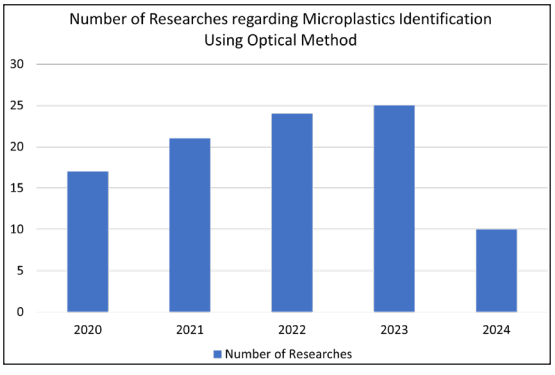


Figure 2. Number of research papers about MPs identification using optical methods

Microplastics

Microplastics (MPs), a term that has gained widespread recognition in recent years, are plastic particles ranging from 5 mm to 25 mm is considered mesoplastics, and the bigger ones (after 25 mm) are macroplastics (Grini et al., 2022; Recabarren et al., 2025). Such minute pollutants have gathered significant interest due to their level of occurrence in many ecosystems. The damage they can cause to wildlife and humans calls for proper management of the environment. The sources, types, distribution and consequences of MPs are important issues that need to be analysed to comprehend the phenomenon.

MPs have many sources, mostly obtained because of the wear of bigger plastic objects like bags, packaging products and bottles. Also, the fibres produced by the textile during the laundry process are synthetic, personal care products contain microbeads, and industrial processes generate them as well, further increasing the number of these particles. When emitted to the environment, they are long-lasting, do not degrade in nature and build up in the environment in oceans and rivers as well as on land.

Figure 3 shows the prevalent types of MP identified by researchers, which are polyethene terephthalate (PET), Polyvinyl chloride (PVC), polypropylene (PP), Polystyrene (PS), HDPE and LDPE. PVC is commonly used in pipes, window fittings and flooring. PS is a common polymer which is frequently used in foam containers.

Figure 4 shows the structure of MPs, where MPs are divided into two components, primary MPs and secondary MPs.

The distribution and breakdown of larger plastic objects are secondary MP. However, primary MP may be deposited directly into the water bodies in case they are carried in the drainage system to the surface water. The deliberate introduction of MP into consumer goods presents a direct danger to aquatic life, as the organisms consume the particles, which can potentially cause trophic interactions to be broken.

Secondary MP are the apportionment and disintegration of bigger plastic objects. Larger plastics, such as bottles, bags, and containers, are exposed to natural pressures like UV radiation, temperature fluctuations, and mechanical abrasiveness, which have an impact on how quickly these plastics break down into smaller pieces over time (Fakhri et al., 2024). Secondary MP tend to be geometrical anomalies and are found in a size range; hence, contributing to a large portion of MP pollution in aquatic and terrestrial environments. They can live decades or even centuries as secondary MPs and present long-term environmental problems.

A variety of MP types were found in the Penang tropical estuarine mangrove ecosystems of Malaysia. Polymer types that were identified were polyacetal (19%), polybutadiene (6%), polychloroprene (CR) (6%), polyethene (PE) (30%), polyformaldehyde (4%), PS (19%), PP (9%), PVC (2%), polyamide 6/6 (PA6) (2%), and polyethene oxide. Sampling was done in coastal surface waters, coastal sediment cores and estuarine sediment layers (Tan & Mohd Zanuri, 2023). Even though the differences in the sampling sites resulted in the density of MP differing between sites, the composition of the MP types was similar across sites.

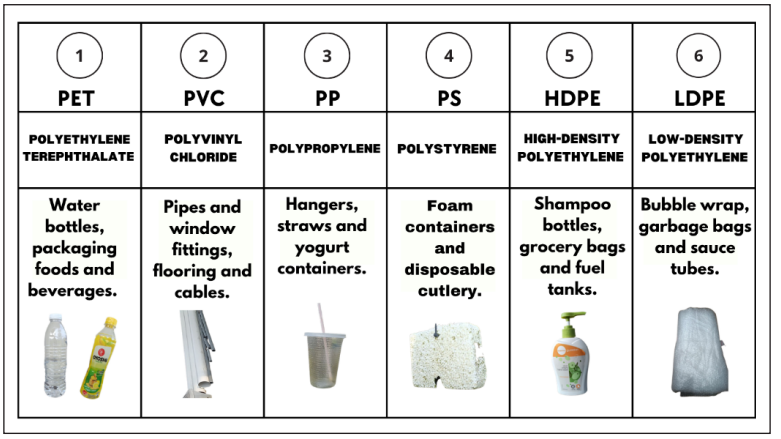


Figure 3. Common types of MPs

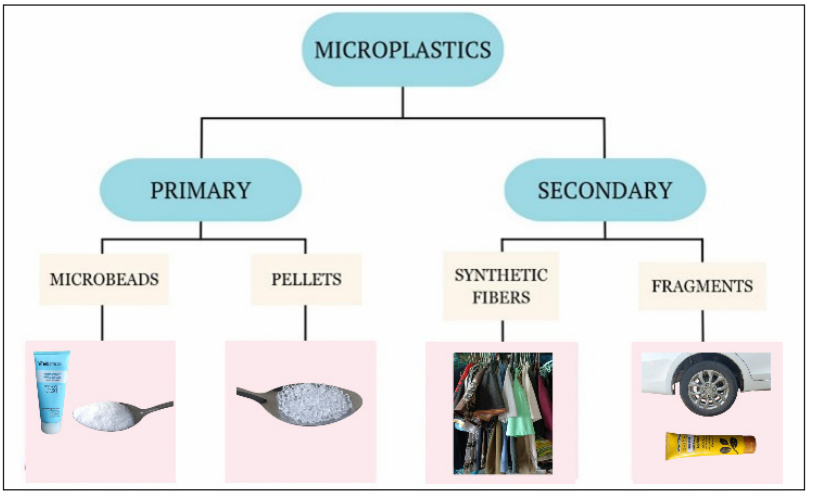


Figure 4. Types of MPs

Present Techniques of Microplastic Detection

The modern MP detection methods are a combination of sophisticated analysis tools. These are Pyrolysis-Gas Chromatography-Mass Spectrometry (Py-GC-MS) that breaks down MP samples to enable chemical interrogation of every element; Nuclear Magnetic Resonance (NMR), which distinguishes the type of MP based on their molecular architecture; X-ray Diffraction (XRD), which identifies and analyses crystalline polymeric structures; Thermogravimetric Analysis (TGA), which measures thermal degradation behaviour; and Scanning Electron Microscopy (SEM), which provides high-resolution images to.

Pyrolysis-Gas Chromatography-Mass Spectrometry (Py-GC-MS)

Pyrolysis-Gas Chromatography-Mass Spectrometry (Py-GC-MS) is an analytical method used to recognise and describe MP. The principal entails exposing an MP sample to high temperatures (pyrolysis), thus breaking it down into smaller pieces (Rosso et al., 2023). Gas chromatography (GC) is used to separate and further analyse the pyrolysis products based on their chemical properties, and mass spectrometry (MS) is used to identify and quantify each component (Kadac-Czapska et al., 2023). Many high-impact studies integrate Py-GC-MS as a confirmatory tool to calibrate optical sensors and minimise false-positive identifications caused by environmental biofilms or organic debris. While optical sensing techniques offer rapid, non-destructive screening, cross-validation remains essential for absolute polymer confirmation. Py-GC-MS allows identification of MP using discrete chemical fingerprints, and has been successful in identifying specific types of MP, such as PVC, PMMA, PP, PS, PE, PET, and PA.

Nuclear Magnetic Resonance (NMR)

The idea underlying the use of Nuclear Magnetic Resonance (NMR) is that certain atomic nuclei have a magnetic moment, which causes them to align with a magnetic field when they are subjected to it. By responding to radiofrequency pulses, the nuclei absorb and consequently release energy, thus giving important information about the molecular environment of the sample (Giaganini et al., 2023). NMR can also be used in the identification and characterisation of different types of polymers in terms of their unique molecular structure within the framework of MP identification. This technique is particularly applicable in the detection of MP, like acrylonitrile-butadiene-styrene MP (ABS-MP), and in such a system, the distinctive polymer chains produce certain NMR spectra.

X-ray Diffraction (XRD)

X-ray diffraction (XRD) is an analytical methodology that is rigorous in nature, based on the interaction of periodic lattice structures of crystalline materials with the collimated X-ray beams. In an XRD experiment, a beam of monochromatic radiation is incident on the specimen, and the resultant diffraction pattern is measured by an array of detectors. The crystal lattice produces constructive and destructive interferences that produce different diffractive peaks that could be quantitatively examined to explain the arrangement of atoms in the sample. When XRD is used to differentiate between crystalline polymers and amorphous ones, it is utilised as part of MP identification, thus improving the specifics of MP shapes.

Examples of representative polymer classes that are routinely interrogated by XRD are PE, high-density polyethene (HDPE), and polyvinyl alcohol (PVA); each of the above has a distinct diffraction signature, which is determined by the crystalline lattice parameters. As a result, XRD is an obligatory instrument for determining MP composition and complements our understanding of their impact on the environment (Dey et al., 2023).

Thermogravimetric Analysis (TGA)

Thermogravimetric analysis (TGA) is one of the thermal analyses, which is used to gauge how a sample's mass changes over a regulated temperature gradient. TGA is used to gauge the thermal degradation characteristic of polymeric particles when it is used to determine MP. In this technique, an MP specimen is subjected to a programmed temperature curve under constant monitoring of its weight loss (Bensharada et al., 2022; Zainuddin & Syuhada, 2020). TGA can be used to identify MPs as the species degrade at discrete degradation rates, where each species has a unique onset temperature at which mass loss initiates. Common polymers that are examined using TGA include PE, PP and PET, and each of them will exhibit a characteristic weight-loss curve that can be utilised to conduct compositional discrimination. The method proves to be, therefore, useful in explaining the thermal stability and the composition of MP in the matrix, thus integrating the strategies to be employed in effective environmental monitoring and mitigation.

Scanning Electron Microscopy (SEM)

Scanning Electron Microscopy (SEM) is an effective imaging technique in the detection of MP in the environmental matrices. SEM is based on a scanning of a narrowly focused electron beam on the specimen surface to excite the emission of secondary electrons, as well as other types of signals that are capable of being processed and yielding images of the surface topography. In MP analyses, the electrons released into the surface of conductively coated surfaces with noble metals like gold or palladium are gathered to form the micrographs. A common set of preparation procedures includes the collection of environmental samples (e.g., aqueous or sedimentary fractions), their filtration and subsequent exposure to dehydration, mounting on stubs, and sputter-coating with a thin conductive layer, usually gold, palladium or carbon, to reduce the charging effects and improve the image resolution (Mateos-Cárdenas et al., 2022). Some of the common polymers studied using SEM are PE, PP, and PET. SEM is the superior method for morphological differentiation. It provides high-resolution imaging of surface topography, allowing researchers to classify primary and secondary plastic through morphological differentiation.

Advantages and Limitations

Nuclear Magnetic Resonance

Nuclear magnetic resonance (NMR) spectroscopy provides useful structural data, which is useful in MP analysis. However, its extensive use is limited by the fact that it requires advanced instrumentation and highly trained people, which can be a hindrance during regular utilisation. The separation of polymeric compounds or the identification of trace amounts of MP may be difficult, which may undermine the quality of analysis. Furthermore, the process of NMR is laborious and time-consuming, as it requires a lot of sample preparation and complicated interpretation of data. The cost of capital involved in the NMR instrumentation also limits its availability in the routine MP characterisation. Spectral data are sometimes subject to fluctuating data, which may require additional extraneous analysis to eliminate inconsistencies, thus compromising the reliability of the results (Seghers et al., 2023).

Nile Red Fluorescence

The detection of MP is commonly done using Nile red fluorescence, but its practical implementation is limited by the complexity of the sample preparation and the interference of the matrix. Environmental samples may have heterogeneous components that may undermine the specificity of Nile red stain, giving false positives or negatives. The method is also quite expensive, requiring the use of special fluorometric equipment and reagents (Mazlan et al., 2022). Moreover, it is time-consuming because it is a series of steps- sample isolation, staining, incubation and microscopic examination. Such accruing limitations reduce the effectiveness and practicability of Nile red fluorescence for use in a complicated environmental matrix.

Thermogravimetric Analysis (TGA)

Thermogravimetric analysis (TGA) is a type of thermal analysis which is used to measure changes in the mass of a sample during a controlled temperature gradient. TGA is used to gauge the thermal degradation characteristic of polymeric particles when it is used to determine MP. In this technique, an MP specimen is subjected to a programmed temperature curve under constant monitoring of its weight loss (Zainuddin & Syuhada, 2020). TGA can be used to identify MPs as the species degrade at discrete degradation rates, where each species has a unique onset temperature at which mass loss initiates. Common polymers that are examined using TGA include PE, PP and PET, and each of them will exhibit a characteristic weight-loss curve that can be utilised to conduct compositional discrimination. The method proves to be, therefore, useful in explaining the thermal stability and the composition of MP in the matrix, thus integrating the strategies to be employed in effective environmental monitoring and mitigation.

Potential of Optical Sensing Technique

One of the main electromagnetic radiation categories that is frequently employed in modern scientific research is optical light, which consists of ultraviolet (UV), visible, and infrared light. UV radiation has shorter wavelengths and thus, contains more photon energy as compared to visible light. The visible light, which is characterised by the wavelengths that can be seen by human eyes, is widely used in microscopy, spectroscopic analysis, and imaging processes. Infrared radiation, characterised by wavelengths longer than those of visible light, has diverse applications including remote sensing, thermography, and spectroscopic study of chemical compositions. The ability of optical light to couple with matter in unique ways is the foundation of its central place in scientific discovery, which allows a very broad range of methodological strategies in a broad range of fields. The current areas of research work still believe in the optical light method in control as a potential area of future research since it is relatively simplified, faster and less intrusive than other more conventional methods.

Fourier-Transform Infrared Spectroscopy (FTIR)

Fourier-Transform Infrared Spectroscopy (FTIR) falls under the category of the most widespread analytical procedures used to determine the presence of MP. This method is achieved by separating aqueous samples with fine mesh sieves or filters of carefully determined pore sizes, thus trapping MP with water and other extraneous contents between the sieves. The retained particles are then irradiated with infrared light to give transmitted spectra, which are then obtained with a detector. The spectral data obtained are computed using the Fourier-transform analysis to identify the identity of the polymer shown in Figure 5. The FTIR method is a fast method of data collection, compared to other traditional methods, because it can scan a wide spectral range in minutes. The technique is non-destructive and leaves samples intact, which is a beneficial characteristic, especially when analysing rare or precious materials. To improve this, the samples need to be dried to reduce water interference (Fomina et al., 2023). Furthermore, for complicated substances, FTIR can only offer a limited amount of information about the whole chemical structure. To get more thorough molecular information, include FTIR with additional methods like mass spectrometry, NMR, or Raman Spectroscopy. Figure 5 shows the working principle of FTIR.

An example result of the FTIR spectrum of PVC is referred from Yaseen et al. (2021). The analysis uses infrared light to monitor how Poly (vinyl chloride) breaks down over time. They observed the growth of some chemical indicators, like carbon and carbonyl double bonds. These are definite indications of structural degradation of the plastic subjected to UV conditions. The experiment shows that new tin complexes can be used to delay degradation by comparing the heights of the peaks to compare them. All spectral measurements were referred to the constant carbon-hydrogen bond.

The highest peak is created at 1722 cm-1, and this shows the presence of carbonyl (C=O) groups. The other important peak is at 1602 cm-1, and it shows the formation of carbon-carbon double bonds (C=C) (Wu et al., 2024). The two peaks are on the rise as the polymer chain breaks and loses the hydrogen chloride in the presence of UV. Another peak that the researchers observed was the 1328 cm-1 peak, which is attributed to C-H bonds in the PVC (Ahmed et al., 2021). This C-H peak was not affected, and therefore it was used as an internal standard to measure degradation indices. With the quantification of these precise spectral variations, the study gives a mathematical demonstration of the efficiency of the stabiliser.

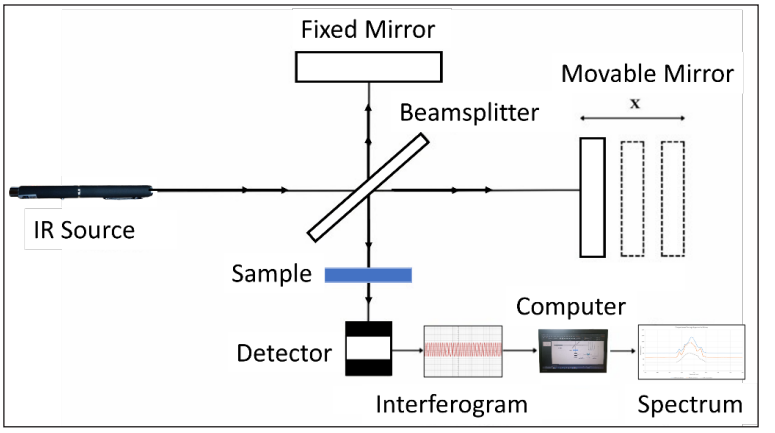


Figure 5. Principle of FTIR

ATR-FTIR (Attenuated Total Reflection Fourier-Transform Infrared Spectroscopy)

Attenuated Total Reflection Fourier-Transform Infrared Spectrophotometry (ATR-FTIR) is also one of the main methods of MP analysis. In this configuration, the beam of infrared is passed through a high-index crystal, which comes in direct contact with the MP surface. The resulting evanescent wave allows one to measure surface absorption, which gives an idea of functional groups existing within the material and thus helps in characterising and identifying it. In contrast, the standard FTIR quantifies the transmission or reflection of infrared light of a specimen. ATR-FTIR enables faster, direct surface probing compared to FTIR. ATR-FTIR can be used for a wide range of sample types, such as gels, films, solids, liquids, powders, and surface analyses, and has increased sensitivity to trace elements, able to detect small amounts of material (Van Haaren et al., 2023). Though it is a widely used and adaptable method, ATR-FTIR has several flaws. ATR-FTIR is quite similar to diamonds in terms of their high index of refraction. The square root term is negative, the ATR impact is absent, and the ratio of indices is close to one if the sample index and crystal index are too identical.

Moreover, compared to conventional transmission FTIR, the ATR approach could have a smaller spectrum range, especially in the low wavenumber range. To cover the entire spectral range, the ATR-FTIR method can be combined with additional IR approaches, if required. Table 1 shows the comparison table between FTIR and ATR-FTIR.

Table 1 shows a comparison table between standard FTIR and ATR-FTIR. In environmental monitoring, the shift from standard FTIR to ATR-FTIR is driven by the need for speed and the ability to analyse particles recovered from the field. Although standard FTIR remains the scientific foundation, the hardware configuration of ATR-FTIR makes it a more practical choice for rapid MP identification.

The samples collected in fields are mixed with types of MP. These MPs can be isolated depending on the type of MP using the method of the Sequential Density Separation (SDS). Sequential density separation is an efficient technique to isolate all forms of MP in sample water. The density of the liquid is gradually increased to suspend some plastic polymers. Low-density plastics such as PE will first be collected on the surface. Adding salt solutions in turn causes the heavier materials, like PS, to rise. It is a methodical method that could be employed to classify various forms of polymers in a proper order. The former advantage is that high-purity plastic fractions can be collected. The process is cost-effective and has high research analytical integrity. It can manage complicated environmental blends without destroying the polymer structures (Rivoira et al., 2020).

Table 1
Comparison between standard FTIR and ATR-FTIR

Feature	Standard FTIR (Transmission)	ATR-FTIR
Physical Mechanism	The IR beam passes directly through the entire sample.	An IR beam reflects inside a crystal, creating an evanescent wave.
Sampling Depth	Probes the entire bulk (thickness) of the sample.	Probes only the surface layer (0.5 to 5 um deep).
Sample Preparation	Requires grinding samples into thin KBr pellets.	Samples are pressed directly onto the crystal surface.
Light Path	The sample must be transparent or very thin to let light pass.	Light never travels through the sample; it only touches the surface.
Data Processing	Produces a "standard" spectrum immediately.	Requires an ATR Correction algorithm to adjust peak intensities.
Key Advantage	Good for bulk chemical analysis of pure materials.	Superior for weathered surfaces and irregular environmental shapes.

Raman Spectroscopy

While different from traditional spectroscopy, Raman spectroscopy is a valuable tool for MP identification. It provides molecular information about the material being analysed. A unique Raman spectrum is produced when a laser beam interacts with an MP particle, which enables the kind of plastic polymer to be identified. The benefits of Raman spectroscopy include high sensitivity and quick result acquisition, making it most appropriate in polymer differentiation and investigation of MPs as small as submicron sizes (Pashaei et al., 2023). Also, Raman spectroscopy can be used to characterise polymers by interrogating specific vibrational modes of the molecular structure. Raman and u-FTIR are excellent in distinguishing primary and secondary MP by their chemical markers and additive detection, as they can detect "oxidative weathering" signatures (Su et al., 2024). Raman instrumentation (portable) can be used to do field analysis with a variety of applications. The method has lower sensitivity to water, thus increasing its aptitude for the analysis of biological tissues and water bodies.

However, there are several drawbacks and difficulties. Raman scattering is generally a very weak process, where only about one in every 10^8 photons undergoes spontaneous scattering. The strength of the achievable Raman signal is constrained by this intrinsic weakness. The sample may get heated when high-intensity laser sources are used. This might result in thermal progression or change the sample's characteristics. For delicate or thermally unstable materials, this is specifically cumbersome. Raman spectroscopy is more sensitive to the analysis of metallic (metal-containing materials or metal oxides such as iron oxide and titanium oxide) and other material components. But pure metals such as pure gold, pure silver, pure aluminium and pure copper normally show no observable Raman lines since their lattice vibrations do not significantly change polarisability. To address these drawbacks, several new sophisticated methodologies have been devised. An example is the Surface Enhanced Raman Spectroscopy (SERS), which can reduce by many orders the interference of fluorescence by enhancing the Raman signal by metallic nanostructures several times. This method also effectively masks organic fluorescence without using harsh chemical treatments.

Nile Red Fluorescence

The effect of fluorogenic dye Nile Red assumes that fluorogenic dyes, including Nile Red, are strongly emitted when interacting with nonpolar components, in particular, the lipidic components of MP (MP). Under this protocol, Nile Red is incubated with MP samples, where it is allowed to adsorb on the hydrophobic surfaces of the plastic particles. The stained MPs can then be visualised and determined in terms of fluorescence after subsequent illumination at certain excitation wavelengths, which can be achieved using a fluorescence microscope. The method is especially effective in the identification of MPs in environmental samples.

Polymers regularly recognised by the Nile Red fluorescence assay include polylactic acid (PLA), PVC, expanded polystyrene (EPS), two versions of HDPE, polyamide-6 (PA6), PE, PET, three types of PP, and two types of PS; these polymers are hydrophobic polymers that are vital in the flu. Nile Red has not gone with use without its limitations. The fluorescence sensitivity of the dye is highly affected by the polarity of the surrounding medium, which may result in great spectral changes and attenuation, complicating the quantitative measurements. Moreover, Nile Red is also photobleached, which is a slow process of the loss of fluorescence due to a long-lasting impact with excitation light (Sha et al., 2023). By using shorter exposure periods and lower light intensity settings to minimise the amount of time and intensity of light exposure during imaging.

