

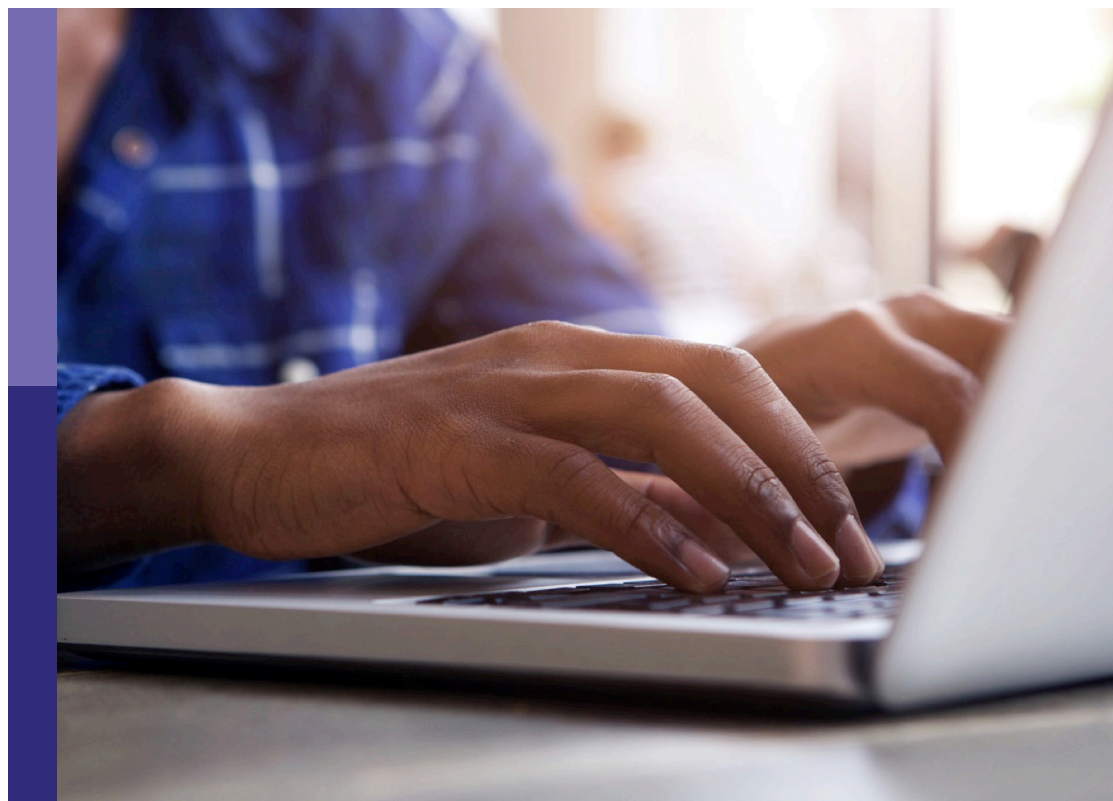
Foundation models for healthcare: innovations in generative AI, computer vision, language models, and multimodal systems

Edited by

Alberto Santamaria-Pang, Shuo Li, Sokratis Makrogiannis,
Paul Nagy and Vaz Zavaletta

Published in

Frontiers in Computer Science
Frontiers in Physiology
Frontiers in Artificial Intelligence
Frontiers in Radiology



FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-7252-8
DOI 10.3389/978-2-8325-7252-8

Generative AI statement

Any alternative text (Alt text) provided alongside figures in the articles in this ebook has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Foundation models for healthcare: innovations in generative AI, computer vision, language models, and multimodal systems

Topic editors

Alberto Santamaria-Pang — Microsoft, United States

Shuo Li — Western University, Canada

Sokratis Makrogiannis — Delaware State University, United States

Paul Nagy — Johns Hopkins University, United States

Vaz Zavaletta — University of Colorado, United States

Citation

Santamaria-Pang, A., Li, S., Makrogiannis, S., Nagy, P., Zavaletta, V., eds. (2025).

Foundation models for healthcare: innovations in generative AI, computer vision, language models, and multimodal systems. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-8325-7252-8

Table of contents

- 05 **Editorial: Foundation models for healthcare: innovations in generative AI, computer vision, language models, and multimodal systems**
Sokratis Makrogiannis
- 08 **Multimodal AI techniques for pain detection: integrating facial gesture and paralinguistic analysis**
Rommel Gutierrez, Joselin Garcia-Ortiz and William Villegas-Ch
- 22 **Multimodal MRI radiomics-based stacking ensemble learning model with automatic segmentation for prognostic prediction of HIFU ablation of uterine fibroids: a multicenter study**
Bing Wen, Chengwei Li, Qiuyi Cai, Dan Shen, Xinyi Bu and Fuqiang Zhou
- 33 **Toward improving precision and complexity of transformer-based cost-sensitive learning models for plant disease detection**
Manh-Tuan Do, Manh-Hung Ha, Duc-Chinh Nguyen and Oscar Tzyh-Chiang Chen
- 51 **Large language models generating synthetic clinical datasets: a feasibility and comparative analysis with real-world perioperative data**
Austin A. Barr, Joshua Quan, Eddie Guo and Emre Sezgin
- 61 **MoNetViT: an efficient fusion of CNN and transformer technologies for visual navigation assistance with multi query attention**
Liliek Triyono, Rahmat Gernowo and Prayitno
- 75 **Breast cancer diagnosis using radiomics-guided DL/ML model-systematic review and meta-analysis**
Nazmul Ahasan Maruf and Abdullah Basuhail
- 96 **CNN based 2D object detection techniques: a review**
Badri Raj Lamichhane, Gun Srijuntongsiri and Teerayut Horanont
- 125 **Leveraging foundation models and goal-dependent annotations for automated cell confluence assessment**
Maximilian Josef Joas, Daniel Freund, Robert Haase, Erhard Rahm and Jan Ewald
- 137 **Assessing the consistency of CT-based ventilation imaging under noise reduction processing with simulated quantum noise using a nonrigid alveoli phantom**
Shin Miyakawa, Hiraku Fuse, Kenji Yasue, Norikazu Koori, Masato Takahashi, Hiroki Nosaka, Shunsuke Moriya, Fumihiro Tomita and Tatsuya Fujisaki

- 148 **Deep learning for vision screening in resource-limited settings: development of multi-branch CNN for refractive error detection based on smartphone image**
Muhammad Syauqie, Harry Patria, Sutanto Priyo Hastono, Kemal Nazaruddin Siregar and Nila Djuwita Farieda Moeloek
- 166 **Heterogeneous ensemble learning: modified ConvNextTiny for detecting molecular expression of breast cancer on standard biomarkers**
Indo Intan, Andrea Stevens Karnyoto, Sitti Harlina, Berti Julian Nelwan, Devin Setiawan, Amalia Yamin and Ririn Endah Puspitasari
- 182 **Research on a machine learning method for predicting discharge time of thyroid cancer patients receiving ^{131}I treatment: a retrospective study**
Chen Zhang, Dandan Zhang, Lijun Tang and Feng Tian



OPEN ACCESS

EDITED AND REVIEWED BY
Marcello Pelillo,
Ca' Foscari University of Venice, Italy

*CORRESPONDENCE
Sokratis Makrogiannis
✉ smakrogiannis@desu.edu

RECEIVED 12 November 2025
REVISED 12 November 2025
ACCEPTED 17 November 2025
PUBLISHED 26 November 2025

CITATION
Makrogiannis S (2025) Editorial: Foundation
models for healthcare: innovations in
generative AI, computer vision, language
models, and multimodal systems.
Front. Comput. Sci. 7:1744581.
doi: 10.3389/fcomp.2025.1744581

COPYRIGHT
© 2025 Makrogiannis. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License \(CC
BY\)](#). The use, distribution or reproduction in
other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Editorial: Foundation models for healthcare: innovations in generative AI, computer vision, language models, and multimodal systems

Sokratis Makrogiannis*

Division of Physics, Engineering, Mathematics, and Computer Science, Delaware State University,
Dover, DE, United States

KEYWORDS

editorial, foundation models, multimodal systems, healthcare, deep learning

Editorial on the Research Topic

[Foundation models for healthcare: innovations in generative AI, computer vision, language models, and multimodal systems](#)

Foundation models—large pre-trained vision, language, and multimodal systems—are reshaping how we approach medical data, from image segmentation to synthetic clinical data generation and multimodal fusion for diagnostics. The Research Topic Foundation Models for Healthcare collects an informative cross-section of recent advances that illuminate the promise and the obstacles of applying foundation models to real-world clinical problems. This editorial synthesizes the Research Topic's contributions, highlights emergent lessons, and outlines priorities for translating foundation models safely and effectively into healthcare practice.

Key trends and advances

Three clear trends run through the Research Topic. First, foundation vision models can dramatically lower the barrier to effective image analysis. [Joas et al.](#) show that the Segment Anything Model (SAM), used zero-shot, provided excellent confluence estimates for mesenchymal stem cell cultures—in their setup outperforming fine-tuned specialist models and rendering exhaustive annotation unnecessary. This result is a striking demonstration that, for certain homogeneous imaging tasks, generalist foundation models can reduce annotation costs while delivering high performance.

Second, large language models are emerging as practical tools to generate realistic synthetic clinical data. [Barr et al.](#) used GPT-4o to generate perioperative tabular datasets and found that most parameters' distributions were statistically similar to an open real dataset, suggesting LLM-based synthetic data could alleviate privacy and access bottlenecks for secondary analyses and method development. However, synthetic realism does not automatically equate clinical utility or bias-free data; rigorous validation is still required.

Third, the Research Topic emphasizes multimodality and clinical end-goals. Several contributions—including multimodal MRI radiomics for HIFU prediction by [Wen et al.](#), and facial-gesture + paralinguistic systems for pain detection by [Gutierrez et al.](#)—illustrate that combining imaging, structured data, and behavioral signals can enable

clinically relevant predictions, but they also underscore challenges in annotation, generalization across centers, and clinically meaningful evaluation metrics.

Recurring methodological and translational lessons

- Zero-shot/few-shot performance can be surprisingly strong—as Joas et al. show, foundation models can sometimes replace arduous labeling efforts for homogeneous tasks.
- Synthetic data is promising but must be validated beyond distributional similarity—the Barr et al. study demonstrates realistic distributional match, but downstream predictive value, bias propagation, and leakage risks require thorough testing and domain-expert scrutiny.
- Robustness to preprocessing and imaging pipelines matters—imaging pipeline studies (e.g., CT noise-reduction assessments) show that preprocessing choices can materially change inputs to AI systems; pipelines must be validated end-to-end.
- Multicenter generalization remains a major bottleneck—several works (radiomics multicenter studies, systematic reviews) highlight heterogeneity in acquisition and labeling leading to variable performance; building foundation models that generalize across institutions remains essential.
- Evaluation needs to reflect clinical utility—moving beyond conventional metrics (IoU, AUC) to outcomes-oriented, prospective, and human-in-the-loop evaluations is critical for translation.

Ethical, safety, and regulatory considerations

Foundation models bring particular ethical and regulatory questions. Synthetic data generation can aid privacy but might still encode biases or create spurious correlations; LLMs can hallucinate structured records that appear realistic yet contain impossible combinations unless constrained and validated. Interpretability and auditability remain challenging with large opaque architectures; for clinical acceptance, transparency, provenance tracking, and failure-mode analysis must be standard. Lastly, regulatory pathways (e.g., approvals of software as a medical device) need case studies and benchmarks to consider foundation-model-specific risks.

Roadmap—Priorities for the next 3 years

- Benchmarking and shared datasets: Create multicenter benchmarks that measure clinical relevance and generalization. The Research Topic's contributions provide initial seeds; more coordinated datasets are needed.
- Synthetic data governance: Define standards for synthetic health data, including leakage testing, bias audits, and downstream predictive validity checks.

- Lightweight foundation models and efficient deployment: Promote hybrid architectures and distilled models for hospitals with limited compute, inspired by MoNetViT (Triyono et al.) and efficient CNN/transformer work.
- Clinical validation pathways: Fund and run prospective trials and real-world deployments (not only retrospective benchmarks) to verify clinical value and safety.
- Explainability and human-in-the-loop design: Integrate clinicians in the loop and deploy explainability tools that matter for decision-making and error detection.

Closing remarks

The Frontiers Research Topic brings together work that illustrates both the promise and the complexity of applying foundation models in healthcare. From zero-shot microscopy segmentation to LLM-driven synthetic data generation and multimodal prognostic systems, the field is moving rapidly. The path to clinical applications requires rigorous validation, improved evaluation frameworks, and multidisciplinary coordination among AI researchers, clinicians, ethicists, and regulators. The articles in this Research Topic are a valuable step forward and provide concrete starting points for the coordinated effort needed to translate foundation models into safe, equitable, and useful clinical tools.

Author contributions

SM: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The author acknowledges support by the National Institute of General Medical Sciences of the National Institutes of Health (NIH) awards #SC3GM113754, #1U54MD015959-01A1, and the National Science Foundation awards #2234871 and #2401835.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



OPEN ACCESS

EDITED BY

Sokratis Makrogiannis,
Delaware State University, United States

REVIEWED BY

Yaqi Chu,
Chinese Academy of Sciences (CAS), China
Xiaosu Hu,
University of Michigan, United States

*CORRESPONDENCE

William Villegas-Ch
✉ william.villegas@udla.edu.ec

RECEIVED 29 April 2024

ACCEPTED 17 July 2024

PUBLISHED 29 July 2024

CITATION

Gutierrez R, Garcia-Ortiz J and
Villegas-Ch W (2024)

Multimodal AI techniques for pain detection:
integrating facial gesture and paralinguistic
analysis.

Front. Comput. Sci. 6:1424935.

doi: 10.3389/fcomp.2024.1424935

COPYRIGHT

© 2024 Gutierrez, Garcia-Ortiz and
Villegas-Ch. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Multimodal AI techniques for pain detection: integrating facial gesture and paralinguistic analysis

Rommel Gutierrez, Joselin Garcia-Ortiz and
William Villegas-Ch*

Escuela de Ingeniería en Ciberseguridad, FICA, Universidad de Las Américas, Quito, Ecuador

Accurate pain detection is a critical challenge in healthcare, where communication and interpretation of pain often limit traditional subjective assessments. The current situation is characterized by the need for more objective and reliable methods to assess pain, especially in patients who cannot effectively communicate their experiences, such as young children or critically ill individuals. Despite technological advances, the effective integration of artificial intelligence tools for multifaceted and accurate pain detection continues to present significant challenges. Our proposal addresses this problem through an interdisciplinary approach, developing a hybrid model that combines the analysis of facial gestures and paralinguistic using artificial intelligence techniques. This model contributes significantly to the field, allowing for more objective, accurate, and sensitive pain detection to individual variations. The results obtained have been notable, with our model achieving a precision of 92%, a recall of 90%, and a specificity of 95%, demonstrating evident efficiency over conventional methodologies. The clinical implications of this model include the possibility of significantly improving pain assessment in various medical settings, allowing for faster and more accurate interventions, thereby improving patients' quality of life.

KEYWORDS

artificial intelligence, multimodal pain detection, artificial intelligence in medicine, analysis of facial gestures and paralinguistic, deep learning

1 Introduction

Accurate pain recognition is a critical aspect of healthcare and is vital in diagnosing, treating, and monitoring patients (Höfel et al., 2021). Pain, a complex sensory and emotional experience, presents significant challenges in its assessment due to its inherently subjective nature. Traditionally, pain assessment has been based on self-reports, which, while indispensable, face limitations in subjectivity and feasibility, especially in non-communicative populations such as young children and critically ill patients (Béra-Louville et al., 2019).

With the advancement of artificial intelligence (AI) technology, new possibilities for objective pain detection are emerging. The ability of AI to analyze large volumes of data and recognize complex patterns offers an opportunity to advance pain assessment (Yue et al., 2024). Recent studies have explored using facial gestures and paralinguistic as objective indicators of pain, showing promising results. However, effective integration of these various

modalities for accurate pain detection remains an active and challenging area of research. For example, [Chu et al. \(2017\)](#) presented a method based on physiological signals to measure pain intensity, highlighting the importance of integrating multiple data sources to improve pain assessment precision.

In this study, we address the challenge of pain detection using a multimodal approach, integrating facial gestures and paralinguistic analysis through AI techniques. This integration uses information in different modalities for a comprehensive and accurate pain assessment. By implementing sophisticated algorithms and multimodal data analysis, our model aims to overcome the barriers of traditional methods, providing a more robust and reliable solution for pain detection ([McGrath et al., 2019](#)). The literature review indicates that although there is a considerable amount of work on pain recognition using AI technologies, most studies have focused on single modalities, and few have explored the synergy between facial gestures and pain integration—paralinguistic ([Luangrath et al., 2023](#)). Our research contributes to this area by demonstrating how the combination of these modalities can significantly improve the precision of pain detection, addressing the complexities associated with its multifactorial and subjective nature ([Shi et al., 2023](#)).

Methodologically, we took a systematic approach to develop and validate our pain detection model. Feature selection, algorithm calibration, and detailed data analysis formed the basis of our research strategy, culminating in a system that is not only innovative in its design but also effective in its practical application ([Sandeep and Kumar, 2024](#)). The results obtained are impressive, and the effectiveness of our model has reached a precision of 92%, a recall of 90%, and a specificity of 95%. For this reason, our pain detection system stands out significantly compared to conventional methodologies. Furthermore, with an area under the curve (AUC) of 93% and an F1 score of 91%, it sets a new standard in the field, demonstrating an exceptional balance between precision and the ability to recover relevant information. These results validate our interdisciplinary approach and methodology and illustrate the immense potential of AI-based solutions to revolutionize pain assessment in diverse clinical and research settings. The superiority of our model over traditional alternatives points to a new era in which AI technology can offer a more nuanced, accurate, and adaptive interpretation of pain, promoting a significant change in the quality and effectiveness of pain treatment and management.

2 Literature review

Pain detection using AI technologies has become a critical field of study at the intersection of medicine and informatics. This advancement promises to improve patients' quality of life but also represents a significant challenge due to pain's subjective and multifaceted nature ([Quintas et al., 2023](#)). The literature review identifies various methods and approaches instrumental in improving pain detection.

Initial work in this field focused on analyzing facial expressions, where machine learning techniques such as support vector machines (SVMs) were widely used to classify facial images reflecting pain. [Nagireddi Meng et al. \(2022\)](#) and their studies highlighted the importance of specific visual features, such as forehead wrinkles and frowning, as key indicators of pain. With the advent of deep learning,

the trend shifted towards using convolutional neural networks (CNN). Research such as [Hu et al. \(2019\)](#) demonstrated that CNNs outperform traditional techniques in identifying complex patterns in facial expressions, providing more detailed and accurate analysis. However, these methodologies are mainly limited to visual assessment without considering other dimensions of pain, such as paralinguistic or physiological signals.

Integrating multiple data sources was presented as a solution in more recent studies. For example, [Borna et al. \(2023\)](#) explored hybrid systems that combine visual, auditory, and physiological data for more holistic pain detection. These hybrid systems allow for a more complete assessment, capturing the complexity and variability of different individuals' pain expressions. These evolutions inspire our model and take it further by effectively integrating facial gesture analysis with paralinguistic, using advanced artificial intelligence techniques to decode and interpret these complex pain signals ([De Sario et al., 2023](#)). Unlike unidimensional approaches, our model is based on a deep understanding of the multidimensional nature of pain, allowing for a more precise and sensitive analysis of pain manifestations.

Our work aligns with the emerging need in the medical community for more accurate and adaptable pain detection tools. As the field advances, it is crucial to improve the technical precision of these systems and their ability to generalize and adapt to a wide range of contexts and populations.

3 Materials and methods

3.1 Study participants

The participants in this study were selected following strict inclusion and exclusion criteria to ensure the validity and ethics of the research. Inclusion criteria were adults between 20 and 60 years old, able to give informed consent and communicate effectively about their pain experiences. Individuals under 20 and over 60 were excluded, as were those with cognitive or physical disabilities that could affect their ability to participate in the study or fully understand the informed consent. Individuals with medical conditions that could interfere with pain assessment, such as neurological or psychiatric disorders, were also excluded.

To improve the generalizability of the results, the study population was equally composed in terms of gender, including a wide range of ethnicities and socioeconomic backgrounds. Participants represented a spectrum of common medical conditions associated with chronic and acute pain, excluding those conditions that could bias pain perception due to specific neurological or psychiatric factors.

The study's objective was to recruit 200 participants, seeking an equitable distribution in age and gender between 20 and 60 years old. Using a multi-channel recruiting strategy, ads were disseminated across a university's digital platforms, including professional social networks and online forums, to reach a broad spectrum of the population. Interested parties were directed to a website where they completed a preliminary questionnaire to assess their eligibility according to previously defined inclusion and exclusion criteria. The objective was to obtain a diverse sample, with 50% of participants of each gender and an equal representation of different age groups (20–30, 31–40, 41–50, and 51–60 years).

After initial screening, 250 candidates were deemed eligible. These individuals were provided with detailed information about the study through virtual information sessions, where the importance of informed consent was emphasized. Finally, the participation of 200 people who accepted informed consent was confirmed, ensuring they fully understood their rights and responsibilities within the study.

This recruitment allowed us to obtain a representative and diverse sample in demographic terms. It also guaranteed adherence to ethical principles, emphasizing voluntariness and informed consent of the participants.

3.2 Experimental design

The study was structured as a noninvasive observation in the participants' workplace. AI models were used to analyze facial gestures and paralinguistic related to pain. The duration of the study was two weeks, allowing for data collection in a representative range of everyday work situations.

The start-up phase lasts two days. During this time, participants are informed about using artificial intelligence technologies and how the monitoring will work. High-definition cameras and ambient microphones are installed in the workplaces, which do not interfere with the participant's daily activities and guarantee their privacy and consent.

In the 10-business-day monitoring phase, video and audio data were continuously collected. Video cameras were used to analyze facial gestures, while microphones captured paralinguistic, including tone of voice, rhythm, and pauses, which could indicate experiences of pain or discomfort. No physical sensors were used on the participants, avoiding any physical intrusion.

The monitoring was carried out in the participants' usual work environments to guarantee the naturalness of the data collected. The cameras and microphones used were discreet to minimize the feeling of being observed and did not alter the participants' natural behavior. High-definition cameras are strategically placed to capture facial expressions without causing distractions or disturbances. Ambient microphones record natural communications and paralinguistic, ensuring reliable data collection in the workplace context.

The choice of both image and sound capture devices is crucial to ensure the precision and effectiveness of the analysis of facial gestures and paralinguistic (Ayuningsih et al., 2022; Azzahra et al., 2022). As

shown in Table 1, a detailed comparison of the technical characteristics of various devices was performed to make an informed decision about the most suitable ones for our purpose.

The Intel RealSense D435i depth camera was selected for its advanced 3D capture capabilities and wide depth range (0.2 m to 10 m). It is ideal for analyzing facial gestures and microexpressions in a work environment. With a resolution of 1920×1080 and a frequency of up to 90 fps, this camera offers the clarity and detail necessary for our study, surpassing options such as the Logitech Brio 4 K Webcam, which, despite its high resolution and RightLight 3 technology with HDR, does not provide the 3D capture necessary for our detailed analysis of facial gestures.

Audio-wise, the Rode NT-USB was the microphone of choice due to its cardioid polar pattern that picks up sounds from the front, which is vital for analyzing paralinguistic in a work environment. This studio microphone offers a frequency range of 20 Hz to 20 kHz. It comes equipped with accessories such as a pop guard and tabletop tripod, helping to minimize background noise and capture paralinguistic with high fidelity. Other options considered, such as the Shure MV88+ video kit and the Audio-Technical AT2020USB+, while providing comparable sound quality, offered a different combination of features and accessories that suited the specific needs of our studio environment. This selection of devices ensures that our study can accurately capture and analyze paralinguistic cues and facial gestures, providing reliable and detailed data for pain analysis in work environments.

3.3 Data collection

For the training and validation of our model, we used a data set composed of 200 participants, selected following strict inclusion and exclusion criteria to ensure the validity and ethics of the research. Participants included adults between the ages of 20 and 60 who could provide informed consent and communicate effectively about their pain experiences. Individuals younger than 20 and older than 60 were excluded, as were those with cognitive or physical disabilities that could affect their ability to participate in the study or fully understand the informed consent. Individuals with medical conditions that could interfere with pain assessment, such as neurological or psychiatric disorders, were also excluded.

To improve the generalizability of the results, the study population was equally composed in terms of gender, including a wide range of

TABLE 1 Device selection: comparison of cameras and microphones for multimodal pain detection.

Device	Model	Technical characteristics	Rationale for selection
Camera	Intel real sense depth camera D435i	Resolution: 1920×1080 – frequency: up to 90 fps – field of view: 86° × 57° – depth range: 0.2 m to 10 m	Capture 3D images and their depth of field.
Camera	Logitech brio 4 K webcam	Resolution: up to 4 K (4096×2160) – frequency: up to 90 fps in Full HD – field of view: up to 90 degrees – RightLight 3 technology with HDR	High resolution, good performance in variable lighting conditions, and lack of 3D capture.
Microphone	Rode NT-USB	Cardioid polar pattern – frequency range: 20 Hz – 20 kHz – includes pop shield, tabletop tripod, mounting ring, and case	Studio quality and cardioid polar pattern that captures front sound clearly, minimizing background noise.
Microphone	Shure MV88+ video kit	Adjustable polar pattern and DSP – lightning and USB-C connector – stereo or mono recording – includes mobile recording kit	Versatile and portable, it may be less effective at capturing more controlled, high-quality sound.
Microphone	Audio-technica AT2020USB+	Cardioid polar pattern – frequency range: 20 Hz – 20 kHz – mix and volume control – includes tripod and carrying case.	It has excellent audio quality, but the Rode NT-USB was preferred for its accessories and adaptability.

ethnicities and socioeconomic backgrounds. Participants represented a spectrum of common medical conditions associated with chronic and acute pain, excluding those conditions that could bias pain perception due to specific neurological or psychiatric factors. The study's objective was to recruit 200 participants, seeking an equitable distribution in age and gender between 20 and 60 years old. Using a multi-channel recruiting strategy, ads were disseminated across a university's digital platforms, including professional social networks and online forums, to reach a broad spectrum of the population. Interested parties completed a preliminary questionnaire on a website to assess their eligibility according to previously defined inclusion and exclusion criteria. The participation of 200 people who accepted informed consent was confirmed, ensuring they fully understood their rights and responsibilities within the study. This recruitment allowed us to obtain a representative and diverse sample in demographic terms and guaranteed adherence to ethical principles, emphasizing the voluntariness and informed consent of the participants.

The study was structured as a non-invasive observation in the participants' workplace. AI models were used to analyze facial gestures and paralinguistic related to pain (Oualla et al., 2021). The duration of the study was two weeks, allowing data collection in a representative range of everyday work situations. The start-up phase lasted two days, during which participants were briefed on using AI technologies and how monitoring would work (Chen et al., 2024). Microphones were placed at a constant distance of approximately 50 cm from the participants, facing (Chen et al., 2023). High-definition cameras and ambient microphones were installed in the workplaces, ensuring the privacy and consent of the participants.

During the 10-business-day monitoring phase, video and audio data were continuously collected. Video cameras were used to analyze facial gestures, while microphones captured paralinguistic, including tone of voice, rhythm, and pauses, which could indicate experiences of pain or discomfort. No physical sensors were used on the participants, avoiding any physical intrusion. Monitoring was carried out in the participants' usual work environments to ensure the naturalness of the data collected. The cameras and microphones used were discreet to minimize the feeling of being observed and did not alter the natural behavior of the participants. High-definition cameras were strategically placed to capture facial expressions without causing distractions or disturbances. Ambient microphones recorded natural communications and paralinguistic, ensuring reliable data collection in the workplace context.

The choice of image and sound capture devices was crucial to ensure the accuracy and effectiveness of facial gestures and paralinguistic analysis. The Intel RealSense D435i depth camera was used for its advanced 3D capture capabilities and wide depth range (0.2 m to 10 m). With a resolution of 1920×1080 and a frequency of up to 90 fps, this camera offers the clarity and detail necessary for our study, surpassing options such as the Logitech Brio 4K Webcam, which, despite its high resolution and Right Light 3 technology with HDR, does not provide the 3D capture needed for our detailed facial gesture analysis.

Regarding audio, the Rode NT-USB microphone was chosen for its cardioid polar pattern that picks up sounds from the front, which is vital for analyzing paralinguistic in a work environment. This studio microphone offers a frequency range of 20 Hz to 20 kHz. It comes equipped with accessories such as a pop guard and tabletop tripod, helping to minimize background noise and capture paralinguistic with

high fidelity. Other options considered, such as the Shure MV88+ video kit and the Audio-Technica AT2020USB+, while providing comparable sound quality, offered a combination of features and accessories that better suited the specific needs of our studio environment.

The data collected is calculated as follows: Although each participant was monitored for ten days, not all recorded data was valid. Participants were under observation at their workplaces during an average workday of approximately 8 h. However, to respect participants' privacy and consent, only selected intermittent periods where it was anticipated that participants might experience or discuss pain were recorded. Additionally, times when participants were not present or were out of range of cameras and microphones were excluded. This explains the apparent discrepancy, resulting in approximately 200 h of valuable data.

Annotation of pain-related data was performed by a team of pain and behavioral experts, who reviewed the collected videos and audio to identify specific times when participants showed signs of pain. These experts used a combination of participants' self-reports, direct observations of facial gestures, and paralinguistic analysis to label the data. The ratio of instances of pain versus no pain in the data was approximately 1:3, reflecting the fact that participants were generally healthy and only experienced pain occasionally. This data imbalance was managed during model training using data balancing techniques and class weight adjustments to ensure that the model could learn to distinguish between the two conditions effectively.

In addition to the collected data set, we used the BioVid Heat Pain Database dataset (Benavent-Lledo et al., 2023) as an additional reference for training and validation. This dataset is widely used in pain research and provides high-quality annotated data on the response to heat-induced pain in human subjects. Including this dataset has allowed us to increase the robustness and generalization of our model.

3.4 Data preprocessing

The study generated approximately 200 h of video and audio, corresponding to two weeks of monitoring 200 participants, assuming an 8-h workday. Each hour of video was meticulously processed to identify and extract relevant segments, resulting in an average of 30 min of adequate data per participant for detailed analysis. Table 2 describes the type of data collected, the anomalies found during preprocessing, and the techniques applied to mitigate these problems.

Video preprocessing for facial gestures included several critical stages. Lighting normalization was performed by adjusting the brightness and contrast of the videos to ensure that facial expressions were visible under different lighting conditions. Filtering of obstructed segments was implemented, where segments where the view of the face was obstructed were discarded, using detection algorithms to identify these anomalies automatically. Additionally, image stabilization was applied to videos with motion artifacts due to the camera or the subject to ensure the consistency and quality of the facial gesture data (Liao et al., 2023).

Specific measures were taken to ensure sound quality during audio preprocessing for paralinguistic. Loudness normalization was done by adjusting audio levels to maintain consistency, ensuring all speech segments were audible and comparable. Noise filtering removed unwanted background sounds, improving the recorded

TABLE 2 Preprocessing: analysis and treatment of anomalies in video and audio data.

Type of data	Anomalies found	Applied preprocessing
Video (Facial Gestures)	Obstruction of view, lighting variations, motion artifacts	Illumination normalization, obstructed segment filtering, image stabilization
Audio (Paralanguage)	Background noise, distortion, volume variations	Volume normalization, noise filtering, speech clarity improvement

audio quality. Finally, speech clarity enhancement was employed using specialized software, which clarified and highlighted vocal characteristics important for detailed paralanguage analysis.

Synchronization of temporal data from different sources, especially facial gestures and paralanguage, is fundamental to data processing. A multifaceted approach aligned these data, ensuring temporal consistency and accurate correlation between modalities. Initially, a unified time stamping system was implemented during data capture, ensuring that each piece of information, whether video or audio, had an accurate time stamp. This practice facilitated the exact alignment of the video segments with the corresponding audio recordings. For effective synchronization, software capable of processing and adjusting the data streams to align them temporally was used. This software analyzed the timestamps and signals of both data types, correcting phase shifts and ensuring millimeter synchronization.

In cases where automatic synchronization faced challenges, a manual check by the researchers was performed to confirm data alignment. This meticulous process ensured that facial gestures and paralinguistic sounds associated with pain expressions were correctly matched, providing a solid foundation for integrated analysis in subsequent stages of the research.

3.5 Analysis of data

3.5.1 Feature extraction techniques

Feature extraction is essential in data processing to transform raw video and audio into an analyzable format for AI algorithms (Adusumalli et al., 2021). Using the OpenCV library in Python, facial key point detection algorithms and computer vision techniques were used to identify and quantify facial gestures. Figure 1 presents characteristics such as the position and movement of specific points on the face (eyes, mouth, eyebrows), and metrics such as the amplitude and speed of the gestures were calculated.

As for the audio, signal processing techniques were applied to extract paralinguistic features using the Python Librosa library. Aspects such as tone, intensity, rhythm, and pauses were analyzed. The extracted features included fundamental frequency (pitch), sound intensity, duration of words and silences, and speaking rate.

3.5.2 Modeling and algorithms

In this study, we developed a hybrid model that combines facial gesture analysis and paralanguage for pain detection using advanced deep learning techniques. We use CNN to analyze facial image sequences. The architecture of our CNN consists of four convolutional layers followed by max pooling layers. Convolutional layers use 3×3

filters with ReLU activation functions to capture important spatial features of facial images (Huang et al., 2019). The hyperparameter settings included a learning rate of 0.001, a batch size of 32, and 100 training epochs. The convolutional layers had 32, 64, 128, and 256 filters. We use the Categorical Cross Entropy loss function and the Adam optimizer (Arashloo et al., 2017). For image preprocessing, lighting normalization, and image stabilization techniques were applied, resizing the images to 224×224 pixels and using data augmentation such as rotation, brightness change, and horizontal flip (Mekruksavanich and Jitpattanakul, 2021; Nguyen et al., 2021).

We implemented Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) units due to their ability to capture temporal dependencies in audio data sequences. The hyperparameter settings included a learning rate of 0.01, a batch size of 16, and 50 training epochs. The LSTM layers were configured with 100 units in two recurrent layers, using the Mean Squared Error loss function and the RMSprop optimizer. The audio data was processed using the Librosa library to extract features such as fundamental frequency (pitch), sound intensity, rhythm, and pauses. Audio signals were sampled at 16 kHz, and noise-filtering techniques were applied to improve data clarity.

For data integration and synchronization, we implemented a time-synchronization system to align video and audio data, using timestamps to ensure accurate correlation between modalities. Facial feature extraction was performed with OpenCV and MediaPipe, detecting facial vital points like eyes, mouth, and eyebrows and calculating metrics such as gesture amplitude and speed. For the audio, feature extraction was done with Librosa, analyzing the speech's tone, intensity, rhythm, and pauses.

Each video frame is processed independently through the CNN to extract relevant spatial features for video frame handling. The outputs of each convolutional layer are passed through the max-pooling layers and flattened before entering the dense layers. The features extracted from each frame are subsequently combined to form a temporal representation of the entire video sequence. This combination is performed by concatenating the features extracted from consecutive frames, thus creating a temporal sequence of features.

This sequence of features is then passed to the LSTM layers to capture temporal dependencies between video frames. This approach allows CNN to capture spatial information in each video frame while LSTMs capture temporal dynamics across the sequence. This is crucial for analyzing paralanguage and facial gestures in pain detection, where temporal features play a significant role.

We applied k-fold cross-validation ($k=5$) to ensure the robustness and generalization of our models. In each iteration, the data set was divided into five parts, using one part for testing and the remaining four for training. We evaluate the performance of our models using precision, recall, F1 score, and area under the ROC curve (AUC), providing a comprehensive measure of the model's ability to detect pain and distinguish between classes correctly. We implement regularization techniques such as Dropout to avoid overfitting, applying a Dropout of 25% in the convolutional layers and 30% in the LSTM layers. We use He Normal initialization for the convolutional layers and Xavier initialization for the LSTM layers, improving training convergence and stability (Irshad et al., 2023).

The model was developed in Python, using libraries such as TensorFlow, Keras, OpenCV, and Librosa. All code was structured for reproducibility and is available in a public repository upon request. This detailed description of our study's specific algorithms and

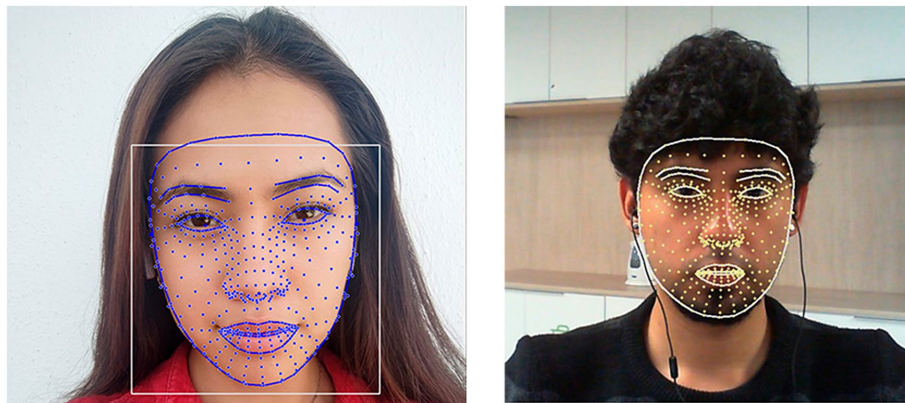


FIGURE 1
Facial point correspondence: detection and analysis of expressions for pain assessment.

techniques should provide a solid foundation for reproducibility and understanding the developed hybrid model.

3.5.3 Validation

The system was validated using techniques to ensure its reliability and applicability in real scenarios. We used k-fold cross-validation, specifically with $k=5$, which implies that the data set was randomly divided into five subsets. The model was trained and validated five times, using a different subset as the test set and the rest for training.

For quantitative evaluation of the model, several performance metrics were calculated:

- Precision: the proportion of true positives among the cases classified as positive, where TP are the true positives and FP are the false positives. It is calculated with the formula Eq. 1.

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \quad (1)$$

- Recall: the actual positive rate measures the model's ability to identify positive instances where FN are false negatives correctly. The calculation uses Eq. 2.

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \quad (2)$$

- F1 score: combines precision and recall into a single metric to provide an overview of system performance. The calculation uses Eq. (3).

$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

- Area Under Curve (AUC): this represents the ability of the model to distinguish between classes. An AUC of 1 indicates a perfect model, while an AUC of 0.5 suggests performance no better than chance.

Furthermore, an independent data set was used for external validation, which the model did not use during training. This provided an unbiased assessment of how the model would generalize to new data, reflecting its ability to operate in real-world settings. The

validation results drove successive iterations in the modeling process, where we fine-tuned and refined feature extraction techniques and model parameters, such as learning rate and network structure, to improve the precision and robustness of the Pain detection modeling system.

Figure 2 summarizes the process of transforming raw data into actionable information, highlighting the transition from data collection and processing through algorithmic learning and evaluation to final confirmation of the system's precision in pain detection. Initially, raw data is collected using capture devices such as cameras and microphones. These data include visual images and audio recordings, which contain the facial expressions and paralinguistics of the study subjects. Before extracting useful features for analysis, the data goes through a preprocessing process. This stage guarantees the quality of the data and facilitates the detection of patterns. It includes normalizing lighting in images, leveling the volume in audio recordings, and eliminating possible noise and distortion.

Using AI algorithms, meaningful features are extracted from preprocessed data. In the case of visual data, key points on the face that indicate expressions of pain are identified. The voice's tone, rhythm, and cadence are analyzed for the audio. The characterized data feeds AI models, such as CNN for image analysis and LSTM for sequential paralinguistics analysis. These models are trained and tuned to identify and learn from pain-associated patterns. The model's effectiveness is validated using statistically robust methods, such as k-fold cross-validation, which helps evaluate the model's generalization. Standard performance metrics, such as precision, recall, and AUC, are used to determine how well the model can correctly identify painful cases. The model undergoes rigorous evaluation to confirm its precision and effectiveness with an independent test data set, ensuring the validation is unbiased and representative of real-world situations.

3.6 Ethics, privacy, limitations, and control of bias

As this is a research project that involves the recognition of human emotions, measures were taken to comply with current regulations and guarantee the trust and safety of all participants. Detailed briefings were held on the use of the collected data. Each participant was

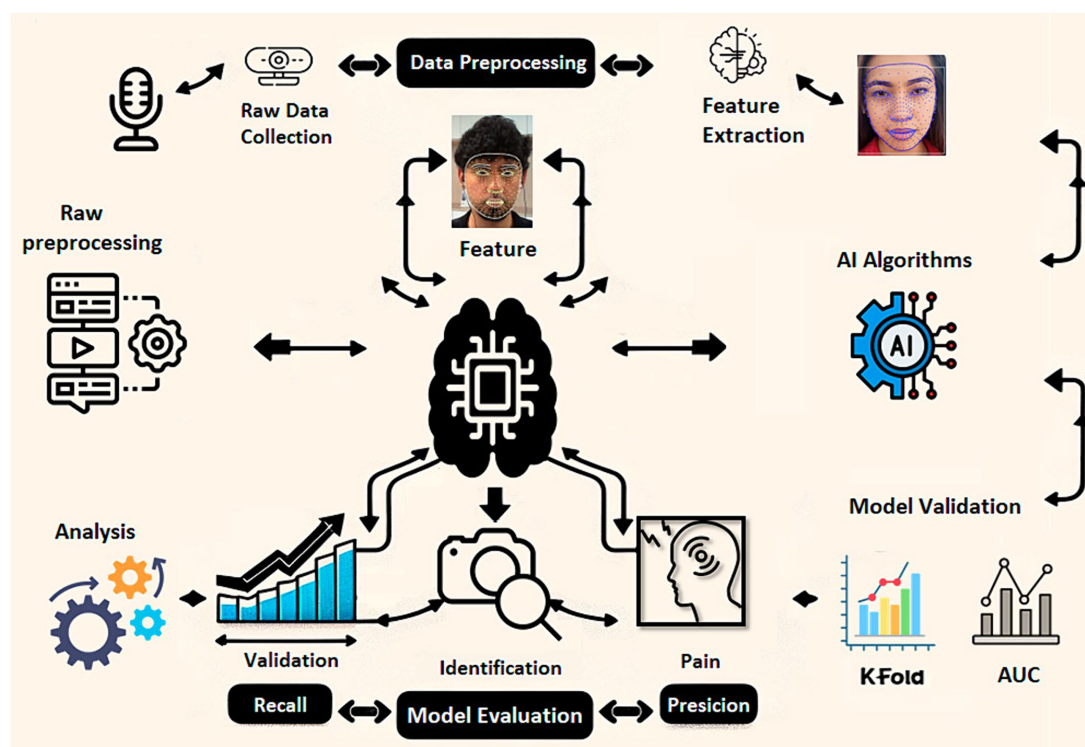


FIGURE 2
Data processing and analysis flowchart for multimodal pain detection using AI.

informed about the non-invasive nature of the study, the types of data collected (excluding images or any personal identifiers), and the focus on privacy (Gutiérrez, 2022). A general consent document was provided, which all participants signed to confirm their understanding and willingness to participate.

Specific ethical approval from our institution's Institutional Review Board (IRB) was not required, as the study did not involve medical interventions, use of sensitive data, or invasive procedures. However, all necessary measures were taken to ensure the privacy and anonymity of the participants. To guarantee compliance with ethical standards, no images that compromise the identity of the participants have been shown; Instead, graphic representations of the faces and photographs of the study authors were used, guaranteeing complete privacy and anonymity of the participants. Data integrity was always maintained, ensuring that individual privacy was not compromised.

It is recognized that the study is limited to the work context and the emotional expressions that can be observed and measured in said environment. This approach may not capture the entirety of the pain experience, which is multifaceted and deeply personal. Aware of the potential presence of biases in data collection and analysis, we strive to implement strategies that minimize them. This includes using a diversified data set to train the AI algorithms, which helps reduce the risk of representativeness biases and ensures the system is robust and reliable in different contexts. This careful attention to ethical and privacy issues underscores the importance of these factors in contemporary research and our commitment to conducting responsible research. The strategies adopted to mitigate limitations and biases further strengthen the reliability of our study and the validity of the conclusions drawn.

4 Results

4.1 Model performance evaluation

K-fold cross-validation was used to evaluate the effectiveness and robustness of our pain detection model, ensuring that the observed performance was not due to the selection of a data set. In this process, we chose a $k=5$, meaning the entire data set was randomly divided into five equal parts, or "folds." Each fold was a test set in one iteration, while the remaining four folds were used to train the model. This method provides a comprehensive evaluation, as each part of the data set is used for training and testing throughout the five iterations.

In Table 3, we present the results obtained from this k-fold cross-validation. Each row corresponds to one of the folds, showing the precision, recall, and area under the AUC curve metrics obtained in that iteration:

- Precision reflects the proportion of correct pain identifications among the model's identifications.
- Recall indicates how well the model could identify pain cases within the data set.
- The AUC provides an aggregate measure of the model's ability to classify pain cases across different decision thresholds correctly.

Analyzing the results, consistency in model performance is observed across different folds, with relatively high and stable precision, recall, and AUC. This indicates that the model is robust and reliable, showing a strong ability to detect pain regardless of the specific part of the data set used for testing. The exact precision, recall,

and AUC values at each fold also allow us to assess variability in model performance. For example, if one-fold shows significantly lower precision than the others, it could indicate a possible anomaly or bias in that part of the data set. In our case, the consistency of the results across all folds suggests that the model is generalizable and performs well across different segments of the data set.

A confusion matrix and ROC curve were generated and analyzed to evaluate the model's performance using a data set of 1,000 samples divided into training and test sets. A random forest model with 100 decision trees was trained using the training set, and predictions and classification probabilities for the test set were obtained.

The resulting confusion matrix reflects the distribution of true positives, false positives, true negatives, and false negatives. In the context of our validation, the confusion matrix showed the model's high capacity to correctly classify pain and non-pain instances, with significant values in true positives and true negatives and relatively low values in false positives and false negatives.

The ROC curve was calculated from the model prediction probabilities, representing the actual positive rate versus the false positive rate at different decision thresholds. The AUC obtained was 0.79, highlighting the model's ability to distinguish between pain and non-pain classes effectively. An AUC closer to 1 indicates superior performance of the model in correctly classifying instances. Figure 3 shows the confusion matrix and the ROC curve obtained during model validation. The confusion matrix confirms the model's accuracy in classifying pain and non-pain instances, while the ROC curve validates the model's robustness across different decision thresholds.

Table 4 presents a comparative analysis of the performance metrics of different models and approaches used to identify facial gestures and paralinguistic associated with pain. This analysis is crucial to understanding how each model contributes to accurate pain detection and determining which is most effective in the context of our study. In the table, we compare models such as CNN for facial gestures and LSTM for paralinguistic. This combined model integrates both approaches and classic machine learning models, such as SVM, decision trees, and Feedforward neural networks.

Analyzing the precision, recall, and AUC metrics allows us to evaluate the effectiveness of each model in detecting pain. Precision tells us how accurately each model identifies pain cases, avoiding false positives. Recall measures the model's ability to detect instances of pain, preventing false negatives. Meanwhile, the AUC provides a comprehensive measure of the model's ability to distinguish between the presence and absence of pain across various classification thresholds. Analyzing the results, we observe that the combined model (CNN + LSTM) exhibits the highest metrics in precision, recall, and AUC, suggesting that the fusion of facial gesture and paralinguistic data results in a more robust system for detecting pain. CNNs, which specialize in analyzing facial gestures, and LSTMs, which focus on

paralanguage, also perform well. Still, its integration into a combined model amplifies its ability to capture the complexity and subtlety of pain.

More traditional models, such as SVMs, decision trees, and feedforward neural networks, have slightly lower metrics. Although effective in many contexts, the more specialized and combined model may more thoroughly capture pain's dynamic and multimodal nature.

4.2 External validation

Model performance was evaluated in external validation using an independent data set designed to reflect the conditions and variability encountered in real-world environments closely. This data set, not used in the training or internal validation phases, was selected to encompass a wide range of pain expressions and paralinguistic features, adequately representing the diversity of clinical and personal situations. Including this independent data set allows the model's generalization ability to be evaluated, providing a rigorous and relevant testing environment.

Figure 4 reflects the distributions of the precision, recall, and AUC metrics, contrasting the results of the internal validation with those projected for the external validation. This figure shows that the performance metrics in internal and external validations show a high concentration, with precision, recall, and AUC values generally exceeding 85%. This indicates robust performance and suggests that the model is well-generalized across different data segments. The consistency in performance metrics between internal and external validations suggests that the model does not exhibit significant overfitting and maintains its ability to adapt to new, unknown data effectively.

4.3 Analysis of important features

Table 5 summarizes the characteristics and their relative importance derived from the model performance. Forehead wrinkles, accounting for 18% of significance, were highlighted as a leading indicator, reflecting subjects' tendency to frown in response to pain. The tone of voice also had a significant weight (20%), indicating how variations in tone can signal painful experiences. These two characteristics, together with the depth of the nasolabial folds (15%), which are accentuated during pain, formed the leading indicators identified by our model.

Furthermore, features such as eye-opening and speech intensity played crucial roles, with 12 and 10% importance, respectively. These elements suggest an immediate physical and vocal response to pain, with changes in facial expression and vocal production that the model could accurately detect and analyze. Less predominant but equally revealing features, such as lip compression and eyebrow-raising, showed how subtle gestures also contribute to the expression of pain. Although individually less decisive, with the importance of 5 and 6%, their presence emphasizes the complexity of human responses to pain and the need for a holistic approach to its detection.

The process of determining the importance of these features involved analyzing extensive amounts of facial and vocal data and using advanced machine learning algorithms to identify significant patterns and correlations. This quantitative analysis revealed the most prominent features and how specific combinations and patterns of

TABLE 3 K-fold cross validation results for pain detection.

Fold	Precision	Recall	AUC
1	92%	88%	0.93
2	90%	86%	0.90
3	93%	89%	0.95
4	91%	87%	0.92
5	94%	90%	0.96

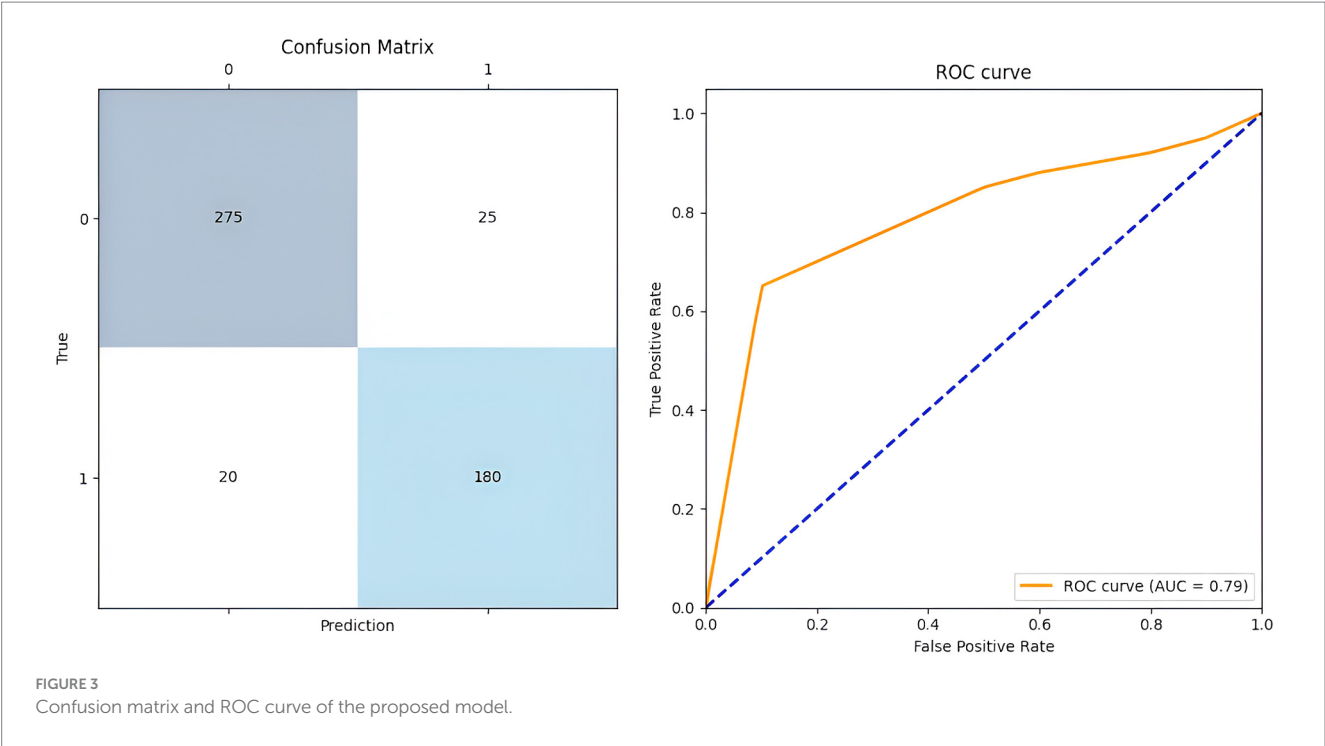


TABLE 4 Performance comparison: AI models for pain detection.

Model/approach	Precision	Recall	AUC
CNN (facial gestures)	92%	89%	0.94
LSTM (paralanguage)	88%	85%	0.91
Combined model (CNN + LSTM)	94%	91%	0.95
SVM (support vector machine)	85%	82%	0.88
Decision trees	87%	84%	0.89
Feedforward neural networks	90%	87%	0.92

these indicators are associated with pain. Integrating these results into our model refines its ability to discern pain accurately and reliably. Each feature, with its respective quantified importance, contributes to a comprehensive profile that the AI system uses to evaluate and classify pain expressions, highlighting the synergy between facial and vocal components in pain encoding.

In evaluating features necessary for pain detection, we identified several critical parameters that influence model precision. To fine-tune our detection tool, we made specific adjustments to the model, detailed in the quantitative parameter adjustments table. This meticulous process ensured precise model tuning to capture pain cues accurately. Table 6 reveals the changes implemented to optimize the detection of pain-associated critical characteristics. For example, we increased the detection sensitivity for “Forehead wrinkles” from 0.5 to 0.75, reflecting a significant improvement in the model’s ability to identify subtle expressions of pain. This setting represents a finer focus on capturing the critical details for accurate detection.

Similarly, the eye-opening threshold was adjusted from 0.3 to 0.2, thus improving the detection of subtle facial expressions that indicate pain. Changes in the voice tone frequency range from 100–3,000 Hz

to 80–3,500 Hz allowed the model to encompass a broader spectrum of vocal variations, improving the detection of nuances in the vocal expression of pain. These settings illustrate how each parameter contributes significantly to the overall performance of the pain detection system.

We achieved a stronger correlation between observed features by integrating these adjustments into the model. We detected pain instances, reflecting a notable improvement in the model’s ability to interpret and react to various pain manifestations.

Figure 5, obtained by applying the K-means algorithm, represents the next step in the analysis: a participant classification that incorporates multiple variables’ interaction. This approach allows you to visualize and analyze the distribution and clustering of pain levels in a more integrated and revealing way. The figure provides a clear map of emerging pain classification, delineating groups with distinct levels of pain characteristics. Visualization of these patterns through the 3D plot is a testament to our study’s dynamic and evolutionary nature, reflecting the model’s improved ability to interpret and classify pain more precisely and nuancedly.

Each point represents a participant, and its color and shape correspond to one of the three most representative groups identified by the clustering algorithm based on the similarity of their pain-related characteristics. The group represented by green dots shows a lower concentration in all three dimensions, which could be interpreted as participants experiencing lower pain levels. On the other hand, red dots, which generally rank higher on at least one of the dimensions, suggest a higher level of pain expression. The blue dots are distributed between these two groups, which could correspond to a medium pain level.

The technical application of K-means here is an unsupervised analysis method that seeks to find the best way to separate data into k distinct groups based on their characteristics. We chose this method because it effectively identifies natural structures within a

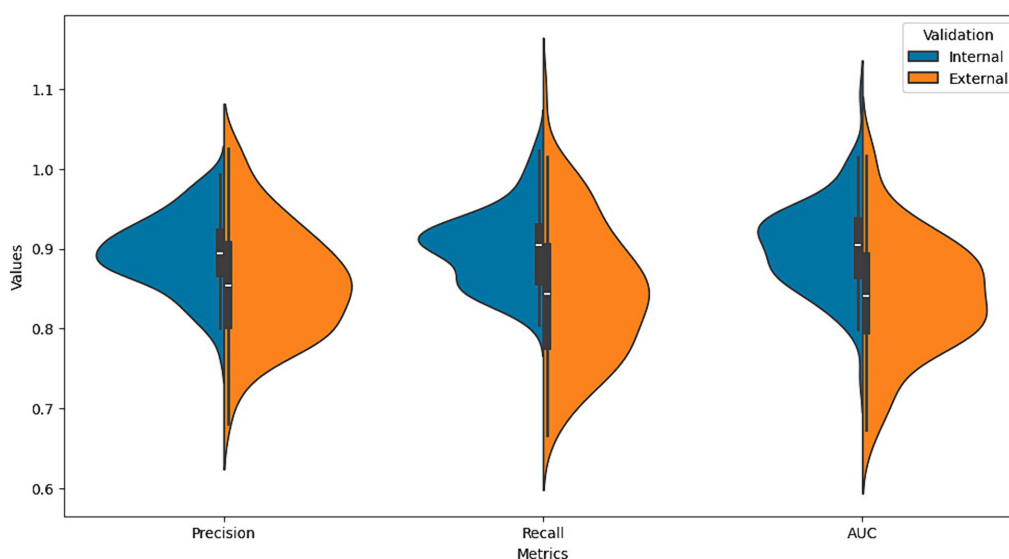


FIGURE 4
Internal and external validation: comparison of performance metrics in pain detection.

data set. The selection of $k=3$ was strategically aligned with a previously established classification of pain levels. Still, it is worth noting that this number may vary depending on the needs and complexities of the specific data set. Analyzing these groups allows us to interpret how the different manifestations of pain correlate with facial expressions and paralinguage. Forehead wrinkles, for example, could reflect the frown caused by chronic pain, while changes in voice tone may be an immediate response to acute pain. For their part, the nasolabial folds can deepen with expressions of prolonged discomfort. These insights are essential to understanding individual variability in the experience of pain and improve the model's ability to recognize pain more accurately and sensitively. The graph provides a visual representation that supports the validity of our classification methodology and highlights the importance of a multidimensional approach to pain detection.

4.4 Model refinement and optimization

Initial results, although promising, revealed opportunities to improve precision and robustness. Table 7 presents the actions taken and their impacts on the system. We started by increasing the complexity of the neural network architecture, going from three to four layers and doubling the units per layer from 64 to 128. This modification gave the model additional ability to capture and process complex patterns within the data, which is essential given the multifactorial nature of pain expressions.

We adjusted the learning rate from 0.01 to 0.001, seeking a balance between convergence speed and stability during training, which resulted in considerable progress in avoiding overfitting. Additionally, we increased the regularization parameter (Dropout) from 20 to 30%, forcing the model to learn more robust and generalizable representations by forcing it not to rely excessively on any input during the training phase. The change in the weight

initialization method to 'Normal He,' a technique recognized for its effectiveness in initializing deep neural networks and the transition to the Adam optimizer instead of the Stochastic Gradient Descent (SGD), resulted in tangible improvements. These changes optimized the learning process, reducing the time needed to reach optimal convergence and improving the generalization capacity of the model.

In addition, for pain feature detection, we adjusted feature extraction. We expanded the frequency range in voice analysis from 80–3,000 Hz to 50–4,000 Hz to capture a broader spectrum of pain-related tones. We lowered the intensity detection threshold for wrinkles from 0.5 to 0.3, which increased the model's sensitivity to more subtle indicators of facial pain. These adjustments were informed by rigorous data analysis and a systematic optimization approach. The results were precise: the model became more efficient at detecting pain and showed an improved ability to operate reliably in various scenarios, bringing us closer to our goal of creating a pain detection system that is accurate and applicable in real-world environments.

4.5 Performance comparison with existing methods

Various approaches have been explored in pain detection, each with strengths and application areas. Our hybrid AI model, which integrates facial gesture and paralinguage analysis, compares favorably to several prominent methodologies based on standard performance metrics.

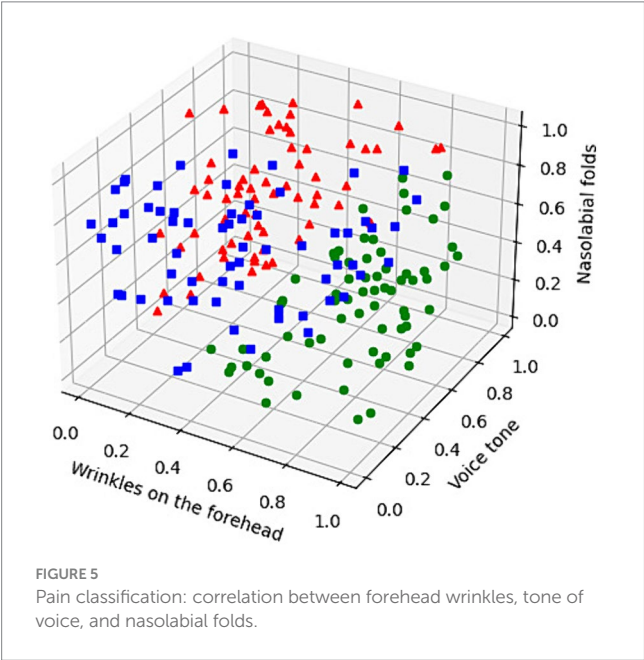
The first method compared, the Traditional SVM classifier, is based on traditional machine learning for facial expression classification. Despite its high precision and specificity, our model outperforms this approach with a better overall balance between all metrics, including a higher AUC, suggesting better discrimination between pain classes. Using CNN for facial images, the following approach is notable for its ability to process complex visual data.

TABLE 5 Characteristics in pain detection: analysis of facial expressions and paralanguage.

Characteristic	Importance (%)	Description
wrinkles on the forehead	18%	Frowning is common in expressions of pain.
Eye-opening	12%	Widely opened eyes may indicate surprise or tension due to pain.
Depth of nasolabial folds	15%	The marking of lines around the nose and mouth is associated with pain.
Voice tone	20%	Acute variations in tone can denote pain.
Speech rhythm	8%	Changes in speech rate may reflect responses to pain.
speech intensity	10%	Increased vocal intensity may indicate pain.
lip compression	5%	Tight lips are a common reaction to pain.
Eyebrow lift	6%	Raising your eyebrows can be an indicator of pain.
Nod	3%	An inclination or rotation of the head can show discomfort.
Duration of facial expressions	3%	The length of time that certain expressions are maintained can indicate the intensity of the pain.

TABLE 6 Tweaking and optimizing features for improved pain detection.

Characteristic	Modified parameter	Previous value	Adjusted value	Impact on the model
wrinkles on the forehead	Detection sensitivity	0.5	0.75	Improved identification of slight puckers
Eye-opening	Opening threshold	0.3	0.2	Increased surprise or tension detection precision
nasolabial folds	Detection depth	0.4	0.6	More excellent discrimination in the intensity of the pain gesture
Voice tone	Frequency range	100-3000 Hz	80-3500 Hz	Improved detection of subtle variations in pain
Speech rhythm	speaking speed	100 words/min	90 words/min	Increased sensitivity to changes in speech rate
speech intensity	Intensity threshold	0.6	0.4	Improved capture of volume fluctuations associated with pain



However, despite CNN's ability to identify detailed image patterns, our hybrid model shows greater sensitivity and specificity, reflecting a remarkable ability to detect and rule out pain instances correctly.

The hybrid sensor system combines biometric data with audio analysis for a multifaceted approach to pain detection. When compared, our model matches in precision and offers greater

sensitivity, indicating a superior ability to correctly identify pain cases, a critical advantage in clinical scenarios. The quantitative results of this comparison are summarized in Table 8.

5 Discussion

The literature review suggests a variety of methodologies applied in the detection and classification of pain. Traditional approaches, such as the use of SVM, have proven to be instrumental in establishing a solid foundation for the computational understanding of pain expressions (Höfel et al., 2021; Wang et al., 2023). However, recent studies, including those employing CNNs, highlight the trend toward more advanced approaches that exploit the capabilities of deep learning to interpret complex visual data. On the other hand, hybrid systems that combine diverse types of data point towards integrating multiple sources of information as the immediate future in pain detection (Vorkachev, 2022; Yue et al., 2024).

Our work aligns with these current trends and expands on them by combining paralanguage and facial gesture analysis techniques. By doing so, our model overcomes the individual limitations of each previous approach and establishes itself as a comprehensive solution. With a precision of 92% and a specificity of 95%, our model demonstrates a balance between sensitivity and generalizability, which is crucial for practical applications in natural and diverse environments. This represents a significant improvement in pain detection, where precision is an ethical imperative and a practical requirement for effective medical interventions.

TABLE 7 Hyperparameter optimization: improvements and adjustments in the neural network for pain detection.

Hyperparameter/process	Previous value/setting	New value/settings	Observed impact on the model
Number of layers in the neural network	3 layers	4 layers	Increased ability to model complex patterns
Number of units per layer	64 units	128 units	Improved detection of subtle features
Learning rate	0.01	0.001	More stable convergence and reduced overfitting
Regularization (dropout)	0.2	0.3	Reducing variability and improving generalization
Weight initialization method	Random	I have normal	Optimization of initial training
Optimization algorithm	SGD	Adam	Improvements in convergence speed and quality
Batch size	32	64	Improved computational efficiency and stability
Frequency range in voice analysis	80-3000 Hz	50-4000 Hz	Expanded detection of pain-related voice tones
Wrinkle detection threshold	Intensity >0.5	Intensity >0.3	Increased sensitivity to less pronounced wrinkles

TABLE 8 Comparison of performance metrics with existing methods.

Method	Precision	Recall	AUC	F1-Score	Specificity
Our model	92%	90%	93%	91%	95%
Traditional SVM	89%	85%	90%	87%	90%
CNN for images	87%	88%	85%	86%	89%
Hybrid sensor system	90%	86%	88%	88%	91%

The innovation and relevance of our model are evidenced by its multifaceted approach and superior performance, especially in recognizing subtle and complex patterns of pain expressions, potentially transforming pain assessment and treatment. With pain detection positioned as a critical challenge in modern healthcare, our model stands out for its high precision and generalizability, addressing one of the most difficult challenges in healthcare: correctly interpreting pain—the full range of human expressions of pain (De Sario et al., 2023).

This research expands the frontiers of automatic pain detection, providing a tool that could radically change pain management, alleviating human suffering through faster, more accurate identification and response (Khalifa and Faddal, 2017; Quintas et al., 2023). The importance of this work lies in its potential integration into patient monitoring systems, telemedicine applications, and clinical practice, thus significantly improving the quality of patient care. In short, our model is not just a technical advance; it is a step forward in humanizing technology in health service.

During the development and evaluation of the proposed model, several limitations and challenges were encountered that deserve to be discussed. One of the main challenges was collecting a sufficiently diverse and representative data set. Although the data set used consisted of 200 participants with an equal distribution regarding gender and socioeconomic background, there are still limitations regarding geographical and cultural diversity. This may affect the generalization of the model to different populations.

Pain is a highly subjective experience and can manifest differently in different people. This variability in facial expressions and paralinguistic presented a significant challenge to the model. Despite efforts to capture a wide range of pain expressions, some signals may not have been adequately represented in the data set, affecting the model's accuracy in some instances.

Integrating facial gestures and paralinguistic data involved significant technical challenges. Temporal synchronization of the two

modalities was crucial to ensure consistency in the analysis but also presented difficulties due to differences in the sampling rate and quality of data collected by the cameras and microphones. Preprocessing large volumes of video and audio data requires extensive computational resources. Data cleaning, denoising, and normalization required advanced techniques to ensure the high quality of the input data. Additionally, the need to perform data augmentation to improve the diversity of the data set added complexity to the process.

Although the model performed well on the test data set, its generalizability to other contexts and populations still needs to be validated. The lack of external data for independent validation is a limitation that should be addressed in future studies. Despite efforts to mitigate biases, the model may still be influenced by inherent biases in the data set. This could affect the model's fairness in detecting pain among different demographic groups. It is essential to continue investigating and addressing these biases to improve the model's fairness.

Despite these limitations and challenges, the proposed model has proven effective in pain detection using a combination of facial gestures and paralinguistic analysis. Future work should address these limitations by collecting more diverse data sets, improving multimodal integration techniques, and externally validating the model.

6 Conclusion

This work has addressed the complex task of detecting and classifying pain using advanced AI technologies. It combines the analysis of facial gestures and paralinguistic. Through a multidimensional approach, we have developed a hybrid AI model that identifies the physical manifestations of pain and interprets the associated vocal signals.

Our study has shown that integrating multiple data sources significantly improves the precision and robustness of the pain detection system. Our model outperforms traditional methods with an overall precision that reaches 92%, a recall of 90%, and an AUC of 93%. This highlights the effectiveness of deep learning techniques in recognizing complex and subtle patterns associated with pain. This high level of performance underlines the model's ability to effectively differentiate between pain and non-pain states, which is crucial in clinical applications where precision can directly affect patient well-being.

Furthermore, the optimization and refinement of the model, based on the exhaustive analysis of the data and the validation results, have allowed a notable improvement in the sensitivity and specificity of the system. Adjustments in hyperparameters, such as the number of layers in the neural network, the learning rate, and regularization methods, have resulted in a more efficient system adapted to pain detection. Comparison with other methodologies revealed that, although there are numerous approaches to pain detection, our model stands out for its integrative approach and generalizability. This comparison highlights the importance of a holistic analysis and the need for systems that can adapt to the variability inherent in human expressions of pain.

Looking to the future, several areas are identified for the continuation of this work. One promising direction is the exploration of explanatory AI (XAI) to improve the transparency and understandability of model decisions. This would facilitate the acceptance of the technology among healthcare professionals and provide valuable insights into the nature of pain and its manifestation. Another avenue of research could focus on the personalization of the model, adjusting its operation to the individual characteristics of the patients. This could include developing pain profiles based on clinical histories, personal preferences, and treatment responses, thereby optimizing pain management interventions.

Additionally, integrating new data modalities, such as physical activity logs, sleep patterns, and biometric responses, could provide a more complete view of the patient's condition, improving the system's ability to detect and respond to pain in a broader context of health and well-being. Interdisciplinary collaboration will also play a crucial role in the evolution of this technology. Working with medical, psychology, neuroscience, and ethics experts will ensure that the solutions developed are technically sound, socially responsible, and ethically viable.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: data supporting this study's findings are available upon request from the corresponding author. The data is not made public to preserve privacy and confidentiality. Any data access request should

be directed to WV-C, who can be contacted at william.villegas@udla.edu.ec.

Ethics statement

Ethical approval was not required for the study involving human samples in accordance with the local legislation and institutional requirements because the study is not considered as medical research. The study's objective is the design of software to help identify gestures and does not use methods or devices that may be invasive for humans or animals. Written informed consent was obtained from the individual(s) for the publication of any identifiable images or data included in this article.

Author contributions

RG: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. JG-O: Data curation, Software, Validation, Visualization, Writing – original draft. WV-C: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Adusumalli, H., Kalyani, D., Sri, R. K., and Pratapjea, M., Rao PVRDP. Face mask detection using open CV. In: Proceedings of the 3rd international conference on intelligent communication technologies and virtual Mobile networks, ICICV 2021. (2021).
- Arashloo, S. R., Kittler, J., and Christmas, W. (2017). An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol. *IEEE Access* 5, 13868–13882. doi: 10.1109/ACCESS.2017.2729161
- Ayuningsih, A. A., Mahmud, M., Salija, K., and Muhayyang, M. (2022). Teachers' paralinguistic in classroom interaction. *Retorika Jurnal Bahasa, Sastra, Dan Pengajarannya* 15, 73–85.
- Azzahra, Y., Muhayyang, M., and Korompot, C. A. (2022). The impacts of Teachers' paralinguistic in EFL Classroom. *Journal of excellence in English language. Education* 1, 1–13.
- Benavent-Lledo, M., Mulero-Pérez, D., Ortiz-Perez, D., Rodriguez-Juan, J., Berenguer-Agullo, A., Psarrou, A., et al. (2023). A comprehensive study on pain assessment from multimodal sensor data. *Sensors* 23, 9675. doi: 10.3390/s23249675
- Béra-Louville, B. L., Barfety-Servignat, B. S., Jm, B., Veys, V., Pickering, P., Catsaros, C., et al. (2019). Diagnostic recommendations and treatment Management for Complex Regional Pain Syndromes: the Lille recommendations. *Douleur et Analgesie*. 32, 155–64. doi: 10.3166/dea-2019-0074

- Borna, S., Haider, C. R., Maita, K. C., Torres, R. A., Avila, F. R., Garcia, J. P., et al. (2023). A review of voice-based pain detection in adults using artificial intelligence. *Bioengineering* 10:500. doi: 10.3390/bioengineering10040500
- Chen, C., Han, D., and Chang, C. C. (2024). MPCCT: Multimodal vision-language learning paradigm with context-based compact Transformer. *Pattern Recognit* 147:110084. doi: 10.1016/j.patcog.2023.110084
- Chen, C., Han, D., and Shen, X. (2023). CLVIN: Complete language-vision interaction network for visual question answering. *Knowl Based Syst.* 275:110706. doi: 10.1016/j.knsys.2023.110706
- Chu, Y., Zhao, X., Han, J., and Su, Y. (2017). Physiological signal-based method for measurement of pain intensity. *Front. Neurosci.* 11:11. doi: 10.3389/fnins.2017.00279
- De Sario, G. D., Haider, C. R., Maita, K. C., Torres-Guzman, R. A., Emam, O. S., Avila, F. R., et al. (2023). Using AI to detect pain through facial expressions: a review. *Bioengineering* 10:548. doi: 10.3390/bioengineering10050548
- Gutiérrez, P. J. (2022). datos personales en el Ecuador como un derecho humano una necesidad de mejoramiento en su regulación. *Revista Jurídica Crítica y Derecho.* 3, 53–66. doi: 10.29166/cyd.v3i5.3950
- Höfel, L., von der Beek, J., Draheim, N., Haas, J. P., and Schramm, A. (2021). Pain disorders in children and adolescents. *Pädiatrische Praxis* 96, 108–129. doi: 10.1192/apt.bp.114.014068
- Hu, X. S., Nascimento, T. D., Bender, M. C., Hall, T., Petty, S., O'Malley, S., et al. (2019). Feasibility of a real-time clinical augmented reality and artificial intelligence framework for pain detection and localization from the brain. *J. Med. Internet Res.* 21:13594. doi: 10.2196/13594
- Huang, J., Shang, Y., and Chen, H. (2019). Improved Viola-Jones face detection algorithm based on Holo Lens. *Eurasip J. Image Video Process.* 2–11. doi: 10.1186/s13640-019-0435-6
- Irshad, R. R., Shaman, F., Alalayah, K. M., Alwayle, I. M., Hazber, M. A. G., Aqlan, A. M., et al. (2023). Convolutional neural network enable optoelectronic system for predicting cardiac response by analyzing auction-based optimization algorithms. *J. Nanoelectron. Optoelectron.* 17, 1274–1282. doi: 10.1166/jno.2022.3329
- Khalifa, E. M., and Faddal, H. (2017). Impacts of using paralinguistic on teaching and learning English language to convey effective meaning. *Studies in English. Lang. Teach.* 5:295. doi: 10.22158/selt.v5n2p295
- Liao, L., Guo, Z., Gao, Q., Wang, Y., Yu, F., Zhao, Q., et al. (2023). Color image recovery using generalized matrix completion over higher-order finite dimensional algebra. *Axioms* 12, 954. doi: 10.3390/axioms12100954
- Luangrath, A. W., Xu, Y., and Wang, T. (2023). Paralinguistic classifier (PARA): an algorithm for automatic coding of paralinguistic nonverbal parts of speech in text. *J. Mark. Res.* 60, 388–408. doi: 10.1177/00222437221116058
- McGrath, H., Flanagan, C., Zeng, L., and Lei, Y. (2019). Future of artificial intelligence in anesthetics and pain management. *J. Biosci. Med.* 7, 111–118. doi: 10.4236/jbm.2019.711010
- Mekruksavanich, S., and Jitpattanakul, A. (2021). Deep convolutional neural network with rnns for complex activity recognition using wrist-worn wearable sensor data. *Electronics (Switzerland)* 10:141685. doi: 10.3390/electronics10141685
- Nagireddi Meng, J. N., Vyas, A. K., Sanapati, M. R., Soin, A., and Manchikanti, L. (2022). The analysis of pain research through the Lens of artificial intelligence and machine learning. *Pain Physician* 25, e211–e243.
- Nguyen, H. D., Tran, K. P., Thomassey, S., and Hamad, M. (2021). Forecasting and anomaly detection approaches using LSTM and LSTM autoencoder techniques with the applications in supply chain management. *Int. J. Inf. Manag.* 57:102282. doi: 10.1016/j.ijinfomgt.2020.102282
- Oualla, M., Ounachad, K., and Sadiq, A. (2021). Building face detection with face divine proportions. *Int J Online Biomed. Eng.* 17, 63–80. doi: 10.3991/ijoe.v17i04.19149
- Quintas, S., González-Martínez, A., and Gago-Veiga, A. B. (2023). Artificial intelligence in migraine and other headaches, vol. 18: Kranion.
- Sandeep, P. V. K., and Kumar, N. S. (2024). Pain detection through facial expressions in children with autism using deep learning. *Soft. Comput.* 28, 4621–4630. doi: 10.1007/s00500-024-09696-x
- Shi, S., Han, D., and Cui, M. (2023). A multimodal hybrid parallel network intrusion detection model. *Conn. Sci.* 35:2227780. doi: 10.1080/09540091.2023.2227780
- Vorkachev, S. G. (2022). Somatic verbalization: a paralinguistic of gratitude (according to Corpus linguistics). *Curr. Issues Philol. Pedagogical Linguist.* 1, 243–250. doi: 10.29025/2079-6021-2022-1-243-250
- Wang, H., Han, D., Cui, M., and Chen, C. (2023). NAS-YOLOX: a SAR ship detection using neural architecture search and multi-scale attention. *Conn. Sci.* 35, 1–32. doi: 10.1080/09540091.2023.2257399
- Yue, J. M., Wang, Q., Liu, B., and Zhou, L. (2024). Postoperative accurate pain assessment of children and artificial intelligence: a medical hypothesis and planned study. *World J. Clin. Cases* 12, 681–687. doi: 10.12998/wjcc.v12.i4.681



OPEN ACCESS

EDITED BY

Alberto Santamaria-Pang,
Microsoft, United States

REVIEWED BY

Imran Iqbal,
New York University, United States
Mladen Andić,
University of Belgrade, Serbia

*CORRESPONDENCE

Fuqiang Zhou,
✉ pandayyzx@163.com

RECEIVED 08 October 2024

ACCEPTED 09 December 2024

PUBLISHED 20 December 2024

CITATION

Wen B, Li C, Cai Q, Shen D, Bu X and Zhou F (2024) Multimodal MRI radiomics-based stacking ensemble learning model with automatic segmentation for prognostic prediction of HIFU ablation of uterine fibroids: a multicenter study.
Front. Physiol. 15:1507986.
doi: 10.3389/fphys.2024.1507986

COPYRIGHT

© 2024 Wen, Li, Cai, Shen, Bu and Zhou. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Multimodal MRI radiomics-based stacking ensemble learning model with automatic segmentation for prognostic prediction of HIFU ablation of uterine fibroids: a multicenter study

Bing Wen¹, Chengwei Li², Qiuyi Cai², Dan Shen¹, Xinyi Bu¹ and Fuqiang Zhou^{1*}

¹Department of Radiology, Yiyang Central Hospital, Yiyang, China, ²Department of Radiology, The Third People's Hospital of Chengdu, Chengdu, China

Objectives: To evaluate the effectiveness of an MRI radiomics stacking ensemble learning model, which combines T2-weighted imaging (T2WI) and contrast-enhanced T1-weighted imaging (CE-T1WI) with deep learning-based automatic segmentation, for preoperative prediction of the prognosis of high-intensity focused ultrasound (HIFU) ablation of uterine fibroids.

Methods: This retrospective study collected data from 360 patients with uterine fibroids who underwent HIFU treatment. The dataset was sourced from Center A (training set: N = 240; internal test set: N = 60) and Center B (external test set: N = 60). Patients were categorized into favorable and unfavorable prognosis groups based on the post-treatment non-perfused volume ratio. Automated segmentation of uterine fibroids was performed using a V-net deep learning models. Radiomics features were extracted from T2WI and CE-T1WI, followed by data preprocessing including normalization and scaling. Feature selection was performed using *t*-test, Pearson correlation, and LASSO to identify the most predictive features for preoperative prognosis Support Vector Machine (SVM), Random Forest (RF), Light Gradient Boosting Machine (LightGBM), and Multilayer Perceptron (MLP) were employed as base learners to construct base predictive models. These models were integrated into a stacking ensemble model, with Logistic Regression serving as the meta-learner to combine the outputs of the base models. The performance of the models was assessed using the area under the receiver operating characteristic curve (AUC).

Results: Among the base models developed using T2WI and CE-T1WI features, the MLP model exhibited superior performance, achieving an AUC of 0.858 (95% CI: 0.756–0.959) in the internal test set and 0.828 (95% CI: 0.726–0.930) in the external test set. It was followed by the SVM, LightGBM, and RF, which obtained AUC values of 0.841 (95% CI: 0.737–0.946), 0.823 (95% CI: 0.711–0.934), and 0.750 (95% CI: 0.619–0.881), respectively. The stacking ensemble learning

model, which integrated these five algorithms, demonstrated a notable enhancement in performance, with an AUC of 0.897 (95% CI: 0.818–0.977) in the internal test set and 0.854 (95% CI: 0.759–0.948) in the external test set.

Conclusion: The DL based automatic segmentation MRI radiomics stacking ensemble learning model demonstrated high accuracy in predicting the prognosis of HIFU ablation of uterine fibroids.

KEYWORDS

artificial intelligence, uterine fibroids, high intensity focused ultrasound, ensemble stacking model, magnetic resonance imaging

1 Introduction

Uterine fibroids are the most prevalent benign tumors of the female reproductive system, characterized by high vascularity, with incidence rates increasing with age. The symptoms associated with fibroids can appreciably impair women's quality of life (Stewart et al., 2017; Bulun, 2013). Traditional treatment modalities primarily encompass hysterectomy, myomectomy, laparoscopic myomectomy, and uterine artery embolization (Lethaby et al., 2017; Donnez et al., 2015; Jacoby et al., 2020; McLucas et al., 2001). Recently, non-invasive high-intensity focused ultrasound (HIFU) has gained prominence in the management of symptomatic uterine fibroids and is now included in treatment guidelines in several countries, with its effectiveness extensively corroborated (Lyon et al., 2020; Chen et al., 2018; Liu et al., 2020). Nevertheless, due to the inherent properties of fibroid tissue and technical constraints, HIFU may not be suitable for all patients (Huang et al., 2019; Zhang et al., 2020; Verpalen et al., 2019; Verpalen et al., 2020). Therefore, precise preoperative evaluation is imperative for the successful application of HIFU. Enhancing the accuracy of preoperative predictions regarding HIFU treatment efficacy is crucial for clinicians to formulate optimal treatment strategies.

Preoperative magnetic resonance imaging (MRI) with HIFU is instrumental in the differential diagnosis of uterine fibroids, as well as in assessing their location, morphology, and potential tissue composition (Liao et al., 2023; Venkatesan et al., 2012). MRI is recommended as an essential preoperative evaluation tool for HIFU treatment of uterine fibroids. The imaging characteristics of fibroids across various MRI sequences exhibit a notable correlation with their tissue heterogeneity. Previous research has elucidated that T2-weighted imaging (T2WI) offers superior visualization of the internal architecture and morphology of uterine fibroids, facilitating the differentiation of cellular hydration levels from fibrous tissue composition (Zhao et al., 2013; Zhao et al., 2015; Funaki et al., 2007). Moreover, contrast-enhanced T1-weighted imaging (CE-T1WI), achieved through the administration of contrast agents, augments the delineation of fibroid boundaries relative to the surrounding tissues, thereby enhancing the assessment of vascularization (Keserci and Duc, 2017; Yoon et al., 2010). Consequently, it assists in evaluating the proliferative activity and invasive potential of the fibroids. These imaging characteristics are instrumental in enabling clinicians to subjectively assess the tissue characteristics of fibroids and forecast the efficacy of ultrasound energy in inducing coagulative necrosis in the targeted region. Despite these advances, challenges remain due

to the variability in clinical experience and the inherent limitations of human visual assessment in accurately evaluating fibroid morphology, spatial distribution, and lesion characteristics (Dou et al., 2024).

Radiomics is revolutionizing medical imaging by enabling the detailed analysis of tumor complexity at a microscopic level through technological advancements (Guiot et al., 2022; Gillies et al., 2016). By utilizing high-throughput imaging data, radiomics overcomes the limitations of traditional imaging, revealing subtle features that are often imperceptible to the naked eye (Lambin et al., 2017). This cutting-edge technology not only enhances the precision of tumor diagnosis but also facilitates personalized treatment, providing unparalleled insight into the potential threats posed by tumors. In the preoperative assessment of HIFU ablation of uterine fibroids, radiomics models have demonstrated significant clinical potential. However, previous studies have primarily relied on radiomics features from single MRI sequences for efficacy prediction (Cheng et al., 2024; Li et al., 2023). While some progress has been achieved, single sequences often fail to comprehensively represent the biological information of fibroids, leading to limited predictive accuracy. By integrating information from different MRI sequences, a more comprehensive understanding of the tissue characteristics and biological properties of fibroids can be achieved, which significantly enhancing the accuracy and stability of prediction models (Zheng et al., 2021). Nonetheless, existing research has not fully incorporated the CE-T1WI sequence, which most accurately reflects fibroid blood supply, and has overlooked the importance of including diverse MRI sequences in the model to enhance its performance. To facilitate this integration, deep learning (DL) based automatic segmentation plays a crucial role by providing precise delineation of fibroid boundaries and relevant anatomical structures across multiple MRI sequences (Wang et al., 2014). As Imran Iqbal demonstrated, deep learning can effectively extract disease-related information and address challenges such as limited annotated data by leveraging pre-trained models, which has proven beneficial in detecting various medical conditions, including joint abnormalities and skin cancer screening (Iqbal et al., 2021; Iqbal et al., 2020). Such methodologies not only enhance the accuracy of segmentation but also provide a solid foundation for the subsequent extraction of radiomic features, ensuring both the efficiency and consistency of the analysis. The application of automated segmentation techniques allows researchers to rapidly process large volumes of imaging data, thereby improving the reliability and predictive power of model construction. However, in the realm of machine learning applications, traditional single algorithms and hyperparameter

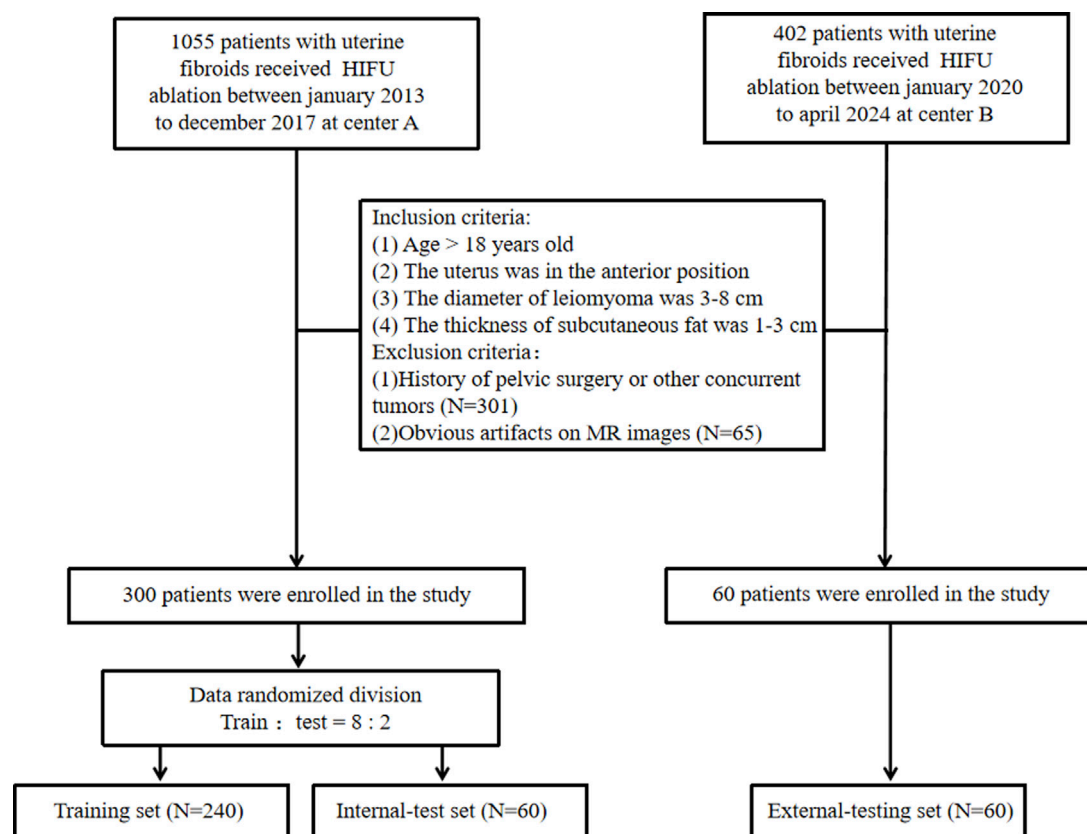


FIGURE 1
Flowchart of patient enrollment and exclusion.

tuning methods, although effective, may not fully exploit the potential of the radiomics features. Stacking, as an advanced ensemble learning method, demonstrates substantial potential in enhancing predictive performance (Naimi and Balzer, 2018). Stacking first trains multiple types of base learners to capture diverse features and patterns within the data. Subsequently, the outputs from these base learners are used as inputs to train a meta-learner, which integrates the strengths of the base learners and addresses their shortcomings, ultimately producing more accurate results. This method leverages the advantages of multiple models to effectively handle complex data, reduce overfitting, and improve prediction accuracy and stability (Sesmero et al., 2015). This study aims to develop a multimodal MRI radiomics analysis method that integrates automated segmentation and stacking ensemble learning techniques to further enhance the predictive performance of HIFU ablation of uterine fibroids.

2 Materials and methods

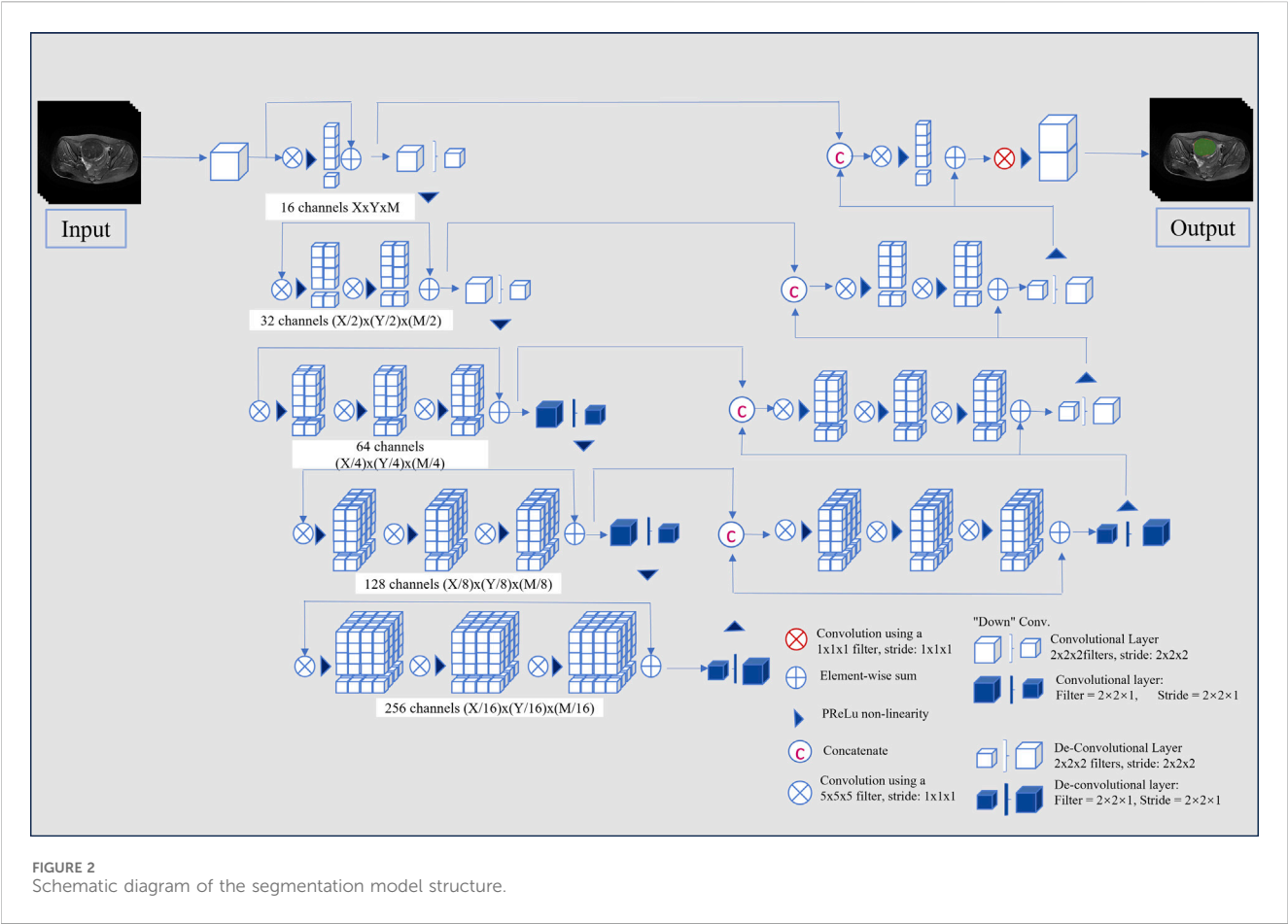
2.1 Patients

We conducted a retrospective analysis involving 1457 consecutive patients diagnosed with uterine fibroids across two medical centers. To accurately assess the efficacy of HIFU ablation for uterine fibroids while controlling for various factors

on efficacy, such as fibroid size, location, tissue composition, abdominal wall thickness, and the presence of abdominal wall scars. We established the following inclusion criteria (Stewart et al., 2017): patients over 18 years old (Bulun, 2013); no prior surgical interventions or relevant medication history (Lethaby et al., 2017); fibroids located in the anterior uterine position (Donnez et al., 2015); fibroids measuring between 3–8 cm in diameter (Jacoby et al., 2020); abdominal fat thickness ranging from 1 to 3 cm; and (McLucas et al., 2001) for patients with multiple fibroids, only the largest fibroid was included. The exclusion criteria were (Stewart et al., 2017): history of pelvic surgery or other concurrent tumors; and (Bulun, 2013) imaging artifacts that interfered with accurate assessment. A non-perfused volume ratio (NPVR) $\geq 80\%$ was used to define treatment efficacy, a threshold validated across different levels of physician expertise (Gong et al., 2022). Thus, an NPVR of $\geq 80\%$ was considered indicative of a favorable prognosis, while an NPVR $< 80\%$ was deemed an unfavorable prognosis. NPVR assessments were independently performed by two radiologists: one with 4 years and the other with 15 years of diagnostic experience. Discrepancies were resolved in favor of the more experienced radiologist. After a stringent screening process, we included 300 patients from Center A and 60 from Center B, dividing them into three groups: a training set (N = 240), an internal test set (N = 60), and an external test set (N = 60). Approval for the study was granted by the Institutional Review Boards of both centers (Center A: approval number 2022-K129,

TABLE 1 MRI acquisition parameters.

Parameters	Signa HDxt		Signa voyager	
	T2WI	CE-T1WI	T2WI	CE-T1WI
Magnetic field strength	3.0 T		1.5 T	
Repetition time (TR)	270	3.84	4,000	6
Echo time (TE)	2.1	1.81	68	2
Feld of view (FOV)	98.1 × 38	68.4 × 26.5	60 × 60	60 × 60
Slice thickness (mm)/gap (mm)	6/2	4/2	6/2	4/2
matrix	512 × 512	512 × 512	256 × 256	256 × 256



Center B: approval number 2024-J-29), and the need for written informed consent was waived. The patient enrollment process is detailed in [Figure 1](#).

8-channel phased array coil for the abdomen. Detailed MRI acquisition parameters are shown in [Table 1](#).