UV-Vis Spectroscopy

Ultraviolet-Visible (UV-Vis) spectroscopy is used to identify the absorption of photons in the ultraviolet and visible fields by new species. As light of distinct wavelengths passes through a sample, characteristic absorption bands form, and there is attenuation of the intensity of light passing through at those wavelengths. Within the framework of MP detection, the UV-Vis spectroscopy can be utilised to characterise the absorption profiles of the polymers that are commonly used in MP. Spectral range 200-780nm on a v-650 spectrometer (Jasco) with a spectral bandwidth of 0.5nm and a spectral resolution of 2nm. Different types of MP, such as low-density polyethene (LDPE), HDPE, PE, PP, and PET, PVC, exhibit distinct absorption spectra in the UV-Vis range (200nm to 780nm). By comparing the absorption patterns of unknown samples to reference spectra of known polymers, scientists can identify and quantify the presence of specific MP in environmental samples using UV-Vis spectroscopy. UV-Vis spectroscopy is a common method in various fields and is widely used for both qualitative and quantitative analysis in laboratories. Due to their broad dynamic range, UV-Vis spectrometers can calculate absorbance across several orders of magnitude (Fakayode et al., 2024). Therefore, samples with different concentrations can be analysed (Paul et al., 2024). UV-Vis spectroscopy is non-invasive towards the sample, hence enabling the sample to be repeatedly examined. Despite this method being commonly used, it also has several issues. Due to UV-Vis spectroscopy having limited sensitivity and detection limitations, it may not be able to detect very low quantities of analytes. To overcome this difficulty, cuvettes with longer path length can be used as they can contain more analytes. The analyte's concentration can also be raised by pre-concentrating materials by methods like solid-phase extraction or evaporation. Moreover, baseline drift, particle scattering, and turbidity in the sample may cause inaccurate absorbance readings. To solve this problem, adjust for baseline drift by using reference samples or baseline correction methods.

To minimise the scattering effects and eliminate particles, either filter or clear the samples. Advanced calibration methods, including partial least squares regression (PLSR), allow the system to adjust for matrix effects and background interference (Nnachi, 2024). Optical methods have been highly used in research as it is non-invasive, rapid, accurate and many more. Optical light methods are trending now as it complies with future characteristics. Figure 6 illustrates the principle of working of UV Vis spectroscopy.

For instance, Klempová et al. (2023) used UV-VIS on PVC to study the effect of degradation on the absorbance. The UV–VIS spectrum of PVC shows characteristic absorption in the near-visible region. A distinct absorbance band appears around 390 nm due to conjugated polyene formation. This peak indicates progressive degradation of PVC under thermal and UV exposure. Increased absorbance intensity reflects the growth of chromophoric sequences within the polymer matrix. These conjugated structures arise from dehydrochlorination reactions during ageing processes (Liu et al., 2026). The spectral shift toward longer wavelengths suggests an extended conjugation length (Bononi et al., 2020). UV–VIS analysis provides a sensitive approach for monitoring optical and chemical changes. This method enables non-destructive evaluation of PVC degradation behaviour.

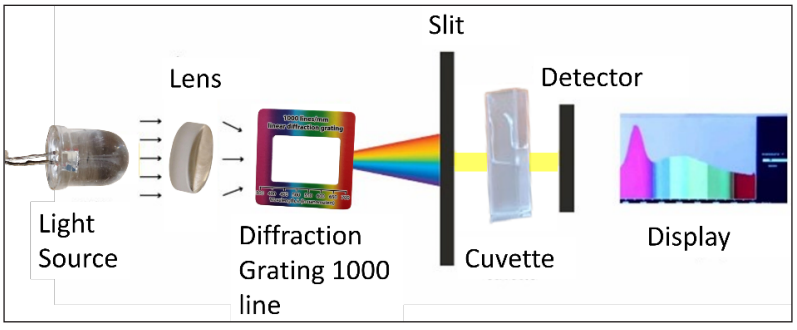


Figure 6. Principle of UV-Vis

Comparison Between Optical Sensing Techniques

Table 2 shows a comparison table of the top common optical sensing techniques for MP identification. No single method fits every research budget or goal perfectly. High-accuracy tools like Raman spectroscopy detect the smallest particles. However, advanced equipment often involves much higher financial costs. Nile Red staining offers a faster, cheaper way to count particles (Shruti et al., 2022). Standard monitoring typically relies on the chemical reliability of FTIR. Researchers must balance precision against available time and resources.

Table 2
Comparative table of optical sensing techniques

Method	Sensitivity (Min. Size)	Specificity (Polymer ID)	Relative Cost	Throughput (Samples/Day)
μ-FTIR	High (~5~10 μm) (0.1~1.0 ppm)	Excellent	Moderate to High	Low (1~5 samples)
Raman Spectroscopy	Very High (<1 μm) (< 0.1 ppm)	Excellent	High	Very Low (1~2 samples)
Nile Red Staining	Moderate (~100 μm) (1.0 – 10 ppm)	Low (non-specific)	Low	High (20~50 samples)
UV-Vis Spectroscopy	Low (5.0 – 50 ppm)	Moderate	Low	Very High (>50 samples)
Fluorescence Microscopy	Moderate (0.5 -5.0 ppm)	Moderate	Moderate	Moderate (10~20 samples)

Recommended Optical Approach for Future Monitoring

Despite the abundance of existing analytical methods, μ-FTIR spectroscopy is recommended as the primary optical method to employ for MP monitoring. Despite other techniques like Raman spectroscopy being in a better position to deliver the optimum spatial resolution of sub-micron-sized particles, μ-FTIR offers the most suitable trade-off between speed and reliability, as well as the availability to environmental researchers. Being sensitive to 5 um to 10 μm, μ-FTIR is a promising technique to detect MP accurately (Olivatto et al., 2024). The high-throughput strategies to analyse complex organic substrates present in tropical ecosystems, such as the Malaysian mangroves, have a huge knowledge gap (Tan & Mohd Zanuri, 2023). The application of μ-FTIR as a reference point has enabled scientists to employ large polymer spectral libraries to obtain high-confidence identification, which can be a bottleneck in large-scale environmental measures. Moreover, external contamination should also be detected with the help of blank filters in the case of MP analysis. Researchers subtract these background counts to have tremendous accuracy in the data. Such a stringent procedure ensures that the findings of environmental research have analytical integrity. The blanks used are also efficient in preventing excess assertion of the plastic levels of ocean samples. This might contribute to more laborious and time-consuming procedures as more procedures and protocols must be followed. One of the examples recommended is by obligating the use of blank filters to increase the level of certainty in the analytical results. This underscores the essential care every researcher must take during their routine procedures.

Careful contamination management is vital for maintaining the integrity of all scientific findings. Precise results depend entirely on avoiding external plastic fibres during the sample processing. The implementation of the Airborne Contamination Control (ACC) Protocol significantly mitigated airborne microplastic (MP) contamination, thereby increasing the confidence that identified particles originated solely from the samples. These findings underscore the necessity of rigorous laboratory hygiene and demonstrate that accessible, low-cost interventions can drastically enhance the reliability and data quality of future environmental and biological MP assessments (Paiva et al., 2022). Hydrogen Peroxide must also be used for oxidative digestion as it is one of the most common and effective methods for removing natural organic matter (NOM), such as microorganisms, algae, and biofilms, from environmental samples. This is needed in the case of Nile Red, because otherwise, this dye will stain organic debris and produce false results (Castillo & Dumale, 2024).

Moreover, the use of a multi-modal optical sensing structure is proposed in future research to solve the weaknesses of the individual sensors. This approach suggests that Nile Red fluorescence staining must be performed as a low-cost and rapid screening protocol to determine the total number of particles in the field in a short time, and μ-FTIR to profile a representative subset chemically (Corami et al., 2021). The hybrid method is the answer to the essential issue of the cost-effective monitoring of the developing regions and the necessity to maintain the high scientific standards of the process of policymaking. This type of monitoring system would allow one to monitor greater geographical regions without the extreme costs of the expensive spectroscopic imaging of individual samples. Finally, the optical sensing MP detection will be capable of integrating with the existing optical hardware and combining it with Machine Learning (ML) and Artificial Intelligence (AI) to apply more functions (Kalatzis et al., 2025). The weakness in the data processing of data, i.e., the time-intensive manual interpretation of spectra, remains a severe obstructive factor to real-time monitoring. Optical sensing accuracy can be enhanced significantly by developing AI-based automated spectral matching and noise-removal algorithms that will specifically work with environmentally damaged plastics (Cowger et al., 2025; Jin et al., 2024; Mikmekova et al., 2024). The technological solution will transform the MP detection technology, which is a tedious and lab-intensive process, into a high-impact and streamlined environmental monitoring tool on a global scale. The implementation of AI in the process of MP identification will form another system of decision-making. To conclude, standardised u-FTIR is the most plausible baseline that can be used to map global MP pollution. Intimate measures, blank filters and oxidative digestion guarantee high levels of analysis. Hybrid Nile Red systems with hydrogen peroxide and spectroscopy offer a rather cheap monitoring system. Additionally, spectral analysis that is time-consuming will eventually be automated with the addition of Artificial Intelligence. Laboratory research becomes a useful tool of environmental policy because of this development. Evidence-based choices will eventually save our delicate ecosystems for the coming generations.

CONCLUSION

Lastly, the high pollution of water bodies with MP has attracted numerous studies that aim at defining and identifying and defining them. Scientists use different methodologies, which can be broadly divided into non-optical and optical, each of which have their drawbacks and merits. The examined methodologies provide priceless information about MP properties, which improves our knowledge regarding their interactions with the environment and biological outcomes. The optical methods that are explained to have little perturbation to the sample will be proven as future-proof against the future need of analysing the sample, and the future techniques will need improvement to combat the negative impact of MP on aquatic life. The recommendations for the future are aimed at the construction of a high-impact, multi-modal model, comprising low-cost screening and highly developed automation. Nile Red staining is proposed by scientists because of its quick and accurate counts of particles in the field. Standardised 3mu-FTIR spectroscopy should then be used to determine the chemical characterisation of representative subsets in detail. Most importantly, the data processing will transform due to the AI and ML integration process. Automated algorithms can be used to spectral match and remove noise on plastics which have been environmentally degraded. This advancement alters the detection techniques of MP that depend on the tedious effort in the laboratory to real-time global monitoring. This type of streamlined approach will give the sound information required in making effective environmental policies.

ACKNOWLEDGEMENT

We would like to acknowledge the Ministry of Higher Education, Malaysia, for the Fundamental of Research Grant Scheme, FRGS (FRGS/1/2023/WAB02/USIM/02/1) (USIM/FRGS/FKAB/KPT/50123) and the Faculty of Engineering and Built Environment, USIM, for the funding and support.

REFERENCES

Abd Rahim, N. H., Cannicci, S., Ibrahim, Y. S., Not, C., Idris, I., Mohd Jani, J., Dahdouh-Guebas, F., & Satyanarayana, B. (2023). Commercially important mangrove crabs are more susceptible to microplastic contamination than other brachyuran species. *Science of the Total Environment, 903*, Article 166271. https://doi.org/10.1016/j.scitotenv.2023.166271

Ahmed, A., Abdallh, M., Al-Mashhadani, M. H., Ahmed, D. S., Bufaroosha, M., Jawad, A. H., & Yousif, E. (2021). Environmental stability of poly(vinyl chloride) modified by Schiff's base under exposure to UV. *Biointerface Research in Applied Chemistry, 11*(5), 13465-13473. https://doi.org/10.33263/BRIAC115.1346513473

Bensharada, M., Telford, R., Stern, B., & Gaffney, V. (2022). Loss on ignition versus thermogravimetric analysis: A comparative study to determine organic matter and carbonate content in sediments. *Journal of Paleolimnology, 67*(2), 191-197. https://doi.org/10.1007/s10933-021-00209-6

Bermúdez, J. R., & Swarzenski, P. W. (2021). A microplastic size classification scheme aligned with universal plankton survey methods. *MethodsX, 8*, Article 101516. https://doi.org/10.1016/j.mex.2021.101516

Bononi, F. C., Chen, Z., Rocca, D., Andreussi, O., Hullar, T., Anastasio, C., & Donadio, D. (2020). Bathochromic shift in the UV–visible absorption spectra of phenols at ice surfaces: Insights from first-principles calculations. *The Journal of Physical Chemistry A, 124*(44), 9288-9298. https://doi.org/10.1021/acs.jpca.0c07038

Castillo, R. T., & Dumale, J. (2024). Rapid detection of microplastic contamination using Nile red fluorescent tagging. *International Journal of Biosciences, 24*(2), 103-109. https://doi.org/10.12692/ijb/24.2.103-109

Contreras-Llin, A., & Díaz-Cruz, M. S. (2024). Microplastic removal in managed aquifer recharge using wastewater effluent. *Environmental Pollution, 342*, Article 122967. https://doi.org/10.1016/j.envpol.2023.122967

Corami, F., Rosso, B., Morabito, E., Rensi, V., Gambaro, A., & Barbante, C. (2021). Small microplastics (<100 μm), plasticisers, and additives in seawater and sediments: Oleo-extraction, purification, quantification, and polymer characterisation using micro-FTIR. *Science of the Total Environment, 797*, Article 148937. https://doi.org/10.1016/j.scitotenv.2021.148937

Cowger, W., Karapetrova, A., Lincoln, C., Chamas, A., Sherrod, H., Leong, N., Lasdin, K. S., Knauss, C., Teofilović, V., Arienzo, M. M., Steinmetz, Z., Primpke, S., Darjany, L., Murphy-Hagan, C., Moore, S., Moore, C., Lattin, G., Gray, A., Kozloski, R., … Maurer, B. (2025). Open Specy 1.0: Automated (hyper) spectroscopy for microplastics. *Analytical Chemistry, 97*(32), 17345-17356. https://doi.org/10.1021/acs.analchem.5c00962

Cutroneo, L., Reboa, A., Besio, G., Borgogno, F., Canesi, L., Canuto, S., Dara, M., Enrile, F., Forioso, I., Greco, G., Lenoble, V., Malatesta, A., Mounier, S., Petrillo, M., Rovetta, R., Stocchino, A., Tesan, J., Vagge, G., & Capello, M. (2020). Microplastics in seawater: Sampling strategies, laboratory methodologies, and identification techniques applied to the port environment. *Environmental Science and Pollution Research, 27*(9), 8938-8952. https://doi.org/10.1007/s11356-020-07783-8

Debraj, D., & Mulky, L. (2025). Convention and beyond: An insight into current methods of microplastic detection. *Environmental Technology Reviews, 14*(1), 259-272. https://doi.org/10.1080/21622515.2025.2475258

Dey, T. K., Jamal, M., & Uddin, M. E. (2023). Fabrication and performance analysis of graphene oxide-based composite membrane to separate microplastics from synthetic wastewater. *Journal of Water Process Engineering, 52*, Article 103554. https://doi.org/10.1016/j.jwpe.2023.103554

Fakayode, S. O., Mehari, T. F., Fernand, V. E., Grant, C., Taylor, M. E., Baker, G. A., Siraj, N., Bashiru, M., Denmark, I., Oyebade, A., Bwambok, D. K., Kuedukey, C., Alonge, T., & Anum, D. (2024). Microplastics: Challenges, toxicity, spectroscopic and real-time detection methods. *Applied Spectroscopy Reviews, 59*, 1183-1277. https://doi.org/10.1080/05704928.2024.2311130

Fakhri, V., Hamzehlouy, A., Janmaleki Dehchani, A., Moradi, E., Tavakoli Dare, M., Jafari, A., & Khonakdar, H. A. (2024). Green solutions for blue waters: Using biomaterials to purify water from microplastics and nanoplastics. *Journal of Water Process Engineering, 65*, Article 105854. https://doi.org/10.1016/j.jwpe.2024.105854

Feld, L., da Silva, V. H., Murphy, F., Hartmann, N. B., & Strand, J. (2021). A study of microplastic particles in Danish tap water. *Water, 13*(15), Article 2097. https://doi.org/10.3390/w13152097

Fomina, P. S., Proskurnin, M. A., Mizaikoff, B., & Volkov, D. S. (2023). Infrared spectroscopy in aqueous solutions: Capabilities and challenges. *Critical Reviews in Analytical Chemistry, 53*(8), 1748-1765. https://doi.org/10.1080/10408347.2022.2041390

Ge, J., Wang, M., Liu, P., Zhang, Z., Peng, J., & Guo, X. (2023). A systematic review on the ageing of microplastics and the effects of typical factors in various environmental media. *TrAC Trends in Analytical Chemistry, 162*, Article 117025. https://doi.org/10.1016/j.trac.2023.117025

Giaganini, G., Cifelli, M., Biagini, D., Ghimenti, S., Corti, A., Castelvetro, V., Domenici, V., & Lomonaco, T. (2023). Multi-analytical approach to characterise the degradation of different types of microplastics: Identification and quantification of released organic compounds. *Molecules, 28*(3), Article 1382. https://doi.org/10.3390/molecules28031382

Grini, H., Metallaoui, S., González-Fernández, D., & Bensouilah, M. (2022). First evidence of plastic pollution in beach sediments of the Skikda coast (northeast Algeria). *Marine Pollution Bulletin, 181*, Article 113831. https://doi.org/10.1016/j.marpolbul.2022.113831

Jansen, M. A. K., Andrady, A. L., Bornman, J. F., Aucamp, P. J., Bais, A. F., Banaszak, A. T., Barnes, P. W., Bernhard, G. H., Bruckman, L. S., Busquets, R., Häder, D.-P., Hanson, M. L., Heikkilä, A. M., Hylander, S., Lucas, R. M., Mackenzie, R., Madronich, S., Neale, P. J., Neale, R. E., … Robinson, S. A. (2024). Plastics in the environment in the context of UV radiation, climate change, and the Montreal Protocol: UNEP Environmental Effects Assessment Panel, update 2023. *Photochemical & Photobiological Sciences, 23*(4), 629-650. https://doi.org/10.1007/s43630-024-00552-3

Jiao, M., Wang, Y., Li, T., Li, R., & Liu, B. (2022). Riverine microplastics derived from mulch film in Hainan Island: Occurrence, source, and fate. *Environmental Pollution, 312*, Article 120093. https://doi.org/10.1016/j.envpol.2022.120093

Jin, H., Kong, F., Li, X., & Shen, J. (2024). Artificial intelligence in microplastic detection and pollution control. *Environmental Research, 262*, Article 119812. https://doi.org/10.1016/j.envres.2024.119812

Jung, H., Shin, G., Kwak, H., Hao, L. T., Jegal, J., Kim, H. J., Jeon, H., Park, J., & Oh, D. X. (2023). Review of polymer technologies for improving the recycling and upcycling efficiency of plastic waste. *Chemosphere, 320*, Article 138089. https://doi.org/10.1016/j.chemosphere.2023.138089

Kadac-Czapska, K., Jutrzenka-Trzebiatowska, P., Knez, E., Zaleska-Medynska, A., & Grembecka, M. (2023). Microplastics in food: A critical approach to definition, sample preparation, and characterisation. *Food Chemistry, 418*, Article 135985. https://doi.org/10.1016/j.foodchem.2023.135985

Kalatzis, D., Katsafadou, A. I., Katsarou, E. I., Chatzopoulos, D. C., & Kiouvrekis, Y. (2025). Explainable artificial intelligence for the rapid identification and characterisation of ocean microplastics. *Microplastics, 4*(3), Article 51. https://doi.org/10.3390/microplastics4030051

Klempová, S., Oravec, M., & Vizárová, K. (2023). Analysis of thermally and UV–Vis aged plasticised PVC using UV-Vis, ATR-FTIR, and Raman spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 293*, Article 122541. https://doi.org/10.1016/j.saa.2023.122541

Landrigan, P. J., Stegeman, J. J., Fleming, L. E., Allemand, D., Anderson, D. M., Backer, L. C., Brucker-Davis, F., Chevalier, N., Corra, L., Czerucka, D., Bottein, M.-Y. D., Demeneix, B., Depledge, M., Deheyn, D. D., Dorman, C. J., Fénichel, P., Fisher, S., Gaill, F., Galgani, F., … Rampal, P. (2020). Human health and ocean pollution. *Annals of Global Health, 86*(1), Article 151. https://doi.org/10.5334/aogh.2831

Lee, S.-Y., An, J., & Kwon, J.-H. (2023). Sequential quantification of the number and mass of microplastics in municipal wastewater using Fourier-transform infrared spectroscopy and pyrolysis-gas chromatography–mass spectrometry. *Environmental Pollution, 336*, Article 122452. https://doi.org/10.1016/j.envpol.2023.122452

Lim, F.-H., Lee, W.-K., Khor, W.-S., Chan, L., Sulaiman, M. S., Shaari, H., & Jalil, M. K. (2023). River basin change index for integrated river basin management of the Langat River, Malaysia. *Water Policy, 25*(11), 1035-1061. https://doi.org/10.2166/wp.2023.035

Liu, L., Zhang, M., Hao, H., You, F., Zhou, D., Duan, J., Li, S., & Jiang, X. (2026). Efficient enhancement of UV ageing resistance in polyvinyl chloride resin using carbon dots. *Polymer Degradation and Stability, 244*, Article 111816. https://doi.org/10.1016/j.polymdegradstab.2025.111816

Mateos-Cárdenas, A., von der G. Moroney, A., van Pelt, F. N. A. M., O'Halloran, J., & Jansen, M. A. K. (2022). Trophic transfer of microplastics in a model freshwater microcosm: Lack of a consumer avoidance response. *Food Webs, 31*, Article e00228. https://doi.org/10.1016/j.fooweb.2022.e00228

Mazlan, N. A., Lin, L., & Park, H. E. (2022). Microplastics in the New Zealand environment. *Processes, 10*(2), Article 265. https://doi.org/10.3390/pr10020265

Nnachi, A. B. (2024). Advances in UV spectroscopy for monitoring the environment. *Research Invention Journal of Biological and Applied Sciences, 3*(1), 82-89. https://www.researchgate.net/publication/381650499_Advances_in_UV_Spectroscopy_for_Monitoring_the_Environment

Olivatto, G. P., Ando, R. A., Fernandes, R. F., de Moraes, N. G., & Tornisielo, V. L. (2024). A critical comparison of the main characterisation techniques for microplastics identification in an accelerated ageing laboratory experiment. *Water Emerging Contaminants & Nanoplastics, 3*, Article 7. https://doi.org/10.20517/wecn.2023.69

Paiva, B. O., de Souza, A. K. M., Soares, P. L., Palma, A. R. T., & Vendel, A. L. (2022). How to control airborne contamination in laboratory analyses of microplastics. *Brazilian Archives of Biology and Technology, 65*, Article e22210399. https://doi.org/10.1590/1678-4324-2022210399

Pashaei, R., Sabaliauskaitė, V., Suzdalev, S., Balčiūnas, A., Putna-Nimane, I., Rees, R. M., & Dzingelevičienė, R. (2023). Assessing the occurrence and distribution of microplastics in surface freshwater and wastewaters of Latvia and Lithuania. *Toxics, 11*(4), Article 292. https://doi.org/10.3390/toxics11040292

Paul, A., Reese, M., Goldhammer, T., Schmalsch, C., Weber, J., & Bannick, C. G. (2024). Spectroscopic evidence for adsorption of natural organic matter on microplastics. *Applied Research, 3*(2), Article e202200126. https://doi.org/10.1002/appl.202200126

Pham, D. T., Kim, J., Lee, S.-H., Kim, J., Kim, D., Hong, S., Jung, J., & Kwon, J.-H. (2023). Analysis of microplastics in various foods and assessment of aggregate human exposure via food consumption in Korea. *Environmental Pollution, 322*, Article 121153. https://doi.org/10.1016/j.envpol.2023.121153

Recabarren, T., Torres, M., Jacobsen, C., Pozo, K., Girones, L., Arias, A., Urrutia, R., Ye, G., Přibylová, P., Klánová, J., Ahumada, R., & Pozo, K. (2025). Land-ocean export of microplastics in central Chile: The contribution of the Biobío River using a particle-tracking model approach. *Environmental Pollution, 374*, Article 126196. https://doi.org/10.1016/j.envpol.2025.126196

Rivoira, L., Castiglioni, M., Rodrigues, S. M., Freitas, V., Bruzzoniti, M. C., Ramos, S., & Almeida, C. M. R. (2020). Microplastic in the marine environment: Reworking and optimisation of two analytical protocols for the extraction of microplastics from sediments and oysters. *MethodsX, 7*, Article 101116. https://doi.org/10.1016/j.mex.2020.101116