2.2 Images acquisition

MRI scans were acquired from two centers: one using a 3.0 T Signa HDxt MRI scanner and the other using a 1.5 T Signa Voyager MRI scanner, both provided by General Electric. Patients were positioned in the supine position and scanned using a dedicated

2.3 Image segmentation and feature extraction

This study employed a V-Net architecture for automatic segmentation of pelvic MRI data, specifically targeting uterine fibroids in T2WI and CE-T1WI ([Milletari et al., 2016](#)). The network is based on an encoder-decoder architecture. The

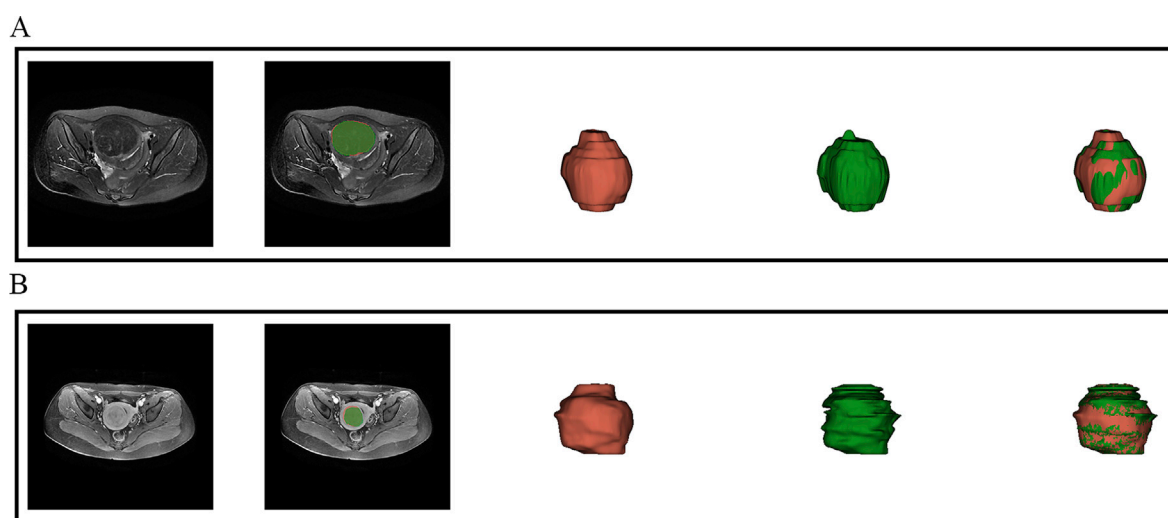


FIGURE 3
Visual inspection of the segmentation results. (A) presents the automatic segmentation results on T2WI for a 48-year-old patient diagnosed with uterine fibroids, while (B) presents the automatic segmentation results on CE-T1WI for a 39-year-old patient also diagnosed with uterine fibroids.

encoder uses 3D convolutional modules to extract features from medical images and adjusts the feature resolution through convolution operations with a stride of 2. In the decoder, 3D transposed convolutions are used to progressively restore the deep semantic features extracted by the encoder to a higher resolution. Skip connections are incorporated between the encoder and decoder to effectively combine low-level and high-level features, thereby enhancing segmentation accuracy. The network includes three auxiliary loss layers and one main loss layer; the main loss layer employs 3D transposed convolutions to recover the feature maps to the original image size, ultimately outputting the automatically segmented target regions. Figure 2 illustrates the architecture of the automatic segmentation network, while Figure 3 presents the segmentation results. We evaluated the model's performance on the validation data set, which yielded an average Dice Similarity Coefficient (DSC) of 0.883 for the T2WI segmentation model and 0.809 for the CE-T1WI segmentation model, indicating good performance in the automatic segmentation for uterine fibroids.

After the automatic segmentation of the remaining MRI scans, two independent radiologists evaluated the segmented regions. For images with inaccurate outlines, manual corrections were made using ITK-SNAP software (version 3.8, <http://www.itksnap.org>) to create complete Regions of Interest (ROIs). Prior to feature extraction, all MRI images and segmentations were resampled to a voxel size of $1 \times 1 \times 1 \text{ mm}^3$ using bilinear interpolation. Features were extracted from the T2WI and CE-T1WI ROIs using Python (version 3.10, <https://www.python.org>), focusing on both low-dimensional and high-dimensional aspects. Low-dimensional features comprised shape and first-order histogram metrics, while high-dimensional features included texture characteristics derived from various matrices: gray-level co-occurrence matrix (GLCM), gray-level run-length matrix (GLRLM), gray-level size-zone matrix (GLSZM), neighborhood gray-tone difference matrix (NGTDM), and gray-level dependence matrix (GLDM). Additionally, texture

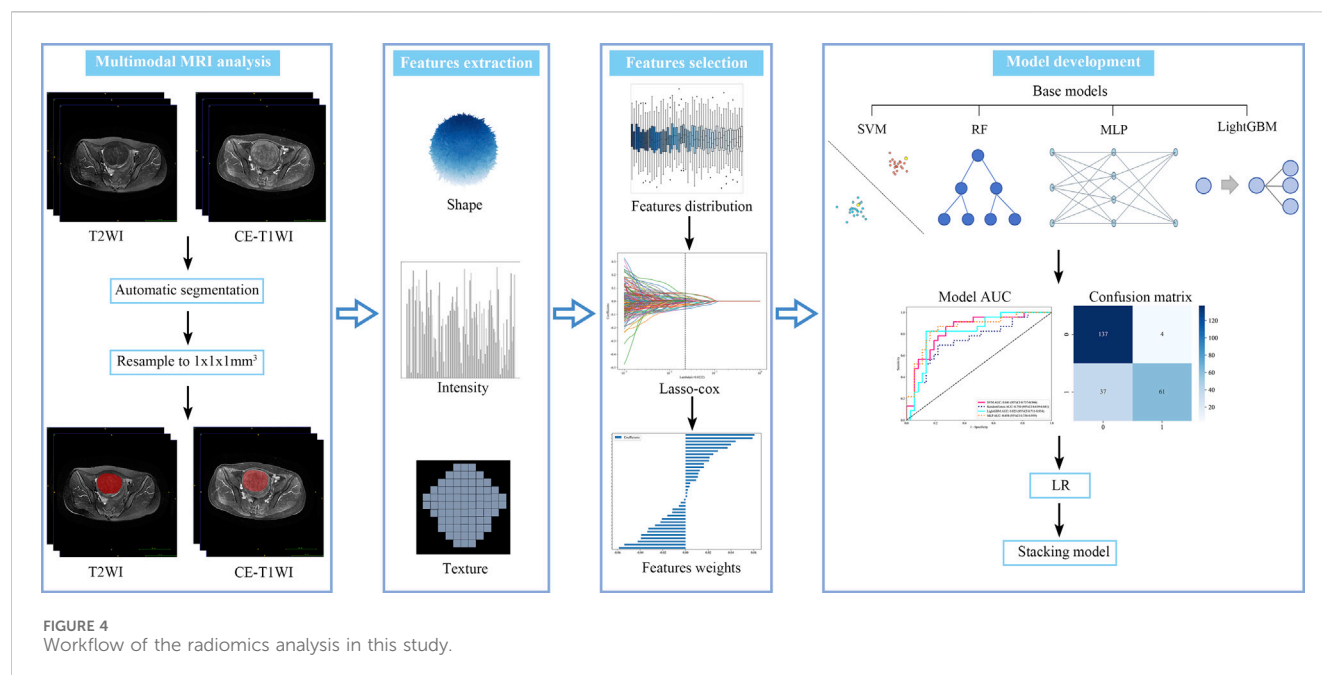
features were analyzed in the Gaussian Laplacian filter domain (core sizes ranging from 2.0 to 5.0 mm) and the wavelet filter domain.

2.4 Feature selection

To assess the interobserver repeatability of radiomic features, the intraclass correlation coefficient (ICC) was calculated using data from 100 randomly selected patients at center A. Features with ICC values greater than 0.8 were deemed highly consistent and were included for further analysis. To harmonize radiomic features across different MRI scanners, the ComBat method was applied (Orlhaac et al., 2022). To address biases introduced by imbalanced data distributions, which could affect model performance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to increase the number of samples in underrepresented classes, thereby improving model robustness (Chawla et al., 2002). Subsequently, feature selection was conducted on the harmonized T2WI and CE-T1WI features. Initially, a *t*-test was employed to filter features with significant correlations to treatment outcomes, retaining those with strong associations. Pearson's correlation coefficient was then computed to analyze the relationships among these features, retaining features with coefficients above 0.8. Finally, the Least Absolute Shrinkage and Selection Operator (LASSO) regression was applied for further refinement, aiming to reduce dimensionality and select the most relevant features for analysis.

2.5 Construction of stacking ensemble learning model

Stacking ensemble learning is an advanced machine learning technique that combines the outputs of multiple base learners using a meta-learner to improve predictive performance. In this study,



four traditional machine learning algorithms were employed as base learners: Support Vector Machine (SVM), Random Forest (RF), Multilayer Perceptron (MLP), and Light Gradient Boosting Machine (LightGBM). These base learners formed the first layer of the ensemble model. A Logistic Regression (LR) algorithm was selected as the meta-learner for the second layer due to its strong generalization capabilities and effectiveness in combining the predictions of the base learners, while addressing their biases and reducing overfitting. Hyperparameters for each base learner were optimized using Bayesian optimization, which efficiently searches the hyperparameter space to enhance model performance (Snoek et al., 2012). The predictive performance of the ensemble model was evaluated by calculating the area under the receiver operating characteristic curve (AUC). In addition, metrics such as accuracy, sensitivity, and specificity were assessed using the maximum Youden's index to provide a comprehensive evaluation of model performance. The radiomics analysis pipeline is shown in Figure 4.

2.6 Statistical analysis

All statistical analysis was performed using SPSS (version 26.0) and Python (version 3.10). The kappa test was used to analyze the consistency of patient grouping results between two radiologists (Kappa values in the range of 0.80–1.00, good consistency; 0.40 to 0.79, fair consistency; less than 0.40, poor consistency). Continuous variables were described as mean $\pm$ standard deviation and compared by a Mann–Whitney *U* test or *t*-test. Categorical variables were summarized as frequencies and percentages using the chi-square test or Fisher's exact test. The AUC of different models were statistically compared using the DeLong test. A *P*-value <0.05 was considered statistically significant.

3 Results

3.1 Patient characteristics and outcome

In the study, a total of 354 patients were initially enrolled at Center A. Out of these, 54 patients were excluded due to either incomplete or artifact-ridden imaging data. The remaining cohort comprised 172 patients with a favorable prognosis and 128 patients with an unfavorable prognosis. At Center B, 65 patients were screened, with 5 exclusions for the same reasons. This left 35 patients with a favorable prognosis and 25 with an unfavorable prognosis. The inter-rater reliability between the two radiologists was evaluated using the Kappa statistic, which yielded a value of 0.894 ($P < 0.001$), reflecting a strong level of agreement. The clinical characteristics of the patients are detailed in Table 2.

3.2 Feature selection

Initially, 2,394 features were initially extracted from the ROIs in T2WI and CE-T1WI. After applying an ICC threshold of 0.8, 2,376 features were retained for further analysis. To address class imbalance, the SMOTE algorithm was employed to increase the number of unfavorable prognosis cases in Center A by 44 instances. The *t*-test was performed to identify features significantly associated with HIFU prognosis, resulting in 491 features. Then, Pearson correlation analysis was used to filter out features with non-significant correlations, reducing the list to 275 features. Finally, LASSO regression was applied for further feature selection and dimensionality reduction, narrowing the list to 36 features from T2WI and CE-T1WI, which were then used to construct the stacking ensemble learning model. The results of feature selection can be found in Supplementary Material S1.

TABLE 2 Clinical and radiological characteristics of patients.

Characteristics	Traning set (n = 240)		P-Value	Internal test set (N = 60)		P-Value	External test set (N = 60)		P-Value
	NPVR ≥ 80% (n = 137)	NPVR < 80% (n = 103)		NPVR ≥ 80% (n = 35)	NPVR < 80% (n = 25)		NPVR ≥ 80% (n = 38) NPVR < 80% (n = 25)	NPV < 80% (n = 22) NPVR < 80% (n = 25)	
Age (years)	39.63 ± 6.22	39.22 ± 6.17	0.496	39.69 ± 6.83	38.36 ± 7.33	0.475	38.18 ± 8.43	40.47 ± 7.67	0.287
Abdominal Fat Thickness (mm)	16.02 ± 7.63	17.11 ± 9.62	0.108	15.89 ± 5.41	17.08 ± 6.31	0.776	20.44 ± 8.71	18.84 ± 5.94	0.69
Fibroid Type (n (%))			<0.001			0.306			0.165
Submucosal	3 (2.19)	2 (1.94)		1 (2.86)	2 (8.00)		3 (7.89)	0 (0.00)	
Intramural	115 (83.94)	57 (55.34)		32 (91.43)	16 (64.00)		23 (60.53)	18 (81.82)	
Subserosal	19 (13.87)	44 (42.72)		2 (5.71)	7 (28.00)		12 (31.58)	4 (18.18)	
T2WI Signal Intensity			0.809			0.141			0.984
Hypointensity	74 (54.01)	57 (55.34)		15 (42.86)	12 (48.00)		29 (76.32)	17 (77.27)	
Isointensity	27 (19.71)	17 (16.50)		2 (5.71)	5 (20.00)		5 (13.16)	3 (13.64)	
Hyperintensity	36 (26.28)	29 (28.16)		18 (51.43)	8 (32.00)		4 (10.53)	2 (9.09)	
T2WI Signal Homogeneity			0.044			1.000			0.951
Homogeneous	98 (71.53)	60 (58.25)		20 (57.14)	15 (60.00)		16 (42.11)	10 (45.45)	
Inhomogeneous	39 (28.47)	43 (41.75)		15 (42.86)	10 (40.00)		22 (57.89)	12 (54.55)	
CE-T1WI Signal Intensity			<0.001			0.006			0.359
Hypointensity	70 (51.09)	67 (65.05)		17 (48.57)	19 (76.00)		21 (55.26)	8 (36.36)	
Isointensity	49 (35.77)	5 (4.85)		14 (40.00)	1 (4.00)		8 (21.05)	6 (27.27)	
Hyperintensity	18 (13.14)	31 (30.10)		4 (11.43)	5 (20.00)		9 (23.68)	8 (36.36)	
CE-T1WISignal Homogeneity			0.869			0.964			0.816
Homogeneous	44 (32.12)	35 (33.98)		14 (40.00)	11 (44.00)		8 (21.05)	6 (27.27)	
Inhomogeneous	93 (67.88)	68 (66.02)		21 (60.00)	14 (56.00)		30 (78.95)	16 (72.73)	

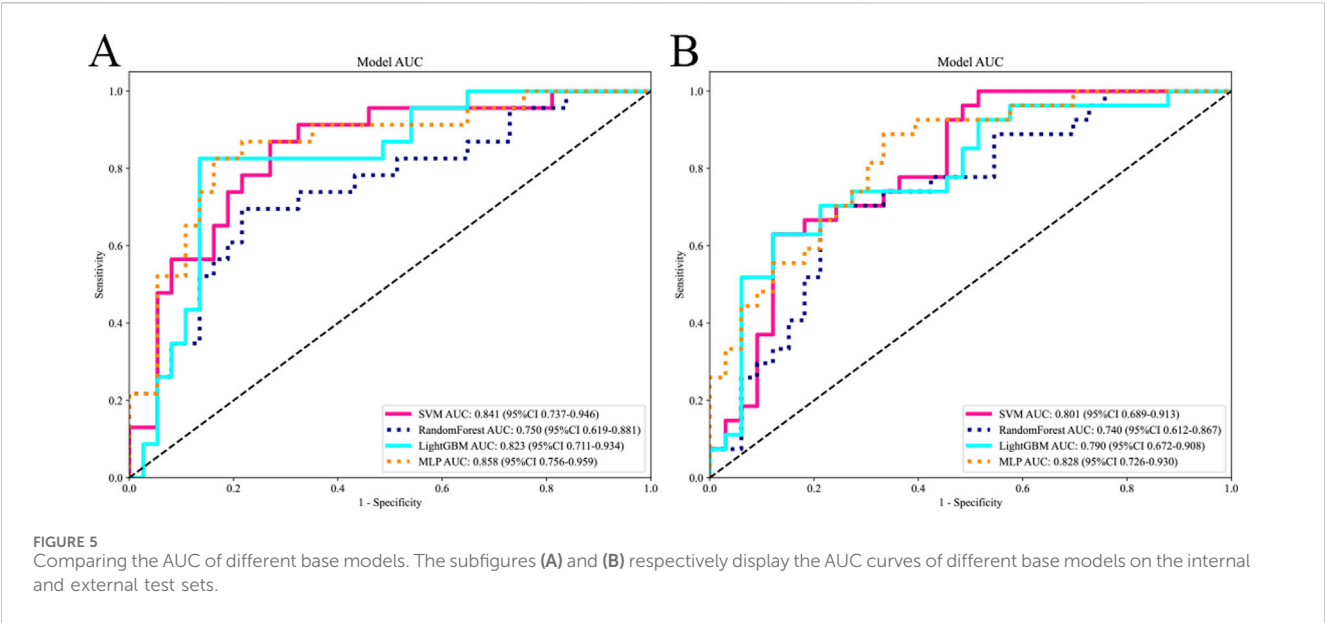


TABLE 3 Performance metrics of base models and stacking ensemble models.

		AUC(95%CI)	Accuracy	Sensitivity	Specificity	Precision	P-Value
SVM	Training set	0.949 (0.737–0.946)	0.879	0.856	0.896	0.864	-
	Internal testing set	0.841 (0.737–0.946)	0.767	0.826	0.730	0.655	0.031
	External testing set	0.801 (0.689–0.913)	0.750	0.593	0.879	0.800	0.017
RF	Training set	0.897 (0.859–0.935)	0.824	0.817	0.830	0.787	-
	Internal testing set	0.750 (0.619–0.881)	0.733	0.652	0.784	0.652	<0.01
	External testing set	0.740 (0.612–0.867)	0.717	0.667	0.758	0.692	<0.01
MLP	Training set	0.927 (0.8967–0.958)	0.849	0.837	0.859	0.821	-
	Internal testing set	0.858 (0.756–0.959)	0.817	0.783	0.838	0.750	0.043
	External testing set	0.828 (0.726–0.930)	0.750	0.852	0.667	0.676	0.028
LightGBM	Training set	0.926 (0.893–0.959)	0.858	0.875	0.844	0.812	-
	Internal testing set	0.823 (0.711–0.934)	0.833	0.783	0.865	0.783	0.026
	External testing set	0.727 (0.598–0.8570)	0.683	0.897	0.484	0.619	0.020
Stacking	Internal testing set	0.897 (0.818–0.977)	0.833	0.850	0.825	0.773	-
	External testing set	0.854 (0.761–0.952)	0.767	0.741	0.788	0.741	-

P values were obtained by performing DeLong test between Stacking ensemble model and base models constructed.

3.3 Performance assessment of different models

The selected features were utilized to construct models with four base learners. Among these, the MLP model showed superior performance, achieving an AUC of 0.858 (95% CI: 0.756–0.959) on the internal test set and 0.828 (95% CI: 0.726–0.930) on the external validation set. This was followed by SVM, LightGBM, and RF, with internal validation set AUCs of 0.841 (95% CI: 0.737–0.946), 0.823 (95% CI: 0.711–0.934), and 0.750 (95% CI: 0.619–0.881), respectively (Figure 5). Detailed performance

evaluations of the models are provided in Table 3. The integration of logistic regression with these four base learners to form an ensemble model led to a substantial enhancement in AUC, reaching 0.897 (95% CI: 0.818–0.977) on the internal test set and 0.854 (95% CI: 0.761–0.952) on the external validation set (Figure 6).

4 Discussion

In this study, a novel multimodal MRI stacking ensemble learning model was developed and independently validated,

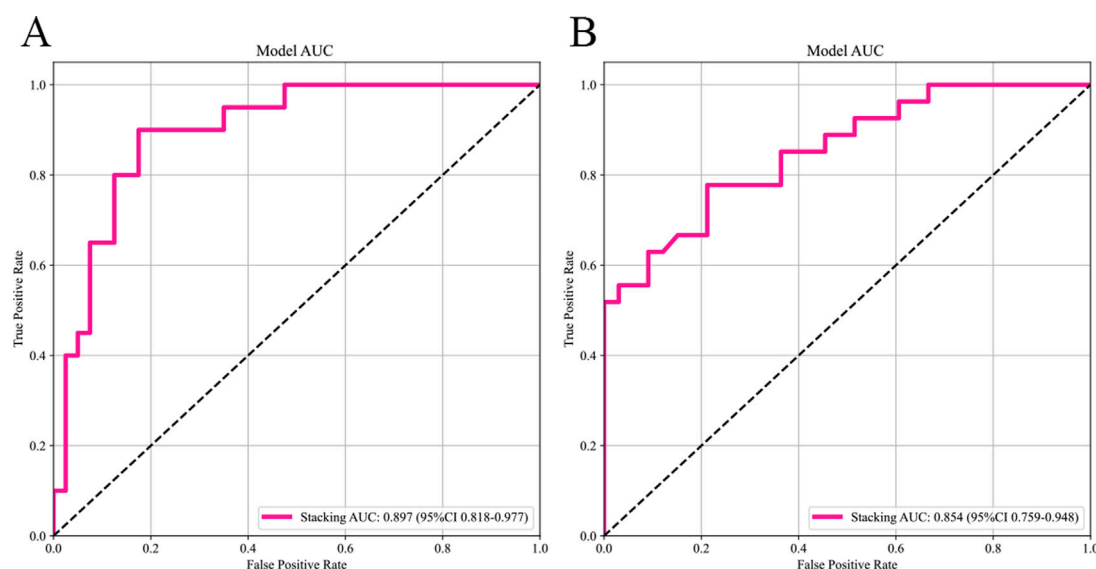


FIGURE 6
The AUC of the stacking ensemble learning models. Subfigures (A) and (B) respectively show the AUC curves of the stacking ensemble learning models on the internal and external test sets.

utilizing DL based automated segmentation and integrating radiomics data from T2WI and CE-T1WI sequences. A diverse array of machine learning algorithms, including SVM, RF, LightGBM, and MLP, served as base learners, while LR was utilized as the meta-learner to construct the ensemble model. This approach notably enhanced the precision of predicting HIFU ablation efficacy for uterine fibroids. Compared to single-algorithm models, the ensemble model exhibited a marked improvement in performance, with AUC values rising to 0.897 (95% CI: 0.818–0.977) in the internal validation cohort and 0.854 (95% CI: 0.759–0.948) in the external validation cohort, thereby underscoring the model's superior capability in managing complex radiomics data.

In the field of radiomics, precise image segmentation is crucial as it enhances the accuracy of data analysis and provides a solid foundation for subsequent predictive modeling. Current studies predicting the effectiveness of HIFU therapy often rely on manual contouring (Li et al., 2023; Zheng et al., 2021). This process is not only time-consuming but also prone to human error. Therefore, incorporating automated segmentation technologies is crucial to enhancing efficiency and minimizing human-related biases. In this study, an automated segmentation model based on V-Net was developed specifically for rapid segmentation of uterine fibroids on T2WI and CE-T1WI. This automated tool substantially reduces the need for human intervention, freeing radiomics analysis from the labor-intensive manual contouring and minimizing biases introduced by human factors. This innovation advances the prognostication of HIFU outcomes towards a more efficient and automated future.

Our study has made notable advancements in the field of multimodal MRI analysis by integrating T2WI and CE-T1WI. This integrated approach shows improved predictive performance with

an AUC value of 0.858, compared to AUC values of 0.822 for T2WI and 0.848 for CE-T1WI when used independently (Li et al., 2023; Zheng et al., 2021). The multimodal integration strategy effectively leverages the complementary information from different MRI sequences, offering a more comprehensive description of fibroid tissue characteristics and thus enhancing the accuracy of predicting HIFU treatment efficacy. T2WI, with its superior contrast resolution, reveals cellular dense areas and fibrotic regions within fibroids. These regions manifest as high signal intensity on T2WI and may affect ultrasound transmission efficiency, while fibrotic tissue might interfere with ultrasound propagation characteristics, impacting the effectiveness of HIFU treatment (Zhao et al., 2013; Gong et al., 2017). Additionally, T2WI can identify degenerative features such as necrosis, ischemia, edema, and calcification within fibroids, all of which significantly influence treatment outcomes (Andrews et al., 2019). Moreover, CE-T1WI assesses the vascular density and blood flow within fibroids, offering insights into ultrasound energy distribution and the thermal effects of HIFU (Liu et al., 2018). Regions with abundant vasculature may lead to the dissipation of ultrasound energy, thereby reducing treatment efficacy (Lénard et al., 2008). Nevertheless, traditional MRI image analysis faces challenges related to subjective interpretation and difficulties in detecting subtle signal intensity variations. This study addresses these limitations by introducing radiomics methods, which overcome these issues by extracting a large number of quantitative features from MRI scans, thereby offering an objective and precise description of fibroid structural heterogeneity (Gillies et al., 2016; Lambin et al., 2017). These features not only enhance the analysis of fibroid morphology but also improve predictive capabilities for HIFU treatment outcomes. Furthermore, the combined use of multimodal MRI radiomic features demonstrates the complementarity among different

MRI modalities, which is crucial for revealing the biological characteristics of fibroids.

In the medical field, where precise treatment is essential, the efficient and comprehensive utilization of multimodal MRI data is critical for enhancing predictive model performance. However, conventional single machine learning algorithms often fail to fully harness the potential of these invaluable multimodal radiomics datasets due to their disparate methodologies in feature handling and focal points. To address this challenge, this study employs four machine learning algorithms, each with a distinctive modeling philosophy: SVM, RF, MLP, and LightGBM, as base learners. SVM excels in managing small sample sizes and nonlinear challenges by identifying optimal hyperplanes within complex feature spaces (Cortes and Vapnik, 1995); RF enhances model stability and mitigates overfitting through ensemble decision trees and a voting mechanism (Breiman, 2001); LightGBM demonstrates exceptional performance in large-scale data processing by iteratively refining weak classifiers (Ke et al., 2017); and MLP leverages deep neural networks to capture intricate nonlinear features, thereby augmenting feature learning capacity (Rumelhart et al., 1986). The study integrated these base models using a meta-learner, LR, to construct a stacked ensemble learning framework. This sophisticated approach not only amalgamates the strengths of diverse models but also rectifies the limitations inherent in individual algorithms with regard to multimodal MRI radiomics features utilization. Consequently, it more effectively utilizes multimodal MRI radiomic features to improve both the accuracy and stability of preoperative predictions for HIFU treatment. This robust and advanced tool improves the precision of preoperative assessments, supports personalized treatment strategies, and is expected to enhance treatment decision-making and patient management in clinical practice.

4.1 Limitations

There are some limitations in this study. First, the sample size is relatively small. Future studies should include a larger cohort of patients to enhance the validity of the findings. Second, the effect of HIFU for uterine fibroids is affected by multiple factors, and future studies should integrate more clinical indicators to further improve the performance and credibility of the model.

5 Conclusion

This study developed a radiomics stacking ensemble model based on multimodal MRI, incorporating automatic segmentation techniques to predict the efficacy of HIFU ablation for uterine fibroids. It offers a comprehensive method for quantifying uterine fibroid heterogeneity and serves as a more precise supplementary tool for clinical practice.

References

Andrews, S., Yuan, Q., Bailey, A., Xi, Y., Chopra, R., Staruch, R., et al. (2019). Multiparametric MRI characterization of Funaki types of uterine fibroids considered for

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

BW: Conceptualization, Writing–original draft. CL: Conceptualization, Investigation, Software, Writing–review and editing. QC: Conceptualization, Data curation, Writing–review and editing. DS: Software, Validation, Writing–review and editing. XB: Data curation, Software, Writing–review and editing. FZ: Conceptualization, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Yiyang Central Hospital Youth Science and Technology Talent Program (2022QNYC10).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fphys.2024.1507986/full#supplementary-material>

MR-guided high-intensity focused ultrasound (MR-HIFU) therapy. *Acad. Radiol.* 26 (4), e9–e17. doi:10.1016/j.acra.2018.05.012

- Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324
- Bulun, S. E. (2013). Uterine fibroids. *N. Engl. J. Med.* 369 (14), 1344–1355. doi:10.1056/NEJMr1209993
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi:10.1613/jair.953
- Chen, J., Li, Y., Wang, Z., McCulloch, P., Hu, L., Chen, W., et al. (2018). Evaluation of high-intensity focused ultrasound ablation for uterine fibroids: an IDEAL prospective exploration study. *BJOG* 125 (3), 354–364. doi:10.1111/1471-0528.14689
- Cheng, Y., Yang, L., Wang, Y., Kuang, L., Pan, X., Chen, L., et al. (2024). Development and validation of a radiomics model based on T2-weighted imaging for predicting the efficacy of high intensity focused ultrasound ablation in uterine fibroids. *Quant. Imaging Med. Surg.* 14 (2), 1803–1819. doi:10.21037/qims-23-916
- Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi:10.1007/bf00994018
- Donnez, J., Arriagada, P., Donnez, O., and Dolmans, M. M. (2015). Current management of myomas: the place of medical therapy with the advent of selective progesterone receptor modulators. *Curr. Opin. Obstet. Gynecol.* 27 (6), 422–431. doi:10.1097/GCO.0000000000000229
- Dou, Y., Zhang, L., Liu, Y., He, M., Wang, Y., and Wang, Z. (2024). Long-term outcome and risk factors of reintervention after high intensity focused ultrasound ablation for uterine fibroids: a systematic review and meta-analysis. *Int. J. Hyperth.* 41 (1), 2299479. doi:10.1080/02656736.2023.2299479
- Funaki, K., Fukunishi, H., Funaki, T., Sawada, K., Kaji, Y., and Maruo, T. (2007). Magnetic resonance-guided focused ultrasound surgery for uterine fibroids: relationship between the therapeutic effects and signal intensity of preexisting T2-weighted magnetic resonance images. *Am. J. Obstet. Gynecol.* 196 (2), 184.e1–e6. doi:10.1016/j.ajog.2006.08.030
- Gillies, R. J., Kinahan, P. E., and Hricak, H. (2016). Radiomics: images are more than pictures, they are data. *Radiology* 278 (2), 563–577. doi:10.1148/radiol.2015151169
- Gong, C., Setzen, R., Liu, Z., Liu, Y., Xie, B., Aili, A., et al. (2017). High intensity focused ultrasound treatment of adenomyosis: the relationship between the features of magnetic resonance imaging on T2 weighted images and the therapeutic efficacy. *Eur. J. Radiol.* 89, 117–122. doi:10.1016/j.ejrad.2017.02.001
- Gong, X., Zhang, X., Liu, D., Yang, C., Zhang, R., Xiao, Z., et al. (2022). Evaluation of physician experience in achieving non-perfused volume ratio of high-intensity focused ultrasound ablation for uterine fibroids: a multicentre study. *J. Int. Med. Res.* 50 (5), 3000605221102087. doi:10.1177/03000605221102087
- Guiot, J., Vaidyanathan, A., Deprez, L., Zerka, F., Danthine, D., Frix, A. N., et al. (2022). A review in radiomics: making personalized medicine a reality via routine imaging. *Med. Res. Rev.* 42 (1), 426–440. doi:10.1002/med.21846
- Huang, H., Ran, J., Xiao, Z., Ou, L., Li, X., Xu, J., et al. (2019). Reasons for different therapeutic effects of high-intensity focused ultrasound ablation on excised uterine fibroids with different signal intensities on T2-weighted MRI: a study of histopathological characteristics. *Int. J. Hyperth.* 36 (1), 477–484. doi:10.1080/02656736.2019.1592242
- Iqbal, I., Shahzad, G., Rafiq, N., Mustafa, G., and Ma, J. (2020). Deep learning-based automated detection of human knee joint's synovial fluid from magnetic resonance images with transfer learning. *IET Image Process.* 14 (10), 1990–1998. doi:10.1049/iet-ipc.2019.1646
- Jacoby, V. L., Parvataneni, R., Oberman, E., Saberi, N. S., Varon, S., Schembri, M., et al. (2020). Laparoscopic radiofrequency ablation of uterine leiomyomas: clinical outcomes during early adoption into surgical practice. *J. Minim. Invasive Gynecol.* 27 (4), 915–925. doi:10.1016/j.jmig.2019.07.025
- Ke, G., Meng, Q., Finley, T., Wang, T., et al. (2017). Lightgbm: a highly efficient gradient boosting decision tree. *Adv. neural Inf. Process. Syst.* 30.
- Keserci, B., and Duc, N. M. (2017). The role of T1 perfusion-based classification in magnetic resonance-guided high-intensity focused ultrasound ablation of uterine fibroids. *Eur. Radiol.* 27 (12), 5299–5308. doi:10.1007/s00330-017-4885-x
- Lambin, P., Leijenaar, R. T. H., Deist, T. M., Peerlings, J., de Jong, E. E. C., van Timmeren, J., et al. (2017). Radiomics: the bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* 14 (12), 749–762. doi:10.1038/nrclinonc.2017.141
- Lénárd, Z. M., McDannold, N. J., Fennessy, F. M., Stewart, E. A., Jolesz, F. A., Hynynen, K., et al. (2008). Uterine leiomyomas: MR imaging-guided focused ultrasound surgery—imaging predictors of success. *Radiology* 249 (1), 187–194. doi:10.1148/radiol.2491071600
- Lethaby, A., Puscasiu, L., and Vollenhoven, B. (2017). Preoperative medical therapy before surgery for uterine fibroids. *Cochrane Database Syst. Rev.* 11 (11), CD000547. doi:10.1002/14651858.CD000547.pub2
- Li, C., He, Z., Lv, F., Liu, Y., Hu, Y., Zhang, J., et al. (2023). An interpretable MRI-based radiomics model predicting the prognosis of high-intensity focused ultrasound ablation of uterine fibroids. *Insights Imaging* 14 (1), 129. doi:10.1186/s13244-023-01445-2
- Liao, L., Xu, Y. H., Bai, J., Zhan, P., Zhou, J., Li, M. X., et al. (2023). MRI parameters for predicting the effect of ultrasound-guided high-intensity focused ultrasound in the ablation of uterine fibroids. *Clin. Radiol.* 78 (1), 61–69. doi:10.1016/j.crad.2022.09.112
- Liu, X., Tang, J., Luo, Y., Wang, Y., Song, L., and Wang, W. (2020). Comparison of high-intensity focused ultrasound ablation and secondary myomectomy for recurrent symptomatic uterine fibroids following myomectomy: a retrospective study. *BJOG* 127 (11), 1422–1428. doi:10.1111/1471-0528.16262
- Liu, Z., Gong, C., Liu, Y., and Zhang, L. (2018). Establishment of a scoring system for predicting the difficulty level of high-intensity focused ultrasound ablation of uterine fibroids. *Int. J. Hyperth.* 34 (1), 77–86. doi:10.1080/02656736.2017.1325015
- Lqbal, I., Younus, M., Walayat, K., Kakar, M. U., and Ma, J. (2021). Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images. *Comput. Med. Imaging Graph* 88, 101843. doi:10.1016/j.compmedimag.2020.101843
- Lyon, P. C., Rai, V., Price, N., Shah, A., Wu, F., and Cranston, D. (2020). Ultrasound-Guided high intensity focused ultrasound ablation for symptomatic uterine fibroids: preliminary clinical experience. *Ultraschall Med.* 41 (5), 550–556. doi:10.1055/a-0891-0729
- McLucas, B., Adler, L., and Perrella, R. (2001). Uterine fibroid embolization: nonsurgical treatment for symptomatic fibroids. *J. Am. Coll. Surg.* 192 (1), 95–105. doi:10.1016/s1072-7515(00)00738-9
- Milletari, F., Navab, N., Ahma, d, et al. (2016). “V-net: fully convolutional neural networks for volumetric medical image segmentation,” in *2016 fourth international conference on 3D vision (3DV)* (Ieee), 565–571.
- Naimi, A. I., and Balzer, L. B. (2018). Stacked generalization: an introduction to super learning. *Eur. J. Epidemiol.* 33, 459–464. doi:10.1007/s10654-018-0390-z
- Orlhac, F., Eertink, J. J., Cottureau, A.-S., Zijlstra, J. M., Thieblemont, C., Meignan, M., et al. (2022). A guide to ComBat harmonization of imaging biomarkers in multicenter studies. *J. Nucl. Med.* 63 (2), 172–179. doi:10.2967/jnumed.121.262464
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature* 323 (6088), 533–536. doi:10.1038/323533a0
- Sesmero, M. P., Ledezma, A. I., and Sanchis, A. (2015). Generating ensembles of heterogeneous classifiers using stacked generalization. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 5 (1), 21–34. doi:10.1002/widm.1143
- Snoek, J., Larochelle, H., and Adams, R. P. (2012). “Practical Bayesian optimization of machine learning algorithms,” in *Paper presented at the 25th international conference on neural information processing systems*.
- Stewart, E. A., Cookson, C. L., Gandolfo, R. A., and Schulze-Rath, R. (2017). Epidemiology of uterine fibroids: a systematic review. *BJOG* 124 (10), 1501–1512. doi:10.1111/1471-0528.14640
- Venkatesan, A. M., Partanen, A., Pulanic, T. K., Dreher, M. R., Fischer, J., Zurawin, R. K., et al. (2012). Magnetic resonance imaging-guided volumetric ablation of symptomatic leiomyomata: correlation of imaging with histology. *J. Vasc. Interv. Radiol.* 23 (6), 786–794. doi:10.1016/j.jvir.2012.02.015
- Verpalen, I. M., Anneveldt, K. J., Nijholt, I. M., Schutte, J. M., Dijkstra, J. R., Franx, A., et al. (2019). Magnetic resonance-high intensity focused ultrasound (MR-HIFU) therapy of symptomatic uterine fibroids with unrestricted treatment protocols: a systematic review and meta-analysis. *Eur. J. Radiology* 2 120, 108700. doi:10.1016/j.ejrad.2019.108700
- Verpalen, I. M., de Boer, J. P., Linstra, M., Pol, R. L. I., Nijholt, I. M., Moonen, C. T. W., et al. (2020). The Focused Ultrasound Myoma Outcome Study (FUMOS); a retrospective cohort study on long-term outcomes of MR-HIFU therapy. *Eur. Radiol.* 30, 2473–2482. doi:10.1007/s00330-019-06641-7
- Wang, H., Jin, Q., Li, S., Liu, S., Wang, M., and Song, Z. (2014). A comprehensive survey on deep active learning in medical image analysis. *Med. Image Anal.* 95, 103201. doi:10.1016/j.media.2024.103201
- Yoon, S.-W., Lee, C., Kim, K. A., and Kim, S. H. (2010). Contrast-enhanced dynamic MR imaging of uterine fibroids as a potential predictor of patient eligibility for MR guided focused ultrasound (MRgFUS) treatment for symptomatic uterine fibroids. *Obstet. Gynecol. Int.* 2010, 834275. doi:10.1155/2010/834275
- Zhang, D. L., Wu, S. S., Chen, S., Liu, X. X., Tang, J. Q., Lin, N., et al. (2020). Differences in the therapeutic effects of high-intensity focused ultrasound (HIFU) ablation on uterine fibroids with different shear wave velocity (SWV): a study of histopathological characteristics. *Int. J. Hyperth.* 37 (1), 1322–1329. doi:10.1080/02656736.2020.1849827
- Zhao, W. P., Chen, J. Y., and Chen, W. Z. (2015). Effect of biological characteristics of different types of uterine fibroids, as assessed with T2-weighted magnetic resonance imaging, on ultrasound-guided high-intensity focused ultrasound ablation. *Ultrasound Med. Biol.* 41 (2), 423–431. doi:10.1016/j.ultrasmedbio.2014.09.022
- Zhao, W. P., Chen, J. Y., Zhang, L., Li, Q., Qin, J., Peng, S., et al. (2013). Feasibility of ultrasound-guided high intensity focused ultrasound ablating uterine fibroids with hyperintense on T2-weighted MR imaging. *Eur. J. Radiol.* 82 (1), e43–e49. doi:10.1016/j.ejrad.2012.08.020
- Zheng, Y., Chen, L., Liu, M., Wu, J., Yu, R., and Lv, F. (2021). Prediction of clinical outcome for high-intensity focused ultrasound ablation of uterine leiomyomas using multiparametric MRI radiomics-based machine learning model. *Front. Oncol.* 11, 618604. doi:10.3389/fonc.2021.618604



OPEN ACCESS

EDITED BY
Sokratis Makrogiannis,
Delaware State University, United States

REVIEWED BY
Nagaraju Yalavarthi,
Central Silk Board, India
Preeta Sharan,
The Oxford College of Engineering, India

*CORRESPONDENCE
Manh-Hung Ha
✉ hunghm@vnu.edu.vn

RECEIVED 14 August 2024
ACCEPTED 31 December 2024
PUBLISHED 22 January 2025

CITATION
Do M-T, Ha M-H, Nguyen D-C and
Tzyh-Chiang Chen O (2025) Toward
improving precision and complexity of
transformer-based cost-sensitive learning
models for plant disease detection.
Front. Comput. Sci. 6:1480481.
doi: 10.3389/fcomp.2024.1480481

COPYRIGHT
© 2025 Do, Ha, Nguyen and Tzyh-Chiang
Chen. This is an open-access article
distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Toward improving precision and complexity of transformer-based cost-sensitive learning models for plant disease detection

Manh-Tuan Do¹, Manh-Hung Ha^{1*}, Duc-Chinh Nguyen¹ and
Oscal Tzyh-Chiang Chen^{1,2}

¹Faculty of Applied Sciences, International School, Vietnam National University, Hanoi, Vietnam,
²Department of Electrical Engineering, National Chung Cheng University, Chiayi, Taiwan

Early and accurate detection of plant diseases is crucial for making informed decisions to increase the yield and quality of crops through the decision of appropriate treatments. This study introduces an automated system for early disease detection in plants that enhanced a lightweight model based on the robust machine learning algorithm. In particular, we introduced a transformer module, a fusion of the SPP and C3TR modules, to synthesize features in various sizes and handle uneven input image sizes. The proposed model combined with transformer-based long-term dependency modeling and convolution-based visual feature extraction to improve object detection performance. To optimize a model to a lightweight version, we integrated the proposed transformer model with the Ghost module. Such an integration acted as regular convolutional layers that subsequently substituted for the original layers to cut computational costs. Furthermore, we adopted the SloU loss function, a modified version of CloU, applied to the YOLOv8s model, demonstrating a substantial improvement in accuracy. We implemented quantization to the YOLOv8 model using ONNX Runtime to enhance to facilitate real-time disease detection on strawberries. Through an experiment with our dataset, the proposed model demonstrated mAP@.5 characteristics of 80.30%, marking an 8% improvement compared to the original YOLOv8 model. In addition, the parameters and complexity were reduced to approximately one-third of the initial model. These findings demonstrate notable improvements in accuracy and complexity reduction, making it suitable for detecting strawberry diseases in diverse conditions.

KEYWORDS

DNN, transformer, Ghost Conv, SloU loss function, pre-trained, quantization, android application

1 Introduction

Detecting diseases in crops, especially in major crops, is a crucial issue in agriculture. Early disease detection and prevention are vital measures to minimize damage to crops and increase productivity. However, detecting diseases on strawberries poses a challenge due to variations in shape, color, and size among different diseases. Traditional methods for strawberry disease detection, reliant on the analyses of agricultural experts, are time-consuming and lack accuracy. With the significant progress in deep learning and computer hardware, advanced image recognition technologies are increasingly employed by scholars for agricultural disease recognition.

Recent applications of deep learning, particularly Convolutional Neural Networks (CNNs), in detecting crop diseases have shown promising results (Mahmud et al., 2019; Jayawardena et al., 2016). Scholars have proposed CNN techniques, such as GoogLeNet (Ferentinos, 2018), for plant leaf identification, achieving recognition rates exceeding 94%, even with partial leaf damage. CNNs have been utilized for discovering crop species and diseases (Ha and Chen, 2021), with reported accuracies of 99.35%. Deep Transfer Learning (DTL) (Chen et al., 2020) has been employed for banana disease detection, reaching an accuracy of 90%. Diverse deep CNN architectures (Cheng et al., 2017), including AlexNet, MobileNet, GoogLeNet, VGGNet, and Xception, have been proposed for inspecting strawberry quality, with reported accuracies of up to 95%. Supervised machine learning technologies (Selvaraj et al., 2019) have also been addressed to recognize strawberry powdery mildew disease with an accuracy of 94.34%. The classification model (Ha et al., 2024) for identifying plant diseases through the integration of local and global features utilizing a transformer-based approach demonstrated very high results with 99.18% and 94.05% accuracy. Object detection techniques, such as Fast RCNN (Mohanty et al., 2016) and Faster-RCNN (Chen et al., 2019; Girshick, 2015; Baweja et al., 2018) have found widespread applications in detecting insect diseases in plants. Additionally, mask R-CNN (Sa et al., 2016) has a demonstrated significant accuracy, especially in fruit discovery for strawberry harvesting.

Beyond strawberries, recent research has significantly expanded the application of object detection models to target plant diseases in various crops. Liu and Wang (2021), in their review "Plant Diseases and Pests Detection Based on Deep Learning," emphasize the advancements in utilizing deep learning techniques to detect diseases such as powdery mildew (*Erysiphe necator*) and black spot (*Alternaria alternata*) in tomatoes and cucumbers, as well as apple scab (*Venturia inaequalis*) and downy mildew (*Plasmopara viticola*). Similarly, Shruthi and Nagaveni (2024) demonstrate a hybrid convolutional neural network (CNN) model using self-regulated layers and inception layer for accurate and efficient diagnosis of tomato diseases with severity levels. These studies highlight the promising capabilities of AI systems to revolutionize plant disease detection and management across diverse agricultural application.

In terms of attention mechanisms and transformer-based architectures, have significantly improved the performance of plant disease detection systems. Transformers, initially popularized in natural language processing, have demonstrated exceptional capability in modeling long-range dependencies and multiscale features, making them highly effective for visual tasks. For instance, Gu et al. (2024) proposed the Multi-Modal Fast Gated Transformer (MFGTN), which integrates spatial and temporal data for improved feature fusion, inspiring potential applications in plant disease detection under diverse environmental conditions. Similarly, Song et al. (2024) introduced CenterFormer, a transformer-based model that enhances segmentation accuracy through a cluster center-guided attention mechanism, offering a promising approach for localizing disease-affected regions in plants. These studies highlight the importance of attention mechanisms in achieving precise feature extraction and localization.

In the context of one-stage object detection, the YOLO (You Only Look Once) family, including YOLOv3, YOLOv4, YOLOv5 (Yu et al., 2019; Zhang et al., 2022; Sozzi et al., 2022), and YOLOv7 (Gallo et al., 2023), has shown promising results in detecting diseases in plants, including strawberries. These models efficiently combine feature extraction and prediction, with fair inference time. The evolution of disease detection methodologies, particularly through the integration of advanced deep learning techniques, provides a strong foundation for AI-mediated disease detection in agriculture.

In line with the prevailing trend and its application to the challenge of strawberry leaf disease detection, this work introduces models designed for accuracy enhancement model complexity minimization based on YOLOv8s released in 2023, a state-of-the-art model in single-stage object detection as well as within the YOLO family. Specifically, our contributions to this paper are as follows:

- A new dataset is proposed and collected from high-quality images on Google with farm settings, as well as from the Ministry of Agriculture, Vietnam. Through rigorous preprocessing and adherence to strict criteria regarding color, area, density of the diseased part, and species shape, we curated 1,000 high-quality images, categorizing them into five classes: Normal, Rubber, Gray Mold, Black Spot, and Powdery Mildew.
- We develop the SC3T module to ameliorate model accuracy. Inspired by the transformer module widely used in natural language processing, the SC3T module employs an attention mechanism designed for multiscale processing, effectively handling feature maps at different scales and ensuring accurate detection of objects of various sizes.
- The loss function of CIoU in YOLOv8s is replaced by the SIoU loss function, a variant of CIoU incorporating angle factors. This establishes the basis for inferring costs related to distance, ratio, and intrusion.
- The Ghost convolution module is specifically devised to address limitations in conventional deep neural network models like YOLOv8s, to successfully establish a lightweight model.
- Quantization through ONNX Runtime, leading to a streamlined ONNX file suitable for deployment, is conducted for aiming at enhancing model performance and efficiency.
- Subsequently, our object detection app was developed and demonstrated on Android devices, showcasing the tangible deployment and practical utility of the proposed deep learning model in real-world object detection scenarios.

The subsequent sections of this paper are organized as follows: Section 2 presents an extensive overview of existing methodologies in the literature employed for plant disease detection. Section 3 details the proposed methodology, offering more specific explanations on enhancement features. In Section 4, we illustrate the dataset, training environment, and both quantitative and qualitative results in three aspects: accuracy complexity, and loss functions. Additionally, comparative analyses of our models are addressed on different datasets. Finally, the conclusion is given in Section 5 with a brief summary of this work.

2 Related previous work

2.1 One-stage object detection

Among the prominent one-stage object detection frameworks (Yao et al., 2021), the YOLO series (Redmon et al., 2016) stands out for its real-time performance and unified, efficient architecture, consistently achieving high accuracy and versatility across various applications through iterative advancements. YOLOv1 and YOLOv2, while groundbreaking, relied on a rigid grid-based prediction mechanism that struggled with localization accuracy for small or occluded objects. To address these limitations, YOLOv3 introduced multi-scale feature detection, significantly enhancing object recognition across varying sizes. However, its deeper architecture increased training time and computational costs. Building on these improvements, YOLOv4 incorporated the CSPDarknet53 backbone, further boosting accuracy but at the expense of higher GPU resource requirements. YOLOv5 shifted focus toward lightweight design, achieving faster inference times but lacking advanced feature fusion capabilities, which limited its performance in cluttered environments. YOLOv7 enhanced training efficiency and detection speed, yet it remained dependent on intricate hyperparameter tuning and showed reduced robustness for objects in motion.

YOLOv8 addresses these challenges with significant architectural advancements by integrating the C2f module to enhance multiscale feature aggregation and reduce computational overhead compared to the C3 module used in YOLOv5 and YOLOv7. In addition, YOLOv8 supports ONNX Runtime and TensorFlow Lite, enabling seamless deployment across diverse platforms and enhancing performance in real-time applications. Its lightweight architecture and optimized inference reduce latency, making it particularly well-suited for time-critical tasks. Based on these strengths, we have utilized and implemented the experiments in YOLOv8s.

2.2 Object detection models with accuracy improvement

From foundational models, numerous studies have sought to improve the accuracy performance, notably in the context of YOLO-related research. For instance, in YOLOv3 (Zhao and Li, 2020), this model was introduced to accelerate the rate of convergence when initializing the width and height of the predicted bounding boxes. This method enables the selection of more representative initial dimensions, leading to a significant increase in mAP. Another study (Yao et al., 2020) employed double K-means to generate anchor boxes, aiming to lift localization accuracy. Several investigations have focused on refining the structures within the YOLO's backbone, such as an introduction of the bottleneck CSP-2 module in Yan et al. (2021) or the incorporation of special modules, as seen in Yao et al. (2021), which introduced SELayer (Xu et al., 2021) and integrated EfficientNet into the YOLO architecture.

One promising module inspired by a transformer, a natural language processing model developed by Google, is gaining attention. With its attention mechanism via matrix computations,

a transformer can effectively link semantically relevant content. This mechanism performs well with image data, facilitating the correlation of related features. Leveraging this advantage, the adoption of the transformer has demonstrated significant effectiveness in models like YOLOv5s (Zhu et al., 2021; Yu et al., 2021). Referring to these enhancement ideas, we propose the SC3T transformer module to improve mAP accuracy in the state-of-the-art YOLO model, YOLOv8.

2.3 Object detection models with light weights

Enhancing the accuracy of YOLO often increases model complexity, leading to higher Floating-Point Operations per Second (FLOPs). Consequently, the challenge of improving the model in a lightweight direction for high-efficiency hardware computing has become a highly prospective research area. For example, tinier-YOLO (Fang et al., 2019) proposed a lightweight solution for tiny YOLOv3. This involved modifying the SqueezeNet module to reduce the number of model parameters and subsequently decrease the overall model size. Another noteworthy approach is found in Lu et al. (2020), named YOLO-compact, which separates the down-sampling layers from all network modules. This model is 3.7, 6.7, and 26 times smaller than tiny-YOLOv3, tiny-YOLOv2, and YOLOv3, respectively. The Ghost module, applied in YOLOv5 (Dong et al., 2022; Liu et al., 2021; Xu et al., 2022), has demonstrated notable efficiency in reducing computational operations and the number of parameters significantly. Such a module can be a good candidate for lightweight model designs. Recognizing its effectiveness, we explored incorporating the Ghost convolution module into the backbone architecture of YOLOv8s, yielding highly favorable results in experimental evaluations.

2.4 Loss functions

The loss functions serve as a crucial aspect for evaluating the overall model accuracy by calculating the deviation between the actual and predicted objects. This dimension associated with maximizing accuracies in object detection tasks is pivotal. In YOLOv5, the loss function comprises three components:

Classification loss: used to compute the deviation between the predicted probabilities of object classes and the actual ones of object classes in an image.

Localization loss: employed to calculate the accuracy of the predicted bounding box positions concerning the actual ones of objects in an image.

Objectness loss (equivalent to IoU loss): used to determine the accuracy of classifying image pixels as objects or non-objects.

This study focuses on object loss in order to heighten the model's accuracy. Specifically, in YOLOv5, the objectness loss utilized is CIoU loss (Zheng et al., 2021), which considers three factors: the distance between the centers of the predicted and actual boxes, the aspect ratio difference, and the diagonal distance ratio. These special factors are beneficial to improve detection accuracy,

especially for small objects. The SIOU loss function, proposed by Gevorgyan (2022) in May 2022, is a variant of CIOU loss, incorporating an angle factor. This addition constructs the basis for inferring distance and ratio factors. Studies associated with replacing CIOU with SIOU have demonstrated promising results in YOLOv5. Inspired by this, we experimented with SIOU in YOLOv8, yielding favorable outcomes.

2.5 Pre-trained model

The early achievements of deep learning in the field of computer vision owe much to transfer learning. The pre-training based on ImageNet played a pivotal role in achieving advancements over state-of-the-art results in various recognition tasks, including object detection (Lin et al., 2017; Liu et al., 2016), semantic segmentation (Chen et al., 2017; He et al., 2017), and scene classification (Zhou et al., 2017; Herranz et al., 2016). The adaptability of pre-trained features has been thoroughly investigated (Azizpour et al., 2015; Cui et al., 2018; Kornblith et al., 2019). For instance, Azizpour et al. (2015) quantified the similarity between tasks using ImageNet classification; Cui et al. (2018) investigated the transfer of insights gained from large classification datasets have been applied to smaller, detailed datasets; Kornblith et al. (2019) explored the relationship among ImageNet pre-training accuracies, transfer accuracies, and network architectures. In this work, the proposed model was trained on a large dataset, as detailed in Afzaal et al. (2021). Subsequently, we utilized these weights as the pre-trained parameters for further learning on our specific dataset.

3 Proposed model

This work proposes and optimizes models for practical use, specifically object detection models in a typical phase like those in the YOLO family. After training, evaluation, and testing, these models undergo quantization and are integrated into the application software. An overview of our system architecture is depicted in Figure 1. Initially, to assess the effectiveness of the proposed models, we trained them on a large dataset, followed by evaluation and selection of the outperformance candidates based on two criteria: accuracy and lightweight nature. These top models are then saved as pre-training models for the Vietnam strawberries dataset. With the proposed accuracy model, we aim to achieve high accuracy without considering its complexity, while with the lightweight model, we emphasize reducing model complexity as well as accuracy. Subsequently, the lightweight model is quantized to optimize its weights, suitably embedded into edge devices. Specifically, we built an Android app for disease detection on strawberries. To delve into the specifics, we will explore the original YOLOv8 architecture.

From an architectural point of view, YOLOv8s doesn't exhibit many differences from YOLOv5s. A key distinction lies in the integration of the C2f module, which supersedes the C3 module originally employed in YOLOv5. In YOLOv5, the C3 module consists of three standard convolutional layers and several bottleneck blocks. This structure incorporates two branches: one

branch utilizes multiple stacked bottleneck blocks and three standard convolutional layers, while the other branch processes a single basic convolutional layer before merging with the first branch. The effective design of the bottleneck module minimizes the number of training parameters and computational load, thus mitigating issues of gradient explosion and vanishing in deep networks, and thereby enhancing the model's learning capabilities. YOLOv7 further refines gradient calculations by introducing multiple parallel gradient branches and implementing the ELAN module, resulting in improved accuracy and more reasonable latency. YOLOv8 designed the C2f module based on the C3 module and ELAN's concept to gather diverse gradient clues while maintaining a lightweight structure.

In Figure 2, YOLOv8s-Transformer is devised based on the basic architecture of YOLOv8s to enhance the accuracy. We have integrated SC3T into the final layer of the backbone to optimize the extracted data at different scales, thereby improving overall model efficiency. Subsequently, these features at various scales are extracted in the head section before moving to the prediction part. Here, we experimented with various loss functions to identify the most suitable one that achieves outstanding performance for our application.

Similar to YOLOv8s-Transformer, as shown in Figure 3, YOLOv8s-Trans-Ghost has been developed with the addition of the transformer module and a loss function similar to those in the YOLOv8s transformer. However, the positions of conventional convolutions are substituted with Ghost convolutions. This modification aims to balance enhanced accuracy with reduced computational load by mapping each input channel to a smaller number of output channels, rather than mapping each input channel to a corresponding output channel as done in typical convolutional layers.

In comparison to YOLOv8, we have implemented the following optimizations:

- Propose the S3CT module into the last layer of the backbone, aiding the model in more accurately extracting and localizing germination features, thereby improving the mAP accuracy of the model.
- Replace convolutional layers with Ghost convolution, enhancing not only the model's performance but also reducing its computational complexity, making it more lightweight and efficient.
- In YOLOv8s, the loss function utilizes CIOU, which has demonstrated significant effectiveness compared to GIoU and DIoU used in the earlier versions of YOLOv8. However, CIOU doesn't handle objects that change in scale whereas SIOU, as a variant of CIOU, addresses this issue by normalizing distance and diagonal distance via the width and height of the ground truth box. This helps the model ameliorate performance in detecting small objects as well as increase the mAP accuracy of the model.

3.1 Proposed SC3T

The proposed SC3T (Do et al., 2023) module which is integrated into the last layer of the backbone is a fusion of the

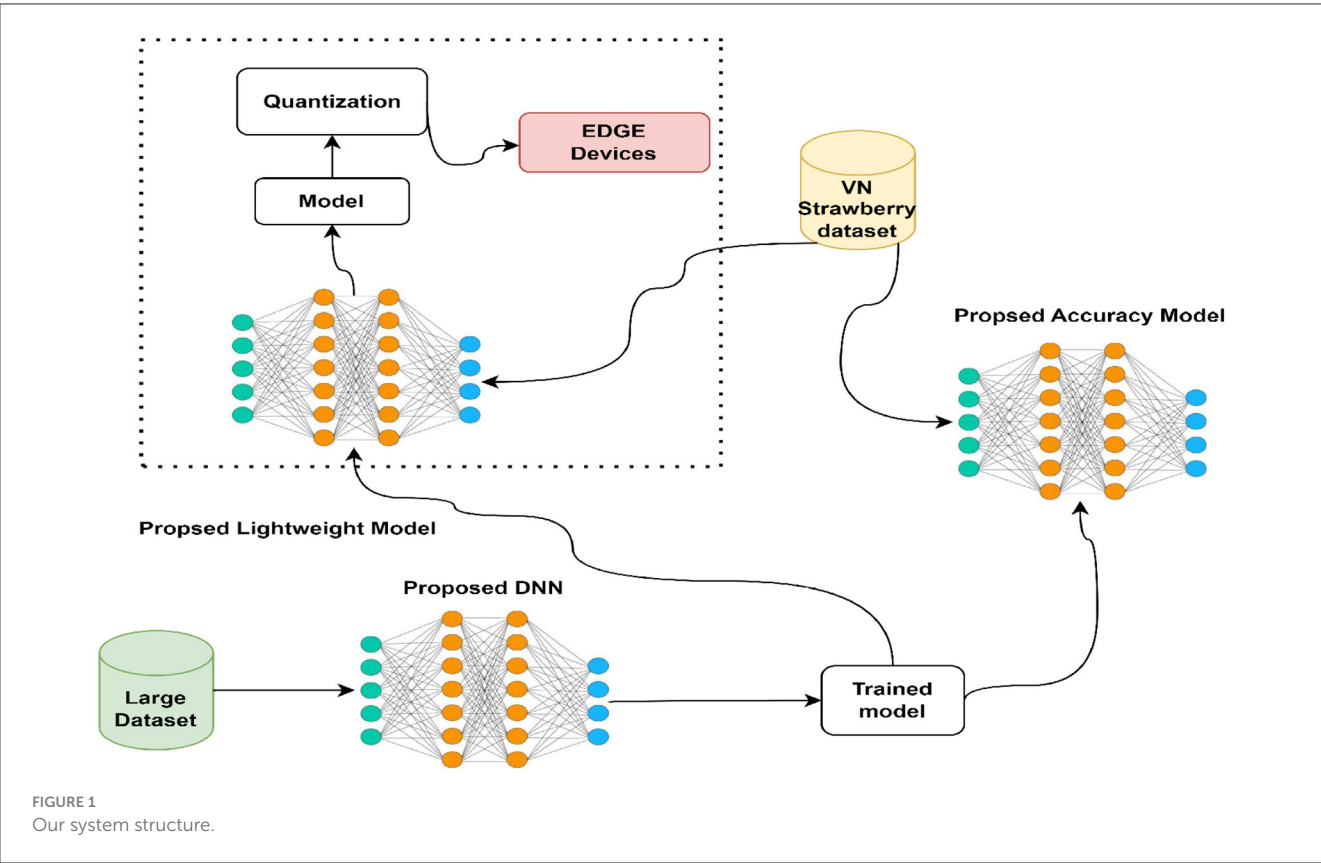


FIGURE 1
Our system structure.

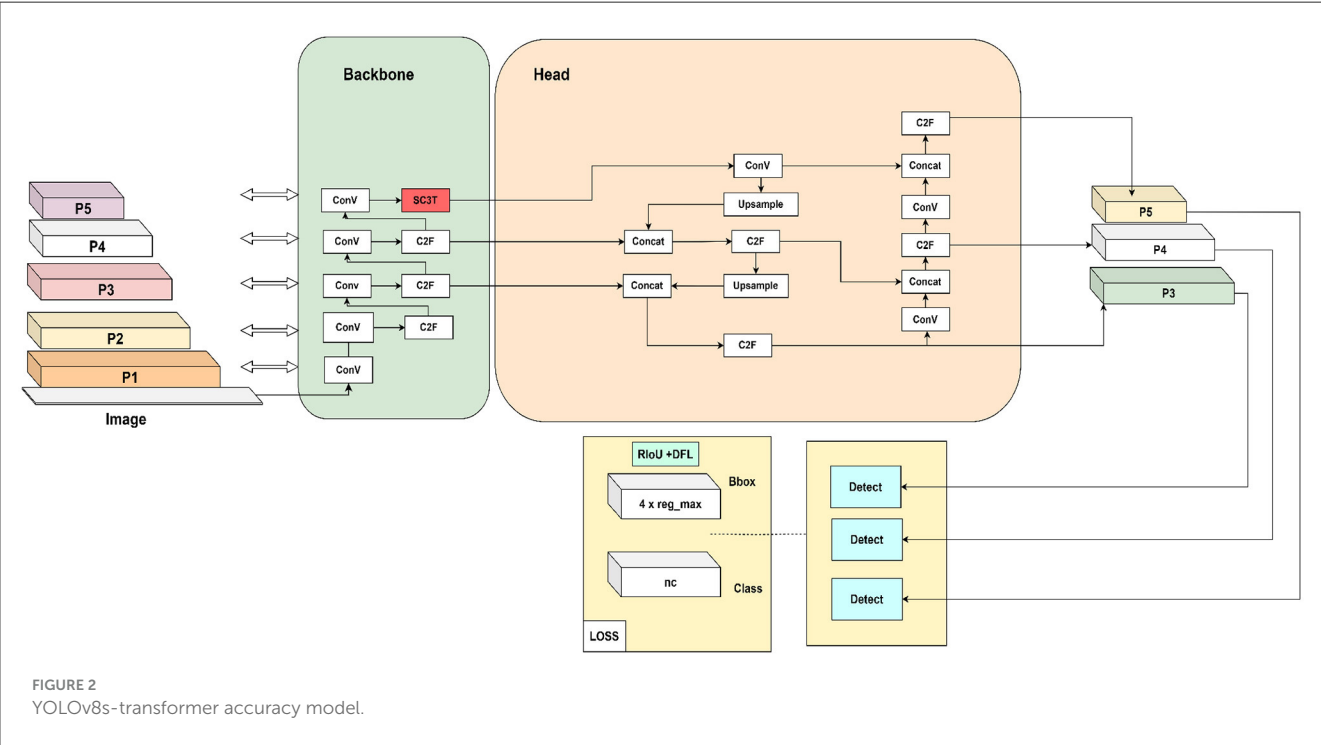
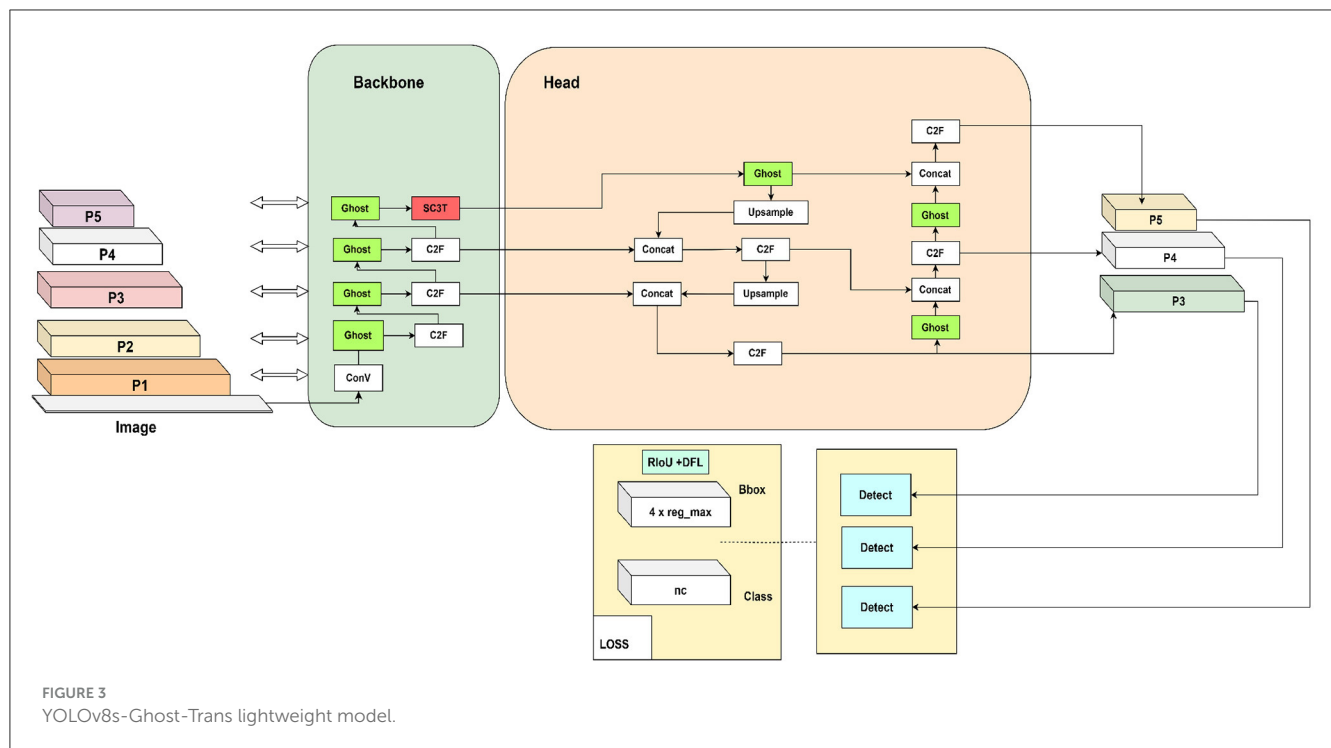


FIGURE 2
YOLOv8s-transformer accuracy model.

SPP and the C3TR modules which is shown in Figure 4. The SPP module addresses the challenge posed by non-uniform input image sizes encountered in object detection tasks. In traditional CNN, the input image size is typically fixed, requiring resizing or cropping

of input images to a predefined one before feeding inputs into the network. However, this approach is not ideal for handling images of different sizes efficiently. To overcome this limitation, the Spatial Pyramid Pooling (SPP) module allows the network to accept input



images of various sizes without the need for resizing. We design SPP using kernels with a uniform stride of 1 but varying sizes (5×5 , 9×9 , and 13×13). This strategy ensures that important features at different scales are captured effectively. After pooling, each sub-region is processed independently to extract distinctive features. The extracted features from all sub-regions are then concatenated along the channel dimension. This feature fusion process through channel concatenation enables the model to combine features of different spatial resolutions efficiently, facilitating robust object detection across images of varying sizes. By incorporating the SPP module into the network architecture, we elevate the model's ability to handle non-uniform input sizes while preserving spatial information effectively, contributing to improved performance in object detection tasks. Subsequently, the output of the SPP module is combined with that from the C3TR module to create the SC3T module in Figure 4, replacing the C3 and the final SPPF layer in the YOLOv5s network to enhance the speed and accuracy of object detection with a smaller network size. The C3TR layer is shown in Figure 4, a combination of a transformer with the C3 layer of CSPDarknet53, which is utilized in the YOLOv5 model. This layer is pivotal for extracting and integrating features from various regions of the image, utilizing the transformer's ability to capture interdependencies among different data segments. Whereas the C3 layer in CSPDarknet53 is dedicated to feature extraction, the transformer is adept at identifying feature relationships. The fusion of these capabilities within the C3TR layer empowers the model to extract features and understand their relationships, which is essential for object detection tasks.

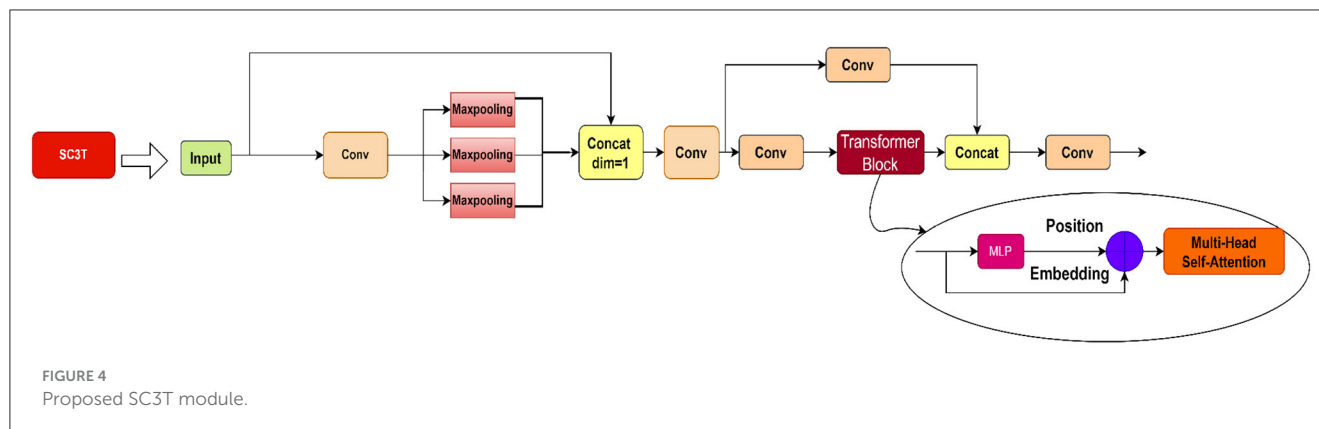
The C3TR layer capitalizes on the transformer's capability to learn hierarchical representations, ranging from low-level to high-level image features. This, in conjunction with CSPDarknet53's

feature extraction prowess, enhances feature learning with a more comprehensive contextual insight. The C3TR layer promotes stable learning by utilizing the transformer's layer normalization and residual connections, thereby alleviating issues of gradient vanishing and explosion during the training process, which ensures stable learning in deep neural networks.

Designed for multiscale processing, the C3TR layer receives and integrates feature maps at different scales, ensuring efficient detection of varied-size objects. The head of the C3TR layer reduces the size of the feature map, integrates it with other feature maps, and channels it through the transformer, enabling the learning of characteristics from different image scales. Serving as the input feature map for channel numbers 512, 1,024, and 2,048, the C3TR layer is a pivotal element in YOLOv5's object detection model.

Transformers have demonstrated exceptional performance in sequence modeling and have proven to be more effective than traditional architectures in non-sequential tasks such as object detection. In particular, the use of transformer-based learning in C3TR capitalizes on the model's ability to capture long-range dependencies. These dependencies involve correlations between elements that are separated by significant distances in a sequence or within spatio-temporal data. Put simply, long-range dependencies reflect the meaningful relationships between a given element and those that come before or after it. For example, this includes the contextual links between words in a sentence, the spatial relationships between objects, and the movements observed across successive frames in a video (Zhao et al., 2019).

C3TR captures these long-term dependencies in input feature maps through a self-attention mechanism, allowing the model to consider a broader context and comprehend global object information, contributing to accurate object detection. The adept



combination of lines (referring to the effective feature extraction pathways in the CNN backbone) and transformers (utilizing a self-attention mechanism) ensures high efficiency with a minimal number of parameters. In the context of our methodology, the term “lines” refers to the efficient feature extraction pathways established within the CNN backbone, such as the DarkNet architecture used in YOLO. These pathways are responsible for extracting hierarchical features from input images. On the other hand, “transformers” denote components of our model that leverage self-attention mechanisms to capture long-range dependencies within the input feature maps. By integrating these effective feature extraction pathways (lines) with the self-attention capabilities of transformers, our model achieves high efficiency. This adept combination grants in the model to process global contextual information while maintaining a compact parameter footprint, leading to improved performance in object detection tasks.

In conclusion, the amalgamation of SPP and C3TR demonstrates effectiveness in YOLOv8 by utilizing concatenation to synthesize features at various sizes. Given the non-uniform input image sizes, this combined with transformer-based long-term dependency modeling and convolution-based visual feature extraction, enhances precise object localization and increases classification performance. Our experiments have consistently shown results aligning with these claims, with an mAP index ~6.0% higher than that from the baseline model.

3.2 Ghost module

Ghost convolution which is depicted in Figure 5 emerges as a structure strategically designed to strike a balance between heightened accuracy and minimal computational overhead. It specifically tackles the limitations found in conventional deep neural network models, known for their excessive intricacy, making them challenging for deployment on resource-limited devices.

In this study, we leverage the potential of the Ghost convolution module to enhance the performance of the YOLOv8 framework, especially in scenarios constrained by resource limitations. The conventional convolution module is replaced with the Ghost convolution module, representing a streamlined alternative to traditional convolutional layers. This transition significantly

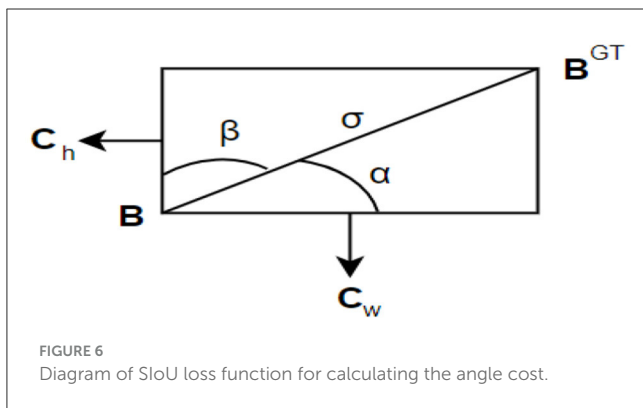
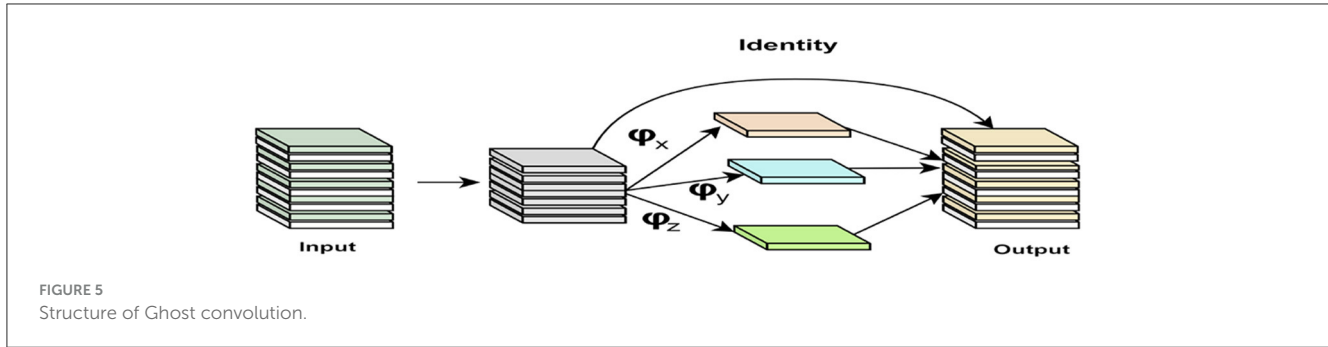
reduces the count of model parameters, facilitating the efficient allocation of computational resources.

3.3 SIoU loss function

After amalgamating models to create a more robust framework, we opted for the YOLOv8s-Transformer model for hardware deployment. However, to lift the system’s performance in accurately identifying smaller objects, especially for search and rescue operations which need extremely high accuracy, thus model optimization is essential. We observed that accuracy could still be improved by modifying or optimizing the loss function, specifically in this case, by utilizing the SIoU loss function that has demonstrated efficacy in the previous version of YOLOv5, and facilitated higher prediction accuracies for small objects due to incorporating a smoothing factor into the computation formula. This approach stabilizes the gradient of the SIoU loss function during model training. As a result, models using the SIoU loss function can better learn from small bounding boxes, thereby improving the accuracy of object localization predictions.

In the loss function officially used in YOLOv8, CIoU is adopted. It evaluates the distance between the centers of the predicted and ground-truth boxes, the aspect ratio difference, and the diagonal distance ratio. These factors aim to enhance the detection accuracy, especially for small objects. The loss function of SIoU (Gevorgyan, 2022) is a variant of CIoU that does not scale with the ratio. In SIoU, the “distance” refers to the distance between the centers of the predicted actual boxes, while the “diagonal distance” is the length of the diagonal line connecting the predicted actual boxes. Both the distance and diagonal distance are normalized by the width and height of the ground-truth box. This normalization allows SIoU to handle objects with different ratios, a common challenge in object detection. The SIoU loss function is defined by the sum of the following costs: angle cost, distance cost, shape cost, and IoU cost. As shown in Figure 6, the expression of angle cost is:

$$L_{\text{ang}} = 1 - 2 \sin^2 \left(\arcsin(x) - \frac{\pi}{4} \right) \quad (1)$$



$$x = \frac{c_h}{\sigma} = \sin(\alpha) \quad (2)$$

where σ represents the distance between the centroid of the ground truth bounding box and the prediction box as follows.

$$c_h = \max(b_{c_y}^g, b_{c_y}) - \min(b_{c_y}^g, b_{c_y}) \quad (3)$$

The distance cost is formulated in the following:

$$L_{\text{dis}} = \sum_{t=x,y} (1 - e^{-\gamma \rho_t}) \quad (4)$$

$$\rho_x = \left(\frac{(b_{c_x}^g t - b_{c_x})}{c_w} \right)^2 \quad (5)$$

$$\rho_y = \left(\frac{(b_{c_y}^g t - b_{c_y})}{c_h} \right)^2 \quad (6)$$

$$\gamma = 2 - L_{\text{ang}} \quad (7)$$

As α approaches 0, the impact of the distance cost diminishes significantly. Conversely, when α is closer to γ , the influence of L_{dis} increases. The difficulty of the problem grows with the angle. Therefore, as the angle increases, γ is given greater priority over the distance value. It is important to note that the distance cost decreases as α approaches 0.

The formula for shape cost is:

$$L_{\text{shape}} = \sum_{t=w,h} (1 - e^{-\omega_t})^\theta \quad (8)$$

$$\omega_w = \frac{|w - w^{gt}|}{\max(w, w^{gt})} \quad (9)$$

$$\omega_h = \frac{|h - h^{gt}|}{\max(h, h^{gt})} \quad (10)$$

Where w and w^{gt} refer to the widths of the prediction and ground-truth boxes, respectively, and h and h^{gt} denote the heights of the prediction and ground-truth boxes, respectively. The total loss function is represented by:

$$L_{\text{box}} = 1 - \text{IoU} + \frac{L_{\text{dis}} + L_{\text{shape}}}{2} \quad (11)$$

In this study, the SIoU loss function is utilized to replace CIoU in the proposed YOLOv8s-Transformer model. We also compare it with the other loss functions such as DIoU, GIoU, and EIoU for an evaluation. Additionally, experimental results indicate that SIoU when integrated with adoption in YOLOv8s, achieves superior outcomes compared to other losses.

4 Experiment result and discussion

4.1 Dataset

Strawberries hold immense agricultural value globally. However, their susceptibility to a diverse range of diseases poses a significant threat, rapidly spreading within short periods. This not only diminishes strawberry yields but also inflicts financial losses on farmers. Consequently, we've developed an Android app for detecting strawberry diseases, utilizing a dataset comprising real-life images of both healthy and diseased strawberries. In this work, we leverage two datasets to validate the proposed model. The first dataset from Afzaal et al. (2021), encompasses images of strawberries afflicted by seven distinct diseases. Unlike datasets gathered from controlled laboratory settings, this dataset, collected from real fields and greenhouses, presents several challenges including variations in background, complex field conditions, and diverse lighting environments. These variations empower our model to be more robust and adaptable after learning. This dataset comprises 2,500 images of strawberry diseases captured in various greenhouses using mobile phones, under natural illumination conditions in South Korea. Expert verification ensured the accuracy of disease labels. The second dataset is built up by capturing the

TABLE 1 Numbers of annotated images for five disease types.

Strawberry dataset	Number of samples		
	Train	Validation	Test
Normal	1414	404	202
Gray mold	420	120	60
Black spot	364	104	52
Powdery mildew	287	82	41
Rubber	252	72	36
Total	630	180	90

intricacies of real-world scenarios. We trained the models using images of strawberry diseases obtained from the Vietnam Ministry of Agriculture. This dataset consists of 1,000 images categorized into five classes: normal strawberries, gray mold disease, black spot disease, powdery mildew disease, and rubber disease as listed in Table 1. Classifications were based on crucial factors like color, area, density of the diseased part, and the species' shape. Rigorous verification, involving two individuals following guidelines, was conducted to minimize labeling errors. Incorrect images, such as non-strawberry entities, from the controlled lab settings, and out-of-scope images, were meticulously removed.

For the object detection task, precise bounding regions encompassing the strawberries in a full image are imperative. To accomplish this, we used Roboflow to annotate the leaves in each image with bounding boxes. Recognizing that real-world images may contain multiple strawberries or a combination of diseased and healthy strawberries, we carefully labeled each leaf with its corresponding class. During the labeling process, we ensured that the bounding box fully encompassed the strawberries and that its area was no less than approximately one-eighth of the image size. After completing the annotations, we divided the dataset into training, validation, and test sets with an 8:1:1 ratio. This division allowed us to train this proposed model on a substantial portion of the data while reserving separate subsets for model validation and final performance evaluation. Subsequently, the model training process commenced, encompassing the essential steps for achieving effective object detection on strawberry images.

4.2 Implementation details

4.2.1 Training details

The proposed models were trained on the strawberry dataset utilizing Google Colab with a high RAM runtime and Tesla V100 GPU configuration. After the training process was completed, we obtained the weight sets for each model. The effectiveness of each model was then evaluated using the test dataset. Lastly, the performance of the proposed models was compared with both the backbone version and alternative methods.

4.2.2 Metrics

In the domains of action recognition and detection, evaluation metrics include average accuracy and video-level mAP at specific

IoU thresholds. For our study, we use an IoU threshold of 0.5 to assess detection performance. Furthermore, we report the computational complexity in terms of FLOPs associated with network inferences. To ensure accurate measurement and avoid unnecessary computations, the model is frozen prior to calculating FLOPs. The total number of operations across all convolution layers is then determined based on factors such as the number of output feature maps, kernel sizes, and both input and output channels.

4.3 Experimental results on the dataset

To verify the effectiveness of the proposed models, YOLOv8s-Transformer and YOLOv8s-Transformer-Ghost, we conducted experiments on the dataset sourced from the study (Afzaal et al., 2021). This dataset was collected in real fields and greenhouses and processed by Korean researchers, encompassing seven different diseases on strawberries, comprising 2,500 samples captured using camera-equipped mobile phones. We compared the proposed models, YOLOv8s-Transformer and YOLOv8s-Trans-Ghost, with the original YOLOv8s and YOLOv5s. The results of the comparative experiments are presented in Table 2.

Regarding accuracy, we observe a significant improvement in models incorporating the transformer in both YOLOv5s and YOLOv8s versions, with mAP values of 89.4 and 91.2%, respectively. These outcomes are higher than those from the original versions and alternatives. However, the number of parameters (params) increases considerably.

In the pursuit of a lightweight model that demands both compactness and accuracy, our YOLOv8s-Trans-Ghost reveals promising results. Combining the Ghost module with the YOLOv8s-Transformer shows favorable outcomes, with a substantial reduction in the number of parameters and GLOPs, at 3.4M and 11.5, respectively. The complexity of the YOLOv8s-Trans-Ghost is even lower than that of the YOLOv5s model, which is considered lightweight. This demonstrates efficiency and responsiveness to the problem of applying the model to the strawberry disease detection system in our context.

From the experiments conducted on the dataset, we observe that the transformer module proves effective in enhancing the model's accuracy. However, it comes with the overhead of an increased number of model parameters. In an effort to mitigate the parameter increase, the implementation of Ghost convolution has demonstrated efficiency by significantly reducing the number of parameters, down to three to four times less compared to that from the original YOLOv8s model. This reduction in parameters is achieved without compromising the overall performance, showcasing the effectiveness of Ghost convolution in optimizing the model's efficiency.

4.4 Experimental results on our dataset

Following the validation of the proposed models on the Afzaal et al. (2021) dataset, we proceeded to train the proposed models on our strawberry dataset.

TABLE 2 Performance of proposed and conventional YOLO models in the Strawberry dataset (Afzaal et al., 2021).

Model	Activation	Loss function	mAP@.5	Params	GLOPS
YOLOv5s	SiLu	CIoU	88.30%	7M	15.8
YOLOv5s-Ghost	SiLu	CIoU	87.30%	5.81M	13.4
YOLOv5s-Transformer	SiLu	CIoU	89.40%	7M	15.8
YOLOv5s-Transformer-Ghost	SiLu	CIoU	86.70%	5.81M	13.2
YOLOv8s	SiLu	CIoU	90.50%	11M	28.5
YOLOv8s-Ghost	Mish	CIoU	88.80%	5.18M	19.9
YOLOv8s-Ghost	SiLu	CIoU	86.10%	5.18M	19.9
YOLOv8s-Transformer	SiLu	SIoU	91.20%	20.5M	30.2
YOLOv8s-Trans-Ghost	Mish	SIoU	88.30%	3.4M	11.5
YOLOv8s-Trans-Ghost	Mish	CIoU	87.60%	3.4M	11.5

The bold values indicate highest value.

4.4.1 Efforts for YOLOv8s-transformer accuracy model

The first approach involves augmenting accuracy. The proposed SC3T module is integrated into the final layer of the backbone, forming the YOLOv8s-Transformer. Subsequently, the Ghost module is incorporated to reduce the model's size and complexity. In this experiment, we compare the results of the improved YOLOv8s models with other integrated models in YOLOv5s with the embedded modules such as ShuffleNetv2, EfficientNet, ShuffleNetv2-Transformer, EfficientNet-Transformer, and the original ones.

As listed in Table 3 the original YOLOv8s model has 11,127,519 parameters, mAP@.5 of 72.3, and GPLOPs of 28.4. In contrast, the YOLOv8s-Transformer model has shown a significant increase improvement of nearly 6% compared to the original model in terms of mAP@.5. Similar results are observed for the YOLOv5s model, where the YOLOv5s-Transformer model with an accuracy of 75.7% outperforms the other integrated models based on YOLOv5s, yielding a 2.1% increase over the original version. However, a limitation arises as the increased accuracy comes with a substantial increase in the number of parameters, computational operations, and model complexity.

The proposed models using the transformer and Ghost modules have demonstrated impressive results and outperformed all other models in all metrics. With precision at 80.4%, recall at 70.3%, mAP@.5 at 80.3%, and a reduced parameter count of 3.4M, half of the size of the YOLOv5s model, alongside a decreased GLOPS of 11.5, the YOLOv8s-Trans-Ghost model exhibits remarkable efficiency. Comparing the nine models in Table 3, it is evident that the YOLOv8s-Trans-Ghost model significantly improves accuracy while maintaining model simplicity. Moreover, the parameter amounts and GLOPS of this model are notably reduced, highlighting the superior performance of the YOLOv8s-Trans-Ghost model.

4.4.2 Efforts for YOLOv8s-trans-ghost lightweight model

With the goal of finding a sufficiently lightweight model applicable to real-time systems while ensuring high accuracy, we

conducted experiments to explore and integrate the embedded modules that have shown promising results for a lightweight model, such as Ghost and CBam in YOLOv5, integrated into the YOLOv8s architecture to become the proposed YOLOv8s-Transformer. We compared the results of these models, considering various activations and loss functions in the experiments.

As illustrated in Table 4, YOLOv8s-Transformer models integrated with Ghost modules exhibited a substantial reduction in the number of parameters (Params), with 3.4M parameters, less than one-third of that from the original model and a half that from the YOLOv5s model. The GLOPS index also reveals a significant improvement, decreasing by over 2.5 times compared to that from the original model. Notably, the YOLOv8s-Trans-Ghost-Pretrain model with precision(P) and mAP@.5 both exceeding 80%, incorporating Silu activation and SIoU loss, not only maintains accuracy but also shows a remarkable increase compared to all other models. This underscores the effectiveness of applying proposed models to hardware systems, ensuring real-time performance while maintaining high accuracy in object detection.

4.5 Experiments for loss functions

To assess the effectiveness of different loss functions on the proposed YOLOv8s-Transformer model, we explored five loss functions—CIoU (2020), DIoU, GIoU, EIoU, and SIoU (2022). These functions are compared based on the mAP values, a critical indicator for evaluating the target detection model's performance, where a higher mAP signifies superior accuracy in detecting target objects.

As illustrated in Table 5, the YOLOv8s-Transformer model with SIoU loss outperforms the other models, with mAP@.5 values of 75.7 and 79.8% for YOLOv5s and YOLOv8s, representing increases of 2.1 and 7.5%, respectively, compared to the original versions. This indicates that employing the SIoU loss function effectively reduces sensitivity to position deviations of small objects, addressing the localization issue of small objects and enhancing training accuracy.

TABLE 3 Performance of the proposed accuracy model, YOLO, and alternatives on our dataset.

Models	Precision	Recall	mAP@.5	Params	GLOPS
YOLOv5s	76.40%	61.80%	73.60%	7M	15.8
YOLOv5s-Transformer	79.20%	62.10%	75.70%	7.5M	16.4
YOLOv5s-ShuffleNetv2	73.00%	68.00%	73.50%	21M	40.4
YOLO5s-EfficientNet	75.10%	57.00%	67.80%	23M	43.9
YOLO5s-shuffle-trans	67.40%	65.00%	67.80%	36M	50.2
YOLO5s-efficient-trans	78.90%	55.00%	65.60%	38M	53.6
YOLOv8s	79.90%	60.90%	72.30%	11M	28.4
YOLOv8s-transformer	74.80%	69.70%	78.10%	20.5M	38.2
YOLOv8s-Trans-Ghost-SIoU-Pretrain	80.40%	70.30%	80.30%	3.4M	11.5

The bold values indicate highest value.

TABLE 4 Performance of the proposed lightweight model, YOLO, and alternatives on our dataset.

Model	Activation	Loss	<i>P</i>	mAP@.5	Params	GLOPS
YOLOv5s	SiLu	CIoU	76.40%	73.60%	7M	15.8
YOLOv8s	SiLu	CIoU	79.90%	72.30%	11M	28.4
YOLOv8s-Ghost	SiLu	CIoU	75.00%	72.10%	5.2M	19.9
YOLOv8s-Trans-Ghost	Mish	CIoU	66.70%	69.80%	3.4M	11.5
YOLOv8s-Trans-Ghost	Mish	SIoU	67.30%	72.60%	3.4M	11.5
YOLOv8s-Trans-Ghost	SiLu	SIoU	70.70%	74.40%	3.4M	11.5
YOLOv8s-Trans-Ghost-CBAM	SiLu	CIoU	75.70%	70.20%	3.5M	15.5
YOLOv8s-Trans-Ghost-Pretrain	SiLu	SIoU	80.40%	80.30%	3.4M	11.5

The bold values indicate highest value.

4.6 Ablation experiments

We proposed four key improvements to the YOLOv8s model: (1) introduction of the SC3T module, (2) addition of the Ghost module to the backbone, (3) utilization of the SIoU loss function, and (4) application of pretraining from dataset (Afzaal et al., 2021) to our YOLOv8s-Trans-Ghost model.

In Table 6, aiming to create a model with higher accuracy, we introduced the transformer module, resulting in a 5.8% increase in mAP@.5. After replacing the loss function with SIoU, accuracy further lifted by 1.7%, reaching 79.8%. To create a lightweight model, the Ghost module is incorporated into the original model, reducing the parameter count by half to 5.2 million compared to the original one, 11 million, and significantly dropping the GLOPS index to 19.9. However, the accuracy decreases, prompting us to combine the two modules to form the YOLOv8s-Trans-Ghost model. This leads to an increase in mAP@.5 to 74.4%, a substantial reduction in parameters to 3.4 million, and a notable improvement in GLOPs to 11.5. This demonstrated an increased accuracy while significantly reducing the model’s weight.

Continuing the pursuit of accuracy enhancement, we pre-trained the YOLOv8s-Trans-Ghost model by the dataset (Afzaal

et al., 2021). The mAP@.5 in the proposed model surges to an astonishing 80.3%, surpassing those from the other models where the model maintains its lightweight structure.

In summary, our YOLOv8s-Trans-Ghost model outperforms the others due to the following key enhancements:

SC3T transformer module: the proposed SC3T module, combining the SPP and C3TR structures, is placed in the last layer of the backbone. The SPP kernel has a consistent stride but varying sizes (5×5 , 9×9 , and 13×13), with feature concatenation through channel concatenation. C3TR consists of a transformer block at the three outputs of the detection network, combined with concat weighted to fuse features obtained from the transformer block with those from other parts of the network, such as the SPP kernel outputs or intermediate features from the backbone. Furthermore, a standard transformer layer is employed to aggregate global information from the final block of the backbone network. The transformer encoder features a Multi-head Self-Attention (MSA) mechanism, which updates and combines query (Q), key (K), and value (V) tensors that encode global features from various spatial locations for linear projection. This self-attention mechanism excels at capturing contextual details and reducing the loss of global information.

Ghost module: this module is incorporated into the YOLOv8s backbone by replacing the original Conv module, compressing input feature layers through non-linear and linear convolution operations, resulting in a reduced parameter count and improved GLOPs index.

SIoU loss function: the SIoU loss function is a variation of CIoU loss that normalizes distance and diagonal distance by the width and height of the ground truth box. This normalization allows SIoU to handle objects with different scales, a common issue in object detection.

TABLE 5 Performance of YOLOv5s, YOLO5s-transformer, YOLOv8, and YOLOv8-transformer using different loss functions on our dataset.

Model	Loss function	mAP@.5	mAP@.95
YOLOv5s	CIoU	73.60%	47.40%
YOLOv5s-Transformer	DIoU	69.70%	44.20%
YOLOv5s-Transformer	EIoU	64.70%	38.10%
YOLOv5s-Transformer	GIoU	64.40%	40.00%
YOLOv5s-Transformer	CIoU	75.60%	48.60%
YOLOv5s-Transformer	SIoU	75.70%	46.10%
YOLOv8s	CIoU	72.30%	49.40%
YOLOv8s-Transformer	DIoU	78.20%	50.20%
YOLOv8s-Transformer	EIoU	76.90%	49.70%
YOLOv8s-Transformer	GIoU	77.90%	48.50%
YOLOv8s-Transformer	CIoU	78.10%	52.90%
YOLOv8s-Transformer	SIoU	79.80%	52.30%

The bold values indicate highest value.

Pre-trained models: initially, we pre-trained the YOLOv8s-Trans-Ghost model on a test dataset. Subsequently, we trained the model on our dataset based on the pre-trained model's weights. The performance was improved significantly as the pre-trained model learned common features from a large dataset, enabling good generalization for different tasks.

These collective improvements make YOLOv8s more accurate and sensitive, especially in real-time detection scenarios, while maintaining a lightweight structure.

4.7 Comparision and discussion on the dataset of the proposed and conventional models

In Table 7, we present a comparative analysis of findings from other datasets with similar characteristics. Ouyang et al. (2013) conducted a study focusing on three types of strawberry diseases. Their approach began with initial segmentation, where diseased strawberries were isolated using digital image processing and pattern recognition techniques. They then compared the performance of a neural network with that of an SVM classifier. Although an exact accuracy figure was not provided, it was concluded that SVM achieved a higher recognition rate than the neural network as a classifier.

Kim et al. (2021) dataset, on the other hand, indicated that the Cascaded Faster R-CNN model, pre-trained on ImageNet with four classes, yielded a result of 78.05%. Another study by Nie et al. (2019) reported a baseline mAP@.5 of 88.05% using Faster R-CNN with ImageNet pre-trained weights. This study further improved performance through a cascaded architecture and additional pre-trained weights from the PlantCLEF dataset. However, it is important to note that their model, designed for coarse-grained object detection, differs from our focus on fine-grained instance segmentation. In a related experiment (Afzaal et al., 2021), tests involving Mask R-CNN with pre-training on the MS COCO dataset achieved a mAP@.5 of 82.43%. In this same study, a comparison with YOLACT was conducted to validate their results. Regarding our accuracy model, the YOLOv8s-Transformer, used in experiments on this dataset, we opted not to use pretraining

TABLE 6 Performance of the ablation analyses of the proposed model based on YOLOv8s on our dataset.

Model	Loss	P	R	mAP@.5	Params	GLOPS
YOLOv8s	CIoU	79.90%	60.90%	72.30%	11M	28.4
YOLOv8s-Ghost	CIoU	75.00%	61.00%	72.10%	5.2M	19.9
YOLOv8s-Ghost	SIoU	77.10%	60.00%	71.50%	5.2M	19.9
YOLOv8s-Transformer	CIoU	74.80%	69.70%	78.10%	20.5M	38.2
YOLOv8s-Transformer	SIoU	78.80%	65.90%	79.80%	20.5M	38.2
YOLOv8s-Trans-Ghost	SIoU	70.70%	69.80%	74.40%	3.4M	11.5
YOLOv8s-Trans-Ghost-Pretrain	SIoU	80.40%	70.30%	80.30%	3.4M	11.5

The bold values indicate highest value.

TABLE 7 Comparison of the proposed and relevant models in dataset (Afzaal et al., 2021).

References	Models	Pretrain datasets	Class no.	Accuracy
Ouyang et al. (2013)	SVM Faster R-CNN	N/A	3	N/A
Nie et al. (2019)	CNN+ Attention	ImageNet	4	78.05%
Kim et al. (2021)	Cascaded Faster R-CNN	PlantCLEF	7	91.62%
Afzaal et al. (2021)	Mask R-CNN	MS-COCO	7	82.43%
	YOLOv8s	MS-COCO	7	79.71%
Our work	YOLOv8s-Transformer	N/A	7	91.2%

The bold values indicate highest value.

but rather trained the model from scratch. Nevertheless, the results were notably high, reaching 91.2%.

4.8 Visualization

Following the quantitative assessments presented above, we proceed to conduct a visual evaluation based on images identified by the models.

4.8.1 Qualitative evaluation of the proposed backbone and original models

To demonstrate the detection performance of the proposed model, we randomly selected images from the test dataset for evaluation. The results are shown in Figure 7, where the highlighted areas represent the network's detection outputs.

According to the experimental findings, it's apparent that the standard YOLOv8 model had difficulty detecting objects in the images, particularly when a strawberry was partially obscured by overlapping ones, which completely hid its visibility. The integration of SC3T significantly improved the network's sensitivity and adaptability for detecting small objects by expanding the receptive field. This enhancement also boosted feature recognition and utilization efficiency, reducing missed detections through effective feature synthesis across different scales. Additionally, the use of Ghost convolution in the model's lightweight optimization has simplified deployment. The enhanced YOLOv8s model shows improved detection performance and confidence compared to the standard YOLOv8 in certain cases. Nevertheless, some missed detections remain, highlighting the need for further optimization to meet practical detection requirements.

4.8.2 Qualitative assessment of model effectiveness with different loss functions

Next, in Figure 8 we visualize the performance improvement of the transformer model by comparing two versions, YOLOv5 and YOLOv8, with five different loss functions: DIOU, GIoU, CIOU, EIOU, and SIOU. Through the images of gray mold disease below, it is evident that most models detect accurately, and YOLOv8s-Trans with SIOU loss demonstrates the highest accuracy at 91.2%, surpassing most other models. This highlights the effectiveness of accurately detecting objects in the proposed models with the SIOU loss function.

4.8.3 Qualitative evaluation of the proposed and conventional models on the dataset

We conduct an inference evaluation on the test dataset (Afzaal et al., 2021). In Figure 9, we can observe that the YOLOv8s-Trans-Ghost model accurately detects various diseases on the leaves, even when the leaves are partially hidden in the image's upper corner. In contrast, the other models either fail to detect the precise frame of the leaves in the image or, if detected, result in a large bounding area with low accuracy. This exhibits that the proposed model is effective in object detection as compared to the other models.

4.8.4 Quantization model for edge devices

In this study, we converted the proposed YOLOv8s-Trans-Ghost-Pretrain model on our dataset to the standard format for Edge device applications. ONNX (Open Neural Network Exchange) serves as an open format designed for the representation of machine learning models, enabling their portability across diverse platforms. Through the process of exporting our model to ONNX, we gain the ability to deploy it across various devices, thereby harnessing hardware acceleration to enhance inference speed in real-time applications. Particularly, we loaded the checkpoint data during training and initialized a YOLOX model. To export the model to ONNX with different input sizes, we set the width and height input axes as dynamic. Finally, when performing inference with ONNX Runtime, we focused on the capabilities needed for inference. ONNX Runtime is a cross-platform inference engine which is a machine-learning accelerator suitable for mobile devices.

4.8.5 Building android apps

Upon completion of the training and quantization processes for the model, our next step involved the development of an Android application. Specifically, we followed seven steps outlined below:

1. Model training: we trained the model to generate a weight file in PyTorch (.pt format), leveraging the robust capabilities of this framework in deep learning research.
2. Conversion to ONNX format: the YOLO model (.pt format) was converted into the ONNX format with FP32 weights, ensuring compatibility and interoperability across different frameworks and platforms.

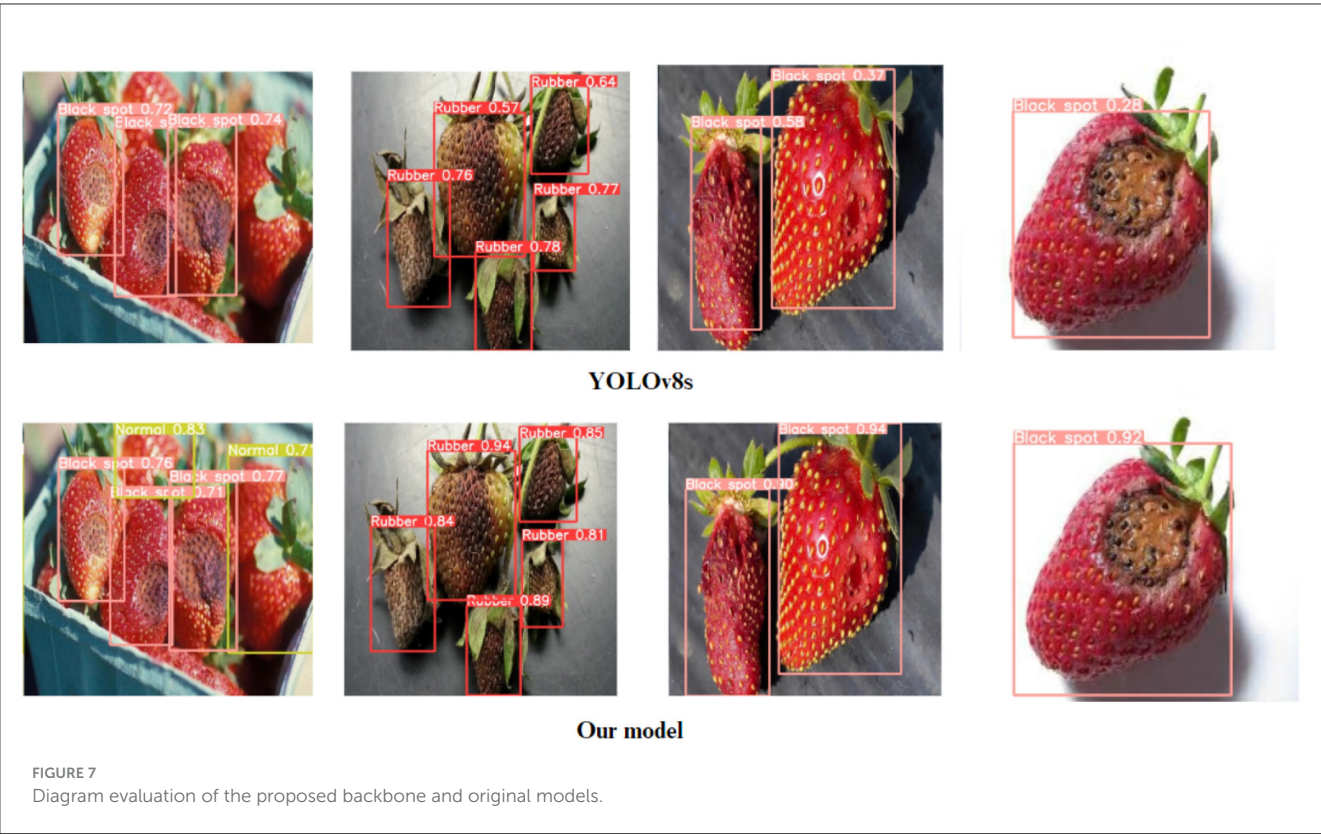


FIGURE 7
Diagram evaluation of the proposed backbone and original models.

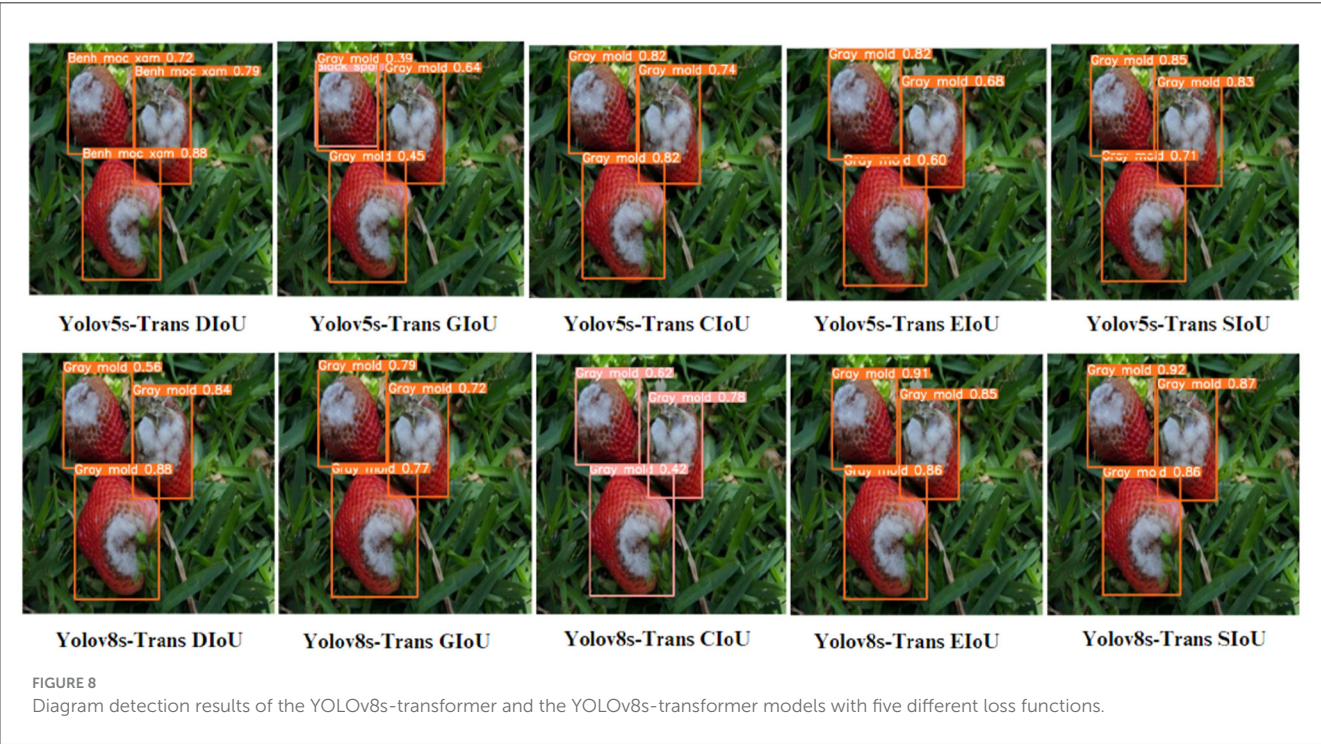


FIGURE 8
Diagram detection results of the YOLOv8s-transformer and the YOLOv8s-transformer models with five different loss functions.

3. Quantization with ONNX runtime: to optimize model performance and efficiency, we performed quantization using ONNX Runtime, resulting in an optimized ONNX file ready for deployment.
4. Conversion to TensorFlow model: the ONNX Runtime model (.onnx file) was further converted into a TensorFlow model (.pb file), facilitating seamless integration with TensorFlow-based applications and frameworks.

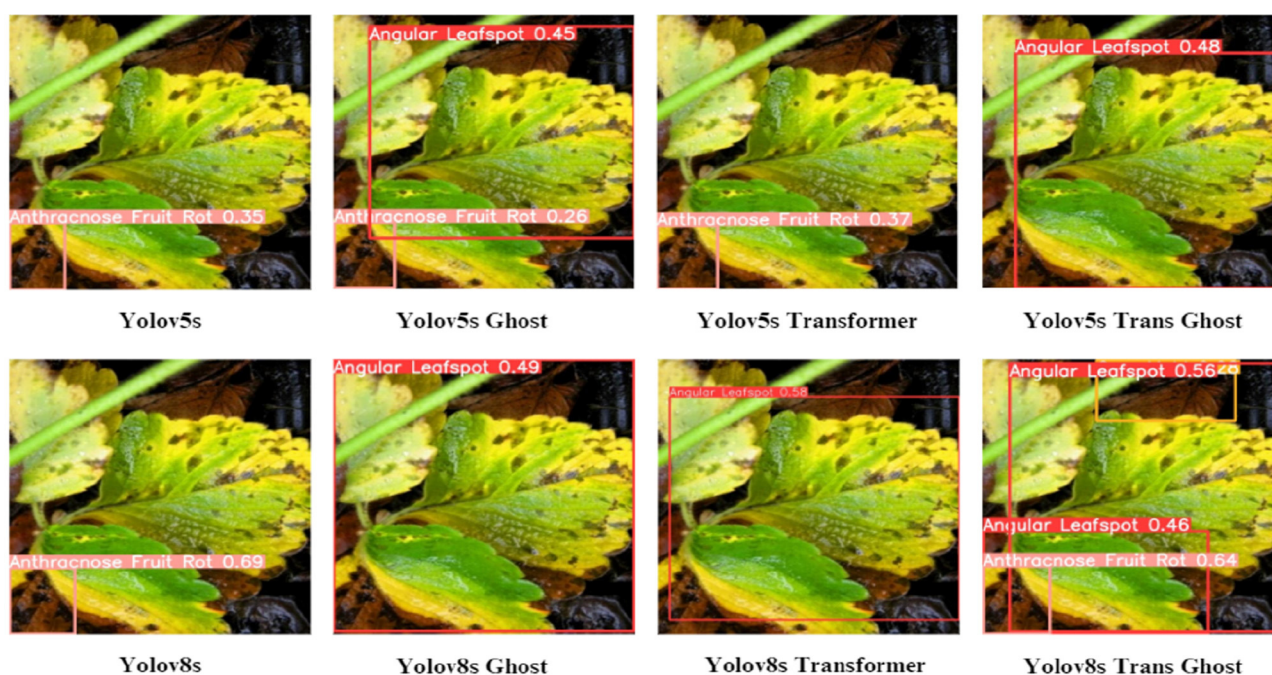


FIGURE 9

Diagram evaluation of the proposed and conventional models on the dataset (Afzaal et al., 2021).

5. Conversion to TensorFlow lite model: subsequently, the TensorFlow model (.pb model) underwent the conversion into a TensorFlow Lite model, suitable for deployment on mobile and edge devices with limited computational resources.
6. Android studio setup: to begin development for the Android platform, we downloaded and installed Android Studio, the official Integrated Development Environment (IDE) for Android application development.
7. App development and deployment: finally, leveraging the features and tools provided by Android Studio, we proceeded to build and run our object detection App on Android devices, demonstrating the practical implementation and real-world applicability of our deep learning model in object detection tasks.

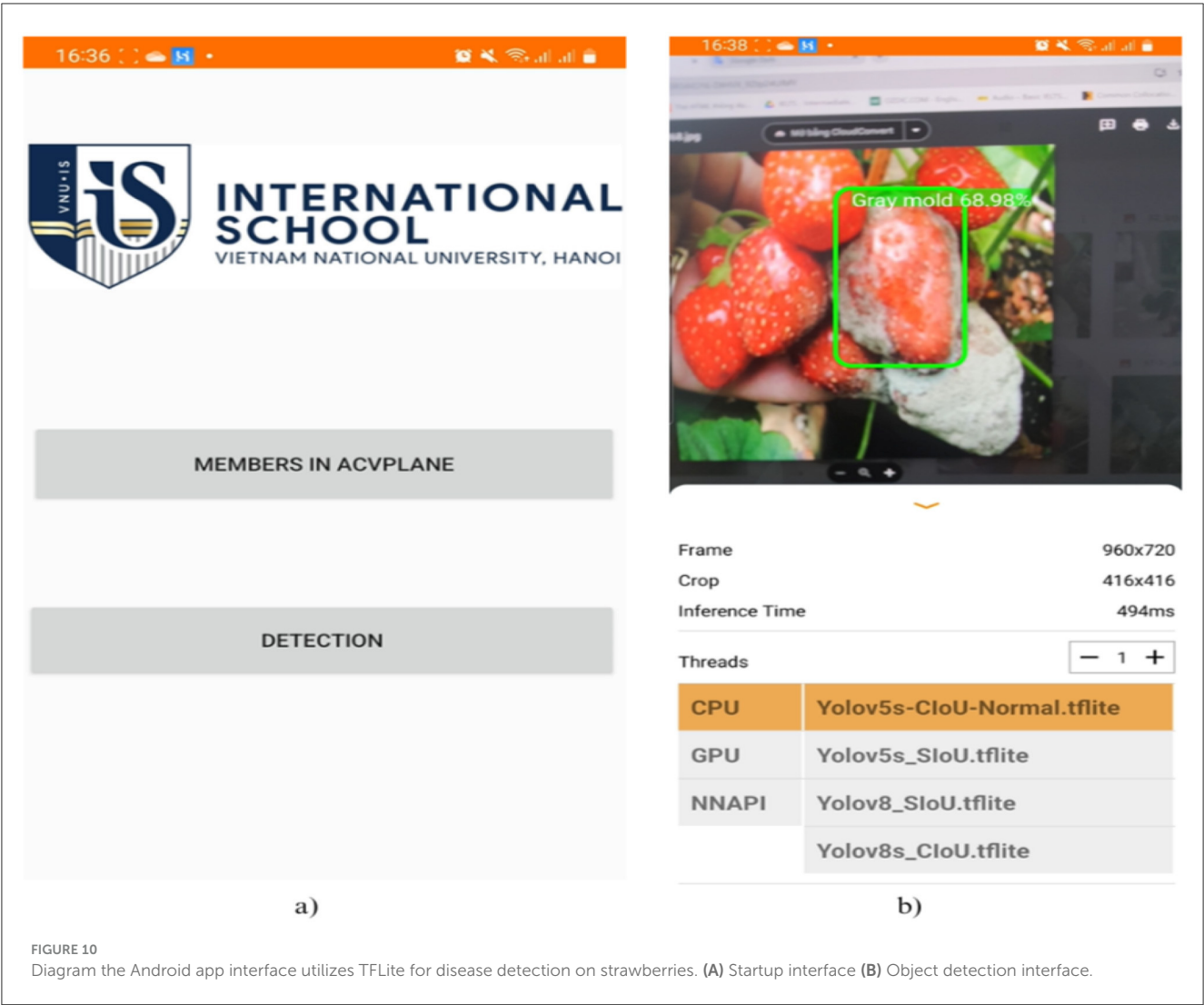
Through these comprehensive steps, we seamlessly transitioned from model training and optimization to the development and deployment of a fully functional Android application, showcasing the practical utility of our deep learning research in a real-world application.

Figure 10 reveals our Android app interface, which is the product created by members of our research team at the International School at Vietnam National University. This app has performed excellently on mobile phones, detecting diseases on strawberries, and can flexibly switch between four different YOLOv5 and YOLOv8 models.

5 Conclusion

This work introduces a valuable dataset focusing on diseases in strawberries. Leveraging this dataset, we conducted experiments

on the proposed YOLOv8s-Trans-Ghost, aiming for fast and accurate target detection. In this model, the SC3T module which is a combination of the SPP and C3TR modules, was proposed. This module excels in feature synthesis across different sizes, incorporating transformer-based long-term dependency modeling and convolution-based visual feature extraction to enhance object localization precision and classification performance. Experiments on the YOLOv8s-Transformer model demonstrated a promising mAP result of 78.1%, with an increase of 5.8% over the baseline model on our dataset. The adoption of the SIOU loss function, replacing CIOU, further increases the mAP up to 79.8%, showcasing the effectiveness of this loss function in precise object detection. However, the model's parameter count doubles to 20.5M, hindering real-time applicability when embedded in a practical hardware platform. To address this, we substituted Conv in the YOLOv8s-Transformer with Ghost Conv, which is proven effective in reducing parameters in YOLOv5 versions. Experimental results are also very promising, with a parameter count reduced to 3.4M, nearly one-third of the original one. Subsequently, we employed the weights of the proposed YOLOv8s-Trans-Ghost model which was trained on the large test dataset (Afzaal et al., 2021) as pretraining for the application on our dataset. The value of mAP significantly increases to 80.3%, accompanied by notable improvements in other metrics, such as precision at 80.4% and recall at 70.3%, surpassing those from the other models. Parameters and GLOPS are substantially lessened reduced to 3.4M and 11.5, respectively. These results reveal that the model proposed offers a lightweight and efficient solution for detecting diseases in strawberries while maintaining high precision, with the potential to significantly promote production efficiency.



Finally, for deploying the proposed model, we adopted the compression technique for model optimization to meet the computation capabilities of mobile devices by ONNX Runtime. Additionally, the Android app was successfully built to effectively spread the proposed model for the applications of detecting strawberry diseases.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

M-TD: Conceptualization, Methodology, Writing – review & editing, Data curation, Software, Validation, Visualization, Writing – original draft. M-HH: Conceptualization, Methodology, Writing – review & editing, Funding acquisition, Investigation,

Project administration, Supervision. D-CN: Conceptualization, Investigation, Methodology, Resources, Visualization, Writing – review & editing. OT-C: Conceptualization, Methodology, Resources, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported by the International School, Vietnam National University, Hanoi (VNU-IS) under the Decree of No. 95/2014/NĐ-CP.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Afzaal, U., Bhattarai, B., Pandeya, Y. R., and Lee, J. (2021). An instance segmentation model for strawberry diseases based on mask R-CNN. *Sensors* 21:6565. doi: 10.3390/s21196565
- Azizpour, H., Razavian, A. S., Sullivan, J., Maki, A., and Carlsson, S. (2015). Factors of transferability for a generic convnet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 1790–1802. doi: 10.1109/TPAMI.2015.2500224
- Baweja, H. S., Parhar, T., Mirbod, O., and Nuske, S. (2018). "Stalknet: a deep learning pipeline for high-throughput measurement of plant stalk count and stalk width," in *Field and Service Robotics*, eds. M. Hutter, and R. Siegwart (Cham: Springer International Publishing), 271–284. doi: 10.1007/978-3-319-67361-5_18
- Chen, L.-C., Papandreu, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 834–848. doi: 10.1109/TPAMI.2017.2699184
- Chen, O. T. C., Ha, M. H., and Lee, Y. L. (2020). Computation-affordable recognition system for activity identification using a smart phone at home. *IEEE Int. Symp. Circuits Syst.* 2020, 1–5. doi: 10.1109/ISCAS45731.2020.9180826
- Chen, Y., Lee, W. S., Gan, H., Peres, N., Fraisse, C., Zhang, Y., et al. (2019). Strawberry yield prediction based on a deep neural network using high-resolution aerial orthoimages. *Remote Sens.* 11:1584. doi: 10.3390/rs11131584
- Cheng, X., Zhang, Y., Chen, Y., Wu, Y., and Yue, Y. (2017). Pest identification via deep residual learning in complex background. *Comput. Electron. Agric.* 141, 351–356. doi: 10.1016/j.compag.2017.08.005
- Cui, Y., Song, Y., Sun, C., Howard, A., and Belongie, S. (2018). "Large scale fine-grained categorization and domain-specific transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 4109–4118. doi: 10.1109/CVPR.2018.00432
- Do, M.-T., Ha, M.-H., Nguyen, D.-C., Thai, K., and Do Ba, Q.-H. (2023). "Human detection based yolo backbones-transformer in UAVs," in *2023 International Conference on System Science and Engineering (ICSSE)* (Ho Chi Minh: IEEE), 576–580. doi: 10.1109/ICSSE58758.2023.10227141
- Dong, X., Yan, S., and Duan, C. (2022). A lightweight vehicles detection network model based on yolov5. *Eng. Appl. Artif. Intell.* 113:104914. doi: 10.1016/j.engappai.2022.104914
- Fang, W., Wang, L., and Ren, P. (2019). Tinier-yolo: a real-time object detection method for constrained environments. *IEEE Access* 8, 1935–1944. doi: 10.1109/ACCESS.2019.2961959
- Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Comput. Electron. Agric.* 145, 311–318. doi: 10.1016/j.compag.2018.01.009
- Gallo, I., Rehman, A. U., Dehkordi, R. H., Landro, N., Grassa, R. L., Boschetti, M., et al. (2023). Deep object detection of crop weeds: performance of yolov7 on a real case dataset from UAV images. *Remote Sens.* 15:539. doi: 10.3390/rs15020539
- Gevorgyan, Z. (2022). Siou loss: more powerful learning for bounding box regression. *arXiv [Preprint]*. arXiv:2205.12740. doi: 10.48550/arXiv.2205.12740
- Girshick, R. (2015). "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision* (Santiago: IEEE), 1440–1448. doi: 10.1109/ICCV.2015.169
- Gu, Y., Hu, Z., Zhao, Y., Liao, J., and Zhang, W. (2024). MFGTN: a multi-modal fast gated transformer for identifying single trawl marine fishing vessel. *Ocean Eng.* 303:117711. doi: 10.1016/j.oceaneng.2024.117711
- Ha, M. H., and Chen, O. T. C. (2021). Deep neural networks using residual fast-slow refined highway and global atomic spatial attention for action recognition and detection. *IEEE Access* 9, 164887–164902. doi: 10.1109/ACCESS.2021.3134694
- Ha, M. H., Nguyen, D. C., Do, M. T., Kim, D. T., Le, X. H., Pham, N. T., et al. (2024). Plant pathology identification using local-global feature level based on transformer. *Indones. J. Electr. Eng. Comput. Sci.* 34, 1582–1592. doi: 10.11591/ijeecs.v34.i3.pp1582-1592
- He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice: IEEE), 2961–2969. doi: 10.1109/ICCV.2017.322
- Herranz, L., Jiang, S., and Li, X. (2016). "Scene recognition with CNNs: objects, scales and dataset bias," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 571–579. doi: 10.1109/CVPR.2016.68
- Jayawardena, R. S., Huang, J. K., Jin, B. C., Yan, J. Y., Li, X. H., Hyde, K. D., et al. (2016). An account of colletotrichum species associated with strawberry anthracnose in china based on morphology and molecular data. *Mycosphere* 7, 1147–1191. doi: 10.5943/mycosphere/si/2c/6
- Kim, B., Han, Y.-K., Park, J.-H., and Lee, J. (2021). Improved vision-based detection of strawberry diseases using a deep neural network. *Front. Plant Sci.* 11:559172. doi: 10.3389/fpls.2020.559172
- Kornblith, S., Shlens, J., and Le, Q. V. (2019). "Do better imagenet models transfer better?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach: CA: IEEE), 2661–2671. doi: 10.1109/CVPR.2019.00277
- Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., et al. (2017). "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 2117–2125. doi: 10.1109/CVPR.2017.106
- Liu, H., Sun, F., Gu, J., and Deng, L. (2021). Sf-yolov5: a lightweight small object detection algorithm based on improved feature fusion mode. *Sensors* 21:7752. doi: 10.3390/s21155817
- Liu, J., and Wang, G. (2021). Plant diseases and pests detection based on deep learning: a review. *Plant Methods* 17:22. doi: 10.1186/s13007-021-00722-9
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., et al. (2016). "SSD: single shot multibox detector," in *Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14* (Cham: Springer International Publishing), 21–37. doi: 10.1007/978-3-319-46448-0_2
- Lu, Y., Zhang, L., and Xie, W. (2020). "Yolo-compact: an efficient yolo network for single category real-time object detection," in *2020 Chinese Control and Decision Conference (CCDC)* (Hefei: IEEE), 1931–1936. doi: 10.1109/CCDC49329.2020.9164580
- Mahmud, M. S., Zaman, Q. U., Esau, T. J., Price, G. W., and Prithiviraj, B. (2019). Development of an artificial cloud lighting condition system using machine vision for strawberry powdery mildew disease detection. *Comput. Electron. Agric.* 158, 219–225. doi: 10.1016/j.compag.2019.02.007
- Mohanty, S. P., Hughes, D. P., and Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Front. Plant Sci.* 7:1419. doi: 10.3389/fpls.2016.01419
- Nie, X., Wang, L., Ding, H., and Xu, M. (2019). Strawberry verticillium wilt detection network based on multi-task learning and attention. *IEEE Access* 7, 170003–170011. doi: 10.1109/ACCESS.2019.2954845
- Ouyang, C., Li, D., Wang, J., Wang, S., and Han, Y. (2013). "The research of the strawberry disease identification based on image processing and pattern recognition," *Computer and Computing Technologies in Agriculture VI: 6th IFIP WG 5.14 International Conference, CCTA 2012, Zhangjiajie, China, October 19–21, 2012, Revised Selected Papers, Part I 6* (Cham: Springer Berlin Heidelberg), 69–77.
- Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 779–788. doi: 10.1109/CVPR.2016.91
- Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., McCool, C., et al. (2016). Deepfruits: A fruit detection system using deep neural networks. *Sensors* 16:1222. doi: 10.3390/s16081222
- Selvaraj, M. G., Vergara, A., Ruiz, H., Safari, N., Elayabalan, S., Ocimati, W., et al. (2019). Ai-powered banana diseases and pest detection. *Plant Methods* 15, 1–11. doi: 10.1186/s13007-019-0475-z
- Shruthi, U., and Nagaveni, V. (2024). Tomseynet: a hybrid CNN model for accurate tomato disease identification with severity level assessment. *Neural Comput. Appl.* 36, 5165–5181. doi: 10.1007/s00521-023-09351-w
- Song, W., Wang, X., Guo, Y., Li, S., Xia, B., Hao, A., et al. (2024). Centerformer: a novel cluster center enhanced transformer for unconstrained dental plaque segmentation. *IEEE Trans. Multimedia* 26, 10965–10978. doi: 10.1109/TMM.2024.3428349

- Sozzi, M., Cantalamessa, S., Cogato, A., Kayad, A., and Marinello, F. (2022). Automatic bunch detection in white grape varieties using yolov3, yolov4, and yolov5 deep learning algorithms. *Agronomy* 12:319. doi: 10.3390/agronomy12020319
- Xu, R., Lin, H., Lu, K., Cao, L., and Liu, Y. (2021). A forest fire detection system based on ensemble learning. *Forests* 12:217. doi: 10.3390/f12020217
- Xu, X., Zhang, X., and Zhang, T. (2022). Lite-yolov5: a lightweight deep learning detector for on-board ship detection in large-scene sentinel-1 SAR images. *Remote Sens.* 14:1018. doi: 10.3390/rs14041018
- Yan, B., Fan, P., Lei, X., Liu, Z., and Yang, F. (2021). A real-time apple targets detection method for picking robot based on improved yolov5. *Remote Sens.* 13:1619. doi: 10.3390/rs13091619
- Yao, J., Qi, J., Zhang, J., Shao, H., Yang, J., Li, X., et al. (2021). A real-time detection algorithm for kiwifruit defects based on yolov5. *Electronics* 10:1711. doi: 10.3390/electronics10141711
- Yao, S., Chen, Y., Tian, X., Jiang, R., and Ma, S. (2020). An improved algorithm for detecting pneumonia based on yolov3. *Appl. Sci.* 10:1818. doi: 10.3390/app10051818
- Yu, Y., Zhang, K., Yang, L., and Zhang, D. (2019). Fruit detection for strawberry harvesting robot in non-structural environment based on mask-RCNN. *Comput. Electron. Agric.* 163:104846. doi: 10.1016/j.compag.2019.06.001
- Yu, Y., Zhao, J., Gong, Q., Huang, C., Zheng, G., Ma, J., et al. (2021). Real-time underwater maritime object detection in side-scan sonar images based on transformer-yolov5. *Remote Sens.* 13:3555. doi: 10.3390/rs13183555
- Zhang, B., Liu, X., Wang, H., and Li, Z. (2022). Enhancing plant disease recognition using deep learning and data augmentation. *J. Plant Dis. Protect.* 129, 45–56.
- Zhao, L., and Li, S. (2020). Object detection algorithm based on improved yolov3. *Electronics*, 9, 537. doi: 10.3390/electronics9030537
- Zhao, Q., Sheng, T., Wang, Y., Tang, Z., Chen, Y., Cai, L., et al. (2019). M2det: a single-shot object detector based on multi-level feature pyramid network. *Proc. AAAI Conf. Artif. Intell.* 33, 9259–9266. doi: 10.1609/aaai.v33i01.33019259
- Zheng, Z., Wang, P., Ren, D., Liu, W., Ye, R., Hu, Q., et al. (2021). Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE Trans. Cybern.* 52, 8574–8586. doi: 10.1109/TCYB.2021.3095305
- Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017). Places: a 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 1452–1464. doi: 10.1109/TPAMI.2017.2723009
- Zhu, X., Lyu, S., Wang, X., and Zhao, Q. (2021). “TPH-yolov5: improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Montreal, BC: IEEE), 2778–2788. doi: 10.1109/ICCVW54120.2021.00312



OPEN ACCESS

EDITED BY

Sokratis Makrogiannis,
Delaware State University, United States

REVIEWED BY

Rocco Zaccagnino,
University of Salerno, Italy
Nikos Kanakaris,
University of Southern California,
United States

*CORRESPONDENCE

Austin A. Barr
✉ austin.barr@ucalgary.ca

RECEIVED 24 November 2024

ACCEPTED 23 January 2025

PUBLISHED 05 February 2025

CITATION

Barr AA, Quan J, Guo E and Sezgin E (2025)
Large language models generating synthetic
clinical datasets: a feasibility and comparative
analysis with real-world perioperative data.
Front. Artif. Intell. 8:1533508.
doi: 10.3389/frai.2025.1533508

COPYRIGHT

© 2025 Barr, Quan, Guo and Sezgin. This is
an open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or reproduction
is permitted which does not comply with
these terms.

Large language models generating synthetic clinical datasets: a feasibility and comparative analysis with real-world perioperative data

Austin A. Barr^{1*}, Joshua Quan¹, Eddie Guo¹ and Emre Sezgin^{2,3}

¹Cumming School of Medicine, University of Calgary, Calgary, AB, Canada, ²The Abigail Wexner Research Institute, Nationwide Children's Hospital, Columbus, OH, United States, ³Department of Pediatrics, The Ohio State University College of Medicine, Columbus, OH, United States

Background: Clinical data is instrumental to medical research, machine learning (ML) model development, and advancing surgical care, but access is often constrained by privacy regulations and missing data. Synthetic data offers a promising solution to preserve privacy while enabling broader data access. Recent advances in large language models (LLMs) provide an opportunity to generate synthetic data with reduced reliance on domain expertise, computational resources, and pre-training.

Objective: This study aims to assess the feasibility of generating realistic tabular clinical data with OpenAI's GPT-4o using zero-shot prompting, and evaluate the fidelity of LLM-generated data by comparing its statistical properties to the Vital Signs DataBase (VitalDB), a real-world open-source perioperative dataset.

Methods: In Phase 1, GPT-4o was prompted to generate a dataset with qualitative descriptions of 13 clinical parameters. The resultant data was assessed for general errors, plausibility of outputs, and cross-verification of related parameters. In Phase 2, GPT-4o was prompted to generate a dataset using descriptive statistics of the VitalDB dataset. Fidelity was assessed using two-sample *t*-tests, two-sample proportion tests, and 95% confidence interval (CI) overlap.

Results: In Phase 1, GPT-4o generated a complete and structured dataset comprising 6,166 case files. The dataset was plausible in range and correctly calculated body mass index for all case files based on respective heights and weights. Statistical comparison between the LLM-generated datasets and VitalDB revealed that Phase 2 data achieved significant fidelity. Phase 2 data demonstrated statistical similarity in 12/13 (92.31%) parameters, whereby no statistically significant differences were observed in 6/6 (100.0%) categorical/binary and 6/7 (85.71%) continuous parameters. Overlap of 95% CIs were observed in 6/7 (85.71%) continuous parameters.

Conclusion: Zero-shot prompting with GPT-4o can generate realistic tabular synthetic datasets, which can replicate key statistical properties of real-world perioperative data. This study highlights the potential of LLMs as a novel and accessible modality for synthetic data generation, which may address critical barriers in clinical data access and eliminate the need for technical expertise, extensive computational resources, and pre-training. Further research is warranted to enhance fidelity and investigate the use of LLMs to amplify and

augment datasets, preserve multivariate relationships, and train robust ML models.

KEYWORDS

synthetic data, large language model, artificial intelligence, machine learning, ChatGPT, big data, education, surgery

1 Introduction

Clinical data is fundamental to advance medical research and enable the development of machine learning (ML) models. This is particularly relevant in surgical care, as procedural medicine increasingly relies on decision-making based on large-scale data (Maier-Hein et al., 2017). However, access to real-world (real) clinical data is constrained by ethical, legal, and logistical barriers (Pavlenko et al., 2020; Wartenberg and Thompson, 2010). Data requests often require institutional review board approval, data sharing agreements, and compliance with various data privacy regulations (e.g., HIPAA, GDPR, PIPEDA) (Bentzen et al., 2021; Ness, 2007). In many institutions, clinical datasets are proprietary and restricted to internal use. Clinical data also requires significant pre-processing and de-identification procedures which is resource intensive and can delay or hinder research projects—particularly for students, trainees, and early career researchers (Tudur et al., 2017; Willemink et al., 2020). These protections, while essential for patient privacy, limit the accessibility of real clinical data.

In addition to regulatory concerns, clinical data is often incomplete (Newgard and Lewis, 2015). Data collected in clinical settings may suffer from missing values, errors, or biases introduced during data entry. Furthermore, data de-identification processes commonly remove or obscure personal health information (e.g., date of birth, date of operation, demographic data, geographic data). These constraints limit the reliability of analyses as well as the accuracy and generalizability of ML models trained on this data. Collectively, these challenges underscore the need for alternative approaches to provide researchers, learners, and developers with necessary data while preserving patient privacy.

Synthetic clinical datasets, which are artificially generated rather than captured as real patient information, offer a potential solution to the challenges associated with accessing and using real patient data (Beaulieu-Jones et al., 2019; Bellovin et al., 2019; van Breugel and van der Schaar, 2023). Synthetic data can be shared, analyzed, and used freely, bypassing the regulatory and logistical obstacles associated with real data use (El Emam et al., 2020). Despite the potential of synthetic data, achieving high utility, fidelity, and privacy remains a significant challenge (Jordon et al., 2022). Current methods of synthetic data generation, including generative adversarial networks (GANs) (Goodfellow et al., 2020) and variational autoencoders (VAEs) (Kingma and Welling, 2022), have demonstrated utility in synthetic data generation (Goncalves et al., 2020; Jacobs et al., 2023; Rajotte et al., 2022). However, privacy concerns have been raised regarding generative models (Chen et al., 2020; Hayes et al., 2018) including data extractions, model inversions, and membership inference attacks (Rajotte et al., 2022). Furthermore, despite some workarounds, there are issues with mode collapse (Thanh-Tung and Tran, 2020) and the applicability of GANs toward generating categorical and binary data (Jacobs et al., 2023). The use of GANs and VAEs is also self-limiting to those with technical expertise (e.g., complex architecture, fine-tuning) as well as access to necessary computational resources and reference datasets for training.

In recent years, large language models (LLMs)—a computational model capable of language generation and other natural language processing tasks—offer new possibilities for generating text that is coherent and contextually relevant (Brown et al., 2020; Nazir and Wang, 2023). A prominent and publicly available LLM, OpenAI's ChatGPT, has shown utility in generating synthetic text-based data (Calvo-Lorenzo and Uriarte-Llano, 2024; Hämäläinen et al., 2023; Li et al., 2021). However, the potential of generating tabular synthetic clinical data with ChatGPT remains largely unexplored. Use of LLMs for synthetic data generation may offer an accessible alternative to GANs and VAEs, reducing the need for specialized knowledge and computational resources, which could broaden the reach of synthetic data use in research and ML model development.

This study aims to assess the feasibility of generating realistic tabular clinical data with OpenAI's GPT-4o (Hurst et al., 2024) using zero-shot prompting, and evaluate the fidelity of LLM-generated data by comparing its statistical properties to the Vital Signs DataBase (VitalDB) (Lee et al., 2022), a real open-source multi-parameter perioperative dataset.

2 Methods

2.1 Overview

We conducted a two-phase study to evaluate the feasibility of generating synthetic clinical datasets with GPT-4o using a single prompt and without pre-training. In both phases, GPT-4o was prompted to generate a synthetic dataset based on 13 clinical parameters derived from VitalDB. In Phase 1, GPT-4o was prompted with high-level qualitative descriptions of the 13 clinical parameters, to assess its ability to generate a complete and contextually relevant tabular dataset without guiding statistics. In Phase 2, GTP-4o was prompted to generate a synthetic dataset using descriptive statistics of the VitalDB dataset. Both Phase 1 and 2 datasets were statistically compared to VitalDB, with Phase 1 data serving as a baseline for comparisons.

2.2 Real dataset

The real clinical dataset used as a comparator in this study is the open-source VitalDB. The VitalDB dataset is a perioperative dataset consisting of multi-parameter data from surgery patients who underwent routine and emergency non-cardiac (general, thoracic, urological, and gynecological) operations at Seoul National University Hospital (Seoul, Korea) from August 2016 to June 2017 (Lee et al., 2022). The dataset included 6,388 de-identified cases encompassing a wide range of clinical parameters including demographic, preoperative, intraoperative, and postoperative parameters.

The VitalDB dataset was selected due to its open-source availability and data completeness for parameters spanning the entire perioperative

period. The VitalDB dataset also included a variety of data formats (i.e., numerical, text), variable types (i.e., continuous, categorical, binary), and distributions (i.e., normal, skewed). These considerations ensured a comprehensive evaluation of GPT-4o's ability to generate and replicate statistical properties of a wide array of clinical data.

2.3 Parameter selection and data cleaning

The VitalDB dataset was reviewed for data completeness and parameters with missing data were excluded. Included parameters ($n = 13$) were chosen based on relevance to perioperative care and to represent a range of data formats and variable types. Remaining parameters with similar data formats, variable types, or clinical information were excluded for redundancy in the context of a feasibility study. Timepoint variables in the VitalDB dataset were recorded as the duration from an assigned case start time in seconds. Two time variables were included and converted to hours: operation duration (difference between operation end time and operation start time) and postoperative length of stay (difference between discharge time and operation end time). Selected parameters are presented in [Table 1](#).

To further ensure quality and relevance of the data used for comparison, case files for patients younger than 18 ($n = 57$), older than 89 ($n = 8$), missing an American Society of Anesthesiologists (ASA) physical status classification ($n = 130$), and with negative discharge times ($n = 27$) were excluded from the dataset. In total, $n = 6,166$ cases were included.

2.4 Generation of synthetic datasets

Prior to the generation of Phase 1 and 2 synthetic datasets, GPT-4o was not pre-trained or provided any patient data from the VitalDB dataset. In Phase 1, GPT-4o was prompted with qualitative descriptions of 13 clinical parameters and asked to generate corresponding data for 6,166 patients. The prompt did not include descriptive statistics, definitions (e.g., ASA physical status classification), or formulas to calculate parameters (e.g., body mass index (BMI)). The prompt used to generate the Phase 1 synthetic dataset is presented in [Box 1](#).

In Phase 2, GPT-4o was prompted with descriptive statistics of the VitalDB dataset. For continuous parameters (age, height, weight, operation duration, postoperative length of stay), descriptive statistics included mean, standard deviation, and range. Descriptive statistics were not provided for BMI, and GPT-4o was instructed to calculate this

parameter using each case file's corresponding height and weight. Descriptive statistics for height were inputted to GPT-4o as centimeters, requiring the LLM to convert the height parameter to meters in order to calculate BMI—this transformation was not specifically instructed within the prompt. For categorical and binary parameters (ASA physical status classification, operation type, biological sex, preoperative hypertension, preoperative diabetes mellitus, intraoperative transfusion), corresponding proportions were provided. GPT-4o was instructed to assign ascending whole number values for each case ID. For time variables, natural log transformations were used to normalize skewed distributions and GPT-4o was provided with descriptive statistics of the log-transformed values. The prompt used to generate the Phase 2 synthetic dataset is presented in [Box 2](#).

GPT-4o application programming interfaces (API) were not used in order to determine feasibility of data generation without further technical expertise and resources.

2.5 Dataset analysis

The Phase 1 dataset was assessed for general errors, plausibility of outputs, and cross-verification of related parameters. Assessment of general errors evaluated missing data, unexpected outputs, and formatting issues in the tabular data output. Plausibility of outputs involved evaluating time variables for positive values, ASA physical status classification for values between 1 and 6, and categorical and binary parameters for expected values (i.e., only including categories provided in the prompt, category proportions add to 100%). Cross-verification of related parameters involved confirming that all BMI values were appropriately calculated given the corresponding height and weight for each case file.

The Phase 1 and 2 datasets were compared to the VitalDB dataset for statistical similarity. Continuous variables were compared using two-sample *t*-tests and 95% CI overlap. Given the large sample size, we used parametric two-sample *t*-tests. The log-transformed values of operation duration and postoperative length of stay were used for statistical testing with the two-sample *t*-tests for Phase 2 data. For each continuous parameter, 95% CI overlap was calculated as the proportion of shared values compared to the entire range of values within both 95% CIs from LLM-generated and VitalDB datasets. The Python library Matplotlib was used to generate figures visualizing the overlap of 95% CI for continuous parameters and proportional alignment of categorical and binary parameters. Categorical and binary variables were compared using two-sample proportion tests. Statistical testing was performed using RStudio v.4.4.2 and statistical significance was set at 0.05. For two-sample *t*-tests and proportion tests, *p*-values above 0.05 indicated statistically insignificant differences in means and proportions, therefore representing an effective replication of descriptive statistical properties from the VitalDB reference dataset.

2.6 Ethical considerations

Datasets generated in this study solely represented fictitious patient data. Use of the VitalDB dataset was used in accordance with the requirements outlined by the study team. No data from the VitalDB dataset was inputted directly into GPT-4o for pre-training and no direct data is included in this paper. Furthermore, the synthetic

TABLE 1 Summary of selected parameters from the VitalDB dataset.

Category (<i>n</i>)	Parameters (units)
Demographic data (6)	Case ID, age (years), biological sex (M/F), height (cm), weight (kg), BMI (kg/m ²)
Preoperative morbidity (3)	ASA physical status classification (1–6), preoperative hypertension (yes/no), preoperative diabetes mellitus (yes/no)
Intraoperative data (3)	Operation type, operation duration (hours), intraoperative transfusion (yes/no)
Postoperative outcomes (1)	Postoperative length of stay (hours)

BMI, body mass index, ASA, American Society of Anesthesiologists.

Create a table of realistic patient data for non-cardiac (general, thoracic, urological, and gynecological) surgery patients with the following columns populated for 6166 patients who underwent surgery. Ensure that values for each patient make sense in the context of other values in that row:

Column 1: Case ID (count up from 1)
Column 2: Operation time (duration of the surgery in hours)
Column 3: Post-operative length of stay (duration of patient's length of stay in hospital following surgery in hours)
Column 4: Age of the patient (years)
Column 5: Height of the patient (cm)
Column 6: Weight of the patient (kg)
Column 7: BMI of the patient (calculated using the patient's height and weight)
Column 8: Biological sex of the patient ("M" or "F")
Column 9: ASA physical status classification
Column 10: Operation type ("biliary/pancreas", "breast", "colorectal", "hepatic", "major resection", "minor resection", "others", "stomach", "thyroid", "transplantation", "vascular")
Column 11: Whether the patient had preoperative hypertension (0 or 1)
Column 12: Whether the patient had preoperative diabetes (0 or 1)
Column 13: Whether the patient received intraoperative transfusion (0 or 1)

Re-check the data to ensure that all conditions are met before displaying. Every condition must be met. Provide a downloadable excel file of the dataset

BOX 1

Prompt input to generate the Phase 1 synthetic dataset with GPT-4o.

datasets generated by GPT-4o were evaluated solely for research purposes and not used in any form of clinical decision-making.

3 Results

In Phase 1, GPT-4o generated a complete and structured tabular dataset comprising 6,166 case files. All 13 expected columns were present, complete, and appropriately labeled; no missing data, unexpected outputs, or formatting issues were present within the generated dataset. Furthermore, all generated time variables were positive, ASA physical status classifications were within the appropriate range (1–6), and categorical and binary parameters only included expected values and proportions all added to 100%. Each case also included a correctly calculated BMI corresponding to the appropriate height and weight.

Review of calculated means and ranges for continuous variables included operation duration (6.46 h; 1.00–12.00), postoperative length of stay (154.84 h; 12.00–299.90), age (53.52 years; 18.00–89.00), height (174.51 cm; 150.00–199.00), weight (97.43 kg; 45.00–149.00), and BMI (32.62 kg/m²; 11.40–66.20). All continuous parameters showed plausible means and ranges for a perioperative dataset. However, proportions among categorical and binary variables did not differ based on context. Proportions were evenly spread across categories for parameters which are likely to demonstrate uniform distributions (e.g., sex, operation type) as well as parameters that may have skewed distributions (e.g., ASA physical status classification, preoperative

comorbidities, intraoperative transfusions). A percent stacked bar plot displaying proportional alignment of categorical and binary parameters can be seen in [Figure 1](#). Overall, generated data in Phase 1 was realistic, displayed appropriate ranges, included correct calculations without the provision of descriptive statistics, formulas, or unit conversions (e.g., BMI), and maintained definitional boundaries of parameters without explicit instructions (e.g., ASA physical status classification).

The Phase 1 and 2 datasets were statistically compared to the VitalDB dataset. The results of the statistical testing ([Tables 2, 3](#)) revealed that 12/13 (92.31%) parameters from the Phase 2 dataset did not show statistically significant differences from VitalDB, including 6/6 (100.00%) of the categorical and binary parameters and 6/7 (85.71%) of the continuous parameters. The only continuous parameter which demonstrated statistically significant differences in Phase 2 was the BMI parameter, which was calculated based on each case's height and weight rather than generated based on descriptive statistics in the prompt. For the Phase 1 dataset, 2/13 (15.28%) parameters did not show statistically significant differences from VitalDB, one of which was the Case ID parameter. Overlap of 95% CI was observed in 6/7 (85.71%) of the Phase 2 continuous parameters. The measured 95% CI overlaps were as follows: case ID (100.0%), weight (85.93%), height (61.31%), age (43.12%), postoperative length of stay (34.84%), operation duration (15.17%), and BMI (0.0%). The Phase 1 dataset only showed 95% CI overlap in the case ID parameter (100.0%). Visualization of the 95% CI overlaps of each continuous parameter is displayed in [Figure 2](#). Overall, 12/13 (92.31%) of the Phase 2 parameters met the predefined threshold for statistical

Generate a table with 13 columns with the following headers:

Column 1: "Case ID"
 Column 2: "log(Operation Time)"
 Column 3: "log(Post-Operative LOS)"
 Column 4: "Age"
 Column 5: "Height"
 Column 6: "Weight"
 Column 7: "BMI"
 Column 8: "Sex"
 Column 9: "ASA Classification"
 Column 10: "Operation Type"
 Column 11: "Preoperative HTN"
 Column 12: "Preoperative DM"
 Column 13: "Intraoperative Transfusion"

Where the table contains realistic patient data for 6166 patients, based on the following conditions and ensure that values for each patient make sense in the context of other values in that row:

Column 1: count up from 1
 Column 2: mean = 0.533, standard deviation = 0.804, range = -3.746 to 2.767
 Column 3: mean = 4.527, standard deviation = 1.13, range = -1.08 to 8.585
 Column 4: mean = 57.73, standard deviation = 14.276, range = 18 to 89
 Column 5: mean = 162.51, standard deviation = 8.608, range = 130.5 to 188.6
 Column 6: mean = 61.747, standard deviation = 11.518, range = 24.4 to 139.7
 Column 7: BMI calculated based on patient's corresponding height and weight
 Column 8: 50.81% ("M"), 49.19% ("F")
 Column 9: 28.77% (1), 59.58% (2), 10.88% (3), 0.57% (4), 0.19% (6)
 Column 10: 12.86% ("biliary/pancreas"), 6.91% ("breast"), 21.38% ("colorectal"), 4.07% ("hepatic"), 9.28% ("major resection"), 8.73% ("minor resection"), 12.23% ("others"), 10.83% ("stomach"), 4.12% ("thyroid"), 5.77% ("transplantation"), 4.15% ("vascular")
 Column 11: 31.25% (1), 68.75% (0)
 Column 12: 10.43% (1), 89.57% (0)
 Column 13: 5.29% (1), 94.71% (0)

All numbers in the table must be positive, except for in columns 2 and 3. Re-check the data to ensure that all conditions are met before displaying. Every condition must be met exactly. Re-iterate until exactly correct. Provide a downloadable excel file of the dataset

BOX 2

Prompt input to generate the Phase 2 synthetic dataset with GPT-4o.

similarity, demonstrating the parameter effectively replicated statistical properties of corresponding data from the VitalDB dataset.

The Excel files containing the Phase 1 and 2 synthetic datasets generated with GPT-4o are available in the [Supplementary data](#).

4 Discussion

4.1 Overview

The present study evaluated the feasibility of generating tabular synthetic clinical data with GPT-4o using zero-shot prompting, and

assessed the fidelity of the generated data by comparing it to a real clinical dataset, VitalDB. By examining two phases of data generation—one using qualitative prompts (Phase 1) and another incorporating descriptive statistics from VitalDB (Phase 2)—we explored GPT-4o's capacity to generate plausible data and replicate statistical properties, variable distributions, and contextual characteristics typical of clinical data. Generated data included various formats (numerical, text), variable types (continuous, categorical, binary), and distributions (normal, skewed), covering demographic, preoperative, intraoperative, and postoperative data. The results indicate: (1) GPT-4o can produce realistic synthetic data without descriptive statistics or reference data, and (2) GPT-4o can generate

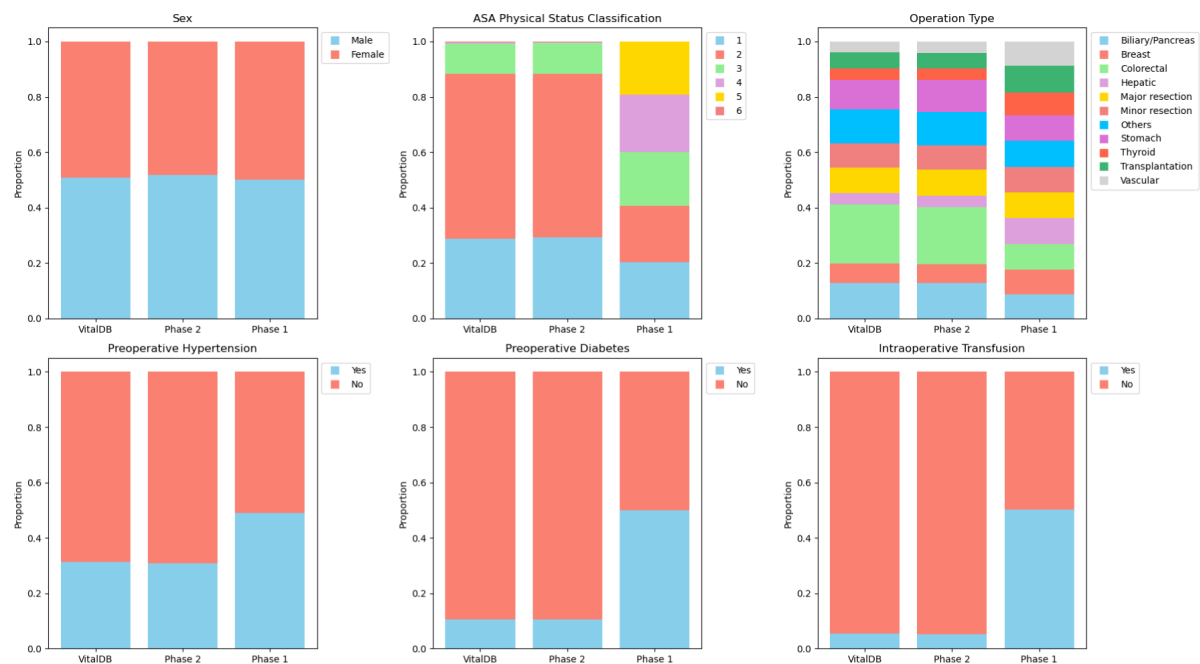


FIGURE 1

Percent stacked bar plot displaying proportional alignment of categorical and binary parameters between VitalDB, Phase 1, and Phase 2 datasets.

datasets that align closely with real clinical data, when provided with statistical guidance.

4.2 Principal findings and implications

The use of LLMs to generate structured tabular data using zero-shot prompting is a novel concept. Consistent with other modalities of synthetic data generation, LLM-generated data has the potential to address many of the challenges of accessing and using real clinical data (Beaulieu-Jones et al., 2019; El Emam et al., 2020; van Breugel and van der Schaar, 2023). While GPT-4o has yet to demonstrate equivalent fidelity and utility to GANs and VAEs, LLMs may offer solutions to some of their shortcomings. The generation of data using GANs and VAEs requires technical expertise and computational resources. However, data generation using LLMs is accessible to anyone with an internet connection, and can produce clean and ready-to-use datasets, outputted in a downloadable Excel file, through plain-language prompting. This holds substantial implications for democratizing data access in research, educational contexts, and ML model development (Rajotte et al., 2022).

Given that no reference data was required in Phase 1 or inputted for pre-training in Phase 2, LLMs may overcome privacy concerns associated with current approaches to synthetic data generation (Rajotte et al., 2022). The Phase 1 data, which generated realistic clinical data in the absence of guiding statistics, definitions, and formulas, further emphasizes the contextual relevance of outputs from GPT-4o (Brown et al., 2020; Nazir and Wang, 2023). This is particularly useful in educational contexts, whereby learning opportunities for students and trainees can be enhanced by practicing data analysis using synthetic data generated with desired statistical properties and without requiring a reference dataset. LLM-generated datasets can also include synthetic

personal health information which would otherwise be removed or de-identified. Since synthetic data can be used without restriction, data can also be re-inputted into LLMs for analysis, which further broadens prospects and scope of future research.

Clean and structured synthetic datasets, with the statistical similarity Phase 2 data demonstrated to VitalDB, has vast implications for data-driven medicine and ML model development. Perioperative data, in particular, is inherently heterogeneous, encompassing a wide variety of sources, formats, and qualities (Maier-Hein et al., 2017). By generating synthetic data which can replicate real-world data distributions, researchers can bypass additional challenges associated with the use of raw clinical data. In this way, synthetic datasets can accelerate the development of predictive tools and surgical decision support systems, ultimately contributing to patient care and surgical outcomes.

4.3 Limitations

While this feasibility study demonstrated remarkable preservation of within-column statistical properties and simple relationships between variables, there are some limitations and challenges to consider. First, this study focused solely on GPT-4o, and it remains uncertain whether similar results would be achieved using other LLMs. Similarly, direct comparisons in performance between LLMs, GANs, and VAEs are necessary to assess their relative utility, fidelity, and privacy preservation, which may uncover additional limitations and strengths. At present, it is unclear whether bivariate and multivariate relationships were retained by the LLM-based approach, as this was not directly assessed. Demonstrating the preservation of correlations and other nuanced interdependencies, present in clinical data, is necessary before meaningful comparisons can

TABLE 2 Means with associated 95% CI intervals, *p*-values from associated two-sample *t*-tests, and 95% CI overlap values comparing case ID, operation duration, postoperative length of stay, age, height, weight, and BMI between VitalDB and Phase 1 and 2 datasets.

Parameter	VitalDB mean (95% CI)	VitalDB vs Phase 1 <i>p</i> -value	VitalDB vs Phase 1 95% CI overlap	Phase 1 mean (95% CI)	VitalDB vs Phase 1 <i>p</i> -value	Phase 2 mean (95% CI)	VitalDB vs Phase 2 <i>p</i> -value	VitalDB vs Phase 2 95% CI overlap
Case ID	3083.5 (3039.06–3127.94)	1.000	100.0%	3083.5 (3039.06–3127.94)	1.000	3083.5 (3039.06–3127.94)	1.000	100.0%
Operation duration (hours)	2.27 (2.22–2.31)	<0.001	0.0%	6.46 (6.38–6.54)	<0.001	2.33 (2.28–2.39)	0.116*	15.17%
Postoperative length of stay (hours)	167.19 (160.48–173.90)	<0.001	0.0%	154.84 (152.78–156.90)	<0.001	173.6 (166.99–180.31)	0.845*	34.84%
Age (years)	57.73 (57.37–58.09)	<0.001	0.0%	53.52 (53.00–54.04)	<0.001	58.01 (57.66–58.37)	0.819	43.12%
Height (cm)	162.51 (162.30–162.73)	<0.001	0.0%	174.51 (174.16–174.87)	<0.001	162.41 (162.20–162.62)	0.200	61.31%
Weight (kg)	61.75 (61.46–62.04)	<0.001	0.0%	97.43 (96.67–98.19)	<0.001	61.70 (61.42–61.99)	0.327	85.93%
BMI (kg/m ²)	23.32 (23.23–23.41)	<0.001	0.0%	32.62 (32.33–32.91)	<0.001	23.60 (23.47–23.73)	0.023	0.0%

**p*-values were calculated using log-transformed values.

be made between the performance of LLMs and other data generation methods.

This study also revealed the importance of prompt design in generating accurate and relevant synthetic data outputs. In Phase 1, where prompts lacked descriptive statistics, generated data deviated significantly from the reference dataset. While this underscores the notable results in Phase 2, it also suggests that without explicit guidance, LLMs may produce plausible but statistically unaligned data. Therefore, the quality of outputted data is reliant on effective prompting, and improper prompt design can introduce bias or errors into generated data. It is also unknown whether an iterative approach to prompting may result in greater fidelity.

Continued improvements in LLMs have been previously associated with greater accuracy in a variety of generative and clinically associated tasks (Meyer et al., 2024; Rosol et al., 2023), and further iterations of GPT-4o may improve upon these results and current limitations.

4.4 Future directions

This study’s use of an open-source dataset (VitalDB) as a comparator was intentional to facilitate reproducibility and encourage follow-up research. Future research should continue to investigate the capabilities of LLMs in generating tabular datasets, with particular focus on capturing complex interdependencies between parameters and further assessing reproducibility of results. This should involve using existing and robust frameworks to assess the fidelity and privacy preservation of LLM-generated synthetic datasets (El Emam, 2020; El Emam et al., 2022; Platzer and Reutterer, 2021; Vallevik et al., 2024). Direct comparisons should be made between the performance of GPT-4o and other prominent LLMs. Following further refinement of this zero-shot approach, direct comparisons in utility and privacy should also be conducted between LLMs, GANs, and VAEs, using a systematic benchmarking approach (Yan et al., 2022).

Future work should assess the potential of LLMs in data enhancement, including data amplification and augmentation (El Emam, 2023)—particularly in domains with missing data or limited data availability (e.g., rare diseases, underrepresented patient populations) (Rajotte et al., 2022). By supplementing existing datasets with synthetic data that preserves statistical properties, LLMs could mitigate data scarcity and enable more robust research in data-constrained fields. Applications of LLM-generated synthetic data toward ML model development and validation, predictive tools, and surgical decision support systems should also be explored.

Applications in educational contexts may be evaluated. Surveys, qualitative interviewing, or randomized trials involving students who have used LLM-generated datasets may reveal whether supplementing educational programs with synthetic data can enhance learning for students training toward careers in disciplines which analyze clinical data (e.g., statisticians, data scientists, epidemiologists). It is also recommended to assess whether LLMs can be used to effectively summarize and analyze outputted synthetic data.

5 Conclusion

This study demonstrates that zero-shot prompting with GPT-4o can generate realistic tabular synthetic datasets that replicate key

TABLE 3 Number and proportion of patients by sex, ASA physical status classification, operation type, preoperative hypertension status, preoperative diabetes mellitus status, and intraoperative transfusion status for VitalDB, Phase 1, and Phase 2 datasets with *p*-values from associated two sample proportion tests comparing distributions between VitalDB and synthetic datasets for each parameter.

Parameter	VitalDB <i>n</i> , (%)	Phase 1 <i>n</i> , (%)	VitalDB vs Phase 1 <i>p</i> -value	Phase 2 <i>n</i> , (%)	VitalDB vs Phase 2 <i>p</i> -value
Sex					
Male	3133 (50.81%)	3087 (50.06%)	0.407	3195 (51.82%)	0.263
Female	3033 (49.19%)	3079 (49.94%)		2971 (48.18%)	
ASA physical status classification					
1	1774 (28.77%)	1249 (20.26%)	<0.001	1810 (29.35%)	0.478
2	3674 (59.58%)	1254 (20.34%)	<0.001	3637 (58.98%)	0.497
3	671 (10.88%)	1198 (19.43%)	<0.001	683 (11.08%)	0.726
4	35 (0.57%)	1291 (20.94%)	<0.001	26 (0.42%)	0.246
5	0 (0.0%)	1174 (19.04%)	<0.001	0 (0.0%)	-
6	12 (0.19%)	0 (0.0%)	<0.001	10 (0.16%)	0.667
Operation type					
Biliary/Pancreas	793 (12.86%)	534 (8.66%)	<0.001	782 (12.68%)	0.764
Breast	426 (6.91%)	558 (9.05%)	<0.001	419 (6.80%)	0.803
Colorectal	1318 (21.38%)	556 (9.02%)	<0.001	1274 (20.66%)	0.332
Hepatic	251 (4.07%)	587 (9.52%)	<0.001	260 (4.22%)	0.682
Major resection	572 (9.28%)	569 (9.23%)	0.928	577 (9.36%)	0.881
Minor resection	538 (8.73%)	563 (9.13%)	0.430	536 (8.69%)	0.952
Others	754 (12.23%)	585 (9.49%)	<0.001	750 (12.16%)	0.912
Stomach	668 (10.83%)	571 (9.26%)	0.004	710 (11.51%)	0.230
Thyroid	254 (4.12%)	509 (8.25%)	<0.001	265 (4.30%)	0.624
Transplantation	356 (5.77%)	596 (9.67%)	<0.001	332 (5.38%)	0.347
Vascular	236 (4.15%)	538 (8.73%)	<0.001	261 (4.23%)	0.254
Preoperative hypertension					
Yes	1927 (31.25%)	3028 (49.11%)	<0.001	1899 (30.80%)	0.582
No	4239 (68.75%)	3138 (50.89%)		4267 (69.20%)	
Preoperative diabetes mellitus					
Yes	643 (10.43%)	3075 (49.87%)	<0.001	645 (10.46%)	0.952
No	5523 (89.57%)	3091 (50.13%)		5521 (89.54%)	
Intraoperative transfusion					
Yes	326 (5.29%)	3098 (50.24%)	<0.001	314 (5.09%)	0.624
No	5840 (94.71%)	3068 (49.76%)		5852 (94.91%)	

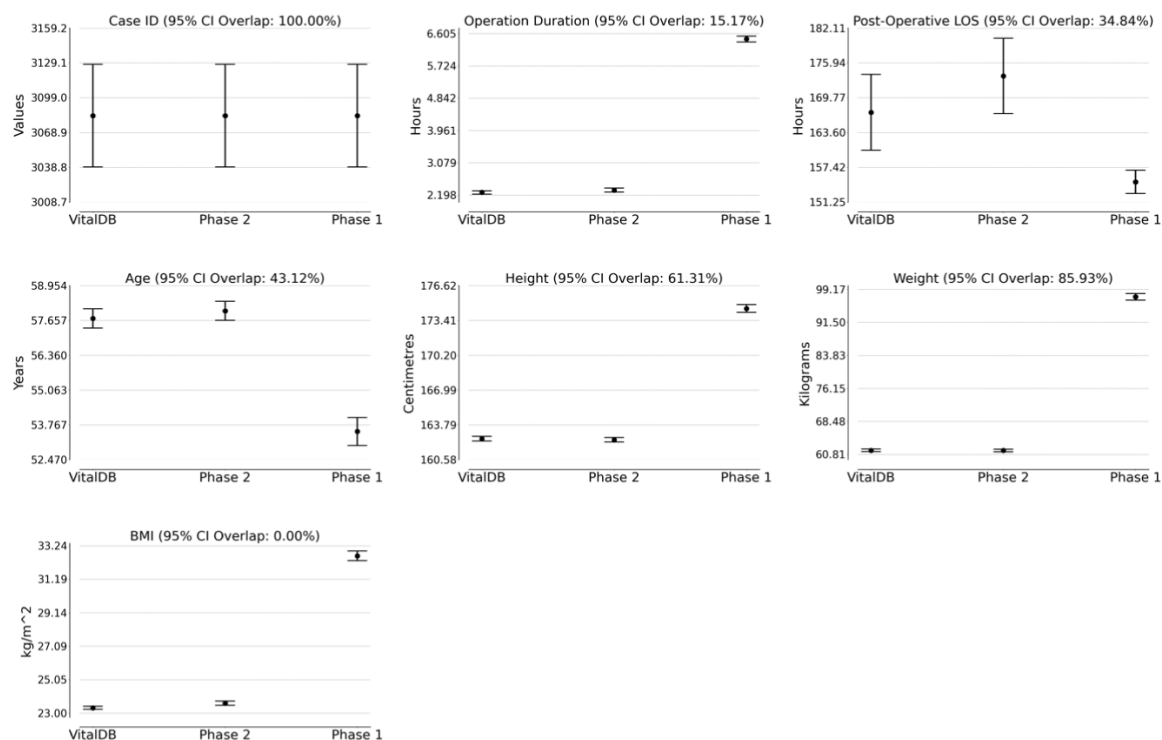


FIGURE 2

95% confidence intervals for VitalDB, Phase 1, and Phase 2 datasets with labeled percentage of 95% CI overlap between VitalDB and Phase 2 data. LOS, Length of Stay; BMI, body mass index.

statistical properties of real perioperative data. By eliminating the need for technical expertise, extensive computational resources, and pre-training in synthetic data generation, LLMs can offer an accessible modality to address critical barriers associated with clinical data access. Collectively, these findings highlight the broad implications of LLM-generated synthetic data in democratizing data access and enhancing educational opportunities. Future research should focus on enhancing fidelity and investigating the application of LLMs in data amplification and augmentation, replication of multivariate relationships, and ML model development.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary material](#), further inquiries can be directed to the corresponding author.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

AB: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. JQ: Formal analysis, Investigation, Methodology, Writing – review & editing. EG: Methodology, Visualization, Writing – review & editing. ES: Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Generative AI was used in the creation of this manuscript. This project made use of GPT-4o, a language

model developed by OpenAI, to generate synthetic data. We acknowledge the contribution of GPT-4o in the present work.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product

that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frai.2025.1533508/full#supplementary-material>

References

- Beaulieu-Jones, B. K., Wu, Z. S., Williams, C., Lee, R., Bhavnani, S. P., Byrd, J. B., et al. (2019). Privacy-preserving generative deep neural networks support clinical data sharing. *Cardiovasc. Qual. Outc.* 12:e005122. doi: 10.1161/CIRCOUTCOMES.118.005122
- Bellovin, S. M., Dutta, P. K., and Reitinger, N. (2019). Privacy and synthetic datasets. *Stan. Tech. L. Rev.* 22:1.
- Bentzen, H. B., Castro, R., Fears, R., Griffin, G., ter Meulen, V., and Ursin, G. (2021). Remove obstacles to sharing health data with researchers outside of the European Union. *Nat. Med.* 27, 1329–1333. doi: 10.1038/s41591-021-01460-0
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Arxiv*. doi: 10.48550/arXiv.2005.14165
- Calvo-Lorenzo, I., and Uriarte-Llano, I. (2024). Massive generation of synthetic medical records with ChatGPT: an example in hip fractures. *Med. Clin.* 162, 549–554. doi: 10.1016/j.medcle.2023.11.033
- Chen, D., Yu, N., Zhang, Y., and Fritz, M. (2020). GAN-leaks: A taxonomy of membership inference attacks against generative models, in Proceedings of the 2020 ACM SIGSAC conference on computer and communications security, (New York, NY, USA: Association for Computing Machinery), 343–362.
- El Emam, K. (2020). Seven ways to evaluate the utility of synthetic data. *IEEE Sec. Privacy* 18, 56–59. doi: 10.1109/MSEC.2020.2992821
- El Emam, K. (2023). Status of synthetic data generation for structured health data. *JCO* e2300071:7. doi: 10.1200/CCI.23.00071
- El Emam, K. E., Mosquera, L., and Bass, J. (2020). Evaluating identity disclosure risk in fully synthetic health data: model development and validation. *J. Med. Internet Res.* 22:e23139. doi: 10.2196/23139
- El Emam, K. E., Mosquera, L., Fang, X., and El-Hussuna, A. (2022). Utility metrics for evaluating synthetic health data generation methods: validation study. *JMIR Med. Inform.* 10:e35734. doi: 10.2196/35734
- Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., and Sales, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Med. Res. Methodol.* 20:108. doi: 10.1186/s12874-020-00977-1
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2020). Generative adversarial networks. *Commun. ACM* 63, 139–144. doi: 10.1145/3422622
- Hämäläinen, P., Tavast, M., and Kunnari, A. (2023). Evaluating large language models in generating synthetic HCI research data: A case study., in Proceedings of the 2023 CHI conference on human factors in computing systems, (New York, NY, USA: Association for Computing Machinery), 1–19.
- Hayes, J., Melis, L., Danezis, G., and De Cristofaro, E. (2018). LOGAN: Membership inference attacks against generative models. *Arxiv*. doi: 10.48550/arXiv.1705.07663
- Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., et al. (2024). GPT-4o system card. *Arxiv*. doi: 10.48550/arXiv.2410.21276
- Jacobs, F., D'Amico, S., Benvenuti, C., Gaudio, M., Saltalamacchia, G., Miggiano, C., et al. (2023). Opportunities and challenges of synthetic data generation in oncology. *JCO* 7:45. doi: 10.1200/CCI.23.00045
- Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., et al. (2022). Synthetic data - what, why and how? *Arxiv*. doi: 10.48550/arXiv.2205.03257
- Kingma, D. P., and Welling, M. (2022). Auto-encoding Variational Bayes. *Arxiv*. doi: 10.48550/arXiv.1312.6114
- Lee, H. C., Park, Y., Yoon, S. B., Yang, S. M., Park, D., and Jung, C.-W. (2022). VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *Sci. Data* 9:279. doi: 10.1038/s41597-022-01411-5
- Li, J., Zhou, Y., Jiang, X., Natarajan, K., Pakhomov, S. V., Liu, H., et al. (2021). Are synthetic clinical notes useful for real natural language processing tasks: a case study on clinical entity recognition. *J. Am. Med. Inform. Assoc.* 28, 2193–2201. doi: 10.1093/jamia/ocab112
- Maier-Hein, L., Vedula, S. S., Speidel, S., Navab, N., Kikinis, R., Park, A., et al. (2017). Surgical data science for next-generation interventions. *Nat. Biomed. Eng.* 1, 691–696. doi: 10.1038/s41551-017-0132-7
- Meyer, A., Riese, J., and Streichert, T. (2024). Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written German medical licensing examination: observational study. *JMIR* 10:e50965. doi: 10.2196/50965
- Nazir, A., and Wang, Z. (2023). A comprehensive survey of ChatGPT: advancements, applications, prospects, and challenges. *Meta Radiol.* 1:100022. doi: 10.1016/j.metrad.2023.100022
- Ness, R. B. (2007). Influence of the HIPAA privacy rule on Health Research. *JAMA* 298, 2164–2170. doi: 10.1001/jama.298.18.2164
- Newgard, C. D., and Lewis, R. J. (2015). Missing data: how to best account for what is not known. *JAMA* 314, 940–941. doi: 10.1001/jama.2015.10516
- Pavlenko, E., Streh, D., and Langhof, H. (2020). Implementation of data access and use procedures in clinical data warehouses. A systematic review of literature and publicly available policies. *BMC Med. Inform. Decis. Mak.* 20:157. doi: 10.1186/s12911-020-01177-z
- Platzer, M., and Reutterer, T. (2021). Holdout-based empirical assessment of mixed-type synthetic data. *Front. Big Data* 4:679939. doi: 10.3389/fdata.2021.679939
- Rajotte, J.-F., Bergen, R., Buckeridge, D. L., El Emam, K., Ng, R., and Strome, E. (2022). Synthetic data as an enabler for machine learning applications in medicine. *iScience* 25:105331. doi: 10.1016/j.isci.2022.105331
- Rosol, M., Gąsior, J. S., Łaba, J., Korzeniewski, K., and Młyńczak, M. (2023). Evaluation of the performance of GPT-3.5 and GPT-4 on the polish medical final examination. *Sci. Rep.* 13:20512. doi: 10.1038/s41598-023-46995-z
- Thanh-Tung, H., and Tran, T. (2020). Catastrophic forgetting and mode collapse in GANs., In 2020 International joint conference on neural networks (IJCNN).
- Tudur, C., Nevitt, S., Appelbe, D., Appleton, R., Dixon, P., Harrison, J., et al. (2017). Resource implications of preparing individual participant data from a clinical trial to share with external researchers. *Trials* 18:319. doi: 10.1186/s13063-017-2067-4
- Vallevik, V. B., Babic, A., Marshall, S. E., Elvatun, S., Brøgger, H. M. B., Alagaratnam, S., et al. (2024). Can I trust my fake data – a comprehensive quality assessment framework for synthetic tabular data in healthcare. *Int. J. Med. Inform.* 185:105413. doi: 10.1016/j.ijmedinf.2024.105413
- van Breugel, B., and van der Schaar, M. (2023). Beyond privacy: Navigating the opportunities and challenges of synthetic data. *Arxiv*. 10.48550/arXiv.2304.03722
- Wartenberg, D., and Thompson, W. D. (2010). Privacy versus public health: the impact of current confidentiality rules. *Am. J. Public Health* 100, 407–412. doi: 10.2105/AJPH.2009.166249
- Willemsink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., et al. (2020). Preparing medical imaging data for machine learning. *Radiology* 295, 4–15. doi: 10.1148/radiol.2020192224
- Yan, C., Yan, Y., Wan, Z., Zhang, Z., Omberg, L., Guinney, J., et al. (2022). A multifaceted benchmarking of synthetic electronic health record generation models. *Nat. Commun.* 13:7609. doi: 10.1038/s41467-022-35295-1



OPEN ACCESS

EDITED BY

Sokratis Makrogiannis,
Delaware State University, United States

REVIEWED BY

Karisma Putra,
Muhammadiyah University of Yogyakarta,
Indonesia
Mosiur Rahaman,
Asia University, Taiwan

*CORRESPONDENCE

Liliek Triyono
✉ liliek.triyono@polines.ac.id

RECEIVED 12 October 2024

ACCEPTED 21 January 2025

PUBLISHED 10 February 2025

CITATION

Triyono L, Gernowo R and Prayitno (2025)
MoNetViT: an efficient fusion of CNN and
transformer technologies for visual navigation
assistance with multi query attention.
Front. Comput. Sci. 7:1510252.
doi: 10.3389/fcomp.2025.1510252

COPYRIGHT

© 2025 Triyono, Gernowo and Prayitno. This
is an open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other forums is
permitted, provided the original author(s) and
the copyright owner(s) are credited and that
the original publication in this journal is cited,
in accordance with accepted academic
practice. No use, distribution or reproduction
is permitted which does not comply with
these terms.

MoNetViT: an efficient fusion of CNN and transformer technologies for visual navigation assistance with multi query attention

Liliek Triyono^{1,2*}, Rahmat Gernowo¹ and Prayitno²

¹Doctoral Program of Information System, Diponegoro University, Central Java, Semarang, Indonesia,
²Department of Electrical Engineering, Politeknik Negeri Semarang, Semarang, Indonesia

Aruco markers are crucial for navigation in complex indoor environments, especially for those with visual impairments. Traditional CNNs handle image segmentation well, but transformers excel at capturing long-range dependencies, essential for machine vision tasks. Our study introduces MoNetViT (Mini-MobileNet MobileViT), a lightweight model combining CNNs and MobileViT in a dual-path encoder to optimize global and spatial image details. This design reduces complexity and boosts segmentation performance. The addition of a multi-query attention (MQA) module enhances multi-scale feature integration, allowing end-to-end learning guided by ground truth. Experiments show MoNetViT outperforms other semantic segmentation algorithms in efficiency and effectiveness, particularly in detecting Aruco markers, making it a promising tool to improve navigation aids for the visually impaired.

KEYWORDS

indoor navigation, computer vision, markers, assistive technology, mobile devices

1 Introduction

Navigating independently is a major challenge for individuals with visual impairments. It affects their ability to perform daily tasks and limits their involvement in social and economic activities. This challenge reduces their personal autonomy and impacts their overall quality of life. Indoor navigation issues for visually impaired individuals persist, as current solutions frequently do not overcome critical limits. Conventional GPS is inadequate inside due to signal interference, requiring alternative systems such as beacon-based technologies and smartphone applications that employ digital maps (Theodorou et al., 2022; Kubota, 2024). Nonetheless, these systems frequently depend on pre-existing maps, which are not universally accessible, hence constraining their efficacy (Kubota, 2024).

The intricacy of interior environments exacerbates the problem, as visually impaired individuals encounter challenges in traversing unfamiliar settings due to ambiguous aural or tactile signals and various impediments (Jeamwathanachai et al., 2019; Fernando et al., 2023). Although systems such as Snap&Nav provide navigation solutions through the creation of node maps, their dependence on sighted aid diminishes user autonomy (Kubota, 2024). Moreover, the computing requirements of contemporary models provide difficulties. Advanced algorithms and machine learning techniques enhance obstacle identification and route planning but frequently necessitate substantial processing resources, rendering them impractical for mobile applications (Tao and Ganz, 2020; Shah et al., 2023). The integration of IoT and cloud computing introduces additional complexity, emphasizing the necessity for

lightweight, dependable systems designed for visually impaired users (Messaoudi et al., 2020). Rectifying these deficiencies is essential for the advancement of inclusive indoor navigation solutions.

Recent research has increasingly focused on developing advanced navigation aids to support visually impaired individuals in both indoor and outdoor environments. Technologies such as deep learning, machine vision, wearable devices, and mobile applications have been leveraged to enhance navigation capabilities, offering promising solutions to this pervasive issue (Bai et al., 2019; El-taher et al., 2021; Kuriakose et al., 2021; Martínez-Cruz et al., 2021).

The importance of developing an effective navigation assistance system for visually impaired individuals, particularly in obstacle-filled indoor environments, cannot be overstated. These environments present unique challenges that require sophisticated solutions capable of providing accurate and real-time guidance. The integration of advanced AI models, such as MobileNetV2 and MobileViTV2, along with multi-query attention mechanisms, has shown potential in creating robust and efficient navigation systems. These models aim to empower visually impaired individuals, allowing them to navigate unfamiliar spaces with confidence and independence (Wang et al., 2019).

The main research problem addressed in this study is the development of an independent navigation model that improves the accuracy and efficiency of detecting and interpreting navigation markers under extreme conditions for people with visual impairment. Traditional navigation systems often fall short in complex, obstacle-filled indoor environments, necessitating the need for a more advanced solution. This research proposes integrating MobileNetV2 and MobileViTV2 methods, enhanced by multi-query attention mechanisms, to develop a navigation model that provides precise and reliable assistance, thereby improving the quality of life for visually impaired individuals.

The integration of MobileNetV2 and MobileViTV2 methods represents a cutting-edge approach to developing an independent navigation model for the visually impaired. MobileNetV2, introduced by Sandler et al. (2018), is designed to operate efficiently on mobile and embedded devices, making it highly suitable for real-time applications. Its architecture employs inverted residuals and linear bottlenecks, which help maintain high accuracy while reducing computational demands. This efficiency is crucial for applications requiring portability and immediate response, such as navigation aids for visually impaired individuals.

On the other hand, MobileViTV2, as explored by Chen et al. (2021), utilizes vision transformers to enhance the model's capability to understand visual contexts. Vision transformers are adept at capturing long-range dependencies within images, providing a more comprehensive interpretation of complex scenes. The integration of these technologies, combined with multi-query attention mechanisms as highlighted by Mehta and Apple (2022), allows the model to focus on multiple aspects of the visual input simultaneously. This multifaceted attention mechanism is instrumental in improving the accuracy and timeliness of navigation instructions, thus providing a significant advancement over existing models.

Existing research on navigation aids for visually impaired individuals has explored a variety of technological solutions, ranging from GPS-based applications to wearable devices and computer vision techniques. Studies like those by Martínez-Cruz et al. (2021) and El-taher et al. (2021) have highlighted the limitations of GPS in indoor

environments and the bulkiness of wearable devices, respectively. These limitations underscore the need for more refined and user-friendly solutions. The integration of deep learning models, such as those using CNNs, has shown promise; however, these models often require substantial computational resources, limiting their practicality in mobile settings (Bai et al., 2019).

The recent development of efficient models like MobileNetV2 and vision transformers like MobileViTV2 addresses some of these challenges by offering high accuracy with reduced computational demands. However, a gap remains in the effective integration of these technologies to develop a comprehensive navigation system that is both lightweight and capable of real-time processing. Additionally, the potential benefits of multi-query attention mechanisms in enhancing the focus and accuracy of these models have not been fully explored. This gap presents an opportunity to develop a novel, integrated solution that leverages these advanced techniques for improved navigation assistance.

The aim of this research is to develop a new navigation model for visually impaired individuals by combining MobileNetV2 and MobileViTV2 with multi-query attention mechanisms. The hypothesis is that this AI model will improve both the accuracy and efficiency of Aruco marker detection under challenging conditions compared to existing models. This advancement is expected to offer reliable navigation assistance in indoor environments, enhancing the quality of life for visually impaired users. The study focuses on designing, implementing, and evaluating the model, with future work aimed at refining fusion mechanisms, reducing model complexity, and exploring transfer learning to maintain high accuracy while minimizing computational demands.

Following a brief introduction of the problem statement and the proposed method, the rest of the paper is structured as follows: Section 2 outlines the research methodology used to conduct the study. In Section 3, we present the search results obtained from the research. Section 4 discusses the findings related to multi-scale features and the various combinations of MQA and FFM used to improve model segmentation of multi-class ArUco markers. Finally, the conclusion of the paper is provided in Section 5.

2 Methods

2.1 Transformer and CNNs

CNNs have demonstrated exceptional performance in various image segmentation tasks, showcasing their robust feature representation capabilities. However, despite these strengths, CNN-based methods frequently encounter limitations in modeling long-range relationships. A primary issue is their inefficiency in capturing global context information. Methods that rely on stacking receptive fields necessitate continuous downsampling convolution operations, leading to deeper networks. Training such deep neural networks on small datasets can present significant challenges, including training instability and overfitting. Overfitting is particularly common in deep learning models due to their strong expressive ability relative to traditional models (Zhang et al., 2023). Non-local attention mechanisms have been increasingly utilized in various fields to address challenges related to capturing long-range dependencies and global information (Mei et al., 2020; Huang et al., 2022; Abozeid et al.,

2023; Zhou et al., 2023). While these mechanisms can enhance the network's ability to capture global context, they also introduce considerable computational complexity. This complexity, which is quadratic in relation to the input size, often renders these methods impractical for high-resolution images.

Attention mechanisms were utilized in numerous research that focused on integrating Convolutional Neural Networks. Especially to further enhance the output processing of CNNs. Various visual tasks were implemented with integrated approaches, including video processing (Qi and Zhang, 2023; Sun et al., 2022; Mujtaba et al., 2022), image classification (Dosovitskiy et al., 2021; Liu et al., 2021), and object detection (Benmouna et al., 2023; Wen et al., 2023).

The transformer in natural language processing used transformation tasks (Vaswani et al., 2023). Several natural language processing activities have since shifted to using it. Some natural language processing activities have switched to using ViT. Pre-training on very large datasets is required for ViT (Chen et al., 2023; Misawa et al., 2024). To the State of the Art in the natural image segmentation task, Imagenet replaced the encoder component of the decoding network with a transformer (Doppalapudi, 2023; Xia and Kim, 2023).

Although transformer-based models have demonstrated impressive skills in diverse visual tasks, they have not yet attained acceptable results when compared to traditional CNNs. Transformer designs still exhibit worse performance in visual tasks compared to similarly-sized CNNs, such as EfficientNet (Thakur et al., 2023). The computational cost of transformers based on the mechanism of self-attention is $O(N^2C)$, in contrast to the convolution-based CNNs ($O(NC^2)$) (Zhou et al., 2024). Therefore, employing the transformer for image-related activities will unavoidably need a substantial amount of GPU resources.

2.2 Image segmentation using transformer and CNNs

The present cutting-edge architecture in computer vision predominantly depends on complete CNNs, with UNet (Chen et al., 2021) and its variations being notable instances. The current state-of-the-art (SOTA) framework in computer vision primarily relies on full CNNs, with UNet and its variants being prominent examples. UNet (Chen et al., 2021) employs an encoding-decoding network architecture. This architecture utilizes cascaded convolutional layers to extract various levels of visual characteristics. The decoder utilizes skip connections to recycle high-resolution feature maps generated by the encoder, enabling the retrieval of crucial feature information (Petit et al., 2021).

2.3 Lightweight networks

Deep learning, although powerful, often requires extensive training data to effectively enhance model learning. However, challenges arise in scenarios like the ArUco dataset due to limitations in data collection related to factors such as lighting conditions, capture angles, and distances. Moreover, the availability of large, publicly accessible datasets is limited, further complicating model training (Lee et al., 2019). To address these challenges, the development of lightweight deep learning models becomes imperative.

Research in deep learning has demonstrated that supervised training of deep learning models heavily relies on large labeled datasets (Karimi et al., 2020). This requirement poses a significant challenge, especially in scenarios where data collection is constrained. Techniques such as model optimization, pruning, quantization, and knowledge distillation have been explored to create lightweight deep-learning models suitable for mobile terminals (Wang et al., 2022). These approaches aim to reduce the computational burden while maintaining model performance.

A self-attention-based vision transformer (ViT), known as MobileViT, is employed to learn the global representation of images. MobileViT (Mehta and Apple, 2022) stands out as the initial lightweight, general-purpose transformer designed for mobile devices. An approach integrating a transformer with a CNN-based lightweight model was investigated, with a particular focus on assessing the feasibility of this lightweight network model for the challenging task of ArUco marker segmentation.

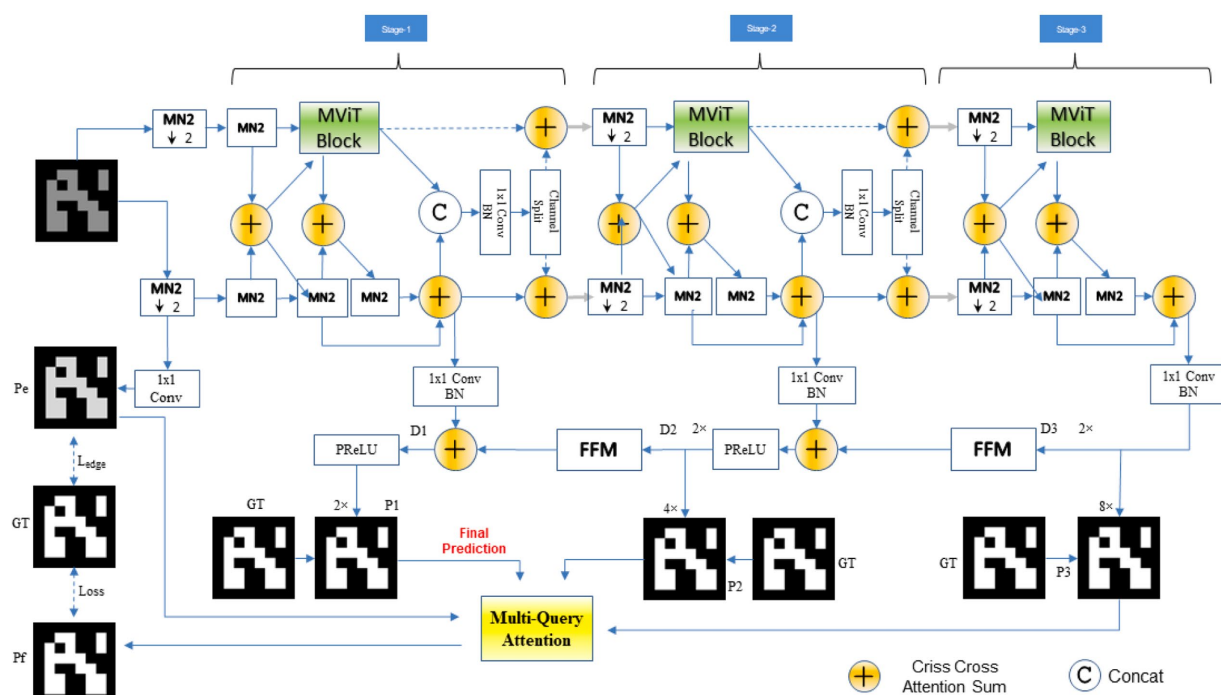
2.4 Mini-MobileNet-MobileViT network

In this part, the Mini-MobileNet-MobileViT (MoNetViT) network architecture and its principal network components were introduced. The system's backbone structure follows to the architecture of an encoder and decoder, as represented in Figure 1A. Section 2.4.1 offers a more detailed explanation sub-network of the encoder, while Section 2.4.2 focuses on the exploration sub-network of the decoder. The MobileViT module, a crucial part of the network's encoder architecture, is introduced in Section 2.4.3. This section covers the architecture of the MobileViT module, its primary calculation process internally, and the comparisons between this module and CNN. Furthermore, the MQA module that is suggested in this study is described in Section 2.4.4. The Globalized Block and the Asymmetrical Globalized Block, as well as the justification for their adoption, are part of this module.

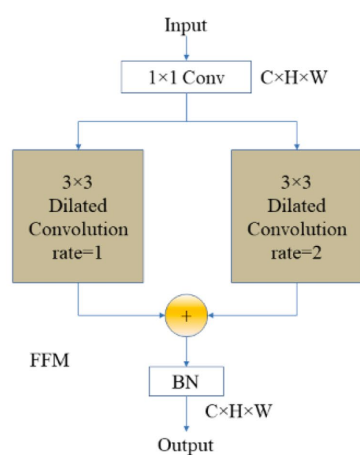
2.4.1 Encoder sub-network

The proposed model will be developed using an encoder-decoder structure, where the encoder will build two parallel paths connected by a series of attention additions, improving the model's ability to capture spatial and channel dependencies. The encoder will use MobileNet v2 (MN2 block) (Sandler et al., 2018) and MobileViT block as the base module. $I \in \mathbb{R}^{3 \times H \times W}$, is the representation of the input image, where H and W stand for the input image's height and width, respectively. The input image undergoes resolution degradation through three consecutive stages. In each stage, the size of the feature map is reduced by a factor of 2. As a result, the output feature maps are reduced in size to one-half, one-fourth, and one-eighth of the initial feature map. The MobileViT block is one of the essential components used in the encoder. The input and output sizes of the MobileViT block are the same, indicating that this module does not change the spatial dimensions of the feature map. The MN2 block is another basic module used in the encoder. Stride 1 implies that the module does not perform resolution degradation, and the input and output sizes remain the same.

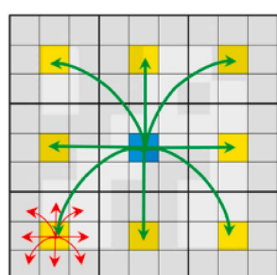
At the i th stage in Equation 1, it is assumed that $M_j^i(\cdot)$ represents the transformation function of the j th MV2-Block. For example, M_4^i denotes the result produced by the 4th MN2-Block in the i th stage.



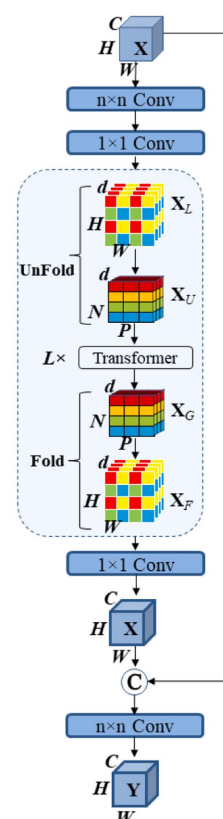
(a)



(b)



(c)



(d)

FIGURE 1

MoNetViT Main Architectural diagram and the essential network components. (A) MoNetViT deep CNN encoder and a few basic modules (B) Feature Fusion Modul (FFM) (C) Illustration calculation between pixels in MobileViT (D) MobileViT-Block.