Rosso, B., Gregoris, E., Litti, L., Zorzi, F., Fiorini, M., Bravo, B., Barbante, C., Gambaro, A., & Corami, F. (2023). Identification and quantification of tyre wear particles by employing different cross-validation techniques: FTIR-ATR, micro-FTIR, Pyr-GC/MS, and SEM. *Environmental Pollution, 326*, Article 121511. https://doi.org/10.1016/j.envpol.2023.121511

Rume, T., & Islam, S. M. D.-U. (2020). Environmental effects of the COVID-19 pandemic and potential strategies for sustainability. *Heliyon, 6*(9), Article e04965. https://doi.org/10.1016/j.heliyon.2020.e04965

Seghers, J., Günther, M., Breidbach, A., Peez, N., Imhof, W., & Emteborg, H. (2023). Feasibility of using quantitative ^{1}H-NMR spectroscopy and ultra-microbalances for investigation of a PET microplastic reference material. *Analytical and Bioanalytical Chemistry, 415*(15), 3033–3040. https://doi.org/10.1007/s00216-023-04567-0

Sha, H., Li, H., Zhang, Y., & Hou, S. (2023). Deep learning-enhanced single-molecule spectrum imaging. *APL Photonics, 8*(9), Article 096102. https://doi.org/10.1063/5.0156793

Shruti, V. C., Pérez-Guevara, F., Roy, P. D., & Kutralam-Muniasamy, G. (2022). Analysing microplastics with Nile red: Emerging trends, challenges, and prospects. *Journal of Hazardous Materials, 423*, Article 127171. https://doi.org/10.1016/j.jhazmat.2021.127171

Singh, A., & Mishra, B. K. (2023). Microplastics in personal care products: An overlooked environmental concern. *Journal of Cleaner Production, 427*, Article 139082. https://doi.org/10.1016/j.jclepro.2023.139082

Su, L., Liu, Z., He, J., Wu, Y., & Wang, Q. (2024). The ageing behaviour of polyvinyl chloride microplastics under UV/sodium percarbonate oxidation: Efficiency and mechanism. *Water, 16*(11), Article 1529. https://doi.org/10.3390/w16111529

Tan, E., & Mohd Zanuri, N. B. (2023). Abundance and distribution of microplastics in tropical estuarine mangrove areas around Penang, Malaysia. *Frontiers in Marine Science, 10*, Article 1148804. https://doi.org/10.3389/fmars.2023.1148804

Van Haaren, C., De Bock, M., & Kazarian, S. G. (2023). Advances in ATR-FTIR spectroscopic imaging for the analysis of tablet dissolution and drug release. *Molecules, 28*(12), Article 4705. https://doi.org/10.3390/molecules28124705

Wu, J., Papanikolaou, K. G., Cheng, F., Addison, B., Cuthbertson, A. A., Mavrikakis, M., & Huber, G. W. (2024). Kinetic study of polyvinyl chloride pyrolysis with characterisation of dehydrochlorinated PVC. *ACS Sustainable Chemistry & Engineering, 12*(19), 7402–7413. https://doi.org/10.1021/acssuschemeng.4c00564

Yaseen, A. A., Yousif, E., Al-Tikrity, E. T. B., El-Hiti, G. A., Kariuki, B. M., Ahmed, D. S., & Bufaroosha, M. (2021). FTIR, weight, and surface morphology of poly(vinyl chloride) doped with tin complexes containing aromatic and heterocyclic moieties. *Polymers, 13*(19), Article 3264. https://doi.org/10.3390/polym13193264

Zainuddin, Z., & Syuhada. (2020). Study of analysis methods for microplastic identification in bottled drinking water. *Macromolecular Symposia, 391*(1), Article 1900195. https://doi.org/10.1002/masy.201900195

Zambrano-Pinto, M. V., Tinizaray-Castillo, R., Riera, M. A., Maddela, N. R., Luque, R., & Díaz, J. M. R. (2024). Microplastics as vectors of other contaminants: Analytical determination techniques and remediation methods. *Science of the Total Environment, 908*, Article 168244. https://doi.org/10.1016/j.scitotenv.2023.168244

Zhao, T., Shen, L., Ye, X., Bai, G., Liao, C., Chen, Z., Peng, T., Li, X., Kang, X., & An, G. (2023). Prenatal and postnatal exposure to polystyrene microplastics induces testis developmental disorder and affects male fertility in mice. *Journal of Hazardous Materials, 445*, Article 130544. https://doi.org/10.1016/j.jhazmat.2022.130544

Pertanika J. Sci. & Technol. 34 (3): 1905 - 1932 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Quantitative Evaluation of Machine Learning-based Land Use Classification with Multispectral Satellite Data

Krati Bansal[1]*, Anindita Nath[2], Tanupriya Choudhury[1], Bappaditya Koley[2], and Jitendra Rajpurohit[3]

[1]Department of Computer Science and Engineering, School of Computer Science, University of Petroleum and Energy Studies (UPES), 248007 Dehradun, Uttarakhand, India
[2]Department of Geography, Bankim Sardar College, 743329 South 24 Parganas, West Bengal, India
[3]Department of Computer Science and Engineering, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Lavale Campus, 412115 Pune, Maharashtra, India

ABSTRACT

Land use and land cover modification have been observed at an intense level in the Rupnarayan basin, West Bengal, India, due to the rapid expansion of fallow land and built-up area, and the decrease in water bodies and vegetation area. However, monitoring and quantifying the effects of rural catchment area modification is challenging. The heterogeneous landscape of the Rupnarayan catchment area was assessed using multispectral satellite data for 2000, 2010 and 2020 to identify land-cover change. Quantitative Evaluation of various Machine Learning techniques is adopted for the current research work, i.e., Support Vector Machine (SVM), Random Forest (RF) and Maximum Likelihood Classifier (MLC). The novelty of this work lies in benchmarking a parametric (MLC) classifier against two non-parametric classifiers (RF and SVM) within a single Google Earth Engine workflow and validating every annual map with both the Kappa statistic and AUC-ROC analysis for a previously unassessed rural reach of the basin. Five features have been taken for assessing the change detection over the region, i.e., built-up, agricultural land, fallow land, vegetation and water body. After performing each ML algorithm, SVM performed the best result among others with the overall accuracy rate of 97% (2000), 95% (2010) and 95% (2020) and a Kappa value of 0.95, while the other two scored in kappa 0.89 and 0.87 in MLC and RF, respectively. The analysis has been cross-checked by the AUC-ROC curve model. In 2000, the AUC value obtained 0.778 in MLC, RF AUC is 0.689, and SVM AUC is 0.946. In 2010, the MLC AUC is 0.875, the RF AUC value is 0.766, and the SVM value is obtained as 0.969.

ARTICLE INFO

Article history:
Received: 28 November 2025
Accepted: 08 June 2026
Published: 25 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.26

E-mail addresses:
kratibansal1992@gmail.com (Krati Bansal)
aninditan286@gmail.com (Anindita Nath)
tanupriya@ddn.upes.ac.in (Tanupriya Choudhury)
bappadityakoley2012@gmail.com (Bappaditya Koley)
jitendra.rajpurohit@sitpune.edu.in (Jitendra Rajpurohit)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

Finally, the 2020 map of each algorithm has been applied for the AUC curve, where the MLC curve value is 0.898, the RF value is 0.794, and the SVM obtained an AUC of 0.986. From the entire validation model, it has been revealed that SVM acquired the maximum AUC value for each year. This research will be beneficial for future researchers, policymakers, and land planners in enhancing management strategies for sustainable development.

Keywords: AUC-ROC, kappa coefficient, landsat images, LULC, maximum likelihood classifier, random forest, support vector machine

INTRODUCTION

Landscape contains heterogeneous factors with diverse geological, biological, social, ecological, hydrological and environmental characteristics. Land cover with local ecosystems is the primary component of the Earth's surface (Forman, 1995). Land cover is the biophysical cover of the Earth's surface, that is, vegetation and non-vegetation cover. A Land use/land cover (LULC) analysis denotes the contrast of several anthropogenic activities, i.e., built-up, mining, agriculture, etc. Spatial distribution and changes have been assessed by the quantitative LULC analysis (Ramachandra & Bharath, 2012), which also indicates the human interaction with the environmental phenomena.

Natural landscapes have changed due to anthropogenic activities, the growth of the population and climatic change over the globe. Anthropogenic activities have gradually risen over the biosphere, resulting in a large-scale alteration of the land surface of the earth, which impacts the global environment (Lambin et al., 2001). Changing of the land use is significant for effective environmental and watershed management (Twisa & Buchroithner, 2019). In developing countries, rapid land use and land cover changes (Hegazy & Mosbeh, 2015) have caused various vital natural resources to be reduced, such as vegetation, water and soil (Cheruto et al., 2016). Changes like urbanisation, expansion of the agricultural land, deforestation have a directly impacts on infiltration of groundwater, runoff and evapotranspiration (Dires & Temesgen, 2020; Ullah et al., 2024a). Climate, natural disasters and patterns of socio-economic dynamics at the global and local levels negatively affect land use and land cover change (Chakilu & Moges, 2017; Mahata et al., 2024; Ullah et al., 2024b). Ecosystem services are broadly defined as the functions that humans directly and indirectly derive from ecosystems, with terrestrial carbon storage and its consequences being widely studied and one of the most important ecosystem services (IPCC, 2006). In recent decades, land use change resulting from anthropogenic activities, namely timber and food production, has led to changes in the land use ecosystem (Fan et al., 2016).

LULC modify the land surface, fragmenting the contiguous vegetation into small patches that destroy the sustenance of carbon sedimentation, biodiversity and several geo-environmental sequences at the regional to global scale (Liu et al., 2019; Riitters et al., 2012; Ramachandra & Kumar, 2011). The Earth's surface has faced significantly decreasing

the vegetation cover due to a large amount of anthropogenic activities. Consequently, the human activities gradually increased along with river catchment areas, which caused changing are observed in land cover dynamics and the budget of surface temperature, producing changes in global, regional and local climates (Islam & Islam 2013). These dynamic changes have impacts on agriculture, the ecosystem and human society in the river basin. Present research has been carried out on the dynamic changes of the Land use and land cover (LULC) and its relations to the distributions of surface temperature (Ahmed et al. 2013). Several methods have been applied for the analysis and estimation of LULC and Land Surface Temperature (LST) (Rozenstein et al., 2014).

In recent decades, a wide range of Remote sensing and GIS classifiers technologies have been used for changes of Land used and Land covers dynamics and mapping. Multispectral spatial and temporal satellite data provide information about the land cover condition of any specific landscape, which is significant to manage the resources, monitor the modification of land cover and ecosystem, as well as mapping of LULC with change detection (Chowdhury, 2023; Farhan et al., 2024a; Manjunatha & Basavarajappa, 2020). From the recent past, various machine learning (ML) approaches (Khan et al., 2021; Shafi et al., 2023) are employed to monitor and assess the modification of LULC, i.e., Random Forest (RF) (Fu & Zhang, 2022), Support Vector Machine (SVM) (Liu et al., 2022), Maximum Likelihood Classifier (MLC) (Batar et al., 2017), Decision Trees (DT) (Zenouzi et al., 2022), Artificial Neural Network (ANN) (Raczko & Zagajewski, 2017). Among the other ML algorithms, SVM has reflected the most accurate result in limited training samples (Guo et al., 2019). The current study area is a very dynamic and rural developing catchment. In this landscape condition, SVM is a more suitable algorithm for assessing the LULC classification. But the study also applies two other ML algorithms to compare the accuracy of SVM, i.e., RF and MLC. Research in Malaysia mangrove forest has proven that SVM accuracy was 90% in overall accuracy, which is better than the MLC with 70% accuracy in overall accuracy (Deilmai et al., 2014; Halder et al., 2021). The current study aims to analyse and quantify the spatial and temporal modification of LULC in the diverse catchment area of the Rupnarayan catchment for 20 years of time span (2000 - 2020). Such land surface is the best classifier of the SVM algorithm to assess the LULC (Yousefi et al., 2022; Ge et al., 2020). The SVM algorithm has the best potential to classify the features with limited training samples (Bouslihim et al., 2022). Three classifiers were purposely chosen to represent the principal supervised families - a parametric statistical classifier (MLC), an ensemble tree-based classifier (RF) and a margin-based kernel classifier (SVM). CART is not treated separately because RF is itself an ensemble of CART trees and therefore subsumes it, while ANN, gradient boosting and deep-learning classifiers were not adopted because the limited and partly historical (2000 and 2010) training samples available for this rural basin make such data-hungry models prone to overfitting and difficult to tune reliably within the GEE environment (Fe & Zhang, 2022).

The rest of these methods, supervised classifiers, are most significant and widely applied for approaches of LULC dynamics changes (Chowdhury, 2023; Farhan et al., 2024a; Kogo et al., 2019; Langat et al., 2019). Remote-sensing classification models draw on multi-temporal, spatial, spectral and multi-sensor information, but every classified image inevitably carries some misclassification error (Chowdhury, 2026; Farhan et al., 2024b). Accuracy assessment, or verification, is performed as a validation step in remote sensing data processing, as it determines the informational value of the generated maps for end users. Effective utilisation of geodata is feasible only when the quality of the data is understood. The overall accuracy of the classified image evaluates how each pixel is classified compared to the actual land cover conditions derived from corresponding ground truth data. The producer's accuracy assesses errors of omission, indicating how effectively real-world land cover types can be classified. Conversely, commission error in measuring user's accuracy, reflecting the probability that a classified pixel corresponds accurately to the actual land cover type at its real-world location. Error metrics and the kappa coefficient have become ideal for evaluating image classification accuracy. The current accuracy assessment has also adopted the AUC-ROC curve model for cross-validation to validate the generated result by three specific supervised algorithms for each year 2000, 2010 and 2020. In the ROC curve, a statistical graphical representation has been prepared to show the accuracy of algorithms.

In the Rupnarayan catchment, very little research work has been conducted to assess the LULC using several ML algorithms. As the study is based on the Quantitative Evaluation of ML algorithms, the chief objectives of the current research have aimed i) to identify the LULC modification over the basin area; ii) to understand the change detection of the study area; iii) finally, to evaluate the performance of classified algorithms. "To the best of knowledge, no prior study of the Rupnarayan catchment has benchmarked a parametric classifier (MLC) against two non-parametric machine-learning classifiers (RF and SVM) within a single cloud-based (Google Earth Engine) workflow and then cross-validated every annual map with the AUC-ROC technique. The scientific contribution of this work is therefore three-fold: (i) a reproducible, GEE-based multi-classifier benchmark for a data-scarce rural tidal-fluvial reach; (ii) dual validation combining the Kappa confusion-matrix statistic with AUC-ROC analysis, which is rarely reported together in regional Indian LULC literature; and (iii) a 20-year (2000-2020) change-detection record quantifying the conversion of water body and vegetation into fallow and built-up land in the lower basin."

STUDY AREA CHARACTERISTICS

West Bengal is surrounded by a drainage system with many interlinking channels, rivulets, streams, and torrents, thus forming a very complex topography. In southern West Bengal, rivers such as the Damodar, Dwarakeswar, Silabati, Rupnarayan, Kangsabati and Haldi

are seen to be rapidly getting choked up. Based on the drainage networks and water availability for agriculture, water retention, and an excess resulting in floods, which is a major problem in these river basins (Mukhopadhyay & Dasgupta, 2010). The Rupnarayan River basin suffers from the highest depositional threat at its lower reaches. The Silabati and Dwarakeswar flows from opposite directions in Bandar (Ghatal), and their swollen mass is known as the Pangar, which joins the Hoogly River at Geonkhali after passing through 788 km with a watershed area of 10,797 sq. km. The Rupnarayan River is again subdivided into three reaches depending on the channel pattern of the river and associated riverine characteristics. The upper reach is a 28 km stretch beginning from Bandar to Jasar. The 10 km elongation of the middle reach is between Jasar and Kolaghat. The climate and meteorological characteristics of this basin are mainly humid subtropical in nature. However, the maximum rainfall occurs only during the rainy season, and its amount varies between 1804 mm/y to 1077 mm/y (IMD). As a result, this basin has gradually become a drought-prone area due to high summer temperatures, lack of water retention capacity in the lateral layer, and high evaporation and transpiration rates, respectively (Mittal et al., 2016). The Rupnarayan River basin reach, which is the main focus of this investigation, covers a distance of 40 km from Kolaghat (Dainan) to Geonkhali with the approximate latitudes of 22°01'20" N to 22°02'60" N and longitudes of 87°05'30" E to 88°00'30" E as shown in Figure 1. This lower reach is further divided into five sub-reaches: Geonkhali to Dhanipur, Dhanipur to Pyratungi, Pyratungi to Anantapur, Anantapur to Soyadighi, and Soyadighi to Dainan. This is done based on the needs of the study.

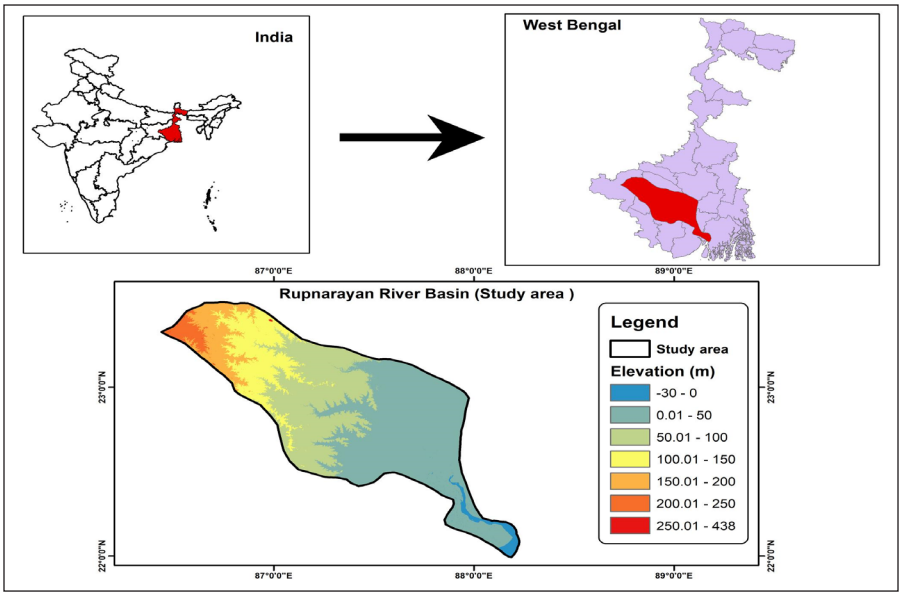


Figure 1. Location of the Rupnarayan river basin

METHODS AND MATERIALS

Data Used and Methodology

To monitor changes in land use and land cover dynamics, it is necessary to have data comparisons from at least two time periods. The LULC analysis of the present area has been performed based on a couple of steps as expressed in Figure 2. Spatial data has been acquired from a space-borne satellite sensor of the public domain and pre-processed the data through radiometric and geometric rectification to analyse the LULC of the corresponding study area (Phiri et al., 2018). LULC analyses have been implemented through machine learning supervised algorithms applying the Google Earth Engine system as a cloud-based open-source platform. The analyses of spatial data of LULC have involved i) finding out the heterogeneous areas through FCC (False Colour Composite) acquired from the wavelength of NIR, Red and Green. ii) Polygons in the virtual data were separated into polygons corresponding to heterogeneous patches in the FCC, and polygons evenly distributed throughout the region were separated from the FCC and complementary. The polygons have been covered at least with a spatial extent of 15% throughout the region, as uniformly distributed. iii) LULC analyses have been executed by the parametric and non-parametric supervised classifiers, i.e., maximum likelihood classifier (MLC), random forest (RF) and Support Vector Machine (SVM). This all-classifier method has been executed on the Google Earth Engine platform. After all the analysis, each algorithm has been validated through the Kappa coefficient statistics and the AUC-ROC curve model. LULC classifications were generated by supervised classification and the maximum likelihood method in the years 2000, 2010, and 2020 in the study area, as shown in Table 1. Scenes were selected primarily from the post-monsoon dry season, November to December, to ensure cloud-free conditions and comparable low-moisture, low-flow surface states across years; March and May scenes were used only where a cloud-free post-monsoon scene was unavailable for a given path/row. Because the basin is strongly monsoon-driven, residual seasonal differences in agricultural, fallow-land and water-body extent cannot be fully eliminated, and this is acknowledged as a limitation.

Table 1
List and sources of data collected

Data	Date Acquired	Path/Row	Resolution	Source
Landsat 5 TM	2000.11.09 2000.03.21 2000.11.23			
Landsat 5 TM	2010.12.23 2010.12.14 2010.02.20	138/045 139/044 140/044	30 m (Cloud Free)	USGS: https://earthexplorer.usgs.gov/
Landsat 8/9 (OLI) /(TIRS)	2020.12.18 2020.11.23 2020.05.22			
Base Map	Topographical Map	Study area	1:50000	Survey of India

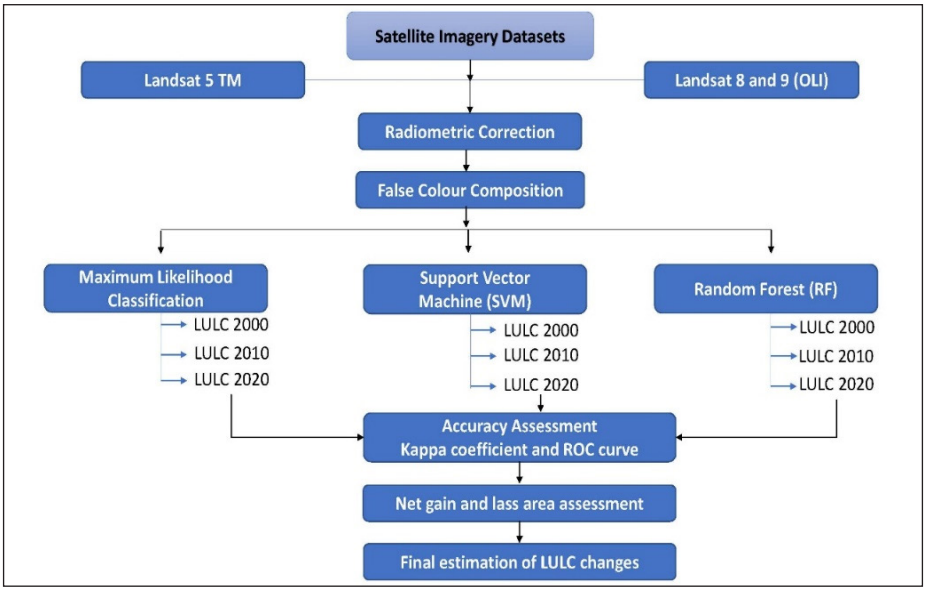


Figure 2. Flowchart of the adopted methodology

Maximum Likelihood Classification and Accuracy Assessment

Maximum likelihood classification (MLC) is a technique used to identify a known class of distributions by maximising a specific statistic. The origins of MLC can be traced back to electrical engineering. This method assumes that the training samples follow a normal distribution (Nilsson, 1956). The algorithm constructs probability density functions for each category. During the classification process, unclassified pixels are assigned to categories based on the relative likelihood (or probability) of each pixel fitting within the probability density function of each category (Hagner & Reese, 2007).

Each class spectral reflectance has been evaluated based on the qualitative method of variance and covariance as per Equation 1.

$$L = \frac{1}{(2\lambda)^{\frac{n}{2}}(\sum_k \frac{1}{2})} \exp \frac{1}{2} (X - \mu_k) \sum_k^{-1} (X - \mu_k) t \quad [1]$$

Here, L represent the likelihood of X that belongs to the K class, n denotes the spectral resolution, μ_k reflects the K class mean vector, $\sum K$ represent the matrix of the K class of variance and covariance.

Random Forest (RF)

Google Earth Engine (GEE) is the most significant tool that provides various machine learning algorithm applications through a cloud computing online platform, including random forest (RF), support vector machine (SVM), decision trees (DT), maximum likelihood classifier (MLC), etc. (Shafi et al., 2023) to analyse the land use/ land cover change. In the current analysis, the random forest classifier method of 2000, 2010 and 2020 was also adopted for a specific region.

Random Forest is a combination of non-parametric methods in machine learning that is excellent for solving the pre-direction problem as well as classification. An accumulation of decision trees, the algorithm has been generated for each of the different subsets of input data. Each subset has been generated by random choices of instances from the provided dataset with replacement and using the bootstrap sampling method; the decision trees are generated (Mateen et al., 2023). To enhance the generalising model, a random subset of data has been considered for splitting at each node of the decision trees. To decrease the correlation between trees and obtain the maximum accuracy in the outcome, a random subset of the available input dataset to make the splits during the training stage. The decision trees can be used to classify the new dataset by applying the class probability of each tree after the trees are trained well and then averaging the results over all the trees. To reduce the noise, the diverse tree facility and random features combined selection has been mitigating the overfitting of subsets. For accurate outcomes, each feature has been extracted to assess its contribution. The cloud GEE open platform is one of the most accurate and reliable algorithms applying through large scale remote sensing data (Ge et al., 2020).

Training samples have been collected through GPS-based ground truth field investigation during 2020 for LULC analysis. Maximum field data were collected according to the accessibility and portion of the selected LULC region, randomly. Previous years' field investigation was done by high-resolution images of Google Earth Proando visual interpretation system. Then uploaded the training samples in GEE as assets from ArcGIS Pro, where the field data was stored. Near bout 3000 reference points have been generated to assess the LULC in the GEE platform. After that, 80% of the dataset was randomly set for the training sample, and 20% considered as testing or validation to verify the accuracy of the RF classifier, as Feng et al. (2022) proposed. Reference data for 2020 were obtained from a GPS-based ground-truth survey, whereas reference polygons for 2000 and 2010 were generated by visual interpretation of the False-Colour-Composite of the corresponding Landsat-5 TM scenes, supported by the nearest-date archival imagery available in Google Earth Pro and by Survey of India 1: 50,000 topographic sheets. High-resolution Google Earth coverage contemporaneous with the 2000 acquisition was sparse; for such cases, interpretation relied on the Landsat FCC, and the temporally closest archival image, and ambiguous polygons were cross-checked against the topographic sheets.

Approximately 3,000 reference points were distributed across the basin and randomly partitioned so that the 20% testing subset was spatially held out and never used during training. A single random 80/20 hold-out partition was applied to all three classifiers for each year; the study did not use k-fold cross-validation and repeated random subsampling.

Support Vector Machine

In machine learning, SVM is a non-parametric binary supervised classification process. From structural risk minimisation (SRM), the SVM method was formed by dividing and maximising the data closest to the hyperplanes of spectral angle mapper (SAM). The method involves the division of the feature point into several land cover classes. The vector width of the margin line has been maximised while processing the SVM. To acquire a more precise result from SVM, a kernel function was required while performing the SVM algorithm. To enhance the distance in the vector boundary of support vectors is the actual tendency of the SVM algorithm. SVM is applied for LULC classification as well as regression assessment (Vapnik, 1999). Some binary formula, Equation 2, has been adopted for the classification:

$$(x) = \mathrm{Sin}(\sum_{i=1}^{n} \propto i\, \gamma i K(x, xi) + b) \qquad [2]$$

Here, f(x) denotes the decision function which predicts the input classes, x, αi is the multiplier of language that was acquired from the training sample, yi is the training sample of xi class label, and K (x, xi) is the kernel function. The bias term is represented by b.

Accuracy Assessment

Accuracy assessment is crucial for supervised classification within a GIS environment. The Kappa coefficient and overall accuracy are key techniques for analysing accuracy and indicating the quality of any classified image algorithm. The Kappa coefficient is derived from a confusion matrix and is applied to current land cover mapping (Chowdhury & Hafsa, 2022; McHugh, 2012). In this study, various Kappa indices were calculated, including overall accuracy, user accuracy, producer accuracy, and the Kappa coefficient itself. Overall accuracy reflects the true values of the entire Land Use and Land Cover (LULC) map, while user accuracy estimates the false values for a specific class. Producer accuracy assesses the true values of a particular class. The Kappa coefficient and its associated metrics are determined using a specific applied Equation 3 (Das et al., 2021).

$$K = \frac{[n \sum_{i=1}^{r} x_{ij} - \sum_{i=1}^{r} x_i x_j]}{[n^2 - \sum_{i=1}^{r} (x_i x_j)]} \qquad [3]$$

Where K indicates Kappa Coefficient; n indicates the total number of samples; x_{ij} is the total corrected sample; x_i is the producer total, and x_j is the user total

Overall accuracy is estimated using Equation 4:

$$\text{OA} = \frac{\sum CP_{ii}}{N} \quad [4]$$

OA indicates the Overall accuracy. ΣCP_{ii} indicates the total observed samples classified correctly for each feature. N indicates the total number of samples observed.

Model Validation by AUC-ROC

The present analysis of LULC by various methods, i.e., SSVM, RF, and MLC, has been validated by a true statistical method, AUC (area under the curve), and ROC (receiver operating characteristics), which is one of the most suitable and mathematical approaches for model validation, also applied by several researchers in the past. In the model, the actual position rate has been assessed by the false position rate of the hypothesis. The following formula, Equation 5, has been applied here to calculate the AUC-ROC:

$$\text{AUC} = \sum_{i=1}^{n=100} \frac{(x_1 + x_2)}{2(y_2 + y_1)} \quad [5]$$

RESULT AND DISCUSSION

Comparative Assessment of LULC Change Dynamics by MLC, RF and SVM

Temporal changes in land use/land cover analysis between the years 2000 and 2020 have been performed through supervised classification algorithms of maximum likelihood classifier (MLC), random forest (RF) and support vector machine (SVM) systems. The analysis of each algorithm has been calculated for the years of 2000, 2010 and 2020 to depict the land use and land cover condition of the Rupnarayan river basin area. First, the MLC has been generated over the area for 2000, 2010 and 2020, where five distinct features have been selected, i.e., built up. Vegetation, fallow land, agricultural land and water body. From the LULC map through MLC, it has been revealed that 164.2 sq.km area was found under built-up area in the year 2000, where it has been expanded in 2010 to 3207.06 sq.km and 1976.46 sq.km area was covered with built-up in 2020.

From the analysis of MLC agricultural land area, in 2000 it covered only 308.8 sq.km, but in 2010 it increased to 2167.89 sq.km and decreased to 693.2 sq.km in 2020. Fallow land in the study area reveals the major changes where a 5001.1 sq.km area fell under fallow land in 2000 and 1893.54 sq.km in 2010; however, increases in 2020 to 7867.55 sq.km, that denote a very positive expansion of fallow land. Vegetation area by MLC in 2000, showing the 1365.2 sq.km and expanded to 3237.89 sq.km in 2010, but extremely decreased in the year 2020 to 45.46 sq.km. In the LULC assessment, major modification has been observed in the water body where a 3957.7 sq.km area in the year 2000 reduced to 290.66 sq.km and 214.37 sq.km in the years 2010 and 2020, respectively, as shown in Figure 12. The map of land use/land cover has been generated to assess the change detection of each year through three different supervised algorithms and spatial data derived from the prepared map to identify the loss and gain areas within the period, as shown in Figures 3 to 11.

From the analysis of the random forest, it has been observed that some spatial changes were made in each feature area with the temporal condition. In case of built-up, 1931.00 sq.km area in 2000 has been enhanced in 2010 to 5839.88 sq.km, which again reduced in 2020 to 3569.12 sq.km. A very positive change has been found in agricultural land, where a 338.11 sq.km area in 2000 expanded to 1366.81 sq.km in 2020. In fallow land, a 3703.42 sq.km area in 2000 shifted to a 330.59 sq.km area in 2010, which was modified to 53.1539 sq.km in 2020. A drastic change has been observed in the vegetation area, where a 1286.27 sq.km area was found in 2000, which increased to 4164.59 sq.km in 2010, but was extremely reduced in 2020 to 71.197 sq.km by RF. Water bodies represent a negative shift over 20 years, where 3538.21 sq.km in 2000 shifted to 305.46 sq.km and 474.51 sq.km in 2010 and 2020, respectively, as shown in Figure 12.

Finally, the study generates a map of support vector machine (SVM) for each year to show the land use/land cover condition. In the year 2000, a map of the SVM built-up area observed in 469.24 sq.km, which expanded to 6227.37 sq.km in 2010 and again reduced in 2020 to 2121.09 sq.km. Agricultural area has been shifted from 471.377 sq.km in 2000 to 2392.68 sq.km and 1529.66 sq.km area in 2010 and 2020, respectively. After the analysis of SVM, the fallow land has been denoted as the major modification where 5609.54 sq.km area was modified to 197.62 sq.km and 6170.21 sq.km in 2010 and 2020, respectively. Similarly, the vegetation area in 2000 was revealed to be 2160.74 sq.km, which shifted to 1614.40 sq.km and 658.43 sq.km in 2010 and 2020, respectively. The water body has been gradually decreasing through SVM analysis, where 2086.13 sq.km area in 2000 regressed to 364.97 sq.km and 317.64 sq.km in 2010 and 2020, respectively, as shown in Figure 12.

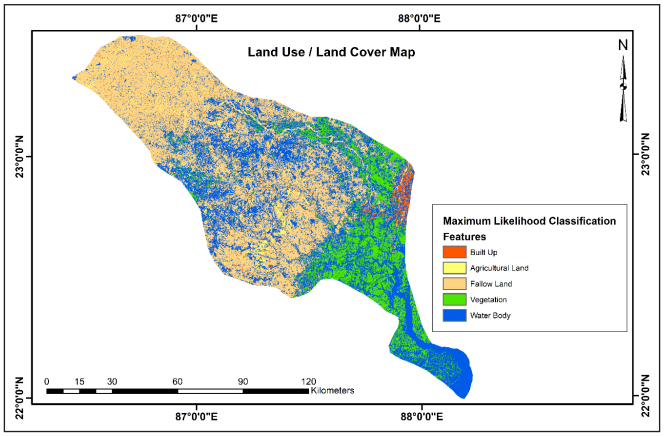


Figure 3. LULC map by maximum Likelihood Classifier (MLC), year 2000

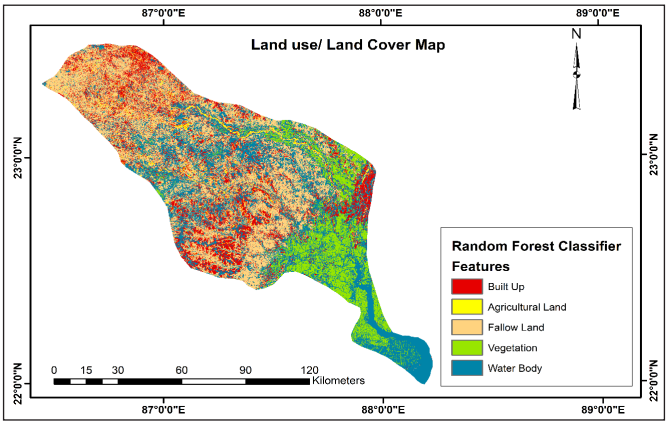


Figure 4. LULC map by Random Forest (RF), year 2000

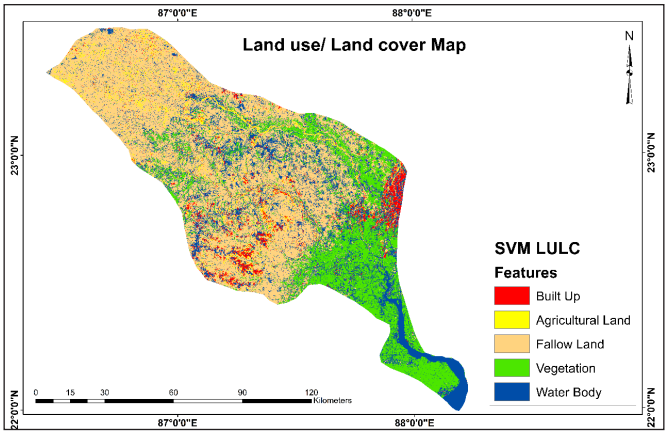


Figure 5. LULC map by Support Vector Machine (SVM), year 2000

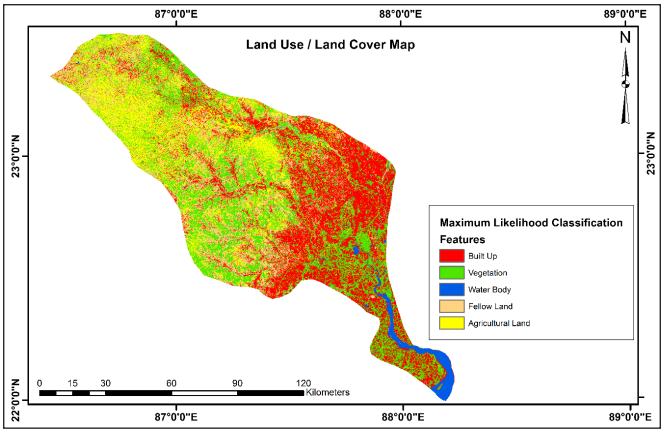


Figure 6. LULC map by Maximum Likelihood Classifier (MLC), year 2010

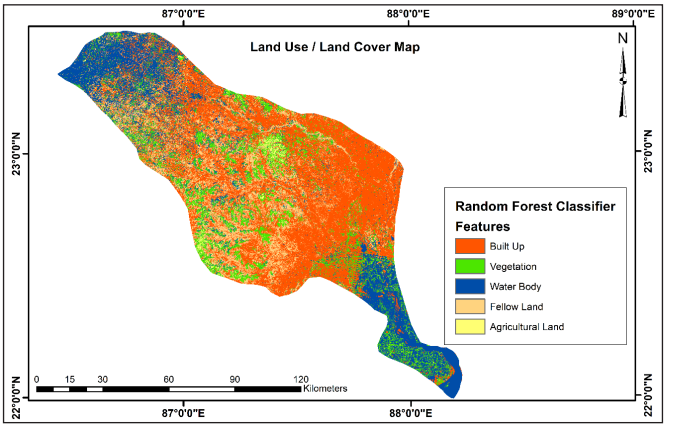


Figure 7. LULC map by Random Forest (RF), year 2010

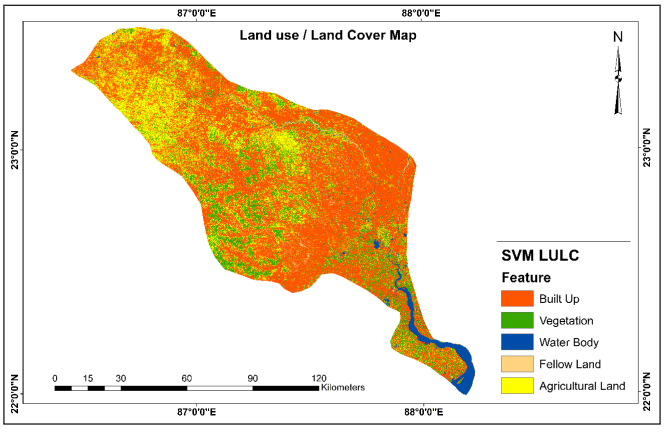


Figure 8. LULC map by Support Vector Machine (SVM), year 2010

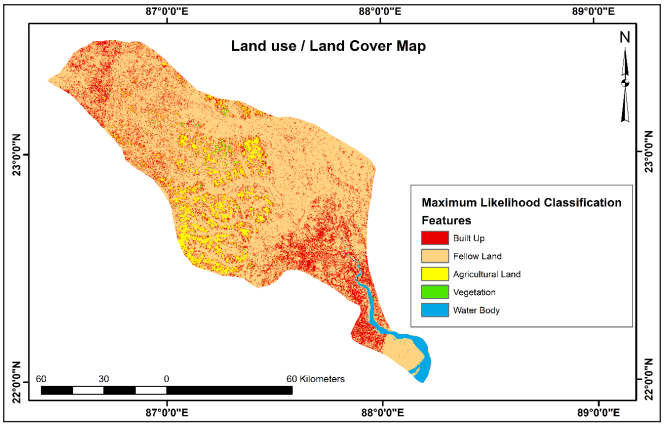


Figure 9. LULC map by Maximum Likelihood Classifier (MLC), year 2020

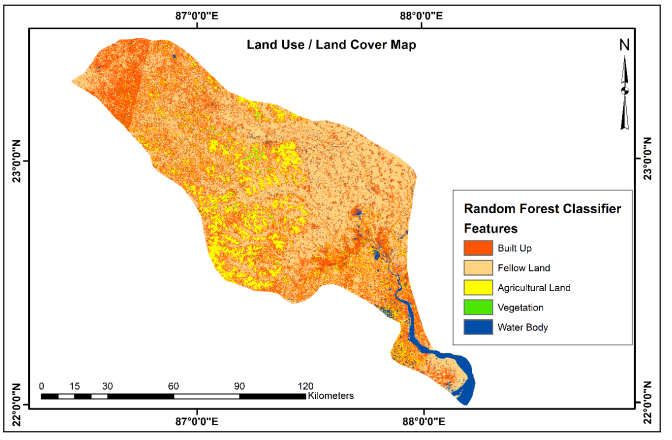


Figure 10. LULC map by Random Forest (RF), year 2020

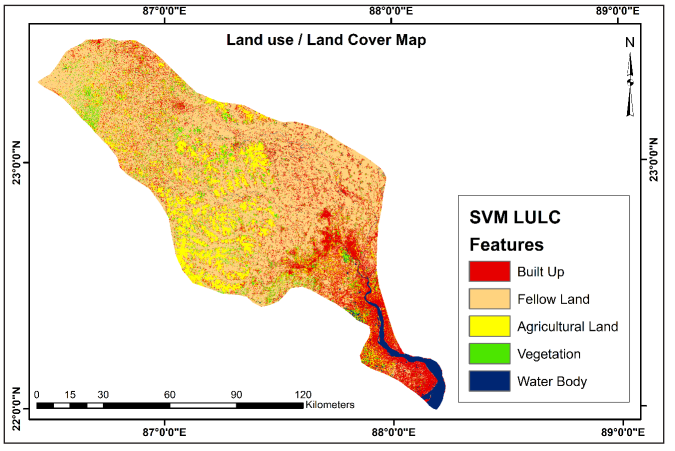
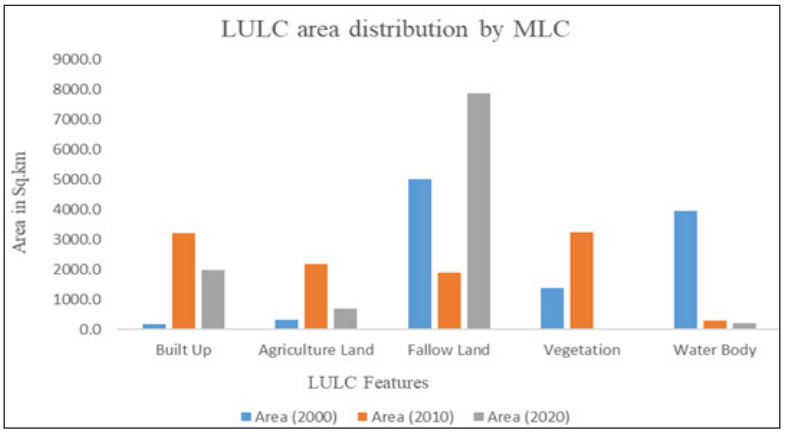
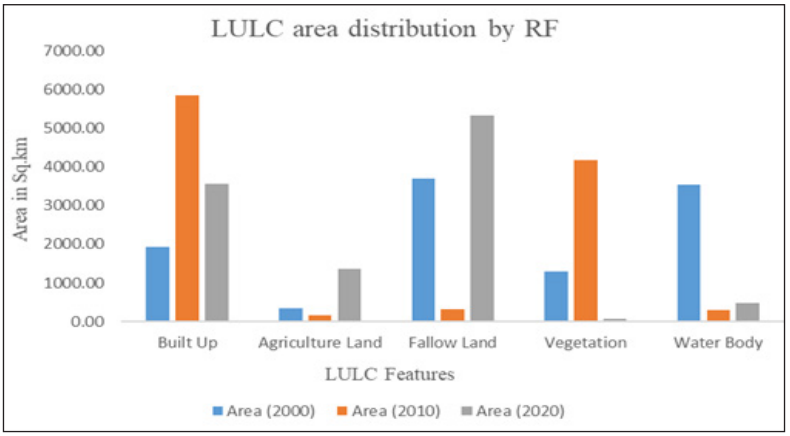


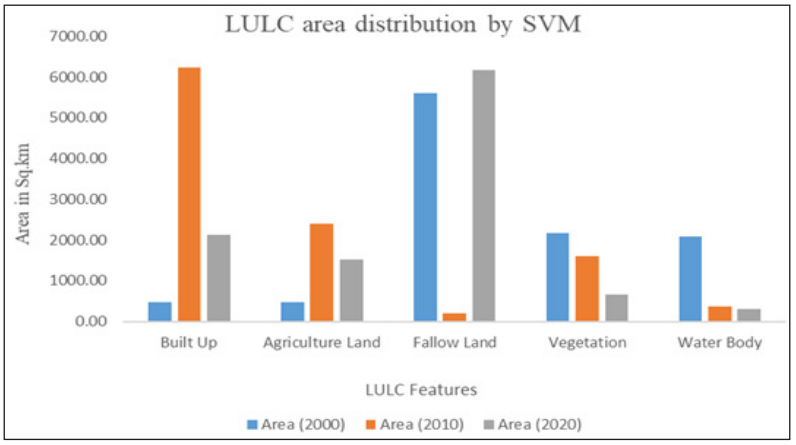
Figure 11. LULC map by Support Vector Machine (SVM), year 2020



(a)



(b)



(c)

Figure 12. LULC area distribution by (a) MLC; (b) RF; (c) SVM

Change Detection of LULC from 2000 to 2010 and 2010 to 2020 by MLC RF SVM

After the analysis, the loss and gain areas have been calculated for each algorithm from 2000 to 2010 and from 2010 to 2020. Table 2 and Figure 13 (a) show the modification of each feature through the MLC method, where the built-up area positively changed from 2000 to 2010 to 3042.85 sq.km, but built up reduce in the next 10 years (2010 to 2020) to -1230.61 sq.km. On the other hand, in the RF method, as shown in Table 3 and Figure 13 (b), it has been depicted that the built-up area increased between the first 10 years, 3908.88 sq.km, but reduced in the next 10 years to -2270.76 sq.km. But maximum built-up reduction has been observed by the SVM method -4106.29 sq.km as shown in Table 4 and Figure 13 (c). In agricultural land -1474.68 sq.km area disappeared from the MLC method from 2010 to 2020, but 1210.31 sq.km area was enhanced in the RF model, where -863.01 sq.km area was again lost in the SVM method from 2010 to 2020. In the case of fallow land, the 2000 to 2010 period depicts the negative change where -3107.55 sq.km area lost by MLC, -3372.83 sq km and -5411.93 sq km by RF and SVM accordingly within that time. But the next 10 years showed the positive change where 5974.01, 4984.79 and 5972.59 sq.km by MLC, RF and SVM, respectively. In the case of vegetation, MLC and RF models remarked the positive change in the 2000 to 2010 period, but the SVM model depicts the negative change in both time spans (-546.34 sq.km in 2000 – 2010 and -955.98 sq.km in 2010 to 2020). A proper negative shifting has been observed in the water body by the MLC and SVM model, where -3232.75 sq.km and -1721.16 sq.km were lost in 2000 to 2010 and -76.27 sq.km and -47.32 sq.km in 2010 to 2020, as shown in Tables 2 and 4 and in Figure 3.