The MobileViT-block module at the i th stage has a transformation function denoted as $V^i(\cdot)$. It is important to emphasize that there is a singular MobileViT-block present at each stage.

Moreover, if we represent the output generated by the j th MN2-Block module during the i th stage as (M_j^i) , it is important to highlight that just the first two MN2-Block components reduce the resolution of the original feature map. Thus, M_j^i belongs to the set of elements in $\mathbb{R}^{C_i \times \frac{H}{2^i} \times \frac{W}{2^i}}$, where i belong to the collection $\{1,2,3\}$, and j is an element of the set $\{1,2,3,4\}$. C_i represents the numerical value assigned to the feature channel at the i th stage.

$$\begin{aligned} V^i &= V^i \left(M_1^i + \left(CAM \left(M_2^i + M_1^i \right) \right) \right) \\ M_3^i &= M_3^i \left(M_1^i + \left(CAM \left(M_2^i + M_1^i \right) \right) \right) \\ M_4^i &= M_4^i \left(CAM \left(V^i + M_1^i \right) \right) \end{aligned} \quad (1)$$

The channel attention module is denoted by $CAM(\cdot)$. It is reasonable to assume that at this stage i , the output of the left path is L^i and the output of the right path is R^i . The formulas Equation 2 can be used to compute R^i and L^i :

$$\begin{aligned} L^i &= \text{Split} \left(F^{1 \times 1} \left(\text{Concat} \left(V^i, CAM \left(M_3^i + M_4^i \right) \right) \right) \right) \\ R^i &= \text{Split} \left(F^{1 \times 1} \left(\text{Concat} \left(V^i, CAM \left(M_3^i + M_4^i \right) \right) \right) \right) \end{aligned} \quad (2)$$

Here, $F^{1 \times 1}$ represents a 1×1 convolution operation, V^i represents the feature map, and M_3^i and M_4^i represent the outputs of the third and fourth MN2-Block modules at stage i , respectively. The *Concat* function concatenates the feature map V^i with the output of the channel attention module applied to the sum of M_3^i and M_4^i . The *Split* function splits the resulting tensor into multiple parts.

2.4.2 Decoder sub-network

In the context of a decoder sub-network shown in Figure 1B, the function $D_i(\cdot)$ represents the operation of the Feature Fusion Module (FFM) like figured in Equation 3. The module takes input I and processes it through a series of transformations involving convolutional operations and batch normalization. The formula provided is as follows:

$$D_i(I) = \text{BatchNorm}(\hat{F}^{3 \times 3} \left(F^{1 \times 1}(I) \right) + \hat{F}_2^{3 \times 3} \left(F^{1 \times 1}(I) \right)) \quad (3)$$

$\hat{F}^{3 \times 3}$ represents a divided convolution functional with a kernel size of 3×3 and an increase rate of 1, which is equal to a conventional convolution. $\hat{F}_2^{3 \times 3}$ represents a dilated convolution operation with a kernel size of 3×3 and an expansion rate of 2. This indicates a convolution with a dilation process using a 3×3 kernel area and an expansion rate that is 1, equivalent to a conventional. BatchNorm refers to the batch normalization operation that standardizes the inputs to a layer for each mini-batch.

In the decoder stage of the network, the feature maps are represented as $D_i \in \mathbb{R}^{C_i \times \frac{H}{2^i} \times \frac{W}{2^i}}$, where $i \in \{1,2,3\}$. After the encoder phase, the operation on the feature map D_3 (at the third stage of the decoder) is defined $D_3 = \text{BatchNorm} \left(F^{1 \times 1} \left(M_3^3 + M_3^4 \right) \right)$. The two main steps for calculating D_i (for $i = 2,1$) as described. This process entails increasing the resolution of the data and then combining it with the results from the earlier stage of the encoding process. The initial step involves performing up-sampling and Feature Fusion Mapping (FFM). Up-sampling is intended to adjust the feature size to match the output size of the encoder from the previous stage, as illustrated in Equation 4. This process yields an intermediate variable, denoted as S_i . The subsequent step entails a feature fusion operation with the encoder's output from the prior stage, as demonstrated in Equation 5:

$$S_i = D_i \left(\text{Upsample} \left(CAM \left(D_{i+1} \right), 2 \right) \right) \quad (4)$$

$$D_i = \text{PReLU} \left(S_i + \text{BatchNorm} \left(F^{1 \times 1} \left(M_3^i + M_4^i \right) \right) \right) \quad (5)$$

Upsample($\cdot$, t) denotes the procedure of augmenting the data map in accordance with the parameter t by bilinear interpolation. The PReLU function of activation is denoted as PReLU($\cdot$), whereas batch normalization is indicated as BatchNorm($\cdot$). Once all D_i ($i = 1,2,3$) are calculated, a final prediction is obtained through a 1×1 convolution.

$$P_i = \text{Softmax} \left(\text{Upsample} \left(F^{1 \times 1} \left(D_i \right), 2^i \right) \right) \quad (6)$$

Where i is an element of the set $\{1,2,3\}$, Softmax($\cdot$) denotes the function that activates softmax. $P_i \in \mathbb{R}^{H \times W}$ denotes the predicted class label map, with P_i being the final output prediction, as demonstrated in Equation 6.

2.4.3 MobileViT block

Vision Transformers (ViTs) can achieve comparable accuracy to Convolutional Neural Networks (CNNs) in image identification tasks, especially when trained on extensive datasets (Dosovitskiy, 2021). On the other hand, unlike CNNs, ViTs are difficult to optimize and require a large amount of data for training. Research indicates that the suboptimal performance of ViTs is due to a lack of inductive biase (Lee et al., 2019; Petit et al., 2021; Zhou et al., 2024). Inductive biases, while beneficial, also have drawbacks for CNNs; they enable CNNs to capture local spatial information but can limit the network's overall performance.

However, the transformer's self-attention system has the capacity to collect global data. Numerous transformers and CNNs combinations have been investigated to overcome their respective deficiencies. ConViT (d'Ascoli et al., 2022) uses gated positional self-attention soft convolutional inductive biases. Semantic segmentation models such as ACNET (Hu et al., 2019) and CMANet (Zhu et al., 2022) have been developed; however, many of these models are computationally intensive. The possibility of leveraging the strengths of both CNNs and ViTs to construct a lightweight network for visual tasks remains an area of ongoing exploration. MobileViT suggests that

such an approach is indeed feasible. In this paper, we first examine the calculations involved in MobileViT.

The MobileViT Block, seen in Figure 1D, shares an identical structure with the MobileViT Unit (Mehta and Apple, 2022). The following four phases are applied to a given source tensor $X \in \mathbb{R}^{H \times W \times C}$:

- The input tensor X is first passed through an $N \times N$ standard convolution layer, followed by a 1×1 convolution layer to generate $X_L \in \mathbb{R}^{H \times W \times d}$. The $N \times N$ convolution stage captures and represents nearby spatial details, whereas the 1×1 convolution transforms the tensors into higher-dimensional spaces (with d dimensions, where d is more than c) by acquiring knowledge of a linear combination of the input channels.
- In order to incorporate spatial inductive bias into MobileViT's learning process, the input X_L is divided into N non-overlapping flattening patches $X_U \in \mathbb{R}^{(P \times N \times d)}$, where $P = hw$. The total number of patches is represented by the formula $N = \frac{HW}{P}$, where h and w are the physical dimensions of every single patch, individually.
- The transformer is then applied to encode the relationships between the patches through the following operation, as demonstrated in Equation 7

$$X_G(p) = \text{Transformer}(X_U(p)), 1 \leq p \leq P \quad (7)$$

- The resulting $X_G \in \mathbb{R}^{P \times N \times d}$ is then folded back to obtain $X_F \in \mathbb{R}^{H \times W \times d}$.
- Ultimately, X_F is transformed into a space with fewer dimensions (C dimensions) using point-by-point convolution and then merging with X using concatenation.

The second phase contains the algorithm's core. The input image X , which has dimensions $H \times W \times C$, is separated into patches in the standard ViT structure. Subsequently, every patch undergoes a linear transformation to convert it into a vector. These vectors are then encoded with positional information. Furthermore, the interconnections between the patches are acquired by employing N transformer blocks.

Contrary to ViT, the MobileViT algorithm preserves both the patch order and the physical order of pixels inside each of the patches during its second stage. It is crucial to emphasize that the values of w and h must be exact divisors of W and H , respectively.

Local information can be encoded by the relationship X_U . The yellow pixels inside a patch have the ability to aggregate data from the pixels that surround them in that patch, as seen in Figure 1C. X_G accomplishes the worldwide data encoding of the transformer by encoding inter-patch connections at the p -th place of every patch. The red pixel keeps track of every one of the pixels that encode the full image since, as Figure 1C illustrates, it can identify the yellow pixel that is at the same location in other patches. The lightweight aspect of the model is enhanced by the dot product procedure, which selects only pixels that are in the same position.

According to Mehta and Apple (2022), ordinary convolution can be broken down into three steps: unfolding, matrix multiplication, and

folding. Based on the previously indicated computation, a layer of convolution and the Unfold operation carry out the local feature modeling, providing them with convolution-like inductive biases. Next, global feature modeling is carried out using the Transformer $\rightarrow$ Fold sequence, which gives the MobileViT block global processing power.

2.5 Multi query attention

Figure 2A displays the configuration of a standard non-local block (Wang et al., 2018). The non-local block (Wang et al., 2018) first requires the computation of the similarity between all places. This is accomplished by performing matrix multiplication on an input $X \in \mathbb{R}^{C \times H \times W}$. The primary computational procedure in the non-local block can be succinctly described as consisting of the following five steps:

- The source feature X is subjected to three 1×1 convolutions, denoted as W_ϕ , W_θ , and W_γ , resulting in the transformation of X into $\phi \in \mathbb{R}^{\hat{C} \times H \times W}$, $\theta \in \mathbb{R}^{\hat{C} \times H \times W}$, and $\gamma \in \mathbb{R}^{\hat{C} \times H \times W}$. The three numbers correspond to the query, key, and value, respectively. They are used to change the total amount of streams from C to $\hat{C}$.
- A similarity matrix M is created by flattening the query, key, and value to size $\hat{C} \times N$, where $N = HW$. The matrix M is calculated to determine the similarity, as demonstrated in Equation 8:

$$M = \phi^T \times \theta, M \in \mathbb{R}^{N \times N} \quad (8)$$

- The matrix M is normalized using a normalization function such as softmax: $\bar{M} = \text{softmax}(M)$.
- The matrix of attention A is then derived, as demonstrated in Equation 9:

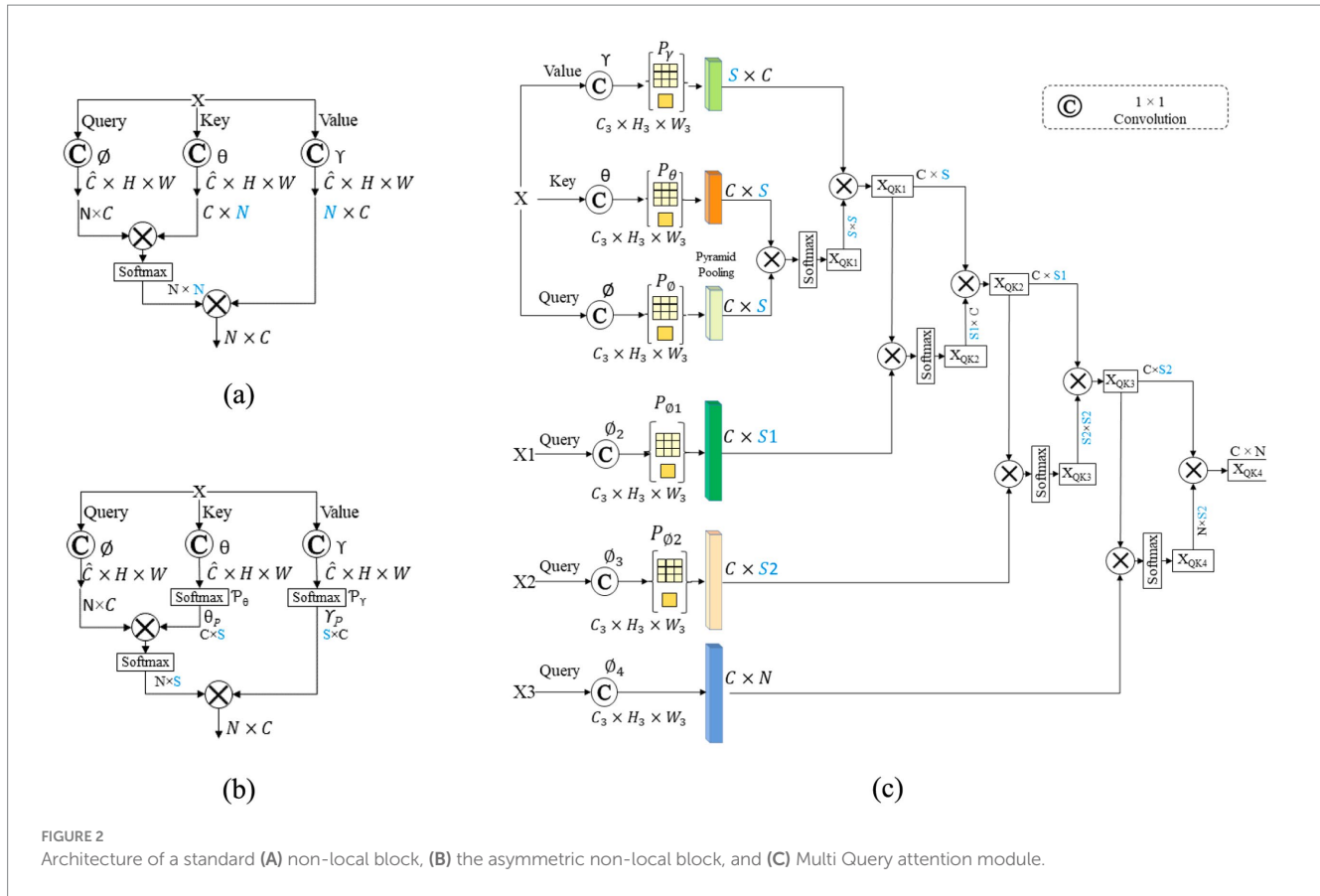
$$A = \bar{M} \times \gamma^T, A \in \mathbb{R}^{N \times \hat{C}} \quad (9)$$

- The final result is computed as: $Y = W_A(A^T) + X$. The channel dimension is adjusted from $\hat{C}$ back to C by a 1×1 convolution, denoted as W_A .

Reassessing the Non-local Asymmetric Block

The computational complexity of the global attention block can be described as $O(\hat{C}N^2) = O(\hat{C}H^2W^2)$. The calculation efficiency is mostly affected by N 's size. To fix this, reduce N to $S(S \ll N)$ without altering output size. Zhu et al. (2019) offers the asymmetric non-local block, whose construction is shown in Figure 2B, to tackle this problem.

The asymmetrical pyramid non-local block (APNB) is a modified version of this block that incorporates pyramid pooling within the non-local block in order to decrease computational expenses. One more thing is added after θ and γ : a spatial pyramid pooling function (Lazebnik et al., 2006) to pick out a few good anchor points. When the spatial pyramid pooling modules are P_n^θ



and P_n' , with n denoting the pooling layer's output size (width or height) and $n \in \{1, 3, 6, 8\}$ as per [Zhu et al. \(2019\)](#), the overall amount of sampling anchor points is $S = \sum_{n \in \{1, 3, 6, 8\}} n^2 = 110$. If we assume

that $H = 224$ and $W = 256$, the number of calculations will be reduced by an amount $\frac{H \times W}{S} = \frac{256 \times 256}{110} \approx 595$. This adjustment efficiently decreases the value of N to a lower value, S , by selectively sampling a few sample data from θ and γ , instead of utilizing all of the points.

The primary computational procedure in the asymmetrical non-local block entails the subsequent modifications, as demonstrated in [Equations 10, 11](#):

$$\underbrace{\mathbb{R}^{N \times \hat{C}} \times \mathbb{R}^{\hat{C} \times N}}_{Eq.(8)} \rightarrow \underbrace{\mathbb{R}^{N \times N} \times \mathbb{R}^{N \times \hat{C}}}_{Eq.(9)} \rightarrow \mathbb{R}^{N \times \hat{C}}. \quad (10)$$

$$\underbrace{\mathbb{R}^{N \times \hat{C}} \times \mathbb{R}^{\hat{C} \times S}}_{Eq.(12)} \rightarrow \underbrace{\mathbb{R}^{N \times S} \times \mathbb{R}^{S \times \hat{C}}}_{Eq.(13)} \rightarrow \mathbb{R}^{N \times \hat{C}}. \quad (11)$$

The computational technique for the APNB module may be delineated as follows:

Introduce sampling modules P_θ and P_γ after θ and γ , accordingly, to sample multiple sparse anchor points. These anchor points are designated as $\theta_P \in \mathbb{R}^{\hat{C} \times S}$ and $\gamma_P \in \mathbb{R}^{\hat{C} \times S}$, respectively.

Generating a similarity matrix M_P , as demonstrated in [Equation 12](#):

$$M_P = \phi^T \times \theta_P, M_P \in \mathbb{R}^{N \times S} \quad (12)$$

M_P is normalized. $\tilde{M}_P = f(M_P)$

$f(\cdot)$ stands for normalization function.

The attention matrix AP is subsequently computed, as demonstrated in [Equation 13](#):

$$AP = \tilde{M}_P \times \gamma_P^T, AP \in \mathbb{R}^{N \times \hat{C}} \quad (13)$$

The final result is obtained by adding the product of the variables W_A and A_T to the variable X , and assigning it to the variable Y_P . W_A is a one-by-one convolution. $Y_P = W_A(A_P^T) + X.W_A$ represents a 1×1 convolutions.

2.6 Motivation

The APNB discussed earlier operates with a single data stream as input, while the asymmetrical fusion nonlocal block typically utilizes two data sources: the top-level feature map and the lower-level feature map. In contrast, the proposed MQA mechanism extends this approach by incorporating four input sources. As shown in [Figure 1A](#), MQA integrates the fundamental value ($P1$), the elevated characteristics $P2$ and $P3$, and the edge content (Pe), allowing for the explicit acquisition of multiple levels of feature representation. By

incorporating edge information throughout the semantic segmentation process, the model imposes valuable constraints, enhancing segmentation precision. The cross-entropy loss function further refines this process by measuring the difference between the ground truth (GT) and the feature aggregation map (Pf), ensuring robust alignment of predictions with the actual data.

Commence the acquisition of a parameter map of features $X \in \mathbb{R}^{C \times H \times W}$, and additional feature maps $X1 \in \mathbb{R}^{C_1 \times H_1 \times W_1}$, $X2 \in \mathbb{R}^{C_2 \times H_2 \times W_2}$, and $X3 \in \mathbb{R}^{C_3 \times H_3 \times W_3}$. Five 1×1 convolutions, denoted as $W_\phi, W_\theta, W_\gamma, W_{\phi_2}$, and W_{ϕ_3} , are applied to transform these input maps into new feature maps: $\phi \in \mathbb{R}^{C \times H \times W}$, $\gamma \in \mathbb{R}^{C \times H \times W}$, $\phi1 \in \mathbb{R}^{C \times H1 \times W1}$, $\phi2 \in \mathbb{R}^{C \times H2 \times W2}$, and $\phi3 \in \mathbb{R}^{C \times H3 \times W3}$, as demonstrated in Equation 14:

$$\begin{aligned}\phi &= W_\phi(X), \theta = W_\theta(X), \gamma = W_\gamma(X), \\ \phi1 &= W_{\phi1}(X1), \phi2 = W_{\phi2}(X2), \phi3 = W_{\phi3}(X3)\end{aligned}\quad (14)$$

The parameters X , $X1$, $X2$, and $X3$ in this experiment correspond to the results P_e , P_l , P_2 , and $P3$ from MoNetViT. The sample as well as main computation methodologies within the MQA module were equivalent to those in APNB. The selection units $P_\theta, P_\gamma, P_\phi, P_{\phi1}$, and $P_{\phi2}$ are utilized to sample multiple sparse anchor points. These anchor points are represented as $\theta_P \in \mathbb{R}^{C \times S}$, $\gamma_P \in \mathbb{R}^{C \times S}$, $\phi_P \in \mathbb{R}^{C \times S}$, $\phi1_P \in \mathbb{R}^{C \times S1}$ and $\phi2_P \in \mathbb{R}^{C \times S2}$, where S , $S1$, and $S2$ denote the number of sampled anchor points. S is less than $S1$, which is less than $S2$, and $S2$ is much less than N . Mathematically, this is computed using the following Equation 15:

$$\begin{aligned}K &= P_\theta(\theta), V = P_\gamma(\gamma), Q = P_\phi(\phi) \\ Q1 &= P_{\phi1}(\phi1), Q2 = P_{\phi2}(\phi2)\end{aligned}\quad (15)$$

The correlation matrix M_P for Q and anchoring K is displayed here, as demonstrated in Equation 16:

$$M_P = Q^T \times K \quad (16)$$

The dimensions of the M_P are $S \times S$, where S is significantly smaller than N . Next, the process of normalization is carried out on M_P , which enables the calculation of X_{QK1} .

The ultimate result of the initial layer is, as demonstrated in Equation 17:

$$X_{QKV1} = X_{QK1} \times V, X_{QKV1} \in \mathbb{R}^{C \times S} \quad (17)$$

The value of X_{QKV1} is equal to the product of X_{QK1} and V . X_{QKV1} belongs to the set of $\mathbb{R}^{C \times S}$. The similarity matrix for levels 2, 3, and 4 is computed using an analogy, as demonstrated in Equation 18:

$$X_{QK2} = Q_1^T \times X_{QKV1}$$

$$X_{QK3} = Q_2^T \times X_{QKV2}$$

$$X_{QK4} = Q_3^T \times X_{QKV3} \quad (18)$$

The equation X_{QK2} is equal to the product of Q_1^T and X_{QKV1} . The equation X_{QK3} is equal to the product of Q_2^T and X_{QKV2} . The equation X_{QK4} is equal to the product of Q_3^T and X_{QKV3} .

The ultimate result of levels 2, 3, and 4 is determined in the following manner Equation 19:

$$\begin{aligned}X_{QKV2} &= X_{QK2} \times X_{QKV1}, X_{QKV2} \in \mathbb{R}^{C \times S1} \\ X_{QKV3} &= X_{QK3} \times X_{QKV2}, X_{QKV3} \in \mathbb{R}^{C \times S2} \\ X_{QKV4} &= X_{QK4} \times X_{QKV3}, X_{QKV4} \in \mathbb{R}^{C \times N}\end{aligned}\quad (19)$$

The value of X_{QKV2} is equal to the product of X_{QK2} and X_{QKV1} , where X_{QKV2} belongs to the set $\mathbb{R}^{C \times S1}$. The value of X_{QKV3} is equal to the product of X_{QK3} and X_{QKV2} , where X_{QKV3} belongs to the set $\mathbb{R}^{C \times S2}$. The value of X_{QKV4} is equal to the product of X_{QK4} and X_{QKV3} , where X_{QKV4} belongs to the set $\mathbb{R}^{C \times N}$.

The final result is represented as X_{QKV4} belonging to the set of $\mathbb{R}^{C \times N}$. The temporal complexity can be represented as $O(C(S \times S + S1 \times S + S2 \times S1 + S3 \times S2))$, which is significantly lower than $O(CN^2)$ in the conventional non-local block.

2.7 Loss function

The function that measures loss is defined like the ones used by Yeung et al. (2022). The loss function comprises two components: L_{edge} and L_{seg} . Equations 20, 21 (Yeung et al., 2022) display the loss function.

$$L_{edge} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (20)$$

where ground-truth (GT) and the anticipated edge map P_e 's coordinates for each pixel point are represented by $(y_i, \hat{y}_i)$.

IoU loss and a conventional cross-entropy loss make up the two components of the L_{seg} loss function.

$$L_{seg} = \lambda_1 L_{CE}^w + \lambda_2 L_{IoU}^w \quad (21)$$

2.8 Experiment setup

The NVIDIA Tesla T4 GPU was used to train the model within the PyTorch framework for this project. The model underwent training for 50 epochs using a batch size of 16, with the Adam optimizer and a starting learning rate of 1e-3. A learning rate

TABLE 1 Network hyperparameter.

Hyperparameter	Options
Resize the images	224×224
Epochs	50
Batch size	16
Optimizer	Adam
Learning rate (Lr)	1e – 3
Factor	0.5
Patience	5
λ_1, λ_2	0.2, 0.8

reduction factor of 0.5 was applied when no improvement was observed for 5 consecutive epochs (patience set to 5). The specific hyperparameters utilized in this investigation are outlined in Table 1. In addition, the experiment aimed to compare the proposed MoNetViT model with several state-of-the-art methods, including TransFuse (Zhang et al., 2021), Inf-Net (Fan et al., 2020), U-Net (Ronneberger et al., 2015), U-Net++ (Kwak and Sung, 2021), Mini-Seg (Kim et al., 2023), and DeepLabV3+ (Asadi Shamsabadi et al., 2022), for multi-class segmentation in ArUco marker identification.

The researchers conducted experiments using freely available datasets for ArUco manual labeling. The dataset employed in this study is an open-source resource that has been fully labeled to indicate various classes of ArUco markers, making it ideal for training models in identification and classification tasks. The dataset preparation involved several pre-processing steps to enhance model generalizability and ensure reproducibility. Images were resized to a resolution of 224×224 , and pixel values were normalized to the range 0 and 1. Additionally, data augmentation techniques such as random rotation, flipping, and contrast adjustment were applied to improve the model's ability to generalize across varied conditions. The complete dataset, including all necessary images and labels for training and testing models in ArUco marker identification and classification tasks, can be accessed and downloaded from the following link: <https://universe.roboflow.com/loliktry/dataarucomustofa/dataset/5>. It was specifically utilized in the multi-class segmentation experiments. Table 2 provides a comprehensive overview of the specific details of this dataset.

Furthermore, the present study utilized identical methodologies as Fan et al. (2020) to assess the performance of the model. Standard criteria, such as accuracy, specificity, sensitivity, and Dice similarity coefficient, comprise the assessment metrics. In addition, it uses several metrics from object recognition evaluation methods, such as the design measure, the enhanced alignment value (Fan et al., 2018), and the mean absolute error.

3 Results

The dataset utilized in this study is notably large-scale, comprising a total of 4,144 slices. Consequently, the following sections will focus exclusively on a detailed analysis of the results, offering an in-depth examination of the findings and their implications.

TABLE 2 Specifications of datasets.

Marker class	No. training	No. valid.	No. testing	Total images
1	853	188	289	1,330
2	904	238	269	1,411
3	895	237	271	1,403
Total	2,652	663	829	4,144

3.1 Three-class ArUco marker labeling results

The segmentation data findings for ArUco Marker on the dataset are displayed in Figure 3, demonstrating that MoNetViT in this study exhibits superior performance compared to other baseline models. U-Net and U-Net++ have low Dice scores and sensitivities, resulting in large unsegmented areas. Inf-Net and Mini-Seg show slight improvements but still lack accurate boundary detection. While TransFuse, a CNN + Transformer model, was evaluated, it is not included in Figure 3 due to its very low Dice score. DeepLabV3+ performs reasonably well but falls short of MoNetViT. Overall, MoNetViT achieves the highest Dice, sensitivity, and specificity scores, along with the lowest MAE, indicating its superior segmentation precision.

The MoNetViT model shows superior performance over other state-of-the-art models, including U-Net, U-Net++, Mini-Seg, Inf-Net, TransFuse, and DeepLabV3+, across key evaluation metrics. As summarized in the Table 3, MoNetViT achieves the highest Dice score (0.9584), sensitivity (0.9424), specificity (0.9424), structural similarity S_α (0.9923), and mean edge accuracy E_θ^{mean} (0.9381), while also attaining the lowest Mean Absolute Error (MAE) of 0.0077.

This outstanding performance is attributed to the integration of Convolutional Neural Networks (CNNs) with a transformer module, which captures both local and global semantic features. The transformer component is crucial for calculating global semantic relationships, while the CNN module extracts local contextual features, resulting in a more robust feature representation. Additionally, the multi-query attention (MQA) module enriches feature diversity through supervised learning, enhancing overall model performance.

An analysis of false positives (FP) and false negatives (FN) further emphasizes MoNetViT's robustness. With high sensitivity (0.9424) and specificity (0.9424), MoNetViT significantly reduces both FP and FN compared to other models. For instance, while DeepLabV3+ achieves a Dice score of 0.6351, its sensitivity (0.6392) and specificity (0.9655) indicate a higher FN rate relative to MoNetViT. Similarly, Mini-Seg's balanced sensitivity (0.6268) and specificity (0.9537) suggest that it is more prone to FP and FN, impacting its reliability. In contrast, MoNetViT's ability to minimize FP and FN contributes to its high Dice score and overall segmentation accuracy.

Despite Inf-Net having the smallest model size (0.073 MB) and TransFuse having the smallest parameter count (0.038 M), MoNetViT outperforms these models in critical metrics. Its effectiveness stems from model design rather than sheer training data volume, highlighting MoNetViT's robustness and capacity to generalize effectively on the dataset. Paired *T*-tests indicated that MoNetViT's improvements were statistically significant ($p < 0.05$) compared to all models except U-Net++, Mini-Seg, and DeepLabV3+, suggesting that while MoNetViT is generally superior,



FIGURE 3 Comparison of segmentation results of three-class labeling.

TABLE 3 Result of three-class labeling.

Methods	Param.(M)	Size(Mb)	Dice	Sen.	Spec.	S_{α}	F_{θ}^{mean}	MAE
U-Net	1.953	7.438	0.5622	0.6469	0.9803	0.9676	0.5622	0.0344
U-Net++	7.783	29.69	0.6082	0.5744	0.9013	0.9816	0.6082	0.0351
Inf-Net	0.076	0.291	0.5889	0.5585	0.8865	0.9762	0.5889	0.0458
Mini-Seg	0.038	0.145	0.6238	0.6268	0.9537	0.9845	0.6238	0.0292
TransFuse	0.019	0.073	0.3231	0.3333	0.6667	0.9403	0.3231	0.1177
DeepLabV3+	13.324	50.508	0.6351	0.6392	0.9655	0.9881	0.6351	0.0221
MoNetViT(ours)	1.014	3.869	0.9584	0.9424	0.9424	0.9923	0.9381	0.0077

Bold value indicates the best value.

these models are still competitive in certain tasks. The addition of FP and FN analysis strengthens these findings, demonstrating that MoNetViT achieves its exceptional performance by addressing key limitations in segmentation errors observed in other models.

3.2 Component impact analysis

Several experiments were conducted to validate the functionality of the Multi Query Attention (MQA) and Fusion Feature Module (FFM), two crucial elements of the MoNetViT. Figure 1A depicts an architecture consisting of three stages. The segmentation performance of the MoNetViT model is considerably improved by the MQA module component, as the findings displayed in Table 4 reveal. Employing MQA and FFM in conjunction with the baseline improves segmentation performance. Specifically, integrating both MQA and FFM with the baseline resulted in improvements of 1.5 and 1.7% in the Dice coefficient, respectively. The findings indicate that using MQA and FFM enhances the encoder’s and decoder’s clarity, thereby further improving segmentation performance. The analysis shows that combining Backbone, FFM, and MQA results in the best performance across all metrics. This model has the lowest loss (0.0115) and the highest Dice coefficient (0.9731), indicating more

accurate segmentation and fewer errors. It also achieves the best sensitivity (0.9686) and specificity (0.9686), accurately identifying both positive and negative cases. Additionally, it has the highest structural alignment (S_{α} = 0.9951) and precision (F_{θ}^{mean} = 0.9599), along with the lowest Mean Absolute Error (0.0049). In comparison, other methods like Backbone alone, Backbone+FFM, and Backbone+MQA perform worse in various metrics, highlighting the benefits of using both FFM and MQA together.

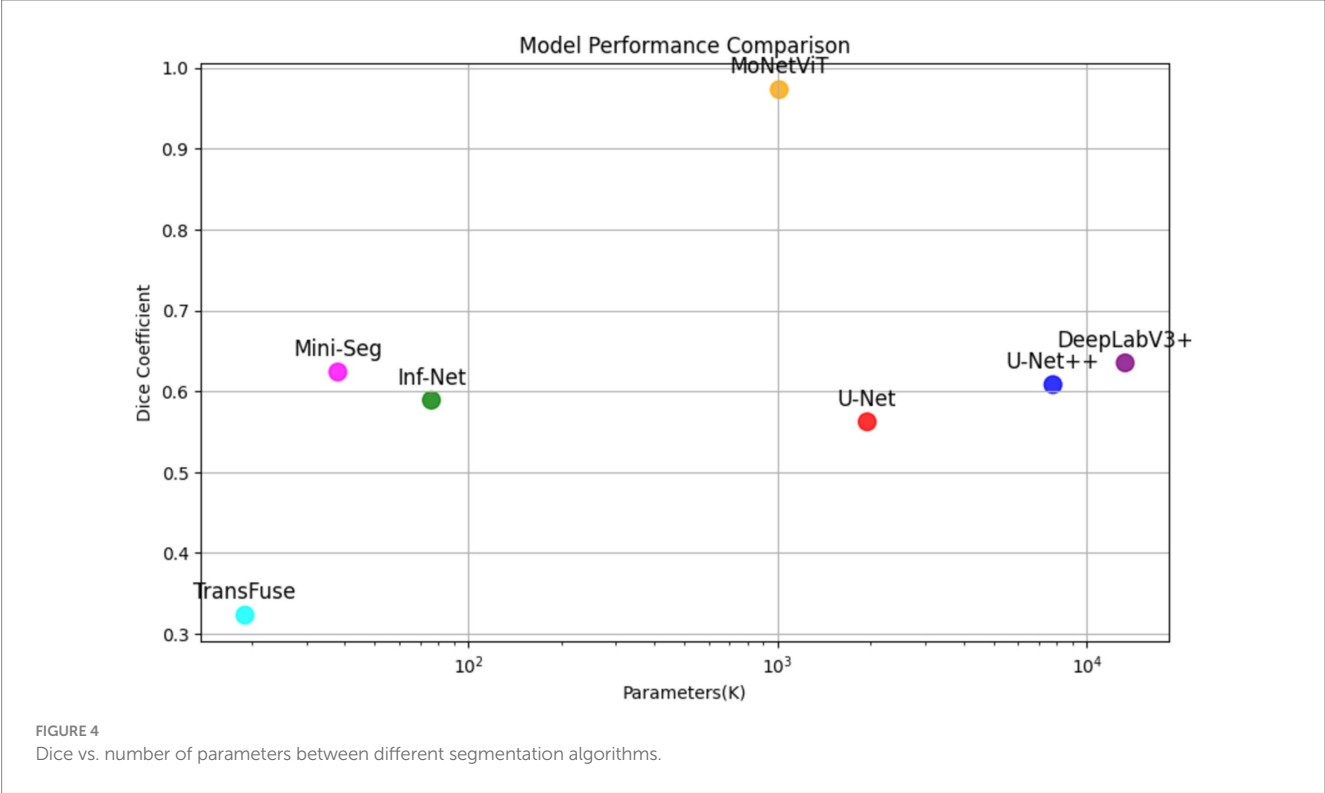
3.3 Parameter comparison

Figure 4 provides a detailed overview of the model performance. MoNetViT achieves superior accuracy while maintaining a relatively small parameter count. Specifically, it uses about half the parameters of U-Net (1.014 M vs. 1.953 M) and significantly fewer parameters than U-Net++ (7.783 M) and DeepLabV3+ (13.324 M). While Inf-Net (0.076 M) and Mini-Seg (0.038 M) have smaller parameter counts, and TransFuse uses the fewest (0.019 M), MoNetViT achieves a much higher Dice score (0.9584) compared to these models. This demonstrates that MoNetViT not only optimizes model size and complexity but also outperforms other models, including transformer-based models like TransFuse, in terms of accuracy (Figure 5).

TABLE 4 MoNetViT ablation study.

Methods	Loss	Dice	Sen.	Spec.	S_{α}	E_{θ}^{mean}	MAE
Backbone	0.0183	0.9584	0.9424	0.9424	0.9923	0.9381	0.0077
Backbone+MQA	0.0192	0.9558	0.9360	0.9360	0.9919	0.9342	0.0081
Backbone+FFM	0.0189	0.9566	0.9322	0.9322	0.9921	0.9354	0.0079
Backbone+MQA + FFM	0.0115	0.9731	0.9686	0.9686	0.9951	0.9599	0.0049

Bold value indicates the best value.



4 Discussion

4.1 Comparison of multi-scale features

Numerous models for combining multi-scale features use conventional networks for object identification and semantic segmentation in image processing. For example, architectures like Feature Pyramid Network (FPN) and U-Net rely on three primary network paths: bottom-up, top-down, and horizontal connections. These paths allow the integration of high-level semantic information with low-level geometric data. In FPN, the bottom-up path extracts high-level features, while the top-down path applies upsampling to enhance semantic details at higher resolutions. Horizontal connections fuse low-level convolution features with high-level features, resulting in a more detailed representation of semantic information.

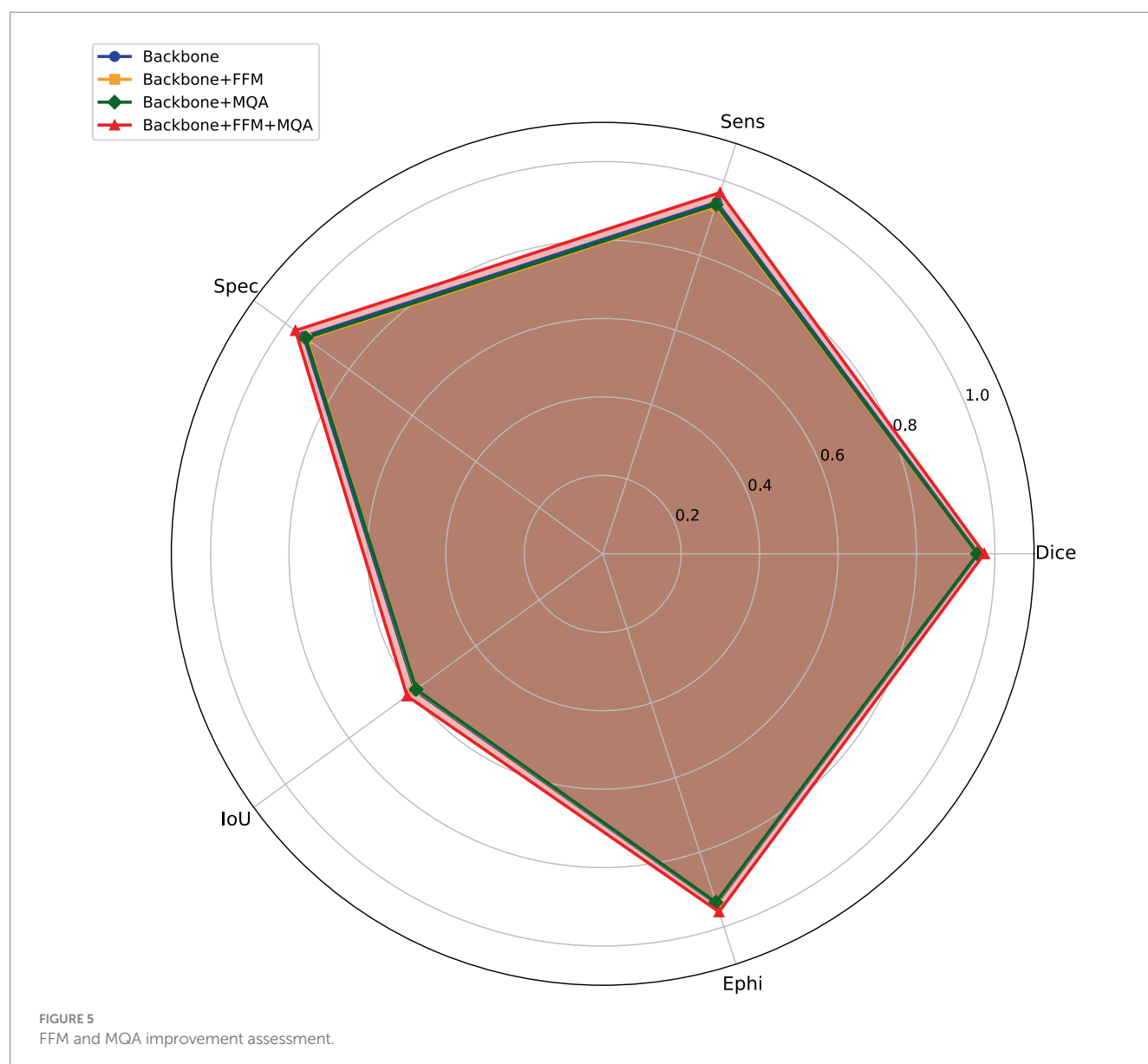
Nevertheless, FPN faces challenges due to its complex hierarchical structure. The computation of intermediary layers relies heavily on the higher-level layers, requiring the analysis of preceding layers to be completed before passing information to subsequent layers. This dependency can lead to inefficiencies in computation and integration. The proposed MQA approach addresses these limitations by enabling

the simultaneous integration of lower-level map attributes, higher-level attribute maps, and edge attribute maps, streamlining the process and enhancing feature representation.

The current work offers a direct and efficient MQA approach and introduces a cascading multi input computational framework. The system employs a mechanism for attention to iteratively compute and use feature maps of varying sizes, directing the ultimate semantic segmentation process. The proposed MQA has the ability to combine multiple input and multiple scale features, and can be trained end-to-end with ground truth supervision. The proposed technique effectively leverages both low and high-resolution features and integrates a method of attention to successfully accomplish the segmentation job on ArUco markers.

4.2 Comparison of different combination MQA and FFM

To evaluate the impact of MQA and FFM on MoNetViT's performance, a series of tests were conducted. The experiments utilized the same network backbone and implementation details to ensure consistency with previous studies. The results, as shown in the



radar chart and table, compare the baseline “Backbone,” “Backbone+MQA,” “Backbone+FFM,” and “Backbone+MQA + FFM” configurations. The radar chart illustrates that the region representing “Backbone+MQA + FFM” (in red) is larger than those of other configurations, indicating superior performance across key metrics. Similarly, the table reinforces these findings, showing that “Backbone+MQA + FFM” achieves the highest Dice score (0.9731), sensitivity (0.9686), and specificity (0.9686), along with the lowest MAE (0.0049). These results suggest that integrating both MQA and FFM significantly enhances the backbone’s performance, as the areas with MQA and FFM have a notably larger magnitude than those without these modules.

The contributions of the Multi-Query Attention (MQA) and Feature Fusion Module (FFM) to the segmentation performance were further validated through an ablation study. Table 4 highlights the significant improvements achieved by integrating these modules into the baseline model. Specifically, the Dice coefficient increased from 0.9584 for the baseline to 0.9558 (+1.7%) with MQA alone and 0.9566

(+1.9%) with FFM alone. When both modules were combined, the Dice coefficient reached 0.9731 (+3.4%), demonstrating their synergistic effect. Furthermore, sensitivity and specificity improved from 0.9424 each in the baseline to 0.9686 with the combined MQA and FFM setup, while the Mean Absolute Error (MAE) decreased from 0.0077 to 0.0049. These findings emphasize the critical role of MQA in enhancing multi-scale feature integration and FFM in refining feature clarity, resulting in superior segmentation performance. This robust improvement across key metrics underscores the effectiveness of the proposed MoNetViT architecture in addressing complex segmentation tasks.

MoNetViT’s architecture demonstrates significant advantages through its dual-path encoder, which effectively balances local feature extraction using CNNs and global feature extraction via Transformers. This design allows the model to capture both fine-grained details and long-range dependencies, improving segmentation performance. Additionally, the integration of the Multi-Query Attention (MQA) module enhances multi-scale feature integration, enabling the model

to better aggregate features across varying spatial scales. This contributes to improved segmentation accuracy, particularly in complex scenarios. Furthermore, MoNetViT's lightweight design minimizes computational demands, making it highly efficient for real-time applications without sacrificing performance. Compared to models like DeepLabV3+, which rely on more resource-intensive architectures, MoNetViT achieves a superior balance of accuracy and efficiency, reinforcing its suitability for deployment in resource-constrained environments.

While MoNetViT demonstrates strong performance, scalability to larger datasets may require optimization strategies, and its adaptability to diverse marker types needs further evaluation across different styles and conditions. Future experiments will focus on enhancing scalability and generalizability through transfer learning and broader dataset evaluations.

5 Conclusion

This study introduces a novel model called MoNetViT, which utilizes fused CNNs and transformers to create a segmentation model for ArUco marker-infested regions. The picture features are extracted simultaneously utilizing Convolutional Neural Networks (CNNs) and transformers, resulting in a reduction in computing burden and model complexity, while enhancing the segmentation performance. Furthermore, this work introduces the multi-query attention (MQA) module as a means to enhance performance. The empirical findings demonstrate that MoNetViT outperforms the other approaches on the ArUco dataset. Further studies will concentrate on the influence of the merging of each model and on discovering strategies for decreasing the level of detail within the model. Future research will focus on enhancing the capabilities of MoNetViT to achieve even more robust outcomes. This includes exploring additional fusion methods to further optimize feature integration and segmentation accuracy. Another key direction is adapting the model for outdoor environments by incorporating GPS data, thereby extending its applicability to diverse navigation scenarios. Leveraging transfer learning techniques will also be prioritized to reduce training times and improve scalability, enabling the model to handle larger and more diverse datasets effectively. These advancements aim to broaden the utility and efficiency of MoNetViT, ensuring its suitability for a wide range of real-world applications.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

References

- Abozeid, A., Taloba, A., Faiz Alwaghid, A., Salem, M., and Elhadad, A. (2023). An efficient indoor localization based on deep attention learning model. *Comput. Syst. Sci. Eng.* 46, 2637–2650. doi: 10.32604/csse.2023.037761
- Asadi Shamsabadi, E., Xu, C., Rao, A. S., Nguyen, T., Ngo, T., and Dias-da-Costa, D. (2022). Vision transformer-based autonomous crack detection on asphalt and concrete surfaces. *Autom. Constr.* 140:104316. doi: 10.1016/j.autcon.2022.104316
- Bai, J., Liu, Z., Lin, Y., Li, Y., Lian, S., and Liu, D. (2019). Wearable travel aid for environment perception and navigation of visually impaired people. *Electronics* 8:697. doi: 10.3390/electronics8060697
- Benmouna, B., Pourdarbani, R., Sabzi, S., Fernandez-Beltran, R., García-Mateos, G., and Molina-Martínez, J. M. (2023). Attention mechanisms in convolutional neural networks for nitrogen treatment detection in tomato leaves using hyperspectral images. *Electronics* 12:22706. doi: 10.3390/electronics12122706

Author contributions

LT: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. RG: Conceptualization, Formal analysis, Funding acquisition, Investigation, Supervision, Validation, Writing – review & editing. Prayitno: Conceptualization, Formal analysis, Funding acquisition, Investigation, Supervision, Validation, Writing – review & editing, Data curation.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The authors thank the Ministry of Education, Culture, Research, and Technology for the research grant and support provided for this doctoral dissertation. Additionally, we extend our thanks to Diponegoro University, especially the Doctoral Program of Information Systems, for their continuous support and valuable resources that have significantly contributed to the success of this research.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). TransUNet: transformers make strong encoders for medical image segmentation, arXiv. 1–13. doi: 10.48550/arXiv.2102.04306
- Chen, Y., Wang, T., Tang, H., Zhao, L., Zhang, X., Tan, T., et al. (2023). CoTrFuse: a novel framework by fusing CNN and transformer for medical image segmentation. *Phys. Med.* 68:175027. doi: 10.1088/1361-6560/acde8
- d'Ascoli, S., Touvron, H., Leavitt, M. L., Morcos, A. S., Biroli, G., and Sagun, L. (2022). ConViT: improving vision transformers with soft convolutional inductive biases. *J. Stat. Mech. Theor. Exp.* 2022, 2286–2296. doi: 10.1088/1742-5468/ac9830
- Doppalapudi, S. A. I. K. (2023). Semantic image segmentation using transformers.
- Dosovitskiy, A. (2021). “An image is WORTH 16X16 WORDS: transformers for image recognition at scale,” in *ICLR 2021 - 9th International Conference on Learning Representations*.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). “An image is Worth 16x16 Words: transformers for image recognition at scale,” arXiv. doi: 10.48550/arXiv.2101.11929
- El-taher, F., Taha, A., Courtney, J., and Mckeever, S. (2021). A systematic review of urban navigation systems for visually impaired people. *Sensors* 21:3103. doi: 10.3390/s21093103
- Fan, D. P., Tao, Z., Ge-Peng, J., Yi, Z., Geng, C., Huazhu, F., et al. (2018). Enhanced-alignment measure for binary foreground map evaluation. *IJCAI*. 12, 698–704. doi: 10.24963/ijcai.2018/97
- Fan, D.-P., Zhou, T., Ji, G. P., Zhou, Y., Chen, G., Fu, H., et al. (2020). Inf-net: automatic COVID-19 lung infection segmentation from CT images. *IEEE Trans. Med. Imaging* 39, 2626–2637. doi: 10.1109/tmi.2020.2996645
- Fernando, N., McMeekin, D. A., and Murray, I. (2023). ‘Route planning methods in indoor navigation tools for vision impaired persons: a systematic review’, disability and rehabilitation. *Assist. Technol.* 18, 763–782. doi: 10.1080/17483107.2021.1922522
- Huang, Q., Lei, Y., Xing, W., He, C., Wei, G., Miao, Z., et al. (2022). Evaluation of pulmonary edema using ultrasound imaging in patients with COVID-19 pneumonia based on a Non-Local Channel attention ResNet. *Ultrasound in Medicine & Biology [Preprint]*. 48, 945–953. doi: 10.1016/j.ultrasmedbio.2022.01.023
- Hu, X., Yang, K., Fei, L., and Wang, K. (2019). ACNET: attention based network to exploit complementary features for RGBD semantic segmentation. *IEEE* 21, 1440–1444. doi: 10.1109/icip.2019.8803025
- Jeamwathanachai, W., Wald, M., and Wills, G. (2019). Indoor navigation by blind people: behaviors and challenges in unfamiliar spaces and buildings. *Br. J. Vis. Impair.* 37, 140–153. doi: 10.1177/0264619619833723
- Karimi, D., Dou, H., Warfield, S. K., and Gholipour, A. (2020). Deep learning with Noisy labels: exploring techniques and remedies in medical image analysis. *Med. Image Anal.* 65:101759. doi: 10.1016/j.media.2020.101759
- Kim, H. W., Lee, S., Yang, J. H., Moon, Y., Lee, J., and Moon, W. J. (2023). Cortical Iron accumulation as an imaging marker for neurodegeneration in clinical cognitive impairment Spectrum: a quantitative susceptibility mapping study. *Korean J. Radiol.* 24, 1131–1141. doi: 10.3348/kjr.2023.0490
- Kubota, M. (2024). Snap: smartphone-based indoor navigation system for blind people via floor map analysis and intersection detection. *Proc. ACM Hum. Comput. Int.* 8, 1–22. doi: 10.1145/3676522
- Kuriakose, B., Shrestha, R., and Sandnes, F. E. (2021). Towards independent navigation with visual impairment: a prototype of a deep learning and smartphone-based assistant. *Association for Computing Machinery*. 113–114. doi: 10.1145/3453892.3464946
- Kwak, J., and Sung, Y. (2021). DeepLabV3-refiner-based semantic segmentation model for dense 3D point clouds. *Remote Sens.* 13:165. doi: 10.3390/rs13081565
- Lazebnik, S., Schmid, C., and Ponce, J. (2006). “Beyond bags of features: spatial pyramid matching for recognizing natural scene categories,” in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2169–2178.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., et al. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 1234–1240. doi: 10.1093/bioinformatics/btz682
- Liu, Y., Zhang, Z., Liu, X., Lei, W., and Xia, X. (2021). Deep learning based mineral image classification combined with visual attention mechanism. *IEEE Access* 9, 98091–98109. doi: 10.1109/access.2021.3095368
- Martinez-Cruz, S., et al. (2021). An outdoor navigation assistance system for visually impaired people in public transportation. *IEEE Access* 9, 130767–130777. doi: 10.1109/access.2021.3111544
- Mehta, S., and Apple, M. R. (2022). MobileViT: light-weight, general-purpose, and Mobile-friendly vision transformer. *ICLR* 22:3. doi: 10.48550/arXiv.2110.02178
- Mei, Y., Fan, Y., Zhou, Y., Huang, L., Huang, T. S., and Shi, H. (2020). Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. arXiv. doi: 10.48550/arxiv.2006.01424
- Messaoudi, M. D., Menelas, B. A. J., and Mcheick, H. (2020). Autonomous smart white cane navigation system for indoor usage. *Technologies* 8:37. doi: 10.3390/technologies8030037
- Misawa, N., Yamaguchi, R., Yamada, A., Wang, T., Matsui, C., and Takeuchi, K. (2024). Design methodology of compact edge vision transformer CiM considering non-volatile memory bit precision and memory error tolerance. *Jap. J. Appl. Phys.* 63:03SP05. doi: 10.35848/1347-4065/ad1bbd
- Mujtaba, G., Malik, A., and Ryu, E. (2022). LTC-SUM: lightweight client-driven personalized video summarization framework using 2D CNN. *IEEE access* 10, 103041–103055. doi: 10.1109/access.2022.3209275
- Petit, O., Thome, N., Rambour, C., Themyr, L., Collins, T., and Soler, L. (2021). “U-net transformer: self and cross attention for medical image segmentation” in *Machine learning in medical imaging*. eds. C. Lianet al. (Cham: Springer International Publishing), 267–276.
- Qi, M., and Zhang, H. (2023). Dimensional emotion recognition based on two stream CNN fusion attention mechanism. *Third International Conference on Sensors and Information Technology (ICSIT 2023)*. (Eds.). H. Kannan and J. Hemanth. International Society for Optics and Photonics.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. Cham: Springer.
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). “MobileNetV2: inverted residuals and linear bottlenecks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4510–4520.
- Shah, P. A., Hessah, A. R. A. A., and Rayan, H. M. A. A. (2023). Machine learning-based smart assistance system for the visually impaired. 139–144. doi: 10.1109/ITT59889.2023.10184257
- Sun, S., Yue, X., Zhao, H., Torr, P. H. S., and Bai, S. (2022). ‘Patch-based separable transformer for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* 22, 1–8. doi: 10.1109/tpami.2022.3231725
- Tao, Y., and Ganz, A. (2020). Simulation framework for evaluation of indoor navigation systems. *IEEE access* 8, 20028–20042. doi: 10.1109/ACCESS.2020.2968435
- Thakur, P. S., Chaturvedi, S., Khanna, P., Sheorey, T., and Ojha, A. (2023). Vision transformer meets convolutional neural network for plant disease classification. *Eco. Inform.* 77:102245. doi: 10.1016/j.ecoinf.2023.102245
- Theodorou, P., Tsiligkos, K., Meliones, A., and Tsigris, A. (2022). An extended usability and UX evaluation of a Mobile application for the navigation of individuals with blindness and visual impairments indoors: an evaluation approach combined with training sessions. *Br. J. Vis. Impair.* 42, 86–123. doi: 10.1177/02646196221131739
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2023). Attention is all you need. arXiv. doi: 10.48550/arXiv.1706.03762
- Wang, X., Girshick, R., Gupta, A., and He, K. (2018). “Non-local neural networks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7794–7803.
- Wang, Y., Yi, B., and Yu, K. (2022). Exploration and research about key technologies concerning deep learning models targeting Mobile terminals. *J. Phys. Conf. Series* 2303:012086. doi: 10.1088/1742-6596/2303/1/012086
- Wang, Z., Jianbo, W., Pengfei, M., Fuyong, W., and Wenbao, D. (2019). “Evaluation for parachute reliability based on fiducial inference and Bayesian network,” in *Proceedings of the 2018 International Conference on Mathematics, Modeling, Simulation and Statistics Application (MMSSA 2018)*. London: Atlantis Press.
- Wen, F., Wang, M., and Hu, X. (2023). DFAM-DETR: deformable feature based attention mechanism DETR on slender object detection. *IEICE Trans. Inf. Syst.* E106, 401–409. doi: 10.1587/transinf.2022edp7111
- Xia, Z., and Kim, J. (2023). Enhancing mask transformer with auxiliary convolution layers for semantic segmentation. *Sensors* 23:20581. doi: 10.3390/s23020581
- Yeung, M., Sala, E., Schönlieb, C. B., and Rundo, L. (2022). Unified focal loss: generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Comput. Med. Imaging Graph.* 95:102026. doi: 10.1016/j.compmedimag.2021.102026
- Zhang, Y., Liu, H., and Hu, Q. (2021) “TransFuse: fusing transformers and CNNs for medical image segmentation,” in *BruijneM. de Medical Image Computing and Computer Assisted Intervention -- MICCAI 2021*. Cham: Springer International Publishing, 14–24
- Zhang, Z., Gao, Q., Liu, L., and He, Y. (2023). A high-quality Rice leaf disease image data augmentation method based on a dual GAN. *IEEE access* 11, 21176–21191. doi: 10.1109/access.2023.3251098
- Zhou, C., Zhang, X., and Zhong, Y. (2024). Are transformers more suitable for plant disease identification than convolutional neural networks?. doi: 10.21203/rs.3.rs-4284240/v1
- Zhou, L., Zhu, M., Xiong, D., Ouyang, L., Ouyang, Y., and Zhang, X. (2023). MR image reconstruction via non-local attention networks. *Fourteenth International Conference on Graphics and Image Processing (ICGIP 2022)*. (Eds.). L. Xiao and J. Xue. International Society for Optics and Photonics.
- Zhu, L., Kang, Z., Zhou, M., Yang, X., Wang, Z., Cao, Z., et al. (2022). CMANet: cross-modality attention network for indoor-scene semantic segmentation. *Sensors* 22:8520. doi: 10.3390/s22128520
- Zhu, Z., Mengde, X., Song, B., Tengting, H., and Xiang, B. (2019). “Asymmetric non-local neural networks for semantic segmentation,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 593–602.



OPEN ACCESS

EDITED BY

Sokratis Makrogiannis,
Delaware State University, United States

REVIEWED BY

Domenico Pomarico,
University of Bari Aldo Moro, Italy
Amrutanshu Panigrahi,
Siksha O Anusandhan University, India
Meng Wang,
Liaoning Technical University, China

*CORRESPONDENCE

Nazmul Ahasan Maruf
✉ nmaruf0001@stu.kau.edu.sa

RECEIVED 09 June 2024

ACCEPTED 10 March 2025

PUBLISHED 01 April 2025

CITATION

Maruf NA and Basuhail A (2025) Breast cancer diagnosis using radiomics-guided DL/ML model-systematic review and meta-analysis. *Front. Comput. Sci.* 7:1446270. doi: 10.3389/fcomp.2025.1446270

COPYRIGHT

© 2025 Maruf and Basuhail. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Breast cancer diagnosis using radiomics-guided DL/ML model-systematic review and meta-analysis

Nazmul Ahasan Maruf* and Abdullah Basuhail

Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

Cancer is one of the leading causes of death on a global scale, whereas breast cancer is the type of cancer that affects the most women. Early detection and accurate staging are essential for effective cancer treatment and improved patient outcomes. Recent developments in medical imaging and artificial intelligence (AI) have created new opportunities for breast cancer detection and staging. Medical image analysis techniques, including radiomics, machine learning and deep learning, have shown promise for breast cancer detection and stage estimation. The goal of the systematic review and meta-analysis is to evaluate and examine the state-of-the-art implications of radiomics-guided deep learning (DL) approaches for breast cancer early detection utilizing different medical image modalities. The selection criteria were established on the basis of the PRISMA statement. Our research employs a PICO structure and text mining technique (Topic Modeling) using Latent Dirichlet allocation (LDA) approach. The primary objective of the search was to conduct a thorough evaluation of the literature related to radiomics analysis and breast cancer in the fields of medical informatics, computer vision, and cancer research. Subsequently, the investigation concentrated on the fields of medical science, artificial intelligence, and computer science. The inquiry encompassed the years 2021 to 2024. The QUADAS-2 instrument is employed to evaluate the articles to ensure their quality and eligibility. Feature extraction methods that employ radiomics and deep learning are extracted from each study. The sensitivity value was pooled and transformed using a random-effects model to estimate the performance of DL techniques in the classification of breast cancer. The systematic review comprised 40 studies, while the meta-analysis consisted of 23 studies. The research studies employed a variety of image modalities, radiomics, and deep learning models to diagnose breast cancers. Ultrasound and DCI-MRI are the most frequently employed image modalities. The pyradiomics python package is employed to extract the radiomic features, and CNNs, ResNet, and DenseNet models are employed to extract the deep features. The LASSO (13) and *T*-test (9) statistical models are the most commonly used for feature selection. The most widely used deep learning models for breast cancer classification are ResNet and VGG. This systematic review and meta-analysis examined the feasibility of employing radiomics-guided deep learning/machine learning models for identifying breast cancer. The studies yielded positive results, as specific models demonstrated remarkable precision in distinguishing between malignant and benign breast tumors. However, there is a wide variety of variations in the designs of studies, the architectures of models, and the methodologies used for validation. Further research is required to verify the results of this study and to

investigate the potential of deep learning models guided by radiomics in the early detection of breast cancer.

KEYWORDS

breast cancer, medical image processing, radiomics, deep learning, AI, systematic review and meta-analysis

1 Introduction

Cancer is one of the primary causes of mortality worldwide, whereas breast cancer is the type of cancer that affects the greatest number of women worldwide (Sung et al., 2021). In 2020, approximately 2.3 million new cases of breast cancer were diagnosed, making it the most prevalent cancer among women globally (Lei et al., 2021). Although the death rates have decreased due to innovations in breast cancer early detection and diagnosis, it remains the second-leading cause of death for women (Arnold et al., 2022). However, effective cancer treatment and enhanced patient outcomes are contingent upon early detection and accurate staging. Staging is a comprehensive procedure that involves assessing the extent of cancer progression, including whether it has progressed from the breast to other regions of the body. Accurate staging could be useful in choosing the most effective treatment and determining the patient's prognosis.

In the past decade, there has been a growing topic of discussion on the significance of medical imaging techniques in breast cancer staging (Balkenende et al., 2022). Mammography is one of the most widely used image modalities for detecting breast cancer, and recently, with enhanced imaging approaches, it's become a potential tool for detection and diagnosis (Tsarouchi et al., 2023). A follow-up examination is crucial for detecting abnormalities identified during mammography and assessing the degree of dense breast tissue (Gatta et al., 2023). Mammography may be supplemented with other imaging modalities like computed tomography (CT), ultrasound, magnetic resonance imaging (MRI), and positron emission tomography (PET), as mammography has some limitations (Ha et al., 2023). Dynamic MRI (DCE-MRI) is a process of contrast enhancement based on the usage of images, wherein the time-dependent transformation in contrast enhancement is observed precisely (Ingrisch and Sourbron, 2013). This technique can help identify areas of the body where blood vessels associated with cancer are growing and determine the size and depth of the tumor. MRI also has one main negative aspect concerning unsatisfactory specificity that might lead to probable false positives, in addition to raising the overall cost compared to mammography and ultrasound (Zhang et al., 2023). Besides, it takes a long evaluation period (Wang L. C. et al., 2023). Elastography with ultrasound is a technique that is used to determine the body tissues' stiffness (Ditunno et al., 2023), which might be the symbol of malignant growth. Image processing is an essential part of the ultrasound screening technique used to examine and expose soft tissues. It enables the detection and description of breast abnormalities (Cè et al., 2023). Studies indicate that ultrasound

imaging techniques are perfectly safe for their frequent use and have the unique property of being free from radiation (Abhisheka et al., 2023).

Recent advances in medical imaging and artificial intelligence (AI) have created new opportunities for the detection and staging of breast cancer. Techniques of medical image analysis, such as radiomics and machine learning, have demonstrated promise for breast cancer detection and stage estimation (Wang Q. et al., 2023). Radiomics as a field of study is about obtaining quantitative features from medical images, which help to visualize and identify the patterns and biomarkers related to cancer growth (Rizzo et al., 2018). These attributes process information about the shape, texture, intensity levels, and spatial relations between the data in the images (Gupta et al., 2024). Radiomics analysis embraces most of the medical imaging modalities like computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, positron emission tomography (PET), and others (Peng et al., 2023). Those modalities should be an integral part of any analysis that seeks to establish the impact radiomics has on imaging for diagnosis, prognosis, treatment planning, or disease monitoring, specifically in cancer.

Deep learning (Guo et al., 2016; Dhar et al., 2023; Li X. et al., 2023), a subfield of machine learning, has shown enhanced performance in medical image analysis. In the research of Yuan et al. (2023), they emphasize that the CNN (convolutional neural network) can be trained to accurately recognize patterns and markers associated with breast cancer using medical images. Based on radiomics characteristics, machine learning and deep learning algorithms can be trained to develop predictive models for breast cancer detection and staging. Such models have demonstrated high accuracy rates, outperforming conventional diagnostic techniques. Early detection and accurate staging of breast cancer are crucial for selecting the optimal treatment and enhancing patient outcomes. Through using deep learning methods for medical image analysis, we can make our approach more accurate and go into the higher stages too, which in turn can lead to improved treatment outcomes and survival rates.

Despite the fact that a substantial amount of work has been done in this area in recent years, there is still potential for improvements, as the accuracy of radiomics-guided breast cancer detection and stage prediction can significantly impact the diagnosis of this disease. Hence, to lead our research on radiomics-guided breast cancer detection, we are now assessing the current state-of-the-art literature through a systematic review and meta-analysis. Our systematic review focused on the following questions:

- What are the most prominent deep learning and machine learning models that have been developed in state-of-the-art literature, and how are those models analyzed with performance metrics?
- What are the widely used feature extraction and selection processes in breast cancer detection-based medical image processing?
- How is radiomics-guided feature extraction made in cutting-edge research?

The forthcoming sections of the paper will be presented in the following arrangement: Section 2 describes the methodology of a systematic review taken. In the result Section 3, the findings of the study are provided. Section 4 discusses the findings. Lastly, Section 5 concludes the inferences from the analysis.

2 Methodology

Systematic reviews and meta-analyses serve as an essential conduit in breast cancer research for appraising the efficacy of radiomics analysis using deep learning techniques. This systematic approach carefully weighs and combines ongoing research to get an overall verdict on the effectiveness of deep learning radiomics in improving the early detection, diagnostic, and treatment functions of breast cancer.

2.1 Search strategy

In order to conduct a comprehensive and systematic search, we devised a search strategy to locate pertinent material. The search technique was customized for five databases: Scopus, Web of Science, Science direct, IEEE and Google scholar. The search terms used were “Radiomics” AND “Radiomics Analysis” AND “Deep learning” AND “Breast Cancer.” The search included all data from the database’s establishment until 2021, specifically focusing on journal articles, conferences and review articles that were published only in English.

2.2 Eligibility criteria

To ensure the eligibility of studies for our present study, we established precise criteria for inclusion and exclusion. Additionally, we created a PICO structure specific to our research, which comprises the following components: P, Breast Cancer Patient; I, Deep Learning and Image Processing Approaches; C, Radiomics Analysis; Radiomics Guided; O, Image Classification, Segmentation, Prediction, Detection, and Medical Image Analysis. In order to minimize publication bias, we also implement the text mining technique (topic modeling) for knowledge discovery using the latent Dirichlet allocation (LDA) approach.

2.3 Inclusion criteria

In our systematic review and meta-analysis, we established inclusion criteria based on many factors. All the research conducted was centered on the English language. The studies specifically targeted breast cancer, and the reported findings were based on the analysis of deep learning and radiomics features in breast cancer images.

2.4 Exclusion criteria

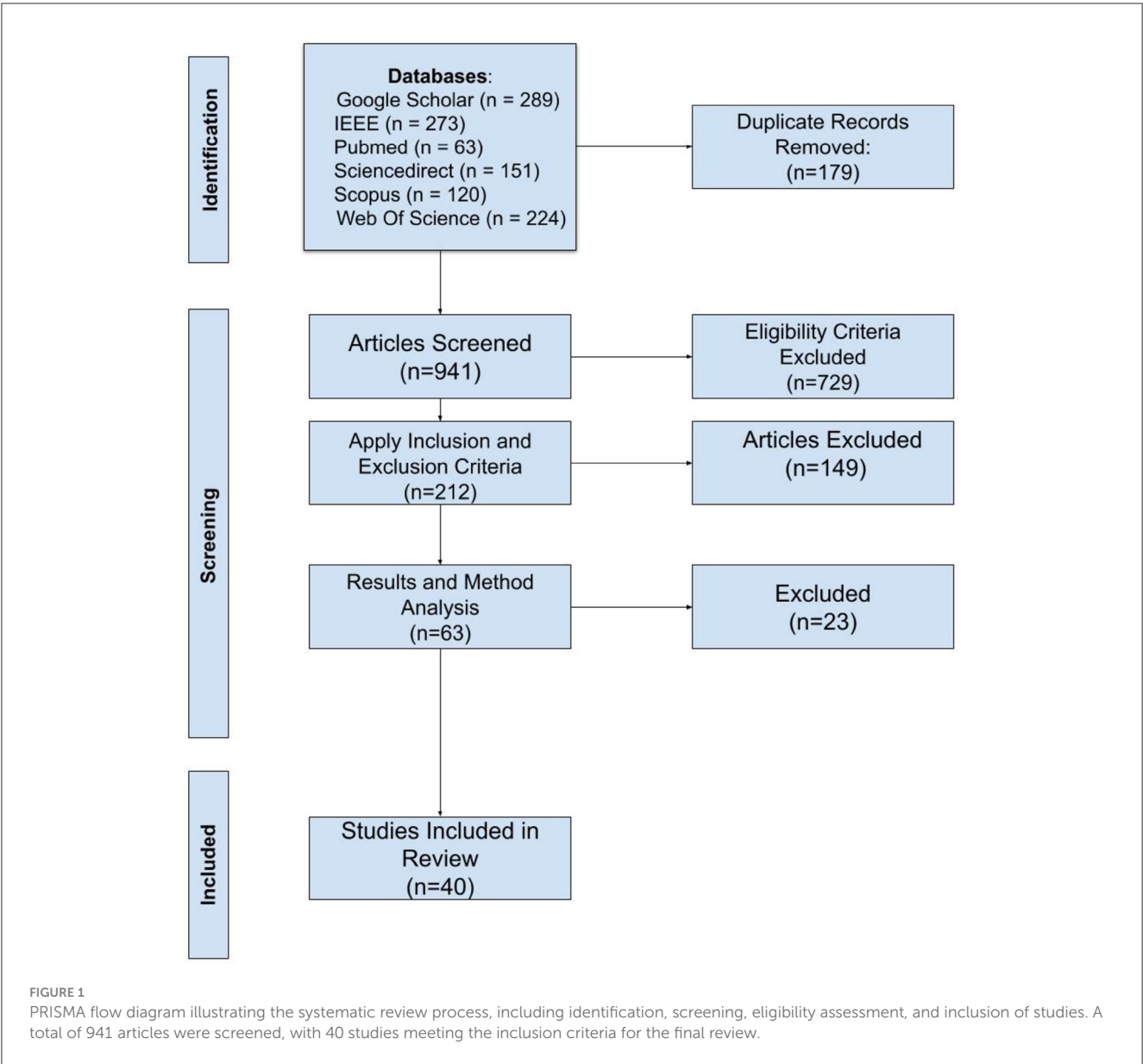
We have excluded from our systematic review and meta-analysis any studies that meet particular criteria. These requirements include research that does not analyze the results of deep learning approaches, case report studies, book chapters, conference abstracts, comments, letters to the editor, or studies published in languages other than English. In addition, research focusing on topics other than breast cancer is also eliminated.

2.5 Selection criteria

The selection criteria were derived from the PRISMA Statement (Moher et al., 2009). The search primarily focused on a comprehensive assessment of the literature pertaining to Radiomics Analysis and Breast cancer within the domains of cancer research, medical informatics, and computer science. The inquiry thereafter focused on the subject areas of computer science, Artificial Intelligence, and medical science. The search included the time period from 2021 to 2024. Articles published before to 2021 were omitted from the search results. The search mostly focused on certain keywords such as Deep learning, Radiomics, Radiomics Analysis, Breast cancer, and image processing. The selection method started by eliminating any redundant items from the study selection process. The studies underwent a two-step procedure, which included screening the titles and abstracts, followed by an in-depth review of the entire texts. Only pertinent papers that satisfied the predetermined criteria for inclusion were selected for comprehensive evaluation after conducting a thorough examination of the title and abstract. Figure 1 presents the PRISMA flow chart, which visually represents the process of selecting studies for the research.

2.6 Quality assessment

A quality appraisal tool called QUADAS-2 was used to assess the quality of studies on diagnostic tests (Whiting, 2011). This tool contains four questions, which are categorized into four different domains. The categories include patient selection, index tests, time and flow, and reference standards. Here are the questions: (1). Could the selection of patients have introduced bias? (2). Was the threshold value pre-specified, if threshold value used? (3). Is the reference standard likely to correctly classify the target condition? (4). Were all patients included in the analysis?



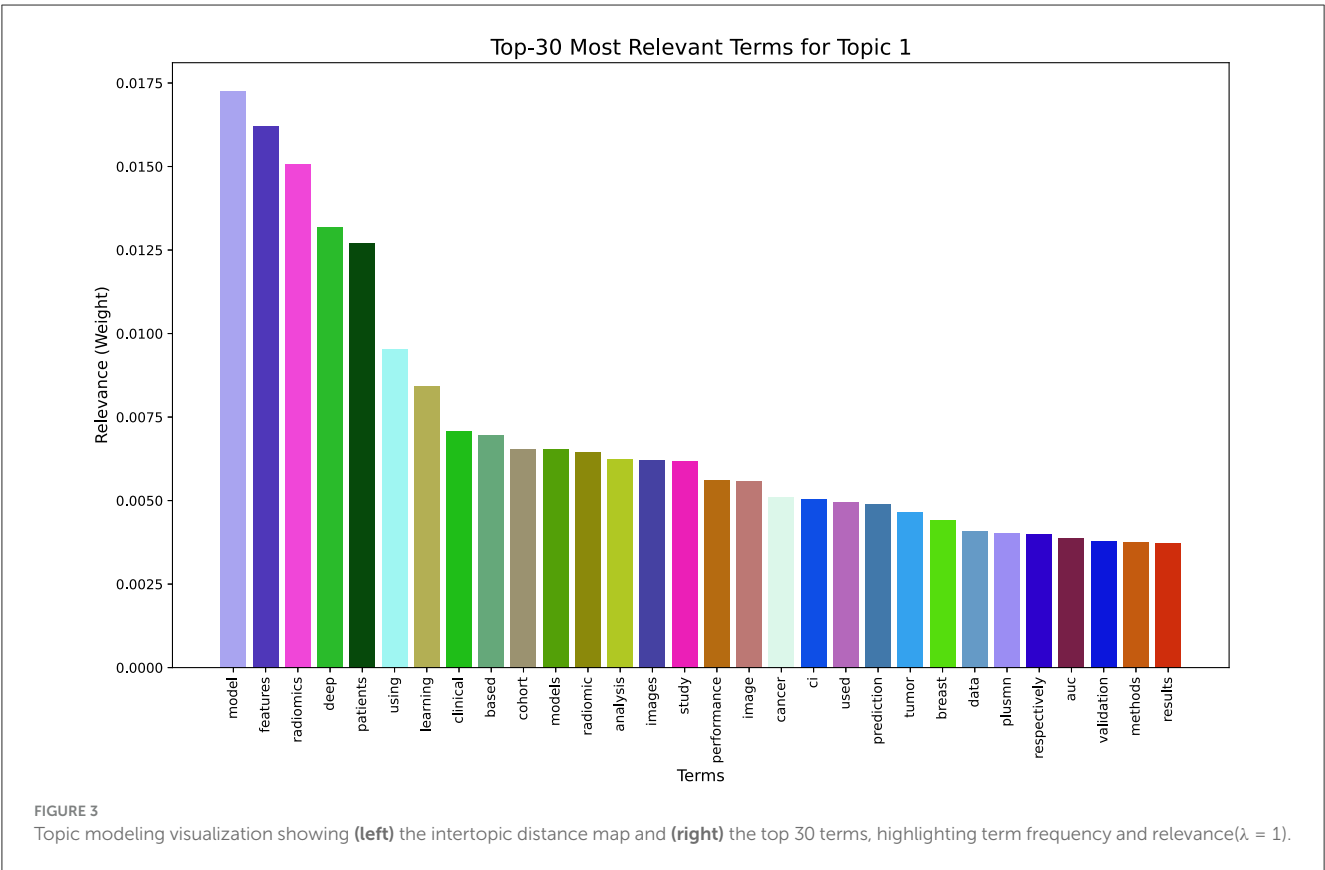
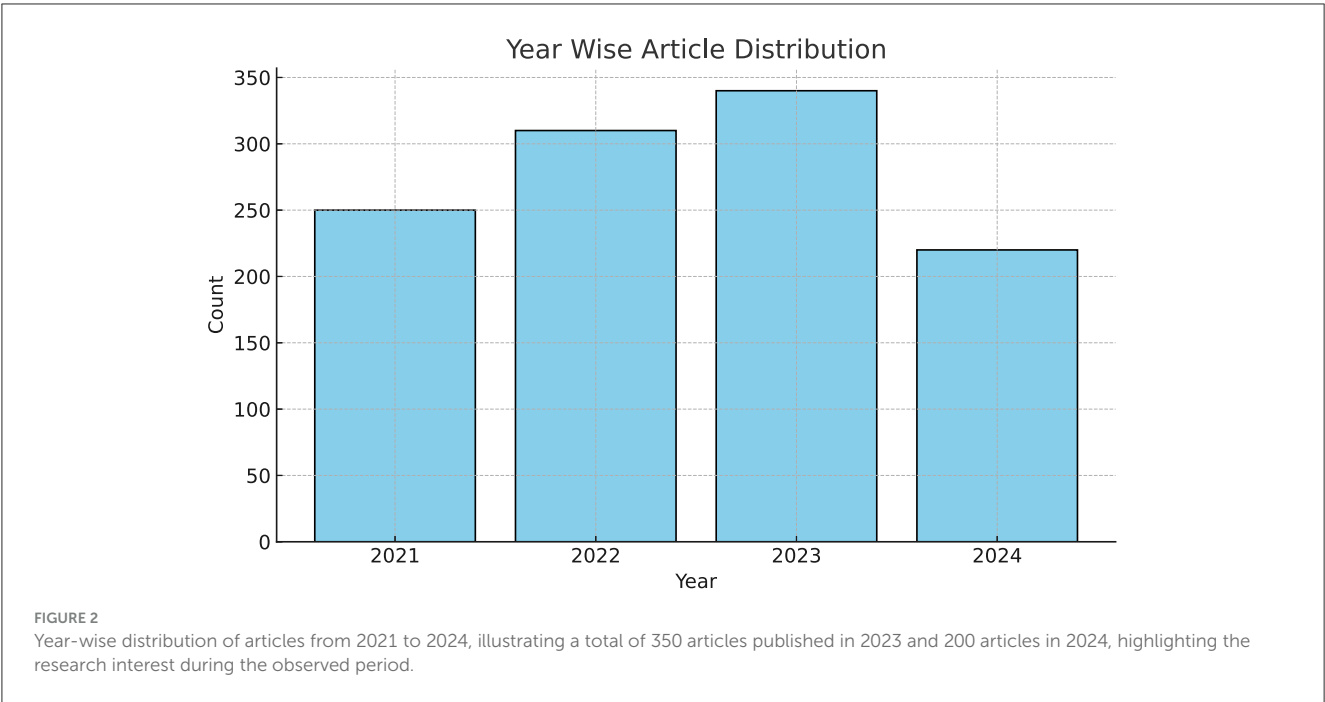
2.7 Meta-analysis

We employed a correlation coefficient-specific random effect model for correlation data with the application of sensitivity and sample size analysis. The web application [Jamovi (Version 2.5), 2024; Hornik, 2012; Viechtbauer, 2010] was used to effect and conduct the meta-analysis of all the data.

3 Result

The initially conducted electronic systematic search yielded 1120 studies from databases such as PubMed, Google Scholar, IEEE, Science Direct, Web of Science, and Scopus. Following the elimination of the duplicate, we discovered 941 articles. Figure 2 depicts the year-wise distribution of articles. The x-axis

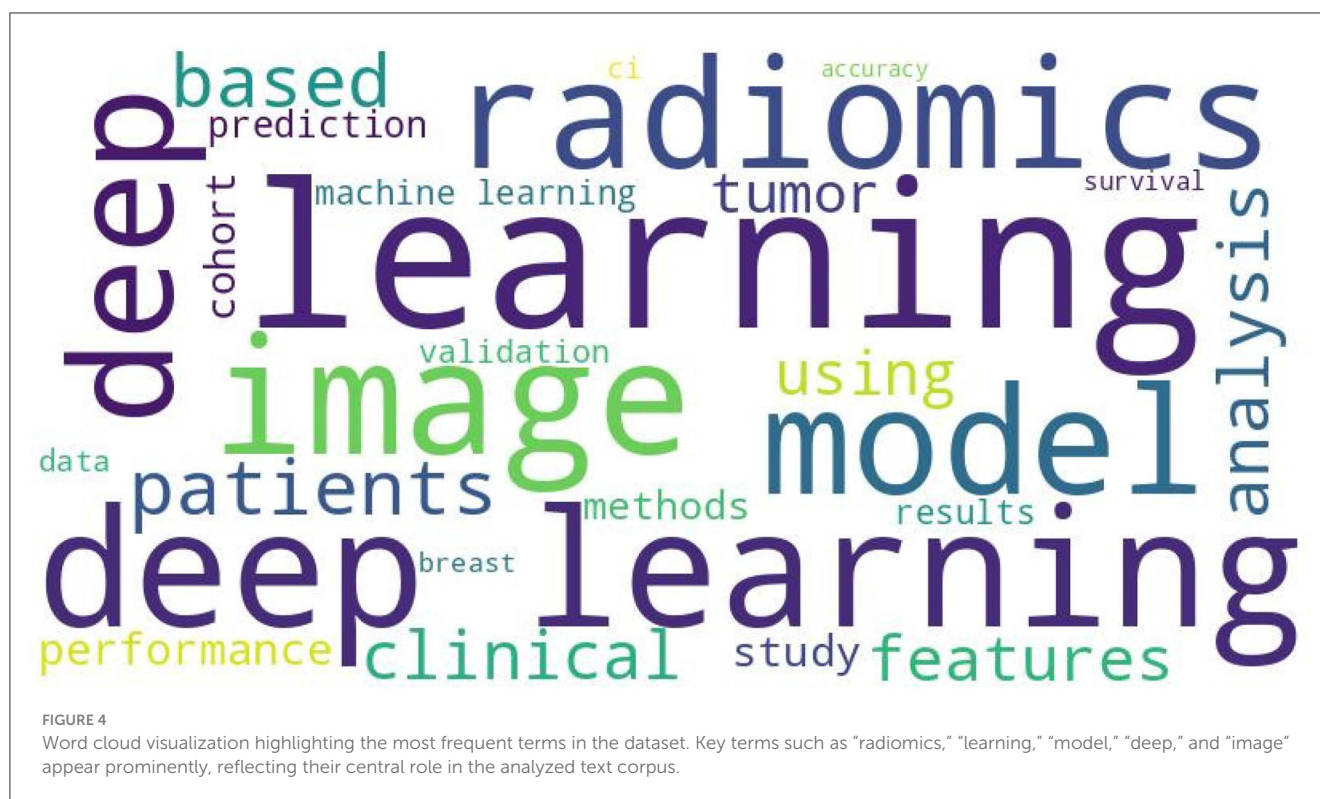
represents the time from 2021 to 2024, with each year being shown. The vertical axis with blue bars represents the number of papers published each year. There has been a consistent growth in the quantity of published papers over the last four years. Subsequently, we use inclusion and exclusion criteria to assess all titles, abstracts, and keywords using our unique keywords. As a result, we identified 63 papers for additional investigation. We use a custom Python script to implement screening criteria and conduct the actions. Out of the total number of articles screened, 23 were removed because they did not have sufficient research relevance to our research and did not provide clear information on the sensitivity, specificity, accuracy, precision, and AUC of outcomes. Ultimately, 40 studies satisfied all the requirements and were subsequently added for additional examination. Figure 1 depicts the PRISMA flowchart used for the selection procedure.



3.1 Knowledge discover from literature abstract

Topic modeling is a text mining technique that has demonstrated its effectiveness as a method for doing systematic

literature reviews (Asmussen and Muller, 2019). The most common topic modeling technique for knowledge discovery, Latent Dirichlet Allocation (LDA), finds meaningful topics in multiple literature by calculating the probability of words from each topic. LDA reveals latent topics for papers by extracting



a set of words with high probabilities (Jelodar et al., 2019). In our systematic review and meta-analysis, we extract the abstracts from selected 40 literature and identify the latent knowledge discoveries from each paper by analyzing the frequency of words and their probabilities using LDA. The results from Figure 3 shows that certain words like radiomics, image, breast, deep, learning appear more frequently, and have higher probabilities. In our topic modeling we followed a specific methodology. After extracting the abstracts we first preprocess the text by removing stop words and punctuation after that we convert the abstract to lower case. We create a dictionary and document-term matrix to identify the most frequent words and topics that appear in the abstracts. Then, we use LDA for topic modeling. The Figure 4 shows most frequent words used in abstract. This word cloud supports the relevance of our paper’s focus on deep learning, machine learning, and radiomics analysis by highlighting these topics as prominent themes in recent research. The prominence of terms like “deep learning,” “machine learning,” and “radiomics” reflects a significant trend in applying advanced learning techniques to medical imaging, which aligns closely with our study’s approach. Additionally, terms such as “patients,” “clinical,” and “cohort” underscore a focus on real-world clinical applications and patient outcomes, validating our research’s emphasis on improving healthcare through radiomics-based predictive modeling. The presence of words like “model,” “validation,” “performance,” and “prediction” shows that evaluating model accuracy is crucial in this field, which supports the significance of our work in assessing and validating predictive models. Furthermore, the word “image” emphasizes the role of medical imaging, demonstrating the importance of our findings in enhancing diagnostic capabilities through radiomics analysis. Overall, the word cloud illustrates that our study not only

aligns with high-interest areas but also contributes meaningfully to the ongoing discourse in medical machine learning, reinforcing our research’s novelty and practical impact.

3.2 Study classification

A collection of 40 papers, with publication dates ranging from 2021 to 2024, was used in this study. Overall, 12,685 patients were in the study, which trained, tested, and validated deep learning models for processing outcomes. In this review, the participants were between 48 and 70. Figure 5a shows the distribution of patients among different publications. The articles that were selected included an average of 507 patients. The imaging modalities highlighted in the literature focus on DCI-MRI, a commonly utilized technique. The Figure 5b showcases the imaging techniques frequently used in papers where ultrasound ranked as the second most popular modality. Twelve studies utilized DCI-MRI image modalities. Out of these, five studies utilize various techniques to forecast breast cancer. Various methods involve the prediction of cancer states, classification of sentinel lymph nodes (SLN) and metastasis (SLNM), estimation of HER2 expression in breast cancer, and forecasting preoperative axillary lymph node (ALN) status. Other imaging techniques are commonly employed to predict the condition of axillary lymph nodes (ALN). In addition, researchers use these techniques to predict the probability of achieving a pathological complete response (pCR) to neoadjuvant chemotherapy (NAC), evaluate lymphovascular invasion (LVI) in patients, and make predictions based on the proliferation marker Ki-67. Figure 5c in this study illustrates the frequency distribution of cancer biomarker values and treatment

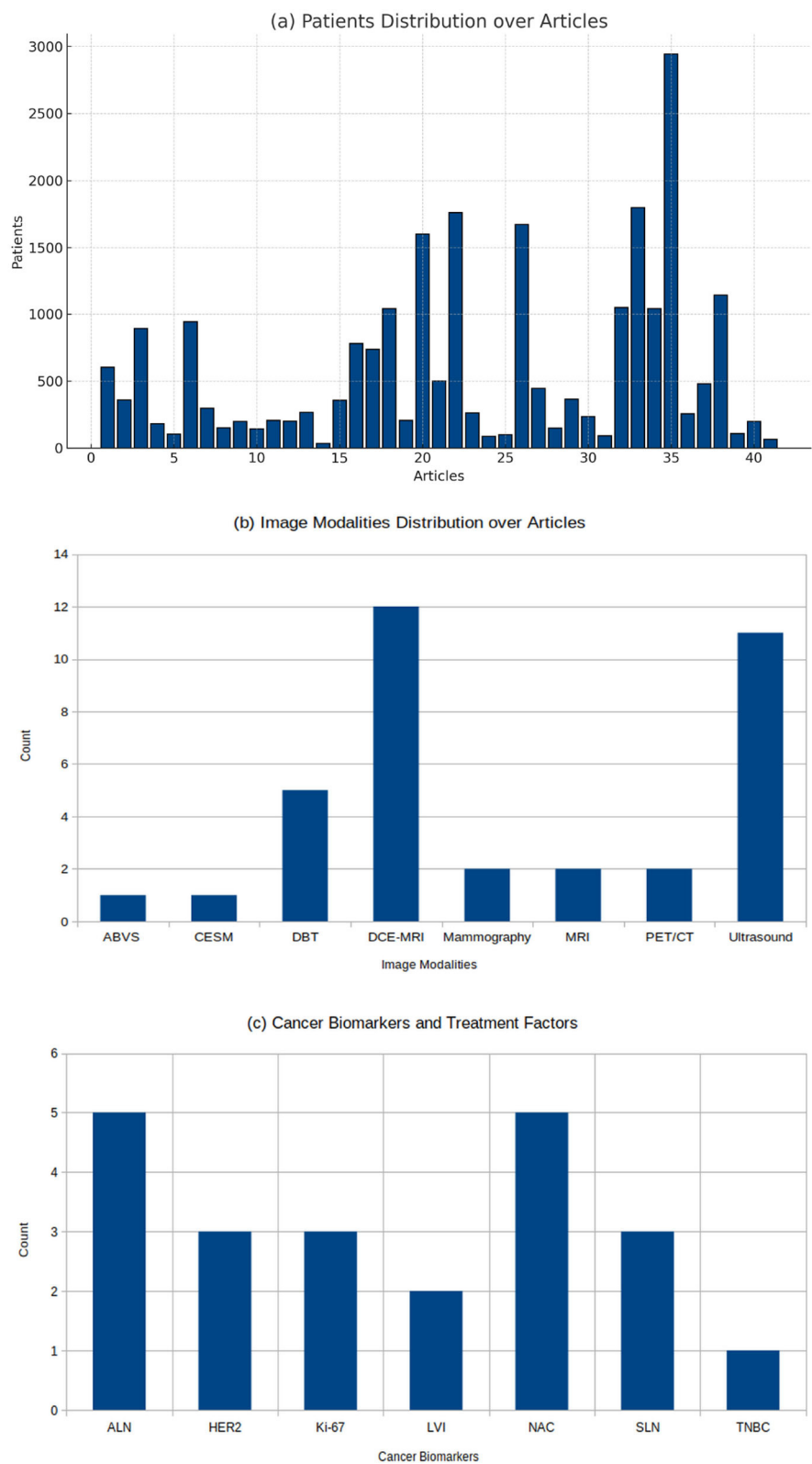


FIGURE 5 Illustrating key distributions across analyzed articles: **(a)** Patient distribution over 40 articles, with the maximum patient count reaching approximately 3000 in individual studies. **(b)** Image modality distribution, where DCE-MRI and Ultrasound are the most frequently utilized techniques. **(c)** Distribution of cancer biomarkers and treatment factors, highlighting NAC and ALN as the most reported, followed by HER2 and SLN, reflecting the emphasis on these factors in the studies.

factors in the selected literature. The 13 reviewed studies centered their analyses on ALN, SLN, and NAC biomarkers. Nine of the articles used specific cancer biomarkers, including HER2, Ki-67, triple-negative breast cancer (TNBC), and LVI, employing diverse methodologies. [Table 1](#) explains the processing method for particular biomarkers and the process for choosing radiomics features using feature selection. In radiomics analysis, several easily detectable changes are qualitative characteristics extracted from CT, MRI, or PET scans. Different data types are sent through various stages of analysis to seek patterns and correlation ([Kumar et al., 2012](#)). Likewise, such patterns provide information that is vital for the diagnosis and prediction of prognosis. They act closely with the clinical outcomes, aligning with them ([Traverso et al., 2018](#)). The major part of the chosen articles employs the usual features of radiomics, which are first-order Features, Shape features (3D and 2D), Gray Level Co-occurrence Matrix too (GLCM), Gray Level Run Length Matrix (GLRLM), and Gray Level Size Zone Matrix (GLSZM), Neighboring Gray Tone Difference Matrix (NGTDM), and Gray Level Dependence Matrix (GLDM). After reviewing 40 studies ([Table 1](#)), DCI-MRI and Ultrasound emerged as the most widely used imaging modalities for breast cancer detection. Key cancer biomarkers included ALN, NAC, and SLN, central to various diagnostic and prognostic analyses. Advanced radiomics features were frequently selected to enhance predictive accuracy, demonstrating the essential role of these tools in personalized cancer assessment.

3.3 Feature extraction and selection model

The [Table 2](#) presented all extraction-feature and selection methods used in the selected studies. This knowledge will be a facilitator for determining the ways in which case data analysis can be conducted and the results can be deduced. This table provides an in-depth review through which we can learn trend details of feature selection methods and also get a grasp of the significance of feature features in different tasks. The existing radiomics method is PyRadiomics, a highly reliable approach used for feature extraction with the CNNs, ResNet, and DenseNet models following. There were twenty four studies that used pyradiomics in order to extract radiomics features for different imaging modalities. Furthermore, out of the five papers that utilized deep learning techniques, three of them used CNN, ResNet, or VGG-16 to extract their features. The number of features extracted varies significantly across studies, from as few as 25 to over 11,000, highlighting differences in feature granularity and study focus. The table indicates that many studies in radiomics and breast cancer research utilize privately collected datasets from different hospitals, reflecting a common approach where institutions collect and analyze their imaging data. This reliance on private datasets allows for tailored data that fits specific study objectives. It also introduces variability, making it challenging to replicate results or compare findings across studies due to differences in data acquisition protocols, equipment, and patient demographics. On the other hand, the Duke-Breast-Cancer-MRI ([Saha et al., 2018](#)) dataset is one of the most frequently used publicly available datasets in this domain. The technique of feature selection generally involves the use

of an array of methods, with machine learning approach like LASSO being the most popular option that is frequently used. Statistical methods like the ANOVA, t-test, the Spearman rank correlation coefficient, and correlation analysis are the methods often selected for feature selection. Other advanced selection models include neural networks, ensemble methods like Random Forest and LightGBM, as well as ranking algorithms such as Mutual Information, Gain Ratio, and Information Gain. These techniques help manage complex data relationships effectively. This was observed as 15 studies applied the LASSO, 5 studies adopted the ANOVA, and other studies used the Spearman rank correlation coefficient, correlation analysis, T-test, PCA, U-Test. This overview highlights the wide array of tools and methods utilized in radiomics research, reflecting a trend toward integrating statistical rigor with machine learning capabilities for effective feature selection.

3.4 Characteristics of DL/ML models

The [Table 3](#) presents a comprehensive overview of the performance of various deep learning (DL) and machine learning (ML) models utilized in studies, highlighting their sensitivity, specificity, accuracy, and area under the curve (AUC) metrics. The performance metrics presented here provide a rounded evaluation of each model's effectiveness. Sensitivity, or recall, reflects a model's ability to correctly identify positive cases, making it essential in scenarios where missing a positive result is costly, such as medical diagnoses. Specificity measures how accurately a model identifies negative cases. Reducing false positives and avoiding misclassifying negative cases as positive is crucial. Conversely, accuracy assesses the model's overall efficacy by determining the correct positive and negative prediction ratio to the total predictions made. Lastly, the Area Under the Curve (AUC) provides a holistic measure of the model's discriminatory power across various threshold settings, with higher values indicating improved performance in differentiating between positive and negative outcomes. These metrics comprehensively understand each model's performance, highlighting their strengths and limitations in various contexts. The [Table 3](#) indicates that ResNet (10 of the studies) is a preferred model in breast cancer analysis, with numerous studies employing its different variations due to its effectiveness in extracting detailed imaging features. For instance, [Beuque et al. \(2023\)](#) use ResNet101 along with Mask R-CNN, achieving a strong balance of high sensitivity and AUC, highlighting ResNet's robustness in capturing intricate details in imaging data. The study employs a data split rather than cross-validation, dividing the dataset into training, test, and external validation sets. Specifically, the study allocated 850 patients to the training set, 212 patients to the test set, and 279 patients to an external validation set from a separate institution. This setup does not involve cross-validation folds; instead, using an independent external dataset is an additional test for the model's generalizability. The external validation provides a strong benchmark for testing the model on unseen data from a different population. The wide adoption of ResNet's variations, including ResNet18, ResNet34, ResNet50, and ResNet101, across studies reassures researchers about its adaptability to specific data sizes and computational resources. It also found that authors leverage

TABLE 1 Summary of imaging modalities, cancer biomarkers, and radiomics feature extraction techniques reported across reviewed studies, highlighting the diversity in methodologies and patient cohorts.