Table 2
LULC change detection by MLC (From 2000 to 2010 and 2010 to 2020)

Feature	Area (Year 2000)	Area (Year 2010)	Area (Year 2020)	Change Detection (Year 2000 to 2010)	Change Detection (Year 2010 to 2020)
Built Up	164.2	3207.06	1976.45	3042.85	-1230.61
Agriculture Land	308.8	2167.88	693.19	1859.11	-1474.68
Fallow Land	5001.1	1893.54	7867.55	-3107.55	5974.01
Vegetation	1365.2	3237.89	45.45	1872.6426	-3192.44
Water Body	3957.7	290.66	214.38	-3667.06	-76.27
			10797.04		

Table 3
LULC change detection by RF (From 2000 to 2010 and 2010 to 2020)

Feature	Area (Year 2000)	Area (Year 2010)	Area (Year 2020)	Change Detection (Year 2000 to 2010)	Change Detection (Year 2010 to 2020)
Built Up	1931.01	5839.88	3569.12	3908.88	-2270.76
Agriculture Land	338.12	156.50	1366.81	-181.62	1210.31

Table 3 (*continued*)

Feature	Area (Year 2000)	Area (Year 2010)	Area (Year 2020)	Change Detection (Year 2000 to 2010)	Change Detection (Year 2010 to 2020)
Fallow Land	3703.43	330.59	5315.39	-3372.83	4984.79
Vegetation	1286.27	4164.59	71.18	2878.32	-4093.39
Water Body	3538.22	305.47	474.51	-3232.75	169.05
			10797.04		

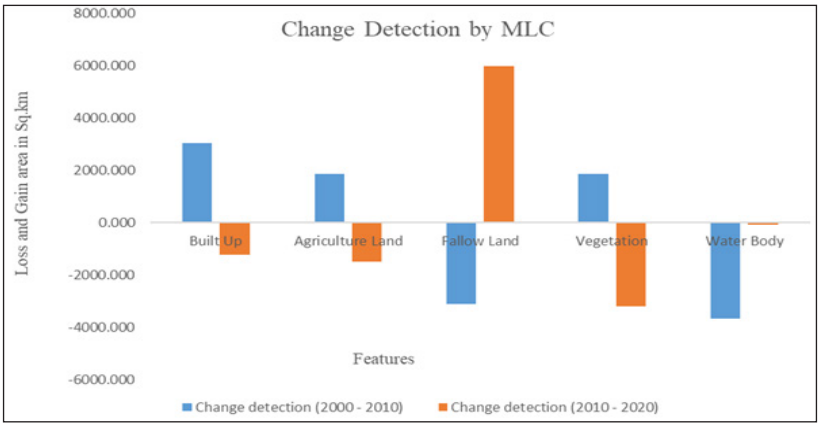
Table 4
LULC change detection by SVM (From 2000 to 2010 and 2010 to 2020)

Feature	Area (Year 2000)	Area (Year 2010)	Area (Year 2020)	Change Detection (Year 2000 to 2010)	Change Detection (Year 2010 to 2020)
Built Up	469.24	6227.36	2121.07	5758.12	-4106.29
Agriculture Land	471.38	2392.68	1529.66	1921.34	-863.01
Fallow Land	5609.54	197.62	6170.21	-5411.93	5972.59
Vegetation	2160.74	1614.40	658.42	-546.34	-955.98
Water Body	2086.13	364.97	317.64	-1721.16	-47.320
			10797.04		

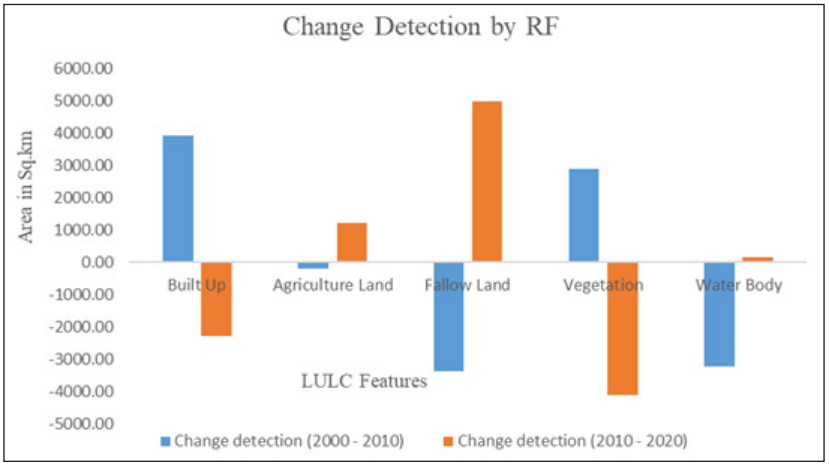
Comparative Accuracy Assessment of MLC RF SVM

The per-class areas estimated by the three classifiers differ in magnitude, which is expected because each algorithm applies a different decision rule to the same imagery; these differences should be read as a comparative evaluation of classifier behaviour rather than as data contradictions. Importantly, all three classifiers agree on the direction of change - a decline of water bodies and vegetation and an expansion of fallow and built-up land over 2000 to 2020, and they differ mainly in how sharply they resolve the spectrally mixed classes. The following paragraphs summarise each classifier in turn, with full figures reported in Tables 2 to 4 and Figures 12 to 13.

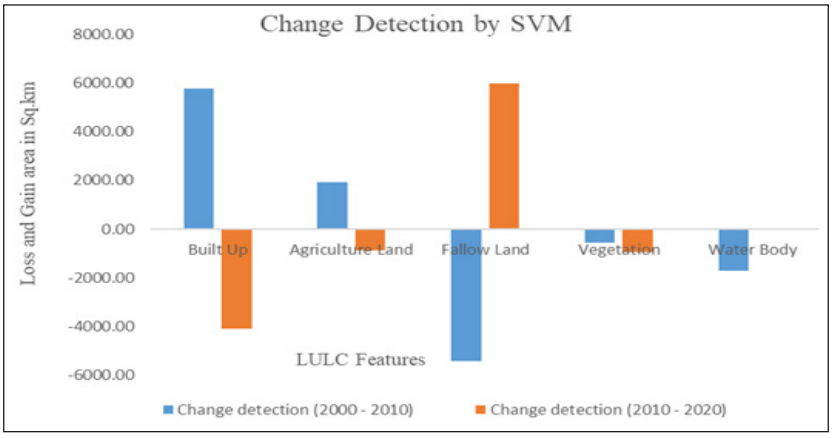
The LULC classification has been evaluated based on the accuracy assessment as shown in Tables 5, 6 and 7. The comparative supervised algorithm reveals that the SVM machine learning algorithm has been more accurate and precise for the current study. The kappa value has been obtained in 0.97 (2000), 0.95 (2010) and 0.95 (2020) in SVM. On the other hand, the second relevant method has been depicted in the MLC classifier, where an accuracy of 0.85 was obtained in the year 2000, 0.88 in 2010 and 0.89 in 2020. The RF model revealed the least accuracy in the current research work, where a 0.85 value was acquired in 2000, 0.85 in 2010 and 0.87 in 2020. From the comparative evaluation, it has been seen that built-up area, agriculture and water body misplaced in the omission error while performing the model.



(a)



(b)



(c)

Figure 13. Change statistics of LULC features by (a) MLC; (b) RF; (c) SVM

Similarly, in the MLC model, built-up and vegetation areas have been misplaced while operating the algorithm. But the SVM algorithm portrayed the maximum accuracy in omission and commission error in the current LULC assessment. Tables 5 to 7 provide the category-wise accuracy assessment of each algorithm through the Kappa coefficient.

Table 5
Accuracy assessment of MLC by the kappa coefficient

LULC Classes	User's Accuracy			Producer's Accuracy			Overall Accuracy			Kappa Coefficient (K)		
	2000	2010	2020	2000	2010	2020	2000	2010	2020	2000	2010	2020
Vegetation	85	90	95	87	88	87	85%	88%	90%	0.85	0.88	0.89
Water Body	87	90	92	92	90	85						
Build up	85	87	90	82	85	85						
Agricultural Land	92	85	88	85	88	87						
Fallow Land	85	88	90	90	85	85						

Table 6
Accuracy assessment of RF by the kappa coefficient

LULC Classes	User's Accuracy			Producer's Accuracy			Overall Accuracy			Kappa Coefficient (K)		
	2000	2010	2020	2000	2010	2020	2000	2010	2020	2000	2010	2020
Vegetation	88	90	87	85	88	92	85%	85%	87%	0.85	0.85	0.87
Water Body	89	85	85	84	89	89						
Build up	88	85	82	87	87	85						
Agricultural Land	89	92	87	90	85	92						
Fallow Land	90	85	92	85	83	88						

Table 7
Accuracy assessment of SVM by the kappa coefficient

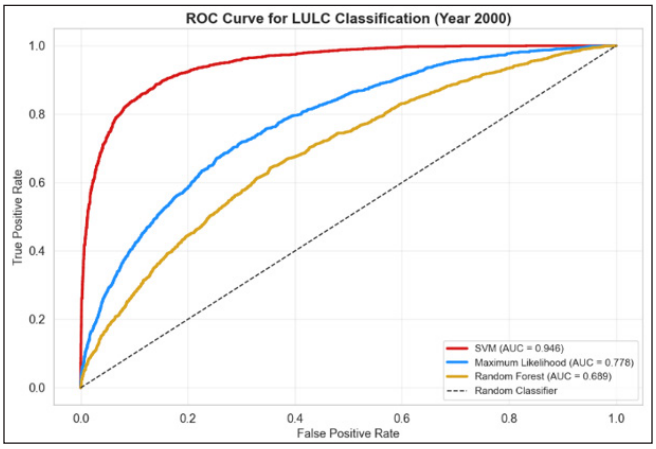
LULC Classes	User's Accuracy			Producer's Accuracy			Overall Accuracy			Kappa Coefficient (K)		
	2000	2010	2020	2000	2010	2020	2000	2010	2020	2000	2010	2020
Vegetation	92	100	100	95	92	100	97%	95%	95%	0.97	0.95	0.95
Water Body	90	92	95	100	100	100						
Build up	92	100	100	90	100	100						
Agricultural Land	90	90	90	95	100	100						
Fallow Land	92	95	100	95	95	100						

Model Validation by AUC-ROC

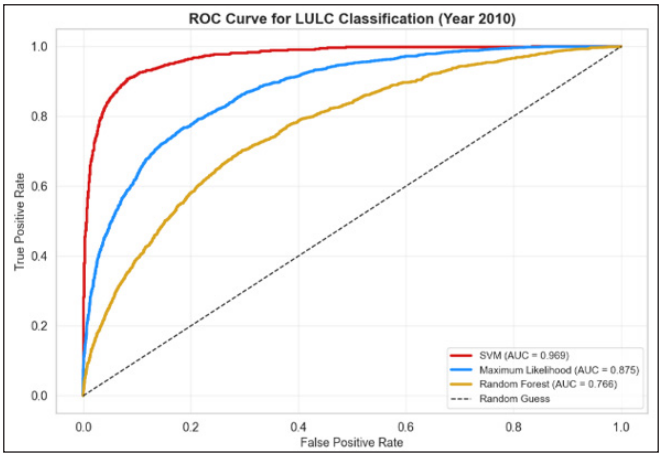
It is significant to ensure the model validation with the real ground conditions of LULC. The spatial position of actual land cover features is required to match the generated land false colour map by SVM, RF and MLC of each year (Ogato et al., 2020; Tadesse et al., 2022). The applied model of each algorithm of each year has been cross checked the accuracy through the AUC-ROC model. All the land use spatial conditions have been cross-verified by the ground position through Global Positioning System (GPS) and high-resolution Google Earth imagery to validate the model. From the ArcSDM extension tool of ArcGIS, the curve has been drawn. In the analysis of 2000, the AUC value obtained 0.778 in MLC, RF AUC is 0.689, and SVM AUC is 0.946. In 2010, the MLC AUC is 0.875, the RF AUC value is 0.766, and the SVM value is obtained as 0.969. Finally, the 2020 map of each algorithm has been applied for the AUC curve, where the MLC curve value is 0.898, the RF value is 0.794, and the SVM obtained an AUC of 0.986. From the entire validation model, it has been revealed that SVM acquired the maximum AUC value for each year, while the other two (MLC and RF) algorithms obtained the lowest range of AUC value, as shown in Figure 14. Moreover, a 0.85 value in the model has been considered to validate the algorithm.

DISCUSSION

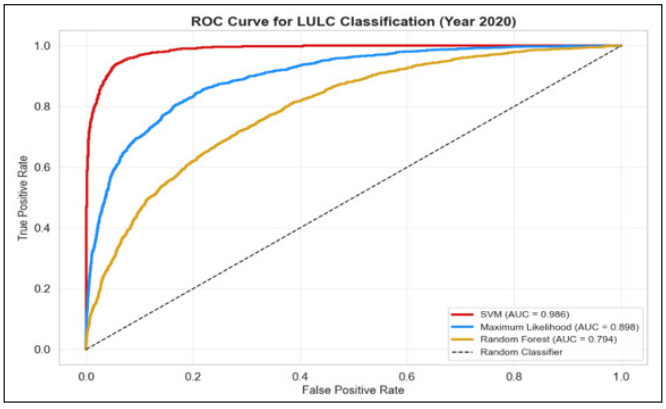
Three different algorithms have been applied for assessing the LULC in the Rupnarayan catchment area, i.e., MLC, RF and SVM. Three model was depicting the spatial extension of the built-up area during the years 2000 to 2010. The spatial extension of fallow land increased, and 5974.01 sq.km area gained during 2010 to 2020 in MLC, 4984.79 sq.km increased by RF, and 5972.59 sq.km area enhanced through the SVM method due to the loss of agricultural land, which reduced -863.01 sq.km during the last ten years. After generating the LULC map, it has been observed that extreme modification depicted in the water body has been reduced within two specific time spans due to the fallow land expansion over the catchment area. In the SVM method, the water body area reduced by -1721.16 sq.km during the years 2000 to 2010 and again -47.32 sq.km was lost between 2010 and 2020. The evaluation application of random forest, maximum likelihood classifier and support vector machine classifier for assessing the land cover change in Rupnarayan River catchment area, it has been revealed that the most accurate outcomes were acquired through the SVM classifier method with a high accuracy of Kappa coefficient value and AUC-ROC model validation of each model. SVM consistently outperformed MLC and RF (Figure 14) because its margin-based decision boundary separates spectrally overlapping classes (fallow vs. agricultural land, built-up vs. bare soil) more effectively than the Gaussian-distribution assumption of MLC or the variance of individual trees in RF, and because it is comparatively robust when training samples are limited - a key advantage given the scarce historical reference data for 2000 and 2010.



(a)



(b)



(c)

Figure 14. Classifier algorithm validation by specificity and sensitivity, AUC-ROC curve of (a) Year 2000; (b) Year 2010; (c) Year 2020

Practically, the consistent loss of water bodies and vegetation against expanding fallow and built-up land signals reduced water-retention capacity and rising land-surface stress in the lower basin, with direct implications for flood management and sustainable land planning.

CONCLUSION

The current study indicates that the Rupnarayan River Basin has undergone significant dynamic changes in LULC from 2000 to 2020. While land use is increasing, this study highlights a decline in certain areas. Across all three years, the SVM classifier returned the highest accuracy (Kappa 0.97 / 0.95 / 0.95 and AUC 0.946 / 0.969 / 0.986 for 2000 / 2010 / 2020), consistently outperforming MLC and RF, which confirms that a margin-based classifier is best suited to this heterogeneous, data-scarce catchment. The dominant change detected was a marked decline of water bodies and vegetation and a corresponding expansion of fallow and built-up land between 2000 and 2020. The current research work three specific supervised algorithms, i.e., MLC, RF and SVM. After generating the algorithm in high-resolution multispectral satellite data, it has been remarked that the SVM algorithm provides the maximum accuracy on ground truth data and Google Earth imagery. A ROC curve was also generated based on the true and false data to validate the model of each year, which also reflects the highest SVM AUC value in each year. This finding is consistent with earlier comparative studies in which SVM outperformed parametric classifiers (Darabi et al., 2019; Fatah et al., 2025; Javidan et al., 2019; Jodhani et al., 2025; Ruidas et al., 2023; Tripathi et al., 2023), and it extends that evidence to a data-scarce tidal-fluvial Indian catchment validated jointly by Kappa and AUC-ROC.

From the analysis, it has been observed that the most notable decline was observed in the water body. Additionally, vegetation cover decreased in both time periods between 2000 and 2020 through SVM. Conversely, the built-up area increased significantly, primarily at the expense of Fallow land. It is essential to develop innovative land systems tailored to the local context, framed within the global socio-ecological system. These systems should explicitly consider land governance as a key driver of land change and agricultural production.

This study has several limitations. Scenes were drawn from slightly different months, so residual seasonal variation in fallow, agricultural and water classes cannot be fully excluded. Accuracy was assessed with a single random 80/20 hold-out rather than k-fold cross-validation or repeated sub-sampling, so the reported accuracies should be read as indicative. Reference data for 2000 and 2010 relied partly on visual interpretation of archival imagery, which introduces interpretation uncertainty, and the 30 m Landsat resolution produces mixed pixels along class boundaries. Future work should (i) extend the comparison to CART, gradient boosting and deep-learning (CNN) classifiers; (ii) fuse finer-resolution multi-sensor data (e.g. Sentinel-2) to reduce mixed pixels;

(iii) adopt k-fold cross-validation with repeated random sampling to quantify accuracy stability; and (iv) link the observed LULC trajectories to their socio-hydrological drivers.

The significance of this study is that it delivers a validated, reproducible GEE-based workflow and a 20-year change record for the lower Rupnarayan basin, providing policymakers, land planners and watershed managers with a reliable evidence base for sustainable land- and water-resource management in similar rural river catchments

AUTHOR CONTRIBUTIONS

KB, AN, TC and BK contributed to conceptualising the study and writing the introduction, literature review, data analysis, modelling, and interpretation. BK formally analyses the data and provides the figure representation.TC, BK and JR handled formatting, language editing, review the manuscript, and AN and TC supervised this research. Manuscript confirming that all authors have approved the final version of the manuscript.

DECLARATIONS

Ethical Approval and Consent to Participate

No human subjects are involved in this research.

Consent for Publication

Not applicable.

Competing Interests

The authors declare that they do not have any competing interests.

Data Availability

The datasets used in this study were derived from publicly available satellite imagery. All outcomes and generated datasets were cross-verified through ground-truth field surveys to ensure accuracy and reliability. Spatial data processing, analysis, and mapping were conducted using the ArcGIS platform and Google Earth Engine (GEE). The processed data and related materials are available from the corresponding author upon reasonable request.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to all individuals and organisations that contributed to the successful completion of this research. The authors also thank the anonymous reviewers and the editor for their constructive comments and suggestions, which helped improve the quality of the manuscript. Additionally, we gratefully acknowledge the

Here Technologies Lab, University of Petroleum and Energy Studies (UPES), Dehradun, Uttarakhand, India, for providing the necessary research facilities.

REFERENCES

Ahmed, B., Kamruzzaman, M., Zhu, X., Rahman, M. S., & Choi, K. (2013). Simulating land cover changes and their impacts on land surface temperature in Dhaka, Bangladesh. *Remote Sensing, 5*(11), 5969-5998. https://doi.org/10.3390/rs5115969

Batar, A. K., Watanabe, T., & Kumar, A. (2017). Assessment of land-use/land-cover change and forest fragmentation in the Garhwal Himalayan region of India. *Environments, 4*(2), Article 34. https://doi.org/10.3390/environments4020034

Bouslihim, Y., Kharrou, M. H., Miftah, A., Attou, T., Bouchaou, L., & Chehbouni, A. (2022). Comparing pan-sharpened Landsat-9 and Sentinel-2 for land-use classification using machine learning classifiers. *Journal of Geovisualization and Spatial Analysis, 6*, Article 35. https://doi.org/10.1007/s41651-022-00130-0

Chakilu, G., & Moges, M. (2017). Assessing the land use/cover dynamics and their impact on the low flow of the Gumara Watershed, Upper Blue Nile Basin, Ethiopia. *Hydrology: Current Research, 7*(2), Article 268. https://doi.org/10.4172/2157-7587.1000268

Cheruto, M. C., Kauti, M. K., Kisangau, P. D., & Kariuki, P. (2016). Assessment of land use and land cover change using GIS and remote sensing techniques: A case study of Makueni County, Kenya. *Journal of Remote Sensing & GIS, 5*, Article 175. https://doi.org/10.4172/2469-4134.1000175

Chowdhury, M. S. (2023). GIS-based method for mapping actual LULC by combining seasonal LULCs. *MethodsX, 11*, Article 102472. https://doi.org/10.1016/j.mex.2023.102472

Chowdhury, M. S. (2026). Incorporation of seasonal land cover to develop a seasonally integrated land cover map for change detection in Dhaka Mega City. *Scientific Reports, 16*(1), Article 468. https://doi.org/10.1038/s41598-025-17478-0

Chowdhury, M. S., & Hafsa, B. (2022). Multi-decadal land cover change analysis over Sundarbans Mangrove Forest of Bangladesh: A GIS and remote sensing-based approach. *Global Ecology and Conservation, 37*, Article e02151. https://doi.org/10.1016/j.gecco.2022.e02151

Darabi, H., Choubin, B., Rahmati, O., Haghighi, A. T., Pradhan, B., & Kløve, B. (2019). Urban flood risk mapping using the GARP and QUEST models: A comparative study of machine learning techniques. *Journal of Hydrology, 569*, 142-154. https://doi.org/10.1016/j.jhydrol.2018.12.002

Das, N., Mondal, P., Sutradhar, S., & Ghosh, R. (2021). Assessment of variation of land use/land cover and its impact on land surface temperature of Asansol subdivision. *The Egyptian Journal of Remote Sensing and Space Sciences, 24*(1), 131-149. https://doi.org/10.1016/j.ejrs.2020.05.001

Deilmai, B. R., Ahmad, B. B., & Zabihi, H. (2014). Comparison of two classification methods (MLC and SVM) to extract land use and land cover in Johor, Malaysia. *IOP Conference Series: Earth and Environmental Science, 20*(1), Article 012052. https://doi.org/10.1088/1755-1315/20/1/012052

Fan, S., Guan, F., Xu, X., Forrester, D. I., Ma, W., & Tang, X. (2016). Ecosystem carbon stock loss after land use change in subtropical forests in China. *Forests, 7*(7), Article 142. https://doi.org/10.3390/f7070142

Farhan, M., Wu, T., Amin, M., Tariq, A., Guluzade, R., & Alzahrani, H. (2024a). Monitoring and prediction of the LULC change dynamics using time series remote sensing data with Google Earth Engine. *Physics and Chemistry of the Earth, Parts A/B/C, 136*, Article 103689. https://doi.org/10.1016/j.pce.2024.103689

Farhan, M., Wu, T., Anwar, S., Yang, J., Naqvi, S. A. A., Soufan, W., & Tariq, A. (2024b). Predicting land use land cover dynamics and land surface temperature changes using CA-Markov chain models in Islamabad, Pakistan (1992-2042). *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17*, 16255-16271. https://doi.org/10.1109/JSTARS.2024.3441241

Fatah, K. K., Mustafa, Y. T., & Hassan, I. O. (2025). Assessing the influence of climate and land use/land cover changes on groundwater levels in arid and semi-arid regions: A geoinformatics-based machine learning approach in the Iraqi Kurdistan Region. *GeoJournal, 90*(3), Article 125. https://doi.org/10.1007/s10708-025-11374-w

Feng, S., Li, W., Xu, J., Liang, T., Ma, X., Wang, W., & Yu, H. (2022). Land use/land cover mapping based on GEE for the monitoring of changes in ecosystem types in the Upper Yellow River Basin over the Tibetan Plateau. *Remote Sensing, 14*(21), Article 5361. https://doi.org/10.3390/rs14215361

Forman, R. T. T. (1995). Some general principles of landscape and regional ecology. *Landscape Ecology, 10*(3), 133-142. https://doi.org/10.1007/BF00133027

Fu, Y., & Zhang, Y. (2022). Research on temporal and spatial evolution of land use and landscape pattern in Anshan City based on GEE. *Frontiers in Environmental Science, 10*, Article 988346. https://doi.org/10.3389/fenvs.2022.988346

Ge, G., Shi, Z., Zhu, Y., Yang, X., & Hao, Y. (2020). Land use/cover classification in an arid desert-oasis mosaic landscape of China using remotely sensed imagery: Performance assessment of four machine learning algorithms. *Global Ecology and Conservation, 22*, Article e00971. https://doi.org/10.1016/j.gecco.2020.e00971