References	Year	Image modality	Patients	Cancer bio markers	Radiomics features
Yang et al., 2023	2023	DCE-MRI	206	Lymphovascular Invasion(LVI)	7 Radiomics features
Abbasian Ardakani et al., 2023	2023	Ultrasound	200	None	3 Radiomics features
Wang C. et al., 2023	2023	Ultrasound	359	Sentinel Lymph Node(SLN)	DL-Grayscale DL-CDFI DL-elastography
Yu et al., 2023	2023	Ultrasound	603	Neo Adjuvant Chemotherapy (NAC)	Deep learning radiomics
Wei et al., 2023	2023	Ultrasound	892	Axillary Lymph Node(ALN)	5 radiomics features
Zheng et al., 2023	2023	PET/CT	105	Neo adjuvant Chemotherapy (NAC)	7 Radiomics features
Gao et al., 2023	2022	DCE-MRI	941	Axillary lymph Node(ALN)	7 Radiomics features
Xu et al., 2022	2022	Ultrasound	144	Human epidermal Growth factor receptor 2 (HER2)	None
Peng et al., 2022	2022	DCE-MRI	356	Neo adjuvant Chemotherapy (NAC)	7 Radiomics features
Vigil et al., 2022	2022	Ultrasound	780	None	7 Radiomics features
Rashid et al., 2022	2022	MRI	35	None	7 Radiomics features
Chen et al., 2022	2022	Digital breast Tomosynthesis (DBT)	298	None	7 Radiomics features
Jiang et al., 2022	2022	Digital breast Tomosynthesis (DBT)	266	K-67	7 Radiomics features
Wang et al., 2022	2022	DCE-MRI	151	Axillary Lymph Node(ALN)	7 Radiomics features
Cattell et al., 2022	2022	DCE-MRI	198	Sentinel Lymph Node(SLN)	5 Radiomics features
Wu et al., 2022	2022	Contrast-enhanced Cpectral Mammography(CESM)	182	Sentinel lymph Node(SLN)	7 Radiomics features
Bong et al., 2023	2023	Ultrasound	1,024	None	None
Sharmin et al., 2023	2023	Histology images	None	None	Deep feature extraction
Beuque et al., 2023	2023	Mammography	850	None	Handcrafted radiomics
Liu et al., 2024	2024	MRI	1,760	None	7 Radiomics features
Wang J. et al., 2023	2023	Ultrasound	263	Ki-67	7 Radiomics features
Ferre et al., 2023	2023	Ultrasound	88	HER2	7 Radiomics features
Haraguchi et al., 2023	2023	Histology images	100	ALN	7 Radiomics features
Jailin et al., 2023	2023	Mammography	1673	None	None
Quan et al., 2023	2023	Ultrasound	445	None	7 Radiomics features
Murtas et al., 2023	2023	DBT	150	None	7 Radiomics features
Nicosia et al., 2023	2023	Ultrasound	365	None	7 Radiomics features
Chen W. et al., 2023	2023	PET/CT	236	None	Deep radiomics
Sun et al., 2023	2023	DCE-MRI	91	TNBC	7 Radiomics features
Wei Li et al., 2023	2023	DCE-MRI	1,048	NAC	7 Radiomics features
Xiang et al., 2023	2023	Histology images	8,797	None	None
Bangalore et al., 2024	2024	Histology images	None	None	Deep feature extraction
Romeo et al., 2023	2023	DCE-MRI	248	HER-2	7 Radiomics features
Li Y. et al., 2023	2023	DCE-MRI	95	None	7 Radiomics features
Caballo et al., 2023	2023	DCE-MRI	922	NAC	7 Radiomics features
Chen Y. et al., 2023	2023	DCE-MRI	479	ALN	Deep feature extraction
Oladimeji et al., 2024	2024	DBT	1,140	None	7 Radiomics features
Gamal et al., 2024	2024	DCE-MRI	109	NAC	7 Radiomics features
Jiang et al., 2024	2024	DCE-MRI	198	LVI	7 Radiomics features
Del Corso et al., 2024	2024	ABVS and DBT	66	None	7 Radiomics features

TABLE 2 Overview of feature extraction methods, extracted feature counts, and feature selection techniques used across reviewed studies, emphasizing the diverse approaches to data processing and analysis.

References	Feature extraction	Extracted features count	Feature selection
Yang et al., 2023	PyRadiomics	2,478	T-tests, Correlation analysis, LASSO
Abbasian Ardakani et al., 2023	PyRadiomics CNN	27 features for each Image	Kolmogorov-Smirnov test, <i>t</i> -test ROC analysis and curve
Wang C. et al., 2023	Annotation Tool		ResNet-18
Yu et al., 2023	Annotation Tool		DCNN
Wei et al., 2023	PyRadiomics	1,403	Spearman's rank correlation coefficient, LASSO
Zheng et al., 2023	PyRadiomics ResNet101	7,740	Boruta-shap Method, Univariate analysis Multivariate Cox analysis
Gao et al., 2023	PyRadiomics	1,409	ANOVA, LASSO
Xu et al., 2022	DenseNets		
Peng et al., 2022	PyRadiomics	851	LASSO
Vigil et al., 2022	Radiomics Deep Radiomics	354	Deep Neural Network Spectral embedding method Laplacian Eigenmaps
Rashid et al., 2022	PyRadiomics	75	
Chen et al., 2022	PyRadiomics	107 (Saha et al., 2018)	Multiple SVM classifier
Jiang et al., 2022	Radiomics Deep Radiomics	2,479	Intraclass correlation coefficient (ICC) Mann-Whitney U-test LASSO
Wang et al., 2022	PyRadiomics VGG-16	120	T-test LASSO
Cattell et al., 2022	PyRadiomics VGG-16	1,105	z-score Normalization Spearman, U-test Principal Component Analysis(PCA), LASSO
Wu et al., 2022	PyRadiomics ResNet-18	3,738	Correlation Analysis LASSO ANOVA
Bong et al., 2023	ResNet101	Deep features	Clinical factors and R-CNN
Sharmin et al., 2023	ResNet50	Deep Features (Janowczyk and Madabhushi, 2016)	Random forest and LightGBM
Beuque et al., 2023	Handcrafted radiomics ResNet101		
Liu et al., 2024	PyRadiomics	11,342	LASSO ANOVA
Wang J. et al., 2023	PyRadiomics	849	LASSO
Ferre et al., 2023	PyRadiomics		
Haraguchi et al., 2023	PyRadiomics		LASSO
Jailin et al., 2023	YOLO		
Quan et al., 2023	PyRadiomics	91	
Murtas et al., 2023	LIFEx	58	K-best, ANOVA, Sequential Embedded method, Random Forest
Nicosia et al., 2023	LIFEx		LASSO
Chen W. et al., 2023		Deep Features	Attention-based aggregate CNN
Sun et al., 2023	PyRadiomics	2,832	ANOVA Recursive feature elimination
Wei Li et al., 2023	PyRadiomics		
Xiang et al., 2023		Deep Features	DenseNet-121
Bangalore et al., 2024	Pruned CNN	550 (Janowczyk and Madabhushi, 2016 ; Balkenende et al., 2022)	Information Gain (IG) Supervised Relative Reduct (SRR) Gain-Ratio (GR), OneR mRMR Chi-Squared (CS)

(Continued)

TABLE 2 (Continued)

References	Feature extraction	Extracted features count	Feature selection
Romeo et al., 2023	PyRadiomics	1,288 (Saha et al., 2018)	ICC Analysis
Li Y. et al., 2023	PyRadiomics	1,130	LASSO
Caballo et al., 2023	PyRadiomics	348 (Saha et al., 2018)	U-test
Chen Y. et al., 2023	DenseNet121	1,000 (Saha et al., 2018)	mRMR LASSO SMOT
Oladimeji et al., 2024	PyRadiomic	857 (Buda et al., 2020)	Mutual information U Test
Gamal et al., 2024			Genetic algorithm (GA) Standardization
Jiang et al., 2024	PyRadiomics	6,195	mRMR LASSO
Del Corso et al., 2024	PyRadiomics	25	PCA Adaptive feature selection

two deep learning models like (Yu et al., 2023), leveraging VGG16 and ResNet50 models where authors use a split-sample validation approach rather than cross-validation. Specifically, author divides the data from three medical centers, using patients from two centers as a training cohort (420 patients) and patients from the third center as an external validation cohort (183 patients), attaining robust accuracy metrics highlighting deep learning capacity for breast cancer classification. Other famous models, such as Inception V3, DenseNet, and SqueezeNet, are frequently used by researchers. In contrast, traditional ML models are often used as benchmarks or in cases of limited data availability; for instance, Liu et al. (2024) utilize Support Vector Machine (SVM) and Gaussian Process models where authors employ a split-sample approach across five distinct cohorts to evaluate the model's robustness. They divide the data into training (775 patients) and validation (518 patients) cohorts for model development. For further testing, they use three independent testing cohorts: an internal retrospective cohort (167 patients), an internal prospective cohort (188 patients), and an external retrospective cohort (112 patients), reporting moderate sensitivity and specificity values. Ensemble methods, particularly XGBoost, are frequently implemented, with Quan et al. (2023) demonstrating XGBoost's impressive high specificity and accuracy. The authors apply a 4:1 split ratio, dividing patients into a training set (357 patients) and an independent test set (88 patients), highlighting the advantages of ensemble learning in achieving better generalizability. Studies using hybrid approaches, such as Nicosia et al. (2023), incorporate attention-based CNNs where authors apply 10-fold cross-validation to select radiomics features using the LASSO (Least Absolute Shrinkage and Selection Operator) logistic regression model. Additionally, they split the data into a training set (70%, with 255 lesions) and a test set (30%, with 110 lesions) to further validate model accuracy and robustness, achieving high AUC values. Moreover, studies such as those by Ferre et al. (2023) and Gamal et al. (2024) show the continued relevance of logistic regression and random forests, particularly when paired with feature selection techniques, to achieve competitive performance. In the study by Rashid et al.

(2022), CNN-SVM achieved the highest AUC of 0.974. Notably, this model also achieves higher accuracy (98.83%) compared to other models. Yang et al. (2023) achieved the highest sensitivity of 0.889 using 3DResNet. This model also demonstrates a remarkable level of lower specificity, with a value of 0.692. The study conducted by Wu et al. (2022) attained the nearest sensitivity of 0.88 utilizing a radiomics model. Accuracy is crucial in cases where detecting all positive instances, such as cancerous tumors, is vital. In a study by Rashid et al. (2022) and Bangalore et al. (2024), the CNN-SVM model and EfficientNet-Transformer models respectively achieved the highest accuracy of 0.9883 and 0.9884. Specificity is vital in situations where there is a chance that a negative prediction is truly negative. Cattell et al. (2022) demonstrated that VGG-16 achieved the highest level of specificity, with a score of 0.87. However, the AUC value for this study is 0.83, which is notably higher than other models.

3.5 Literature methodology analysis

The primary objective of the study will be the detailed analysis of the methods sections of 16 selected papers. We evaluated and studied different research designs for our examine the methodology used in recent literature. This approach will ensure the exploration of each methodology, developing a more detailed understanding of the research process. Choosing an appropriate image dataset is usually the first step in research methodology where researcher typically use one specific type of image modality. After doing several image processing steps the features are extracted using radiomics models or deep learning models. Then, the next stage will be selecting the best features using several appropriate statistical models. Finally, different DL/ML models are proposed or evaluated, either individually or in combination. The authors of those studies (Yang et al., 2023; Wu et al., 2022; Abbasian Ardakani et al., 2023; Gao et al., 2023) implemented new deep learning or radiomics models in their research. Yang

TABLE 3 Performance analysis of DL/ML models across studies, including metrics such as sensitivity, specificity, accuracy, and area under the curve (AUC), highlighting variations in model effectiveness and application.

References	ML/DL	Sensitivity	Specificity	Accuracy	Area under curve (AUC)
Yang et al., 2023	3DResNet	0.889	0.692	0.743	0.857
Abbasian Ardakani et al., 2023	SqueezeNet, InceptionV3 VGG16, VGG19				0.861, 0.843 0.847, 0.867
Wang C. et al., 2023	ResNet18	DL-grayscale - 0.82, DL-CDFI - 0.77 DL-elastography - 0.88	DL-grayscale - 0.84, DL-CDFI - 0.71 DL-elastography - 0.82	DL-grayscale - 0.83, DL-CDFI - 0.76 DL-elastography - 0.86	DL-grayscale - 0.85, DL-CDFI - 0.76 DL-elastography - 0.87
Yu et al., 2023	ResNet34, ResNet50 VGG16, DenseNet121	ResNet34 - 0.788, ResNet50 - 0.773 VGG16 - 0.606, DenseNet121 - 0.697	ResNet34 - 0.684, ResNet50 - 0.855 VGG16 - 0.786, DenseNet121 - 0.684	ResNet34 - 0.721, ResNet50 - 0.825 VGG16 - 0.716, DenseNet121 - 0.689	ResNet34 - 0.785, ResNet50 - 0.879 VGG16 - 0.707, DenseNet121 - 0.724
Wei et al., 2023	ResNet50, ResNet101 Inception_v3, VGG19	ResNet50 - 0.71, ResNet101 - 0.57 Inception_v3 - 0.71, VGG19 - 0.71	ResNet50 - 0.67, ResNet101 - 0.80 Inception_v3 - 0.74, VGG19 - 0.49	ResNet50 - 0.72, ResNet101 - 0.76 Inception_v3 - 0.72, VGG19 - 0.83	ResNet50 - 0.69, ResNet101 - 0.69 Inception_v3 - 0.71, VGG19 - 0.58
Zheng et al., 2023	ResNet101				0.8595
Gao et al., 2023	ResNet	0.792	0.853	0.828	0.852
Xu et al., 2022	DenseNets	0.727	0.84	0.805	0.84
Peng et al., 2022	ResNext50	0.566	0.556	0.558	0.554
Vigil et al., 2022	Unet			0.853	
Rashid et al., 2022	Inception_V3, CNN-SVM			Inception_V3 - 0.9196, CNN-SVM - 0.9883	Inception_V3 - 0.932, CNN-SVM - 0.974
Chen et al., 2022	ResNet50	0.68	0.78	0.74	0.82
Jiang et al., 2022	ResNet50	0.864	0.500	0.500	0.714
Wang et al., 2022	VGG-16			0.78	0.87
Cattell et al., 2022	VGG-16	0.58	0.87	0.77	0.83
Wu et al., 2022	Radiomics Model	0.88	0.68	0.75	0.85
Bong et al., 2023	Mask R-CNN ResNet101	0.78	0.85		
Sharmin et al., 2023	ResNet50V2			0.95	
Beuque et al., 2023	Mask R-CNN ResNet101	0.90	0.86		0.95
Liu et al., 2024	Gaussian Process Support vector machine(SVM)			0.84	
Wang J. et al., 2023	Logistic regression Random forests Support vector machine Extreme gradient boosting (XGBoost)				0.88
Ferre et al., 2023	Logistic regression k-nearest neighbor Näve Bayes	0.818	0.742		0.824
Haraguchi et al., 2023	Support vector machine				0.765
Jailin et al., 2023	YOLO				0.964
Quan et al., 2023	Logistic regression Random forests Support vector machine Extreme gradient boosting (XGBoost)	0.991	0.996		
Murtas et al., 2023	Random forest			0.72	

(Continued)

TABLE 3 (Continued)

References	ML/DL	Sensitivity	Specificity	Accuracy	Area under curve (AUC)
Nicosia et al., 2023	Multivariable logistic				0.914
Chen W. et al., 2023	Attention-based aggregate CNN	0.809	0.85	0.83	0.886
Sun et al., 2023	Support vector machine			0.82	0.85
Wei Li et al., 2023	Logistic regression (LR) Linear discriminant analysis (LDA) Support vector machine (SVM)			0.843	0.893
Xiang et al., 2023	Attention-based CNN	0.864	0.866	0.865	0.95
Bangalore et al., 2024	Triplet attention-based Efficient network (TAENet)	0.9718	0.9812	0.9884	
Romeo et al., 2023	Logistic regression	0.80	0.43	0.63	0.66
Li Y. et al., 2023	Deep learning radiomic (DLR) Logistic regression	0.875	0.90	0.893	0.914
Caballo et al., 2023	Logistic regression model with leave-one-out cross-validation (LOOCV)				0.707
Chen Y. et al., 2023	DenseNet121	0.65	0.80	0.75	0.80
Oladimeji et al., 2024	Random forest Extreme gradient boosting (XGBoost) Logistic regression Decision Tree			0.92	
Gamal et al., 2024	LightGBM Extreme gradient boosting (XGBoost)	0.88	0.98	0.88	
Jiang et al., 2024	Logistic regression (LR) Support vector machine (SVM) Classification and Regression Trees (CART) K-Nearest Neighbors (KNN) Gradient Boosting Machine (GBM)	0.90	0.72	0.87	
Del Corso et al., 2024	Logistic regression (LR) Support vector machine (SVM) Logistic regression	0.842	0.807	0.807	0.921

[et al. \(2023\)](#) introducing three innovative deep-learning models that leverage the power of radiomics: DL-gray scale, DL-CDFI, and DL-elastography. [Wu et al. \(2022\)](#) developed two radiomics nomograms capable of reliably showing the presence or absence of NSLN metastasis and the extent of axillary tumoral burden. A new filter based on deep learning and adaptive residual learning has been proposed by [Abbasian Ardakani et al. \(2023\)](#). [Gao et al. \(2023\)](#) explore the capabilities of their attention-based DL model in distinguishing ALN metastasis in breast cancer using dynamic contrast-enhanced MRI (DCE-MRI) before surgery. The remaining authors of the papers utilize various cancer bio-markers to make predictions about breast cancer. Authors employ feature

extraction and selection model that is based on radiomics and deep learning. The authors also conducted various statistical analyses to select features. To develop their DL/ML model, authors created separate models for radiomics and deep learning. Furthermore, several of them utilized radiomics and deep learning features to train DL/ML models. In the end, the authors provided an affirmation of accuracy by performing both internal and external validation techniques. Research from [Jiang et al. \(2022\)](#) exploited radiomics and deep learning features from different tumor areas. These features were then sequentially selected by LASSO regression and finally produced radiomics signatures. AUROC, accuracy, sensitivity, and specificity were calculated to estimate

the radiomics signatures. After doing extensive research on 40 articles we developed a widely used methodology which is shown in Figure 6.

3.6 Risk of bias

The quality assessment of 16 publications was assessed using QUADAS-2. We prioritized studies that provided high-quality and comprehensive data on patient selection and target classification for breast cancer to ensure the robustness and applicability of our findings. The selected studies represented the most relevant and rigorous evidence available for quality assessment. For the risk of bias, there is a higher number of unclear answers in the reference standard selection due to the specific modality used. The flow and timing of high-risk responses may suggest different reference standards being used and ambiguous intervals with index tests. There are some concerns regarding the applicability of the reference standard due to the use of a particular modality, which has resulted in a higher number of unclear answers. The increased risk of bias in index tests primarily stems from the design of the validation process. There is some ambiguity when it comes to patient selection due to the lack of clear inclusion and exclusion criteria. Figure 7 illustrates the process of quality assessment, utilizing the robvis tool (McGuinness and Higgins, 2021).

3.7 Correlation coefficient

Correlation coefficients are thus pivotal in the distinction between correlation types, revealing whether two variables show synchronized movements, diverge in opposite directions, or are completely unrelated. By adopting the random effect model, we endeavor to find out the values of the correlation coefficient (n , r) between the sample size and sensitivity in our meta-analysis study.

This meta-analysis investigates the correlation coefficients across 23 studies using a random-effects model. Fisher's r -to- z transformed correlation coefficient is used as the primary outcome measure, allowing for a standardized effect size across studies with varying sample sizes. The model uses the restricted maximum likelihood (REML) estimator for study-level variability. The Table 4 and Figure 8 describe our meta analysis. The random-effects model estimated an average effect size of 1.20 with a standard error of 0.0869. A Z-test yielded a value of 13.8, with a p -value of <0.001 , indicating statistical significance. The 95% confidence interval (CI) for the effect size ranges from 1.033 to 1.373, suggesting a robust, statistically significant association. Tau τ , representing the standard deviation of true effect sizes, is estimated at 0.412, and Tau² τ^2 , the variance of effect sizes, is 0.1694 (with a standard error of 0.0524). I^2 is very high at 98.63%, indicating substantial heterogeneity across studies. This indicates that almost all variability in observed effect sizes is due to differences between studies rather than random chance. The Q-test for heterogeneity also supports this high level of heterogeneity, with a Q value of 1,574.190 ($df = 22$) and a p -value < 0.001 . These metrics imply that the included studies are highly heterogeneous, and the overall effect size should be interpreted with caution, as it may not represent a single underlying population

effect. The funnel plot shows asymmetry, suggesting the potential for publication bias or small-study effects. Studies appear unevenly distributed around the mean effect size, especially toward the left side, which might indicate an underrepresentation of studies with small or null effects. The forest plot displays individual effect sizes for each study along with their 95% confidence intervals. Effect sizes range considerably across studies, from approximately 0.5 to 2.7, reinforcing the high heterogeneity ($I^2 = 98.63\%$). The pooled effect size obtained from the random-effects model is 1.20, with a 95% confidence interval of 1.03 to 1.37, displayed at the bottom of the forest plot. Log-likelihood, Deviance, AIC, BIC, and AICc values are reported for both the maximum-likelihood and restricted maximum-likelihood models. The REML model has a slightly better fit, with lower values across these criteria (AIC = 27.743, BIC = 29.925, AICc = 28.374), suggesting it is more appropriate for this data due to better handling of the high heterogeneity. The meta-analysis reveals a significant average effect size, suggesting a meaningful correlation across studies. However, the high heterogeneity, as evidenced by I^2 and the Q-test, indicates that effect sizes vary widely among the studies, possibly due to differences in study populations, methodologies, or contextual factors. The presence of funnel plot asymmetry further suggests the possibility of publication bias, which should be considered when interpreting these findings.

4 Discussion

In the discussion section of this systematic review and meta-analysis, we explore the detailed insights derived from the comprehensive body of evidence examined. By carefully analyzing the data, we aim to present an in-depth understanding of the research field and learn the research patterns.

After surveying the studies, DL/ML technology shows remarkable accuracy in forecasting results. Through an in-depth and methodical review, we have assessed the efficacy of existing techniques, identified promising areas for future research, and acquired valuable insights into their accuracy. The radiomics-guided DL/ML model has shown promising potential for improving accuracy. In some instances, the radiomics model alone has shown the ability to achieve the highest accuracy in identifying breast cancer. In our research, 12,685 patients were enrolled, and the authors used several strategies to identify these individuals. Since our review question primarily pertains to image processing methods using radiomics-guided deep learning models, we will only concentrate on image processing using radiomics-guided deep learning models. Ultrasound is a widely used tool for imaging cancer biomarkers. Sentinel lymph nodes (SLN) and axillary lymph nodes (ALN) are mainly used as biomarkers. Most of the articles we picked used seven radiomics attributes for radiomics analysis. The radiomics characteristics are extracted using the pyradiomics python package. Deep learning models are famous for extracting deep learning information from images. Following the extraction process, writers often use LASSO statistical analysis for feature selection. In addition, the researchers used several statistical analyses, such as U-test, T-test, and correlation coefficient, in their study. The authors used many statistical approaches

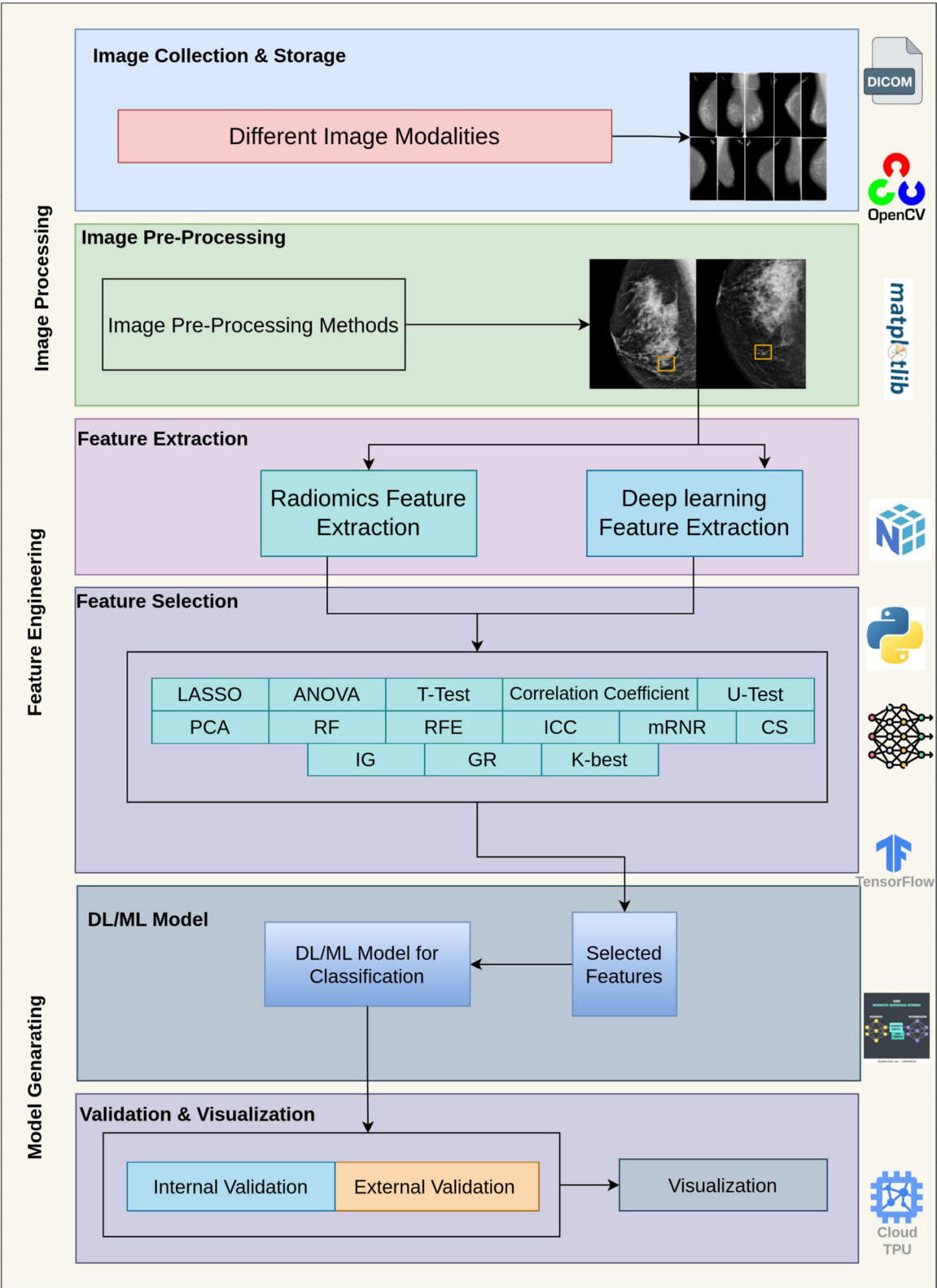


FIGURE 6
Workflow for medical image analysis: image collection, pre-processing, feature extraction, feature selection, model generation, and validation.

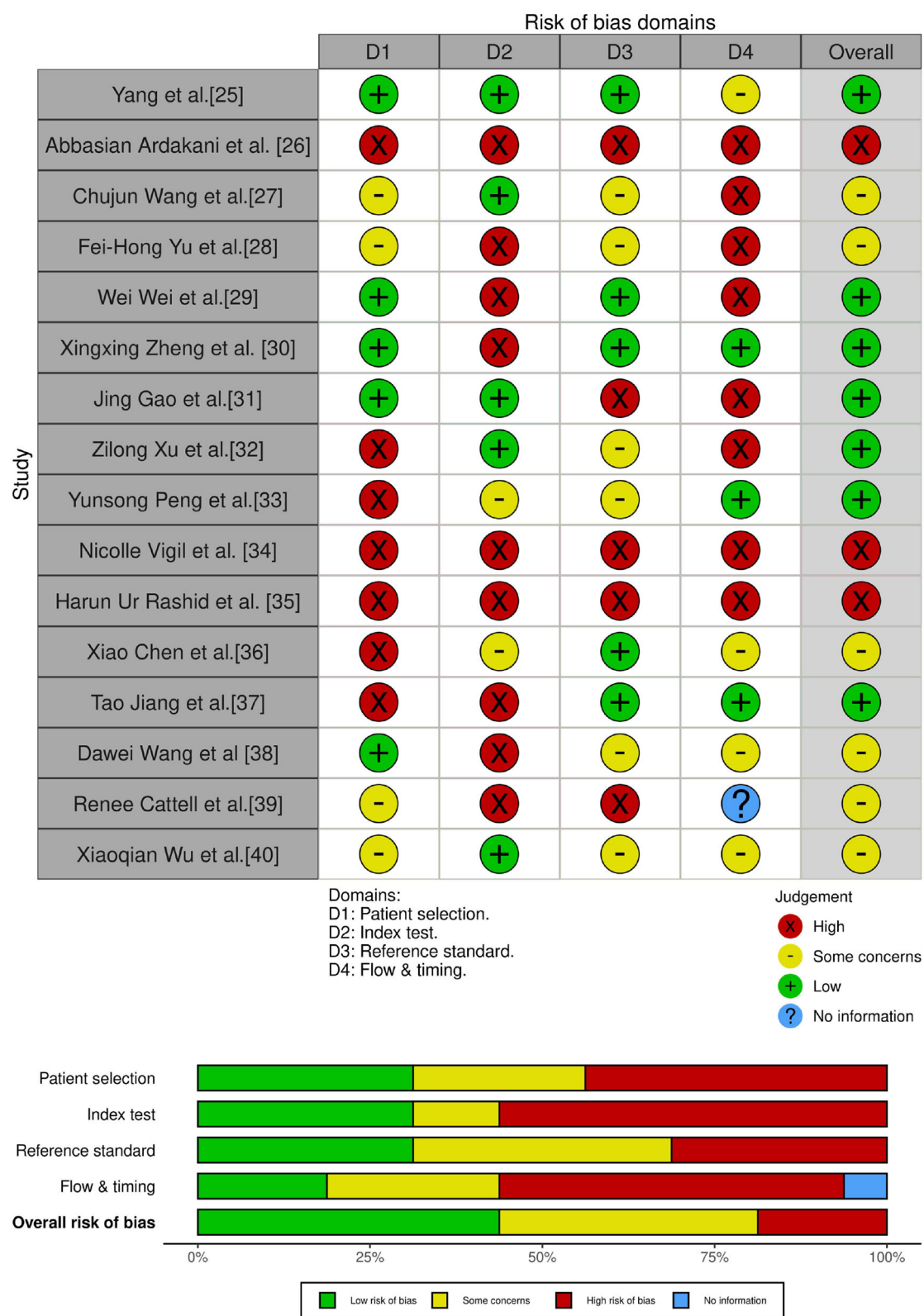


FIGURE 7
Quality assessment according To QUADAS-2.

collectively for their investigations. Our research revealed that authors independently use statistical approaches for selection in their radiomics and deep learning models. When generating DL/ML models, authors often use various variations of ResNet, including ResNet 18, ResNet 34, ResNet 50, and ResNet 101. Even though the CNN-based model achieved the highest AUC of

TABLE 4 Random-effects model summary (R = 2.5, K = 23) with heterogeneity statistics and model fit metrics for ML and REML methods.

	Estimate	se	Z	p	CI lower bound	CI upper bound
Random-effects model (k = 23)						
Intercept	1.14	0.0869	13.8	0.001	1.033	1.373
Tau	τ^2	I^2	H^2	IR^2	df	Q
Heterogeneity Statistics						
0.412	0.1694 (SE = 0.00524)	98.63%	72.766	-	22.000	1,574.190
	Log-likelihood	Deviance	AIC	BIC	AICc	
Model Fit Statistics and information Criteria						
Maximum-Likelihood		-11.915	113.448	27.831	30.102	28.431
Restricted Maximum-Likelihood		-11.871	23.743	27.743	29.925	28.374

97.4%, with the same model, we achieved an impressive accuracy of 98.83%.

A previous systematic review conducted by Taddese and Tilahun (2024) consisted of 48 studies, while their meta-analysis comprised 24. The studies utilized various images and models to diagnose various gynecological cancers. The authors emphasized that DL algorithms demonstrated higher sensitivity but lower specificity than machine learning (ML) methods.

This systematic review thoroughly examined the topic using a well-defined methodology and strict inclusion criteria. We also customized the quality assessment tools to suit the included studies. We explore image feature extraction and selection methods, utilize radiomics and deep learning models, and compare their performance. It is worth noting that previous studies and current guidelines have highlighted the importance of internal validation. This involves training and validating models using the same dataset, often through techniques like cross-validation. However, it is essential to be cautious with the results obtained through internal validation, as they tend to overestimate accuracy and may not be easily generalized due to overfitting. Therefore, only studies that utilized external validation of test sets were considered during the initial phase of literature identification. Thus, our research offers valuable insights into the use of DL for diagnosing breast cancer. However, this systematic review did not include publications in languages other than English, which could introduce bias in the selection process. Furthermore, the lack of sufficient data hindered the calculation of comprehensive diagnostic measures.

We recommend conducting more precise research on feature extraction and selection based on radiomics and deep learning models. After extracting the features, we should find the most valuable features by conducting statistical analysis using both categories of features combined to obtain more accurate prospective studies. We recommend using externally validated data to conduct a more thorough assessment of the DL/ML model for both ruptured and unruptured aneurysms. To ensure that the results of this promising technology can be replicated and applied to a broader range of cases, we suggest developing standardized research guidelines for further investigations.

This systematic review investigated the potential of utilizing radiomics-guided deep learning/machine learning models to identify breast cancer. The studies provided encouraging findings,

as specific models showed impressive accuracy in distinguishing between malignant and benign breast tumors. Nevertheless, there is a wide range of variations in the designs of studies, architectures of models, and techniques used for validation. In literature, a variety of imaging methods are employed. Upon evaluation, we found that the most frequent imaging modality is ultrasound imaging. In addition, DBT, CT/PET, and DCI-MRI are utilized for early identification of breast cancer. Extracting features mainly combines radiomics and deep learning feature extraction methods. The Pyradiomics Python package extracts Radiomics characteristics, and the most used deep learning model for extracting medical image features is ResNet. In this review, we observed that while there is a considerable use of other statistical models such as T-test, ANOVA, and correlation analysis, researchers mainly utilize LASSO for feature selection. The most popular deep-learning models for classifying breast cancer are ResNet and VGG. Additional research is necessary to establish uniform techniques, enhance applicability, and explore the practical implications of these models. Future research in radiomics-guided deep learning (DL) and machine learning (ML) for breast cancer detection should prioritize several key areas to build upon the promising yet varied findings highlighted in this review. First, standardized model development, validation, and evaluation guidelines are crucial. The wide variability in model architectures, feature extraction techniques, and validation methods across studies has led to inconsistent performance metrics, challenging the generalizability of findings. Establishing a common framework will allow researchers to compare results more effectively and ensure that the models developed are robust, reproducible, and clinically applicable. Second, the future of our research should be built on a foundation of external validation through multi-center prospective trials. Many current studies rely on internal validation, which may introduce overfitting and overestimate model accuracy. It's urgent that we evaluate model performance in real-world clinical environments by conducting trials across diverse populations and imaging settings. This step is crucial in increasing confidence in the models' diagnostic accuracy and generalizability, and it's a key part of our journey toward more reliable and effective breast cancer detection. Additionally, the field would benefit from further investigation into optimal feature selection techniques that combine radiomics and DL features. Current methods, such

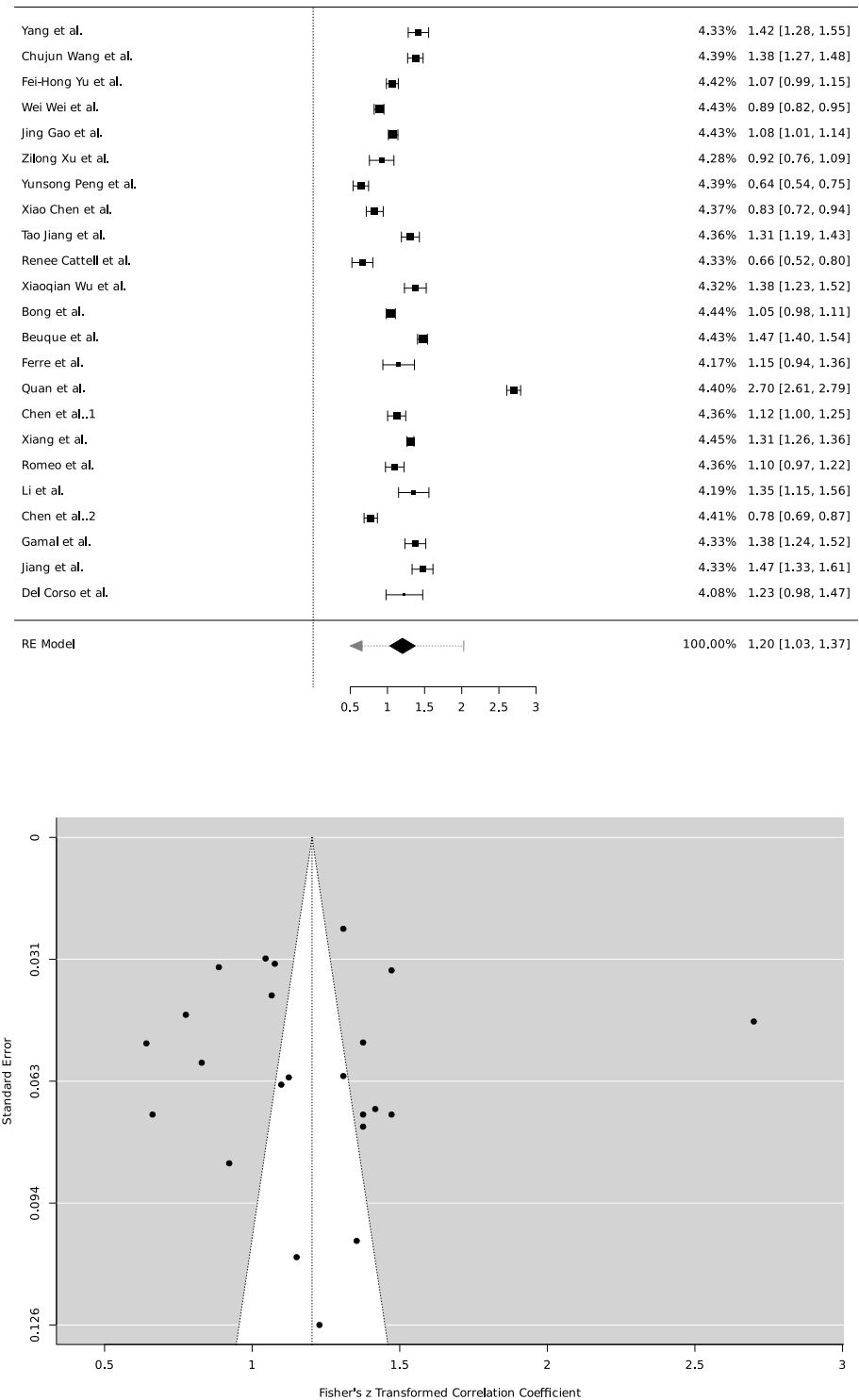


FIGURE 8 Forest plot and funnel plot presenting the meta-analysis results. The forest plot illustrates individual study effect sizes and 95% confidence intervals, with a pooled effect size of 1.20 (95% CI: 1.03, 1.37) obtained using a random-effects model. The funnel plot indicates an asymmetric distribution, suggesting potential publication bias or small-study effects.

as LASSO, U-test, and *T*-test, show promise, but additional methods that integrate both categories of features could enhance the predictive power of DL/ML models. Exploring new feature selection algorithms or hybrid approaches could yield insights into the most predictive attributes for distinguishing between benign and malignant tumors. Lastly, our research should explore the practical application of these advanced imaging techniques with existing diagnostic workflows. We should study the potential

roles of DL/ML models not as standalone diagnostic methods, but as complementary tools. This approach could facilitate their practical application in early breast cancer detection, monitoring, and treatment planning, making them an integral part of our clinical workflows.

5 Conclusion

This systematic review investigated the potential of utilizing radiomics-guided deep learning/machine learning models to identify breast cancer. The studies provided encouraging findings, as certain models showed impressive accuracy in distinguishing between malignant and benign breast tumor. Nevertheless, there is a wide range of variations in the designs of studies, architectures of models, and techniques used for validation. In literature, a variety of imaging methods are employed. Upon evaluation, we found that the most frequent imaging modality is ultrasound imaging. In addition, DBT, CT/PET, and DCI-MRI are utilized for early identification of breast cancer. Extracting features are mostly done with the combination of radiomics and deep learning feature extraction methods. The Pyradiomics Python package is used to extract Radiomics characteristics, and the most used deep learning model for extracting medical image features is ResNet. In this review, we observed that while there is a considerable use of other statistical models such as *T*-test, ANOVA, and correlation analysis, researchers mostly utilize LASSO for feature selection. The most popular deep learning models for classifying breast cancer are ResNet and VGG. The review identifies notable challenges, such as variability in model architectures, feature selection techniques, and validation approaches across studies, which have led to inconsistencies in model performance and generalizability. Although widely used, internal validation methods are limited in assessing accurate diagnostic accuracy due to the risk of overfitting. The recommendation is to prioritize external validation through multi-center, prospective trials, enabling more accurate and generalizable assessments and supporting broader clinical applicability. Furthermore, combining radiomics and DL/ML features through optimized feature selection techniques, such as LASSO, and exploring hybrid approaches could enhance model precision. Standardizing model development, validation, and evaluation protocols is essential to improve the comparability and reliability of findings across studies. Looking forward, radiomics-guided DL/ML models show great promise as complementary diagnostic tools rather than standalone methods, potentially enhancing early breast cancer detection, monitoring, and treatment planning. However, the findings underscore the need for standardized guidelines, external validation, and more rigorous prospective studies to realize the full potential of these models

in clinical settings. This review provides a foundation for future research, which should continue to refine and integrate DL/ML methodologies into diagnostic pathways, ultimately improving patient outcomes in breast cancer care.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

Author contributions

NM: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. AB: Supervision, Funding acquisition, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The Deanship of Scientific Research (DSR) at King Abdu-laziz University in Jeddah, Saudi Arabia, provided funding for this project under grant number (GPIP: 1821-611-2024). The authors therefore acknowledge with thanks DSR for technical and financial support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Abbasian Ardakani, A., Mohammadi, A., Vogl, T. J., Kuzan, T. Y., and Acharya, U. R. (2023). AdaRes: a deep learning-based model for ultrasound image denoising: results of image quality metrics, radiomics, artificial intelligence, and clinical studies. *J. Clin. Ultrasound* 52, 131–143. doi: 10.1002/jcu.23607
- Abhisheka, B., Biswas, S. K., Purkayastha, B., Das, D., and Escargueil, A. (2023). Recent trend in medical imaging modalities and their applications in disease diagnosis: a review. *Multim. Tools Applic.* 83, 43035–43070. doi: 10.1007/s11042-023-17326-1

- Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., et al. (2022). Current and future burden of breast cancer: global statistics for 2020 and 2040. *Breast* 66, 15–23. doi: 10.1016/j.breast.2022.08.010
- Asmussen, C. B., and Möller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review. *J. Big Data* 6:93. doi: 10.1186/s40537-019-0255-7
- Balkenende, L., Teuwen, J., and Mann, R. M. (2022). Application of deep learning in breast cancer imaging. *Semin. Nucl. Med.* 52, 584–596. doi: 10.1053/j.semnucmed.2022.02.003
- Bangalore, M., Hiremath, B. S., and Polnaya, A. M. (2024). Classification of breast cancer using ensemble filter feature selection with triplet attention based efficient net classifier. *Int. Arab J. Inf. Technol* 21, 17–31. doi: 10.34028/iajit/21/1/2
- Beuque, M. P., Lobbes, M. B., van Wijk, Y., Widaatalla, Y., Primakov, S., Majer, M., et al. (2023). Combining deep learning and handcrafted radiomics for classification of suspicious lesions on contrast-enhanced mammograms. *Radiology* 307:221843. doi: 10.1148/radiol.221843
- Bong, J. H., Kim, T. H., and Jeong, S. (2023). Deep learning model for the diagnosis of breast cancers smaller than 1 cm with ultrasonography: integration of ultrasonography and clinical factors. *Quant. Imaging Med. Surg.* 13, 2486–2495. doi: 10.21037/qims-22-880
- Buda, M., Saha, A., Walsh, R., Ghate, S., Li, N., Swiecicki, A., et al. (2020). *Breast cancer screening digital breast tomosynthesis (bcs-dbt)*. Data obtained from Duke Health System.
- Caballo, M., Sanderink, W. B., Han, L., Gao, Y., Athanasiou, A., and Mann, R. M. (2023). Four-dimensional machine learning radiomics for the pretreatment assessment of breast cancer pathologic complete response to neoadjuvant chemotherapy in dynamic contrast-enhanced MRI. *J. Magn. Reson. Imag.* 57, 97–110. doi: 10.1002/jmri.28273
- Cattell, R., Ying, J., Lei, L., Ding, J., Chen, S., Serrano Sosa, M., et al. (2022). Preoperative prediction of lymph node metastasis using deep learning-based features. *Visual Comput. Ind. Biomed. Art* 5:8. doi: 10.1186/s42492-022-00104-5
- Cè, M., D'Amico, N. C., Danesini, G. M., Foschini, C., Oliva, G., Martinenghi, C., et al. (2023). Ultrasound elastography: basic principles and examples of clinical applications with artificial intelligence—a review. *BioMedInformatics* 3, 17–43. doi: 10.3390/biomedinformatics3010002
- Chen, W., Liu, F., Wang, R., Qi, M., Zhang, J., Liu, X., et al. (2023). End-to-end deep learning radiomics: development and validation of a novel attention-based aggregate convolutional neural network to distinguish breast diffuse large B-cell lymphoma from breast invasive ductal carcinoma. *Quant. Imaging Med. Surg.* 13, 6598–6614. doi: 10.21037/qims-22-1333
- Chen, X., Zhang, Y., Zhou, J., Wang, X., Liu, X., Nie, K., et al. (2022). Diagnosis of architectural distortion on digital breast tomosynthesis using radiomics and deep learning. *Front. Oncol.* 12:991892. doi: 10.3389/fonc.2022.991892
- Chen, Y., Wang, L., Dong, X., Luo, R., Ge, Y., Liu, H., et al. (2023). Deep learning radiomics of preoperative breast MRI for prediction of axillary lymph node metastasis in breast cancer. *J. Digit. Imaging* 36, 1323–1331. doi: 10.1007/s10278-023-00818-9
- Del Corso, G., Germanese, D., Cudai, C., Anastasi, G., Belli, P., Formica, A., et al. (2024). Adaptive machine learning approach for importance evaluation of multimodal breast cancer radiomic features. *J. Imag. Inform. Med.* 37, 1642–1651. doi: 10.1007/s10278-024-01064-3
- Dhar, T., Dey, N., Borra, S., and Sherratt, R. S. (2023). Challenges of deep learning in medical image analysis improving explainability and trust. *IEEE Trans. Technol. Soc.* 4, 68–75. doi: 10.1109/TTTS.2023.3234203
- Dittonno, F., Franco, A., Manfredi, C., Vecchia, A., Valerio, M., Bukavina, L., et al. (2023). Novel non-MRI imaging techniques for primary diagnosis of prostate cancer: micro-ultrasound, contrast-enhanced ultrasound, elastography, multiparametric ultrasound, and PSMA PET/CT. *Prostate Cancer Prostatic Dis.* 27, 29–36. doi: 10.1038/s41391-023-00708-9
- Ferre, R., Elst, J., Senthilnathan, S., Lagree, A., Tabbarah, S., Lu, F. I., et al. (2023). Machine learning analysis of breast ultrasound to classify triple negative and HER2+ breast cancer subtypes. *Breast Dis.* 42, 59–66. doi: 10.3233/BD-220018
- Gamal, A., Sharafeldien, A., Alnaghy, E., Alghandour, R., Saleh Alghamdi, N., Ali, K. M., et al. (2024). A novel machine learning approach for predicting neoadjuvant chemotherapy response in breast cancer: integration of multimodal radiomics with clinical and molecular subtype markers. *IEEE Access* 12, 104983–105003. doi: 10.1109/ACCESS.2024.3432459
- Gao, J., Zhong, X., Li, W., Li, Q., Shao, H., Wang, Z., et al. (2023). Attention based deep learning for the preoperative differentiation of axillary lymph node metastasis in breast cancer on DCE-MRI. *J. Magn. Reson. Imag.* 57, 1842–1853. doi: 10.1002/jmri.28464
- Gatta, G., Somma, F., Sardù, C., De Chiara, M., Massafra, R., Fanizzi, A., et al. (2023). Automated 3D ultrasound as an adjunct to screening mammography programs in dense breast: literature review and meta-analysis. *J. Pers. Med.* 13:1683. doi: 10.3390/jpm13121683
- Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., and Lew, M. S. (2016). Deep learning for visual understanding: a review. *Neurocomputing* 187, 27–48. doi: 10.1016/j.neucom.2015.09.116
- Gupta, H., Singh, H., and Kumar, A. (2024). Texture and radiomics inspired data-driven cancerous lung nodules severity classification. *Biomed. Signal Process. Control* 88:105543. doi: 10.1016/j.bspc.2023.105543
- Ha, S. M., Yi, A., Yim, D., Jang, M.-J., Kwon, B. R., Shin, S. U., et al. (2023). Digital breast tomosynthesis plus ultrasound versus digital mammography plus ultrasound for screening breast cancer in women with dense breasts. *Korean journal of radiology* 24, 274–283. doi: 10.3348/kjr.2022.0649
- Haraguchi, T., Kobayashi, Y., Hirahara, D., Kobayashi, T., Takaya, E., Nagai, M. T., et al. (2023). Radiomics model of diffusion-weighted whole-body imaging with background signal suppression (DWIBS) for predicting axillary lymph node status in breast cancer. *J. Xray. Sci. Technol.* 31, 627–640. doi: 10.3233/XST-230009
- Hornik, K. (2012). *The Comprehensive R Archive Network*. John Wiley & Sons, Ltd, 394–398. doi: 10.1002/wics.1212
- Ingrisch, M., and Sourbron, S. (2013). Tracer-kinetic modeling of dynamic contrast-enhanced MRI and CT: a primer. *J. Pharmacokinet. Pharmacodyn.* 40, 281–300. doi: 10.1007/s10928-013-9315-3
- Jailin, C., Mohamed, S., Iordache, R., Milioni De Carvalho, P., Ahmed, S. Y., Abdel Sattar, E. A., et al. (2023). AI-based cancer detection model for contrast-enhanced mammography. *Bioengineering* 10:974. doi: 10.3390/bioengineering10080974
- Jamovi (Version 2.5). (2024). *The jamovi project*. Available online at: <https://www.jamovi.org>
- Janowczyk, A., and Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J. Pathol. Inform.* 7. doi: 10.4103/2153-3539.186902
- Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., et al. (2019). Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimed. Tools Appl.* 78, 15169–15211. doi: 10.1007/s11042-018-6894-4
- Jiang, T., Jiang, W., Chang, S., Wang, H., Niu, S., Yue, Z., et al. (2022). Intratumoral analysis of digital breast tomosynthesis for predicting the Ki67 level in breast cancer: a multi-center radiomics study. *Med. Phys.* 49, 219–230. doi: 10.1002/mp.15392
- Jiang, Y., Zeng, Y., Zuo, Z., Yang, X., Liu, H., Zhou, Y., et al. (2024). Leveraging multimodal MRI-based radiomics analysis with diverse machine learning models to evaluate lymphovascular invasion in clinically node-negative breast cancer. *Heliyon* 10:e23916. doi: 10.1016/j.heliyon.2023.e23916
- Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S. A., Schabath, M. B., et al. (2012). Radiomics: the process and the challenges. *Magn. Reson. Imag.* 30, 1234–1248. doi: 10.1016/j.mri.2012.06.010
- Lei, S., Zheng, R., Zhang, S., Wang, S., Chen, R., Sun, K., et al. (2021). Global patterns of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020. *Cancer Commun.* 41, 1183–1194. doi: 10.1002/cac2.12207
- Li, X., Li, M., Yan, P., Li, G., Jiang, Y., Luo, H., et al. (2023). Deep learning attention mechanism in medical image analysis: basics and beyonds. *Int. J. Netw. Dyn. Intell.* 2, 93–116. doi: 10.53941/ijndi0201006
- Li, Y., Fan, Y., Xu, D., Li, Y., Zhong, Z., Pan, H., et al. (2023). Deep learning radiomic analysis of DCE-MRI combined with clinical characteristics predicts pathological complete response to neoadjuvant chemotherapy in breast cancer. *Front. Oncol.* 12:1041142. doi: 10.3389/fonc.2022.1041142
- Liu, Y., Jia, X., Zhao, J., Peng, Y., Yao, X., Hu, X., et al. (2024). A machine learning-based unenhanced radiomics approach to distinguishing between benign and malignant breast lesions using T2-weighted and diffusion-weighted MRI. *J. Magn. Reson. Imag.* 60, 600–612. doi: 10.1002/jmri.29111
- McGuinness, L. A., and Higgins, J. P. T. (2021). Risk-of-bias visualization (robvis): An R package and shiny web app for visualizing risk-of-bias assessments. *Res. Synth. Methods* 12, 55–61. doi: 10.1002/jrsm.1411
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 339:b2535. doi: 10.1136/bmj.b2535
- Murtas, F., Landoni, V., Ordóñez, P., Greco, L., Ferranti, F. R., Russo, A., et al. (2023). Clinical-radiomic models based on digital breast tomosynthesis images: a preliminary investigation of a predictive tool for cancer diagnosis. *Front. Oncol.* 13:1152158. doi: 10.3389/fonc.2023.1152158
- Nicosia, L., Pesapane, F., Bozzini, A. C., Latronico, A., Rotili, A., Ferrari, F., et al. (2023). Prediction of the malignancy of a breast lesion detected on breast ultrasound: radiomics applied to clinical practice. *Cancers* 15:964. doi: 10.3390/cancers15030964
- Oladimeji, O. O., Ayaz, H., McLoughlin, I., and Unnikrishnan, S. (2024). Mutual information-based radiomic feature selection with SHAP explainability for breast cancer diagnosis. *Results Eng.* 24:103071. doi: 10.1016/j.rineng.2024.103071
- Peng, J., Wang, W., Song, Q., Hou, J., Jin, H., Qin, X., et al. (2023). 18f-fdg-pet radiomics based on white matter predicts the progression of mild cognitive impairment to Alzheimer disease: a machine learning study. *Acad. Radiol.* 30, 1874–1884. doi: 10.1016/j.acra.2022.12.033

- Peng, Y., Cheng, Z., Gong, C., Zheng, C., Zhang, X., Wu, Z., et al. (2022). Pretreatment DCE-MRI-based deep learning outperforms radiomics analysis in predicting pathologic complete response to neoadjuvant chemotherapy in breast cancer. *Front. Oncol.* 12:846775. doi: 10.3389/fonc.2022.846775
- Quan, M. Y., Huang, Y. X., Wang, C. Y., Zhang, Q., Chang, C., and Zhou, S. C. (2023). Deep learning radiomics model based on breast ultrasound video to predict HER2 expression status. *Front. Endocrinol.* 14:1144812. doi: 10.3389/fendo.2023.1144812
- Rashid, H. U., Ibricki, T., Payda, S., Binokay, F., and Çevik, U. (2022). Analysis of breast cancer classification robustness with radiomics feature extraction and deep learning techniques. *Expert Syst.* 39, 1–12. doi: 10.1111/exsy.13018
- Rizzo, S., Botta, F., Raimondi, S., Origgi, D., Fanciullo, C., Morganti, A. G., et al. (2018). Radiomics: the facts and the challenges of image analysis. *Eur. Radiol. Exper.* 2:36. doi: 10.1186/s41747-018-0068-z
- Romeo, V., Cuocolo, R., Sanduzzi, L., Carpentiero, V., Caruso, M., Lama, B., et al. (2023). MRI radiomics and machine learning for the prediction of oncotype dx recurrence score in invasive breast cancer. *Cancers* 15:1840. doi: 10.3390/cancers15061840
- Saha, A., Harowicz, M. R., Grimm, L. J., Kim, C. E., Ghate, S. V., Walsh, R., et al. (2018). A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. *Br. J. Cancer* 119, 508. doi: 10.1038/S41416-018-0185-8
- Sharmin, S., Ahammad, T., Talukder, M. A., and Ghose, P. (2023). A hybrid dependable deep feature extraction and ensemble-based machine learning approach for breast cancer detection. *IEEE Access* 11, 87694–87708. doi: 10.1109/ACCESS.2023.3304628
- Sun, K., Zhu, H., Chai, W., and Yan, F. (2023). Multimodality MRI radiomics analysis of TP53 mutations in triple negative breast cancer. *Front. Oncol.* 13:1153261. doi: 10.3389/fonc.2023.1153261
- Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 71, 209–249. doi: 10.3322/caac.21660
- Taddese, A. A., and Tilahun, B. C. (2024). Deep-learning models for image-based gynecological cancer diagnosis : a systematic review and meta- analysis. *Front Oncol.* 13:1216326. doi: 10.3389/fonc.2023.1216326
- Traverso, A., Wee, L., Dekker, A., and Gillies, R. (2018). Repeatability and reproducibility of radiomic features: a systematic review. *Int. J. Radiat. Oncol.* 102, 1143–1158. doi: 10.1016/j.ijrobp.2018.05.053
- Tsarouchi, M. I., Hoxhaj, A., and Mann, R. M. (2023). New approaches and recommendations for risk-adapted breast cancer screening. *J. Magn. Reson. Imag.* 28, 1–24. doi: 10.1002/jmri.28731
- Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. *J. Stat. Softw.* 36, 1–48. doi: 10.18637/jss.v036.i03
- Vigil, N., Barry, M., Amini, A., Akhloufi, M., Maldague, X. P. V., Ma, L., et al. (2022). Dual-intended deep learning model for breast cancer diagnosis in ultrasound imaging. *Cancers* 14:2663. doi: 10.3390/cancers14112663
- Wang, C., Zhao, Y., Wan, M., Huang, L., Liao, L., Guo, L., et al. (2023). Prediction of sentinel lymph node metastasis in breast cancer by using deep learning radiomics based on ultrasound images. *Medicine* 102:E35868. doi: 10.1097/MD.00000000000035868
- Wang, D., Hu, Y., Zhan, C., Zhang, Q., Wu, Y., and Ai, T. (2022). A nomogram based on radiomics signature and deep-learning signature for preoperative prediction of axillary lymph node metastasis in breast cancer. *Front. Oncol.* 12:940655. doi: 10.3389/fonc.2022.940655
- Wang, J., Gao, W., Lu, M., Yao, X., and Yang, D. (2023). Development of an interpretable machine learning model for Ki-67 prediction in breast cancer using intratumoral and peritumoral ultrasound radiomics features. *Front. Oncol.* 13:1290313. doi: 10.3389/fonc.2023.1290313
- Wang, L. C., Rao, S., Schacht, D., and Bhole, S. (2023). Reducing false negatives in biopsy of suspicious MRI findings. *J. Breast Imag.* 5, 597–610. doi: 10.1093/jbi/wbad024
- Wang, Q., Xu, J., Wang, A., Chen, Y., Wang, T., Chen, D., et al. (2023). Systematic review of machine learning-based radiomics approach for predicting microsatellite instability status in colorectal cancer. *Radiol. Med.* 128, 136–148. doi: 10.1007/s11547-023-01593-x
- Wei Li, H., Huang, Y., Zhu, T., Wu, Z., Lin, Y., Ye, G., and Wang, K. (2023). Abstract P6-01-19: multiparametric MRI-based Longitudinal-radiomics analysis for early prediction of treatment response of breast cancers to neoadjuvant chemotherapy: a multicenter study. *Cancer Res.* 83, P6-01-19. doi: 10.1158/1538-7445.SABCS22-P6-01-19
- Wei, W., Ma, Q., Feng, H., Wei, T., Jiang, F., Fan, L., et al. (2023). Deep learning radiomics for prediction of axillary lymph node metastasis in patients with clinical stage T1–2 breast cancer. *Quant. Imaging Med. Surg.* 13, 4995–5011. doi: 10.21037/qims-22-1257
- Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., et al. (2011). Quadas-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann. Int. Med.* 155, 529–536. doi: 10.7326/0003-4819-155-8-201110180-00009
- Wu, X., Guo, Y., Sa, Y., Song, Y., Li, X., Lv, Y., et al. (2022). Contrast-enhanced spectral mammography-based prediction of non-sentinel lymph node metastasis and axillary tumor burden in patients with breast cancer. *Front. Oncol.* 12:823897. doi: 10.3389/fonc.2022.823897
- Xiang, H., Wang, X., Xu, M., Zhang, Y., Zeng, S., Li, C., et al. (2023). Deep learning-assisted diagnosis of breast lesions on US images: a multivendor, multicenter study. *Radiology* 5:e220185. doi: 10.1148/ryai.220185
- Xu, Z., Yang, Q., Li, M., Gu, J., Du, C., Chen, Y., et al. (2022). Predicting HER2 status in breast cancer on ultrasound images using deep learning method. *Front. Oncol.* 12:829041. doi: 10.3389/fonc.2022.829041
- Yang, X., Fan, X., Lin, S., Zhou, Y., Liu, H., Wang, X., et al. (2023). Assessment of lymphovascular invasion in breast cancer using a combined MRI morphological features, radiomics, and deep learning approach based on dynamic contrast-enhanced MRI. *J. Magn. Reson. Imaging* 102:e35868. doi: 10.1002/jmri.29060
- Yu, F.-H., Miao, S.-M., Li, C.-Y., Hang, J., Deng, J., Ye, X.-H., et al. (2023). Pretreatment ultrasound-based deep learning radiomics model for the early prediction of pathologic response to neoadjuvant chemotherapy in breast cancer. *Eur. Radiol.* 33, 5634–5644. doi: 10.1007/s00330-023-09555-7
- Yuan, F., Zhang, Z., and Fang, Z. (2023). An effective CNN and Transformer complementary network for medical image segmentation. *Pattern Recognit.* 136:109228. doi: 10.1016/j.patcog.2022.109228
- Zhang, M., Mesurolle, B., Theriault, M., Meterissian, S., and Morris, E. A. (2023). Imaging of breast cancer beyond the basics. *Curr. Probl. Cancer* 47:100967. doi: 10.1016/j.currprobcancer.2023.100967
- Zheng, X., Huang, Y., Lin, Y., Zhu, T., Zou, J., Wang, S., et al. (2023). 18F-FDG PET/CT-based deep learning radiomics predicts 5-years disease-free survival after failure to achieve pathologic complete response to neoadjuvant chemotherapy in breast cancer. *EJNMMI Res.* 13:105. doi: 10.1186/s13550-023-01053-7



OPEN ACCESS

EDITED BY

Sokratis Makrogiannis,
Delaware State University, United States

REVIEWED BY

Aili Wang,
Harbin University of Science and Technology,
China
Manoj Kumar,
University of Wollongong in Dubai,
United Arab Emirates

*CORRESPONDENCE

Teerayut Horanont
✉ teerayut@siit.tu.ac.th

RECEIVED 24 May 2024

ACCEPTED 03 March 2025

PUBLISHED 09 April 2025

CITATION

Lamichhane BR, Srijuntongsiri G and
Horanont T (2025) CNN based 2D object
detection techniques: a review.
Front. Comput. Sci. 7:1437664.
doi: 10.3389/fcomp.2025.1437664

COPYRIGHT

© 2025 Lamichhane, Srijuntongsiri and
Horanont. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

CNN based 2D object detection techniques: a review

Badri Raj Lamichhane, Gun Srijuntongsiri and
Teerayut Horanont*

School of Information, Computer and Communication Technology, Sirindhorn International Institute
of Technology, Thammasat University, Bangkok, Thailand

Significant advancements in object detection have transformed our understanding of everyday applications. These developments have been successfully deployed in real-world scenarios, such as vision surveillance systems and autonomous vehicles. Object recognition technologies have evolved from traditional methods to sophisticated, modern approaches. Contemporary object detection systems, leveraging high accuracy and promising results, can identify objects of interest in images and videos. The ability of Convolutional Neural Networks (CNNs) to emulate human-like vision has garnered considerable attention. This study provides a comprehensive review and evaluation of CNN-based object detection techniques, emphasizing the advancements in deep learning that have significantly improved model performance. It analyzes 1-stage, 2-stage, and hybrid approaches for object recognition, localization, classification, and identification, focusing on CNN architecture, backbone design, and loss functions. The findings reveal that while 2-stage and hybrid methods achieve superior accuracy and detection precision, 1-stage methods offer faster processing and lower computational complexity, making them advantageous in specific real-time applications.

KEYWORDS

CNN, hybrid, object detection, 1-stage, 2-stage, review

1 Introduction to object detection

Access to information occurs through a diverse array of channels, encompassing both traditional and digital sources. Traditional sources include newspapers, magazines, television, radio, books, libraries, and billboards, while digital sources comprise websites, blogs, social media platforms, mobile applications, streaming services, and search engines. When individuals encounter visual stimuli such as advertisements or traffic signs, the ability to accurately identify the objects depicted and extract pertinent information is crucial. Effective information extraction guides individuals along appropriate pathways and mitigates the risks associated with confusion or misinformation that may lead to erroneous conclusions. Consequently, meticulous and precise extraction of information from images is of paramount importance in ensuring informed decision-making (Ardia et al., 2020).

Image detection represents an advanced computational technology that processes visual data to identify and locate specific objects within images. This methodology differs from image classification, which categorizes entire images without delineating object locations. Image detection focuses on recognizing the presence and spatial positioning of objects, often utilizing bounding boxes to indicate their locations within a given frame. The significance of image detection spans multiple domains due to its ability to automate and enhance processes that previously relied on human visual assessment. For instance,

image detection improves operational efficiency in manufacturing quality control by enabling the rapid and accurate inspection of numerous components. In autonomous driving applications, it is essential for detecting pedestrians, traffic signs, and other vehicles, thereby ensuring safety on roadways (Dong et al., 2018). In healthcare settings, image detection is critical for identifying tumors or abnormalities in medical imaging, leading to improved diagnostic accuracy and timely interventions. Furthermore, surveillance systems leverage image detection technologies to monitor environments for unauthorized access or suspicious activities, thereby enhancing security measures. Therefore, image detection provides vital insights that enable systems to respond appropriately to their visual contexts, thereby highlighting its crucial role in contemporary technological applications and societal functions (Hammoudeh et al., 2022). This technology uses advanced machine learning and deep learning algorithms to improve safety across various domains by accurately detecting objects and their environments (Guan, 2017).

1.1 Object identification in image

For a comprehensive understanding of visual data, the classification of images and object detection methodologies constitute critical paradigms in computer vision. The precise identification and spatial localization of objects within digital images or video streams enable a nuanced understanding of content across diverse computational applications. This fundamental interpretive framework encompasses multifaceted computational processes, including but not limited to object trajectory analysis, pose estimation, instance-level object segmentation, and sophisticated inventory management strategies (Dong et al., 2018; Hammoudeh et al., 2022; Dundar et al., 2016). Traditional object detection methodologies are characterized by their ability to operate without necessary historical training data, rendering them predominantly unsupervised. Seminal approaches such as the Viola-Jones algorithm (Viola and Jones, 2001; Li et al., 2012), the Scale-Invariant Feature Transform (SIFT) (Lowe, 1999), and histogram-based techniques (Freeman and Roth, 1995; Dalal and Triggs, 2005) exemplify this methodological category. Contemporary research, however, demonstrates the exponential efficacy of supervised learning paradigms leveraging sophisticated deep learning architectures, which have become predominant in real-world computational scenarios. Within this context, machine learning and advanced artificial intelligence techniques are strategically deployed to comprehend and interpret visual information (Hammoudeh et al., 2022). These sophisticated computational tools enable precise object localization and identification, finding critical applications in multidimensional domains such as intelligent traffic monitoring systems, comprehensive surveillance and navigation frameworks, and advanced biometric recognition technologies in smartphones and autonomous vehicular systems (Guan, 2017; Tang et al., 2023; He et al., 2015a). Figure 1 provides a schematic representation of the intricate object detection and classification methodological landscape.

The paradigmatic evolution of object detection methodologies is significantly driven by the sophisticated integration of advanced

deep learning architectures with supervised learning algorithmic frameworks. This intricate symbiosis represents a pivotal mechanism for optimizing object detection methodologies, substantially augmenting the computational capacity to precisely identify and spatially localize objects within digital imagery and video streams. Contemporary deep learning neural networks, exemplified by Convolutional Neural Networks (CNN) (Sun, 2024), Region-based Convolutional Neural Networks (R-CNN) (Girshick et al., 2014a), You Only Look Once (YOLO) (Redmon et al., 2016), and Residual Networks (ResNet) (He et al., 2016b,a), have made transformative contributions to the field of computer vision. These advanced computational models demonstrate exceptional performance by strategically integrating multi-scale feature representations (Zhang et al., 2019) and iteratively refining candidate object bounding box delineations (Yao et al., 2022) during object identification processes. Neural network algorithmic frameworks, which systematically build upon established architectural paradigms (Dundar et al., 2016) and advanced learning systems, have initiated a global transformation in computational object detection capabilities (Hammoudeh et al., 2022; Guo et al., 2019). Despite these remarkable advancements, significant computational challenges remain in fully recognizing objects across heterogeneous imaging contexts, which include varying illumination conditions, diminutive object dimensions, partial occlusions, diverse viewing angles, complex poses, and varying spatial configurations. Consequently, the scholarly discourse has increasingly focused on object localization methodologies, with researchers actively seeking innovative solutions to address these inherent computational complexities (Guan, 2017). The primary scholarly objective is to achieve unprecedented accuracy in object identification through cutting-edge computational tools. To enable a comprehensive understanding, the conventional object detection framework can be systematically outlined in three fundamental computational stages: the strategic identification of salient informative regions, sophisticated feature extraction mechanisms, and subsequent probabilistic classification processes. This modular computational architecture significantly enhances object recognition capabilities by employing a rigorous, multi-stage approach to identifying and taxonomically classifying objects within digital imagery (Girshick et al., 2014b).

The computational process of object detection involves a sophisticated, multi-staged methodological framework, with each stage playing a crucial role in the precise identification and spatial localization of objects in digital imagery. Informative region selection represents the first computational phase, strategically focusing on identifying spatially salient regions with a high probabilistic likelihood of containing target objects. This critical stage is enabled through advanced computational methodologies such as selective search algorithms and Region Proposal Networks (RPNs), which generate candidate regions through a comprehensive analysis of image chromatic intensities, textural characteristics, and spatial configurations (Girshick et al., 2014a,b). Bounding box representations are systematically employed to outline these potential object zones, providing a precise cartographic representation of the object's spatial positioning within the digital image. The subsequent phase, feature extraction, involves the sophisticated retrieval and transformation of relevant computational data from the selected bounding box



FIGURE 1
Object detection and classification.

regions. Convolutional Neural Networks (CNNs) function as computational transformative mechanisms, converting raw image features into standardized representational matrices that enable the extraction of intricate spatial and hierarchical characteristics essential for effective object detection. Established computational techniques, including Scale-Invariant Feature Transform (SIFT) (Lowe, 1999), HOG (Dalal and Triggs, 2005), and Haar-like feature extraction methodologies (Cristianini and Shawe-Taylor, 2000), are strategically deployed during this computational phase to enhance feature representation and discriminative capabilities. The conclusive stage, Classification, involves the probabilistic assignment of taxonomical class labels to candidate object regions predicated upon the extracted computational attributes. This process entails identifying specific object categories such as anthropomorphic entities, vehicular structures, arboreal organisms, and urban infrastructural elements. The classification mechanism is realized through sophisticated classifiers embedded within fully connected neural network architectures, which leverage the extracted computational characteristics as input to determine the most probable taxonomical designation for each detected object. Canonical classification algorithms, including Support Vector Machines (SVM) (Cristianini and Shawe-Taylor, 2000; Awad et al., 2015), AdaBoost ensemble learning frameworks (Freund et al., 1999), and Deformable Part-based Model (DPM) networks (Viola and Jones, 2001), represent pivotal computational paradigms employed in this sophisticated classificatory process. These meticulously orchestrated computational stages collectively constitute a comprehensive and robust methodological framework for executing sophisticated object detection across diverse computational and visual analysis applications (Chhabra et al., 2024).

This observation highlights the critical need for improvement to increase the effectiveness of real-time object detection systems. Solutions tailored for hardware compatibility rely on discriminant feature descriptors that involve minimal computational overhead and shallow, easily trainable architectures. This is achieved by adopting a pragmatic methodology grounded in reality. However, these techniques may become less reliable when recognizing and predicting essential items is crucial. Finding the right balance between accuracy and efficiency remains essential for successfully using these methods. Deep learning approaches have

seen significant advancements through a result-driven focus. In this study, we focus on object recognition methods using CNNs, which are renowned for their ability to replicate human visual intelligence. We examine one-stage, two-stage, and hybrid approaches to image recognition, localization, classification, and identification to gain a better understanding of the methodologies used by CNN-based object detection systems. We illustrate the benefits of two-stage and hybrid methods regarding accuracy and detection precision while acknowledging the effectiveness of one-stage methods concerning processing speed and computational simplicity. This analysis considers the architecture, backbone structure, and performance metrics of these approaches, emphasizing the need to strike a balance between accuracy, efficiency, and resource usage. Our review aims to facilitate informed decisions when designing and implementing CNN-based object identification systems (Zhao et al., 2024; Aggarwal and Kumar, 2021).

This study focuses on a brief discussion of object detection techniques based on CNN. It begins with key milestones illustrating the developmental process, then dives into several deep-learning object detection techniques utilizing a variety of benchmark datasets. The study explores the hierarchical growth of CNN-based detection strategies, focusing on the architecture of deep-learning CNN models for both one-stage and two-stage object detection. Additionally, it compares approaches based on computational cost, time efficiency, accuracy, algorithmic adaptability, and significance within and across detection stages for both CNN-based generic and salient object detection architectures. The “Challenges and Future Opportunities” section discusses potential ways to overcome the existing obstacles in object detection, while the ‘Conclusion and Future Works’ section summarizes the study’s conclusions and provides guidance for future research directions aimed at advancing CNN-based object detection methodologies.

2 Key milestones in object detection development

Detecting objects in images is a crucial step in the transition from hand-crafted templates to advanced deep learning models. Initially, there was a template-matching approach where image patches were compared to predefined templates. Subsequently, the

idea of designing manual features emerged, such as edges, colors, and textures, which were developed for object identification. After this period, the use of statistical methods for object recognition gained popularity, which proved valuable in certain applications as well (Zou et al., 2023). HOG (Dalal and Triggs, 2005), and SIFT (Lowe, 1999) were prominent during this time. HOG divides an image into blocks to calculate gradients, subsequently combining these blocks with adjacent ones to produce gradient orientation histograms that capture light and maintain invariance over broader areas. While this method is effective for low-cost geometric modifications and varying lighting, it is less effective in identifying small objects or multiple objects within the same image. On the other hand, SIFT examines an object's surrounding areas and spatial context, using edge detection or Laplacian filtering to identify unique key points. The construction of SIFT descriptors relies on histogram computation, with the Gaussian window defining core regions, while key-point matching is performed using Euclidean distance. SIFT provides robustness by carefully selecting key points that generate descriptors, but it is susceptible to issues such as occlusion and background clutter. Therefore, it must be used with caution.

The early manual methods that relied on designed elements such as colors and textures were constrained and rigid; therefore, improvements were required. Furthermore, obstacles to accurate object identification within images include overfitting, which arises from issues with training data for algorithms, such as large datasets and computational resources mentioned by Chen et al. (2017). Deep learning became more famous for overcoming these limitations after 2006 (Zou et al., 2023; Elgendy, 2020), as it fully harnesses the extensive learning capacity of a network structure. Thus, after 2010, there was a revolution in deep learning methods, marking a pivotal shift toward robust convolutional neural networks (CNNs). CNNs utilize deep learning-enabled features to learn complex patterns from massive data, delivering significant accuracy directly. The techniques employed, such as AlexNet (Krizhevsky et al., 2017), GoogleNet (Szegedy et al., 2016), and VGGNet (Simonyan and Zisserman, 2014), in the two-stage detectors, such as R-CNN (Girshick et al., 2014a), Fast R-CNN (Girshick, 2015), and Faster R-CNN (Ren et al., 2015), were enhanced to improve accuracy and performance, making them possible to work in real-time applications. The milestone chart shown in Figure 2 presents the development year alongside the corresponding architecture name. As deep learning approaches expand their applicability in the real world, they diversify techniques that address real challenges without being constrained by their previous limitations. These diverse fields include autonomous vehicles, robotics, medical imaging, security, and more. Advanced deep learning methods revolutionize object detection techniques, ushering in a new era of possibilities. However, some challenges remain to unlock its full potential (Elgendy, 2020).

3 Understanding deep neural networks

Deep Neural Networks (DNNs) are computational models inspired by the human brain, characterized by multiple

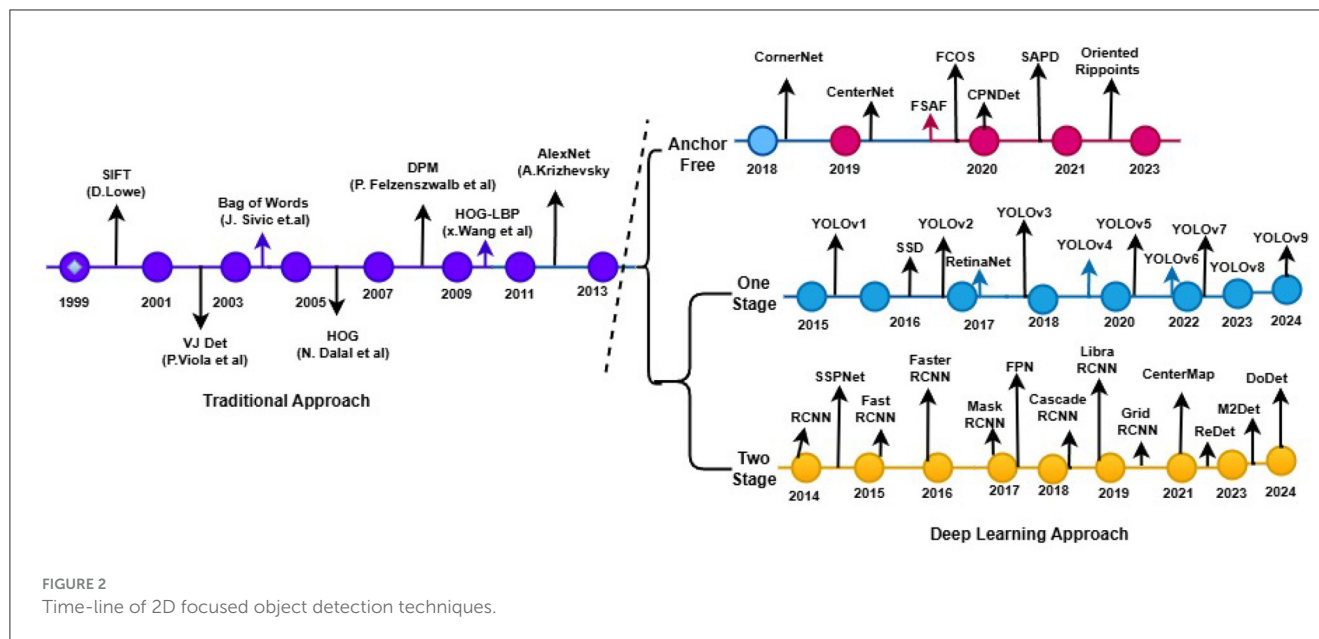
interconnected layers that excel at capturing complex data patterns. These hidden layers enable hierarchical learning of intricate relationships within data, making DNNs powerful tools for addressing diverse and challenging tasks. Neurons in these layers dynamically adjust weights and biases during training to improve feature abstraction. Discrepancies between actual and predicted values are minimized through gradient descent optimization, ensuring improved performance (Krizhevsky et al., 2012).

Various types of DNNs have been developed for specific applications: Feedforward Neural Networks (FNN) are commonly utilized for identification and recognition tasks (Ben Braiek and Khomh, 2023); Convolutional Neural Networks (CNN) excel at image processing tasks (Sun, 2024); Recurrent Neural Networks (RNNs) are primarily used for time series data (Girshick et al., 2014a); Long Short-Term Memory (LSTM) networks address issues related to vanishing gradients in longer sequences (Hochreiter and Schmidhuber, 1997); and Transformer Networks have gained prominence in natural language processing tasks.

CNNs specifically target image recognition and processing. In this context, models are trained using labeled datasets, enabling them to effectively extract relevant information from test images. The feature map technique highlights detected features and serves as input for subsequent layers that progressively build a hierarchical image representation. The fully connected layer then integrates features from earlier layers, mapping them to specific classes and playing a crucial role in image classification (Elgendy, 2020; Zou et al., 2023).

Region-based CNNs (R-CNN) (Girshick et al., 2014a) combine the strengths of CNNs with region-based approaches, significantly enhancing object detection accuracy. Initially, R-CNN segments an image into multiple proposals that may contain objects. A CNN then processes these regions for feature extraction and classification, further refining the final classification through this process. Key variants include Fast R-CNN (Girshick, 2015), which improves processing speed by sharing convolutional features across regions, and Faster R-CNN (Ren et al., 2018), which directly generates proposals to enhance both speed and accuracy. Additionally, Mask R-CNN (He et al., 2017) extends this framework to predict object masks for segmentation tasks. LSTM networks (Hochreiter and Schmidhuber, 1997) enhance learning dependencies within sequential data, thereby improving performance on time series tasks through a network-controlling mechanism that regulates information flow (Amjoud and Amrouch, 2023).

In the realm of object detection, both generic object detection (Girshick et al., 2014b) and salient object detection (Liu et al., 2015; Vig et al., 2014) methodologies aim to identify and understand objects within images. Generic object detection relies on deep learning models to meticulously detect objects of varying sizes when trained on labeled datasets; this includes applications such as pedestrian and traffic sign recognition in autonomous vehicles. Conversely, salient object detection mimics human attention by prioritizing visually distinct objects based on contrast and spatial arrangement attributes. This approach enhances tasks such as robust vision, image compression, and segmentation by emphasizing captivating elements within an image. Implementing these complex algorithms requires extensive training on large



datasets to continually improve object detection capabilities. Consequently, deep learning methodologies have revolutionized image editing techniques and significantly impacted various fields reliant on image classification and object detection (Zhao et al., 2019).

4 Exploring the architecture and functionality of CNN

CNNs represent a powerful deep learning architecture commonly used in various domains, including artificial intelligence, natural language processing, computer vision, and autonomous vehicles. As a dependable foundation for image recognition and analysis, CNNs utilize a structured layer arrangement that collaboratively processes and extracts meaningful information from input data, particularly images. As illustrated in Figure 3, the architecture encompasses several key components: the input layer, pooling layers, convolutional layers, and fully connected layers, each playing a role in extracting abstract features from the image while facilitating sophisticated data interpretation and analysis. This intricate interplay of layers enhances the model's capability to identify complex patterns, positioning CNNs at the forefront of technological advancements in visual recognition tasks.

4.1 Input layer

The input layer is the gateway that receives raw data, establishing the foundation for subsequent network processing. The input data can encompass various categories based on availability and specific requirements. This may include image data in a 3D map format with pixel values indicating width and height, time series data such as stock market values in a 2D format corresponding to time steps, or textual data fed into the network for

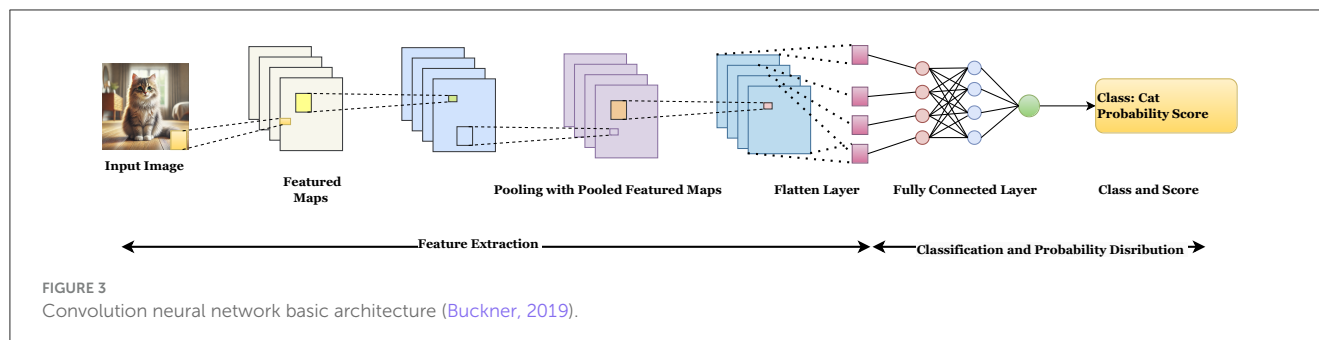
desired outputs. Pre-processing steps, including normalization, are crucial for preparing the data to ensure it aligns with the network's processing capabilities. The primary role of the input layer is to enable meaningful insights derived from the transfer of data during the input stage.

4.2 Convolutional layers

The convolutional layer is a fundamental component where most computations occur. It employs small grids, filters, or kernels to detect specific patterns such as lines, curves, or shapes within the receptive field. This multi-layered architecture of convolutional layers progressively interprets the visual information embedded in raw image data. To detect complex patterns and objects, each successive layer extracts feature maps that inform the deeper layers of the network.

4.3 Pooling layers

Pooling layers generate summary statistics for adjacent layers by downsampling data while retaining essential information. This process enhances object detection capabilities by providing invariance to rotations and translations. Additionally, pooling reduces memory consumption, manages computational costs and weights, and mitigates overfitting. The most common pooling methods include max pooling and average pooling. In max pooling, the highest value within a specified region of the input feature is selected as the output for that region, thereby emphasizing prominent features. Conversely, average pooling calculates the average value from a specific region of the input feature map to produce a smoother representation of features within that region. This approach aids in locating objects in images while considering their overall appearance.



4.4 Activation layers

Activation layers are critical components that enable CNNs to learn non-linear transformations of complex patterns for object detection tasks. These layers capture subtle relationships between features, leading to advanced models with enhanced generalization capabilities. As illustrated in Table 1, popular activation functions include ReLU (Fukushima, 1975), Tanh (Hereman and Malfliet, 2005), Leaky ReLU (Bai, 2022), ELU (Clevert et al., 2015), Sigmoid (Ramachandran et al., 2017), and SELU (Zhu et al., 2023), each with unique characteristics (Nwankpa et al., 2018). By leveraging diverse activation functions, CNNs can effectively address more challenging object detection tasks while retaining resilience and adaptability. These activation layers are essential for enhancing the network's ability to recognize intricate patterns and generate accurate predictions, ultimately improving performance in object detection by modeling and interpreting complex relationships within the data.

4.5 Fully connected layers

Fully Connected Layers (FCL) are integral components of CNNs, designed to connect neurons across different layers. Comprising neurons, weights, and biases, these dense layers serve as essential mechanisms that transform extracted information into a format that can be meaningfully interpreted. The FCL facilitates complex information sharing by integrating features according to the specific nature of the task, whether classification or regression. Ultimately, a single neuron representing the expected output emerges as the final result of a fully connected layer. The structure and functionality of FCLs are illustrated in Figure 3. Fully connected layers enhance the network's ability to comprehend intricate patterns by acting as a bridge between feature extraction and decision-making. The close interconnectivity of neurons within these layers enables CNNs to excel in object detection tasks, effectively synthesizing information from various features to inform predictions (Alzubaidi et al., 2021).

4.6 Architectural backbone network

The backbone architecture of a neural network is its fundamental structure, forming the basis for models, particularly

in deep learning applications utilized for tasks such as image processing. The core of CNNs is composed of layers designed for hierarchical feature extraction. These layers may include pooling, normalization, and convolutions. The backbone architecture captures precise representations of incoming data as it moves through the system, enhancing the network's capability to understand and handle complex information. CNNs employ several well-known backbone networks, including AlexNet (Krizhevsky et al., 2017), VGGNet (Simonyan and Zisserman, 2014), ResNet (Residual Networks) (Choi et al., 2018), InceptionNet (GoogLeNet) (Szegedy et al., 2016), MobileNet (Sandler et al., 2018), and DenseNet (Huang et al., 2017).

4.7 VGGNet architecture

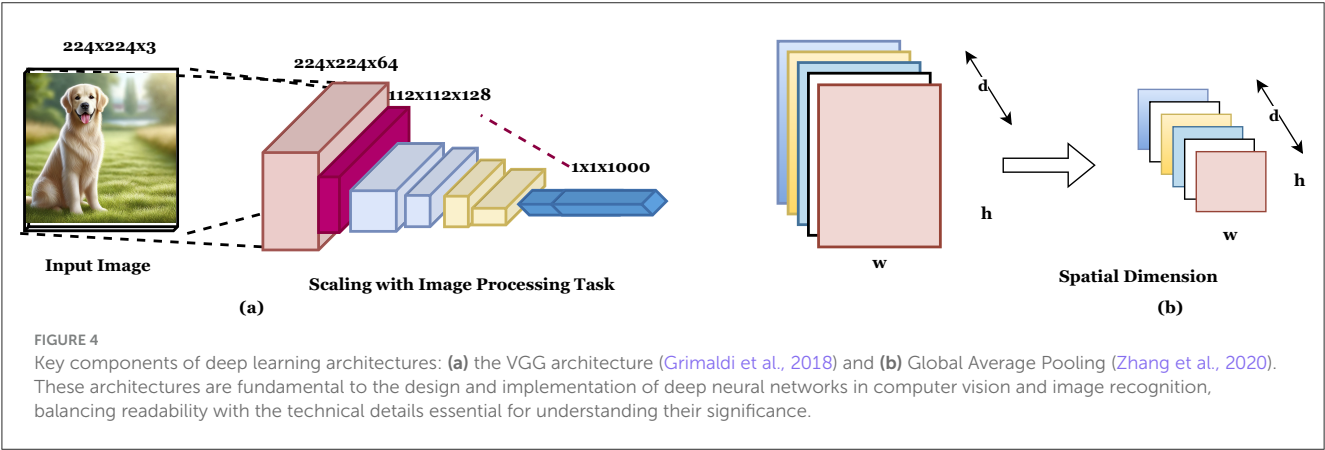
Simonyan and Zisserman (2014) proposed the VGG architecture by significantly enhancing traditional CNN models. This refined design achieved an impressive top-5 accuracy of 92.7% on the widely recognized ImageNet benchmark dataset, demonstrating its effectiveness in large-scale image classification tasks. A general diagram is shown in Figure 4. A key innovation of the VGG architecture is the consistent use of 3x3 convolutional filters throughout the network, reducing the overall parameter count and ensuring architectural simplicity and coherence while maintaining the ability to capture intricate features. The authors presented two variants of this architecture, namely VGG16 and VGG19, which comprise 16 and 19 layers of deep neural networks, respectively (Nash et al., 2018).

4.8 InceptionNet (GoogLeNet)

A groundbreaking deep learning architecture, InceptionNet, also known as GoogLeNet, was introduced by Szegedy et al. (2016). This architecture addressed a critical bottleneck in traditional models by allowing images of varying resolutions to be fed directly into the network without extensive preprocessing. Designed with computational efficiency in mind, InceptionNet achieves superior performance in image classification tasks while optimizing resource utilization. A defining feature of this architecture is the introduction of inception modules, which integrate multi-scale convolutions within a single layer and concatenate their outputs. This approach facilitates the effective capture of local and global features, enhancing the network's ability to learn complex

TABLE 1 Comparative analysis of activation functions in deep learning architectures.

Activation function	Mathematical representation	Benefits	Limitations	Optimal use cases
ReLU (Fukushima, 1975)	$f(x) = \max(0, x)$	Computationally efficient, Alleviates vanishing gradient issues	Susceptible to Dying ReLU phenomenon	Widely adopted in various architectures
Sigmoid (Ramachandran et al., 2017)	$f(x) = \frac{1}{1+e^{-x}}$	Produces binary-outputs (0 and 1)	Vulnerable to vanishing gradients, Saturation effects	Primarily for binary classification tasks
Tanh (Hereman and Malfliet, 2005)	$f(x) = \tanh(x)$	Outputs centered between -1 and 1, Superior gradient behavior compared to Sigmoid	Saturation effects at extremes	Viable alternative to Sigmoid function
Leaky ReLU (Bai, 2022)	$f(x) = \max(\lambda x, x)$ (λ is a small positive constant)	Mitigates Dying ReLU issue, Reduces vanishing gradient risk	More intricate than standard ReLU	Effective in preventing neuron inactivity
ELU (Clevert et al., 2015)	$f(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha(e^x - 1) & \text{otherwise} \end{cases}$ (α is the leakiness parameter)	Produces smoother outputs compared to ReLU, Prevents Dying ReLU issue	Increased complexity relative to ReLU	Suitable for deep neural networks requiring robustness
SELU (Zhu et al., 2023)	$f(x) = \lambda \cdot \alpha \cdot e^x - \alpha$ (λ and α are constants)	Self-normalizing properties, Scaled variant of ELU function	More complex implementation than ELU	High gain scenarios (requires careful tuning)
Swish (Ramachandran et al., 2017)	$f(x) = x \cdot \text{sigmoid}(\beta \cdot x)$ (β is a tunable hyperparameter)	Facilitates smooth gradient propagation, Non-monotonic behavior enhances expressiveness	More computationally intensive than ReLU alternatives	Balances performance and efficiency in various applications
Mish (Ramachandran et al., 2017)	$f(x) = x \cdot \tanh(\ln(1 + e^x))$	Provides smooth outputs akin to ReLU, Enhances training stability through improved gradients	Complex computational overhead compared to simpler functions	Optimal for scenarios requiring superior optimization performance



representations. Furthermore, compared to conventional deep neural networks, InceptionNet significantly reduces the number of parameters while maintaining state-of-the-art accuracy, making it both innovative and efficient.

4.9 ResNet

The Residual Network (ResNet) architecture revolutionized deep learning by addressing the challenges associated with training complex neural networks. ResNet incorporates residual connections, also known as skip connections, which allow the direct flow of information and gradients between layers. This innovative approach effectively mitigates the vanishing gradient problem, a common issue in deep networks, and

facilitates training exceptionally deep architectures comprising hundreds or even thousands of layers. ResNet models, such as ResNet-50, ResNet-101, and ResNet-152, are available in varying depths, with the numbers indicating the total layers in the network. These architectures have demonstrated state-of-the-art performance across various computer vision tasks, including image classification, object detection, and segmentation, establishing ResNet as a foundational model in deep learning research (Choi et al., 2018).

4.10 Output layer

The output layer is the final decision-making component in object detection with CNNs, providing results after thorough data

processing and analysis. It adapts to the specific requirements of regression or classification tasks. In classification scenarios, for instance, the output layer may consist of individual neurons corresponding to different classes, estimating the probability that an input belongs to each category. Conversely, in regression tasks, a single neuron may represent the predicted value or utilize an activation function to convey the learned prediction, with evaluation metrics such as mean squared error or absolute error assessing accuracy. As a critical element of the neural network architecture, the output layer embodies the system's capability to accurately detect and classify objects based on predefined criteria, ensuring optimal performance in object detection tasks and reflecting the complexity of the problems addressed.

5 Generic object detection techniques

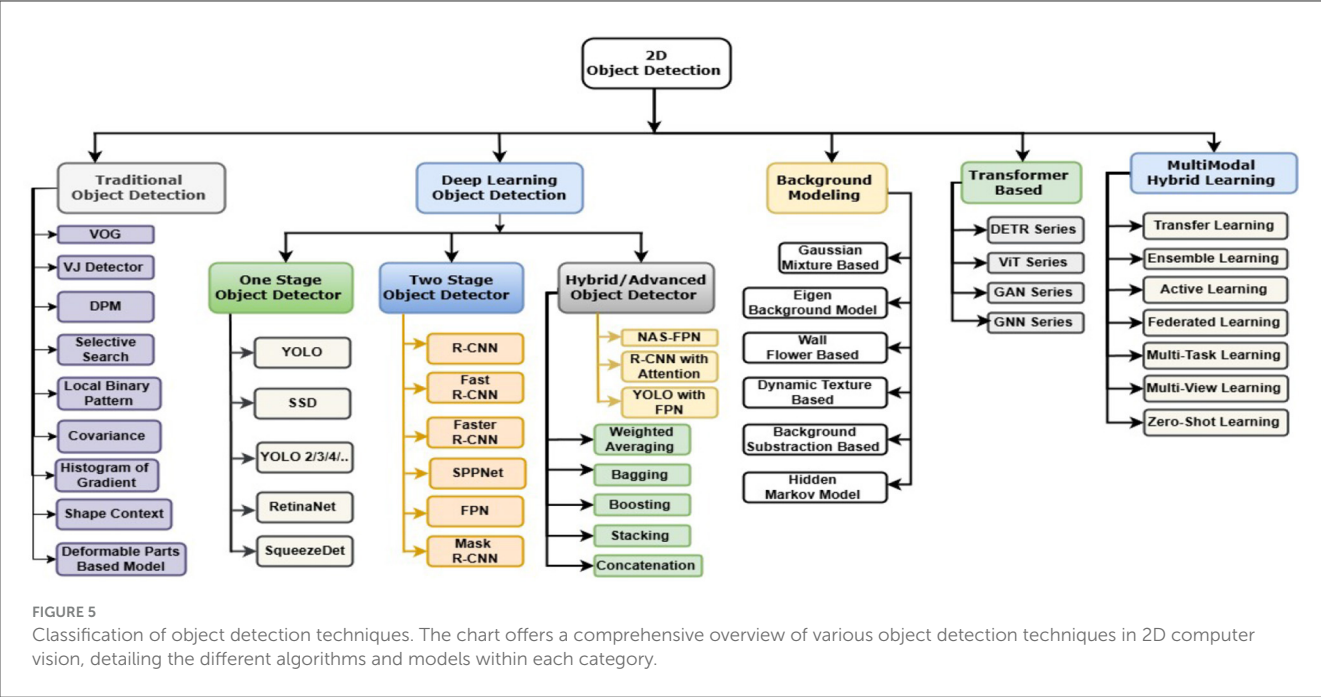
Generic object detection techniques are recognized for their adaptability and versatility in the realm of CNN-based object detection. These methods can detect and classify a wide range of objects in images, including those that do not fit into predefined categories. They effectively identify and classify objects in complex visual environments by generating bounding boxes that outline object locations and provide confidence scores for their detection. A key feature of generic object detection is its ability to function without prior knowledge of specific object categories, relying instead on universal attributes such as color, shape, texture, and edge patterns to facilitate detection. The field includes various methodologies tailored to different requirements and scenarios, integrating advanced strategies to address varying challenges. These techniques achieve accurate and comprehensive object detection through innovative algorithms and robust feature extraction processes, even in intricate and unstructured visual data. As the field evolves, researchers continue to explore novel approaches to feature extraction and detection strategies, enhancing the efficiency, accuracy, and applicability of generic object detection techniques in real-world applications.

The One-Stage framework represents a regression-based approach designed to prioritize speed by predicting object attributes directly, eliminating the need for a separate region proposal step. This architecture aims to simultaneously predict object locations and bounding boxes in a single forward pass through the network, making it particularly suitable for real-time applications. However, the challenge of accurately detecting objects in a single pass often hinders its ability to achieve the same level of precision as more complex frameworks. Prominent implementations of the One-Stage framework include Single Shot Multibox Detectors (SSD) (Erhan et al., 2014; Van de Sande et al., 2011), You Only Look Once (YOLO) (Redmon et al., 2016), AttentionNet (Yoo et al., 2015), G-CNN (Najibi et al., 2016), and Differentiable Single Shot Detector (DSSD) (Fu et al., 2017). These models have garnered widespread popularity due to their ability to streamline object detection tasks and deliver rapid inference times, making them ideal for scenarios where computational efficiency is critical. Despite their speed advantages, substantial research

is underway to enhance their accuracy while maintaining efficiency. As a result, the One-Stage framework remains an active and significant area of exploration in CNN-based object detection, offering both practical applications and opportunities for further innovation.

The Two-Stage framework employs a region-based approach, operating in two distinct stages, making it particularly effective for accurately handling objects with complex shapes and varying poses. In the first stage, the framework identifies potential object regions within the image. The second stage refines the bounding boxes and classifies the objects within these proposed regions. Compared to the One-Stage framework, the Two-Stage approach consistently achieves higher accuracy, although it comes at the cost of increased computational complexity and longer inference times. Prominent implementations of the Two-Stage framework include R-CNN (Girshick et al., 2014a), SPPnet (He et al., 2015a), Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2015), R-FCN (Dai et al., 2016), and Mask R-CNN (He et al., 2017). The evolution of this framework has been marked by several key innovations. R-CNN introduced the concept of a region proposal phase, enabling the detection of potential object locations, while SPPnet incorporated spatial pyramid pooling to effectively handle objects of varying scales. Faster R-CNN enhanced this approach by introducing the Region Proposal Network (RPN), which generates region proposals more efficiently. Fast R-CNN improved upon R-CNN by sharing convolutional features across proposals, which significantly reduces computational overhead. R-FCN introduced position-sensitive score maps, allowing for more precise localization and classification, while Mask R-CNN extended Faster R-CNN by adding instance segmentation capabilities. Despite its superior accuracy, the Two-Stage framework's computational demands and extended processing time render it less suitable for real-time applications. However, it remains a pivotal area of research in object detection using CNNs. Ongoing advancements aim to balance the framework's high accuracy with the need for improved computational efficiency, ensuring its continued relevance in academic and applied settings (Shah and Tembhurne, 2023). The simple classification diagram is shown in Figure 5.

The goal of the **hybrid approach** is to balance speed, computational complexity, and accuracy by combining elements of both one-stage and two-stage frameworks. This strategy enables the creation of object detection systems that leverage the advantages of both methods, maximizing their respective strengths. NAS-FPN (Ghiasi et al., 2019) is a well-known hybrid approach that uses multi-scale representations to enhance object detection at various scales. Other implementations of the hybrid approach use the focal loss technique to address class imbalance issues commonly encountered in object detection applications, such as Mask R-CNN with attention and YOLO with FPN. Several R-CNN variants also adopt a hybrid approach to enhance performance, integrating components of both one-stage and two-stage frameworks. The capability of hybrid approaches to capitalize on the benefits of both frameworks while addressing their limitations has contributed to their increasing popularity in recent years.



5.1 Comprehensive review of one-stage (regression based) networks object detection model

The objects in the image can be recognized quickly and efficiently using single-stage object detection techniques. Single-stage detection methods, such as SSD (Elgendy, 2020) and the YOLO series, predict an object’s approximate bounding box in a single neural network run, enabling quick and effective object recognition. Although this comes at the expense of lower accuracy rates, these methods demonstrate excellent reliability compared to two-stage detectors. Typically, greater accuracy is achieved in identifying larger items compared to smaller and closely spaced objects. Later iterations of YOLO (Redmon et al., 2016), such as YOLOv2 (Redmon and Farhadi, 2017), YOLOv3 (Redmon and Farhadi, 2018), and others, have successfully incorporated deeper neural networks to enhance performance. This modification has yielded more effective results while maintaining strong accuracy and low processing complexity. An in-depth examination of the YOLO and SSD architectures, along with their comparative outcomes across various datasets, will illuminate their advantages and disadvantages. These investigations contribute to a better understanding of the practical performance of these one-stage detection methods. Single-stage detection methods hold promise in applications requiring fast response times as they provide a trade-off between speed and accuracy. Further improvements in these methods will lead to even higher efficiency and accuracy.

5.2 You Only Look Once (YOLO)

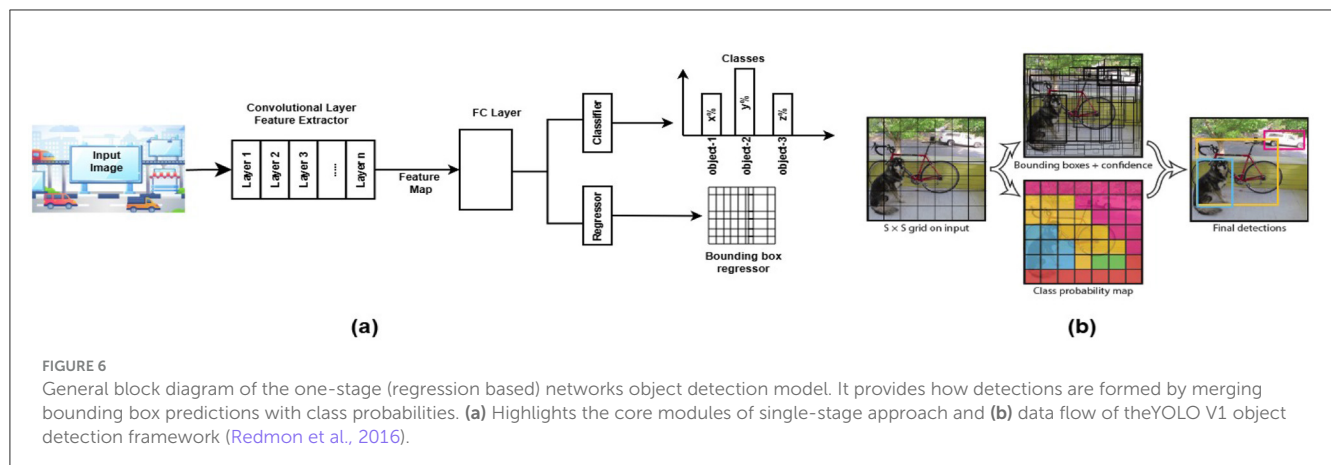
Redmon et al. (2016) proposed a unique and improved single-stage detector called YOLO for object detection and image

verification. YOLO enhances object recognition by combining high-level feature mapping with a reliable evaluation of various item categories, resulting in precise predictions represented in bounding boxes, as shown in Figure 6a. This innovative design, shown in Figure 6b, divides the input image into SxS cells using a grid-based method, with each cell providing bounding box features and essential confidence scores. The probability of an object’s presence and the accuracy of the bounding box location are carefully combined to yield the confidence score. $Pr(\text{project})=1$ indicates the target object’s presence, while $Pr(\text{object})=0$ indicates its absence. YOLO ensures substantial accuracy in object localization evaluations using the Intersection over Union (IoU) metric, a crucial measure of alignment between the actual and predicted boxes. The method of calculating confidence involves a rigorous multiplication of dimensions (x, y, w, h), which represent the size and position of the bounding box. This illustrates YOLO’s commitment to precision and effectiveness in recognizing object tasks (Redmon et al., 2016).

$$Pr(\text{Class}_i|\text{Object}) \times Pr(\text{Object}|\text{IOU}_{pred}^{\text{truth}}) = Pr(\text{Class}_i) \times \text{IOU}_{pred}^{\text{truth}} \tag{1}$$

5.3 YOLOv2

Redmon and Farhadi (2017) developed YOLOv2 in 2017 based on the foundation of YOLOv1. The authors aim to enhance both the speed and accuracy of object detection in a real-time system. The concept of anchor boxes and predefined boxes of varying sizes and aspect ratios is introduced to better estimate the location of objects in the image. Additionally, it randomly resizes input images during the training phase to improve the network’s capacity to detect objects at different scales. In contrast to YOLOv1, which



relies solely on down-sampling to bolster the high-resolution classifier for small object detection, YOLOv2 utilizes multiple layers to transmit high-resolution features to the detector.

5.4 YOLOv3

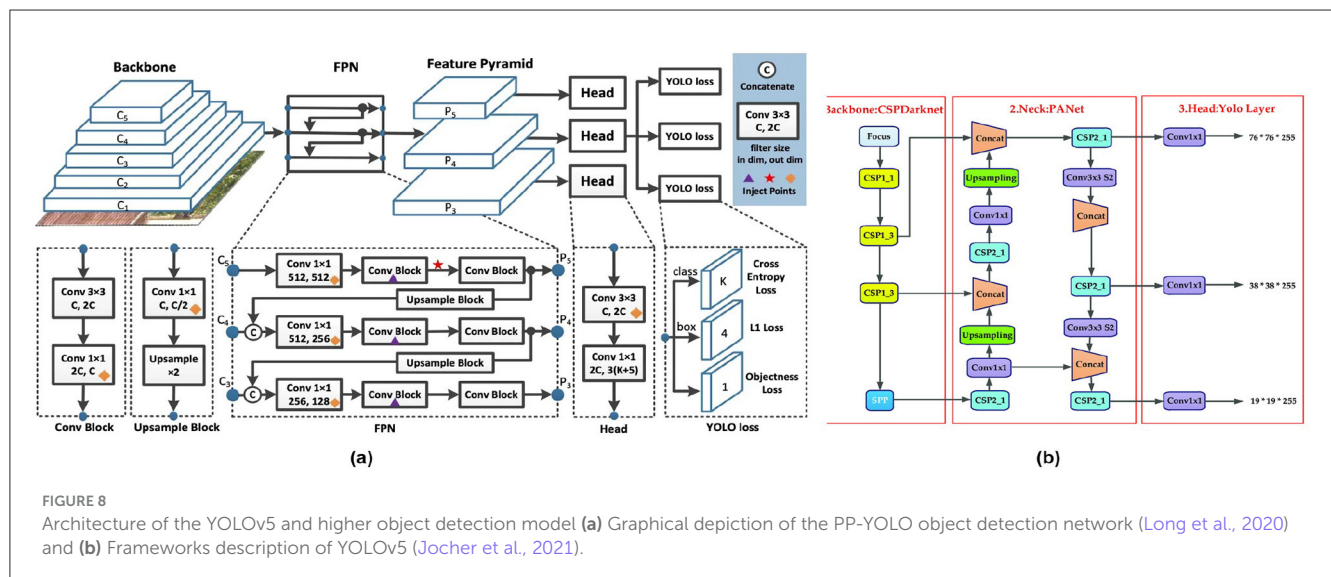
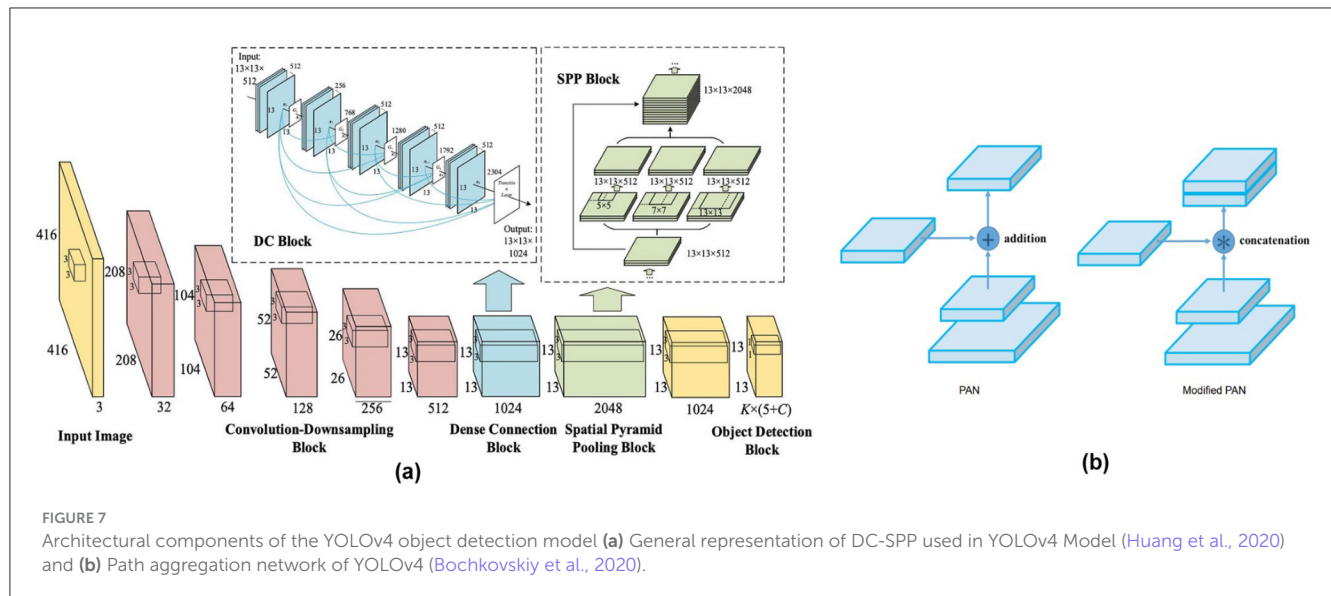
With the release of YOLOv3 (Redmon and Farhadi, 2018), real-time object identification has significantly advanced, showcasing impressive improvements in both speed and accuracy. To improve feature extraction performance and mitigate deterioration in deeper neural networks, a robust Darknet-53 architecture is implemented, consisting of 53 convolutional layers with residual connections. The multi-class probabilistic classifier, featuring independent classes and pyramidal forecasts, integrated into YOLOv3's innovative design, revolutionizes object recognition precision. It improves bounding box creation by applying distance penalties using an aggregated intersection over union technique, further improving the model's ability to accurately locate objects within images. Due to its speed and accuracy, YOLOv3 is one of the top choices for real-time object detection applications, such as detection systems and automated vehicles. It can identify objects at large scales and aspect ratios with minimal processing overhead, making it an effective tool for resource-constrained object detection tasks. Its versatile detection capabilities and economical approach enhance its usability.

5.5 YOLOv4 and higher version

Several significant enhancements are included in YOLOv4 (Bochkovskiy et al., 2020) compared to the earlier version for improving object detection performance. First, CSPDarknet53, a more effective backbone, extracts rich features while maintaining a lower computational load. Second, a smoother activation function ensures improved gradient flow and training stability. Thirdly, by focusing on key points and combining data from different sizes, the Spatial Attention Module and Path Aggregation

Network improve feature representation. In addition, different anchor boxes and optimized loss functions address variations in object size and enhance localization; a focus method prioritizes high-confidence objects during inference, improving real-time performance. Figure 7 is the representation of DC-SPP in YOLOv4. It highlights its use of spatial pyramid pooling with dilated convolutions to enhance receptive fields and capture multi-scale features for robust object detection. The main enhancements of YOLOv5's (Jocher et al., 2021) small object detection are responsible for its success. Initially, prioritizing areas with high feature values, the layer-focusing technique enables the model to efficiently allocate processing power and provide sharper, more accurate detection, especially for smaller objects. Second, the model's ability to identify and locate small objects, which may appear dim or fuzzy at a single scale, is improved by the "Multi-Scale Feature Fusion" technique, which effectively merges information from various feature maps generated at multiple scales. These improvements highlight YOLOv5's dedication to addressing the difficulties related to recognizing small objects in the object detection domain, significantly improving its overall accuracy, particularly in small object identification. Figure 8a presents the PP-YOLO object detection network, visually illustrating its architecture and showcasing key components such as the backbone, feature pyramid, detection head, and post-processing steps for efficient object detection. Figure 8b presents the architecture of the YOLOv5 object detection model.

The new versions of the object detection model, YOLOv6 (Li et al., 2022) and YOLOv7 (Wang et al., 2023), not only predict objects but also estimate their poses. One of the key features of YOLOv6 and YOLOv7 is their ability to forecast an object's presence and pose. This means that the models provide detailed information about the detected objects, making them essential for applications that require a thorough understanding of the spatial orientation of objects within an image. In particular, YOLOv7 incorporates advanced techniques such as position encoding, level smoothing, and data augmentation. These improvements result in more accurate and versatile object identification systems by improving the models' ability to manage real-world data, reducing noise, and enhancing spatial understanding.



YOLOv8 (Reis et al., 2023) primarily focuses on pose estimation through image segmentation. Higher versions of YOLO represent a continual development process, offering improved accuracy, better identification of small objects, enhanced pose detection, and more accurate detection of cropped images. The table below summarizes the different versions of YOLO, the architecture used, the techniques implemented during development by the respective authors, and their performance evaluated on various standard datasets. Detection accuracy, computational time, and complexity with resources are the key factors distinguishing the different YOLO versions.

YOLOv9 (Wang and Liao, 2024) introduced feedback initialization, attention-based modules, and improved feature pyramids, enhancing the detection of small and distant objects while optimizing multi-scale feature learning for faster, more robust inference. Building on this foundation, YOLOv10 (Wang et al., 2024) incorporated dynamic task prioritization and

transformer-based feature extraction, improving the management of complex object interactions and strengthening robustness in challenging scenarios such as occlusions. YOLOv11 (Jocher and Qiu, 2024) further advanced the detection pipeline by implementing cross-domain learning, refining loss functions for better localization, and applying knowledge distillation techniques, enabling efficient training with limited data. Additionally, YOLOv11 achieved state-of-the-art performance with reduced computational overhead, making it highly effective for edge and real-time applications. Collectively, these developments illustrate a trajectory of innovation focused on enhanced feature extraction, robustness, and computational efficiency, positioning YOLOv11 as a versatile model for diverse detection tasks. Table 2 provides a comprehensive overview of various YOLO versions, detailing their advancements in one-stage object detection techniques. It highlights key aspects such as backbone architecture, loss functions, datasets used, and

TABLE 2 Understanding of different YOLO versions of 1-stage object detection techniques.

YOLO version	Researcher	Backbone architecture	Loss function	Used dataset	Accuracy (mAP)	Detection head	References
YOLOv1	Redmon et al. (2016)	GoogleNet	Square Error	VOC 2012 VOC 2007	57.90% 63.40%	Connected Layers, Linear regression	Redmon et al., 2016
YOLOv2	Redmon et al. (2017)	Darknet-19	Logistic regression	COCO dataset VOC 2012 (0.5 IoU threshold) VOC 2007	44.00% 78.60% 78.80%	Multi scale prediction, Connected Layers	Redmon and Farhadi, 2017
YOLOv3	Redmon et al. (2018)	Darknet-53	Binary cross Entropy	COCO VOC 2007 COCO dataset	80.50% 57.90% 33.00%	FPN	Redmon and Farhadi, 2018
YOLOv4	Bochkovskiy et al. (2020)	CSPDarknet-53	Consolidated IoU	COCO	40–61% 62.8% (Ap-50 at over 96 FPS)	SPP PANet	Bochkovskiy et al., 2020
YOLOv5	Glenn Jocher et al. Ultralytics (2021)	CSPDarknet-53	CIoU	COCO	55.8% (YOLO-v5s) 62.4% (YOLO-v5m) 65.4% (YOLO-v5l)	SPPF, CSP-PAN	Jocher et al., 2021
YOLOv6	Li et al. at Meituan (2022)	Efficient REP	CIoU	COCO	43.5% (YOLO-v6-S) 49.7% (YOLO-v6-M) 51.7% (YOLO-v6-L-ReLU)	REPPAN	Li et al., 2022
YOLOv7	Wang et al. (2023)	Extended ELAN	CIoU	COCO	52.8% (YOLO-v7-tiny) 69.7% (YOLO-v7) 71.1% (YOLO-v7-X) 70-84.5%	PAN and SPPCSPC	Wang et al., 2023
YOLOv8	Reis et al. (2023)	modified CSPDarknet53	CIoU with weight loss calculation	COCO and VOC	53.98% (on COCO)	PAN and SPP	Reis et al., 2023
YOLOv9	Wang and Liao (2024)	GELAN	PGI	COCO	Improved accuracy to SOTA	Anchor-free mechanism	Wang and Liao, 2024
YOLOv10	Wang et al. (2024)	Enhanced version of CSPNet	Combines elements of classification loss, localization loss, and objectness loss	COCO	Improved accuracy to SOTA	Adaptive anchor assignment and dynamic label assignment	Wang et al., 2024

accuracy metrics across different iterations from YOLOv1 to YOLOv10.

5.5.1 Single Shot MultiBox Detector (SSD)

Liu et al. (2016) introduced the SSD concept in 2015, utilizing a CNN as the backbone architecture for object detection. SSD enables fast object classification and localization in a single forward path. It employs a set of predefined boxes called “anchor boxes” with different sizes and aspect ratios to detect objects at different locations in an image. The network attempts to predict offsets and adjusts these anchor boxes to accurately fix detected objects according to their size and location using the deep learning techniques. For each object class, the anchor box receives a confidence score to indicate the likelihood of the presence of an object within the box. SSD architecture is divided into two major parts: firstly, the backbone model, a pre-trained classification model, which is a feature map extractor, and secondly, the SSD head, which is moved to the top of the backbone model. This SSD head will provide the bounding box as output over any detected object, resulting in a fast and efficient object detection model. Compared to YOLO, where the object detection method is used to run on different layers at different scales, SS and D run only on the top layers. Similarly, relying on the COCO7 dataset, the tiny SSD has performed better with reliability than the tiny SSD (Wong et al., 2018).

5.5.2 RetinaNet

Lin et al. (2017b) revolutionized object detection by introducing a groundbreaking concept that enhances both accuracy and efficiency through a novel loss function. Rather than using the traditional cross-entropy loss, they proposed the Focal Loss function, which is specifically designed to address the challenges associated with class imbalance during training (Lin et al., 2017b). This innovation allows the single-stage RetinaNet object detector to achieve exceptional accuracy, particularly for small and densely packed objects in images. The model employs a robust backbone network architecture along with two specialized subnetworks, which operate seamlessly at multiple scales to detect objects with precision. The backbone processes input images of varying sizes to compute convolutional feature maps, while the subnetworks manage object classification: one is embedded within the backbone for feature extraction, and the other focuses on the bounding boxes of detected objects. Collectively, these components synergistically improve detection performance within this single-stage framework.

RetinaNet incorporates two pivotal upgrades: Focal Loss and Feature Pyramid Networks (FPN), which redefine its capabilities. Focal loss mitigates the impact of class imbalance caused by the prevalence of background classes or numerous anchor boxes, effectively diminishing the loss contributions of easy-to-classify samples while focusing on more complex cases. Meanwhile, by leveraging a multi-scale feature extraction strategy, FPN enables RetinaNet to excel across varying object scales. Constructing an image pyramid captures critical features at different layers, allowing for precise detection of objects

regardless of size. However, the convolutional process within the CNN architecture naturally reduces feature map sizes at deeper layers, forming a hierarchical, pyramid-like structure that is ideal for multi-scale detection. Figure 9 provides a detailed illustration of the RetinaNet architecture, showcasing its innovative design.

In addition to YOLO, SSD, and RetinaNet, other similar one-stage architectures are used in object detection. Popular examples include SqueezeDet (Wu et al., 2017), DSSD (Fu et al., 2017), DenseNet (Huang et al., 2017), and CornerNet (Law and Deng, 2018). SqueezeNet enhances accuracy in large object detections with its fire module backbone architecture but is limited on mobile platforms. Deconvolutional layer SSD (DSSD) is more efficient for dense and smaller objects in images, utilizing multi-scale predictions for higher accuracy results. However, it consumes significant memory and has slower performance. DenseNet SSD incorporates the feature reuse concept within the SSD framework, serving as a balance between accuracy and resource utilization. CornerNet represents a unique style of object detection, identifying objects through keypoint estimation techniques that can accommodate rotated objects in images; nonetheless, its computational complexity is excessively high. PAA-SSD is the latest model that focuses on one specific type of system, integrating with various SSD model-based platforms to improve accuracy through the probabilistic anchor assignment technique. The results depend on the dataset used and the chosen backbone architecture. Additionally, it may increase the computational complexity of training based on the chosen platform, necessitating careful assignment. Table 3 illustrates different one-stage object detection techniques along with a comparative analysis. The variation in datasets for various object detections, taking into account not only size but also purpose, is considered for the results obtained.

5.6 Comprehensive review of two-stage (region based) networks object detection model

Two-stage or region-based deep learning approaches rank among the most prominent models for achieving accurate object detection in images. These methods excel by employing a two-step process. The Region Proposal Network (RPN) focuses solely on areas containing objects, thereby avoiding an exhaustive search across the entire image. Unlike brute-force methods, RPNs enhance detection accuracy by training on data relevant to object-specific regions, facilitating precise and efficient classification. This method is especially beneficial for real-time applications, as RPNs identify potential object regions, allowing the classification stage to refine bounding boxes for accurate localization. Both stages present opportunities for customization, with advanced network architectures designed to meet the specific requirements of RPN and classification. This section summarizes and analyzes popular two-stage object detection models, comparing them across factors such as speed, accuracy, computational complexity, and advancements proposed by various researchers.

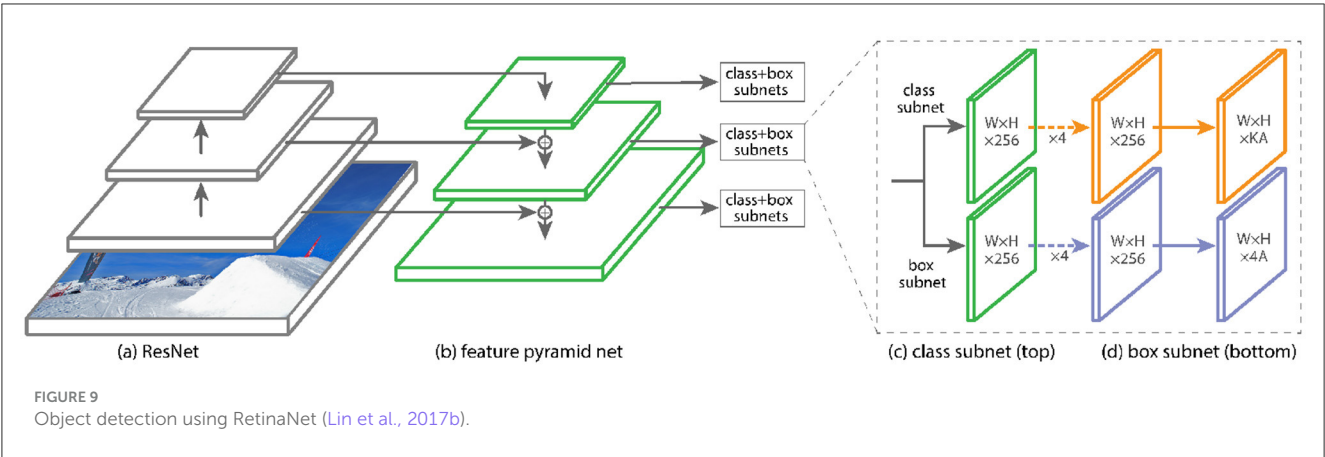


TABLE 3 Understanding of Different 1-Stage Object Detection Techniques beside YOLO.

Model	Researcher	Backbone	Key Parameter	Dataset	Input Size	References
SSD	Liu et al.	VGG-16	Simple, 1-stage	VOC 2007	300x300	Liu et al., 2016
SqueezeDet	Wu et al.	Fire modules	Better accuracy	COCO 2017	Various	Wu et al., 2017
RetinaNet	Lin et al.	Resnet-50	Improved accuracy	COCO 2017	800x800	Lin et al., 2017b
DSSD	Fu et al.	VGG-16	Small Objects prediction	VOC 2007	300x300	Fu et al., 2017
DenseNet	Huang et al.	DenseNet	Higher accuracy	VOC 2007	300x300	Huang et al., 2017
MobileNet	Sandler et al.	MobileNet	Lightweight	VOC 2007	300x300	Sandler et al., 2018
Mobiledets	Xiong et al.	MobileNet	Fast and real-time	COCO	300x300	Xiong et al., 2021
CornerNet	Law and Deng	Hourglass network	Key-point detection	COCO 2017	Various	Law and Deng, 2018
NAS-FPN	Ghiasi et al.	Neural Architecture	Optimized FPN	COCO 2017	Various	Ghiasi et al., 2019

5.6.1 Region-based Convolutional Neural Network

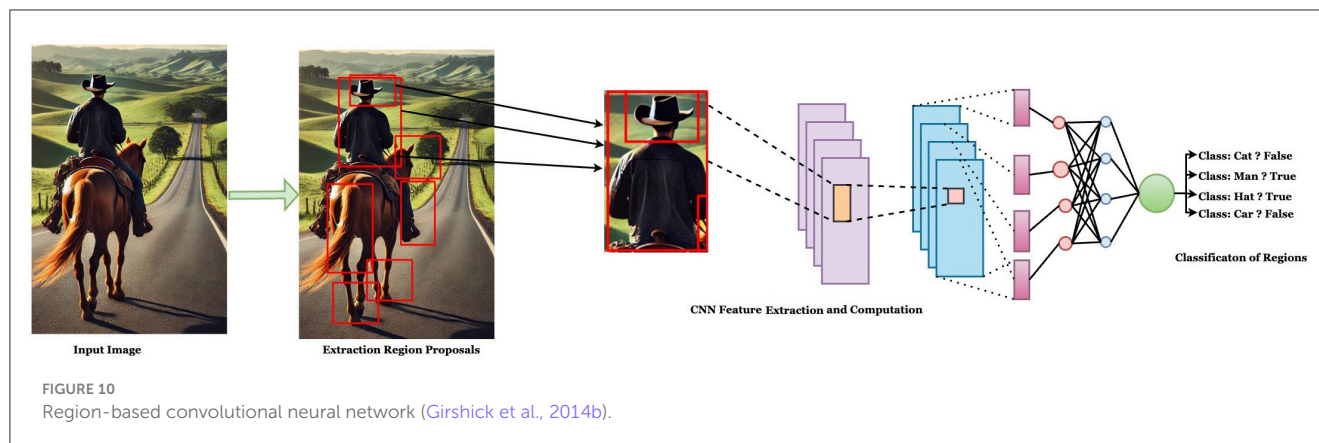
In 2014, Girshick et al. (2014b) introduced a seminal approach to object detection by incorporating CNNs to enhance detection accuracy and improve bounding box quality through deep feature extraction. This method achieved a significant milestone, attaining a mean Average Precision (mAP) of 53.4%, a remarkable improvement over contemporaneous models. The model was trained on the PASCAL VOC 2012 benchmark dataset, setting a new standard for object detection tasks. A simplified representation of the RCNN process is illustrated in Figure 10. The RCNN process comprises two primary stages: region proposal and feature extraction with classification. In the region proposal stage, the entire image is scanned using a selective search algorithm, which evaluates features such as color, texture, position, and location to generate candidate regions likely to contain objects. These candidate regions are resized to conform to the input dimensions the applied CNN requires.

In the feature extraction and classification stage, the resized regions are processed using a pre-trained CNN model to extract high-level features such as color, shapes, textures, and edges. The extracted features are then input into two distinct support vector machines (SVMs): one for object classification and the other for bounding box refinement. The classification SVM predicts the object class (e.g., car, airplane, chair, person, cat), while the bounding box refinement SVM fine-tunes the proposed bounding box to ensure better localization of the detected object. The SVM

assigns a score to each class through non-maximum suppression while maintaining the Intersection over Union (IoU) below a predefined threshold, further enhancing detection precision.

RCNN pioneered object detection by leveraging deep neural networks to extract hierarchical features from images, capturing multi-scale information across layers for precise detection. The model classifies objects and generates bounding boxes around detected regions. However, RCNN also has notable limitations. The fully connected layers in the CNN necessitate resizing images to a fixed size of 277×277 , which increases computational overhead. The selective search algorithm generates thousands of potential regions, resulting in inefficiencies and high computation time. Additionally, processing these regions individually leads to redundant computations, and the SVM-based classification introduces bottlenecks that hinder speed and optimization. These challenges limit RCNN's performance in real-time applications, complex image backgrounds, and small object detection.

To address these issues, several advancements have been proposed. For instance, Zhang et al. (2015) tackled inaccurate localization in RCNN by introducing three key improvements: (1) Bayesian optimization to refine bounding boxes by evaluating classification scores and localization accuracy; (2) structured loss to penalize inaccuracies in predicted bounding boxes; and (3) class-specific CNNs to enhance accuracy for diverse object categories (Zhang et al., 2015). Furthermore, adopting superpixel classification can refine object boundaries and improve efficiency in handling complex scenes and small objects. However, careful



implementation is needed to mitigate potential segmentation inaccuracies.

The fixed filter sizes used in CNN training, along with challenges such as object rotation, deformation, and pose variation, were addressed by Ouyang et al. using deformation-constrained pooling layers. Their approach used guided deformable filters to adaptively adjust shape and size, predicting offset values for local object alignment while applying geometric penalties to promote meaningful deformations. This method, integrated into the DeepID-Net, demonstrated improved accuracy on the ISVRC 2014 dataset (Ouyang et al., 2015).

The limitations of anchor-based bounding boxes, which affect small object detection and computational efficiency, were mitigated by the DeepBox anchor-free design proposed by Zhang et al. This method detects objects without predefined anchors, facilitating better localization of small objects with flexible bounding box shapes and orientations, although it remains sensitive to hyperparameter tuning. Additionally, Pinheiro et al. (2016) introduced SharpMask for refining critical regions, providing superior performance in managing overlapping objects and complex scenes, albeit at the expense of increased computational requirements due to the attention mechanism. Understanding these advancements and their trade-offs provides valuable insights into the suitability of various RCNN-based models for specific tasks. Models such as SPPNet (He et al., 2015a) and Fast RCNN (Girshick, 2015) build upon the foundation of RCNN, delivering significant improvements in efficiency and performance, which will be explored in subsequent sections.

5.6.2 Spatial Pyramid Pooling Network (SPPNet)

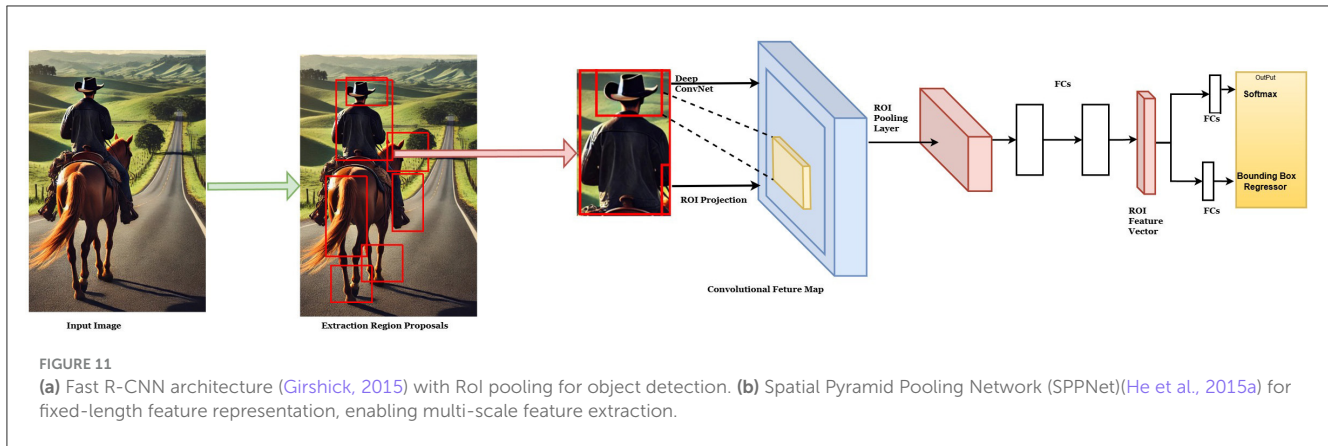
In the context of RCNN, the CNN model requires input images to be of a fixed size, creating challenges when handling images of varying dimensions. This necessitates resizing, which can lead to information loss, reduced accuracy, and increased computational overhead during the scaling process. The Spatial Pyramid Pooling Network (SPPNet) was introduced to address this limitation, enabling the processing of variable-sized input images. SPPNet employs a spatial pyramid pooling mechanism that divides the input image into pyramids of subregions, extracts features from each subregion, and pools them into a fixed-size representation that is independent of the original image

dimensions. This approach enhances the model's flexibility and scale invariance. Furthermore, SPPNet allows feature extraction from multiple convolutional layers at different resolutions, facilitating improved object localization and mitigating the resolution reduction issue inherent in RCNN. By integrating multi-scale feature extraction and enabling efficient handling of variable image sizes, SPPNet significantly improves object detection models' flexibility, scalability, and robustness. It is particularly effective in managing diverse image sizes and complex backgrounds, making it a valuable advancement over traditional RCNN methods. Figure 11 illustrates pivotal architectures in object detection. Fast R-CNN enhances region-based convolutional networks by integrating RoI pooling and shared convolutional features, while SPPNet introduces spatial pyramid pooling to efficiently handle input images of varying sizes (Kaur and Singh, 2023).

Lazebnik et al. (2006) introduced the groundbreaking concept of SPM for object detection, which captures the spatial information of an image by dividing it into multiple subregions and extracting features at various scales. This innovative approach enables the representation of spatial relationships within the image, enhancing feature extraction and localization. Building on this concept, SPPNet incorporates several key advancements. SPPNet efficiently handles images of various sizes, providing scale-invariant detection capabilities. Leveraging multi-scale features extracted from different subregions significantly improves localization accuracy and enhances overall detection performance. These features collectively make SPPNet a robust and accurate model for object detection, particularly in scenarios involving diverse image sizes and complex spatial structures.

5.6.3 Fast R-CNN

The primary limitation of R-CNN lies in its slow processing speed and high computational cost, primarily due to its dependence on selective search for region proposals. Addressing this issue, Girshick (2015) introduced Fast R-CNN, a model that significantly enhances detection speed while maintaining high accuracy. In Fast R-CNN, the RPN directly generates region proposals from image features, eliminating the inefficiency of exhaustive region searches and reducing computational overhead. This approach accelerates



the process and ensures accurate object detection, as presented in Figure 11a.

Fast R-CNN employs a multi-task learning strategy, jointly training the RPN and the classifier to optimize region proposal generation and object detection in a unified pipeline. This integrated framework improves accuracy compared to traditional pipeline-based methods. To further enhance efficiency, Fast R-CNN leverages pre-trained CNN models, such as VGG or ResNet, which are trained on large-scale datasets such as ImageNet. These pre-trained networks capture hierarchical features, ranging from low-level patterns to high-level abstractions, enabling precise analysis of specific regions. The challenge of processing variable-sized regions of interest in fully connected layers, prevalent in R-CNN, is addressed in Fast R-CNN by introducing ROI pooling. ROI pooling divides each ROI into fixed-size subgrids, extracting uniformly sized features for the fully connected layers, thus ensuring consistency and improving detection accuracy.

For region proposal generation, ROI pooling utilizes features extracted from the RPN, avoiding external algorithms such as selective search. The softmax layer classifies objects within the image by predicting the probability of each object class, resulting in a $K + 1$ -dimensional vector for K object classes, with the additional dimension representing the background category. The class with the highest probability is assigned to the detected object. The bounding box regression branch also employs linear regression to refine the predicted bounding box coordinates. Offset values, derived from ROI-pooled features, are added to the initial ROI coordinates to improve bounding box precision, ensuring accurate localization of objects. Fast R-CNN represents a significant advancement over its predecessor, achieving superior speed and accuracy in object detection tasks (Girshick, 2015).

The multi-task loss L of Fast R-CNN is jointly expressed with the two output layers, specifically training for classification and bounding box regression for each labeled ROI. For the trained model, the discrete probability distribution p , computed by a softmax over $K + 1$ categories per ROI from a fully connected layer, is defined by

$$p = (p_1 + p_2 + \dots + p_n) \quad (2)$$

the output bounding box regression offset is given by

$$t^k = (t_x^k, t_y^k, t_w^k, t_h^k). \quad (3)$$

Where K is the object class indexed by k . The Iverson bracket indicator function $[u \geq 1]$ is employed to omit all background ROIs (Girshick et al., 2014b).

$$L(p, u, t^u, v) = L_{\text{cls}}(p, u) + \lambda [u \geq 1] L_{\text{loc}}(t^u, v) \quad (4)$$

in which $L_{\text{cls}}(p, u)$ is log loss for true class u .

The second task loss, L_{LOC} . The bounding-box regression targets for the class $(u, v) = (v_x, v_y, v_w, v_h)$ and the predicted tuple $tu = (tu_x, tu_y, tu_w, tu_h)$. For the bounding box regression, we use the loss

$$L_{\text{loc}}(t^u, v) = \sum_{i \in \{x, y, w, h\}} \text{smooth}_{L_1}(t_i^u - v_i), \quad (5)$$

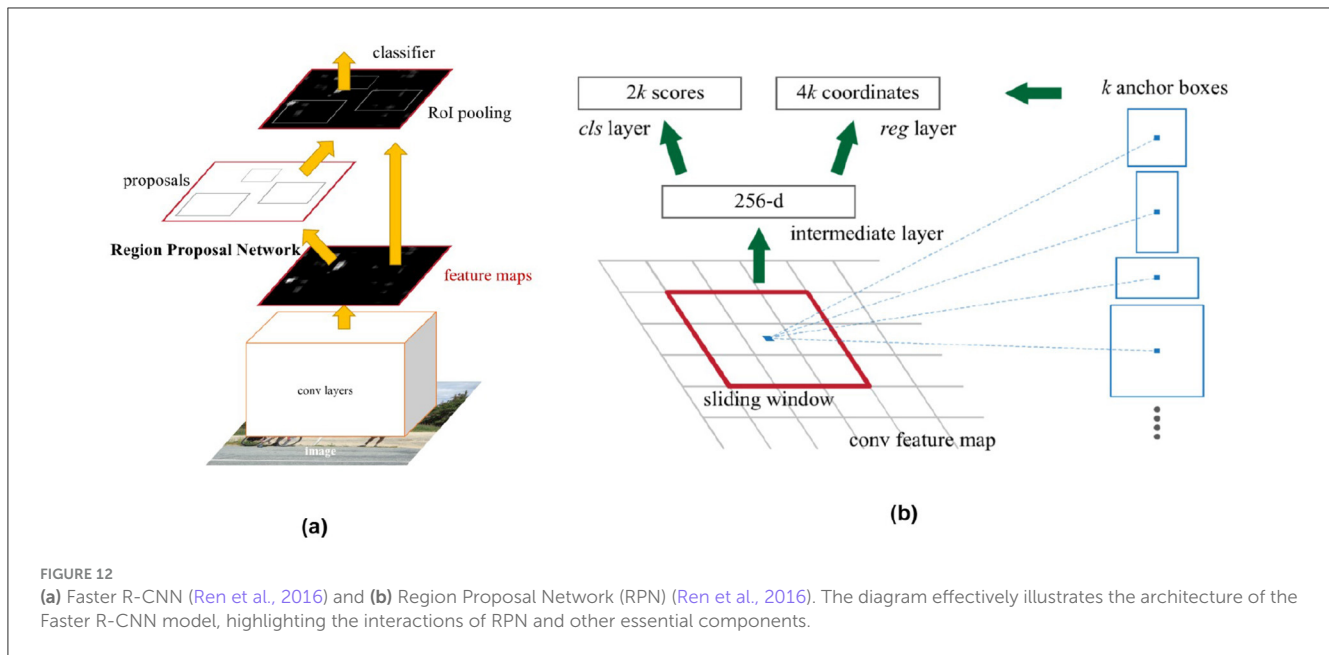
in which

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise,} \end{cases} \quad (6)$$

Here the L_1 loss is less sensitive to the outliers than the L_2 loss used in R-CNN and SSPNet.

5.6.4 Faster R-CNN

Although Fast R-CNN improves detection speed and accuracy compared to its predecessor, it still has limitations in terms of optimization and efficiency. Fast R-CNN relies on pre-trained feature extraction and external ROI pooling mechanisms that utilize fixed-size features and a softmax bounding box classifier. This dependence on external algorithms for region proposals can introduce inefficiencies, including slower processing and potential inaccuracies. Additionally, the requirement for separate stages in the training process decreases overall efficiency. Faster R-CNN was introduced to address these issues by integrating RPN with the CNN architecture into a unified framework. This design eliminates the dependence on external algorithms, resulting in significantly higher speeds than Fast R-CNN. The model trains the entire pipeline jointly, enhancing efficiency, accuracy, and effectiveness. Furthermore, Faster R-CNN is better equipped to handle diverse



datasets, improving its performance in real-time applications (Ren et al., 2015).

In the architecture of Faster R-CNN, the RPN slides a small spatial window over the feature map, connecting to an $n \times n$ spatial region. For instance, with VGG16, a low-dimensional vector of size 512 is extracted within the sliding window and passed to two fully connected (FC) layers: one for box classification (cls) and another for box regression (reg). This architecture incorporates an $n \times n$ convolutional layer connected to two 1×1 convolutional layers, as depicted in the corresponding (Figure 12).

Bounding box regression is achieved by refining the proposals in relation to the reference boxes. The model utilizes anchors of three different scales and three aspect ratios, which improve detection for objects of various sizes and shapes. The loss function in Faster R-CNN is similar to that of Fast R-CNN, maintaining a balance between classification accuracy and bounding box regression.

The loss function is given by,

$$L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \quad (7)$$

Where

p_i : Predicted probability of anchor box i containing an object (foreground).

t_i : Predicted bounding box coordinates (4 values: x, y, width, height) for anchor box i .

p_i^* : Ground truth label for anchor box i (1 for foreground, 0 for background).

t_i^* : Ground truth bounding box coordinates for the object associated with positive (foreground) anchor box i .

N_{cls} : Number of anchor boxes per image in the mini-batch during training.

N_{reg} : Number of positive (foreground) anchor boxes in the image.

λ : Hyperparameter balancing the importance of classification

and regression tasks.

$L_{cls}(p_i, p_i^*)$: Classification loss for anchor box i , often binary cross-entropy.

$L_{reg}(t_i, t_i^*)$: Regression loss for the predicted bounding box of anchor box i , often Smooth L1 loss.

5.6.5 Mask R-CNN

Mask R-CNN is a robust deep learning framework for object detection and instance segmentation. During object detection, it identifies and localizes objects within an image while incorporating instance-level context, which enables precise recognition of what the objects are and their locations. In the segmentation phase, Mask R-CNN goes beyond bounding boxes to create pixel-level masks for individual objects, providing superior precision. It can accurately segment various objects, such as cars, cats, bicycles, or billboards, even in challenging conditions such as partial occlusion or shadowed regions.

Mask R-CNN addresses key limitations of Faster R-CNN, such as the inability to segment individual objects within the same class or differentiate between multiple instances (e.g., distinguishing people in a crowd). It also reduces the computational overhead of storing intermediate features, thereby improving efficiency. By incorporating fine-tuning mechanisms, Mask R-CNN enhances detection accuracy and optimization, making it a robust solution for tasks such as autonomous navigation. Additionally, its end-to-end network training provides better optimization and performance compared to Fast and Faster R-CNN, establishing it as a versatile and reliable tool for image analysis.

Introduced by He et al. (2017), Mask R-CNN enhances the capabilities of Faster R-CNN by adding instance segmentation at the pixel level. Its innovation lies in integrating a mask prediction branch alongside the bounding box classifier, addressing the spatial limitations of Faster R-CNN's bilinear interpolation with a novel sampling technique that preserves spatial information.

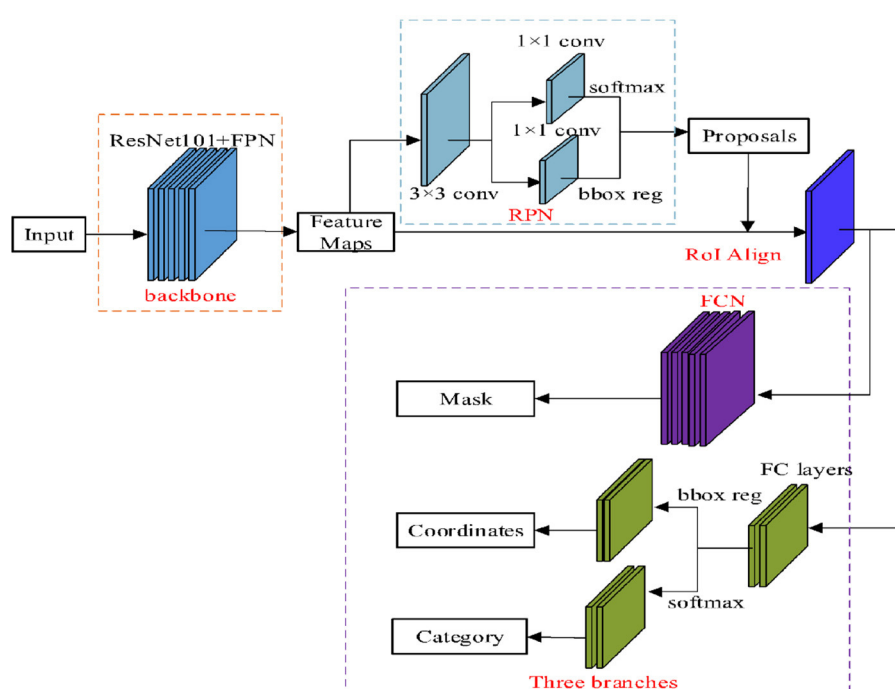


FIGURE 13
Architecture of mask R-CNN for instance object detection (He et al., 2017).

By combining region proposal and classification, this unified network architecture improves training efficiency by eliminating intermediate feature storage and optimizing the entire network through end-to-end learning. This architecture is depicted in Figure 13.

The core of Mask R-CNN is built upon Faster R-CNN, incorporating additional components such as ROI Align, shared feature pooling, and a mask prediction branch. Key architectural elements include the backbone feature extractor, RPN, shared pooling layers, detection and segmentation branches, and a multi-task loss function for joint optimization. Table 4 summarizes the performance of various models, with segmentation showing notable results. The findings demonstrate significant improvements in AP across different backbone architectures, highlighting the effectiveness of Mask R-CNN in this domain.

The foundational architecture of Mask R-CNN is Faster R-CNN. The in-depth architectural components of Mask R-CNN include the backbone feature extractor, RPN, shared feature pooling, detection, segmentation, and the calculation of the training and loss functions. ROI Align, shared features, and end-to-end training are additional components of Mask R-CNN compared to Faster R-CNN.

The backbone extractor, typically ResNet (50/101) or a VGG variant, captures complex features to enhance detection accuracy. Feature maps derived from this backbone provide rich semantic information about the input image. Anchors within the RPN are adapted to the objects' shape and size, and convolutional layers predict the presence of objects and bounding boxes. Non-maximum suppression (NMS) ensures efficient processing by suppressing redundant regions and selecting high-confidence proposals. Shared feature pooling, specifically ROI Align, preserves

spatial information while resizing features for consistent mask prediction. For each ROI, the classification branch predicts object classes using fully connected layers and a softmax activation function, while the bounding box regression branch refines localization. The mask branch generates binary masks for ROIs, and skip connections enhance the network's ability to capture object shapes and extents. A multi-task loss function optimizes classification, bounding box regression, and mask prediction simultaneously, enabling robust performance through end-to-end training. Despite its computational complexity and high hardware requirements, Mask R-CNN remains a state-of-the-art tool for computer vision applications (He et al., 2017).

5.6.6 Feature Pyramid Network

The concept of Feature Pyramid Networks (FPN) was developed by researchers (Lin et al., 2017a) to address the challenges associated with traditional object detection methodologies. Specifically, FPN aims to resolve two primary issues: the loss of spatial information due to down-sampling and the limited semantic information that can hinder accurate object detection. Furthermore, FPN partially mitigates the limitations inherent in Mask R-CNN, which employs traditional CNNs that often experience reduced spatial resolution, complicating the precise localization of small objects within images. FPN integrates bottom-up and top-down pathways to produce multi-scale feature maps while maintaining semantic information. This architecture enhances detection capabilities for small objects across various resolutions and sizes. The semantic gap in feature maps derived from different levels in Mask R-CNN can significantly degrade detection accuracy, particularly in cluttered scenes. To address

TABLE 4 Mask R-CNN performance for instance segmentation (He et al., 2017).

Model	Backbone	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
MNC	ResNet-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
FCIS +OHEM	ResNet-101-C5-dilated	29.2	49.5	-	7.1	31.3	50
Mask RCNN	ResNet-101-C4	33.1	54.9	34.8	12.1	35.6	51.1
Mask RCNN	ResNet-101-FPN	35.7	58	37.8	15.5	38.1	52.4
Mask RCNN	ResNetXt-101-FPN	37.1	60	39.4	16.9	39.9	53.5

this, FPN employs lateral connections that bridge this gap by injecting high-level semantic information from deeper layers into the feature maps. This ensures accurate object identification along with location and class information (Chhabra et al., 2023).

A notable advantage of FPN is its ability to reuse features computed within the CNN backbone, which minimizes computational overhead. This resource-efficient design enables the construction of multi-scale features without creating pyramids from scratch, thereby enhancing computational efficiency. As a result, FPN demonstrates improved accuracy while maintaining low computational complexity across diverse applications. It excels at detecting small objects of varying sizes with high accuracy and efficiency. During the detection process, FPN effectively integrates both bottom-up and top-down approaches. The bottom-up pathway captures fine spatial details with high semantic value using existing convolutional networks, although it may lack semantic richness and exhibit lower resolution. Conversely, the top-down pathway begins with the deepest feature map and progressively upsamples it while merging it with shallower maps through lateral connections. This synthesis results in a comprehensive representation that combines high-level semantic information with preserved low-level spatial details. Figure 14 presents how features are extracted at multiple scales, resulting in a feature pyramid that captures information at different levels of detail and abstraction.

5.6.7 CentripetalNet

CentripetalNet demonstrates higher prediction accuracy than the bounding box approach in FPN. It achieves fine-grained localization of potential objects within an image and delivers superior performance in challenging scenarios, such as dense or crowded scenes and partially visible objects. The architecture of CentripetalNet, illustrated in Figure 15, leverages key points for object detection. Kivee (Dong et al., 2020) developed CentripetalNet to pursue high-quality keypoint pairs for object detection, addressing issues related to inaccurate keypoint matching and limited feature integration, which often result in the loss of spatial context and essential information for effective object detection.

This recent object detection approach relies on key points instead of bounding boxes, predicting primary objects based on the location and relationships of corner key points. Initially, the model predicts the corner key points associated with each object and utilizes a shift vector, known as the centripetal shift, to guide these points toward the object's center. To pair

corresponding key points within the same object, it employs predicted shift values in a process called shift matching, which is particularly useful when the points are initially scattered. Corner pooling extracts features from the area surrounding each corner point with sufficient precision to represent them as detected objects. Finally, deformable convolutions are employed to refine the exact shape of the object in real-time. Table 5 provides a comparative evaluation of object detection performance on the MS-COCO test-dev dataset, focusing on various detection methodologies and their performance metrics. Key indicators, including Average Precision (AP), AP at different Intersection over Union (IoU) thresholds (AP₅₀, AP₇₅), and performance across small, medium, and large object scales, are presented. The findings indicate that multi-scale approaches, particularly those employing CentripetalNet and CentripetalNet, exhibit enhanced performance across all assessed metrics, highlighting their efficacy in object detection tasks.

5.6.8 Dual-path aggregation network (D2Det)

Cao et al. (2020) introduced an aggregation network for object detection that significantly enhances accuracy while maintaining high processing speed, making it suitable for real-time applications. D2Det addresses the limitations of traditional methods by employing a dual-path aggregator that integrates high spatial detail from low-resolution features with rich semantic information from high-resolution features. This design balances the trade-off between accuracy and classification efficiency. D2Det selectively applies deformable convolutions at specific stages to improve feature learning, optimizing the balance between system performance and computational efficiency. Furthermore, lightweight layers ensure faster processing speeds, making the architecture highly suitable for real-time tasks. The simplified yet advanced design of D2Det has led to its adoption in various real-time domains, demonstrating robust performance and scalability.

5.6.9 TridentNet

Li et al. (2019) developed TridentNet, a multi-branch architecture featuring a three-branch structure aimed at addressing scale variations in object detection. Each branch processes distinct field parameters to specialize in detecting objects of various sizes. The low-resolution branch captures fine-grained details of small objects, the middle-resolution branch balances detail and semantic information for medium-sized objects, and the high-resolution branch focuses on the semantic representation of large objects.

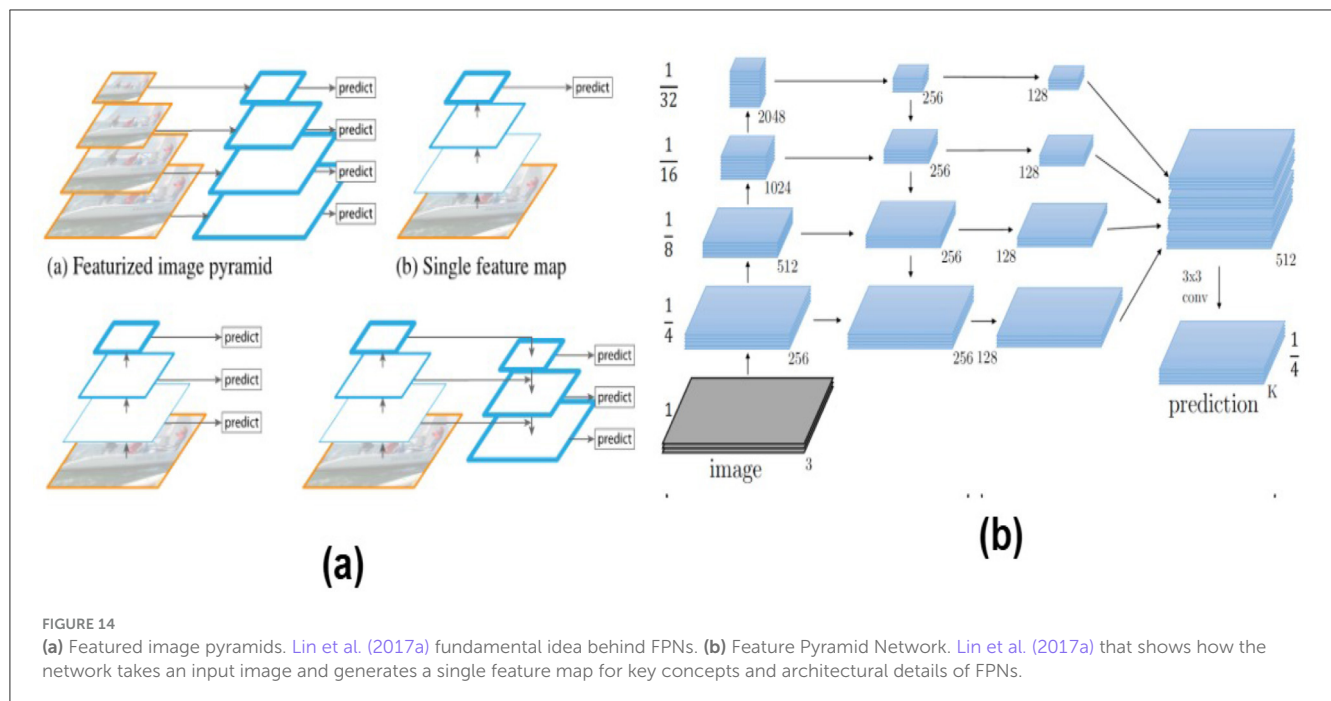


FIGURE 14

(a) Featurized image pyramids. Lin et al. (2017a) fundamental idea behind FPNs. (b) Feature Pyramid Network. Lin et al. (2017a) that shows how the network takes an input image and generates a single feature map for key concepts and architectural details of FPNs.

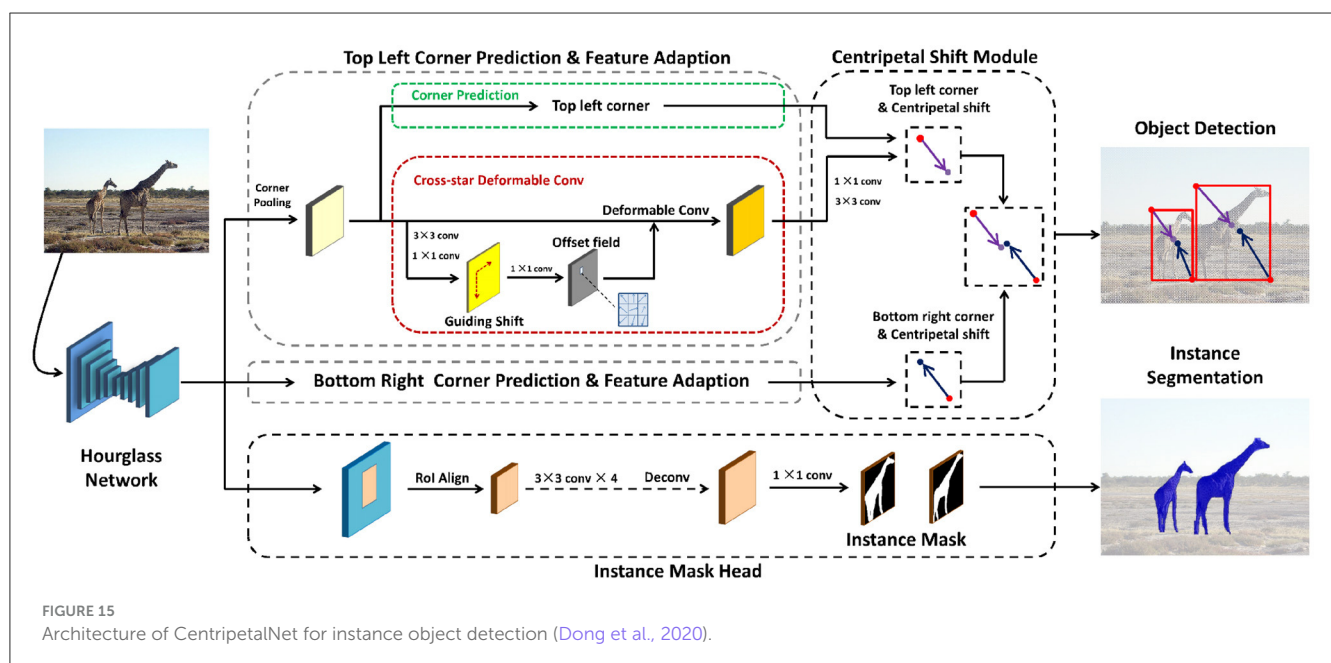


FIGURE 15

Architecture of CentripetalNet for instance object detection (Dong et al., 2020).

During the training phase, TridentNet segments the image based on the size of the objects, with each branch functioning on its corresponding scale. To achieve an efficient design with low computational complexity, TridentNet shares weights among the branches, setting it apart from conventional multi-scale approaches (Alzubaidi et al., 2021). This architecture provides improved accuracy and effectiveness for detecting objects across diverse scales, as demonstrated in its performance comparison with other architectures in Table 6.

The methods discussed above emphasize multitasking, multi-scaling, and contextual detection to manage objects of varying sizes and complexities in images. In multitask learning, they

detect objects using bounding boxes of regular shapes, classify the detected objects, and estimate key points, particularly corner points, while assessing object depth. Multi-scale representations facilitate the detection of objects at different scales by extracting and integrating features from various resolutions within an image. Contextual modeling focuses on understanding the relationships between objects and their surrounding backgrounds, enabling the differentiation of overlapping objects through spatial information. This approach enhances scene comprehension, leading to more accurate object detection in complex environments. Addressing these aspects improves detection accuracy, particularly in challenging scenarios.

TABLE 5 Comparison of object detection performance on the MS-COCO test-dev dataset for various methods, highlighting metrics such as Average Precision (AP), AP₅₀, AP₇₅, and performance across small (AP_S), medium (AP_M), and large (AP_L) object scales for single-scale and multi-scale evaluations (Dong et al., 2020).

Methods	Backbone	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
ExtremeNet (single-scale)	Hourglass -104	40.2	55.5	43.2	20.4	43.2	53.1
CornerNet511(multi-scale)	Hourglass -104	42.1	57.8	45.3	20.8	44.8	56.7
ExtremeNet (multi-scale)	Hourglass -104	43.7	60.5	47	24.1	46.9	57.6
Centernet511(single-scale)	Hourglass -104	43.7					
Centernet511(multi-scale)	Hourglass -104	47	64.5	50.7	28.9	49.9	58.9
CetripetalNet w./o mask (single-scale)	Hourglass -104	45.8	63	49.3	25	48.2	58.7
CetripetalNet w./o mask (multi-scale)	Hourglass -104	47.8	65	51.5	28.9	50.2	59.4
CetripetalNet (single-scale)	Hourglass -104	46.1	63.1	49.7	25.3	48.7	59.2
CetripetalNet (Multi-scale)	Hourglass -104	48	65.1	51.8	29	50.4	59.9
Instance segmentation performance comparison							
CetripetalNet (single-scale)	Hourglass -104	38.8	60.4	41.7	19.7	41.3	51.3
CetripetalNet (Multi-scale)	Hourglass -104	40.2	62.3	43.1	22.5	42.6	52.1

The bold values (with highest score) indicate the best performance.

TABLE 6 Two-stage object detection architectures summary.

Architecture	Authors (Year)	Region proposal	Bounding box prediction	Loss function	References
R-CNN	Girshick et al. (2014)	Selective Search	SVM classification	SVM hinge loss	Girshick et al., 2014a
Fast R-CNN	Girshick et al. (2015)	RPN (CNN-based)	Shared CNN layers with region-specific features	Multi-task learning with SGD: Smooth L1 loss, softmax loss	Girshick, 2015
Faster R-CNN	Ren et al. (2015)	RPN (CNN-based)	Shared CNN layers with RoI Pooling	Multi-task learning with SGD: Smooth L1 loss, softmax loss	Ren et al., 2015
Mask R-CNN	He et al. (2017)	RPN (CNN-based)	Shared CNN layers with RoI Pooling and mask prediction branch	Multi-task learning with SGD: Smooth L1 loss, softmax loss, binary cross-entropy loss	He et al., 2017
Cascade R-CNN	Cai and Nuno (2018)	RPN (CNN-based)	Multi-stage refinement with residual connections	Multi-stage learning with SGD: Smooth L1 loss, softmax loss	Cai and Nuno, 2018
RetinaNet	Lin et al. (2017)	FPAN-based anchor generation	FPN-based multi-level prediction with focal loss	Focal loss, multi-task learning with SGD: Smooth L1 loss, softmax loss	Lin et al., 2017b
PolarMask	Xie et al. (2020)	RPN (CNN-based)	Shared CNN layers with RoI Pooling and mask refinement branch	Multi-task learning with Adam: Smooth L1 loss, softmax loss, binary cross-entropy loss	Xie et al., 2020
NAS-FPN	Ghiasi et al. (2019)	RPN (CNN-based)	Shared CNN layers with NAS-designed FPN	Multi-task learning with SGD: NAS-optimized loss function	Ghiasi et al., 2019
Deformable DETR	Zhu et al. (2021)	Transformer-based proposal generation	Set Transformer-based bounding box prediction	Hungarian loss	Zhu et al., 2021

5.7 Hybrid approach of object detection model

In addition, algorithms such as Cascade R-CNN (Cai and Nuno, 2018) utilize a cascaded framework based on Faster R-CNN, merging the strengths of both stages. In the first stage, it generates feature maps for proposal generation and coarse classification, while the second stage refines these feature maps to improve accuracy. This combination of features enhances detection precision and achieves superior performance on benchmarks.

Similarly, Mask R-CNN, which includes a cascade head, integrates both stage concepts by utilizing a cascading structure for mask prediction, thus refining bounding boxes and improving prediction accuracy. This method excels in instance segmentation by delivering accurate mask predictions. Similarly, Libra R-CNN (Pang et al., 2019) utilizes a hybrid approach that alternates between YOLOv2 (one-stage) and Faster R-CNN (two-stage) depending on the confidence score. YOLOv2 manages initial predictions, while Faster R-CNN provides further refinement, balancing speed and accuracy. Another notable method is RetinaNet-RegNet

citepxu2022regnet, a hybrid object detection technique that incorporates the RegNet backbone into the RetinaNet architecture. This method improves the detection of small objects while ensuring robust performance in multi-scale and multi-class object detection. Table 7 summarizes various hybrid approaches in object detection algorithms, highlighting their region proposal methods, use of softmax, and loss functions.

5.8 CornerNet and CornerNet lite

CornerNet (Law and Deng, 2018) and CornerNet-Lite (Law et al., 2019) are complex object detection algorithms designed for irregularly shaped objects. These architectures utilize a keypoint-based approach to predict objects rather than relying on anchor boxes. This design enhances robustness to object orientation by being rotation-invariant and achieves higher accuracy while eliminating the hyperparameter tuning associated with anchor boxes. CornerNet employs a single-stage architecture consisting of three key steps: corner heatmap prediction to identify the corner points of objects throughout the image, embedding prediction to locate key points with associated class information, and box refinement to finalize bounding boxes along with class probabilities. CornerNet-Lite addresses the computational limitations of CornerNet while enhancing accuracy. It introduces two key innovations: CornerNet Saccade, which reduces unnecessary computations through an attention mechanism focused on key points, and CornerNet Squeeze, which ensures efficient feature extraction compatible with the backbone architecture. These advancements make CornerNet-Lite suitable for real-time applications.

5.9 Datasets

Datasets are crucial in testing and training object detection models, enabling researchers to create more accurate and adaptable algorithms. These datasets feature detailed annotations, such as segmentation masks and bounding boxes, which enable precise object localization and classification. They encompass a variety of object classes and contexts, making them suitable for numerous computer vision applications. The availability of such datasets has significantly advanced object detection, resulting in the development of more sophisticated and reliable detection models.

5.9.1 General purpose datasets

The COCO dataset comprises over 200,000 images captured in diverse environments, encompassing various objects and scenarios. Its primary goal is to enhance object recognition and segmentation models by providing a comprehensive benchmark and encouraging algorithms to manage diverse categories and complex scenarios. COCO includes detailed annotations for object instances, segmentation masks, and key points. These annotations are invaluable for training models to identify and differentiate objects, particularly in situations involving partial occlusion or intricate shapes. Furthermore, these annotations are especially

beneficial for developing models that perform reliably in real-world scenarios.

The Pascal VOC dataset, available in two editions (2007 and 2012), consists of over 20,000 images spanning various object categories and backgrounds. Pascal VOC has been a cornerstone in the development and evaluation of object detection models, serving as a benchmark for early detection methods. Its annotated bounding boxes support tasks such as object recognition and localization. These annotations are useful for assessing the performance of detection models in realistic settings where objects may be partially obscured or exhibit complex geometries.

ImageNet, one of the largest image datasets, contains over 14 million meticulously labeled images across numerous object categories. It serves as the foundation for training and evaluating large-scale object detection and recognition algorithms. ImageNet has played a crucial role in advancing deep learning in computer vision, providing extensive resources for the development of innovative algorithms. Its detailed annotations allow models to learn and identify a wide range of objects with high accuracy, significantly enhancing their performance and robustness across numerous applications. Table 8 summarizes popular object detection datasets, detailing their training, validation, and testing statistics, including the number of images and objects in each dataset.

5.9.2 Domain-specific dataset

Domain-specific datasets, which provide specialized data tailored to particular application domains, are vital for advancing object detection research. These curated datasets enable practitioners and researchers to develop highly accurate and effective object detection models by addressing the unique requirements and challenges of specific industries or scenarios. They include targeted annotations, diverse object classes, and relevant contextual information, making them essential for training and evaluating object detection algorithms in real-world environments. Domains such as autonomous driving, medical imaging, retail, and agriculture benefit significantly from this customized approach, which enhances model performance within specific domains and fosters innovation in specialized object detection research.

The KITTI Vision Benchmark Suite is a prominent dataset for autonomous driving. It offers annotations for objects such as cars, pedestrians, and bicycles alongside images and LiDAR data from diverse scenarios. Similarly, BDD100K is another extensive dataset for autonomous driving, featuring detailed object labels and a wide range of driving conditions. NuScenes, designed for urban scene understanding, provides large-scale object annotations across complex urban landscapes.

In medical imaging, specialized datasets address tasks such as organ segmentation and tumor detection using modalities such as computed tomography (CT) scans and X-rays. Notable examples include datasets developed under the Medical Image Computing and Computer-Assisted Intervention Society (MICCAI). These datasets have significantly contributed to research and development in medical imaging analysis, leading to advancements in diagnostic accuracy and improved treatment planning.

Table 9 provides a comparative analysis of various object detection models evaluated on the Microsoft COCO dataset,

TABLE 7 Hybrid approaches of object detection algorithms summary.

Algorithm	Researchers (Year)	Region proposal	Notes	Softmax incorporated	Loss function	References
Cascade R-CNN	Zhaowei Cai, Tsung-Yi Lin, Wei Wei, Songtao Xu (2018)	Two-stage: uses Faster R-CNN for region proposals in the first stage. Subsequent stages refine proposals based on previous predictions.	True hybrid, combining one-stage (initial proposals) and two-stage (refinement).	Yes	Multi-stage learning with SGD: Smooth L1 loss for bounding box regression, softmax loss for classification	Cai and Nuno, 2018
Mask R-CNN with Cascaded Head	Kailin He, Georgia Gkioxari, Piotr Dollar, Ross Girshick (2017)	Two-stage: Uses standard Mask R-CNN for region proposals (no cascade for proposals). Cascaded structure applies only for the mask prediction branch.	Not necessarily a hybrid in the context of region proposals.	Yes	Multi-stage learning with SGD: Smooth L1 loss for bounding box regression, softmax loss for classification, binary cross-entropy loss for mask prediction	He et al., 2017
Libra R-CNN	Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, Dahua Lin (2019)	Hybrid: Chooses between YOLOv2 (one-stage) and Faster R-CNN (two-stage) proposals based on confidence scores.	True hybrid, dynamically using both types of proposals.	Yes	Multi-task learning with SGD: Focal loss for classification, Smooth L1 loss for bounding box regression	Pang et al., 2019
RegNet	Jing Xu, Yu Pan, Xinglin Pan, Steven Hoi (2021)	One-stage: No explicit region proposals. Uses anchor boxes for candidate object locations.	Not a hybrid in terms of region proposals.	Yes	Multi-task learning with SGD: Focal loss for classification, smooth L1 loss for bounding box regression	Xu et al., 2021

TABLE 8 Popular object detection datasets and their statistics (Zou et al., 2023).

Dataset	Train		Validation		TrainVal		Test	
	Image	Objects	Image	Objects	Image	Objects	Image	Objects
VOC-2007	2,501	6,301	2,510	6,307	5,011	12,608	4,952	14,976
VOC-2007	5,717	13,609	5,823	13,841	11,540	27,450	10,991	-
ILSVRC-2014	456,567	478,807	20,121	55,502	476,688	534,309	40,152	-
ILSVRC-2017	456,567	478,807	20,121	55,502	476,688	534,309	65,500	-
MS-COCO-2015	82,783	604,907	40,504	291,875	123,287	896,782	81,434	-
MS-COCO-2017	118,287	860,001	5,000	36,781	123,287	896,782	40,670	-
Objects 365-2019	600,000	9,623,000	38,000	479,000	628,000	10,102,000	100,000	170,000
OID-2020	1,743,042	14,610,229	41,620	303,980	1,784,662	14,914,209	125,436	937,327

detailing key performance metrics such as AP and AP at different intersections over union thresholds. The results highlight advancements in object detection technologies, demonstrating that newer models outperform their predecessors across multiple performance metrics.

6 Salient object detection

Salient object detection, also called visual saliency detection, is a domain of computer vision dedicated to

identifying the most significant or visually distinctive regions in an image. These regions often correspond to areas that naturally capture human attention, similar to how our eyes instinctively focus on specific elements within a scene. By leveraging advanced algorithms, salient object detection identifies these visually unique areas, facilitating applications such as image understanding, object recognition, and scene analysis. This capability enhances tasks such as content-based image editing, image retrieval, segmentation, and cropping, making it a vital area of computer vision research and application development.

TABLE 9 Comparison of Object Detection Models in Microsoft COCO (Zhao et al., 2019).

Models	Backbone architecture	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
Fast R-CNN (Girshick, 2015)	ResNet	20.5	39.9	19.4	4.1	20.0	35.8
ION (Bell et al., 2016)	-	23.6	43.2	23.6	6.4	24.1	38.2
OHEM+FRCN (Shrivastava et al., 2016)	VGG16	22.6	42.5	22.2	5.0	23.7	34.6
Faster R-CNN (Ren et al., 2015)	ResNet	24.2	45.3	23.5	7.7	26.4	37.1
YOLOv2 (Redmon and Farhadi, 2017)	Darknet	21.6	44.0	19.2	5.0	22.4	35.5
SSD300 (Liu et al., 2016)	VGG16	23.2	41.2	23.4	5.3	23.2	39.6
SSD512 (Liu et al., 2016)	ResNet-50	26.8	46.5	27.8	9.0	28.9	41.9
R-FCN (Dai et al., 2016)	ResNet101	29.2	51.5	-	10.8	32.8	45.0
R-FCN (multi-scale training) (Dai et al., 2016)	ResNet101	29.9	51.9	-	10.4	32.4	43.3
FPN (Lin et al., 2017a)	ResNet101	36.2	59.1	39.0	18.2	39.0	48.2
Mask R-CNN (He et al., 2017)	ResNet101+FPN	38.2	60.3	41.7	20.1	41.1	50.2
Mask R-CNN (He et al., 2017)	ResNeXt101+FPN	39.8	62.3	43.4	22.1	43.2	51.2
DSSD513 (Fu et al., 2017)	ResNet101	33.2	53.3	35.2	13.0	35.4	51.1

Identifying salient objects can be likened to solving a complex mystery. The bottom-up (BU) approach (Tu et al., 2016) acts like a meticulous investigator, analyzing local features such as edges and spatial information. However, its limited perspective often results in low-contrast and blurry "saliency maps," resembling vague shadows rather than well-defined objects. Conversely, the top-down (TD) approach (Yang and Yang, 2016) functions as a strategic analyst, utilizing prior knowledge about object types to refine the saliency map and emphasize the object's key features. For instance, in semantic segmentation tasks, where individual pixels are classified, the TD approach enhances the clarity and accuracy of BU-detected details, ensuring that the proper structure and boundaries of the object are effectively captured (Gao et al., 2009).

6.1 Deep learning for salient object detection

CNNs are pivotal in high-level and multi-scale feature representation within salient object detection. These architectures have proven effective in various computer vision tasks, including edge detection, object recognition, and semantic segmentation. Eleonora (Vig et al., 2014) pioneered a data-driven approach, leveraging deep networks with diverse layers and parameters to maximize feature extraction. Similarly, Kümmerer et al. (2014) introduced Deep Gaze, which utilized AlexNet to create a high-dimensional feature space for saliency mapping, addressing challenges posed by limited training data. Extending this idea, Huang et al. (2015) fine-tuned pre-trained object recognition deep networks using saliency evaluation metrics such as Similarity and KL-Divergence. Numerous strategies have since been developed to enhance the integration of local and global visual cues for salient object detection. For instance, Wang et al. employed two separate deep CNNs to capture both local and global features,

while Cholakal et al. (2018) proposed a weakly supervised system that fuses top-down and bottom-up saliency maps, refining them through multi-scale superpixel averaging. Additionally, Zhao et al. (2015) designed a multi-context deep learning framework using superpixel segmentation to combine local and global contextual modeling.

Efforts to incorporate context modeling and semantic information into salient object detection have also shown promising results. For example, Li et al. (2016) proposed a multi-task deep saliency model that creates intrinsic connections between saliency detection and semantic segmentation. In contrast, He et al. (2015b) introduced SuperCNN, a superpixel-based CNN aimed at enhancing performance.

The integration of multi-scale feature maps has proven crucial for improving detection accuracy. Liu et al. (2015) utilized CNNs for fixation prediction by jointly learning visual saliency components, while Wang et al. (2015) introduced RegionNet, which preserves edges and incorporates multi-scale contextual modeling for salient object detection. The evolution of deep learning techniques in salient object detection demonstrates a continuous trajectory toward more accurate and efficient solutions, solidifying its importance in computer vision research (Gao et al., 2009).

6.2 Benchmark datasets and evaluation metrics

ECSSD (Yan et al., 2013), HKU-IS (Li and Yu, 2016b), PASCALS (Li et al., 2014), and SOD (Movahedi and Elder, 2010) are widely recognized benchmark datasets for evaluating the performance of salient object detection methods. ECSSD contains over 4,000 challenging images characterized by low contrast and

TABLE 10 Comparison between the state-of-art methods in salient object detection.

Dataset	Metrics	CHM (Li et al., 2013)	RC (Cheng et al., 2014)	DRFI (Jiang et al., 2013)	MC (Zhao et al., 2015)	MDF (Li and Yu, 2016b)	DSR (Tang et al., 2016)	DCL (Li and Yu, 2016a)	ELD (Lee et al., 2016)	NLDF (Luo et al., 2017)	DSSC (Hou et al., 2017)
PASCALS	MAE	0.222	0.225	0.221	0.147	0.145	0.128	0.108	0.121	0.099	0.080
ECSSD	MAE	0.195	0.187	0.166	0.107	0.108	0.037	0.071	0.098	0.063	0.052
HKU-IS	MAE	0.058	0.165	0.143	0.098	0.129	0.040	0.048	0.071	0.048	0.039
SOD	MAE	0.249	0.242	0.215	0.184	0.155	-	0.126	0.154	0.143	0.118

The bold values (lowest score) indicate the best performance.

multiple salient objects, while HKU-IS comprises over 1,000 semantically rich and complex natural images. PASCALS originates from the validation set of the PASCAL VOC 2010 segmentation dataset, consisting of 850 natural images. In comparison, the SOD dataset includes 300 images, each featuring multiple salient objects. Adhering to the standardized training and validation splits proposed by Jiang et al. (2013) ensures a rigorous and consistent evaluation of methodologies.

Saliency map evaluation primarily relies on two metrics: Mean Absolute Error (MAE) and F-measure. The F-measure quantifies saliency map quality through precision and recall, computed based on the intersection of the generated binary mask B with a ground truth Z . These datasets collectively cover diverse image attributes and complexities, enabling a comprehensive assessment of salient object detection techniques.

$$F_{\beta} = \frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \quad (8)$$

where β^2 is set to 0.3 to emphasize how crucial the precision value is. Using the following formula, the MAE score is calculated.

The Mean Absolute Error (MAE) is calculated using the following equation:

$$\text{MAE} = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W |\hat{S}(i, j) - \hat{Z}(i, j)| \quad (9)$$

It represents the average absolute difference between the predicted values $\hat{S}$ denote the value at position (i, j) in matrix S , and $\hat{Z}$ represents the ground truth value at the corresponding position in the matrices. The Mean Absolute Error is calculated by taking the absolute difference between each corresponding pair of values in the matrices, summing up these absolute differences, and then dividing by the total number of elements ($H \times W$) in the matrices, where the W and H are the width and height of the salient area.

This research looks at some salient feature object detection methods, such as deep learning-focused and classical methods. Notable for their exceptional performance are the Context-aware Hierarchical Model (CHM) (Li et al., 2013), Region Contrast (RC) (Cheng et al., 2014), and Discriminative Region Feature Integration (DRFI) (Jiang et al., 2013). CNN is the foundation of other methods, including Multi-Contextual (MC) (Zhao et al., 2015), Multi-level Deep Feature Integration (MDF) (Li and Yu, 2016b), Deep contrast learning (DCL) (Li and Yu,

2016a) Edge-Loss with Diverse-thresholding (ELD) (Lee et al., 2016), Non-Local Deep Features (NLDF) (Luo et al., 2017), and Deep Scale Selection and Classification (DSSC) (Hou et al., 2017).

In general, CNN-based techniques outperform classical approaches; the Table 10 shows the evaluation metrics, F-measure, and Mean Absolute Error (MAE). In particular, MC and MDF make better saliency forecasts by utilizing data from both local and global settings. ELD provides additional information by taking advantage of low-level custom features. LEGS employs generic region recommendations for the first salient regions, which might not be sufficient. Future directions for improvement are suggested by integrating semantic segmentation and recurrent networks in DSR and MT. Multi-scale representations and superpixel segmentation are necessary for DCL, NLDF, and DSSC to produce highly salient regions and smooth boundaries. Among these, DCL, NLDF, and DSSC show the best performance on all four datasets; scale-to-scale short connection modeling allows DSSC to show the best performance.

Most CNN-based techniques require using superpixel segmentation to simulate visual saliency along area boundaries because CNN primarily provides salient information in small regions. Measuring local conspicuity requires extracting multi-scale deep CNN features. Strengthening local connections between several CNN layers and incorporating complementing data from local and global contexts is considered vital.

7 Challenges and future opportunities

The use of CNN examines potential advancements in object detection. This study highlights the importance of enhancing object identification methods to strike a balance between speed and accuracy. Two-stage and hybrid detection systems provide greater precision at the cost of increased computational complexity, while one-stage alternatives prioritize quicker data processing with a certain level of accuracy. Future research will creatively address this trade-off by developing systems that are both precise and efficient.

However, there are several challenges in object detection, such as occlusion, where items may be hidden by other objects, leading to inaccurate detection. Additionally, the less noticeable characteristics of small or distant objects complicate recognition. Moreover, object detection algorithms encounter difficulties in situations with overlapping objects and issues related

to illumination and viewpoint. Furthermore, the limited availability of data and the lack of annotated data hinder effective model training. Additionally, the incorporation of multi-modal detection affects the overall performance of object identification systems. However, the resolution and image processing techniques impose restrictions on this integration.

It is critical to ensure the stability and dependability of object-detecting systems in various situations. Models can be made more general using domain adaptation and transfer learning strategies, which will help them function well in novel environments. Combining many modalities of information, including textual or temporal signals, might increase the accuracy of complicated scene interpretation and improve contextual understanding. To balance this trade-off and create object identification algorithms that are both extremely precise and computationally economical, researchers frequently employ multi-task loss functions to penalize misclassifications and localization errors.

8 Conclusion

The rapid evolution of object detection algorithms marks a transformative era in image and pattern recognition, enabling groundbreaking advancements in visual perception and interaction. This study has comprehensively reviewed object detection methodologies, ranging from single-stage to two-stage and hybrid approaches. While single-stage methods excel in speed and computational efficiency, two-stage and hybrid approaches demonstrate superior accuracy and detection precision, making them highly suitable for real-world applications. By analyzing architectural frameworks, backbone structures, and loss functions, this study emphasizes the critical importance of iterative improvement to address the growing demands of modern technological applications. The developments discussed pave the way for revolutionary advancements in domains such as autonomous vehicles, surveillance systems, and broader image recognition tasks, fundamentally reshaping how humans interact with the visual environment.

Future research must prioritize the integration of multimodal data by combining textual, contextual, and visual signals to improve robustness and contextual sensitivity in object detection models. This multidisciplinary approach holds promise for innovative applications in multimedia analysis, augmented reality, and human-computer interaction. Furthermore, scalable and parallelizable object detection systems are essential for meeting the growing demand for real-time processing of large image and video datasets. Advances in distributed computing, edge computing, and hardware acceleration will be crucial for deploying these systems in resource-constrained environments.

Equally important is the need to address object detection technologies' social and ethical implications. Privacy, bias, and fairness concerns must be rigorously examined to ensure responsible and equitable deployment across diverse societal contexts. Future research should strive to develop frameworks and policies that safeguard these principles, fostering the ethical adoption of object detection systems. By aligning technological innovation with ethical accountability, the field can ensure its advancements serve humanity responsibly while

unlocking unprecedented opportunities for creative and practical applications.

Author contributions

BL: Conceptualization, Investigation, Methodology, Resources, Writing – original draft, Writing – review & editing. GS: Formal analysis, Investigation, Project administration, Supervision, Writing – original draft. TH: Conceptualization, Methodology, Project administration, Supervision, Validation, Funding acquisition, Writing – original draft.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The research was financially supported by Center of Excellence in Digital Earth and Emerging Technology (CoE:DEET), Thammasat University, Thailand.