Guo, Y., Yin, X., Zhao, X., Yang, D., & Bai, Y. (2019). Hyperspectral image classification with SVM and guided filter. *EURASIP Journal on Wireless Communications and Networking, 2019*, Article 56. https://doi.org/10.1186/s13638-019-1346-z

Hagner, O., & Reese, H. (2007). A method for calibrated maximum likelihood classification of forest types. *Remote Sensing of Environment, 110*(4), 438-444. https://doi.org/10.1016/j.rse.2006.08.017

Halder, B., Bandyopadhyay, J., & Banik, P. (2021). Monitoring the effect of urban development on urban heat island based on remote sensing and geospatial approach in Kolkata and adjacent areas, India. *Sustainable Cities and Society, 74*, Article 103186. https://doi.org/10.1016/j.scs.2021.103186

Hegazy, I. R., & Mosbeh, R. K. (2015). Monitoring urban growth and land use change detection with GIS and remote sensing techniques in Daqahlia Governorate, Egypt. *International Journal of Sustainable Built Environment, 4*(1), 117-124. https://doi.org/10.1016/j.ijsbe.2015.02.005

Intergovernmental Panel on Climate Change. (2006). *2006 IPCC guidelines for national greenhouse gas inventories*.https://www.ipcc-nggip.iges.or.jp/public/2006gl/vol4.html

Islam, M. S., & Islam, K. S. (2013). Application of thermal infrared remote sensing to explore the relationship between land use-land cover changes and urban heat island effect: A case study of Khulna City. *Journal of Bangladesh Institute of Planners, 6*, 49-60. https://doi.org/10.3329/jbip.v6i1.76962

Javidan, N., Kavian, A., Pourghasemi, H. R., Conoscenti, C., & Jafarian, Z. (2019). Gully erosion susceptibility mapping using multivariate adaptive regression splines: Replications and sample size scenarios. *Water, 11*(11), Article 2319. https://doi.org/10.3390/w11112319

Jodhani, K. H., Patel, D., Madhavan, N., Gupta, N., Singh, S. K., & Pandey, M. (2025). ML-based land use and land cover classification: Assessing performance and predicting future changes. *Journal of Hydrologic Engineering, 30*(4), Article 05025011. https://doi.org/10.1061/JHYEFF.HEENG-6416

Khan, A. R., Khan, A., Masud, S., & Rahman, R. M. (2021). Analysing the land cover change and degradation in Sundarbans Mangrove Forest using machine learning and remote sensing techniques. In I. Rojas, G. Joya, & A. Català (Eds.), *Advances in computational intelligence* (Lecture Notes in Computer Science, Vol. 12862, pp. 429-438). Springer. https://doi.org/10.1007/978-3-030-85099-9_35

Kogo, B. K., Kumar, L., & Koech, R. (2019). Analysis of spatiotemporal dynamics of land use and land cover changes in Western Kenya. *Geocarto International*, 1-16. https://doi.org/10.1080/10106049.2019.1608594

Lambin, E. F., Turner, B. L., Geist, H. J., Agbola, S. M., Angelsen, A., Bruce, J. W., Coomes, O. T., Dirzo, R., Fischer, G., & Folke, C. (2001). The causes of land-use and land-cover change: Moving beyond the myths. *Global Environmental Change, 11*(4), 261-269. https://doi.org/10.1016/S0959-3780(01)00007-3

Langat, P. K., Kumar, L., Koech, R., & Ghosh, M. K. (2019). Monitoring of land use/land-cover dynamics using remote sensing: A case of Tana River Basin, Kenya. *Geocarto International, 36*(13), 1470-1488. https://doi.org/10.1080/10106049.2019.1655798

Liu, J., Coomes, D. A., Gibson, L., Hu, G., Liu, J., Luo, Y., Wu, C., & Yu, M. (2019). Forest fragmentation in China and its effect on biodiversity. *Biological Reviews, 94*(5), 1636-1657. https://doi.org/10.1111/brv.12519

Liu, J., Xu, Q., Yi, J., & Huang, X. (2022). Analysis of the heterogeneity of urban expansion landscape patterns and driving factors based on a combined multi-order adjacency index and Geodetector model. *Ecological Indicators, 136*, Article 108655. https://doi.org/10.1016/j.ecolind.2022.108655

Mahata, B., Sahu, S. S., Sardar, A., Laxmikanta, R., & Maity, M. (2024). Spatiotemporal dynamics of land use/land cover (LULC) changes and their impact on land surface temperature: A case study in New Town Kolkata, eastern India. *Regional Sustainability, 5*(2), Article 100138. https://doi.org/10.1016/j.regsus.2024.100138

Manjunatha, M. C., & Basavarajappa, H. T. (2020). Mapping of land units and its change detection analysis in Chitradurga Taluk of Karnataka State, India, using geospatial technology. *International Advanced Research Journal in Science, Engineering and Technology, 7*(7), 61-68. https://doi.org/10.17148/IARJSET.2020.7711

Mateen, S., Nuthammachot, N., Techato, K., & Ullah, N. (2023). Billion Tree Tsunami forests classification using image fusion technique and random forest classifier applied to Sentinel-2 and Landsat-8 images: A case study of Garhi Chandan, Pakistan. *ISPRS International Journal of Geo-Information, 12*(1), Article 9.

McHugh, M. L. (2012). Interrater reliability: The kappa statistics. *Biochemia Medica, 22*(3), 276-282.https://doi.org/10.11613/BM.2012.031

Mukhopadhyay, S. C., & Dasgupta, A. (2010). *River dynamics of West Bengal: Physical aspects* (Vol. 1).

Nilsson, N. J. (1965). *Learning machines: Foundations of trainable pattern-classifying systems*. McGraw-Hill Book Company.

Ogato, G. S., Bantider, A., Abebe, K., & Geneletti, D. (2020). Geographic information system (GIS)-based multicriteria analysis of flooding hazard and risk in Ambo Town and its watershed, West Shoa Zone, Oromia Regional State, Ethiopia. *Journal of Hydrology: Regional Studies, 27*, Article 100659. https://doi.org/10.1016/j.ejrh.2019.100659

Phiri, D., Morgenroth, J., Xu, C., & Hermosilla, T. (2018). Effects of preprocessing methods on Landsat OLI-8 land cover classification using OBIA and random forests classifier. *International Journal of Applied Earth Observation and Geoinformation, 73*, 170-178. https://doi.org/10.1016/j.jag.2018.06.008

Raczko, E., & Zagajewski, B. (2017). Comparison of support vector machine, random forest and neural network classifiers for tree species classification on airborne hyperspectral APEX images. *European Journal of Remote Sensing, 50*(1), 144-154. https://doi.org/10.1080/22797254.2017.1299557

Ramachandra, T. V., & Bharath, H. A. (2012). Spatiotemporal pattern of landscape dynamics in Shimoga, Tier II City, Karnataka State, India. *International Journal of Emerging Technology and Advanced Engineering, 2*(9), 563-576.

Ramachandra, T. V., & Kumar, U. (2011). Characterisation of landscape with forest fragmentation dynamics. *Journal of Geographic Information Systems, 3*(3), 242-252. https://doi.org/10.4236/jgis.2011.33021

Riitters, K. H., Coulston, J. W., & Wickham, J. D. (2012). Fragmentation of forest communities in the eastern United States. *Forest Ecology and Management, 263*, 85-93. https://doi.org/10.1016/j.foreco.2011.09.022

Rozenstein, O., Qin, Z., Derimian, Y., & Karnieli, A. (2014). Derivation of land surface temperature for Landsat-8 TIRS using a split-window algorithm. *Sensors, 14*(4), 5768-5780. https://doi.org/10.3390/s140405768

Ruidas, D., Pal, S. C., Towfiqul Islam, A. R. M., & Saha, A. (2023). Hydrogeochemical evaluation of groundwater aquifers and associated health hazard risk mapping using ensemble data-driven model in a water-scarce plateau region of eastern India. *Exposure and Health, 15*(1), 113-131. https://doi.org/10.1007/s12403-022-00480-6

Shafi, A., Chen, S., Waleed, M., & Sajjad, M. (2023). Leveraging machine learning and remote sensing to monitor long-term spatiotemporal wetland changes: Towards a national Ramsar inventory in Pakistan. *Applied Geography, 151*, Article 102868. https://doi.org/10.1016/j.apgeog.2022.102868

Tadesse, D., Suryabhagavan, K. V., Nedaw, D., & Hailu, B. T. (2022). A model-based flood hazard mapping in Itang District of the Gambella Region, Ethiopia. *Geology, Ecology, and Landscapes*. https://doi.org/10.1080/24749508.2021.2022833

Tripathi, V., Mohanty, M. P., & Singh, H. (2023). Fidelity of machine learning models in capturing flood inundation through geomorphic descriptors over Ganga Sub-basin [Conference paper]. *12th World Congress on Water Resources and Environment (EWRA 2023)*, Thessaloniki, Greece.

Twisa, S., & Buchroithner, M. F. (2019). Land-use and land-cover (LULC) change detection in Wami River Basin, Tanzania. *Land, 8*(9), Article 136. https://doi.org/10.3390/land8090136

Ullah, S., Khan, M., & Qiao, X. (2024a). Examining the impact of land use and land cover changes on land surface temperature in Herat City using machine learning algorithms. *GeoJournal, 89*(5), Article 225. https://doi.org/10.1007/s10708-024-11217-0

Ullah, S., Qiao, X., & Abbas, M. (2024b). Addressing the impact of land use land cover changes on land surface temperature using machine learning algorithms. *Scientific Reports, 14*(1), Article 18746. https://doi.org/10.1038/s41598-024-68492-7

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer. https://doi.org/10.1007/978-1-4757-3264-1

Yousefi, S., Mirzaee, S., Almohamad, H., Al Dughairi, A. A., Gomez, C., Siamian, N., Alrasheedi, M., & Abdo, H. G. (2022). Image classification and land cover mapping using Sentinel-2 imagery: Optimisation of SVM parameters. *Land, 11*(7), Article 993. https://doi.org/10.3390/land11070993

Zenouzi, S. A., Yenneti, K., Teimouri, R., Abbasiyan, F., & Palme, M. (2022). Analysis of changes in vegetation index during the rapid urban spatial development period (1990-2020) in Tehran Metropolis, Iran. *Atmosphere, 13*(12), Article 2010. https://doi.org/10.3390/atmos13122010

Pertanika J. Sci. & Technol. 34 (3): 1933 - 1951 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Multi-task Deep Learning Pipeline for Rice Field Classification and Growth Monitoring Using Drone Imagery

Youssef Yasser Salaheldin Elhammamy[1], Chung Gwo Chin[2]*, Gan Ming Tao[3], Chan Kah Yoong[4], and Pang Wai Leong[5]

[1]Department of Information Technology, Nityo Infotech Services Sdn. Bhd., Q Sentral, No. 2A, Jalan Stesen Sentral 2, Kuala Lumpur Sentral, 50470 Kuala Lumpur, Malaysia
[2]Department of Electrical and Electronic Engineering, Centre of Excellence for Intelligent Networks, Faculty of Artificial Intelligence and Engineering, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia
[3]Department of Artificial Intelligence, Centre of Excellence for Artificial Intelligence, Faculty of Artificial Intelligence and Engineering, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia
[4]Department of Electrical and Electronic Engineering, Centre of Excellence for Robotics and Sensing Technologies, Faculty of Artificial Intelligence and Engineering, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia
[5]Department of Electrical and Electronic Engineering, School of Engineering, Faculty of Innovation and Technology, Taylor's University, 47500 Subang Jaya, Selangor, Malaysia

ABSTRACT

A booming population around the world raises the concern of shortages of food resources in this new era. Thus, monitoring and managing crop production is extremely essential, especially rice crops, as they are the fundamental food source for most countries. Several challenges need to be addressed in this case, such as the classification of farmland from various land usages, precise monitoring of rice seedlings, and segmentation of rice growth. By leveraging advanced technologies such as drone imagery and machine learning, this paper proposed a new integrated pipeline for rice field classification and growth monitoring: a combination of convolutional neural networks (CNNs), You Only Look Once (YOLO), and modified U-Net models. These models were used in stages, specifically for paddy field classification, rice seedling detection, and rice growth segmentation. Substantial measurements and analysis have been carried out to verify the performance of the proposed system, including an accuracy of at least 85%, low classification/segmentation loss below 0.35, and high detection recall above 0.9. Thus, the findings highlight how combining different machine learning models with aerial photography can revolutionise conventional farming methods for better efficacy.

Keywords: Convolutional neural network, deep learning, drone, modified U-Net, rice field classification, YOLOv8

ARTICLE INFO

Article history:
Received: 08 June 2025
Accepted: 20 May 2026
Published: 25 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.27

E-mail addresses:
yousefyassersalah@gmail.com (Youssef Yasser Salaheldin Elhammamy)
gcchung@mmu.edu.my (Chung Gwo Chin)
mtgan@mmu.edu.my (Gan Ming Tao)
kychan@mmu.edu.my (Chan Kah Yoong)
waileong.pang@taylors.edu.my (Pang Wai Leong)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

INTRODUCTION

With a rapidly changing climate and a fast-growing global population, the presence of food security and agricultural sustainability is highly demanded in this new era. Among other crops, rice stands out as a staple food, especially in developing nations like Malaysia. According to the statistics provided by the Food and Agriculture Organisation of the United Nations (FAOSTAT), rice is one of the main crops grown in the world (Food and Agriculture Organisation of the United Nations, n.d.) and is gradually increasing yearly, underscoring its importance and the need for better management practices. However, the task of monitoring and managing the rice crop presents major obstacles, especially considering the vast expanses of land dedicated to its cultivation, in addition to the reliance on huge manpower and time usage (Sinha & Dhanalakshmi, 2022). The first challenge is the categorisation of land use, particularly differentiating between regions with rice plantations and other uses. The complexity of this undertaking stems from the variations in spectral fingerprints resulting from various farming techniques and geographical features. Local knowledge, like crop management and phonological details, greatly influences the subtleties of these spectral signatures. The second challenge is the precise monitoring of rice seedlings. Although comprehensive imagery can be obtained from satellite and high-resolution optical data, its implementation is hindered by inaccuracies in the early phases of crop growth and the challenge of mapping crops across wide areas. The third challenge is segmenting rice crops' growth stages. Effective segmentation aids in tracking crop development and optimising agricultural interventions. However, limited motivations have been found in precise farming for this purpose.

To overcome these challenges, a modern automated system that does not require more manpower and setup costs is a better solution. Hence, the development of drone technology has opened new avenues for agricultural study and surveillance (Velusamy et al., 2021). The cost of installing drone equipment is much lower than that of satellite and surveillance cameras on the farmland, especially in rural areas. Crop pictures are also more adaptable and easier to capture than optical data. Drone imagery offers an airborne perspective on crop fields that are difficult to observe from the ground. This form of application is known as photogrammetry (Jiménez-Jiménez et al., 2021), and it appears to be a better way to obtain image data for precision agriculture. Photogrammetry has been demonstrated in research to be effective in monitoring agricultural fields over time with high spatial and temporal resolution. For instance, researchers in Costa Rica employed it to observe bean crop advancement and degradation with accuracies in the cm range (Arriola-Valverde et al., 2020). Another paper has used this technology too to create a 3D map of fruit crop canopies for precision input control (Sinha et al., 2022).

Despite the capability of drone imagery to deliver high-resolution crop images from multiple angles, it requires sophisticated image processing, including computer vision,

to accurately extract significant features from the crops due to the intricate structures of the geographical terrain and vegetation (Velusamy et al., 2021). Hence, machine learning has been proposed to mitigate this problem. A comprehensive review was presented in (Parashar et al., 2024) to analyse the effectiveness of various input parameters considered in crop yield prediction models. Convolutional neural networks (CNNs) are one of the simplest and most efficient machine learning models for performing landscape classification. For example, Pan et al. (2020) suggested a CNN-based multi-spectral Light Detection and Ranging (LiDAR) land-cover classification framework. This system incorporated pre-processing of multi-spectral 3D LiDAR data into 2D pictures, reducing the number of layers utilised in the CNN model while maintaining good performance. Another approach using CNN, as reported by (Kussul et al., 2017), achieved an accuracy of more than 85% for agricultural land-use and crop classification. In addition, Hsieh and Kiang (2020) developed different versions of CNNs to categorise hyperspectral images (HSIs) of crop and mixed vegetation agricultural fields. The CNN model achieved the maximum accuracy with only a one-dimensional (1D) CNN, demonstrating the efficacy of CNN in picture categorisation. Meanwhile, a CNN variant (AlexNet) was also presented in (Ajayi et al., 2024), and its performance outperformed conventional CNN for mixed-crop classification.

Aside from that, machine learning has demonstrated amazing success in detecting or identifying objects in pictures for precision agriculture. CNN and the You Only Look Once (YOLO) algorithm are the most widely used machine learning models for object recognition. In (Vishnoi et al., 2022), a CNN model was claimed to be well-suited to identify apple leaf diseases and achieved 98% classification accuracy. On the other hand, a survey was conducted in (Badgujar et al., 2024), demonstrating how YOLO provides real-time detection performance with high accuracy in a variety of agricultural jobs such as monitoring, surveillance, sensing, automation, and robotics operations. Meanwhile, a unique weed detection method using YOLOv5 was described in (Alrowais et al., 2022) by allowing Internet of Things (IoT) devices to record farm photographs and transfer them to a cloud server for model training. Furthermore, YOLOv4 with Mish activation was implemented for rice seedling detection, counting, and location labelling, as documented in (Yeh et al., 2024), which was effective for small-object detection in UAV imagery. More variants of YOLOs (YOLOv8n, YOLOv9t, and YOLOv10n) were presented in (Chen et al., 2025) to analyse the performance of robust seedlings' detection in paddy fields.

In addition, more advanced machine learning models, such as the U-Net architecture, can address the difficulty of crop segmentation using pictures. U-Net is a deep learning model that is widely used in medical image segmentation because it can extract very exact characteristics from an image using several convolutional layers (Siddique et al., 2021).

Hence, it has also been used in agriculture for various purposes, such as plant disease detection and growth stage identification. An improved U-Net (MU-Net) was constructed for plant diseased leaf image segmentation (Zhang & Zhang, 2023). Another paper has also applied the U-Net for the segmentation of wheat farmland using drone images (Liu et al., 2022). The study showed that the modified U-Net achieved significantly better segmentation results compared to classical models.

Finally, the Internet of Things (IoT) is another latest technology integrated with machine learning for sustainable precision agriculture. As reported by Ramu et al. (2026), a lightweight CNN, long short-term memory (LSTM) model, along with whale optimisation algorithm (WOA), was designed to categorise the health status of crops, optimise irrigation and spreading of fertilisers, and enhance sustainability performance based on spatiotemporal field information. Another approach proposed a machine learning-based fertiliser recommendation methodology according to the real-time soil fertility context captured through IoT-assisted soil fertility mapping to improve the accuracy of the fertiliser recommendation system (Khan et al., 2022a). This study has been extended to include context-aware monthly evapotranspiration (ET) estimations for saline soil reclamation with the effective leaching process using LSTM models (Khan et al., 2022b) and to determine the raw reference evapotranspiration (Eto) with only temperature and optional optimisation of the raw ETo according to the availability using an artificial neural network (ANN) model (Bashir et al., 2023).

Therefore, machine learning has become a key technique in the fast-developing field of precision farming, especially when integrated with drone imagery. Table 1 summarises the benefits of machine learning above conventional approaches (without machine learning) from earlier research. This paper aims to analyse the performance of machine learning-based rice field classification and growth monitoring using drone-captured imagery. The proposed system integrated a new deep learning-based pipeline into three stages of rice field analysis: paddy field classification, rice seedling detection, and rice growth segmentation. In the first stage, a CNN model is applied to distinguish between rice plantation land and other usages of land. In the second stage, a YOLOv8 model is used to detect and identify individual rice seedlings after the paddy field has been recognised. In the last stage, a modified U-Net model is utilised to determine various growth stages of rice crops. Finally, the performance of these models is then evaluated extensively to measure the accuracy, consistency, and reliability of the system. With the contribution of this research, rice crop monitoring can be gradually transformed into automated management, resulting in better decision-making and sustainable farming methods.

Table 1
Comparison of machine learning and traditional methods

Aspects	Machine Learning Techniques	Traditional Methods
Accuracy and Precision	Higher accuracy and precision in classifying elements, ability to learn complex data patterns.	Lower accuracy and precision may miss complex patterns.
Complex Visual Data	Excellent at processing complex visual data, including variability in lighting and crop density.	Struggle with complex visual data variability.
Automation and Scalability	Can automate tasks and process large datasets quickly once trained, scalable for extensive agricultural areas.	Manual or less efficient automated processes are not as scalable.
Adaptability	Models can be updated with new data to refine predictions and adapt to changes, enhancing accuracy and relevance.	May require significant adjustments to handle new data types or improve accuracy.
Multi-task Learning	Simultaneous multi-task analysis (classification, segmentation, detection), providing comprehensive agricultural insights.	Often single-task focused, requiring separate analyses for different tasks.
Reduction in Manual Labour and Subjectivity	Reduces reliance on manual labour and minimises subjectivity, leading to more objective outcomes.	Rely more on manual surveying and interpretation, subject to human error and bias.

METHODOLOGY

Figure 1 describes the system flow of the proposed methodology. It began with the collection of drone images of rice crops. The selected paddy field was at Sekinchan, Malaysia, due to its well-known and widespread rice plantation. There were nearly 1000 images collected online freely (Roboflow) or captured using the DJI Mavic Air 2 drone during the seedling period in February 2024 and the transplanting period in March 2024. These images have a maximum size of 1280 × 720 pixels. The altitude of the drone was maintained at around 10m above the ground to obtain roughly a ground solution distance (GSD) of 5 cm/pixel. The images were captured at different spots on different days for 2 months (the actual date varied due to the need for clear weather). However, only 250 images (around 125 from online and 125 from the drone) were selected at several view angles for the rice crops' detection after pre-processing of unwanted and low-quality images. The distribution of images obtained from drones and online sources was comparable under imaging conditions, such as resolution, lighting, and type of paddies. Although the online images did not have the altitude specification, image filtering was eventually done to match the size of the paddy with the drone imagery. The selected images were then resized to a resolution of 416 × 416 pixels, which is important for the machine learning models to train the dataset efficiently and quickly for all stages. Figure 2 shows some samples of the image dataset. To further reduce noise, open-source software named LabelMe was used to perform image filtering and processing, such as image masking and augmentation. By using augmentation approaches, the machine learning models are more resistant to overfitting and are trained on a variety of data circumstances. In this case, alterations such as rotation, flipping, scaling, brightness

adjustment, noise injection, etc., generated modified versions of the original images while preserving their class label.

After that, the drone images were split into a 60:20:20 ratio (training data, validation data, and test data). The data size was limited due to time limitations and environmental factors, such as cultivation events that have already begun in April 2024 in the study field. However, data augmentation was applied to enrich the dataset and generate slightly altered versions of images for each epoch during the training process. The proposed method divided the rice field analysis into three continuous stages: paddy field classification, rice seedling detection, and rice growth segmentation. Different machine learning models were proposed at different stages based on their capability and suitability. At the first stage, a CNN model was used. All 250 images were used in this case (200 paddy fields and 50 non-paddy fields). The classified paddy land was then trained with the YOLOv8 model for rice seedling detection. The last stage was to determine the rice growth using a modified U-Net model. For the last 2 stages, only 200 images with paddy fields were used. All models utilised the same dataset for the purpose of fair performance comparison. They were trained using a computer with an Intel Core i7 processor and 16 GB RAM, and Python 3.10. A performance analysis was then conducted to determine the effectiveness of these models on rice crop detection, which includes training and validation loss, distributional focal loss, mean average precision (mAP), and recall. The next part goes into detail about the proposed steps, which are based on three different machine learning models in this study.