Acknowledgments

BL acknowledges the Faculty-Quota Scholarship awarded by SIIT, Thammasat University. The authors would like to thank the Advance Geospatial Technology Research Unit, SIIT, and the Center of Excellence in Digital Earth and Emerging Technology (CoE: DEET), Thammasat University, for providing a technical environment and supporting information. We gratefully acknowledge the financial support provided by the Thammasat University Research Fund. The authors acknowledge the assistance of ChatGPT developed by OpenAI, for helping to refine the content and generate some figures for this work.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fcomp.2025.1437664/full#supplementary-material>

References

- Aggarwal, A., and Kumar, M. (2021). Image surface texture analysis and classification using deep learning. *Multimed. Tools Appl.* 80, 1289–1309. doi: 10.1007/s11042-020-09520-2
- Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., et al. (2021). Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *J. Big Data* 8, 1–74. doi: 10.1186/s40537-021-00444-8
- Amjoud, A. B., and Amrouch, M. (2023). Object detection using deep learning, cnns and vision transformers: a review. *IEEE Access* 11, 35479–35516. doi: 10.1109/ACCESS.2023.3266093
- Ardia, D., Ringel, E., Ekstrand, V. S., and Fox, A. (2020). “Addressing the decline of local news, rise of platforms, and spread of mis- and disinformation online,” in *UNC Center for Media Law and Policy*.
- Awad, M., Khanna, R., Awad, M., and Khanna, R. (2015). “Support vector machines for classification,” in *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, 39–66.
- Bai, Y. (2022). “Relu-function and derived function review,” in *SHS Web of Conferences* (Les Ulis: EDP Sciences), 02006.
- Bell, S., Zitnick, C. L., Bala, K., and Girshick, R. (2016). “Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 2874–2883.
- Ben Braiek, H., and Khomh, F. (2023). Testing feedforward neural networks training programs. *ACM Trans. Softw. Eng. Methodol.* 32, 1–61. doi: 10.1145/3529318
- Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*. doi: 10.48550/arXiv.2004.10934
- Buckner, C. (2019). Deep learning: a philosophical introduction. *Philosophy Comp.* 14e12625. doi: 10.1111/phc3.12625
- Cai, Z., and Nuno, N. (2018). “Cascade R-CNN: delving into high quality object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 6154–6162. doi: 10.1109/CVPR.2018.00644
- Cao, J., Cholakal, H., Anwer, R. M., Khan, F. S., Pang, Y., and Shao, L. (2020). “D2det: Towards high quality object detection and instance segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 11485–11494.
- Chen, G., Choi, W., Yu, X., Han, T., and Chandraker, M. (2017). “Learning efficient object detection models with knowledge distillation,” in *Advances in Neural Information Processing Systems*, 30.
- Cheng, M.-M., Mitra, N. J., Huang, X., Torr, P. H., and Hu, S.-M. (2014). Global contrast based salient region detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 569–582. doi: 10.1109/TPAMI.2014.2345401
- Chhabra, M., Ravulakollu, K. K., Kumar, M., Sharma, A., and Nayyar, A. (2023). Improving automated latent fingerprint detection and segmentation using deep convolutional neural network. *Neural Comp. Appl.* 35, 6471–6497. doi: 10.1007/s00521-022-07894-y
- Chhabra, M., Sharan, B., Elbarachi, M., and Kumar, M. (2024). “Intelligent waste classification approach based on improved multi-layered convolutional neural network” in *Multimedia Tools and Applications*, Vol. 83, 84095–84120. doi: 10.1007/s11042-024-18939-w
- Choi, H., Ryu, S., and Kim, H. (2018). “Short-term load forecasting based on resnet and lstm,” in *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)* (Aalborg: IEEE), 1–6.
- Cholakal, H., Johnson, J., and Rajan, D. (2018). Backtracking spatial pyramid pooling-based image classifier for weakly supervised top-down salient object detection. *IEEE Trans. Image Proc.* 27, 6064–6078. doi: 10.1109/TIP.2018.2864891
- Clevert, D.-A., Unterthiner, T., and Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (ELUS). *arXiv preprint arXiv:1511.07289*. doi: 10.48550/arXiv.1511.07289
- Cristianini, N., and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge: Cambridge University Press.
- Dai, J., Li, Y., He, K., and Sun, J. (2016). “R-FCN: object detection via region-based fully convolutional networks,” in *Advances in Neural Information Processing Systems*, 29.
- Dalal, N., and Triggs, B. (2005). “Histograms of oriented gradients for human detection,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)* (San Diego, CA: IEEE), 886–893.
- Dong, X., Zheng, L., Ma, F., Yang, Y., and Meng, D. (2018). Few-example object detection with model communication. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 1641–1654. doi: 10.1109/TPAMI.2018.2844853
- Dong, Z., Li, G., Liao, Y., Wang, F., Ren, P., and Qian, C. (2020). “Centripetalnet: pursuing high-quality keypoint pairs for object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10519–10528. doi: 10.1109/CVPR42600.2020.01053
- Dundar, A., Jin, J., Martini, B., and Culurciello, E. (2016). Embedded streaming deep neural networks accelerator with applications. *IEEE Trans. Neural Netw. Learn. Syst.* 28, 1572–1583. doi: 10.1109/TNNLS.2016.2545298
- Elgendy, M. (2020). *Deep Learning for Vision Systems*. New York: Simon and Schuster.
- Erhan, D., Szegedy, C., Toshev, A., and Anguelov, D. (2014). “Scalable object detection using deep neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH: IEEE), 2147–2154.
- Freeman, W. T., and Roth, M. (1995). “Orientation histograms for hand gesture recognition,” in *International Workshop on Automatic Face and Gesture Recognition* (Princeton, NJ: Citeseer), 296–301.
- Freund, Y., Schapire, R., and Abe, N. (1999). A short introduction to boosting. *J.-Jap. Soc. Artif. Intellig.* 14:1612.
- Fu, C.-Y., Liu, W., Ranga, A., Tyagi, A., and Berg, A. C. (2017). DSSD: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*. doi: 10.48550/arXiv.1701.06659
- Fukushima, K. (1975). Cognitron: A self-organizing multilayered neural network. *Biol. Cybern.* 20, 121–136. doi: 10.1007/BF00342633
- Gao, D., Han, S., and Vasconcelos, N. (2009). Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 989–1005. doi: 10.1109/TPAMI.2009.27
- Ghiasi, G., Lin, T.-Y., and Le, Q. V. (2019). “NAS-FPN: Learning scalable feature pyramid architecture for object detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (Long Beach, CA: IEEE), 7036–7045. Available online at: <https://arxiv.org/abs/1904.07392>
- Girshick, R. (2015). “Fast R-CNN,” in *Proceedings of the IEEE International Conference on Computer Vision* (Santiago: IEEE), 1440–1448.
- Girshick, R., Donahue, J., Darrell, T., Berkeley, U., and Malik, J. (2014a). “R-CNN: region-based convolutional neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2–9.
- Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014b). “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH: IEEE), 580–587. Available online at: <https://arxiv.org/abs/1311.2524>
- Grimaldi, M., Tenace, V., and Calimera, A. (2018). Layer-wise compressive training for convolutional neural networks. *Future Internet* 11(1):7. doi: 10.3390/fi11010007
- Guan, L. (2017). *Multimedia Image and Video Processing*. Boca Raton, FL: CRC Press.
- Guo, C., Fan, B., Gu, J., Zhang, Q., Xiang, S., Prinnet, V., et al. (2019). “Progressive sparse local attention for video object detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul: IEEE), 3909–3918.
- Hammoudeh, M. A. A., Alsaykhan, M., Alsalamah, R., and Althwaibi, N. (2022). Computer vision: a review of detecting objects in videos-challenges and techniques. *Int. J. Online Biomed. Eng.* 18:27577. doi: 10.3991/ijoe.v18i01.27577
- He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). “Mask R-CNN,” in *Proceedings of the IEEE International Conference on Computer Vision* (Venice: IEEE), 2961–2969. doi: 10.1109/ICCV.2017.322
- He, K., Zhang, X., Ren, S., and Sun, J. (2015a). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 1904–1916. doi: 10.1109/TPAMI.2015.2389824
- He, K., Zhang, X., Ren, S., and Sun, J. (2016a). “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 770–778.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016b). “Identity mappings in deep residual networks,” in *Computer Vision-ECCV 2016: 14th European Conference* (Amsterdam: Springer), 630–645.
- He, S., Lau, R. W., Liu, W., Huang, Z., and Yang, Q. (2015b). Supercnn: a superpixelwise convolutional neural network for salient object detection. *Int. J. Comput. Vis.* 115, 330–344. doi: 10.1007/s11263-015-0822-0
- Hereman, W., and Malfliet, W. (2005). “The tanh method: a tool to solve nonlinear partial differential equations with symbolic software,” in *Proceedings 9th World Multi-Conference on Systems, Cybernetics and Informatics* (Orlando, FL: IEEE), 165–168.
- Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

- Hou, Q., Cheng, M.-M., Hu, X., Borji, A., Tu, Z., and Torr, P. H. (2017). "Deeply supervised salient object detection with short connections," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 3203–3212.
- Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 4700–4708.
- Huang, X., Shen, C., Boix, X., and Zhao, Q. (2015). Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 262–270. doi: 10.1109/ICCV.2015.38
- Huang, Z., Wang, J., Fu, X., Yu, T., Guo, Y., and Wang, R. (2020). Dc-spp-yolo: dense connection and spatial pyramid pooling based yolo for object detection. *Inf. Sci.* 522, 241–258. doi: 10.1016/j.ins.2020.02.067
- Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., and Li, S. (2013). "Salient object detection: a discriminative regional feature integration approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 9 Portland, OR: IEEE), 2083–2090.
- Jocher, G., and Qiu, J. (2024). *Ultralytics Yolo11*. Available online at: <https://github.com/ultralytics/yolov5>
- Jocher, G., Stoken, A., Borovec, J., Chaurasia, A., Changyu, L., Hogan, A., et al. (2021). *Ultralytics/yolov5: v5.0-yolov5-p6 1280 Models, AWS, Supervise. LY and Youtube Integrations*. Geneva: Zenodo.
- Kaur, R., and Singh, S. (2023). A comprehensive review of object detection with deep learning. *Digit. Signal Process.* 132:103812. doi: 10.1016/j.dsp.2022.103812
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 25. Available online at: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi: 10.1145/3065386
- Kümmerer, M., Theis, L., and Bethge, M. (2014). Deep gaze I: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*. doi: 10.48550/arXiv.1411.1045
- Law, H., and Deng, J. (2018). "Cornersnet: detecting objects as paired keypoints," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Cham: Springer), 734–750.
- Law, H., Teng, Y., Russakovsky, O., and Deng, J. (2019). Cornernet-lite: efficient keypoint based object detection. *arXiv preprint arXiv:1904.08900*. doi: 10.48550/arXiv.1904.08900
- Lazebnik, S., Schmid, C., and Ponce, J. (2006). "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (New York, NY: IEEE), 2169–2178.
- Lee, G., Tai, Y.-W., and Kim, J. (2016). "Deep saliency with encoded low level distance map and high level features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 660–668.
- Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., et al. (2022). Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*. doi: 10.48550/arXiv.2209.02976
- Li, G., and Yu, Y. (2016a). "Deep contrast learning for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 478–487.
- Li, G., and Yu, Y. (2016b). Visual saliency detection based on multiscale deep cnn features. *IEEE trans. Image Proc.* 25, 5012–5024. doi: 10.1109/TIP.2016.2602079
- Li, Q., Niaz, U., and Merialdo, B. (2012). "An improved algorithm on viola-jones object detector," in *2012 10th International Workshop on Content-Based Multimedia Indexing (CBMI)* (Annecy: IEEE), 1–6.
- Li, X., Li, Y., Shen, C., Dick, A., and Van Den Hengel, A. (2013). "Contextual hypergraph modeling for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision* (Sydney: IEEE), 3328–3335.
- Li, X., Zhao, L., Wei, L., Yang, M.-H., Wu, F., Zhuang, Y., et al. (2016). Deepsaliency: Multi-task deep neural network model for salient object detection. *IEEE Trans. Image Proc.* 25, 3919–3930. doi: 10.1109/TIP.2016.2579306
- Li, Y., Chen, Y., Wang, N., and Zhang, Z. (2019). "Scale-aware trident networks for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul: IEEE), 6054–6063.
- Li, Y., Hou, X., Koch, C., Reh, J. M., and Yuille, A. L. (2014). "The secrets of salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH: IEEE) 280–287.
- Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017a). "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE) 2117–2125. Available online at: <https://arxiv.org/abs/1612.03144>
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017b). "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice: IEEE), 2980–2988. Available online at: <https://arxiv.org/abs/1708.02002>
- Liu, N., Han, J., Zhang, D., Wen, S., and Liu, T. (2015). "Predicting eye fixations using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston: IEEE), 362–370.
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., et al. (2016). "SSD: Single shot multibox detector," in *Computer Vision-ECCV 2016: 14th European Conference* (Amsterdam: Springer), 21–37.
- Long, X., Deng, K., Wang, G., Zhang, Y., Dang, Q., Gao, Y., et al. (2020). Pp-yolo: An effective and efficient implementation of object detector. *arXiv preprint arXiv:2007.12099*. doi: 10.48550/arXiv.2007.12099
- Lowe, D. G. (1999). "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision* (Kerkyra: IEEE), 1150–1157.
- Luo, Z., Mishra, A., Achkar, A., Eichel, J., Li, S., and Jodoin, P.-M. (2017). "Non-local deep features for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 6609–6617.
- Movahedi, V., and Elder, J. H. (2010). "Design and perceptual validation of performance measures for salient object segmentation," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops* (San Francisco, CA: IEEE), 49–56.
- Najibi, M., Rastegari, M., and Davis, L. S. (2016). "G-CNN: an iterative grid based object detector," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 2369–2377.
- Nash, W., Drummond, T., and Birbilis, N. (2018). A review of deep learning in the study of materials degradation. *NPJ Mater Degrad.* 2:37. doi: 10.1038/s41529-018-0058-x
- Nwankpa, C., Ijomah, W., Gachagan, A., and Marshall, S. (2018). Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint arXiv:1811.03378*. doi: 10.48550/arXiv.1811.03378
- Ouyang, W., Wang, X., Zeng, X., Qiu, S., Luo, P., Tian, Y., et al. (2015). "Deepid-net: Deformable deep convolutional neural networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA: IEEE), 2403–2412.
- Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., and Lin, D. (2019). "Libra R-CNN: towards balanced learning for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 821–830. Available online at: <https://arxiv.org/abs/1904.02701>
- Pinheiro, P. O., Lin, T.-Y., Collobert, R., and Dollár, P. (2016). "Learning to refine object segments," in *Computer Vision-ECCV 2016: 14th European Conference* (Amsterdam: Springer), 75–91.
- Ramachandran, P., Zoph, B., and Le, Q. V. (2017). Searching for activation functions. *arXiv [preprint] arXiv:1710.05941*. doi: 10.48550/arXiv.1710.05941
- Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 779–788.
- Redmon, J., and Farhadi, A. (2017). "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 7263–7271. doi: 10.1109/CVPR.2017.690
- Redmon, J., and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv [preprint] arXiv:1804.02767*. doi: 10.48550/arXiv.1804.02767
- Reis, D., Kupec, J., Hong, J., and Daoudi, A. (2023). Real-time flying object detection with yolov8. *arXiv [preprint] arXiv:2305.09972*. doi: 10.48550/arXiv.2305.09972
- Ren, S., He, K., Girshick, R., and Sun, J. (2015). "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 28.
- Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. *arXiv [preprint] arXiv:1506.01497*. doi: 10.1109/TPAMI.2016.2577031
- Ren, Y., Zhu, C., and Xiao, S. (2018). Small object detection in optical remote sensing images via modified faster r-cnn. *Appl. Sci.* 8:813. doi: 10.3390/app8050813
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 4510–4520.
- Shah, S., and Tembhurne, J. (2023). Object detection using convolutional neural networks and transformer-based models: a review. *J. Elect. Syst. Inform. Technol.* 10:54. doi: 10.1186/s43067-023-00123-z
- Shrivastava, A., Gupta, A., and Girshick, R. (2016). "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 761–769. doi: 10.1109/CVPR.2016.89

- Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. doi: 10.48550/arXiv.1409.1556
- Sun, Z. (2024). *Parallelization of Hybrid Multi-Objective Evolutionary Algorithm on Multi-Core Architectures*. Available online at: <https://mspace.lib.umanitoba.ca/items/01bac995-c4a2-4f19-9ac7-a13e5fd1bbfc>
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 2818–2826.
- Tang, G., Ni, J., Zhao, Y., Gu, Y., and Cao, W. (2023). A survey of object detection for uavs based on deep learning. *Remote Sens.* 16:149. doi: 10.3390/rs16010149
- Tang, Y., Wu, X., and Bu, W. (2016). "Deeply-supervised recurrent convolutional neural network for saliency detection," in *MM '16: Proceedings of the 24th ACM international conference on Multimedia* (New York, NY: IEEE), 397–401. doi: 10.1145/2964284.2967250
- Tu, W.-C., He, S., Yang, Q., and Chien, S.-Y. (2016). "Real-time salient object detection with a minimum spanning tree," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 2334–2342.
- Van de Sande, K. E., Uijlings, J. R., Gevers, T., and Smeulders, A. W. (2011). "Segmentation as selective search for object recognition," in *2011 International Conference on Computer Vision* (Barcelona: IEEE), 1879–1886.
- Vig, E., Dorr, M., and Cox, D. (2014). "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH: IEEE), 2798–2805.
- Viola, P., and Jones, M. (2001). "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001* (Kauai, HI: IEEE).
- Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al. (2024). Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*. doi: 10.48550/arXiv.2405.14458
- Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2023). "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Vancouver, BC: IEEE), 7464–7475.
- Wang, C.-Y., and Liao, H.-Y. M. (2024). *Yolov9: Learning What You Want to Learn Using Programmable Gradient Information*. Available online at: <https://arxiv.org/abs/2402.13616>
- Wang, L., Lu, H., Ruan, X., and Yang, M.-H. (2015). "Deep networks for saliency detection via local estimation and global search," in *Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition* (Boston, MA: IEEE), 3183–3192.
- Wong, A., Shafiee, M. J., Li, F., and Chwyl, B. (2018). "Tiny SSD: a tiny single-shot detection deep convolutional neural network for real-time embedded object detection," in *2018 15th Conference on computer and robot vision (CRV)* (Toronto, ON: IEEE), 95–101.
- Wu, B., Iandola, F., Jin, P. H., and Keutzer, K. (2017). "Squeezenet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (Honolulu, HI: IEEE), 129–137. Available online at: <https://arxiv.org/pdf/1612.01051>
- Xie, E., Sun, P., Song, X., Wang, W., Liu, X., Liang, D., et al. (2020). "Polarmask: Single shot instance segmentation with polar representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 12193–12202.
- Xiong, Y., Liu, H., Gupta, S., Akin, B., Bender, G., Wang, Y., et al. (2021). "Mobilelets: Searching for object detection architectures for mobile accelerators," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Nashville, TN: IEEE) 3825–3834. doi: 10.1109/CVPR46437.2021.00382
- Xu, J., Pan, Y., Pan, X., Hoi, S., Yi, Z., and Xu, Z. (2021). RegNet: self-regulated network for image classification. *arXiv [Preprint]*. arXiv: 2101.00590. doi: 10.48550/arXiv.2101.00590
- Yan, Q., Xu, L., Shi, J., and Jia, J. (2013). "Hierarchical saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Portland, OR: IEEE) 1155–1162. doi: 10.1109/CVPR.2013.153
- Yang, J., and Yang, M.-H. (2016). Top-down visual saliency via joint crf and dictionary learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 576–588. doi: 10.1109/TPAMI.2016.2547384
- Yao, Y., Cheng, G., Wang, G., Li, S., Zhou, P., Xie, X., et al. (2022). On improving bounding box representations for oriented object detection. *IEEE Trans. Geosci. Remote Sens.* 61, 1–11. doi: 10.1109/TGRS.2022.3231340
- Yoo, D., Park, S., Lee, J.-Y., Paek, A. S., and So Kweon, I. (2015). "Attentionnet: Aggregating weak directions for accurate object detection," in *Proceedings of the IEEE International Conference on Computer Vision* (Santiago: IEEE), 2659–2667.
- Zhang, S., Wen, L., Lei, Z., and Li, S. Z. (2020). Refinedet++: Single-shot refinement neural network for object detection. *IEEE Trans. Circuits Syst. Video Technol.* 31, 674–687. doi: 10.1109/TCSVT.2020.2986402
- Zhang, W., Jiao, L., Liu, X., and Liu, J. (2019). "Multi-scale feature fusion network for object detection in vhr optical remote sensing images," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium* (Yokohama: IEEE), 330–333.
- Zhang, Y., Sohn, K., Villegas, R., Pan, G., and Lee, H. (2015). "Improving object detection with deep convolutional networks via bayesian optimization and structured prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA: IEEE), 249–258.
- Zhao, R., Ouyang, W., Li, H., and Wang, X. (2015). "Saliency detection by multi-context deep learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA: IEEE), 1265–1274.
- Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., and Parmar, M. (2024). A review of convolutional neural networks in computer vision. *Artif. Intellig. Rev.* 57:99. doi: 10.1007/s10462-024-10721-6
- Zhao, Z.-Q., Zheng, P., Xu, S., and Wu, X. (2019). Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 3212–3232. doi: 10.1109/TNNLS.2018.2876865
- Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2021). Deformable DETR: deformable transformers for end-to-end object detection. *arXiv [Preprint]*. arXiv:2010.04159. doi: 10.48550/arXiv.2010.04159
- Zhu, Z., Zhou, Y., Dong, Y., and Zhong, Z. (2023). PWLU: Learning specialized activation functions with the piecewise linear unit. *IEEE Trans. Pattern Anal. Mach. Intellig.* 45, 12269–12286. doi: 10.1109/TPAMI.2023.3286109
- Zou, Z., Chen, K., Shi, Z., Guo, Y., and Ye, J. (2023). Object detection in 20 years: a survey. *Proc. IEEE* 111, 257–276. doi: 10.1109/JPROC.2023.3238524



OPEN ACCESS

EDITED BY
Sokratis Makrogiannis,
Delaware State University, United States

REVIEWED BY
Rajeev Nema,
Manipal University Jaipur, India
Andrey Cherstvy,
University of Potsdam, Germany

*CORRESPONDENCE
Jan Ewald
✉ jan.ewald@uni-leipzig.de

RECEIVED 17 January 2025

ACCEPTED 03 June 2025

PUBLISHED 19 June 2025

CITATION

Joas MJ, Freund D, Haase R, Rahm E and
Ewald J (2025) Leveraging foundation models
and goal-dependent annotations for
automated cell confluence assessment.
Front. Comput. Sci. 7:1562358.
doi: 10.3389/fcomp.2025.1562358

COPYRIGHT

© 2025 Joas, Freund, Haase, Rahm and
Ewald. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Leveraging foundation models and goal-dependent annotations for automated cell confluence assessment

Maximilian Josef Joas¹, Daniel Freund², Robert Haase¹,
Erhard Rahm¹ and Jan Ewald^{1*}

¹Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) Dresden/Leipzig, Leipzig University, Leipzig, Germany, ²Center for Regenerative Therapies, Dresden University of Technology, Dresden, Germany

Mesenchymal stem cell therapy shows promising results for difficult-to-treat diseases, but manufacturing requires robust quality control through cell confluence monitoring. While deep learning can automate confluence estimation, research on cost-effective dataset curation and the role of foundation models in this task is limited. We investigate effective strategies for AI-based confluence estimation by studying active learning, goal-dependent labeling, and foundation models that require no training or labeling effort (zero-shot). Here, we show that zero-shot inference with the Segment Anything Model (SAM) achieves excellent confluence estimation without any task-specific training, outperforming even fine-tuned and specialized models. Moreover, our findings demonstrate that active learning does not significantly enhance training and performance compared to the random selection of training samples in homogeneous cell datasets. We demonstrate that streamlined labeling approaches tailored to specific goals yield results comparable to those of exhaustive, time-consuming annotation methods. Our results challenge common assumptions about dataset curation and model training: neither active learning nor extensive fine-tuning provided significant benefits in our real-world scenarios. Instead, we found that leveraging SAM's zero-shot capabilities and goal-dependent labeling offers the most cost-effective approach for AI-based confluence monitoring. Our study provides practical guidelines for implementing automated cell quality control in MSC manufacturing, demonstrating that extensive dataset curation may be unnecessary when foundation models can effectively handle the task right out of the box.

KEYWORDS

active learning, deep learning, cell segmentation, segment anything model, computer vision

1 Introduction

Mesenchymal stem/stromal cells (MSCs) are powerful Advanced Therapy Medicinal Products (ATMPs) that can treat various conditions. Although MSCs are not yet approved for many applications, they demonstrate promising clinical outcomes in the treatment of degenerative inflammatory diseases, autoimmune disorders, tissue injuries, and chronic degenerative ailments (Strecanska et al., 2024; Galipeau and Sensébé, 2018). This is particularly significant for conditions such as rheumatic arthritis, which affects around 18 million people worldwide, imposes significant health burdens, and lacks sufficient alternative treatments (Shimizu et al., 2023; IHME, 2020).

MSC production, however, relies on limited and hard-to-access sources such as bone marrow or umbilical cord tissue, requiring *ex-vivo* expansion of these cells. During this expansion in cell cultures, the density of the cells (confluence) must be tightly controlled, as it serves as a trigger point for cell differentiation (Fernández-Santos et al., 2022; Kim et al., 2017). Therefore, scientists and laboratory technicians need to harvest the cells before they differentiate and lose their potency as stem cells. Thus, scientists and technicians monitor growth through imaging to ensure quality and optimize yield.

During MSC production, microscopic images are often analyzed manually, and scientists estimate the confluence, i.e., the fraction of area covered by cells. This erroneous and non-standardized process leads to increased labor efforts and fails to optimize yield by missing the optimal harvest point. For AI-based confluence estimation, we need: (a) a method for cell segmentation in live-cell images and (b) data with ground truth, particularly images with known confluence and segmented cells as labels. Traditional image processing techniques include thresholding methods (Shen et al., 2018), such as edge detection, and region-growing approaches (Panagiotakis and Argyros, 2018). However, advances in artificial intelligence (AI) have shown that AI models often outperform traditional methods for cell segmentation (Chen and Murphy, 2023). A typical approach to training AI models involves collecting data, labeling it, and training models from scratch, such as a U-Net model for segmentation (Ronneberger et al., 2015). Recently, pre-trained and large generalist foundation models have gained popularity due to their good performance and broad applicability (Han et al., 2021; Chen and Murphy, 2023). In computer vision, several such models have been developed, including generalists for image segmentation [SAM (Kirillov et al., 2023), Detectron2 (Wu et al., 2019)] and specialists for cell segmentation [Cellpose (Stringer et al., 2020), LiveCell (Edlund et al., 2021)]. Although training custom models has become more accessible to end-users without large computing resources (von Chamier et al., 2021), the main advantage of pre-trained models is that they can be used with little to no labeled images. This is especially important since human labeling is costly and time-consuming. Moreover, MSCs, like other underrepresented cells in live-cell datasets, are non-round and irregularly shaped, making the segmentation task challenging. Additionally, ATMP manufacturing processes prohibit staining for higher contrast. Consequently, generating a sufficiently detailed and diverse training dataset for custom model training would require significant effort, highlighting the utility of pre-trained models.

With both foundation models and untrained models, existing approaches naturally raise the question of how to estimate confluence with minimal labeled data while maintaining sufficient performance. With a large number of images, or data, available, there are three strategies for labeling and applying AI models: (a) zero-shot, which involves no labeling or model training, (b) using all images to label and train models, and (c) active learning, where the n most informative samples are selected for labeling and training. AL is a data-centric approach that reduces the effort of data labeling by selecting the next datapoint(s) not randomly, but based on either uncertainty, diversity, or clustering (Monarch, 2021). Interestingly, using AL to choose only a core set of the entire

dataset can yield results similar to or even better than labeling all the data (Jafari et al., 2024). For our model-driven approach, we focus on uncertainty-based methods due to their broad applicability across various models.

Since current research on cell segmentation (Chen and Murphy, 2023) and AL (Sayin et al., 2021; Monarch, 2021) focus on large datasets in theoretical contexts; we aim to apply AL to real-world small datasets using state-of-the-art models for confluence estimation. In our study, we are interested in four insights:

1. Impact and effectiveness of active learning.
2. Applicability of simplified goal-dependent, i.e., “lazy”, labeling.
3. Active learning selection patterns in a time-resolved cell culture, “movie context.”
4. Performance of zero-shot inference.

To gain these insights, we describe in the following section (cf. Section 2) our experimental setup, which includes three datasets for comparing learning strategies across four models. The selected models range from U-Net (Ronneberger et al., 2015), developed from scratch, to large generalist foundation models such as Detectron2 (Wu et al., 2019) and Meta’s Segment Anything Model (SAM) (Kirillov et al., 2023), as well as the specialist pre-trained model Cellpose (Stringer et al., 2020). By leveraging our experiments, which are graphically summarized in Figure 1, we elaborate on the listed insights point by point in Section 3 and describe their impact, along with their limitations, in the final Section 4.

2 Material and methods

In the following data, we describe models and experiments. To ensure full transparency and reproducibility, we provide all relevant code and scripts in our GitLab repository (<https://git.informatik.uni-leipzig.de/joas/confluence-unet>).

2.1 Data

We utilized three datasets, one of which is labeled with two strategies, resulting in four datasets for model training. Additionally, one dataset was derived from a larger published live-cell imaging dataset [“lc-external” (Edlund et al., 2021)], while three originated from our lab (“internal”). One of these internal datasets contains live-cell imaging data obtained using a CytoSmart Lux microscope (10x magnification; 5-megapixel camera). This dataset was labeled in a standard manner (“lc-internal”) and with goal-dependent labeling (“lc-internal-lazy”) for a direct comparison. In the original LC-internal dataset, each cell was labeled individually, whereas in the “lazy” labeled set, cohesive clusters of cells were labeled as single objects. All images had dimensions of 1280x960 pixels (see Table 1). An additional internal dataset contains standard microscopy (“sc-internal”) images acquired with a ZEISS Axiovert 40 CFL microscope [10x objective Ph1 (phase contrast); AxioCam ERc 5s, 5-megapixel camera]. The images in this dataset measured 512x512 pixels. We filtered the external dataset for the A172 cell line (which has morphological similarities to MSC cells

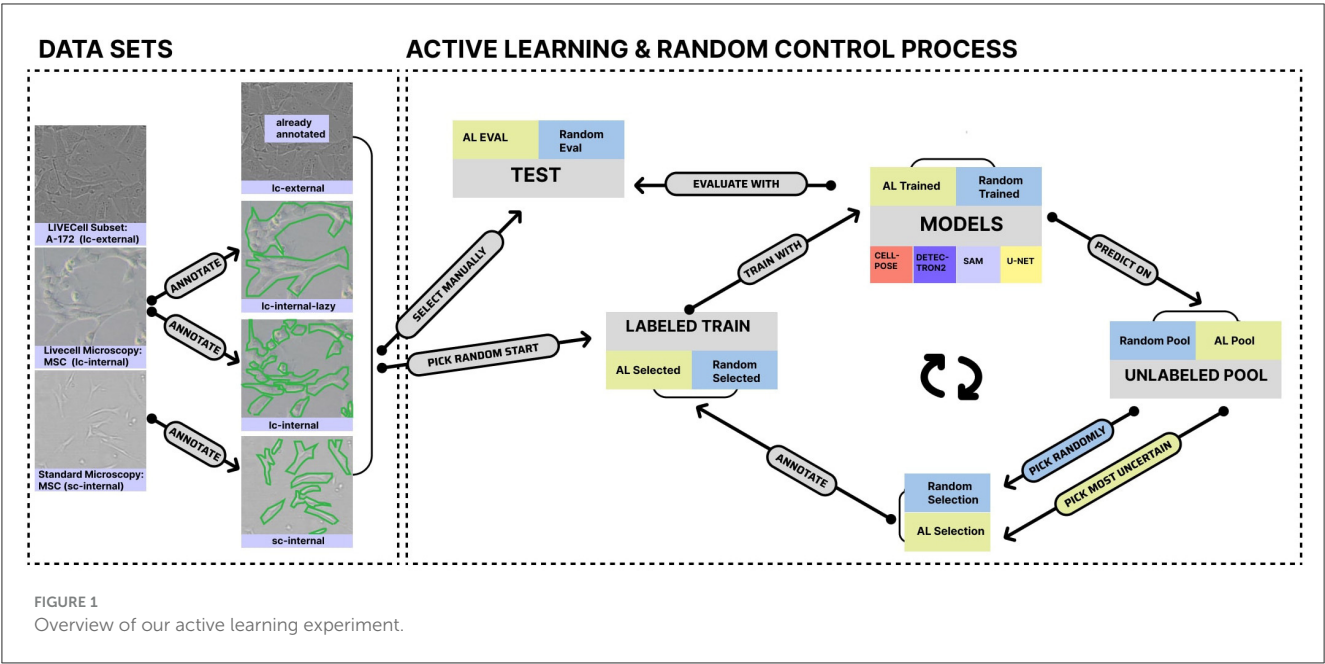


FIGURE 1 Overview of our active learning experiment.

TABLE 1 Dataset characteristics showing the number of regions of interest (ROIs), number of images, and image dimensions for each subset.

Dataset	#ROIs	#Images	Image size
lc-internal-train	2,243	16	1,280 × 960
lc-internal-test	400	2	1,280 × 960
lc-internal-train-lazy	897	16	1,280 × 960
lc-internal-test-lazy	224	4	1,280 × 960
lc-external-train	70,758	333	702 × 520
lc-external-test	30,005	138	702 × 520
sc-train	2,467	12	512 × 512
sc-test	446	3	512 × 512

despite its origin from glioblastomas) to roughly match the cell shapes of the other datasets. Table 1 provides an overview of our dataset, including the number of regions of interest (ROIs), the number of images, and image dimensions. An ROI is defined as a single labeled cell in an image.

We annotated the internal datasets using ImgLab¹ following instructions from wet lab scientists and obtained the annotations in the COCO JSON format (Lin et al., 2014). Since U-Net and Cellpose require instance masks instead of COCO JSON files, we converted the files to masks with custom scripts (https://git.informatik.uni-leipzig.de/joas/confluence/-/blob/main/utills/coco_to_mask.py?ref_type=heads). We divided each dataset into a training set and a testing set to evaluate model performance. To capture a variety of characteristics, we used a combination of datasets, as summarized in Table 1. The LC-external dataset, drawn from the LIVECell paper (Edlund et al., 2021), is the largest, containing the most ROIs. The LC-internal dataset includes full

¹ <https://github.com/NaturalIntelligence/imglab>

and “lazy” subsets and offers the highest resolution. Finally, the sc-internal dataset provides additional data, but with significantly lower resolution and fewer images than the other datasets. Figure 1 shows sample images and annotations for each dataset.

2.2 Active learning for dataset curation

Upon processing the dataset, we manually selected the test set for the internal datasets to ensure that images with approximately 50% confluence were included. This confluence range is the most critical in production, making it essential for evaluating model performance. For AL, we defined a pool set, which can be selected for training, and an initial training set, chosen randomly with a size ranging from two to ten, depending on the total number of images. Selected images in each round were transferred from the pool to the overall train set (either physically on our machine or by filtering the COCO JSON file). We subsequently trained our four models using these initial training sets and in all other rounds for 100 epochs. Figure 1 visually represents this process, and all models were trained on NVIDIA A100 GPUs. To evaluate the performance of our model, we compared the predicted masks with the actual ground-truth masks. We did this by calculating the IoU (Intersection over Union) for each image. IoU measures the overlap between the predicted and actual masks in relation to their combined area. We calculated the mean IoU, standard deviation, and interquartile range across all images in the test set. Additionally, as a second criterion inspired by use cases, we calculated the absolute differences in the confluence between ground truth and predictions.

For our uncertainty-based active learning approach, we created probability maps for all images in the dataset (excluding the Detectron2 model). We then calculated the Shannon entropy on these probability maps to quantify uncertainty. Images in the pool set were ranked according to their entropy values, with those having

the highest entropy (indicating the greatest uncertainty) selected for the next training iteration. To accommodate varying dataset sizes, we chose one image per round from internal datasets and ten images per round from the larger external dataset.

This approach, known as entropy-based sampling (Monarch, 2021; Yin et al., 2023), is a widely used technique in uncertainty-based AL (Zhu et al., 2008). We iteratively retrain the model in each round until all images are transferred from the pool set to the training set. After every training iteration, we assess the model on the predefined test set. To make a direct comparison, we trained another set of models using randomly chosen images. These models adhered to the same training protocol as our uncertainty-based approach, but without utilizing entropy for image selection. This results in ten rounds for the standard microscopy dataset, 14 rounds for the LC-internal datasets, and 34 rounds for the LC-external dataset.

To account for randomness in the image selection, we repeated the previously described process ten times. We aggregated the evaluation metrics from each round across all ten experiments by calculating the mean and the interquartile range of these metrics. To assess whether AL significantly improves training, we compared the random selections with the AL selections and conducted the Mann-Whitney U-test (Mann and Whitney, 1947) for the means of the evaluation metrics in each round across all ten experiments.

2.3 Goal-dependent labeling

Most segmentation tasks are sensitive to the shape of the object. In adherent cell cultures, cells often grow in close proximity to one another or overlap, forming complex shapes that resemble blobs or clusters. Labeling such blobs of cells requires much less effort (“lazy”) than labeling each cell individually. Since confluence estimation necessitates full cell segmentation, we hypothesized that it is sufficient to label blobs of cells instead of labeling them separately. Therefore, we compared the model performance between the lc-internal and lc-internal-lazy datasets.

2.4 Active learning in a microscopy movie context

The images from the “sc-internal” dataset originated from a time-resolved microscopy movie. This means that the later the images were captured in the movie, the more the cells have grown (higher confluence). Therefore, we expect that AL may select images with varying levels of confluence, and subsequently, the movie positions influence the selection process. Based on the results from the AL experiment, we calculated the number of positions at which the selected image differs from the image chosen in the previous step. Again, we aggregated these differences by calculating the mean and determining the interquartile range across all ten random runs. As a control, we compared the image selected by AL with the randomly selected image. We tested for significance using the Mann-Whitney U Test (Mann and Whitney, 1947).

2.5 Fine-tuning compared to zero-shot learning

We evaluated the necessity of fine-tuning by first using all models without additional training and directly feeding the test set into them for inference (zero-shot). In contrast, we fine-tuned the models with all available labeled data, training them for 500 epochs (without early stopping). We performed inference on datasets and their test sets. As before, we used IoU and the absolute delta in confluence as evaluation metrics. To provide better context for the models’ performance, we included a baseline confluence detector that does not incorporate modern deep learning. This baseline algorithm processes grayscale images and detects edges with the Canny edge detector (Canny, 1986). It then fills gaps in the detected edges using binary hole-filling. Subsequently, it removes small objects and detects contours in the processed image with the “marching squares” algorithm, finding iso-valued contours at a specific level. After detecting the contours, any open contours are closed and simplified using polygonal approximation. This baseline algorithm then draws these contours onto a blank mask and interpolates between them to create a filled mask, which can be used to calculate metrics similarly to the other models.

2.6 Models

2.6.1 Cellpose

We utilized the Cellpose model from Stringer et al. (2020), based on version 3.0.0, and added functionality to enable custom names for standard Cellpose log files. This modification was reviewed and incorporated into Cellpose’s code by the authors. We utilized the train function from the Cellpose model with the model type “cyto” for fine-tuning. Cellpose relies on the mean cell diameter as input. Therefore, we calculated the mean cell diameter with a custom script (https://git.informatik.uni-leipzig.de/joas/confluence/-/blob/main/cellpose_main.py?ref_type=heads) using our data. For inference, Cellpose requires two key thresholds: the cell probability threshold and the flow threshold. The cell probability threshold determines the minimum probability for pixels to be classified as part of a cell, while the flow threshold controls the tolerance for errors in detecting cells (Stringer et al., 2020). We observed that the model’s performance is sensitive to these thresholds; thus, we implemented an automatic tuning process to optimize them for our training set. We developed a method that systematically explores different threshold combinations. The function evaluates the model’s performance on the training set using ground-truth masks as a reference. It iteratively tests a range of flow thresholds (from 0 to 3) and cell probability thresholds (from −6 to 6) in 0.5 steps to identify the optimal combination (IoU score). Additionally, a penalty is imposed if no cells are detected in the predicted masks. This ensures that the model does not optimize for precision and only outputs the background without any masks.

Besides the thresholds, we did not alter any other hyperparameters from the default values of the Cellpose model

class and the Cellpose train method. For the AL part, we obtained the cell probabilities directly from the Cellpose model's eval method and calculated the Shannon entropy for each pool image. Furthermore, we trained all models in the AL experiments for 100 epochs.

2.6.2 Detectron2

We employed the Detectron2 framework to perform instance segmentation as published by Meta Research (Wu et al., 2019). The model was configured using a Mask R-CNN architecture with a ResNet-50 backbone and Feature Pyramid Network (FPN) as specified in the `mask_rcnn_R_50_FPN_3x.yaml` configuration file. The model was designed to detect a single class, corresponding to the cells in our images. We configured the model with a base learning rate of 0.00025 and a batch size of 128 images for the region of interest (ROI) heads. To ensure reproducibility, we established a fixed random seed. The model was trained for 10 iterations, with a 5-iteration warmup period. For systems without GPU acceleration, the model defaults to CPU computation automatically.

To determine image selection in the AL process, we cannot obtain the probability masks directly from Detectron2. Instead, the model provides pre-calculated confidence scores for each mask. We use the average of these scores as a measure of uncertainty. In the inference step, we select the image(s) with the lowest score to label next.

2.6.3 Segment anything model

We extended SAM, which does not natively support fine-tuning, by constructing a custom module to enable this functionality, as described in detail in a dedicated blog post (see footnote²) that makes fine-tuning SAM publicly accessible. Our results are based on SAM version 1.0.0 (Kirillov et al., 2023).

We developed a custom wrapper class (`ModelSimple`) around SAM's architecture to enable supervised training on our datasets. The key innovation in our approach was selectively freezing specific components of the network while allowing others to be updated during backpropagation. Specifically, we froze the image encoder and prompt encoder parameters to preserve the model's pre-trained feature extraction capabilities while making only the mask decoder trainable. This strategy significantly reduced computational requirements, allowing the model to adapt to our specific segmentation task.

Since cell segmentation lacks predefined spatial locations, we adjusted the standard SAM inference pipeline to function without explicit prompts or bounding boxes. Our implementation directly processes the input images to generate segmentation masks, eliminating the need for user interaction or a predefined ROI. We maintained SAM's native input resolution of $1,024 \times 1,024$ pixels to leverage the generation of probability mask maps that were upsampled to match the original image dimensions. To optimize our approach, we

employed a combined loss function as specified in the original SAM paper:

$$\mathcal{L}_{total} = 20 \cdot \mathcal{L}_{focal} + \mathcal{L}_{dice} \quad (1)$$

The model was trained using the Adam optimizer, and to ensure reproducibility, we set the random seed to 100 for all random operations (Kirillov et al., 2023). For AL, we calculated the Shannon entropy of the predicted probability maps and selected the image(s) with the highest entropy.

2.6.4 U-Net

We implemented the U-Net architecture according to the original design proposed by Ronneberger et al. (2015). Our implementation includes a symmetric encoder-decoder structure with skip connections to maintain spatial information throughout the network. The encoder pathway consists of four down-sampling blocks, each containing two 3×3 convolutional layers with ReLU activation, followed by a 2×2 max pooling operation. The number of feature channels doubles at each down-sampling step, starting with 64 channels after the initial convolution and expanding to 128, 256, 512, and finally 1,024 channels at the bottleneck (or 512 when using bilinear up-sampling). The decoder pathway mirrors the encoder with four up-sampling blocks. The final layer consists of a 1×1 convolution that translates the 64-channel feature map into the desired number of output classes, resulting in pixel-wise classification for the segmentation mask. In our cell segmentation task, the network produces a single-channel probability map that indicates the likelihood of each pixel belonging to a cell.

For AL, we used the output probability maps directly from the U-Net model. We calculated the Shannon entropy of these predicted probability maps to select the images with the highest entropy, or uncertainty.

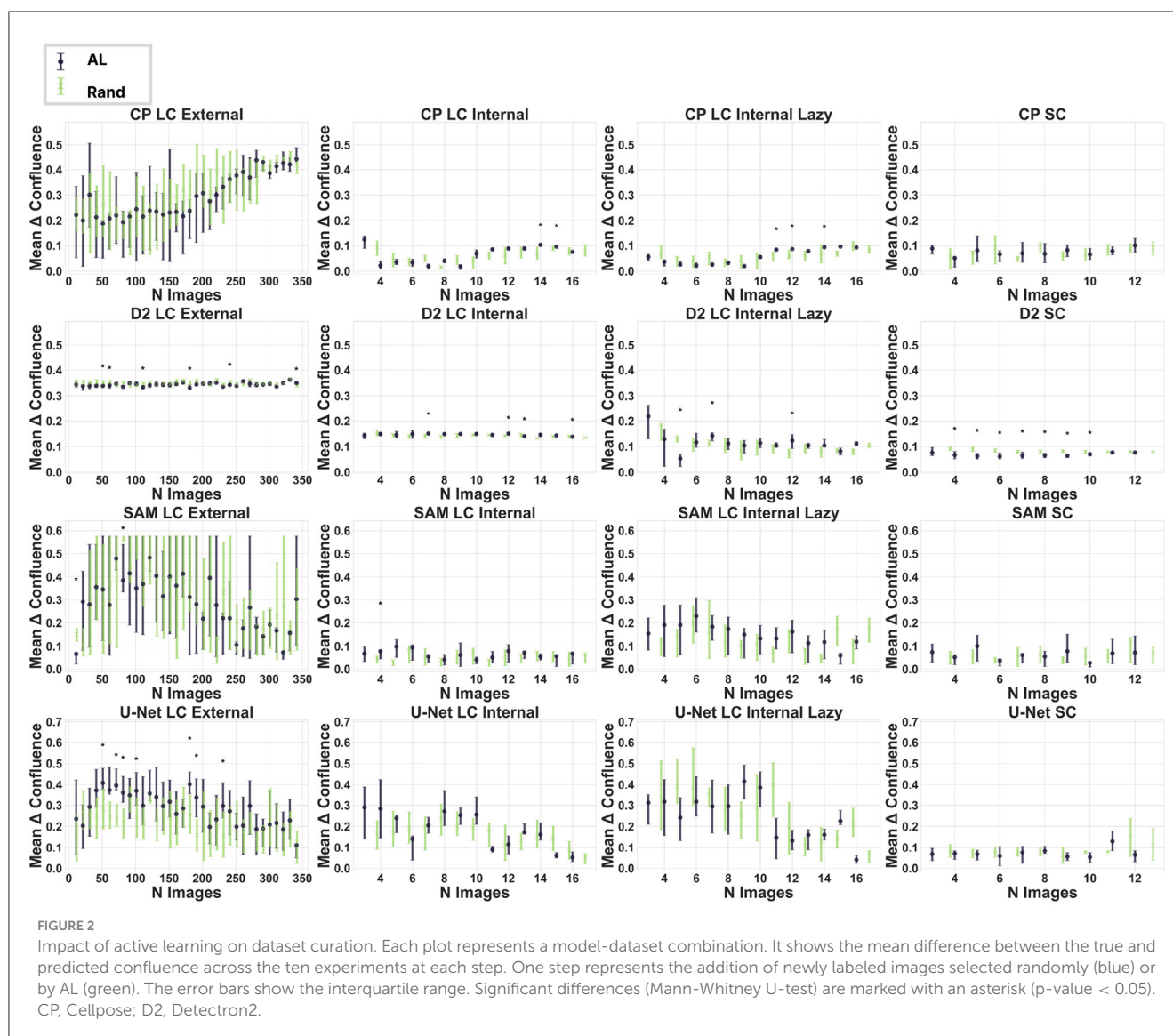
3 Results

3.1 Active learning performance

We analyzed whether AL is an effective approach for improving cell segmentation and reducing labeling effort when using four commonly used models for segmentation. We demonstrate that uncertainty-based AL provides no improvement in confluence prediction performance in our chosen datasets. The experiments with Cellpose and SAM exhibited the smallest differences between random and AL image selection, with only four (Cellpose) and three (SAM) statistically significant steps out of 72 total steps (cf. Figure 2, Supplementary Figure S2). We observe significantly better performance from the U-Net model when using a random dataset curation approach on the external dataset. Detectron2 is the only model where AL improved confluence prediction in the standard microscopy dataset and, to some extent, in the external dataset.

We hypothesized that both specialized and generalist pretrained segmentation models benefit from fine-tuning, given the unique and complex shape of MSC cells. Surprisingly, the impact of fine-tuning is limited and largely depends on the dataset. In seven experiments, we observed the best scores during the early stages of fine-tuning in AL, indicating that more data does not

² <https://maxjoas.medium.com/finetune-segment-anything-sam-for-images-with-multiple-masks-34514ee811bb>



always lead to improved results. Our control experiments, in which random picks were used as the next image for labeling and training, yielded similar results, with six of the best performances occurring in the first half of the fine-tuning. From a model perspective, we observe that Detectron2 does not benefit from fine-tuning, Cellpose gets even worse, and for SAM and U-Net, we do not see a clear trend.

Specifically, nine out of 16 experiments (model and dataset combinations) achieved a minimal absolute delta in confluence of no more than 0.05, while three experiments achieved a minimal absolute delta in confluence of no more than 0.10. When comparing the mean performances across all datasets, SAM predicts the confluence most accurately, with an absolute delta value of 0.05 ± 0.02 , while Detectron2 shows an absolute delta value of 0.15 ± 0.13 . Performance analysis across datasets indicates that the goal-dependent labeled dataset (“lazy”) achieves the best results (mean 0.04 ± 0.02), whereas models perform worst on the external dataset (mean 0.18 ± 0.11). Table 2 provides a detailed aggregation of the best results across models and datasets. Furthermore,

non-active learning (randomized) exhibits similar trends (cf. Supplementary Table S1). While cf. Supplementary Table S2 shows the absolute delta in confluence for all experiments, we observe similar trends when using IoU as a performance metric (cf. Supplementary Tables S4, S5).

3.2 Goal-dependent labeling

For Confluence prediction, accurately segmenting individual cells is unnecessary because a foreground/background classification would suffice. Therefore, we investigated whether faster goal-dependent labeling of cell clusters as single clusters (“lazy labeling”) affects model performance. The results achieved through goal-dependent labeling do not universally enhance the performance of every model. We observe a clear trend with the Detectron2 model, where ten out of 14 steps show significant improvement with the lazy labeling method. In contrast, for U-Net and Cellpose, there is no evident difference. Interestingly, the SAM model demonstrates

TABLE 2 Mean and standard deviation of minimum delta confluence values across models and datasets.

a) Model aggregation		
Model	Min	Convergence point
Cellpose	0.07 ± 0.08	0.29 ± 0.20
Detectron2	0.15 ± 0.13	0.50 ± 0.37
SAM	0.05 ± 0.02	0.56 ± 0.41
U-Net	0.06 ± 0.03	0.95 ± 0.11
b) Dataset aggregation		
Dataset	Min	Convergence point
lc-external	0.18 ± 0.11	0.41 ± 0.45
lc-internal	0.06 ± 0.05	0.75 ± 0.29
lc-internallazy	0.04 ± 0.02	0.63 ± 0.40
sc-internal	0.05 ± 0.01	0.50 ± 0.34

The column *Convergence Point* represents the fraction of Active Learning iterations completed, where 0 indicates no additional data added, and 1 signifies that all available data has been included.

even better results for the precisely labeled images in terms of confluence estimation.

The IoU is significantly higher by a large margin (see [Figure 3](#)) when data are labeled lazily for all models, indicating that this metric is not invariant to labeling strategy, cell size, and cluster shapes.

3.3 Active learning in a microscopy movie context

We expected that uncertainty-based AL would ideally choose images with varying confluence, which sequentially increases in our standard microscopy dataset recorded as a movie of a cell culture. On the contrary, we observe no consistent differences or tendencies in the movie positions of the selected images between the AL and random selection processes across our four segmentation models (see [Figure 4](#)). This suggests that cell density is not a significant factor in the model's uncertainty, likely because cell shape remains relatively unchanged over time. Furthermore, the results align with the observation that AL does not significantly improve model performance for cell segmentation and confluence prediction, as shown in [Figure 2](#).

3.4 Zero-shot inference or full fine-tuning

Since we observed that models exhibited mixed behavior during fine-tuning, such as Cellpose experiencing a decrease in performance, we analyzed the zero-shot capability of the models. The performance of SAM has an absolute delta of 0.05 ± 0.036 in confluence estimation, which is nearly perfect across all datasets, even without fine-tuning. As expected, deep learning-based approaches significantly outperform the algorithmic image segmentation baseline (see [Figure 5](#), [Supplementary Figure S2](#)). We

observe a significant performance improvement for Detectron2 when fine-tuning on our internal datasets. However, we observe a decline in performance when fine-tuning on the external dataset. Additionally, we see small dataset-dependent fluctuations for the Cellpose model.

In summary, our results indicate a strong indication that, for the confluence prediction of MSC-like cells, generalist foundation models, such as SAM, outperform specialized models, such as Cellpose, and fine-tuning is unnecessary. Furthermore, in the case of Cellpose, results indicate that fine-tuning with irregular cell shapes (MSC) may result in decreased performance rather than the expected improvements.

3.5 Qualitative analysis and usability

When deciding how to label a dataset for confluence prediction and which model to choose, many practical considerations arise beyond just performance. We will provide insights on (a) the difficulty of fine-tuning the models, (b) implementing an uncertainty-based AL approach, and (c) overall computational considerations.

Detectron2 ([Wu et al., 2019](#)) is the simplest model to fine-tune, as fine-tuning is a built-in feature. The documentation for Detectron2 is clear and easy to follow. Detectron2 supports input data in COCO annotations, which is a widely used format. While it is easy to use, it offers less customizability and control. Additionally, Detectron2 is not the most cutting-edge model and does not specialize in cell segmentation.

In contrast, Cellpose ([Stringer et al., 2020](#)) specializes in cell segmentation and provides robust fine-tuning options. However, Cellpose predictions are highly sensitive to parameter settings in cell probability and flow thresholds. These thresholds can be adjusted when the ground truth is known, but for automatic segmentation on unknown data, this is not feasible and would require manual intervention to identify the optimal thresholds. Cellpose requires ground-truth masks as input for fine-tuning, which is also a standard practice. With releases in February 2024, Cellpose offers updated models and some customizability regarding cell types.

SAM ([Kirillov et al., 2023](#)) was the most challenging model to fine-tune because this option is not supported. We needed to write custom wrapper classes to enable fine-tuning, which is not possible without significant technical expertise in deep learning. On the other hand, SAM is easy to use for zero-shot learning and is the most powerful of all the models used. SAM supports annotations in the common COCO JSON format.

We trained U-Net from scratch without any fine-tuning. U-Net requires implementation knowledge, such as PyTorch or Keras, and needs to be trained from scratch. Due to its limited performance, it does not provide a good trade-off for efficient confluence prediction.

To combine a dataset with AL, we need to obtain uncertainty measures from the models. Detectron2 was the only model that returned confidence scores for the mask predictions. However, there was no built-in functionality to obtain probability masks. In our custom implementations of U-Net and SAM, we made

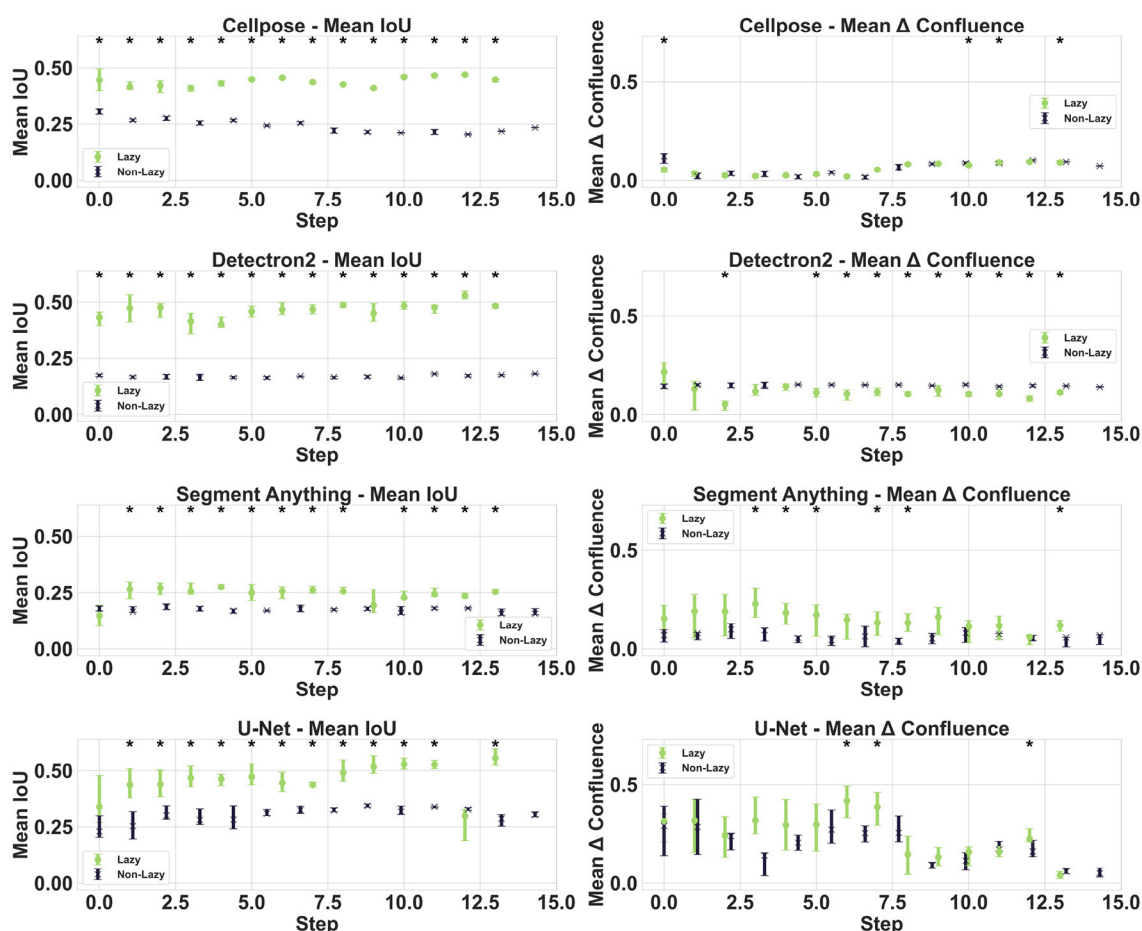


FIGURE 3

Impact of lazy labeling. The plot compares the performance of lazy and exact labeling methods during the dataset curation process for all models. The first column displays the IoU metric, while the second column illustrates the differences in confluence.

it easy to obtain the probability maps directly. Nevertheless, this comes before custom implementation, which requires expertise in deep learning. Cellpose returns the cell probabilities directly, making an uncertainty-based AL approach more straightforward to implement.

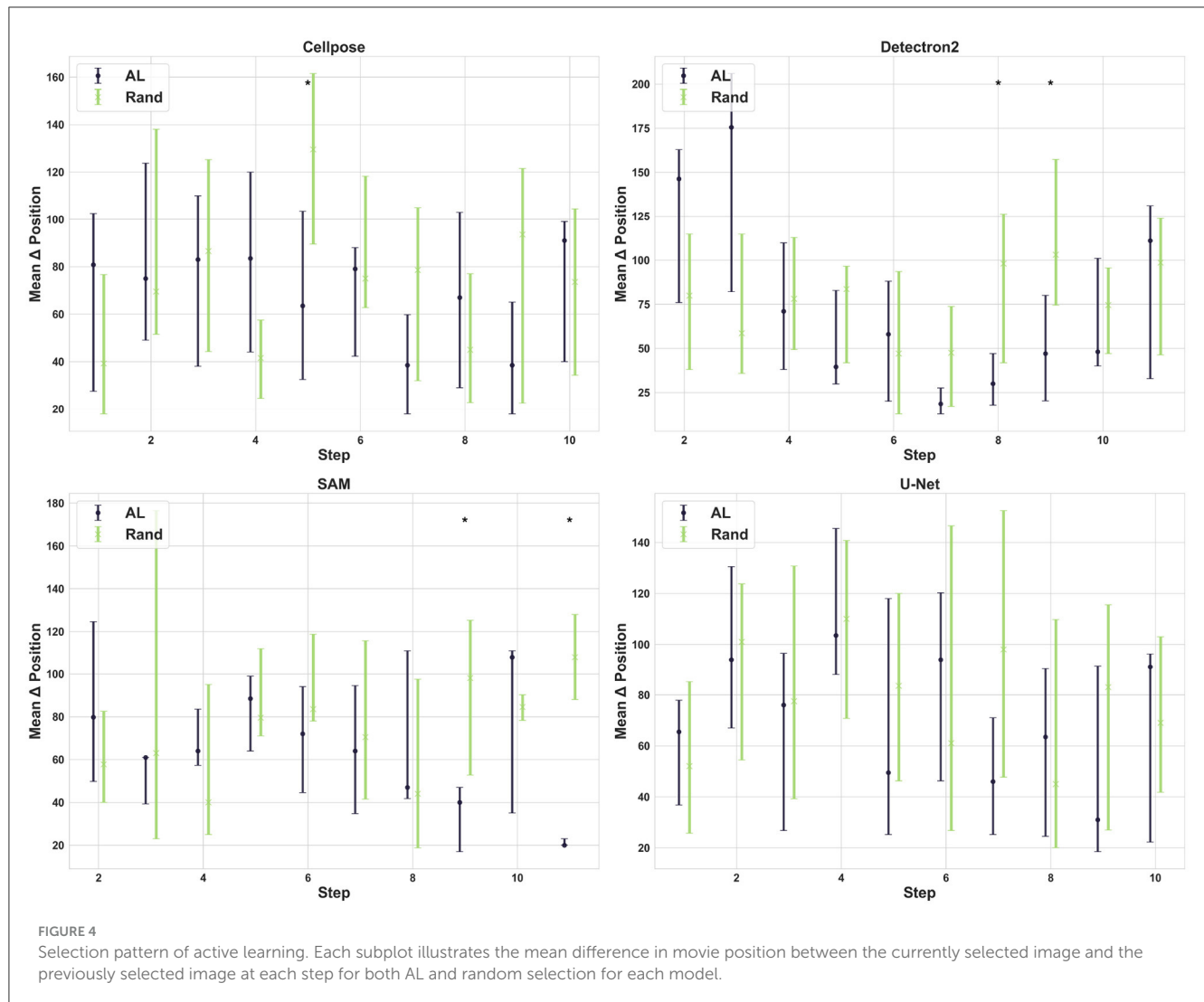
A GPU for model training or fine-tuning is almost essential for all four models. For inference, a CPU is adequate for U-Net, Detectron2, and Cellpose. This is especially important for integrating confluence prediction into real-world automation systems in cell production, where high-performance GPUs may not be available. However, inference with SAM on a CPU is impractical due to the model's size and slow performance.

The fine-tuning process requires a GPU for all models to ensure completion within a reasonable timeframe. Fine-tuning on larger datasets with a CPU is impractical, as it could take weeks and provide minimal benefit compared to zero-shot training. In contrast, fine-tuning with only a few images can be completed within hours for Detectron2 and Cellpose, offering significant performance improvements for these models. Given the high computational cost of fine-tuning and the lack of substantial performance gains, we do not recommend fine-tuning SAM in contexts such as our use case.

4 Discussion

In our study, we compared four models for cell segmentation across various datasets inspired by real-world MSC manufacturing. This comparison aims to gain insights on how to leverage AI-based confluence estimation most efficiently. The results provide actionable strategies applicable to similar contexts. First, we demonstrated that zero-shot inference with SAM achieved near-perfect confluence estimation. Second, we observed that goal-dependent labeling outperformed traditional labeling methods in terms of IoU. Finally, we demonstrate that AL is suboptimal for MSC microscopy images. The limited benefit of AL can be explained by: (a) a lack of diversity within the dataset, (b) the use of large pre-trained models, (c) a basic AL approach, and (d) the simple binary classification (foreground/background) task.

Since our datasets contain only one cell type, the primary variation lies in the growth state or cell density. This does not fundamentally alter the characteristics of the objects to be segmented (i.e., the cells). Consequently, the timing of when a given example is presented during training has a limited impact on model performance. Our analysis revealed that AL did not select images based on their temporal position in the growth



sequence, showing no significant difference from random selection. This finding supports our explanation that cell density alone does not create enough variation that AL strategies typically exploit (Monarch, 2021). The model's uncertainty, which drives the AL selection process, appears to be independent of the growth state. This suggests that once the model learns to segment cells at one density, it can readily generalize to other densities.

We also hypothesize that the benefit of AL in pre-trained models is minimal due to the extensive data exposure these models have already experienced, which diminishes the impact of new data points. Additionally, our observations indicate that U-Nets underperform when faced with a small number of highly diverse instances through AL, while random selection retains a distribution that is more representative of the entire dataset. Furthermore, for usability reasons, we adopt a straightforward maximum entropy approach for uncertainty-based AL. However, capturing the complexities of this data and model may require more complex AL strategies involving combinations of techniques to manage variations more effectively. Although prior research on biomedical images demonstrates that AL techniques can achieve

comparable performance with a reduced sample size (Nath et al., 2021; Kim et al., 2024; Li et al., 2023; Huang et al., 2024), these studies did not incorporate pre-trained models, used diverse data, and employed more complex AL methods. Considering our findings and previous research, we conclude that AL is best utilized when pre-trained models are not appropriate for a given use case and when complex AL algorithms are available for specific problems, ideally in a diverse multi-classification task.

Beyond AL, we explored simplified goal-dependent labeling directly linked to the desired outcome, namely, confluence estimation. Interestingly, for the IoU metric, goal-dependent labeled data significantly outperformed traditional labeling approaches. We attribute this success to simpler shapes that are easier for models to learn. Even when examining the Confluence task directly by calculating the difference from the ground-truth Confluence, we observe no drop in performance when utilizing goal-dependent labeling, or “lazily”. Notably, when employing the SAM model, traditionally labeled data performed slightly better, which we attribute to SAM's extensive pre-training on precisely annotated datasets. Additionally, lazy labeling introduces

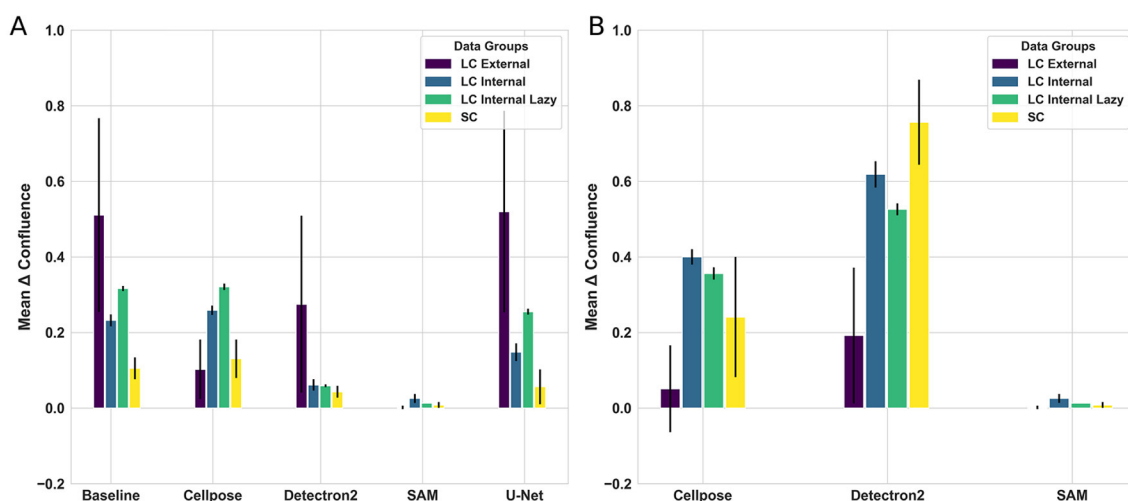


FIGURE 5

Comparison of fine-tuning and zero-shot results for each model and dataset. (A) Shows the differences between true and predicted confluence for four models and one baseline across all datasets during fine-tuning. (B) Shows the performance of zero-shot learning for models where zero-shot is applicable.

irreducible error, acting as a source of noise that is more pronounced when model performance is overall very high, as seen with SAM. Importantly, the substantial reduction in labeling effort makes the lazy labeling approach attractive for confluence estimation since detailed cell segmentation is not required (e.g., to derive individual cell characteristics).

While this specific labeling strategy may not be applicable to all problems, it demonstrates the value of developing task-specific labeling approaches that strike a balance between annotation effort and model performance. To our knowledge, no existing research directly addresses this confluence of estimation-specific annotation. However, recent studies in biomedical imaging have highlighted the use of time-efficient annotation techniques combined with self- and human-supervised learning (human-in-the-loop) to reduce labeling demands. For example, in nucleus segmentation, some approaches focus on selectively annotating only a small subset of critical image patches, utilizing human-supervised methods and data augmentation to match the performance of fully supervised models while minimizing the requirements for labeled data (Lou et al., 2023). Similarly, in cell segmentation, weakly supervised methods use single-point annotations per cell, combined with self- and co-training strategies, to achieve segmentation accuracy close to that of fully supervised methods (Zhao and Yin, 2021). Krishnan et al. (2022)'s review further highlights how efficient annotation methods, when combined with self-supervised learning, enable models to leverage large volumes of unannotated data, thereby enhancing model development while reducing expert annotation time.

Furthermore, tools like LABKIT (Arzt et al., 2022) provide interfaces for efficient annotation and human-in-the-loop processes, combined with supervised deep learning. Although we implemented only a time-efficient annotation strategy without any form of weak human supervision, the relative simplicity of our task suggests that these findings still support our results.

They indicate that efficient annotation methods, particularly when paired with self- and human-supervised learning, can significantly reduce labeling efforts in active learning while maintaining model performance.

While efficient labeling strategies may reduce annotation effort, completely eliminating the need for labeling would be even more desirable. We demonstrate that SAM achieves nearly perfect confluence estimation with zero-shot inference, indicating that the cost-benefit ratio of fine-tuning decreases with larger foundation models. Although other models exhibited marginal improvements with fine-tuning, and even SAM showed slight gains, these advantages were minimal compared to SAM's zero-shot performance. Interestingly, across all models, we discovered that the best performance was achieved using only a subset of the available training data, suggesting that more is not always better. The demands of fine-tuning, including (a) computational resources, (b) technical expertise for model adaptation, and (c) time invested in data annotation, far outweigh modest performance gains. This cost-benefit analysis strongly favors the use of large foundation models like SAM in its zero-shot configuration. Additionally, the recently released SAM 2 (Ravi et al., 2024) may demonstrate even higher accuracy for such tasks and focuses on object tracking in video contexts, which is relevant in time-series data from cellular production systems.

When comparing our results with existing research, we see that zero-shot inference with SAM is powerful even for more complex tasks. However, the optimal approach appears task-dependent, and some amount of fine-tuning or combining it with other models still seems to have a benefit: Baral and Paing (2024) used a fine-tuned object detection model (YOLOv9-E) to generate prompts for zero-shot SAM inference, followed by traditional image processing refinements. This hybrid approach achieved highly accurate performance (94% mAP50) for cell segmentation across varying difficulty levels without fine-tuning SAM directly.

In contrast, more specialized applications require more complex architectures. CryoSegNet (Gyawali et al., 2024) demonstrated that the combination of SAM with a task-specific U-Net significantly improved protein particle detection in cryo-EM images compared to using SAM alone. Similarly, the Segment Anything for Microscopy project (Archit et al., 2023) demonstrated that specialized training for multi-dimensional microscopy data significantly improved segmentation quality across various imaging conditions. In this context, the additional effort is likely warranted due to the complexity of volumetric segmentation and tracking tasks compared to ours. These varying approaches highlight the importance of considering task complexity and resource constraints when choosing between zero-shot applications, hybrid solutions, and full fine-tuning of foundation models.

While our findings provide insights into confluence estimation for MSC production standardization, we acknowledge several limitations of our study. First, our results specifically focus on confluence estimation, reflecting the practical needs of our applied domain. Consequently, our methods are not directly applicable to other tasks that may require tracking, such as spatial colony growth monitoring in bacteria (Kindler et al., 2019). Additional research is necessary to compare specialized tracking tools, such as TRACKASTRA (Gallusser and Weigert, 2025), with generalist foundation models, such as SAM 2 (Ravi et al., 2024).

While not exhaustive, our selected models strategically covered a representative spectrum of approaches: from models trained from scratch to cell-specific models and powerful general-purpose foundation models. This selection allowed us to compare different paradigms in model development while maintaining practical feasibility. Similarly, while our datasets were limited to one cell type, they enabled us to draw important conclusions about the impact of data diversity on active learning effectiveness in real-world scenarios.

Despite these limitations, our study provides several valuable contributions: We demonstrated that for homogeneous cell cultures, (a) SAM delivers excellent results without requiring resource-intensive methods such as active learning or fine-tuning. We provide (b) technical guidelines for implementing active learning in cell imaging applications, demonstrate (c) the importance of goal-specific labeling strategies, and highlight (d) how data homogeneity influences active learning performance. These insights can guide future research in automated cell culture monitoring and quality control. These practical insights significantly lower the barrier to implementing automated quality control in cell manufacturing. A prototype of our confluence detection software is publicly available (<https://livinglab.scaDSai.uni-leipzig.de/cell-confluence/>), enabling immediate community adoption.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: <https://github.com/MaxJoas/confluence-paper>.

Author contributions

MJ: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. DF: Data curation, Resources, Writing – original draft, Writing – review & editing. RH: Conceptualization, Investigation, Resources, Supervision, Writing – original draft, Writing – review & editing. ER: Funding acquisition, Project administration, Resources, Writing – original draft, Writing – review & editing. JE: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. MJ and DF acknowledge support from the German Federal Ministry of Education and Research under the funding codes 03ZU1111NC and 03ZU1111NB (SaxoCellSystems) as part of the Clusters4Future cluster SaxoCell. RH and JE acknowledge the financial support from the Federal Ministry of Education and Research of Germany and the Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus for the Center of Excellence in AI research “Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig,” project identification number: ScaDS.AI. The project also received support from the Open Access Publishing Fund of Leipzig University.

Acknowledgments

The authors gratefully acknowledge the computing time provided on the high-performance computer at the NHR Center of TU Dresden. This center receives joint support from the Federal Ministry of Education and Research and the state governments involved in the NHR (<https://www.nhr-verein.de/unsere-partner>).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. Generative AI was used to improve grammar and language style.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of

their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Archit, A., Nair, S., Khalid, N., Hilt, P., Rajashekar, V., Freitag, M., et al. (2023). *Segment Anything* for microscopy. *bioRxiv*. doi: 10.1101/2023.08.21.554208
- Arzt, M., Deschamps, J., Schmied, C., Pietzsch, T., Schmidt, D., Tomancak, P., et al. (2022). LABKIT: Labeling and segmentation toolkit for big image data. *Front. Comp. Sci.* 4:777728. doi: 10.3389/fcomp.2022.777728
- Baral, S., and Paing, M. P. (2024). Instance segmentation of cells and nuclei from multi-organ cross-protocol microscopic images. *Quant. Imaging Med. Surg.* 14, 6204–6221. doi: 10.21037/qims-24-801
- Canny, J. (1986). A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 8, 679–698. doi: 10.1109/TPAMI.1986.4767851
- Chen, H., and Murphy, R. F. (2023). Evaluation of cell segmentation methods without reference segmentations. *Mol. Biol. Cell* 34:6. doi: 10.1091/mbc.E22-08-0364
- Edlund, C., Jackson, T. R., Khalid, N., Bevan, N., Dale, T., Dengel, A., et al. (2021). LIVECell—a large-scale dataset for label-free live cell segmentation. *Nat. Methods* 18, 1038–1045. doi: 10.1038/s41592-021-01249-6
- Fernández-Santos, M. E., García-Arranz, M., Andreu, E. J., García-Hernández, A. M., López-Parra, M., Villarón, E., et al. (2022). Optimization of mesenchymal stromal cell (MSC) manufacturing processes for a better therapeutic outcome. *Front. Immunol.* 13:918565. doi: 10.3389/fimmu.2022.918565
- Galipeau, J., and Sensebé, L. (2018). Mesenchymal stromal cells: Clinical challenges and therapeutic opportunities. *Cell Stem Cell* 22, 824–833. doi: 10.1016/j.stem.2018.05.004
- Gallusser, B., and Weigert, M. (2025). “TRACKASTRA: transformer-based cell tracking for live-cell microscopy,” in *Computer Vision-ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol (Cham: Springer Nature Switzerland), 467–484.
- Gyawali, R., Dhakal, A., Wang, L., and Cheng, J. (2024). CryoSegNet: accurate cryo-EM protein particle picking by integrating the foundational AI image segmentation model and attention-gated U-Net. *Brief. Bioinform.* 25:bbae282. doi: 10.1093/bib/bbae282
- Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., et al. (2021). Pre-trained models: Past, present and future. *AI Open* 2, 225–250. doi: 10.1016/j.aiopen.2021.08.002
- Huang, J., Farpour, N., Yang, B. J., Mupparapu, M., Lure, F., Li, J., et al. (2024). Uncertainty-based active learning by bayesian U-Net for multi-label cone-beam CT segmentation. *J. Endod.* 50, 220–228. doi: 10.1016/j.joen.2023.11.002
- IHME (2020). “Global burden of disease study 2019 (GBD 2019) disease and injury burden 1990–2019,” in *Technical report, Institute for Health Metrics and Evaluation (IHME)* (Seattle: IHME).
- Jafari, M., Zhang, Y., Zhang, Y., and Liu, S. (2024). “The power of few: Accelerating and enhancing data reweighting with coresets selection,” in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Seoul: IEEE), 7100–7104.
- Kim, D. D., Chandra, R. S., Yang, L., Wu, J., Feng, X., Atalay, M., et al. (2024). Active learning in brain tumor segmentation with uncertainty sampling and annotation redundancy restriction. *J. Imag. Inform. Med.* 37, 2099–2107. doi: 10.1007/s10278-024-01037-6
- Kim, D. S., Lee, M. W., Lee, T.-H., Sung, K. W., Koo, H. H., and Yoo, K. H. (2017). Cell culture density affects the stemness gene expression of adipose tissue-derived mesenchymal stem cells. *Biomed. Reports* 6, 300–306. doi: 10.3892/br.2017.845
- Kindler, O., Pulkkinen, O., Cherstvy, A. G., and Metzler, R. (2019). Burst statistics in an early biofilm quorum sensing model: the role of spatial colony-growth heterogeneity. *Sci. Rep.* 9:1. doi: 10.1038/s41598-019-48525-2
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al. (2023). “Segment Anything,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (Paris: IEEE), 3992–4003.
- Krishnan, R., Rajpurkar, P., and Topol, E. J. (2022). Self-supervised learning in medicine and healthcare. *Nat. Biomed. Eng.* 6, 1346–1352. doi: 10.1038/s41551-022-00914-1
- Li, X., Xia, M., Jiao, J., Zhou, S., Chang, C., Wang, Y., et al. (2023). HAL-IA: A hybrid active learning framework using interactive annotation for medical image segmentation. *Med. Image Anal.* 88:102862. doi: 10.1016/j.media.2023.102862
- Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., et al. (2014). “Microsoft COCO: Common objects in context,” in *Computer Vision-ECCV 2014* (Cham: Springer International Publishing), 740–755.
- Lou, W., Li, H., Li, G., Han, X., and Wan, X. (2023). Which pixel to annotate: A label-efficient nuclei segmentation framework. *IEEE Trans. Med. Imaging* 42, 947–958. doi: 10.1109/TMI.2022.3221666
- Mann, H. B., and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.* 18, 50–60. doi: 10.1214/aoms/1177730491
- Monarch, R. (2021). *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-centered AI*. Shelter Island, NY: Manning.
- Nath, V., Yang, D., Landman, B. A., Xu, D., and Roth, H. R. (2021). Diminishing uncertainty within the training pool: active learning for medical image segmentation. *IEEE Trans. Med. Imaging* 40, 2534–2547. doi: 10.1109/TMI.2020.3048055
- Panagiotakis, C., and Argyros, A. A. (2018). “Cell segmentation via region-based ellipse fitting,” in *2018 25th IEEE International Conference on Image Processing (ICIP)* (Athens: IEEE), 2426–2430. doi: 10.1109/ICIP.2018.8451852
- Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., et al. (2024). SAM 2: Segment anything in images and videos. *arXiv [preprint]* arXiv:2408.00714. doi: 10.48550/arXiv.2408.00714
- Ronneberger, O., Fischer, P., and Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Cham: Springer International Publishing, 234–241.
- Sayin, B., Krivosheev, E., Yang, J., Passerini, A., and Casati, F. (2021). A review and experimental analysis of active learning over crowdsourced data. *Artif. Intell. Rev.* 54, 5283–5305. doi: 10.1007/s10462-021-10021-3
- Shen, S. P., Tseng, H.-a., Hansen, K. R., Wu, R., Gritton, H. J., Si, J., et al. (2018). Automatic cell segmentation by adaptive thresholding (ACSAT) for large-scale calcium imaging datasets. *eNeuro* 5:ENEURO.0056-18.2018. doi: 10.1523/ENEURO.0056-18.2018
- Shimizu, Y., Ntege, E. H., Azuma, C., Uehara, F., Toma, T., Higa, K., et al. (2023). Management of rheumatoid arthritis: Possibilities and challenges of mesenchymal stromal/stem cell-based therapies. *Cells* 12:1905. doi: 10.3390/cells12141905
- Strecanska, M., Sekelova, T., Csobonyeiova, M., Danisovic, L., and Cehakova, M. (2024). Therapeutic applications of mesenchymal/medicinal stem/signaling cells preconditioned with external factors: are there more efficient approaches to utilize their regenerative potential? *Life Sci.* 346:122647. doi: 10.1016/j.lfs.2024.122647
- Stringer, C., Wang, T., Michaelos, M., and Pachitariu, M. (2020). Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* 18, 100–106. doi: 10.1038/s41592-020-01018-x
- von Chamier, L., Laine, R. F., Jukkala, J., Spahn, C., Krentzel, D., Nehme, E., et al. (2021). Democratizing deep learning for microscopy with ZeroCostDL4Mic. *Nat. Commun.* 12:2276. doi: 10.1038/s41467-021-22518-0
- Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. (2019). *Detron2*. Available online at: <https://github.com/facebookresearch/detron2> (accessed September 20, 2024).
- Yin, T., Panapitiya, G., Coda, E. D., and Saldanha, E. G. (2023). Evaluating uncertainty-based active learning for accelerating the generalization of molecular property prediction. *J. Cheminform.* 15:1. doi: 10.1186/s13321-023-00753-5
- Zhao, T., and Yin, Z. (2021). Weakly supervised cell segmentation by point annotation. *IEEE Trans. Med. Imaging* 40, 2736–2747. doi: 10.1109/TMI.2020.3046292
- Zhu, J., Wang, H., Yao, T., and Tsou, B. K. (2008). “Active learning with sampling by uncertainty and density for word sense disambiguation and text classification,” in *Proceedings of the 22nd International Conference on Computational Linguistics* (Coling 2008), D. Scott, and H. Uszkoreit (Manchester, UK: Coling 2008 Organizing Committee), 1137–1144.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fcomp.2025.1562358/full#supplementary-material>



OPEN ACCESS

EDITED BY

Sokratis Makrogiannis,
Delaware State University, United States

REVIEWED BY

Xi Wang,
The Chinese University of Hong Kong, China
Kwang-Hyun Uhm,
Korea University, Republic of Korea

*CORRESPONDENCE

Shin Miyakawa
✉ miyakawashi@aipu.ac.jp

RECEIVED 27 January 2025

ACCEPTED 17 June 2025

PUBLISHED 09 July 2025

CITATION

Miyakawa S, Fuse H, Yasue K, Koori N,
Takahashi M, Nosaka H, Moriya S, Tomita F and
Fujisaki T (2025) Assessing the consistency of
CT-based ventilation imaging under noise
reduction processing with simulated quantum
noise using a nonrigid alveoli phantom.
Front. Radiol. 5:1567267.
doi: 10.3389/fradi.2025.1567267

COPYRIGHT

© 2025 Miyakawa, Fuse, Yasue, Koori,
Takahashi, Nosaka, Moriya, Tomita and
Fujisaki. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Assessing the consistency of CT-based ventilation imaging under noise reduction processing with simulated quantum noise using a nonrigid alveoli phantom

Shin Miyakawa^{1*}, Hiraku Fuse¹, Kenji Yasue¹, Norikazu Koori²,
Masato Takahashi¹, Hiroki Nosaka¹, Shunsuke Moriya³,
Fumihito Tomita⁴ and Tatsuya Fujisaki¹

¹Department of Radiological Sciences, Ibaraki Prefectural University of Health Sciences, Ibaraki, Japan,

²Department of Radiological Technology, Niigata University of Health and Welfare, Niigata, Japan,

³Institute of Medicine, University of Tsukuba, Ibaraki, Japan, ⁴Department of Radiation Oncology,
St. Luke's International Hospital, Tokyo, Japan

Background: Previous studies have reported that quantum noise inherently present in CT images hinders the generation of CT-based ventilation image (CTVI), while quantum noise reduction approaches that do not affect CTVI have not yet been reported.

Aims: The purpose of this study was to evaluate the impact of noise reduction preprocessing on the accuracy and robustness of CTVI in relation to quantum noise present in CT images.

Methods and material: To reproduce the quantum noise, Gaussian noise (SD: 30, 80, 150 HU) was added to each inhalation and exhalation CT image. CTVI_{ref} and CTVI_{noise} was generated from CT_{ref} and CT_{noise}. A median filter and the noise reduction by the CNN were also applied to the CT image, which contained the quantum noise, and CTVI_{med} and CTVI_{cnn} was created in the same manner as CTVI_{ref}. We evaluated whether the regions classified as high, middle, or low in CTVI_{ref} were accurately represented as high, middle, or low in CTVI_{noise}, CTVI_{med} and CTVI_{cnn}. Additionally, to evaluate the ventilation function of each voxel, we compared two-dimensional histograms of CTVI_{ref}, CTVI_{noise}, CTVI_{med} and CTVI_{cnn}.

Statistical analysis used: Cohen's kappa coefficient and Spearman's correlation were used to assess the agreement between CTVI_{ref} and each of the following: CTVI_{noise}, CTVI_{med}, and CTVI_{cnn}.

Results: CTVI_{cnn} significantly improved categorical consistency and voxel-level correlation of CTVI, particularly under high-noise conditions (150 HU), outperforming both CTVI_{noise} and CTVI_{med}.

Conclusions: CNN-based denoising effectively improved the accuracy and robustness of CTVI under quantum noise.

KEYWORDS

computed tomography-based ventilation image, deformable image registration, noise reduction, nonrigid alveoli phantom, radiotherapy

1 Introduction

A variety of imaging modalities exist to assess pulmonary ventilation. Examples include computed tomography (CT), dual-energy CT, magnetic resonance imaging, single-photon emission computed tomography (SPECT), and positron emission tomography (PET). These techniques can accurately assess the three-dimensional (3D) distribution of ventilatory function in a patient's lungs (1–3). In radiation therapy, treatment plans that utilize routine CT imaging and deformable image registration (DIR) to generate CT-based ventilation images (CTVIs) have demonstrated the ability to avoid areas of high ventilatory function within the lungs (4). The approach aims to reduce the dose administered to regions with elevated ventilatory function, creating a more targeted and personalized treatment strategy. This approach not only enhances the accuracy of predicting adverse lung events but also contributes to a more effective reduction in the occurrence of such events during treatment (5, 6). CT scans are used as part of routine radiation treatment procedures for most lung cancer patients and can provide additional functional information about the patient without requiring additional functional imaging equipment or methods. Treatment planning with CTVI is a practical, high-resolution, cost-effective, and time-saving approach that can be performed based on four-dimensional (4D) or expiratory and inspiratory CT images (4–9).

Studies are currently underway to validate the accuracy of CTVI. Radionuclide imaging is widely used to assess pulmonary function and is considered the standard of choice for assessing other functional imaging modalities (7, 8). Recent studies have demonstrated that CT-based and SPECT ventilatory function imaging have good spatial measurement accuracy and correlation (9, 10). In addition, clinical trials demonstrate that radiotherapy using CTVI significantly reduces dose to ventilated lung regions (NCT02528942, NCT02308709, NCT02843568) (11).

Accurate assessment of pulmonary ventilation function is crucial for using CTVI in treatment planning. Small changes in DIR parameters have been reported to cause large relative changes in the CTVI (12). The study noted that DIR-based images may not show accurate ventilatory function even when the spatial accuracy of the deformations is acceptable using target registration error (TRE). The quantum noise in CT images does not significantly affect the accuracy of DIR but may hinder the generation of accurate CTVI (13). A nonrigid alveoli phantom was developed to evaluate the CTVI, based on the assumption that an accuracy validation phantom is required to investigate the causes of these obstacles and improve CTVI accuracy (14). However, various problems related to CTVI methods have not yet been solved. It has been demonstrated that there is a significant difference in the CTVI produced via DIR when different DIR parameters are used, even after meeting the tolerance for DIR accuracy with this phantom (15).

CT images inevitably contain quantum noise, owing to the nature of x-ray images. It is desirable to use high-resolution CT images to create the CTVI. CTVI is used for both treatment planning and tracking pulmonary ventilation function using cone-beam CT acquired during treatment. Therefore, the

accuracy of the CTVI must be independent of the CT image quality. To achieve this, the image quality must be improved using noise reduction and image correction techniques. Consequently, it could improve the accuracy of treatment planning and patient outcomes. Therefore, CTVI plays a crucial role in radiotherapy treatment planning, and it is desirable to improve its reliability and robustness through various methods.

In recent years, in addition to conventional filtering techniques such as median and Gaussian filters (16), image denoising methods using artificial intelligence (AI) have also been increasingly utilized in the field of medical imaging. In particular, deep learning methods based on convolutional neural networks (CNNs) have attracted attention as they can suppress noise while preserving structural details (17–19). Such AI-based preprocessing techniques are being explored as potential means to enhance the robustness and reproducibility of CT-based functional imaging, including CTVI.

In this study, we used a nonrigid alveoli phantom with ventilation functionality, which we developed as the world's first quality control tool for CTVI. We investigated the effect of preprocessing using both a conventional median filter and a deep learning-based denoising model on the accuracy and robustness of CTVIs. The purpose of this study is to evaluate how preprocessing methods, including AI-based denoising, affect the quality of CTVI, and to clarify their potential to improve robustness and accuracy in clinical applications.

2 Subjects and methods

2.1 CT datasets

The expiratory and inspiratory CT images were acquired using a 16-row detector CT scanner (Aquilion LB, Toshiba Medical Systems, Otawara, Japan). Image resolution was set to $0.78 \times 0.78 \times 3$ mm, and a helical scan protocol was used. The scan parameters were set to 120 kVp, 300 mA, rotation time of 0.5 s, and slice thickness of 3.0 mm. The nonrigid alveoli phantom comprised an acrylic cylinder filled with polyurethane foam simulating alveoli, a polyurethane membrane simulating the diaphragm, a metal rod with piston function simulating respiratory muscles, and a polyurethane tube simulating the airway (14). Various motion patterns can be programmed to simulate breathing patterns of various frequencies. Additionally, airflow can be controlled by pressure changes in the vessel owing to diaphragm movement. The phantom was placed horizontally and adjusted to align with the longitudinal axis of the CT system. The respiratory cycle of the phantom was set to 10 s. The normal respiratory cycle is approximately 4 s; however, to focus only on quantum noise, the respiratory cycle of the phantom was set at which the motion artifact was as small as possible.

2.2 Simulation of quantum noise and noise reduction by the median filter

Additional noise was applied to the CT images to simulate the quantum noise in a simplified manner. The amplitude of

quantum noise can be mathematically approximated by a Gaussian distribution (20). In this study, a Python script was developed to add noise with varying standard deviation [0–150 Hounsfield units (HU)] based on a normal distribution. Quantum noise was added using this script to a set of three pairs of expiratory and inspiratory images (30, 80, and

150 HU) to simulate the quantum noise in the CT images, as shown in Figure 1. We developed a script to calculate the median filter for the CT images with simulated quantum noise and fit it to all simulated noise. The filter used a kernel size of 3×3 .

2.3 Noise reduction by the deep learning-based denoising model

We constructed a denoising model based on a two-dimensional U-Net architecture (24) (Figure 2), with a network depth of 3 and an initial number of filters set to 32. The input to the model was a noisy image, and the output was the corresponding denoised image. To enable fair comparison with conventional filtering methods, no normalization was applied to the pixel values. Each image had a resolution of 512×512 pixels, and 131 slices were used per subject. For training, 14 types of Gaussian noise with standard deviations of 10, 20, 50, 60, 70, 100, 110, 120, 130, 140, 170, 180, 190, and 200 HU were added to clean images. For validation, noise levels of 40, 90, and 160 HU were used, and for testing, levels of 30, 80, and 150 HU were selected. The model was trained using the Adam optimizer with a batch size of 8 for up to 500 epochs. The L1 norm loss is defined as shown in Equation 1:

$$l(x, y) = \frac{1}{N} \sum_{n=1}^N |x_n - y_n|, \quad (1)$$

where x_n and y_n denote the predicted and ground-truth pixel values, respectively, and N is the total number of pixels. For the validation dataset, the clean images were used as the ground truth, and the model yielding the lowest loss between the output and the clean images was selected for final testing. All training and evaluation were performed on a workstation equipped with an Intel Core i9-9920X 3.5 GHz twelve-core processor, 32 GB RAM, and an NVIDIA GeForce RTX 2080 Ti GPU running Ubuntu 22.04.4 LTS with NVIDIA Driver 535.183.01, CUDA 12.1, and cuDNN 8.9.7.29-1.

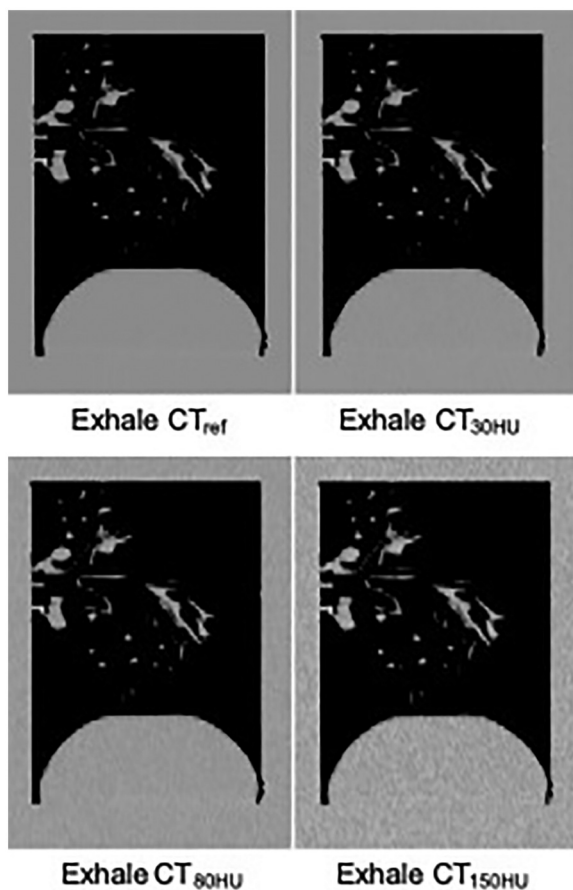


FIGURE 1
Comparison of exhalation CT images simulating different noise levels.

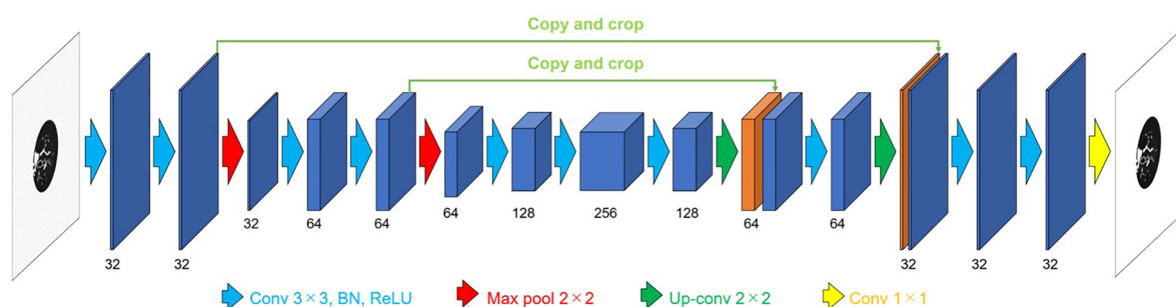


FIGURE 2
The denoising model based on a two-dimensional U-Net architecture.

2.4 Deformable image registration

In this study, the inhalation image was deformed to match the reference expiration image. Deformations were performed on a set of ten pairs of expiratory and inspiratory images: the reference image CT_{ref} without additional noise, CT_{noise} with noise (30, 80, and 150 HU), and CT_{med} and CT_{cnn} , which are denoised versions of CT_{noise} using a median filter and the CNN model, respectively. Deformable image registration was performed using NiftyReg (version 1.4.2), a free and open-source software package for non-rigid image registration. NiftyReg uses a B-spline-based free-form deformation algorithm, which estimates the transformation between moving and reference images by optimizing a normalized mutual information while applying smoothness constraints. The DIR parameters used for these deformations were the optimal parameters reported in a previous study (15): “*bending-energy penalty term*,” introduced in the cost function to smooth deformations; “*max number of iterations*,” which affects the computation time; “*number of levels to perform*,” which refers to the number of optimization calculations; and “*Jacobian-based penalty term*,” which penalizes large local volume changes and prevents folding (21). The deformation was performed in four steps, following the deformation strategy previously reported as optimal in earlier studies (15). Each step was visually checked, and if the deformation was over-deformed, the deformation step was omitted. The deformation vector field was obtained at each step. It was input at the next step and integrated for each deformation. The CT scans in this study were performed in one imaging session and the phantom outline was not moving; therefore, no rigid registration was performed before the deformation process.

2.5 CT-based ventilation imaging

The sum of the deformation vector field acquired for each of the seven paired sets was converted to a Jacobian determinant to obtain the respective $CTVI_{ref}$, $CTVI_{noise}$ (30, 80, and 150 HU), and $CTVI_{med}$ (30, 80, and 150 HU), and $CTVI_{cnn}$ (30, 80, and 150 HU). The DIR-based Jacobian metric was developed by Reinhardt et al. and is a measure of spatial volume change; it ensures that local volume changes do not alter the signal throughout the volume (22). The Jacobian determinant was calculated for each voxel in the phantom using (Equation 2).

$$\text{Jacobian determinant}(x, y, z) = \begin{vmatrix} 1 + \frac{\partial u_x(x, y, z)}{\partial x} & \frac{\partial u_x(x, y, z)}{\partial y} & \frac{\partial u_x(x, y, z)}{\partial z} \\ \frac{\partial u_y(x, y, z)}{\partial x} & 1 + \frac{\partial u_y(x, y, z)}{\partial y} & \frac{\partial u_y(x, y, z)}{\partial z} \\ \frac{\partial u_z(x, y, z)}{\partial x} & \frac{\partial u_z(x, y, z)}{\partial y} & 1 + \frac{\partial u_z(x, y, z)}{\partial z} \end{vmatrix} \quad (2)$$

where u_x , u_y , and u_z are the x, y, and z components of u , respectively. Jacobian determinant measures the expansion and contraction at position (x, y, z) in the image. When Jacobian

determinant is greater than one, local tissue expansion is present, and when Jacobian determinant is less than one, local tissue contraction is present. Jacobian determinant is a relative measure of ventilatory functionality on a voxel-by-voxel basis within the lung.

2.6 Evaluation of spatial deformation accuracy by DIR

Twenty-five landmarks were manually placed by an experienced medical physicist in a volume near the pulmonary vessels and bronchi in a nonrigid alveoli phantom (Figure 3). The target displacement error, that is, the displacement of a landmark due to respiratory motion, was measured as the Euclidean distance between the exhalation and inhalation images. The Euclidean distance was calculated using the formula shown in Equation 3:

$$\sqrt{(x_r - x_t)^2 + (y_r - y_t)^2 + (z_r - z_t)^2}, \quad (3)$$

where (x_r, y_r, z_r) and (x_t, y_t, z_t) are the landmark coordinates of the reference and target images, respectively. To evaluate the spatial accuracy of DIR with added noise, we used the Euclidean distance between the corresponding landmarks defined in the expiratory and deformed inspiratory images to calculate the Euclidean

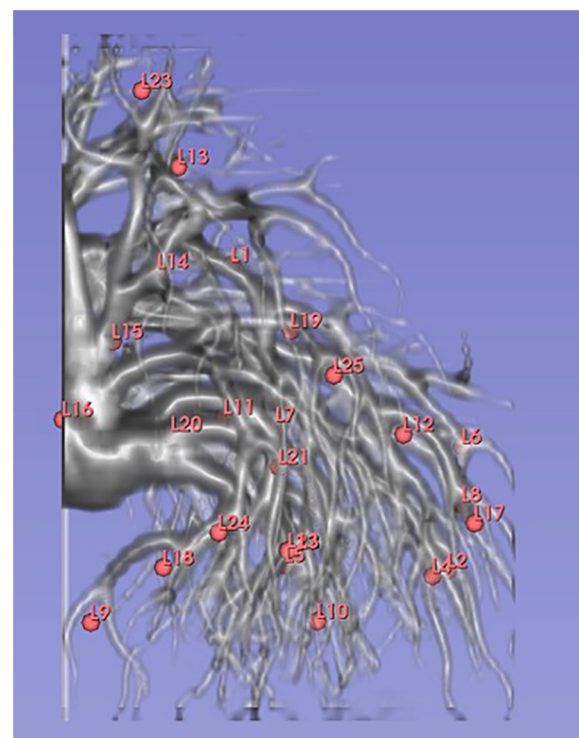


FIGURE 3
Twenty-five landmark setups placed in a volume near the pulmonary vessels and bronchi in a nonrigid alveoli phantom.

distance, which is denoted as TRE. TRE represents the spatial 3D distance discrepancy. When the deformed image perfectly matches the reference expiratory image (Euclidean distance = 0), TRE is equal to zero. Relative spatial accuracy was evaluated and compared with the spatial accuracy of the reference noiseless DIR.

2.7 Global consistency analysis using kappa statistics

To evaluate the clinical consistency of CTVI for treatment planning, each voxel in the CTVI was categorized into three regions—high, middle, and low ventilation—by evenly dividing the range of ventilation values in $CTVI_{ref}$. This categorization reflects a typical clinical scenario where high-ventilation regions are avoided during irradiation. The same classification thresholds were applied to all other CTVIs, including $CTVI_{noise}$ (30, 80, and 150 HU), $CTVI_{med}$ (30, 80, and 150 HU), and $CTVI_{cnn}$ (30, 80, and 150 HU). Although the absolute ventilation values may differ, consistency was defined as the regions categorized as high, middle, or low in $CTVI_{ref}$ being similarly categorized in the compared CTVIs. To quantify consistency, the proportion of voxels in each test CTVI that retained the same categorical label (high, middle, or low) as in $CTVI_{ref}$ was calculated. The degree of agreement was assessed using Cohen's kappa coefficient.

2.8 Voxel-based local evaluation using 2D histograms and spearman correlation

To evaluate the consistency of local ventilation function in each voxel, a two-dimensional (2D) histogram was constructed by plotting the Jacobian determinant value of each voxel in $CTVI_{ref}$ against the corresponding value in $CTVI_{noise}$ (30, 80,

and 150 HU). Similarly, 2D histograms were created for $CTVI_{med}$ (30, 80, and 150 HU) and $CTVI_{cnn}$ (30, 80, and 150 HU), which were generated by denoising CT_{noise} using a median filter and the CNN model, respectively. All histograms were generated based on voxel-wise spatial correspondence with $CTVI_{ref}$, enabling direct comparison of local ventilation values. Spearman's rank correlation coefficients were calculated from the 2D histograms to evaluate the consistency between each CTVI and the reference.