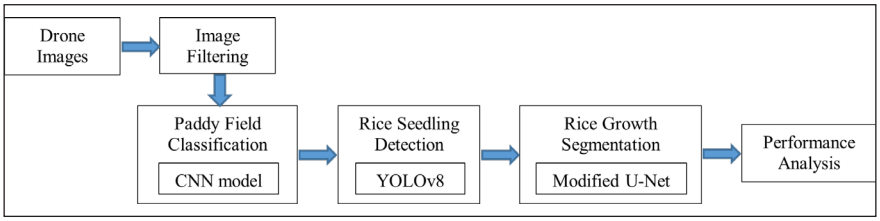


Figure 1. Block diagram

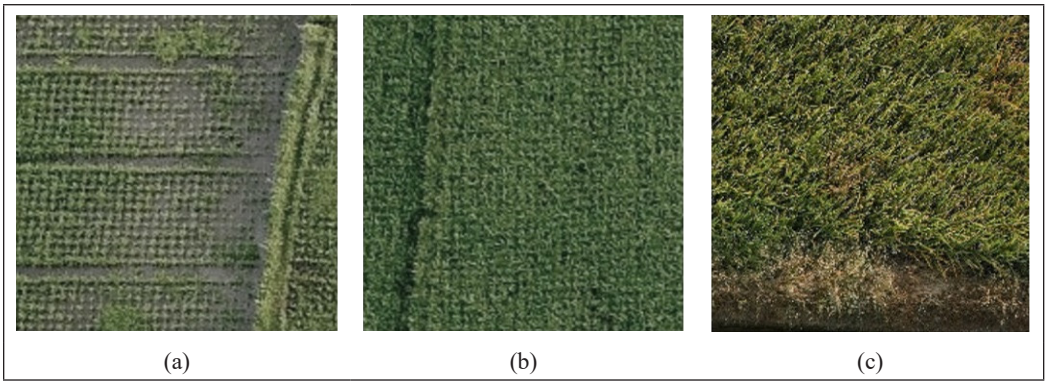


Figure 2. Samples of the image dataset obtained from (a) a drone; (b) online; and (c) online

Rice Field Classification

The first stage is to distinguish rice crops from other uses of land. The CNN model was used as the training model for this classification because of its accuracy and simplicity (Milosevic, 2020). Binary classification was implemented, which consisted of rice field and non-rice field land. Figure 3 shows the architecture of the CNN used to create the classification model in this paper.

The architecture started with an input layer, which supplied raw image data (dimensions of 416 × 416 pixels) into the network. It comprised five convolutional layers with 64, 128, 256, 512, and 512 filters, respectively, each followed by ReLU activation and max-pooling operations. These layers extracted numerous features from pictures. The feature extraction was then followed by two fully connected layers containing 4096 neurons each and a final output layer with two neurons for binary classification. It was expressly decided to employ a limited number of epochs of less than 10 and a dropout rate of 0.5 after each fully connected layer during the training of the model with the intention of avoiding overfitting, a common problem where a model learns too much from the training set, including noise and outliers. Adam was chosen as the optimiser due to its adaptive learning rate, which adjusted automatically during training to achieve faster convergence. To balance training stability and convergence speed, the learning rate was set to 0.001, the batch size was selected as 8, and the early stopping criterion was set to 5 epochs, allowing the model's accuracy to progressively rise over time.

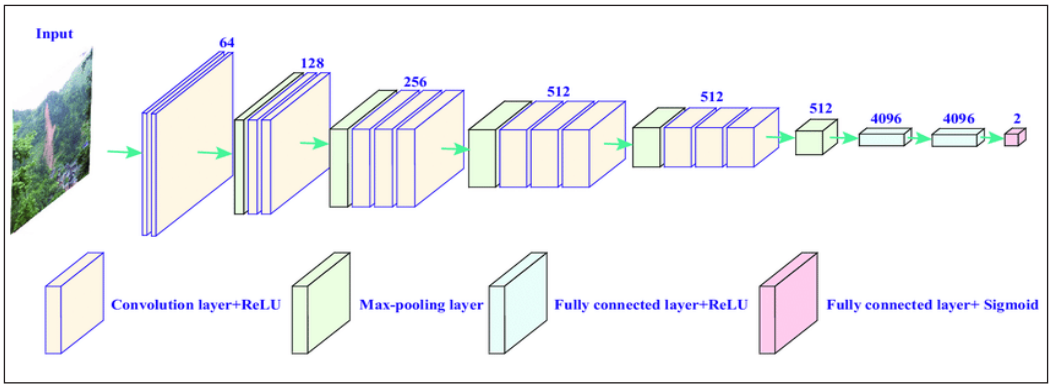


Figure 3. CNN architecture of the classification model

Rice Seedling Detection

Individual seedlings were found and identified in the paddy field during the second stage. Object detection stands as a crucial component in the realm of precision farming, enabling the automated analysis of vast crop fields with speed and accuracy.

As a result, a well-known object recognition deep learning model, YOLOv8, was chosen because it can identify objects in pictures with a variety of backdrops. The architecture of YOLOv8 is very complicated (Jiang et al., 2022) but generally can be described as a series of convolutional layers with batch normalisation and Leaky ReLU activation functions. The real object detection happened in the detection layers at the end of the architecture, in which the bounding boxes, object classes, and confidence ratings for each detected object in the image were predicted using the processed feature maps. The proposed model employs a sophisticated ensemble of layers and approaches designed expressly to tackle the hard problem of detecting objects inside a context, balancing computing efficiency and accuracy in detection. Finally, the dataset for this stage contained 4,500 annotated rice seedlings distributed across 250 drone images, with an average of 21 seedlings per image. Seedling sizes ranged from approximately 15 × 15 to 65 × 65 pixels. Most seedlings were clearly visible, although moderate crowding and partial occlusion were observed in densely planted regions.

Rice Growth Segmentation

A more complex machine learning model is required to identify each developing stage of the rice plant, from pre-booting to maturity in the third stage. As a result, the deep learning model modified U-Net was selected for this task. The U-Net model's architecture is also a convolutional network, with a contracting path to record context and a symmetric expanding path to allow for precise picture localisation. The modified U-Net used to create the segmentation model in this paper was acquired from Siddique et al. (2021). The architecture started with an input layer that received drone photos, which were typically in RGB format. It then went through a few convolutional layers to extract features at different resolutions. These convolutional layers used filters of various sizes to capture a variety of features, with the number of filters growing as the network progressed deeper, as shown by the expanding breadth of the layers Siddique et al. (2021). Activation functions such as ReLU were used after each convolutional process to add non-linearity to the model. To better meet the requirements of the agricultural imagery being analysed, parameter adjustment was made, as reported later in the result section, to the modified U-Net architecture, guaranteeing that the model can segment the images with high precision and effectiveness.

RESULTS AND DISCUSSION

Paddy Field Classification

Figure 4 presents the trajectory of loss values for the training and validation sets across the epochs during the model's learning process. It is a metric used to measure how well the model fits the training data. The training loss is traced by the blue line,

commencing at a higher value and indicating the model's initial inaccuracy in predicting the training data. As the epochs progress, this line shows a steep descent, reflecting a rapid improvement in the model's performance on the training data. Simultaneously, the validation loss, represented by the orange line, begins its course slightly above the training loss. The proximity of the two lines suggests that the model is achieving a similar level of performance on both the training and unseen validation data, which is a desirable outcome indicating good generalisation. As the epochs advance, both lines gradually level off and approach a state of minimal change, indicating that the model has reached a stable state. This plateau signals that the model has learned the dominant patterns in the data to a satisfactory degree and that additional training is unlikely to result in meaningful improvements.

Besides that, Figure 5 shows the binary precision-recall curve of the image classification model for the training set. The graph is designed to evaluate the trade-off between precision, the accuracy of positive predictions, and recall, the ability to identify all the positive instances. The area under the curve (AUC) is annotated as being 1.00, which is the maximum possible value, indicating a perfect precision-recall balance. The curve itself hugs the top right corner of the plot, reflecting an exceptional scenario where both precision and recall metrics are at their optimal level of 1.0 across all thresholds. This means that the model succeeds in identifying all positive instances (recall of 1.0) without any false positives (precision of 1.0), a rare and ideal outcome in practice. To avoid chances of overfitting, data augmentation has been applied during the training process. This graph verifies that the model is exceptionally well-tuned for the binary classification task at hand.

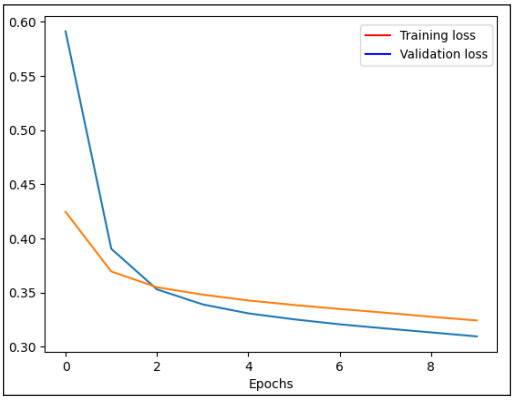


Figure 4. Training and validation loss for the classification model

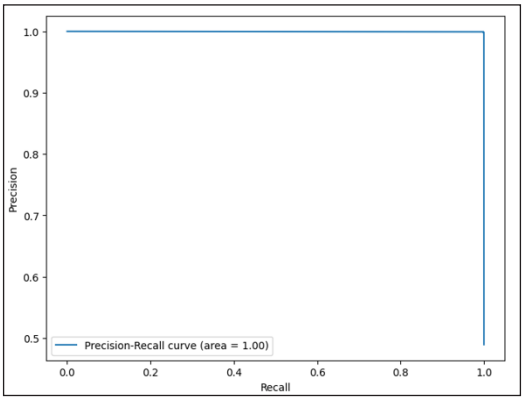


Figure 5. Binary precision-recall curve for classification model

Rice Seedling Detection

Figure 6 depicts the training and validation box loss for the YOLOv8 model used in individual rice seedling detection in this paper. The training box loss, or the difference between the predicted bounding box locations and the ground truth during the training phase, is displayed on the graph on the left. There is a noticeable decreasing trend, which suggests that as training went on, the model's accuracy in predicting the location of rice seedlings improved dramatically. On the other hand, the validation box loss, which shows how well the model's predictions on unseen data match the actual locations, is shown on the graph on the right. The overall trend is downward, despite some noticeable swings, which could be attributable to the inherent instability in the validation data. The decrease in loss for both training and validation highlights the model's learning efficacy, while the ongoing low loss on validation data indicates the model's robustness and potential for real-world applications.

In addition, Figure 7 shows the training and validation classification loss for the YOLOv8 used to detect individual rice seedlings. The classification loss is the model's error in correctly classifying the objects within the bounding boxes during training. The training classification loss (on the left) decreases significantly, as the model rapidly reduces its error rate in the initial epoch. This suggests that the model's predictions on the training data are becoming more accurate with time. The validation classification loss (on the right) has a general decrease in loss but with more pronounced spikes. However, the downward trends in both training and validation classification losses show that the model is successfully learning to classify the rice seedlings within the drone images. The sporadic spikes in the validation loss highlight how crucial it is to use a representative and diverse training dataset to guarantee the model's resilience in an actual agricultural environment.

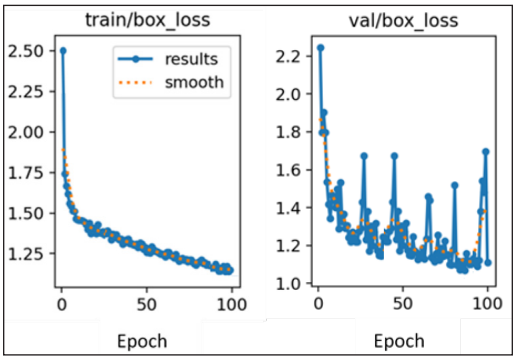


Figure 6. Training and validation box loss

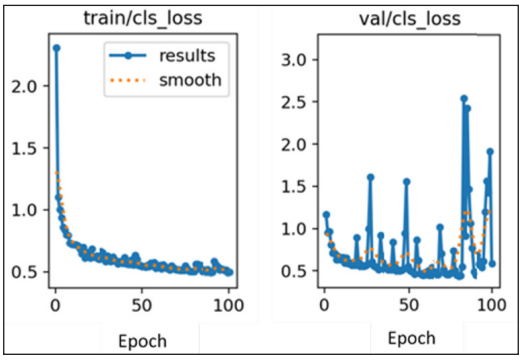


Figure 7. Training and validation classification loss

Besides that, Figure 8 represents the training and validation losses using the distributional focal loss (dfl_loss) function in a machine learning model. The dfl_loss is a part of the standard training framework in YOLOv8. It is particularly relevant when dealing with imbalanced datasets, as it helps to focus the model's learning on harder-to-classify examples, aiming to reduce the imbalance by adjusting the focus towards difficult or misclassified cases within the training process. In the training loss graph, we observe a sharp decline in loss initially, with the model quickly learning and improving its predictions on the training data. The loss values stabilise close to 1.0 as the epochs increase, indicating that the model has effectively learned the primary features and patterns necessary for the task at hand. The validation loss graph shows more variability, which is typical when the model is exposed to previously unseen data. Despite the fluctuations, there is a general downward trend in loss values, suggesting that the model can generalise the patterns it has learned to new data. However, the spikes in validation loss also highlight instances where the model struggles with certain validation examples, which could be areas where the model's predictions are less certain.

In addition, Figure 9 presents two key performance metrics for the YOLOv8 model: mAP at 50% IoU threshold (mAP50) on the left and mAP across IoU thresholds from 50% to 95% (mAP50-95) on the right. The mAP50 metric shown on the left indicates the model's precision at a specific intersection over union (IoU) threshold, a common benchmark for object detection models, where a 50% overlap with the ground truth is considered a correct detection. The model's mAP50 is predominantly high, staying close to 1.0 through most of the training epochs. This tells us that the model is highly accurate when a moderate overlap with the true object is considered sufficient. The mAP50-95 on the right averages the precision across a range of stricter IoU thresholds, from 50% up to 95%. This is a more rigorous evaluation of model performance, as it requires the model's predicted bounding boxes to closely match the ground truth. While the values here are naturally lower due to the increased difficulty, they still show the model performing well, with most values above 0.5. In both metrics, we observe occasional dips, which suggest moments where the model's performance on the validation set fluctuated, possibly due to more challenging data points or anomalies within the dataset.

Other than that, Figure 10 captures the recall metric for the YOLOv8 model across training epochs. Recall is an important measure in object detection scenarios, reflecting the model's ability to identify all relevant objects within the image. The graph indicates that the model's recall remains consistently high, predominantly above 0.9, which means the model is correctly identifying over 90% of the actual rice seedlings present in the images. This level of recall is impressive and suggests that the model is highly sensitive to the presence of seedlings, with few being missed.

However, rare severe drops in recall imply that the model failed to recognise some seedlings in specific epochs. Such changes can be attributable to variations in the complexity of the images in the validation set or slight quirks in the learning process at a few locations. The model's ability to sustain high memory is critical for practical applications, ensuring that fewer seedlings are overlooked, which could be especially useful for complete field study and yield estimation.

Finally, Figure 11 shows an example of detecting individual rice seedlings in a paddy area. Every seedling detected by the model is encompassed in a red bounding box, labelled "RiceSeedling" and accompanied by a confidence score indicating the model's level of conviction. The confidence scores fluctuate, showing that the model is more confident in certain seedlings than others. These scores are most likely dependent on the features that the model learns during training, such as shape, colour, and texture. The model's ability to detect and categorise items in drone images as rice seedlings is demonstrated graphically, and this ability may be useful for agricultural activities such as counting rice seedlings to assess planting success or production estimates.

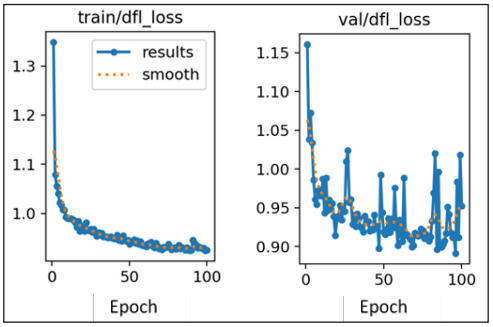


Figure 8. Training and validation of distributional focal loss

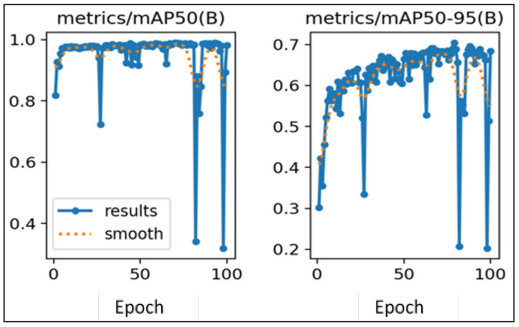


Figure 9. mAP at different IoU thresholds for seedling detection

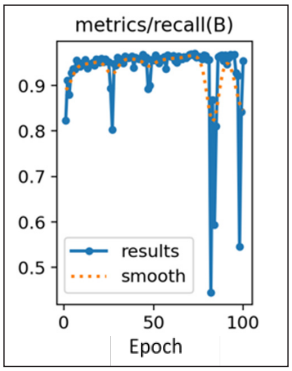


Figure 10. Recall for seedling detection

Figure 11. Example of rice seedling detection

Rice Growth Segmentation

In Figure 12, the loss values for the segmentation model during training are plotted across 10 epochs, applying an Adam optimiser with a learning rate of 0.001. It is observed that both training loss and validation loss decrease in the initial epochs, with the training loss showing a steeper decline than validation loss before gradually plateauing. This convergence of training and validation loss is indicative of a model that is learning effectively without overfitting significantly, as overfitting would typically be represented by a decrease in training loss but an increase in validation loss as epochs progress. The relatively stable and low values of validation loss in later epochs suggest that the model generalises well to new data. However, both training and validation final loss values are still over 0.25, which is considerably high and leads to a lower accuracy. To improve the loss performance, a lower learning rate of 0.0001 is employed. The training and validation losses are plotted over an extended period of 100 epochs as shown in Figure 13. From observation, the loss for both training and validation swiftly declines and levels out, with the training loss slightly lower than the validation loss. The extended training period has allowed the model to thoroughly learn from the data without fitting too closely to the noise within the training set. The reduction in the learning rate and the increased epoch count have evidently contributed to a model that performs well and is stable over time, as seen in the minimal gap between the training and validation lines and lower loss values.

Nevertheless, Figure 14 illustrates a sample output from the second segmentation model, showcasing its ability to predict the growth stages of rice crops from drone imagery. The leftmost image is the input, a high-resolution field image. The centre image is the true mask, which denotes the actual growth stages of the rice crop as identified by experts. The experts consisted of five paddy farmers in Sekinchan, Malaysia. They were assigned to select different stages of rice's growth among all the tested samples. Only those samples with 100 per cent matching would serve as true masks. The stages are colour-coded from 0 to 4, where '0' represents other non-crop elements, and stages '1' to '4' correspond to the increasing growth levels, with '4' being the harvest stage. The rightmost image is the model's predicted mask. It closely mirrors the true mask, indicating that the model can accurately identify and segment the various growth stages. The colours in the predicted mask align well with the true mask, with only minor discrepancies around the edges, reflecting the model's precision and the effectiveness of its learning from the model.

As compared to the state-of-the-art methods mentioned in the previous studies, our proposed methodology can incorporate the paddy field classification, rice seedling detection, and rice growth segmentation in a single system with considerably low prediction loss, high mAP, and high consistency in performance. Table 2 summarises the accuracy of each machine learning model used in this paper.

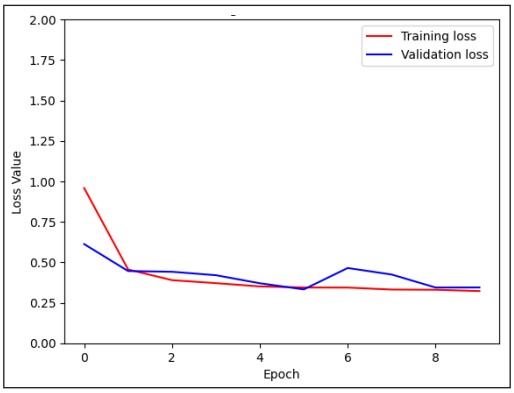


Figure 12. Training and validation loss for the first segmentation model

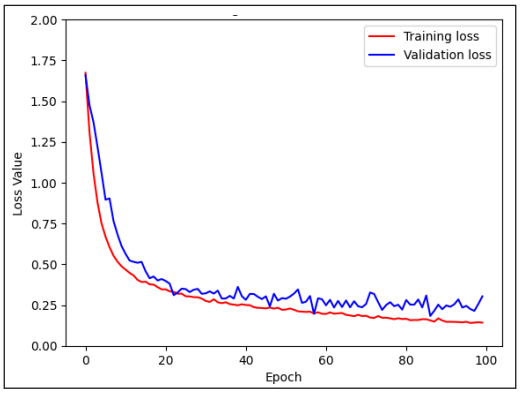


Figure 13. Training and validation loss for the second segmentation model

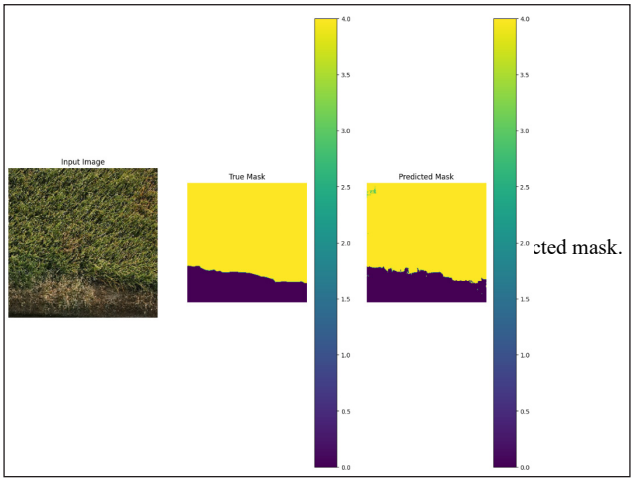


Figure 14. Sample prediction image with true mask and predicted mask

Table 2
Accuracy of the models

Stages	Accuracy
Paddy Field Classification	≈ 99%
Rice Seedling Detection	> 85%
Rice Growth Segmentation	> 90%

Discussion of Findings

Table 3 highlights that existing studies predominantly concentrate on individual tasks, including crop categorisation, seedling detection, or field segmentation. Although these methods demonstrated robust performance, their usefulness is typically confined to a particular phase of agricultural monitoring. The proposed system enhances prior research by integrating CNN-based land-use categorisation, YOLOv8-based seedling identification, and a modified U-Net for growth-stage segmentation into an interconnected UAV-based pipeline. It achieved competitive performance, with accuracies ranging from 85% to 99%, recall values above 0.90, and low classification and segmentation losses. While additional validation on bigger and more diverse datasets is necessary, the findings obtained suggest that the pr method attains comparable performance while offering a more thorough solution for rice field monitoring and precision agricultural applications.

Table 3
Benchmarking of previous and current studies

Studies	Technique Used	Application	Performance Reported	Advantages	Limitations
(Kussul et al., 2017)	CNN	Agricultural land-use and crop classification	Accuracy > 85%	Effective automatic feature learning without manual feature engineering	Require large, labelled datasets for robust performance
(Ajavi et al., 2024)	CNN (AlexNet)	Mixed-crop classification from high-resolution UAV imagery	Accuracy ≈ 72%	Outperformed conventional CNN across all tested effective automatic feature extraction from UAV images epochs;	Validation accuracy remained substantially lower than training accuracy, indicating overfitting; performance decreased after 60 epochs
(Yeh et al., 2024)	YOLOv4 with Mish activation	Rice seedling detection, counting, and location labelling from UAV imagery	F1-score = 0.91; average precision improved by 4.96% compared with the previous YOLOv4	Effective for small-object detection in UAV imagery; integrates detection, counting, and coordinate labelling in a single framework	Limited dataset (44 original UAV images); evaluation focused on a competition dataset rather than multi-location field validation

Table 3 (*continued*)

Studies	Technique Used	Application	Performance Reported	Advantages	Limitations
(Chen et al., 2025)	YOLOv8n, YOLOv9t, and YOLOv10n	Paddy seedling counting and detection from UAV imagery	mAP@50 values of 0.964, 0.936, and 0.944 for YOLOv8n, YOLOv9t, and YOLOv10n	High mAP, robust detection of seedlings; good counting reliability	Limited dataset (100-200 UAV images); performance degrades under severe occlusion and dense planting
(Liu et al., 2022)	Improved U-Net (MGA + Interaction Mechanism)	Wheat land segmentation from UAV imagery	Dice = 93.88%, better than U-Net & DeepLabV3	High segmentation accuracy; improved semantic gap	Require more computation resources; sensitive to data quality
Proposed method	CNN + YOLOv8 + Modified U-Net	Land-use classification, seedling detection, and growth-stage segmentation	Accuracy ranging from 85% to 99%, classification/ segmentation loss < 0.35, and recall > 0.9	Integrated framework addressing three rice-monitoring tasks within a single workflow	Limited dataset size and need for broader external validation

CONCLUSION

This paper successfully analyses the performance of several machine learning models used for paddy field classification, rice seedling detection, and growth stage segmentation using drone-captured images. The functionalities provided by CNN, YOLOv8, and modified U-Net models illustrate very promising performance. Due to 99% accuracy of field classification, the seedling detection was able to achieve at least 85% accuracy. Hence, the accuracy of growth segmentation has been improved to over 90%. However, there are practical challenges and limitations on real-world implementation of the proposed system, including dependency on high-quality drone imagery, environmental influences such as lighting and weather conditions, and the computational cost involved in training machine learning models. Future research might delve into fine-tuning parameters such as dfl_loss or explore its application in conjunction with other techniques to further improve model robustness and generalisation, especially in diverse agricultural settings. Nevertheless, the limited dataset size may affect the robustness of the model, and more evaluation can be done on larger and more diverse external datasets to further validate the model's generalisability. In addition, these models' flexibility would be tested, and their scalability might be revealed by including more real-time images from different regions

and expanding the dataset to cover a range of climates, crop diseases, and growth phases. By doing so, these models can be fully utilised and be instrumental in precision farming, helping farmers to make informed decisions about irrigation, fertilisation, and harvesting times.

ACKNOWLEDGEMENT

This research was not supported by any funding. We thank our colleagues from the Faculty of AI & Engineering who provided insight and expertise that greatly assisted the research, although they may not agree with all the interpretations/conclusions of this paper.

REFERENCES

Ajayi, O. G., Iwendi, E., & Adetunji, O. O. (2024). Optimising crop classification in precision agriculture using AlexNet and high-resolution UAV imagery. *Technology in Agronomy, 4*, Article e011. https://doi.org/10.48130/tia-0024-0009

Alrowais, F., Asiri, M. M., Alabdan, R., Marzouk, R., Hilal, A. M., Alkhayyat, A., & Gupta, D. (2022). Hybrid leader-based optimisation with deep learning-driven weed detection on Internet of Things-enabled smart agriculture environment. *Computers and Electrical Engineering, 104*, Article 108411. https://doi.org/10.1016/j.compeleceng.2022.108411

Arriola-Valverde, S., Villagra-Mendoza, K., Méndez-Morales, M., Solórzano-Quintana, M., Gómez-Calderón, N., & Rimolo-Donadio, R. (2020). Analysis of crop dynamics through close-range UAS photogrammetry. In *2020 IEEE International Symposium on Circuits and Systems (ISCAS)* (pp. 1-5). IEEE. https://doi.org/10.1109/ISCAS45731.2020.9181285

Badgujar, C. M., Poulose, A., & Gan, H. (2024). Agricultural object detection with You Only Look Once (YOLO) algorithm: A bibliometric and systematic literature review. *Computers and Electronics in Agriculture, 223*, Article 109090. https://doi.org/10.1016/j.compag.2024.109090

Bashir, R. N., Khan, F. A., Khan, A. A., Tausif, M., Abbas, M. Z., Shahid, M. M. A., & Khan, N. (2023). Intelligent optimisation of reference evapotranspiration (ETo) for precision irrigation. *Journal of Computational Science, 69*, Article 102025. https://doi.org/10.1016/j.jocs.2023.102025

Chen, S., Li, W., Chen, D., Xie, Z., Zhang, S., Cen, F., Tu, L., Zhao, Q., & Gao, Z. (2025). Recognition of rice seedlings count in UAV remote sensing images via the YOLO algorithm. *Smart Agricultural Technology, 12*, Article 101107. https://doi.org/10.1016/j.atech.2025.101107

Food and Agriculture Organisation of the United Nations. (n.d.). *Rice*. Retrieved July 1, 2024, from https://www.fao.org/markets-and-trade/commodities/rice/en/

Hsieh, T.-H., & Kiang, J.-F. (2020). Comparison of CNN algorithms on hyperspectral image classification in agricultural lands. *Sensors, 20*(6), Article 1734. https://doi.org/10.3390/s20061734

Jiang, P., Ergu, D., Liu, F., Cai, Y., & Ma, B. (2022). A review of YOLO algorithm developments. *Procedia Computer Science, 199*, 1066–1073. https://doi.org/10.1016/j.procs.2022.01.135

Jiménez-Jiménez, S. I., Ojeda-Bustamante, W., Marcial-Pablo, M., & Enciso, J. (2021). Digital terrain models generated with low-cost UAV photogrammetry: Methodology and accuracy. *ISPRS International Journal of Geo-Information, 10*(5), Article 285. https://doi.org/10.3390/ijgi10050285

Khan, A. A., Faheem, M., Bashir, R. N., Wechtaisong, C., & Abbas, M. Z. (2022a). Internet of Things (IoT)-assisted context-aware fertiliser recommendation. *IEEE Access, 10*, 129505-129519. https://doi.org/10.1109/ACCESS.2022.3228160

Khan, A. A., Nauman, M. A., Bashir, R. N., Jahangir, R., Alroobaea, R., Binmahfoudh, A., Alsafyani, M., & Wechtaisong, C. (2022b). Context-aware evapotranspiration (ETs) for saline soils reclamation. *IEEE Access, 10*, 110050-110063. https://doi.org/10.1109/ACCESS.2022.3206009

Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters, 14*(5), 778-782. https://doi.org/10.1109/LGRS.2017.2681128

Liu, G., Bai, L., Zhao, M., Zang, H., & Zheng, G. (2022). Segmentation of wheat farmland with an improved U-Net on drone images. *Journal of Applied Remote Sensing, 16*(3), Article 034511. https://doi.org/10.1117/1.JRS.16.034511

Milosevic, N. (2020). *Introduction to convolutional neural networks: With image classification using PyTorch* [Video course]. Apress. https://doi.org/10.1007/978-1-4842-5648-0

Pan, S., Guan, H., Chen, Y., Yu, Y., Nunes Gonçalves, W., Marcato Junior, J., & Li, J. (2020). Land-cover classification of multispectral LiDAR data using CNN with optimised hyperparameters. *ISPRS Journal of Photogrammetry and Remote Sensing, 166*, 241-254. https://doi.org/10.1016/j.isprsjprs.2020.05.022

Parashar, N., Johri, P., Khan, A., Gaur, N., & Kadry, S. (2024). An integrated analysis of yield prediction models: A comprehensive review of advancements and challenges. *Computers, Materials & Continua, 80*(1), 389-431. https://doi.org/10.32604/cmc.2024.050240

Ramu, S. C., Raman, D., Ramana, K., Krishna, A. V., Khan, A. A., Khan, S., & Abdu, S. M. (2026). A lightweight deep learning and whale optimisation framework for sustainable precision agriculture. *Discover Computing, 29*, Article 73. https://doi.org/10.1007/s10791-026-09952-8

Siddique, N., Paheding, S., Elkin, C. P., & Devabhaktuni, V. (2021). U-Net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access, 9*, 82031-82057. https://doi.org/10.1109/ACCESS.2021.3086020

Sinha, B. B., & Dhanalakshmi, R. (2022). Recent advancements and challenges of Internet of Things in smart agriculture: A survey. *Future Generation Computer Systems, 126*, 169-184. https://doi.org/10.1016/j.future.2021.08.006

Sinha, R., Quirós, J. J., Sankaran, S., & Khot, L. R. (2022). High-resolution aerial photogrammetry-based 3D mapping of fruit crop canopies for precision inputs management. *Information Processing in Agriculture, 9*(1), 11-23. https://doi.org/10.1016/j.inpa.2021.01.006

Velusamy, P., Rajendran, S., Mahendran, R. K., Naseer, S., Shafiq, M., & Choi, J.-G. (2022). Unmanned aerial vehicles (UAVs) in precision agriculture: Applications and challenges. *Energies, 15*(1), Article 217. https://doi.org/10.3390/en15010217

Vishnoi, V. K., Kumar, K., Kumar, B., Mohan, S., & Khan, A. A. (2023). Detection of apple plant diseases using leaf images through a convolutional neural network. *IEEE Access, 11*, 6594-6609. https://doi.org/10.1109/ACCESS.2022.3232917

Yeh, J. F., Lin, K. M., Yuan, L. C., & Hsu, J. M. (2024). Automatic counting and location labelling of rice seedlings from unmanned aerial vehicle images. *Electronics, 13*(2), Article 273. https://doi.org/10.3390/electronics13020273

Zhang, S., & Zhang, C. (2023). Modified U-Net for plant diseased leaf image segmentation. *Computers and Electronics in Agriculture, 204*, Article 107511. https://doi.org/10.1016/j.compag.2022.107511

REFEREES FOR THE PERTANIKA

JOURNAL OF SCIENCE & TECHNOLOGY

Vol. 34 (3) Jun. 2026

The Editorial Board of the Pertanika Journal of Science and Technology wishes to thank the following:

Abdullah Hisam Omar
(UTM, Malaysia)

Abdul Rahman Kram
(UNIMAS, Malaysia)

Abel Adeyi
(ABUAD, Nigeria)

Abou El Kacem Qaiss
(UM6P, Morocco)

Adnan Ashraf Arain
(MUET, Pakistan)

Ahmad Syakirin Ismail @ Rosdi
(UMPSA, Malaysia)

Akshat Agrawal
(AUH, India)

Ali Al-Juboori
(RCNJ, USA)

Ali Malik
(TU Dublin, Ireland)

Andri Saputra
(PATKY, Indonesia)

Arfat Ahmad Khan
(KKU, Thailand)

Arun Sharma
(UMN, USA)

Ashwan A. Abdulmunem
(UOK, Iraq)

Azimbek Khudoyberdiev
(SBU, USA)

Azizul Hakim Lahuri
(UPM, Malaysia)

Azli Abdul Razak
(UiTM, Malaysia)

Balamurugan Ponnambalam
(KARE, India)

Baobao Song
(UTS, Australia)

Charu Gupta
(BPIT, India)

Chia Kuang Lee
(UMPSA, Malaysia)

Dinna @ Nina Mohd Nizam
(UMS, Malaysia)

Ernee Noryana Muhamad
(UPM, Malaysia)

Faieza Abdul Aziz
(UPM, Malaysia)

G. N. Balaji
(VIT, India)

Gökhan Başar
(OKU, Türkiye)

Hamdan Mohamed Yusoff
(UPM, Malaysia)

Hamza Abubakar
(UUM, Malaysia)

Helena V. G. Navas
(NOVA, Portugal)

Hui Shan Lee
(UTAR, Malaysia)

Irulappasamy Siva
(MBU, India)

Joel Than Chia Ming
(SUTS, Malaysia)

K. G. Binu
(SJEC, India)

Kai Chen Goh
(UTHM, Malaysia)

Khan Vun Teong
(UMS, Malaysia)

Krishna Presannakumar
(CHRIST, India)

Leonardo José Agamez Polo
(UNISINÚ, Colombia)

Lili Nurliyana Abdullah
(UPM, Malaysia)

Lindah Roziani Jamru
(UMS, Malaysia)

M. Uthayakumar
(AMRITA, India)

Mahmoud Quweider
(UTRGV, USA)

Maizatul Hayati Mohamad Yatim
(UPSI, Malaysia)

Mine Güngörmüşler
(İEÜ, Türkiye)

Mohamed Thariq Hameed Sultan
(UPM, Malaysia)

Mohd Zaki Sharif
(UMPSA, Malaysia)

Muhammad Noor Ahmad Shukri
(UPM, Malaysia)

Muntadher Sallal
(BU, United Kingdom)

Mustafa Acaroğlu
(SU, Türkiye)

Navneet Kaur
(GNDEC, India)

Nazmi Mat Nawi
(UPM, Malaysia)

Pau Chung Leng
(UTM, Malaysia)

Pitchipoo Pandian
(PSREC, India)

Riyadh Rahef Nuiaa Alogaili
(UOW, Iraq)

Sa'adah Ahmad @ Ahmad Sowi
(IIUM, Malaysia)

Saiful Bahri Mohamed
(UniSZA, Malaysia)

Shahrul Azmi Mohd Yusof
(UUM, Malaysia)

Sheila Belayutham
(UiTM, Malaysia)

Simon Boung Yew Lau
(UNM, Malaysia)

Subba Rao Polamuri
(AU, India)

Syaza Izyanni Ahmad
(UTP, Malaysia)

Vijay Singh Rathore
(AU, India)

Waidah Ismail
(USIM, Malaysia)

Yasmin Abdul Wahab
(UMPSA, Malaysia)

Yonis Gulzar
(KFU, Saudi Arabia)

Yousef Farhaoui
(UMI, Morocco)

Zaiton Abdul Mutalip
(UTeM, Malaysia)

Zeid A. AlOthman
(KSU, Saudi Arabia)

Zuraini Dahari
(USM, Malaysia)

ABUAD	:	*Afe Babalola University*
ADU	:	*Aditya University*
AMRITA	:	*Amrita Vishwa Vidyapeetham*
APU	:	*Apex University, Jaipur*
AUH	:	*Amity University Haryana*
BPIT	:	*Bhagwan Parshuram Institute of Technology*
BU	:	*Bournemouth University*
CHRIST	:	*CHRIST (Deemed to be University)*
GNDEC	:	*Guru Nanak Dev Engineering College*
IEU	:	*İzmir University of Economics*
IIUM	:	*International Islamic University Malaysia*
KARE	:	*Kalasalingam Academy of Research and Education*
KFU	:	*King Faisal University*
KKU	:	*Khon Kaen University*
KSU	:	*King Saud University*
MBU	:	*Mohan Babu University*
MUET	:	*Mehran University of Engineering and Technology*
NOVA	:	*NOVA University Lisbon*
OKU	:	*Osmaniye Korkut Ata University*
PATKY	:	*Politeknik ATK Yogyakarta*
PSREC	:	*P.S.R. Engineering College*
RCNJ	:	*Ramapo College of New Jersey*
SBU	:	*Stony Brook University*
SJEC	:	*St Joseph Engineering College*
SU	:	*Selçuk University*
SUTS	:	*Swinburne University of Technology Sarawak Campus*
TU	:	*Dublin Technological University Dublin*
UM6P	:	*Mohammed VI Polytechnic University*
UMI	:	*Moulay Ismail University*
UMN	:	*University of Minnesota*
UMPSA	:	*Universiti Malaysia Pahang Al-Sultan Abdullah*
UMS	:	*Universiti Malaysia Sabah*
UNIMAS	:	*Universiti Malaysia Sarawak*
UNISINÚ	:	*Universidad del Sinú*
UniSZA	:	*Universiti Sultan Zainal Abidin*
UNM	:	*University of Nottingham Malaysia*
UOK	:	*University of Kerbala*
UOW	:	*University of Wasit*
UPM	:	*Universiti Putra Malaysia*
UPSI	:	*Universiti Pendidikan Sultan Idris*
USIM	:	*Universiti Sains Islam Malaysia*
USM	:	*Universiti Sains Malaysia*
UTAR	:	*Universiti Tunku Abdul Rahman*
UTeM	:	*Universiti Teknikal Malaysia Melaka*
UTHM	:	*Universiti Tun Hussein Onn Malaysia*
UTM	:	*Universiti Teknologi Malaysia*
UTP	:	*Universiti Teknologi PETRONAS*
UTRGV	:	*The University of Texas Rio Grande Valley*
UTS	:	*University of Technology Sydney*
UUM	:	*Universiti Utara Malaysia*
VIT	:	*Vellore Institute of Technology*

While every effort has been made to include a complete list of referees for the period stated above, however if any name(s) have been omitted unintentionally or spelt incorrectly, please notify the Chief Executive Editor, *Pertanika* Journals at executive_editor.pertanika@upm.edu.my

Any inclusion or exclusion of name(s) on this page does not commit the *Pertanika* Editorial Office, nor the UPM Press or the university to provide any liability for whatsoever reason.

Enhanced Deepfake Image Detection and Classification Using Deep Learning 1703
Debasish Samal, Dimple Nagpal, Prateek Agrawal, Vishu Madaan, Chhavi Sharma, and Wou Onn Choo

Occupational Health Implications of High-Frequency Whole-Body Vibration Exposure in Mine Workers 1727
S. Joel Vinoth Singh, I. Siva, JT. Winowlin Jappes, and Chithirai Pon Selvan

A LiDAR and IoT Integrated Approach for Intelligent Monitoring of Lubricant Liquid Levels: A Lab Scale Study on Petronas Products 1745
Nur Shaebah Mat Hussain, Mohd Khairol Anuar Mohd Ariffin, Siti Azfanizam Ahmad, and Zulkifli Mohamed

Human-in-the-Loop Soft Prompt Tuning for Adverse Drug Event Extraction from Clinical Notes 1769
Salisu Modi, Khairul Azhar Kasmiran, Nurfadhlina Mohd Sharef, and Mohd Yunus Sharum

Empirical Analysis of Factors in User Control Model for Cloud Data Migration 1801
Ishak Iskandar, Imbaji Injuwe Danga, Hamidah Ibrahim, and Fatimah Sidi

Towards Inherently Safer Dry Reforming for Carbon Capture: Trends, Gaps, and Insights 1829
Fatin Aliya, Mardhati Zainal Abidin, and Azizul Buang

An IoT-Enabled Decision Support System for Real-time Line Balancing in Semiconductor Manufacturing 1861
Nur Ain Qistina Muhammad Shafee, Effendi Mohamad, Mohd Soufhwee Abd Rahman, Arfauz A. Rahman, and Teruaki Ito

Review Article 1881
Optical Sensing Techniques for Microplastic Identification: Comparative Evaluation, Limitations, and Future Recommendations
Muhammad Nurullah Waliyullah Mohamed Nazli, Irneza Ismail, Fatin Hamimi Mustafa, Wan Zakiah Wan Ismail, Juliza Jamaludin, and Latiffah Karim

Quantitative Evaluation of Machine Learning-based Land Use Classification with Multispectral Satellite Data 1905
Krati Bansal, Anindita Nath, Tanupriya Choudhury, Bappaditya Koley, and Jitendra Rajpurohit

Multi-task Deep Learning Pipeline for Rice Field Classification and Growth Monitoring Using Drone Imagery 1933
Youssef Yasser Salaheldin Elhammamy, Chung Gwo Chin, Gan Ming Tao, Chan Kah Yoong, and Pang Wai Leong

Design and Development of Unmanned Aerial Vehicles' Propellers Based on Biomimetic Design 1451
Saundarya Saran, Rosalam Che Me, Mohamed Thariq Hameed Sultan, and Navaneetha Krishna Chandran

A Zero-shot Foundation Model Approach for Real-time Adaptive Cyber Threat Detection 1469
Muhmmad Al-Khiza'ay and Noora Alallaq

Automated Generation of Calculation-based Physics Questions Using Text-to-Text Transfer Transformer (T5) for Malaysian Upper Secondary Education 1493
Tuong Kiet Ngang and Chih How Bong

D-Xotic: A Deep Learning Improvement for Recognising Different Durian Cultivars Based on Leaves Signature 1525
Aulia Brilliantina, Tri Agus Siswoyo, Yuli Witono, and Bayu Taruna Widjaja Putra

Noise-Resilient and Low Complexity LACO-OFDM Using Approximation Log-Likelihood Ratio (approxLLR) Soft-decision Demodulation with Single-FFT Receiver 1551
Norzalina Othman and Mohd Rashidi Che Beson

Review Article 1579
A Bibliometric Review of Remote Sensing Applications in Water Surge Risk Mapping for Waterfall Areas
Mohammed M. Y. Albarqouni, Syarifuddin Misbari, Nadiatul Adilah Ahmad Abdul Ghani, and Mohd Faizal Jamlos

Carbon Emissions Quantification from Ceramic Tile Production by Using Life-Cycle Assessment (LCA): A Case Study of PT Sinar Karya Duta Abadi, Indonesia 1599
Ransi Pasae, M. Rijal Idrus, Hazairin Zubair, and Darhamsyah

Field Validation of IES-VE for Daylight Assessment in Tropical High-rise Low-cost Housing: Effects of Building Height, Layout Depth, and Façade Design 1625
Nor Izyan Saleh, Mohd Firrdhaus Mohd Sahabuddin, Norafida Ab Ghafar, Zuraina Leily Awaluddin, and Bismiazan Abd.Razak

Gamification for Extended Reality Informative Space (XRIS) Visualisation in the Construction Industry 1655
Changsaar Chai, Anne Ann Yii Chua, and Felix Kai Han Yong

Studying the Effects of Knowledge Distillation and the Fragility of Lightweight BERT-based Models 1685
Allistair Nallie Konsil, Stephanie Chua, Jacey Lynn Minoi, Md Mizanur Rahman, Gerraint Gillan, Rafazila Ramli, Rasitasam Safii, Ahmad Sofian bin Shminan, and Lee Jun Choi

Pertanika Journal of Science & Technology
Vol. 34 (3) Jun. 2026

Content

Foreword i
Mohamed Thariq Hameed Sultan

Review Article 1287
Energy Consumption and Optimisation in Indoor Farming: A Review
Jia Quan Goh, Aimrun Wayayok, Diyana Binti Jamaludin, and Muhamad Saufi Mohd Kassim

Valorisation of Oil Palm Waste into Magnetic Catalysts for Efficient Synthesis of Bio-additive Ethyl Levulinate: Optimisation of Reaction Parameters 1309
Shamala Gowri Krishnan and Pua Fei Ling

YOLOv9-ResCBAM: Enhancing Tomato Ripeness Detection in Greenhouses Through Advanced Object Detection 1331
Jalal Uddin Md Akbar, Syafiq Fauzi Kamarulzaman, Muhammad Danial Mohamad Rizwan, Riadul Islam Rabbi, and Ekramul Haque Tusher

Review Article 1355
Exploring Educational User Engagement: A Systematic Mapping of Serious Games with Gesture-based Interaction
Wan Salfarina Wan Husain, Syadiah Nor Wan Shamsuddin, and Normala Rahim

Utilising Innovative Image Preprocessing Techniques in Vehicle Plate Detection and Recognition 1383
Soheir Noori, Asam Almohamed, Elham Mohammed Thabit A. Alsaadi, and Azhar Abbas

Stability, Thermo-physical, and Tribological Properties of SiO_2 Nanoparticles in a Vegetable Oil-based Biodegradable Lubricant for Minimum Quantity Lubrication Machining 1409
Nurlisa Hamzan, Muhammad Amir Syafiq Mohd Tarmizi, Nurul Huda Abdul Azid, Ahsana Aqilah Ahmad, Afifah Zakiyyah Juri, and Irnie Azlin Zakaria

Silicon-based Fin Field-Effect Transistor (Si-FinFET) Optimisation Based on Gate Length and Operating Temperature 1433
Yousif Atalla, Mohamad Hafiz Mamat, and Yasir Hashim

Pertanika Editorial Office, Journal Division,
Putra Science Park,
1st Floor, IDEA Tower II,
UPM-MTDC Center,
Universiti Putra Malaysia,
43400 UPM Serdang,
Selangor Darul Ehsan
Malaysia

http://www.pertanika.upm.edu.my
Email: executive_editor.pertanika@upm.edu.my
Tel. No.: +603- 9769 1622

http://penerbit.upm.edu.my
Email: penerbit@upm.edu.my
Tel. No.: +603- 9769 3606

e-ISSN 2231-8526