3 Results

3.1 Evaluation of DIR spatial deformation accuracy

Figure 4 presents a comparison of TRE values, indicating the spatial accuracy of DIR between CT images containing noise and CT image pairs with noise removed using the median filter and CNN-based denoising. The average displacement between the expiratory and inspiratory CT images was 14.59 ± 6.42 mm. The mean TRE values of the 25 landmarks were 1.39 ± 0.89 mm (maximum 2.95 mm) for CT_{ref} . The mean TRE values of the 25 landmarks were 1.22 ± 0.65 mm (maximum 2.54 mm), 0.71 ± 0.45 mm (maximum 1.89 mm), and 1.10 ± 0.83 mm (maximum 2.71 mm) for CT_{noise} (30, 80, and 150 HU). The mean TRE values for CT_{med} were 1.78 ± 0.71 mm (maximum 2.93 mm), 1.77 ± 0.78 mm (maximum 2.94 mm), and 1.42 ± 0.65 mm (maximum 2.80 mm) at 30, 80, and 150 HU, respectively. The mean TRE values for CT_{cnn} were 1.34 ± 0.00 mm (maximum 2.55 mm), 1.28 ± 0.00 mm (maximum 2.82 mm), and 1.22 ± 0.00 mm (maximum 2.45 mm) at 30, 80, and 150 HU, respectively. TREs for all conditions, including noisy and denoised images, remained within 3 mm. When comparing mean TRE values between CT_{noise} and denoised

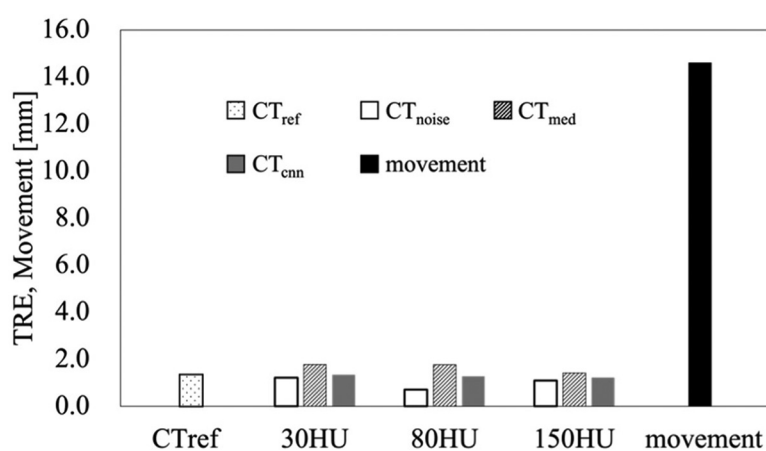


FIGURE 4

Comparison of TRE values was performed for CT_{noise} , CT_{med} , CT_{cnn} , and CT_{ref} at different noise levels (30, 80, and 150 HU), where CT_{ref} represents the DIR results based on CT images without added quantum noise. In this context, movement indicates the displacement caused by respiratory motion between the expiratory and inspiratory phases.

images at each noise level, TREs were higher in CT_{med} , while CT_{cnn} yielded values similar to or slightly higher than CT_{noise} .

3.2 Evaluation of CTVIs

3.2.1 Visual assessment of CTVIs

Figure 5 shows a visual comparison of CTVIs, including $CTVI_{ref}$, $CTVI_{noise}$, $CTVI_{med}$, and $CTVI_{cnn}$ at different noise levels. The visual assessment reveals that, compared to $CTVI_{ref}$, the location of high-functioning regions near the diaphragm remains consistent. However, additional high-functioning regions not observed in $CTVI_{ref}$ are present, and the resolution of ventilatory function distribution is reduced. The visual assessment indicates that $CTVI_{med}$ (30 HU) is closer to $CTVI_{ref}$ than the corresponding $CTVI_{noise}$ shown in Figure 5, suggesting that noise reduction improves the visual accuracy of CTVI under lower noise conditions. $CTVI_{cnn}$ appear visually closer to $CTVI_{ref}$ than $CTVI_{noise}$ at all noise levels, indicating that noise reduction improves the visual accuracy of CTVI. Among the three noise levels, $CTVI_{cnn}$ at 150 HU shows a particularly notable improvement over

$CTVI_{noise}$, suggesting that CNN-based denoising enhances the visualization of ventilation distribution under high noise conditions.

3.2.2 Global consistency analysis using kappa statistics

To evaluate the clinical utility of CTVI for treatment planning, ventilation values in $CTVI_{ref}$ were evenly divided into three regions: high, middle, and low ventilation. The same categorization was applied to $CTVI_{noise}$, $CTVI_{med}$ and $CTVI_{cnn}$ at each noise level. Consistency was defined as regions categorized as high, middle, or low in $CTVI_{ref}$ being similarly categorized in $CTVI_{noise}$, $CTVI_{med}$ and $CTVI_{cnn}$, regardless of absolute ventilation values. If the categorizations were perfectly consistent, the percentage of voxels correctly matching high, middle, and low regions would be 33.3% for each category. Figure 6 summarizes the percentage of voxels that correctly matched high, middle, and low regions between $CTVI_{ref}$ and $CTVI_{noise}$. For example, at 30 HU noise, only 11.32%, 12.64%, and 12.22% of voxels in the high, middle, and low regions of $CTVI_{ref}$ were correctly identified in $CTVI_{noise}$, respectively, with notable mismatches such as 14.36% of voxels categorized as low in

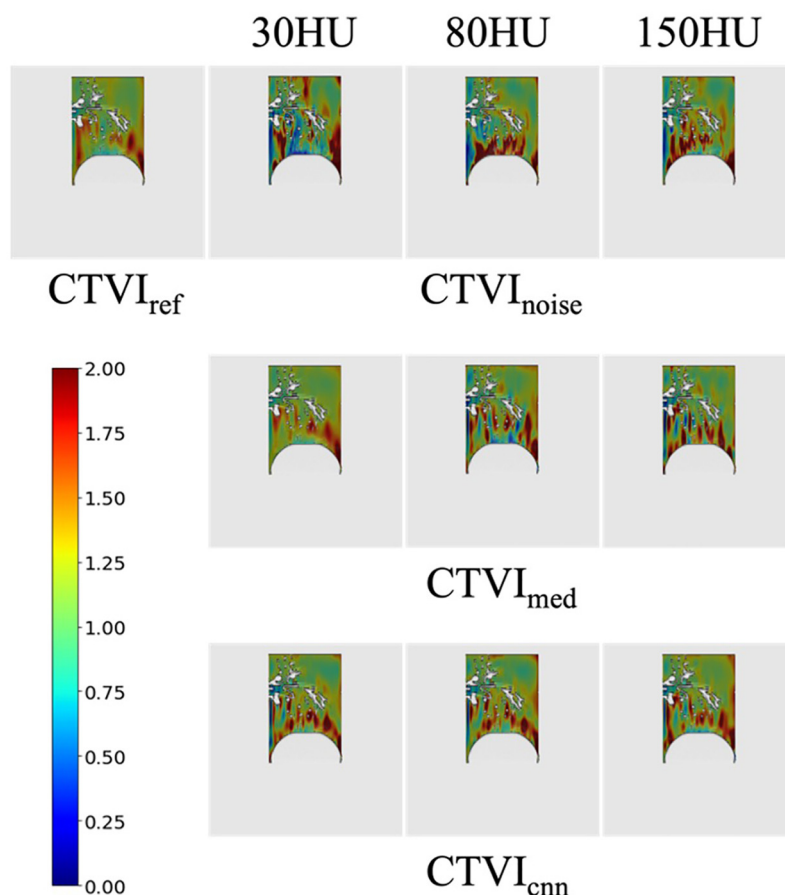


FIGURE 5

$CTVI_{noise}$, $CTVI_{med}$ and $CTVI_{cnn}$ at different noise levels. High-functioning regions are shown in red, while regions with decreased lung ventilation are displayed in blue.

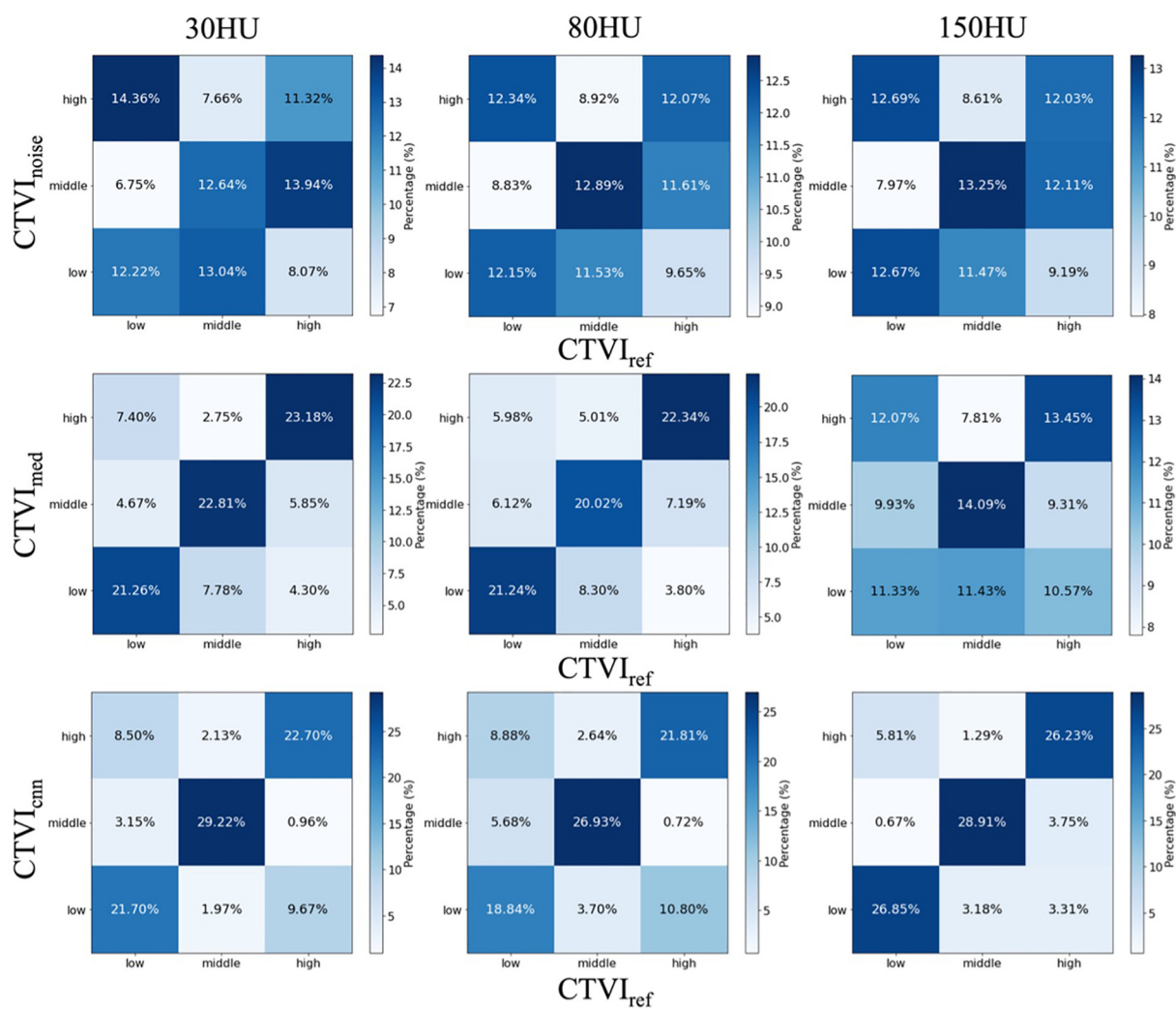


FIGURE 6 Proportion of categorized voxels across high, middle, and low ventilation regions for CTVI_{noise}, CTVI_{med} and CTVI_{cnn} compared to CTVI_{ref}. The horizontal axis represents CTVI_{ref}, while the vertical axis shows the three ventilation categories (high, middle, and low) for CTVI_{noise}, CTVI_{med} and CTVI_{cnn}. Darker colors indicate higher agreement in classification.

CTVI_{ref} being misclassified as high in CTVI_{noise}. In contrast, CTVI_{med} at 30 HU showed substantially improved agreement, with 21.26%, 22.81%, and 23.18% correctly matching the high, middle, and low regions of CTVI_{ref} respectively. Additionally, at 80 HU noise, CTVI_{med} also demonstrated an improvement in consistency compared to CTVI_{noise}. The percentages of voxels correctly categorized as high, middle, and low in CTVI_{med} were 21.24%, 20.02%, and 22.34%, respectively, indicating a better agreement with CTVI_{ref} compared to CTVI_{noise}, where the corresponding percentages were only 12.15%, 12.89%, and 12.07%. This result highlights that the noise reduction processing improved clinical consistency not only at 30 HU but also at 80 HU noise levels. In addition, CTVI_{cnn} showed consistent improvement across all noise levels. Notably, at 150 HU, the consistency between CTVI_{cnn} and CTVI_{ref} was significantly higher than that of CTVI_{noise}. The percentage of correctly matched voxels was 26.23%, 28.91%, and 26.85% for high, middle, and low ventilation regions,

TABLE 1 Cohen's kappa coefficients of CTVIs for CTVI_{ref}.

CTVI methods	30 HU (P value)	80 HU (P value)	150 HU (P value)
CTVI _{noise}	0.043 (<.0001)	0.057 (<.0001)	0.069 (<.0001)
CTVI _{med}	0.51 (<.0001)	0.45 (<.0001)	0.083 (<.0001)
CTVI _{cnn}	0.60(<.0001)	0.51 (<.0001)	0.73 (<.0001)

respectively, compared to only 12.03%, 13.25%, and 12.67% for CTVI_{noise}. These results highlight that CNN-based denoising particularly improves clinical consistency under high-noise conditions. Cohen's kappa coefficients further quantified the agreement between CTVI_{ref} and both CTVI_{noise} and CTVI_{med}. Table 1 presents these results, showing that kappa values for CTVI_{noise} were 0.043, 0.057, and 0.069 at noise levels of 30 HU, 80 HU, and 150 HU, respectively. For CTVI_{med}, the kappa values were significantly higher at lower noise levels (0.51 at 30 HU and

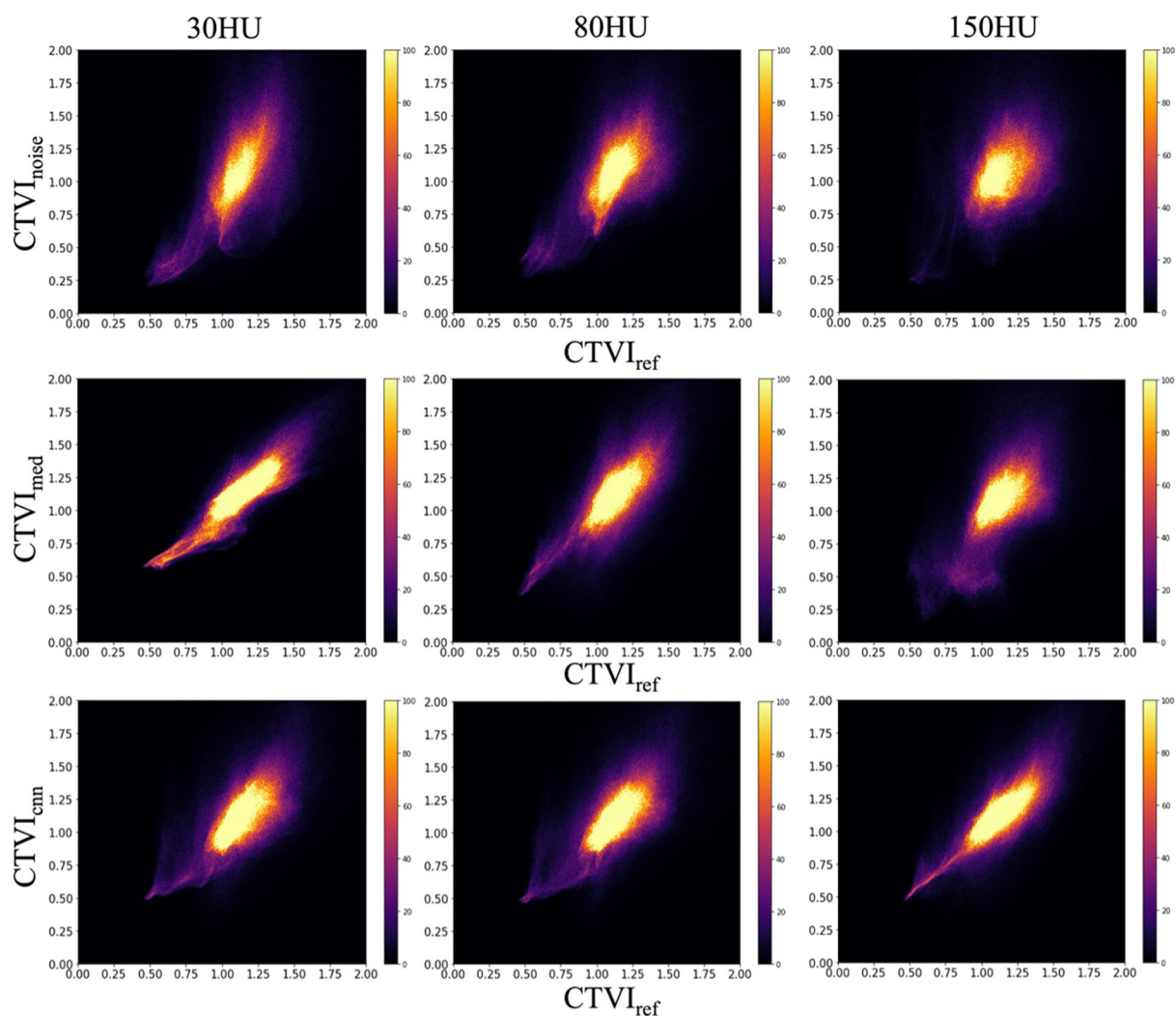


FIGURE 7 Two-dimensional histograms of $CTVI_{noise}$, $CTVI_{med}$ and $CTVI_{cnn}$ at different noise levels with respect to $CTVI_{ref}$. The horizontal axis represents $CTVI_{ref}$, while the vertical axis shows the voxel-based lung ventilation of $CTVI_{noise}$, $CTVI_{med}$ and $CTVI_{cnn}$. Perfect agreement is indicated by the histogram distribution along $y = x$.

0.45 at 80 HU), but the consistency diminished at 150 HU (0.083). $CTVI_{cnn}$ exhibited consistently high kappa values across all noise levels, with 0.60 at 30 HU, 0.51 at 80 HU, and 0.73 at 150 HU, indicating superior categorical agreement with $CTVI_{ref}$ compared to both $CTVI_{noise}$ and $CTVI_{med}$. These results support the effectiveness of CNN-based denoising in preserving clinically relevant ventilation patterns, even under high noise conditions.

3.2.3 Voxel-based local evaluation using 2D histograms and spearman correlation

To assess the consistency of local ventilation function, a voxel-by-voxel comparison between $CTVI_{ref}$ and both $CTVI_{noise}$, $CTVI_{med}$ and $CTVI_{cnn}$ was conducted using 2D histograms (Figure 7). These histograms demonstrate that the distribution improves as it approaches $y = x$, indicating greater consistency with $CTVI_{ref}$. At a noise level of 30 HU, the histogram of $CTVI_{med}$ shows a marked improvement in consistency compared to $CTVI_{noise}$, as the

TABLE 2 Spearman correlation coefficients of CTVIs for $CTVI_{ref}$.

CTVI methods	30 HU (P value)	80 HU (P value)	150 HU (P value)
$CTVI_{noise}$	0.61 (<.0001)	0.59 (<.0001)	0.33 (<.0001)
$CTVI_{med}$	0.87 (<.0001)	0.70 (<.0001)	0.61 (<.0001)
$CTVI_{cnn}$	0.59 (<.0001)	0.65 (<.0001)	0.83 (<.0001)

distribution is more concentrated along $y = x$. However, at noise levels of 80 and 150 HU, no significant improvement was observed. In contrast, $CTVI_{cnn}$ demonstrated a different trend: while there was no notable improvement in the 2D histograms at 30 and 80 HU, the histogram at 150 HU showed greater alignment with the $y = x$ line, suggesting improved voxel-level consistency. Table 2 presents the Spearman correlation coefficients for these comparisons. The results show that $CTVI_{med}$ exhibits better correlation with $CTVI_{ref}$ than $CTVI_{noise}$ at lower noise levels, but the correlation decreases as the

noise level increases. Despite this, a statistically significant trend toward improvement is observed (<0.0001). The Spearman correlation coefficients for $CTVI_{cnn}$ were 0.59, 0.65, and 0.83 at 30, 80, and 150 HU, respectively. This indicates that although $CTVI_{cnn}$ did not improve correlation at lower noise levels, it showed a substantial improvement at 150 HU (<0.0001).

4 Discussion

Despite the extensive body of research on CTVI (4–10), few studies have employed phantoms capable of replicating human-like ventilation in the context of clinical applications. While it is well known that noise significantly affects CT image analysis, the impact of noise reduction preprocessing on the accuracy and consistency of CTVI remains insufficiently explored. In this study, quantum noise was simulated in CT images using a nonrigid alveoli phantom designed to mimic lung motion.

In this study, the spatial accuracy of the DIR was evaluated by simulating noise levels from 0 to 150 HU. For all CT_{ref} , CT_{noise} , CT_{med} and CT_{cnn} , the deformation accuracy was within 3 mm of the tolerances given in TG-132 (23), regardless of the noise level. Although previous studies have investigated noise levels of 200 HU (13), in the initial experiments of this study, the noise level of 200 HU resulted in over-deformation due to DIR for the same deformation parameters, and an accurate CTVI could not be established. This result suggests that noise levels above 200 HU significantly affect the deformation accuracy of the DIR. High noise levels may lead to errors in the DIR algorithm. Therefore, the results of this study suggest that the DIR technique has sufficient accuracy for generating CTVI at quantum noise levels up to 150 HU. Although all TRE values were within the TG-132 tolerance of 3 mm, CT_{med} and CT_{cnn} exhibited slightly higher TREs compared to CT_{noise} . One possible explanation is that the denoising process may have smoothed out anatomical features critical for deformable registration, resulting in slightly reduced precision. Alternatively, mild quantum noise may have enhanced local contrast in CT_{noise} , unintentionally aiding DIR alignment. However, these differences remained within the clinically acceptable margin and are unlikely to affect the final CTVI outcome.

Table 1; Figure 6 illustrate the impact of quantum noise and the application of preprocessing filters on the consistency of CTVI from a clinical perspective. When using CTVI for treatment planning, lung ventilation is categorized into three levels—high, middle, and low—and treatment plans are designed to avoid high-function regions. In $CTVI_{ref}$, approximately 33% of the lung ventilation is classified into each category. Ideally, in cases of accurate classification, the relationships between $CTVI_{ref}$, $CTVI_{noise}$, $CTVI_{med}$ and $CTVI_{cnn}$ should result in high-high, middle-middle, and low-low matches approaching 33%. Focusing on $CTVI_{noise}$, the maximum agreement across all noise levels was only 13.25%, indicating significant misclassification of lung ventilation when quantum noise is present. In contrast, $CTVI_{med}$ achieved over 20% agreement in all ventilation categories at noise

levels below 80 HU. A particularly important observation is the proportion of regions classified as high in $CTVI_{ref}$ but misclassified as low. This proportion was kept below 4.3% at its maximum. In addition, CNN-based denoising further improved the consistency of CTVI at all noise levels. Notably, $CTVI_{cnn}$ achieved the highest agreement with $CTVI_{ref}$, particularly under high-noise conditions. At 150 HU, CNN showed the greatest improvement in both categorical agreement and voxel-wise correlation ($\kappa = 0.73$, Spearman $\rho = 0.83$), outperforming both $CTVI_{noise}$ and $CTVI_{med}$. These results indicate that CNN-based denoising has strong potential to enhance the robustness of CTVI, even under clinically challenging noise conditions.

Table 2; Figure 7 focus on the voxel-level accuracy of CTVI, demonstrating the local effects of quantum noise and the application of preprocessing filters on CTVI accuracy using two-dimensional histograms and Spearman correlation. The results of this study confirmed that as noise levels increase, the accuracy of CTVI decreases, proving that quantum noise is a significant factor that hinders the accuracy of CTVI. However, CNN-based denoising yielded stronger improvements in correlation, especially at 150 HU, suggesting it is a more robust solution in high-noise environments.

Interestingly, while Cohen's kappa coefficients and Spearman correlation coefficients generally showed consistent trends, some discrepancies were noted. For example, at 150 HU, $CTVI_{med}$ yielded a relatively high Spearman correlation ($\rho = 0.61$) but a low kappa value ($\kappa = 0.083$), suggesting that voxel-wise rankings were preserved even though many values crossed categorical thresholds. Conversely, at 30 HU, $CTVI_{cnn}$ showed a high kappa ($\kappa = 0.60$) despite having a lower Spearman correlation ($\rho = 0.59$), indicating that category-level agreement was strong, while voxel value variations limited rank correlation. These findings emphasize that categorical and continuous metrics capture different aspects of agreement, and highlight the need for using both to comprehensively evaluate CTVI accuracy.

Limitations of this study include the lack of comparison with vendor-provided denoising techniques and the absence of hybrid preprocessing strategies. Future work should explore combining CNN-based and conventional filtering approaches and testing these methods in patient datasets.

5 Conclusion

This study quantitatively evaluated the effect of preprocessing on the accuracy and robustness of the CTVI using a nonrigid alveoli phantom with ventilation functionality, developed as the world's first quality control tool for CTVI. We demonstrated that quantum noise significantly impairs the accuracy and consistency of CTVI. While median filtering was shown to be a simple and effective method for mitigating this effect, CNN-based denoising provided superior performance, particularly under high-noise conditions. These findings suggest that both conventional and AI-based preprocessing approaches contribute to improving the

quality of CTVI, with deep learning methods offering strong potential to enhance robustness and accuracy in clinical applications.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

SM: Conceptualization, Investigation, Methodology, Writing – original draft. HF: Methodology, Project administration, Writing – review & editing. KY: Data curation, Methodology, Writing – review & editing. NK: Data curation, Formal analysis, Investigation, Validation, Writing – review & editing. MT: Formal analysis, Investigation, Writing – review & editing. HN: Formal analysis, Investigation, Writing – review & editing. SM: Data curation, Methodology, Writing – review & editing. FT: Investigation, Methodology, Writing – review & editing. TF: Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by JSPS KAKENHI Grant Number JP22K15891.

References

1. Suga K. Technical and analytical advances in pulmonary ventilation SPECT with xenon-133 gas and tc-99m-technegas. *Ann Nucl Med.* (2002) 16:303–10. doi: 10.1007/BF02988614
2. van Beek EJR, Wild JM, Kauczor H-U, Schreiber W, Mugler JP 3rd, de Lange EE. Functional MRI of the lung using hyperpolarized 3-helium gas. *J Magn Reson Imaging.* (2004) 20:540–54. doi: 10.1002/jmri.20154
3. Harris RS, Schuster DP. Visualizing lung function with positron emission tomography. *J Appl Physiol.* (1985). (2007) 102:448–58. doi: 10.1152/jappphysiol.00763.2006
4. Latifi K, Forster KM, Hoffer SE, Dilling TJ, van Elmpot W, Dekker A, et al. Dependence of ventilation image derived from 4D CT on deformable image registration and ventilation algorithms. *J Appl Clin Med Phys.* (2013) 14:4247. doi: 10.1120/jacmp.v14i4.4247
5. Yamamoto T, Kabus S, von Berg J, Lorenz C, Keall PJ. Impact of four-dimensional computed tomography pulmonary ventilation imaging-based functional avoidance for lung cancer radiotherapy. *Int J Radiat Oncol Biol Phys.* (2011) 79:279–88. doi: 10.1016/j.ijrobp.2010.02.008
6. Yaremko BP, Guerrero TM, Noyola-Martinez J, Guerra R, Lege DG, Nguyen LT, et al. Reduction of normal lung irradiation in locally advanced non-small-cell lung cancer patients, using ventilation images for functional avoidance. *Int J Radiat Oncol Biol Phys.* (2007) 68:562–71. doi: 10.1016/j.ijrobp.2007.01.044
7. Eslick EM, Kipritidis J, Gradinscak D, Stevens MJ, Bailey DL, Harris B, et al. CT ventilation imaging derived from breath hold CT exhibits good regional accuracy with galligas PET. *Radiother Oncol.* (2018) 127:267–73. doi: 10.1016/j.radonc.2017.12.010
8. Kipritidis J, Tahir BA, Cazoulat G, Hofman MS, Siva S, Callahan J, et al. The VAMPIRE challenge: a multi-institutional validation study of CT ventilation imaging. *Med Phys.* (2019) 46:1198–217. doi: 10.1002/mp.13346
9. Castillo R, Castillo E, McCurdy M, Gomez DR, Block AM, Bergsma D, et al. Spatial correspondence of 4D CT ventilation and SPECT pulmonary perfusion defects in patients with malignant airway stenosis. *Phys Med Biol.* (2012) 57:1855–71. doi: 10.1088/0031-9155/57/7/1855
10. Yamamoto T, Kabus S, Lorenz C, Mittra E, Hong JC, Chung M, et al. Pulmonary ventilation imaging based on 4-dimensional computed tomography: comparison with pulmonary function tests and SPECT ventilation images. *Int J Radiat Oncol Biol Phys.* (2014) 90:414–22. doi: 10.1016/j.ijrobp.2014.06.006
11. Yamamoto T, Kabus S, Bal M, Keall PJ, Moran A, Wright C, et al. Four-dimensional computed tomography ventilation image guided lung functional avoidance radiation therapy: a single-arm prospective pilot clinical trial. *Int J Radiat Oncol Biol Phys.* (2022) 115:1144–54. doi: 10.1016/j.ijrobp.2022.11.026
12. Castillo E, Castillo R, Vinogradskiy Y, Guerrero T. The numerical stability of transformation-based CT ventilation. *Int J Comput Assist Radiol Surg.* (2017) 12:569–80. doi: 10.1007/s11548-016-1509-x
13. Latifi K, Huang T-C, Feygelman V, Budzevich MM, Moros EG, Dilling TJ, et al. Effects of quantum noise in 4D-CT on deformable image registration and derived ventilation data. *Phys Med Biol.* (2013) 58:7661–72. doi: 10.1088/0031-9155/58/21/7661
14. Miyakawa S, Tachibana H, Moriya S, Kurosawa T, Nishio T, Sato M. Design and development of a nonrigid phantom for the quantitative evaluation of DIR-based mapping of simulated pulmonary ventilation. *Med Phys.* (2018) 45:3496–505. doi: 10.1002/mp.13017
15. Miyakawa S, Tachibana H, Moriya S, Kurosawa T, Nishio T. Evaluation of deformation parameters for deformable image registration-based ventilation imaging using an air-ventilating non-rigid phantom. *Phys Med.* (2018) 50:20–5. doi: 10.1016/j.ejmp.2018.05.016
16. Gonzalez RC, Woods RE. *Digital Image Processing*. 4th ed. Boston, MA: Pearson (2018).

Acknowledgments

The authors would like to thank Editage (<https://www.editage.jp>) for the English language review.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

17. Chen H, Zhang Y, Kalra MK, Lin F, Chen Y, Liao P, et al. Low-dose CT with a residual encoder-decoder convolutional neural network (RED-CNN). *IEEE Trans Med Imaging*. (2017) 36:2524–35. doi: 10.1109/TMI.2017.2715284
18. Zhang K, Zuo W, Chen Y, Meng D, Zhang L. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. *IEEE Trans Image Process*. (2017) 26:3142–55. doi: 10.1109/TIP.2017.2662206
19. Yang Q, Yan P, Zhang Y, Yu H, Shi Y, Mou X, et al. Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. *IEEE Trans Med Imaging*. (2018) 37(6):1348–57. doi: 10.1109/TMI.2018.2827462
20. Herman GT. Image reconstruction from projections. In: *The Fundamentals of Computerized Tomography*. New York: Academic Press (1980). p. 141–2.
21. Modat M. Lung registration using the NiftyReg package. *Med Image Anal Clin Grand Chall*. (2010):33–42.
22. Reinhardt JM, Ding K, Cao K, Christensen GE, Hoffman EA, Bodas SV. Registration-based estimates of local lung tissue expansion compared to xenon CT measures of specific ventilation. *Med Image Anal*. (2008) 12:752–63. doi: 10.1016/j.media.2008.03.007
23. Brock KK, Mutic S, McNutt TR, Li H, Kessler ML. Use of image registration and fusion algorithms and techniques in radiotherapy: report of the AAPM radiation therapy committee task group No. 132. *Med Phys*. (2017) 44:e43–76. doi: 10.1002/mp.12256
24. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing (2015). p. 234–41.



OPEN ACCESS

EDITED BY

Shuo Li,
Western University, Canada

REVIEWED BY

Pranay Goel,
Indian Institute of Science Education and
Research, Pune, India
Sanat Phatak,
KEM Hospital Research Centre, India

*CORRESPONDENCE

Muhammad Syauqie
✉ m.syauqie@med.unand.ac.id

RECEIVED 14 February 2025

ACCEPTED 14 July 2025

PUBLISHED 30 July 2025

CITATION

Syauqie M, Patria H, Hastono SP,
Siregar KN and Moeloek NDF (2025) Deep
learning for vision screening in
resource-limited settings: development of
multi-branch CNN for refractive error
detection based on smartphone image.
Front. Comput. Sci. 7:1576958.
doi: 10.3389/fcomp.2025.1576958

COPYRIGHT

© 2025 Syauqie, Patria, Hastono, Siregar, and
Moeloek. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Deep learning for vision screening in resource-limited settings: development of multi-branch CNN for refractive error detection based on smartphone image

Muhammad Syauqie^{1,2*}, Harry Patria³, Sutanto Priyo Hastono²,
Kemal Nazaruddin Siregar² and Nila Djuwita Farieda Moeloek⁴

¹Department of Ophthalmology, Faculty of Medicine, Universitas Andalas, Padang, Indonesia,

²Department of Biostatistics and Population Studies, Faculty of Public Health, Universitas Indonesia, Depok, Indonesia, ³Strathclyde Business School, University of Strathclyde, Glasgow, United Kingdom,

⁴Department of Ophthalmology, Faculty of Medicine, Universitas Indonesia, Cipto Mangunkusumo Hospital, Jakarta, Indonesia

Introduction: Uncorrected refractive errors are a leading cause of preventable vision impairment globally, particularly affecting individuals in low-resource regions where timely diagnosis and screening access remain significant challenges despite the availability of economical treatments.

Aim: This study introduces a novel deep learning-based system for automated refractive error classification using photorefractive images acquired via a standard smartphone camera.

Methods: A multi-branch convolutional neural network (CNN) was developed and trained on a dataset of 2,139 corneal images collected from an Indonesian public eye hospital. The model was designed to classify refractive errors into four categories: significant myopia, significant hypermetropia, insignificant refractive error, and not applicable to classified. Grad-CAM visualization was employed to provide insights into the model's interpretability.

Results: The 3-branch CNN architecture demonstrated superior performance, achieving an overall test accuracy of 91%, precision of 96%, and recall of 98%, with an area under the curve (AUC) score of 0.9896. Its multi-scale feature extraction pathways were pivotal in effectively addressing overlapping red reflex patterns and subtle variations between classes.

Conclusion: This study establishes the feasibility of smartphone-based photorefractive assessment integrated with artificial intelligence for scalable and cost-effective vision screening. By training the CNN model with a real-world dataset representative of Southeast Asian populations, this system offers a reliable solution for early refractive error detection with significant implications for improving accessibility to eye care services in resource-limited settings.

KEYWORDS

refractive error detection, vision screening, artificial intelligence, convolutional neural network, smartphone, red reflex, photorefractive

Introduction

Uncorrected refractive errors are the leading cause of vision problems in children and adults. Worldwide, among the 338 million individuals with visual impairments, approximately 153 million are affected by uncorrected distance refractive errors, 8 million are blind, and 145 million experience considerable distance vision impairment (Holden, 2007; GBD 2019 Blindness and Vision Impairment Collaborators, 2021). Uncorrected refractive errors can lead to immediate and long-term consequences for children and adults, such as lost educational and job opportunities, reduced economic productivity for individuals, families, and communities, and a lowered quality of life (Resnikoff et al., 2008; Rahi et al., 2014).

Treatments for refractive errors, such as glasses, are economical and easily obtainable in high-income regions. However, refractive errors frequently go undiagnosed or are not referred for treatment, and some obstacles prevent people from accessing these services, especially in low-middle-income countries (Cochrane et al., 2010). Identifying significant refractive errors in communities remains a challenge in both diagnosis and public health. The challenge persists in finding a timely and cost-effective method to screen large groups of people. Screening approaches using professional equipment, such as autorefraction and photo-screening, have been developed to overcome the screening challenges in communities (Braverman, 2007). Such equipment entails high costs, the need for trained eye health professionals, and patient cooperation, which can be challenging to obtain in children. As a result, this poses difficulties for the screening of refractive disorders in communities in resource-limited areas (Yang et al., 2020).

Photorefractive equipment operates on a principle known as photorefractive assessment, which evaluates the red reflex pattern in the fundus. In addition, photorefractive assessment can be adapted to work with a smartphone camera to detect refractive errors. Characteristics, such as the position and size of the red reflex pattern, can provide insights into the specific type of refractive disorder that a person may have (Donahue, 2009; Cooper et al., 1996; Watts et al., 1999). Previous studies (Yang et al., 2020; Wang et al., 2024; Ma et al., 2020; Kiu et al., 2024) have shown the potential of smartphone-based photorefractive assessment in refractive error detection and vision screening initiatives, especially in regions with restricted access to eye care services. Nonetheless, this approach requires trained eye care professionals, particularly optometrists, to analyze the photorefractive images.

Artificial Intelligence (AI) can be a valuable supplementary resource developed to perform specific functions, like analyzing images and handling data processing. Integrating AI into smartphone-based photorefractive assessment could automate the analysis of photorefractive images, enabling accurate predictions of the type of refractive error. Deep learning (DL), a branch of artificial intelligence that employs convolutional neural networks (CNN) with several layers of algorithms, is used for image analysis's advanced feature extraction. This technique enables machines to process and learn autonomously from raw data while examining intricate non-linear relationships. A significant advantage of applying DL algorithms in the medical field is their effectiveness in medical image analysis and screening. Incorporating AI, especially DL, can significantly improve the efficiency and accessibility of eye care

services (Martinez-Perez et al., 2022; Keskinbora and Güven, 2020; Ng et al., 2021; Lim et al., 2022; Panda and Kumar Dhaka, 2022; Latif et al., 2019).

Previous studies have attempted to develop a DL model to predict refractive error. Varadarajan et al. (2018) created a deep-learning prediction model for estimating refractive error using retinal fundus images. The model demonstrates substantial accuracy in forecasting spherical power but lacks precision in predicting cylindrical power. Linde et al. (2023) used portable infrared smartphone-based fundus cameras to obtain red reflex images. However, the CNN model they developed using those images has a modest accuracy in predicting spherical power and a weak accuracy in predicting cylindrical power. Both studies utilized professional equipment or additional devices for smartphones to capture the necessary images for CNN model development. This can incur extra costs and limit their usefulness for vision screening. The development of the CNN model for refractive error prediction using images captured by the built-in smartphone camera has been demonstrated by Do et al. (2022), Fu et al. (2020), Chun et al. (2019), and Yang et al. (2020). These models show promising good accuracy in the validation phase. However, those models did not undergo the testing phase due to the small-size dataset limitation; thus, the actual accuracy score is not yet determined. The models were also trained with the eye image of the East Asian population, mainly of Chinese and Korean ethnicity. Hence, the validity and reliability of the models for use on another ethnicity still need further validation.

South East Asian populations have a high prevalence of refractive error and uncorrected refractive error (Cao et al., 2022; Pan et al., 2013; Hashemi et al., 2018). The vision screening program in this region is hindered by the lack of facilities and the scarcity of human resources. Smartphone-based photorefractive assessment incorporating artificial intelligence (AI) could be an interesting solution to expand the coverage of refractive error screening in this region. Developing an AI model, particularly the CNN model, which is trained with an image dataset derived from the Southeast Asia population, will provide a prompt and timely refractive error detection for this population. A previous study (Linde et al., 2023) shows low accuracy of the pre-trained CNN model for predicting refractive error using red reflex images. Therefore, we chose to develop a multi-branch CNN because single-branch CNNs were not able to differentiate well enough the subtle variations in the morphological patterns of the pupillary red reflex, which are crucial for differentiating the categories of refractive error like significant myopia and hypermetropia. The single-pathway architectures consistently failed to differentiate between subtle variations in pupillary red reflex patterns, resulting in a misclassification in refractive error classes. Finally, we selected the multi-branch CNN because this approach allows us to simultaneously capture features at multiple scales, which is essential for distinguishing crescent-shaped reflexes that often have overlapping characteristics between different refractive error classes. This study aimed to develop an automated deep learning-based prediction system to assess refractive error using photorefractive images of Indonesian patients taken with a smartphone. Another aim of this study is to train the Multi-branch CNN models using photorefractive images to detect different types of refractive errors, as we classified in our previous study (Syauqie et al., 2025), and test the models with different new datasets to establish the validity and reliability of this new CNN model.

Methods

Design and selection of subjects

This study design is a cross-sectional study. Subjects were patients who visited the outpatient clinic for refractive examination at a Public Eye Hospital in Pekanbaru, Riau Province, Indonesia, from January 2023 to June 2024 were invited to participate in this study. Patients with eyelid and ocular abnormalities that obscure the visual axis (cornea, lens, and vitreous), ocular infection or inflammation, and a history of ocular surgery were excluded from this study. A total of 952 patients, most of whom were of Malay ethnicity, participated in this study. The mean age (S.D.) of the patients was 20.06 (11.27) years, and the median was 18 years (range 1–55 years).

The study protocol was reviewed and approved by the Ethics Committee of the Faculty of Public Health Universitas Indonesia (Ket-541/UN2.F10.D11/PPM.00.02/2024). All patients underwent complete ophthalmic examinations, including uncorrected visual acuity (UCVA), slit lamp biomicroscope examination, fundus photography, objective refraction using an autorefractor (Tonoref II, Nidek Co., Japan), and subjective refraction to determine best-corrected visual acuity (BCVA). After complete ophthalmic examinations, the patients underwent both eye images were captured in a dark room with a smartphone camera.

Workflow of the CNN model development

The workflow for the Multi-branch CNN model begins with the acquisition of pupillary red reflex images using a smartphone (Figure 1). These images were preprocessed to ensure uniformity, including resizing to 256×256 pixels and normalization to a range of $[0,1][0,1][0,1]$. Data augmentation techniques, such as rotation, flipping, and zooming, are applied to increase the diversity of the dataset and improve the model's robustness.

The preprocessed images are fed into the selected CNN architecture—1-branch, 2-branch, or 3-branch—for feature extraction and classification. Each architecture processes the input differently, with the 3-branch model providing the most comprehensive feature representations. The classification output predicts one of the following four classes: “Not Applicable to Classified,” “Significant Hypermetropia,” “Insignificant Refractive Error,” and “Significant Myopia.”

The workflow culminates in model evaluation, where metrics such as accuracy, precision, recall, and F1-score are computed. The results are visualized through confusion matrices, ROC curves, and training-validation plots to assess model performance. The successful deployment of the model on a smartphone platform represents the final step, enabling real-time, low-cost refractive error screening in clinical and community settings.

Data acquisition

Eye images were acquired with a 12-megapixel smartphone camera (iPhone 6s, Apple Inc., California, United States) at a 50–60 cm distance from the front of the patient in a dark room (<10 lux) and using an external light source to the camera (Figure 2A). The smartphone was placed straightforward to the face of the patient without angulation. The patient's gaze is slightly upward (15 degrees) during image capture (Ik et al., 2015). The smartphone's built-in flash, which was present next to the left of the camera lens, was used as the light source for image capturing, wherein light entered the eye through the pupil, reached the retinal surface, and then reflected to the camera lens. Reflected light from the eye is shown as a characteristic crescent-shaped red reflex that appears in the pupil, which is called the pupillary red reflex (Chun et al., 2019). All the images were acquired in the same setting (in a dark room, in portrait orientation, and without instilling the cycloplegic agents) and done with one operator (MS) (Figure 2B). All images, each containing images of both eyes with pupillary red reflex, were stored in the database with ophthalmology examination records (Figure 2C).

Data preprocessing and augmentation

Data preprocessing ensures that the input images are standardized for optimal training performance. Each image was cropped to extract only the corneal portion so that each image could produce two corneal images. Each corneal image has a 72×72 pixels resolution. Each corneal image is labeled with one type of refractive error classification based on findings in our previous study (Syauqie et al., 2025). Corneal images with the same refractive error classification were stored in one folder. The images were labeled considering the red reflex pattern in the pupil, which indicates significant myopia, significant hypermetropia, and insignificant refractive error. Corneal images with

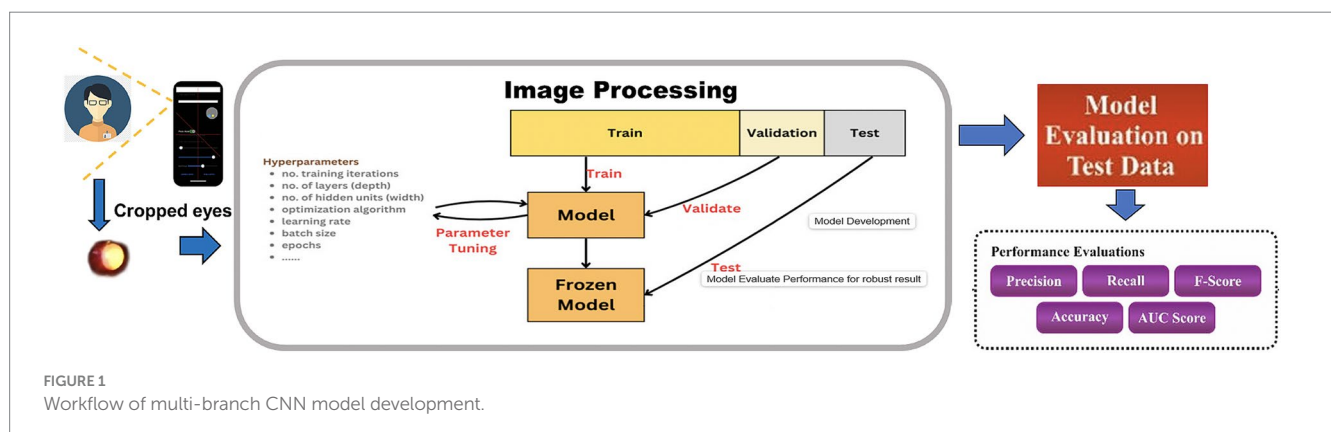


FIGURE 1
Workflow of multi-branch CNN model development.

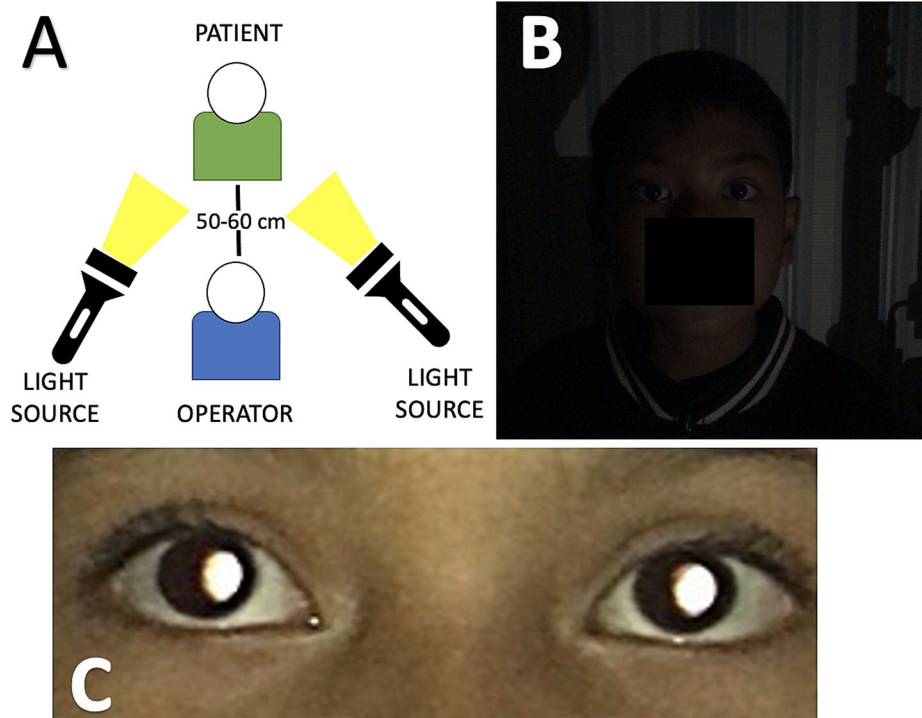


FIGURE 2

(A) Data acquisition process environment setting during image capture, the distance between the operator and the patient is about 50–60 cm. An external light source is positioned beside the operator and faces the patient. (B) Dim lightning conditions and portrait orientation before the image is captured. (C) The image captured by the smartphone shows the pupillary red reflex in both eyes.

the absence of pupillary red reflexes were labeled as not applicable to classification (Figures 3A–D). In addition to the red reflex pattern in the pupil, the ground truth of each type of refractive error classification also considers the spherical equivalent (SE) dioptric power acquired from objective refractive using an autorefractor (Tonoref II, Nidek Co., Japan) and subjective refraction examination. From the findings in our previous study (Syauqie et al., 2025), we found a significant correlation between the red reflex pattern and spherical equivalent dioptric power ($p < 0.001$). The crescent red reflex pattern, which is dominant on the right side of the pupil, has a spherical equivalent of -5.977 ± 3.655 diopters (95% CI: -6.393 to -5.562); the crescent red reflex pattern, which is dominant on the left side of the pupil, has a spherical equivalent of $+1.361 \pm 1.925$ diopters (95% CI: 0.941 – 1.782); and the red reflex pattern, which occupies the entire pupil area, has a spherical equivalent of -1.174 ± 1.325 diopters (95% CI: -1.334 to -1.015). These results support the use of subjective refraction and autorefractor data for refractive error type labeling, which can be further used for model training and testing.

We classified the refractive error type according to its SE dioptric power. Significant myopia was observed if the patient's eye had a dioptric power below -3.00 Diopter and the crescent red reflex pattern was dominant on the right side of the pupil; significant hypermetropia was observed if the dioptric power was above $+1.00$ Diopter and the crescent red reflex pattern was dominant on the left side of the pupil; and insignificant refractive error was observed if the dioptric power was between 3.00 and 0.50 Diopter and between $+0.50$ and $+1.00$ Diopter and the red reflex pattern occupied the entire pupillary area (Figures 3A–C). The dataset labeling was done by one

operator (MS) to avoid bias in labeling the dataset. This procedure ensures that each corneal image is assigned correctly to its refractive error type labeling.

The first step involves preprocessing the raw dataset D_{input} , which comprises corneal images I_k for $k = 1, \dots, N$. The corneal images I_k is normalized to a pixel range of $[0,1]$ to stabilize gradients during training and accelerate convergence (Krizhevsky et al., 2012), represented mathematically as described by the Equation 1 given below.

$$I_k(x,y) = \frac{I_k(x,y)}{255}, \forall (x,y) \in I_k \quad (1)$$

Typical input normalization was applied by rescaling the pixel values in $[0,255]$ to $[0,1]$, as shown in Equation 1. The normalization of corneal images I_k to a pixel range of $[0,1]$ is a standard and essential pre-processing technique in deep learning for medical image analysis. Normalization ensures consistency across images, minimizes the influence of varying imaging conditions, and enables the network to learn more robust and generalizable features by providing a consistent scale for all input pixel intensities. The images are then cropped to focus on the corneal region and resized to 72×72 pixels, ensuring that the model processes inputs of consistent size, balancing computational efficiency and the retention of significant features (Simonyan and Zisserman, 2014).

Data augmentation was performed for the significant hypermetropia class because the number of images is significantly low compared with other classes, which could affect the accuracy of the



FIGURE 3

(A) Corneal images with a specific pupillary red reflex marked certain types of refractive error. (B) Corneal image of significant myopia. (C) Corneal image of significant hypermetropia. Corneal image of an insignificant refractive error. (D) Corneal image showing the absence of pupillary red reflex.

model. Initially, the significant hypermetropia class had 237 images (12.3%); the significant myopia class had 586 images (30.9%); the insignificant refractive error class had 538 images (27.9%), and the not applicable to classified class had 555 images (28.9%). Data augmentation for the significant hypermetropia class was performed by resizing the original image (zoom-in at 20%), resulting in 460 images.

Data augmentation techniques are used to improve model generalization and address class imbalance. This augmentation enhances the dataset's diversity, reducing the risk of overfitting (Shorten and Khoshgoftaar, 2019). Data augmentation introduces synthetic variability into the dataset. Data augmentation techniques, including transformations T such as rotation ($0 \in [-20^\circ, 20^\circ]$), zooming ($z \in [0.8, 1.2]$), and vertical flipping, are applied to the images as described by the Equation 2.

$$T(I) = I_{rotated} \cup I_{zoomed} \cup I_{flipped} \quad (2)$$

where T augments the original image I into multiple augmented variants. This ensures that the model encounters a diverse set of inputs during training, reducing overfitting.

A total of 2,139 images were acquired after data augmentation, and the proportion of images between each class was quite similar; the significant hypermetropia class had 21.5% of total images; the significant myopia class had 27.4% of total images; the insignificant refractive error class had 25.2% of total images, and the not applicable to classified class had 25.9% of total images. The encoding for each class is set as follows: 0 for the not applicable to classified class, 1 for the significant hypermetropia class, 2 for the insignificant refractive error class, and 3 for the significant myopia class (Figure 4). Finally, the dataset is divided into training (70%), validation (15%), and test (15%) subsets following standard practices for model evaluation (Heaton et al., 2018), denoted as D_{train} , D_{val} , and D_{test} , respectively. This partitioning ensures that the model is trained and validated on distinct data while reserving unseen examples for unbiased evaluation. The final distribution of classes in each subset is balanced through augmentation, which is critical for mitigating biases during model training.

Model architecture

The model is designed as a multi-branch CNN with three branches. The multi-branch architecture enhanced the model's capability to process features at multiple resolutions. Each branch independently captured spatial patterns, and their combined outputs improved feature extraction, particularly for classes with subtle distinctions, such as insignificant refractive errors.

The CNN model M_{CNN} is initialized to extract the hierarchical features critical for refractive error classification. The architecture consists of convolutional layers (Conv2D) followed by ReLU activations and MaxPooling layers. The kernel size taken has been 3×3 , uniformly with an input image size of 256×256 and batch size of 32. Convolutional layers learn spatial filters to extract low-level features like edges, and progress to high-level features (LeCun et al., 2015). Batch normalization improves convergence and mitigates internal covariate shifts (Ioffe and Szegedy, 2015). The convolutional layer described here uses a ReLU (i.e., Rectified linear unit (RU)) activation function. The equation for the ReLU activation function is described in Equation 3.

$$ReLU = 0, \text{ for } x < 0 \quad (3)$$

$$ReLU = x, \text{ for } x \geq 0.$$

ReLU has been used to introduce nonlinearity in the form of half-rectified (from the bottom). ReLU gives a zero output for any vector value that is less than 0, while it behaves as an identity function for any vector value that is equal to 0 or greater than 0. It and its derivatives are both monotonic functions. Nevertheless, the fact that all negative values become 0 reduces the functionality of the models that train from data. Hence, it has been used in conjunction with other activation functions. The output layer uses four neurons because four classes can be predicted using the softmax function. The equation for the softmax activation function in the output layer computes class probabilities for multi-class classification and is mathematically described in Equation 4.

$$P(y = k | x) = \frac{\exp(z_k)}{\sum_{j=1}^C \exp(z_j)} \quad (4)$$

where z_k is the output for class k and C is the number of classes.

We use adam as the optimizer and sparse_categorical_crossentropy as the loss function to measure the prediction error because we encoded our dataset using 'tf.keras.utils.image_dataset_from_directory', which returns an integer label instead of one hot encoding (Kingma and Ba, 2014). We also want to use accuracy as the metrics. The Adam optimizer with an initial learning rate ($\eta = 0.001$) provided stable convergence and accelerated training. Training consisted of 10 epochs which is extensible to 100 epochs, and EarlyStopping callback with patience = 20 was used to monitor validation accuracy. EarlyStopping is used to restore the best weights when validation accuracy does not improve after 20 epochs and the training is stopped. The use of EarlyStopping ensured that the model converged efficiently, preventing overfitting. By restoring the best weights after 20 epochs of non-improvement, the model maintained a balance between training

```

Class name: Hasil pemeriksaan tidak dapat dinilai, Integer label: 0
Class name: Hipermetropia signifikan, Integer label: 1
Class name: Kelainan refraksi tidak signifikan, Integer label: 2
Class name: Miopia signifikan, Integer label: 3

```

```
data_iterator = data.as_numpy_iterator()
```

```
batch = data_iterator.next()
```

```
len(batch)
```

```
2
```

```

fig, ax = plt.subplots(ncols=32, figsize=(20,20))
for idx, img in enumerate(batch[0]):
    ax[idx].imshow(img.astype(int))
    ax[idx].title.set_text(batch[1][idx])

```



FIGURE 4

Encoding of corneal images according to their class.

and validation performance. We experimented with learning rates and optimizers (we chose Adam due to stable convergence) and dropout rates (settling on 0.5 to avoid overfitting). We experimented with different layer configurations, finally deciding on three convolutional layers with 32, 64, and 128 filters with MaxPooling, providing the optimal balance between model capacity and generalization.

Model training

The CNN is trained over E epochs using batch size B . Early stopping is applied to monitor validation loss and halt training when no improvement is observed for P consecutive epochs (Prechelt, 2000). This prevents overfitting and ensures efficient training. The model's weights are updated iteratively using the Adam optimization algorithm, which adapts the learning rate based on gradient updates as described by the Equation 5:

$$W_{t+1} = W_t - \eta \cdot \nabla WL \quad (5)$$

where L is the loss function. Training stability and convergence are monitored using accuracy and loss curves for training and validation datasets (Chollet, 2017).

Model evaluation

Model evaluation is performed on the test set D_{test} using standard metrics. The confusion matrix C is computed as described by the Equation 6:

$$C_{i,j} = \sum_{k=1}^N \mathbb{1}[\hat{y}_k = i \wedge y_k = j] \quad (6)$$

where $\mathbb{1}$ is the indicator function. Precision, recall, and F1-scores are derived from this to quantify the model's performance for each class (Powers and Ailab, 2011). Additionally, the ROC-AUC score evaluates the model's discriminatory power in a multi-class setting, defined as described by the Equation 7:

$$AUC = \int_0^1 TPR(FPR) d(FPR) \quad (7)$$

where TPR is the true positive rate and FPR is the false positive rate (Bradley, 1997).

Model explainability

Grad-CAM (Gradient-weighted Class Activation Mapping) is applied to enhance model interpretability by visualizing regions in the input image that contribute most to predictions. Grad-CAM heatmaps illustrate the corneal image regions that contribute most significantly to the predictions. For instance, the heatmaps for significant myopia consistently highlighted crescent-shaped red reflex patterns on the right side of the pupil, aligning with the ground truth. This level of interpretability enhances the model's acceptance in clinical settings, as it provides insights into the decision-making process. Grad-CAM computes the gradients $\frac{\partial S_c}{\partial A_{i,j,c}}$ of class scores S_c with respect to the last convolutional feature map $A_{i,j,c}$, producing class-specific heatmaps as described by the Equation 8 given below:

$$H_{i,j} = \text{ReLU}(c \sum_w w_c A_{i,j,c}), w_c = \frac{1}{Z} \sum_i \frac{\partial S_c}{\partial A_{i,j,c}}$$

$$H_{i,j} = \text{ReLU} \sum_c w_c A_{i,j,c}$$

Where

$$w_c = \frac{1}{Z} \sum_{i,j} \frac{\partial S_c}{\partial A_{i,j,c}} \quad (8)$$

These visualizations help clinicians interpret the model's focus, improving trust, and facilitating adoption in real-world ophthalmic diagnostics (Rs et al., 2020).

Model deployment

The final CNN model M_{CNN} is converted to TensorFlow Lite format M_{TFLITE} , optimizing it for mobile deployment. This lightweight version retains accuracy and enables real-time inference on smartphones. Techniques like quantization further reduce the model size and computational requirements (Abadi et al., 2016). The optimized M_{TFLITE} is integrated into smartphone applications to provide a scalable, low-cost solution for vision screening in

TABLE 1 Pseudocode of corneal image detection.

Step	Description
1. Data preprocessing	Normalize all images I_k such that pixel values fall within $[0,1][0, 1][0,1]$.
	Crop corneal regions from the images and resize to 72×72 pixels.
	Perform data augmentation: rotation ($\theta \in [-20^\circ, 20^\circ]$), flipping, zooming ($z \in [0.8, 1.2]$).
	Split the dataset into training (70%), validation (15%), and test (15%) subsets.
2. Initialize the CNN model	Design architecture with sequential layers:
	$Conv2D \rightarrow ReLU \rightarrow BatchNorm \rightarrow MaxPool \rightarrow Flatten \rightarrow Dens$.
	Use softmax activation in the output layer for multi-class classification:
	Compile with Adam optimizer ($\eta = 0.001$), categorical cross-entropy loss, and accuracy metric.
3. Training	Train the model for E epochs using batch size B .
	Apply early stopping to monitor validation loss with patience P :
	Stop if $L_{val}^{t+1} > L_{val}^t$ for P epochs.
	Save the best-performing model weights.
4. Evaluation	The trained model is evaluated on the test set D_{test} .
	Compute confusion matrix C :
	$C_{i,j} = \sum_k 1[y^k = i \wedge y_k = j]$
	Generate the classification report (Precision, Recall, F1-score):
	$Precision_i = \frac{C_{i,i}}{\sum_j C_{j,i}}$
	$Recall_j = \frac{C_{i,i}}{\sum_j C_{i,j}}$
	$F1 = \frac{2 \cdot Precision_i \times Recall_j}{Precision_i + Recall_j}$
	Generate the ROC-AUC for multi-class classification:
	$AUC = \int_0^1 TPR(FPR) d(FPR)$
5. Explainability	Apply Grad-CAM to visualize important regions in input images that contribute to prediction.
	Compute the Grad-CAM heatmaps using gradients G of class scores:
	$H_{i,j} = ReLU \sum_C w_C A_{i,j,C}$
6. Deployment	Convert the trained model M_{CNN} to TensorFlow Lite format as follows:
	$M_{TFLITE} \leftarrow Convert(M_{CNN})$
	Optimize M_{TFLITE} for mobile deployment and integrate it into smartphone applications.

resource-limited regions. Table 1 summarizes the pseudocode of each step in CNN model development.

Results and discussion

The CNN model demonstrated strong performance across all evaluation metrics, achieving an overall testing accuracy of 91%. The inclusion of three convolutional layers with ReLU activation and MaxPooling, followed by fully connected dense layers, allowed

the network to extract hierarchical features relevant to the classification task. Categorical cross-entropy was used as the loss function to ensure proper gradient propagation for multi-class classification. The model shows excellent results in the testing phase. The model can predict most images according to the type of refractive error, as shown in the confusion matrix (Figure 5, Left). The model also effectively distinguishes one class from another and has a perfect score (AUC score: 0.9896). This excellent score proves that this model could correctly predict the type of refractive error based on corneal images captured by a smartphone camera. The

performance metric of this model shows high scores in all classification metrics.

The confusion matrix revealed that the model performed particularly well in predicting each class (Figure 5, Left). The model accuracy is 91% (95% CI: 87.9–94.1%), which means that the model can correctly predict 91% of all predictions of the refractive error type. The overall precision of the model was 92% (95% CI: 89.0–95.0%), and the precision of the significant myopia class was the highest 96%. This means that the model can correctly predict 96% of 89 corneal images in the testing dataset with the crescent red reflex pattern dominant on the right side of the pupil as significant myopia, indicating that the majority of the predictions for this class were correct.

The overall recall of the model is 92% (95% CI: 87.9–94.1%), and the recall of the not applicable to classified class is the highest 98%, signifying the model's strong ability to identify images where the pupillary red reflex was absent or could not be classified. This means that the model can correctly predict 98% of the 84 corneal images in the testing dataset with an absence of pupillary red reflex as not applicable to classification (Figure 5). The high recall scores achieved by the 3-branch model, for the "Not Applicable to Classified" category, underscore its reliability in identifying cases where refractive errors cannot be determined. This is of critical importance in clinical settings because failure to detect such cases could result in delayed or improper interventions. Moreover, the model's high precision (96%) for "Significant Myopia" and "Significant Hypermetropia" ensures that most positive predictions correspond to true positives, reducing unnecessary follow-up diagnostics.

The model also has a high F1 score of 91% (95% CI: 87.9–94.1%). This means that the model performs well in labeling a patient with refractive error or not. If the model labels a patient with significant myopia, there is a high chance that the patient truly has significant myopia. If the model labels a patient as not applicable to classification, there is a high chance that the patient does not have a refractive error or has another eye disease that impedes the pupillary red reflex.

Figure 5 highlights the confusion matrix, the Receiver Operating Characteristic (ROC) curve, and detailed classification report. The confusion matrix shows the true vs. predicted classifications for the four classes as follows: "Not Applicable to Classified," "Significant Hypermetropia," "Insignificant Refractive Error," and "Significant Myopia." The results reveal strong classification performance, with particularly high recall (98%) for the "Not Applicable to Classified" category, indicating the model's ability to accurately identify the absence of refractive error or any patterns that preclude classification. The classification report below confirms this trend with high precision (96%) and F1-scores (0.92) for "Significant Myopia" and balanced metrics across all categories. The overall accuracy of 91% and weighted precision of 92% further emphasize the reliability of the 3-branch CNN architecture. The results argue in favor of the 3-branch model's capacity to generalize across varying image complexities. The balanced recall and precision scores ensure that the model avoids bias toward any specific class, a challenge often observed in smaller or imbalanced datasets.

The ROC curve in Figure 5, Right, demonstrates near-perfect discrimination, with an AUC score of 0.9896 (95% CI: 0.976–0.996), indicating that the model can effectively and confidently separate positive and negative classes for each category with high confidence. This is critical in medical applications where the cost of false negatives can outweigh that of false positives. The ROC curve shows a

one-vs-rest multi-class classification in our model. "Class 1" refers to the significant hypermetropia class. The ROC illustrates the model's ability to differentiate cases of "significant hypermetropia" (Class 1) from all other conditions (i.e., not applicable, insignificant refractive error, and significant myopia). Class 1 (significant hypermetropia) was treated as the "positive" class, whereas all other classes combined were treated as the "negative" class. The reported ROC area under the curve (AUC) score of 0.9896 quantifies this discriminatory power for Class 1.

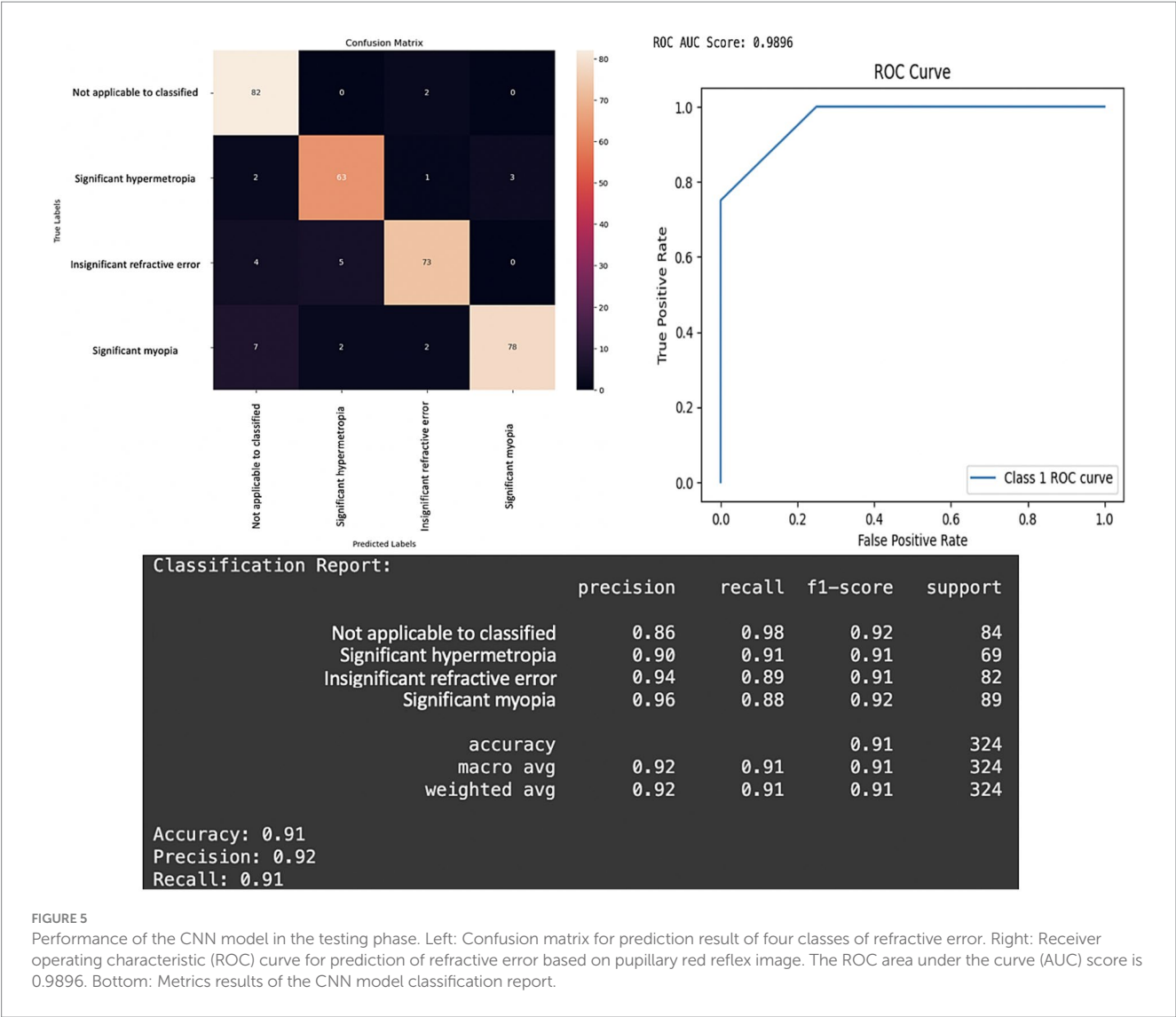
The appearance of only three distinct segments (or roughly four distinct points including the origin and (1,1)) in the ROC curve, giving it a very pronounced "staircase" or "step-like" appearance rather than a smooth curve, arises when the classifier's output probabilities (or confidence scores) for the "positive" class are highly discrete or coarsely quantized, rather than spanning a continuous range. The most probable reason for such a sharply segmented curve is that the model, when predicting the probability of a sample belonging to "Class 1," generates only a very small number of unique confidence scores across the test set. The test set for "Class 1" is relatively small; thus, the number of distinct prediction scores for these samples will naturally be limited, leading to fewer distinct (True Positive Rate, False Positive Rate) points and a more segmented curve. A smooth ROC curve is often expected and implies finely-graded confidence scores. However, a segmented curve does not indicate a fundamental flaw in the model's performance. This reflects the distinct nature of the model's output probabilities for the given test set. The area under the curve (AUC) remains the gold standard metric because this model shows a perfect AUC score and excellent discriminatory power between classes.

The 1-branch CNN architecture is the simplest model in this study, designed to sequentially extract features from the input image (Figure 6). It comprises a single pipeline in which convolutional layers detect spatial patterns in the corneal images. The network's structure is straightforward, starting with a convolutional layer followed by max-pooling. The initial layers capture low-level features, such as edges and textures, which are essential for identifying general patterns in the red pupillary reflex. Higher-level features, such as the symmetry or orientation of the reflex, are captured as the input progresses through additional convolutional and pooling layers.

The simplicity of the 1-branch CNN ensures computational efficiency, as reflected in its relatively small parameter count (approximately 11 million). However, this simplicity comes at a cost: the model is unable to capture the multi-scale features that are often necessary for distinguishing subtle differences between refractive error classes. For instance, overlapping patterns between "Insignificant Refractive Error" and "Significant Hypermetropia" may remain indistinguishable due to the lack of diversity in the feature extraction process.

Although the model demonstrated stable training and validation curves, as shown in Figures 7A,D, its overall performance metrics lagged behind the more complex architectures. The inability to capture diverse and hierarchical features limited its precision and recall, particularly in complex cases. Thus, while the 1-branch model provides a baseline for comparison, its architectural simplicity limits its effectiveness.

The 2-branch CNN introduces additional complexity by splitting the input data into two parallel pipelines (Figure 8). Each branch independently processes the input, capturing distinct features at different spatial scales. This design enables the network to learn



complementary feature representations, which are later combined through a concatenation layer. The branches leverage deeper convolutional and pooling layers, allowing them to detect both localized patterns (e.g., crescent-shaped reflexes) and broader contextual features (e.g., reflex symmetry).

However, the intermediate complexity of the 2-branch model presents challenges. The model struggles to balance feature extraction capacity with generalization, as reflected in its unstable validation accuracy and loss curves (Figures 7B,E). The lack of a third branch to diversify feature learning pathways results in redundant or insufficient feature representations. The validation accuracy of the 2-branch model varied, which indicated the difficulty in trading feature diversity and parameter efficiency. Problems with the two-branch model are due to a lack of diversity in pathways. It could not adequately separate the feature space for our four-class problem and caused errors in distinguishing between classes, leading to inconsistent validation performance. This limitation is particularly evident in the confusion matrix, where the model misclassifies overlapping categories such as “Significant Hypermetropia” and “Insignificant Refractive Error.”

The 2-branch architecture also exhibits a higher tendency to overfit compared to the 1-branch model. Although its parameter

count ($\sim 3.8 \times 10^6 \sim 3.8$ million) is significantly lower than that of the 1-branch CNN, the model’s instability suggests that additional pathways or regularization strategies are necessary to fully realize its potential. Thus, while the 2-branch model demonstrates improvement over the 1-branch model in theory, its practical performance falls short due to architectural limitations. Our 3-branch model overcomes this by concatenating at different scales of features, and a steady validation accuracy of 98.66% and test accuracy of 91% are achieved in this way. The suboptimal performance of the two-branch model highlights the challenges associated with intermediate architectural designs. Although the two-branch model incorporates additional complexity compared to the one-branch model, its limited capacity to diversify feature extraction results in overfitting and unstable validation performance. This limitation is evident in its fluctuating validation loss and low classification metrics. The inability to consistently separate classes with overlapping characteristics, such as “Significant Hypermetropia” and “Insignificant Refractive Error,” further underscores the need for more sophisticated architectural strategies.

The 3-branch CNN represents the most advanced architecture in this study, leveraging three independent feature extraction pipelines

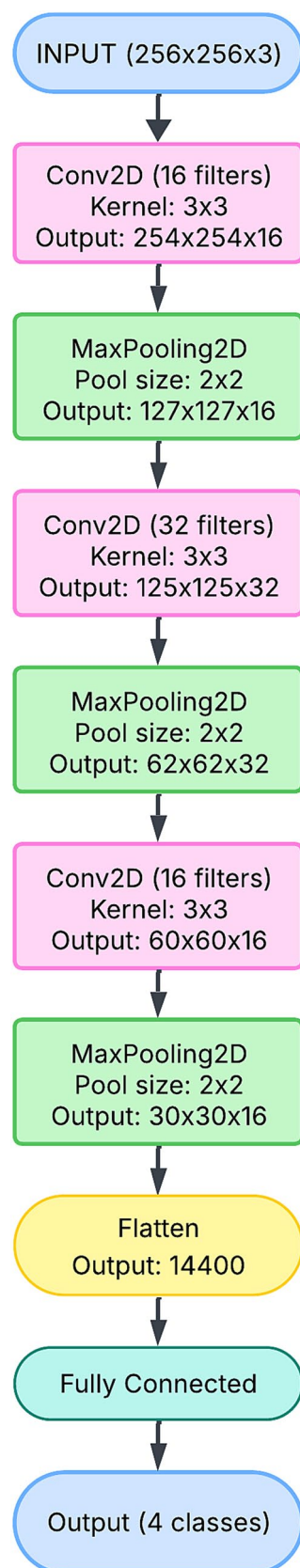


FIGURE 6
One-branch CNN model diagram chart.

(Figure 9). Each branch processes the input image at different resolutions, capturing multi-scale features that are critical for accurately classifying refractive errors. For example, one branch may focus on fine-grained details, such as the edges of the red reflex, whereas another may capture broader patterns, such as symmetry and orientation. The outputs of all three branches are concatenated and passed through dense layers to obtain a comprehensive feature representation.

This architecture demonstrates superior performance, as evidenced by its smooth training and validation accuracy curves (Figure 7C) and consistently low loss (Figure 7F). The 3-branch CNN achieves the highest precision (96%) and recall (98%) for critical classes such as “Significant Myopia,” reflecting its ability to handle complex and overlapping patterns. The high area under the curve (AUC) score of 0.9896 further underscores its robust classification capabilities.

Furthermore, the multi-branch design mitigates overfitting by ensuring that each branch learns unique, complementary features. This diversity enhances generalization, making the model more reliable across unseen data. Despite its higher parameter count ($\sim 274,000$), the 3-branch CNN achieves a balance between complexity and performance, solidifying its role as the optimal architecture for this application. Its success highlights the importance of multi-scale feature extraction in deep-learning models for medical imaging.

Figure 7 shows the accuracy and loss of each CNN model used in this study. The one-branch model ran 20 epochs with a batch size of 32, achieving a training accuracy of 0.9970 and a training loss of 0.0146. The validation accuracy of the one-branch model is 0.9760, and its validation loss is 0.0800. The two-branch model ran 100 epochs with a batch size of 32 and used EarlyStopping, which means that the training is stopped when validation accuracy is not improved after 20 epochs. The two-branch model achieved a training accuracy of 0.9742 and a training loss of 0.0738. The validation accuracy of the two-branch model is 0.9688, and its validation loss is 0.1137. These results show that the performance of the two-branch model is lower than that of the one-branch model. However, the accuracy is improved and the loss is decreased with the three-branch model. The three-branch model ran 100 epochs with a batch size of 32 and used EarlyStopping. The three-branch model achieved training accuracy of 0.9992 and training loss of 0.0035. The validation accuracy of the three-branch model is 0.9866, and its validation loss is 0.0762. Thus, this study shows that the three-branch model achieves the optimum performance.

Architectural strengths of the 3-branch model

The 3-branch CNN model exhibits significant architectural advantages, leveraging its multi-pathway design to process input images at varying resolutions. This approach ensures simultaneous capture of both fine-grained features (e.g., subtle differences in red reflex patterns) and broader spatial structures (e.g., pupil orientation). The superior performance of the 3-branch model, achieving a validation accuracy of 98.66% and a test accuracy of 91%, highlights its robustness in generalizing across complex, multi-class refractive error datasets. The model mitigates information loss by integrating diverse feature representations through its branches, ensuring that

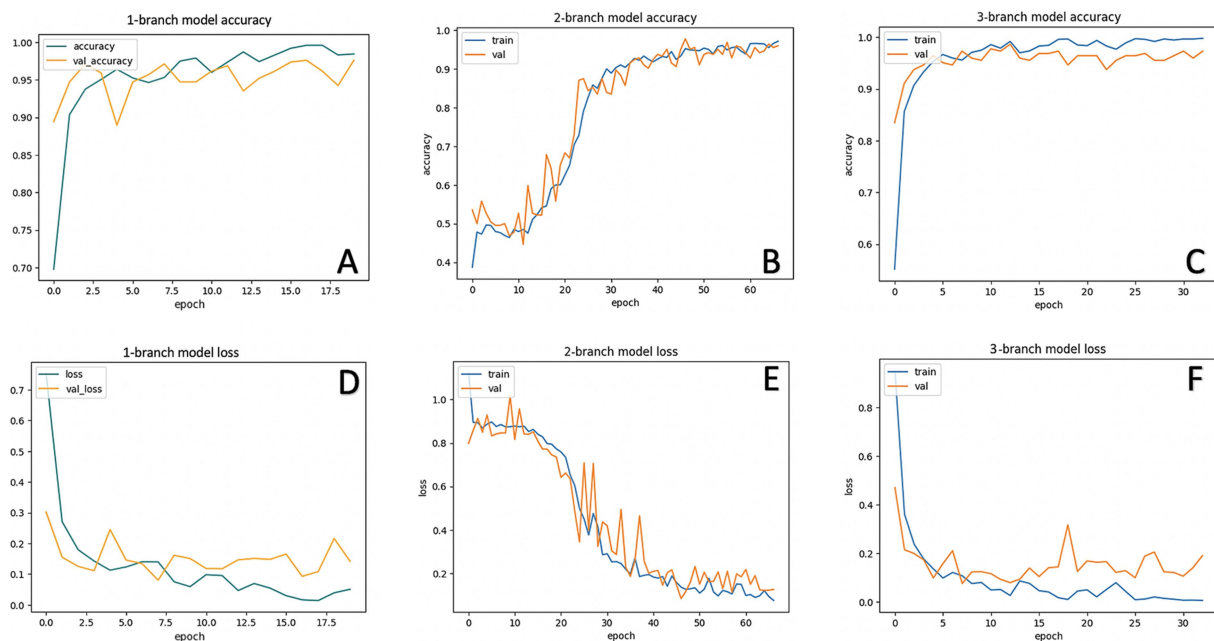


FIGURE 7

Accuracy and loss in each model. Training and validation accuracy of one-branch (A) two-branch (B) and three-branch (C) CNN models. Training and validation loss of one-branch (D), two-branch (E), and three-branch (F) CNN models.

patterns unique to each class are effectively learned. This is particularly crucial for medical applications where misclassifications can have significant implications.

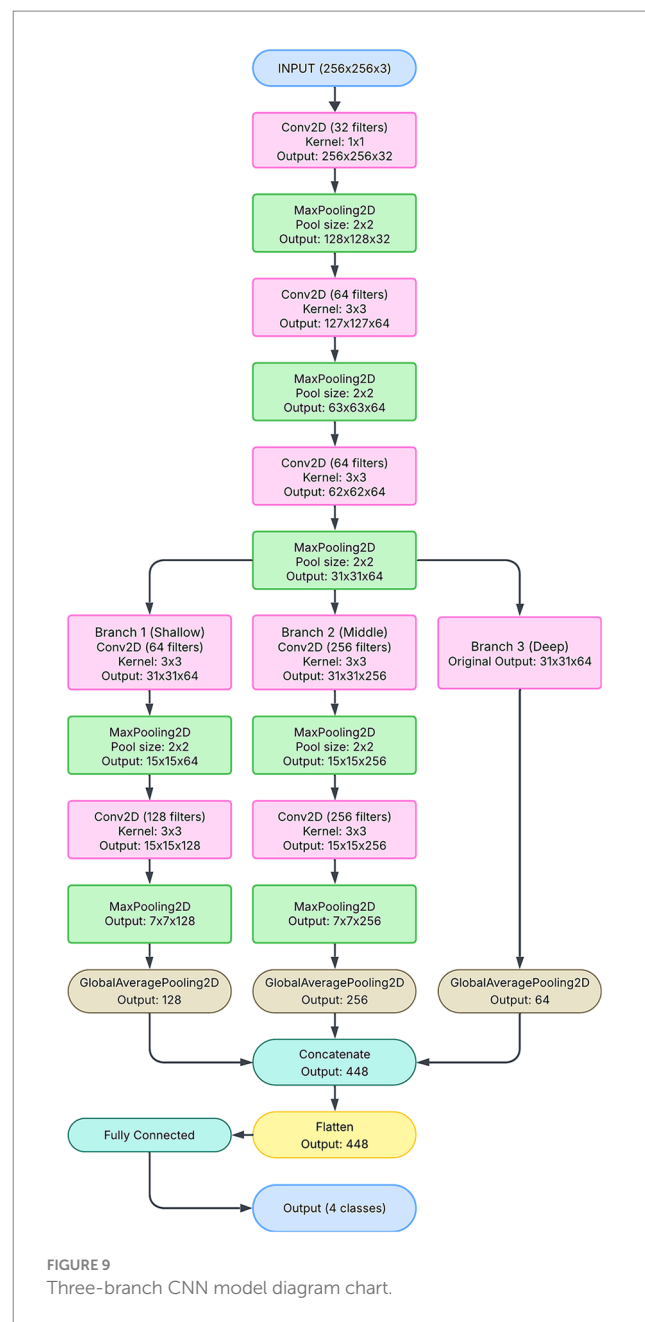
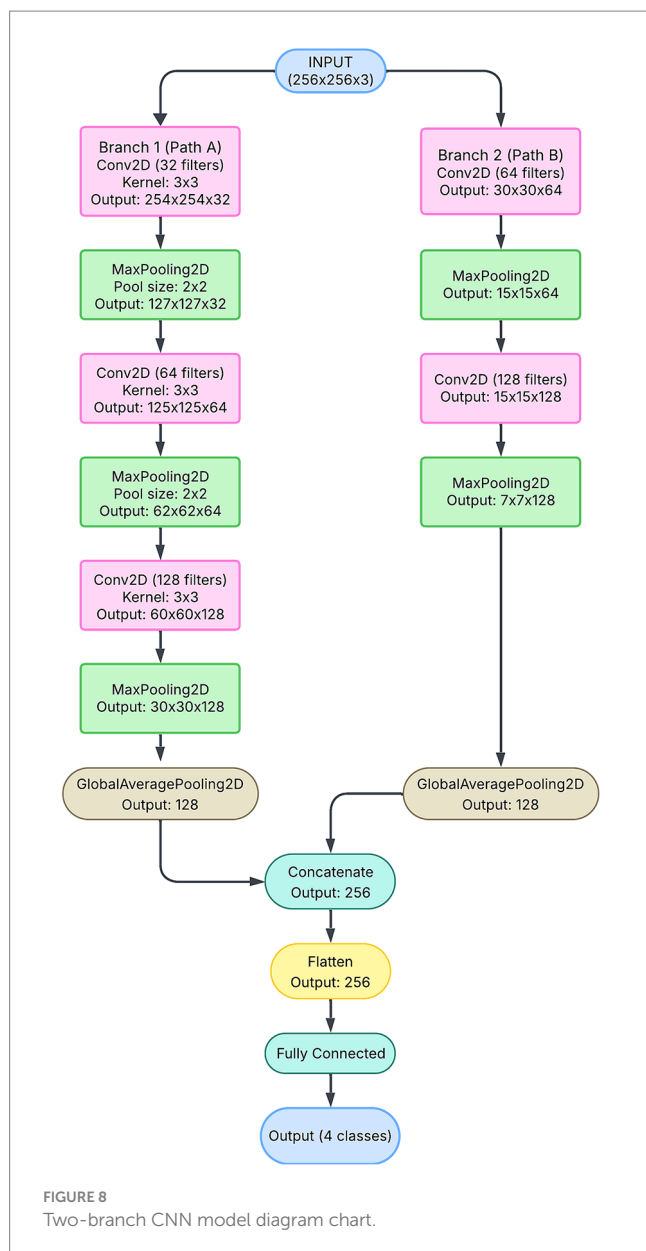
Additionally, the smooth convergence of training and validation loss in the 3-branch model, as shown in Figure 7F, underscores its stability and resistance to overfitting. Unlike the one-branch model, which underfits due to limited feature extraction capacity, and the two-branch model, which suffers from unstable performance, the 3-branch model strikes an optimal balance between architectural complexity and generalization. These findings solidify the importance of multi-branch architectures in high-stakes classification tasks. The image in Figure 10 provides an excellent conceptual diagram of a multi-branch CNN architecture designed for analyzing pupillary red reflex patterns for refractive error classification. The proposed model leverages three distinct branches, each specialized in extracting different levels of features from the input image, resulting in a more robust and comprehensive analysis.

The architecture consists of an initial input image stage, followed by three parallel processing branches (Branch 1, Branch 2, Branch 3) that are then fed into a Feature Fusion stage, culminating in the Classification Output. Branch 1 captures the fundamental, primitive visual characteristics in the red reflex image and focuses on low-level features such as edge detection and basic geometry. For red reflex analysis, this branch is crucial for delineating the pupil boundaries, identifying the edges of any crescent patterns, and potentially localizing the red reflex. It provides foundational structural information. The associated colored blocks (i.e., red, orange, gray) would represent conceptual activations. High activation (red) occurs when strong edges or basic geometric forms are detected. Branch 2 focuses on mid-level features, such as spatial relationships and orientation. This branch helps identify patterns indicative of specific refractive errors, such as the orientation of a crescent in the red reflex.

Branch 3 focuses on high-level features such as contextual patterns and symmetry. This branch integrates information from a wider range of receptive fields to better understand overarching visual themes. This branch is vital for distinguishing subtle differences between types of refractive error that may not be obvious from isolated edges or local orientations. Recognizing the gestalt of the red reflex, such as a “full pupil coverage” for insignificant refractive error or the distinct “crescent patterns” for myopia/hypermetropia, is helpful.

After each branch has extracted its specialized features, their outputs are combined in the Feature Fusion stage, which is achieved through concatenate and fully connected dense layers. The final output of the model is the classification output, which indicates one of 4 classes. The model achieves a more comprehensive understanding of the pupillary red reflex by fusing the insights from all three branches. This integrated approach allows us to make more accurate and nuanced refractive error classifications, leveraging the strengths of each specialized feature extractor.

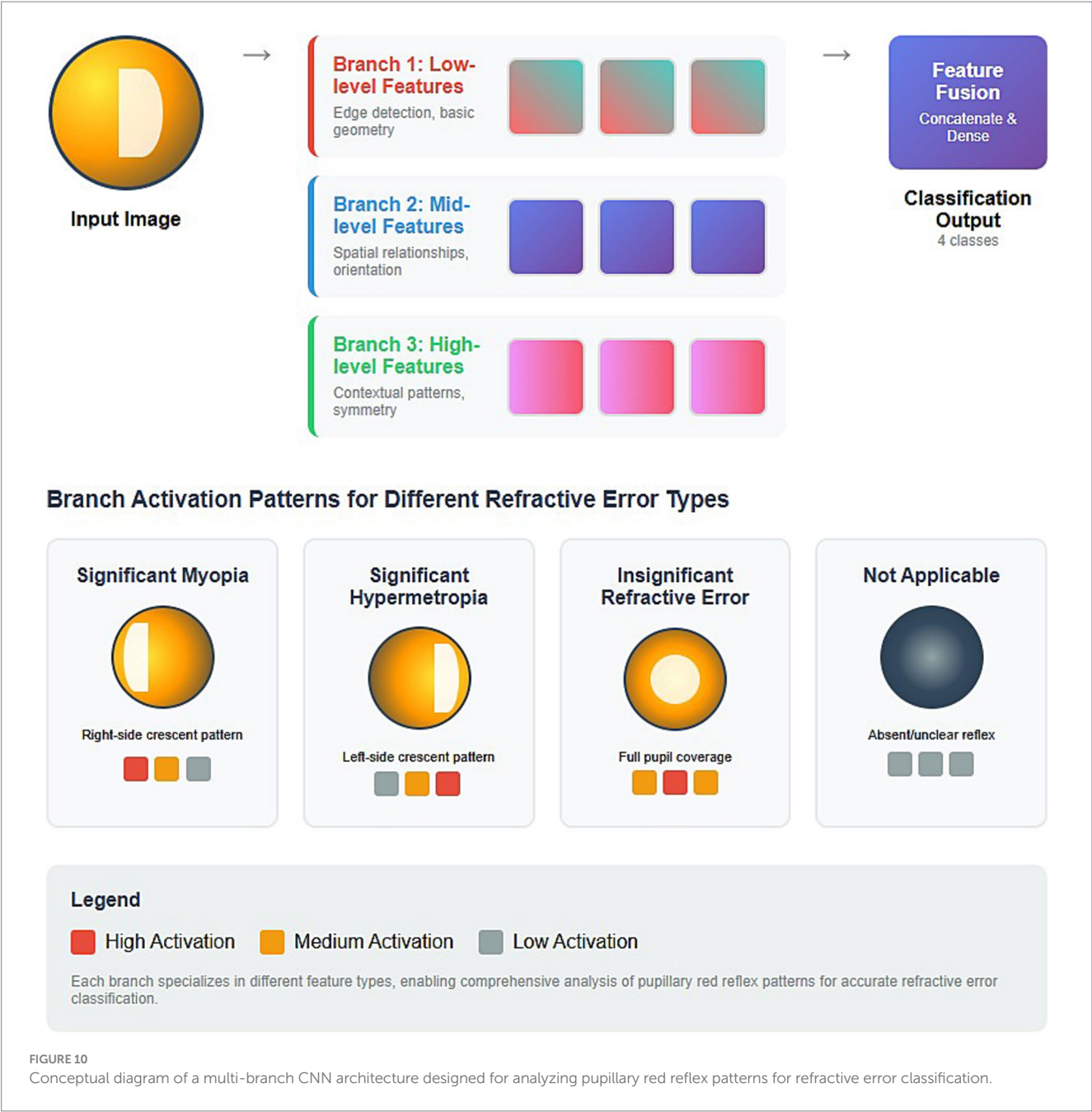
Grad-CAM is a crucial tool for model interpretability and to understand the features driving our multi-branch CNN predictions. Visualizing regions of importance is paramount for a medical application like refractive error detection, because it provides transparency and builds trust in the model's decision-making process. The Grad-CAM heatmaps illustrate the regions of an image that are most important for the model's classification of refractive errors. These illustrations provide an understanding of why a deep learning model makes a certain refractive error prediction. These Grad-CAM visualizations are essential because they validate model behavior and visually confirm that the multi-branch CNN focuses on the diagnostically relevant features (crescent position, full reflex) that are traditionally used by clinicians in photorefractive, rather than spurious correlations. The Grad-CAM heatmaps show red and yellow areas, which indicate regions of high activation that strongly influenced the model's decision for the predicted



class. These areas are the “hotspots” where the CNN is focused. The Grad-CAM heatmaps also show blue and green areas, which indicate regions of low or no activation, indicating that these areas were less influential for refractive error class prediction.

The original image in Figure 11A shows a pupillary red reflex with a prominent crescent of light, primarily on the right side of the pupil, denoting a significant myopia class. The rest of the pupil appears relatively dark or has a very dim reflex. The Grad-CAM heatmap in Figure 11B displays a strong red/yellow activation, precisely concentrated over this right-sided crescent. This indicates that the model relies heavily on the presence, shape, and intensity of this right-sided crescent for its classification. The original image in Figure 11C features a distinct crescent of light, predominantly on the left side of the pupil, denoting a significant class of hypermetropia. Similar to Figure 3A, the remainder of the pupil is less illuminated. The Grad-CAM heatmap in Figure 11D exhibits a strong red/yellow activation, which is concentrated over the left-sided crescent. This indicates that the model’s

decision is strongly driven by the characteristics of this left-sided reflex. The heatmap shows that the model’s attention aligns with this key diagnostic feature. The original image in Figure 11E shows a “Total Red Reflex,” where the red reflex appears to fill nearly the entire pupillary area, indicating a more diffuse and uniform light reflection, denoting an insignificant refractive error class. The Grad-CAM heatmap in Figure 11F shows a broad red/orange activation spread across almost the entire pupillary area, closely matching the extent of the total red reflex. This demonstrates that the model uses the overall presence and distribution of the red reflex across the pupil rather than focusing on a specific crescent. The original image in Figure 11G shows the pupil area with a dark or absent/unclear red reflex, and some reflections or artifacts might be present, but no clear red glow fills the pupil or forms a distinct crescent. This image quality might fall into the “not applicable” or “unclassifiable” category. The Grad-CAM heatmap in Figure 11H shows



very scattered and relatively low-intensity (yellow/blue) activations, primarily around the pupil edges or over reflections/artifacts, with no strong, coherent red hotspot over the central pupillary area. This indicates that the model is struggling to find distinct, meaningful red reflex patterns or is focusing on subtle cues that indicate an unclassifiable image. The diffused and less intense heatmap indicates the absence of clear patterns for the other refractive error types in the model learned.

Comparative analysis of model performance

The comparative results across the one-branch, two-branch, and three-branch models emphasize the interplay between model

complexity and generalization. The one-branch model although stable (Figures 7A,D), demonstrates lower accuracy and higher loss due to its inability to learn intricate patterns. In contrast, the two-branch model shows significant instability in validation performance (Figures 7B,E), with fluctuating validation accuracy and failure to converge. This instability indicates that the intermediate complexity of the two-branch model lacks the structural diversity required to effectively separate overlapping classes, such as “Insignificant Refractive Error” and “Significant Hypermetropia.”

Theoretically, a standard monolithic architecture like ResNet can be adapted for multi-class classification. The decision to employ a multi-branch architecture was deliberate and was fundamentally driven by the inherent complexity and multi-faceted nature of ophthalmic image analysis for refractive error detection, especially in

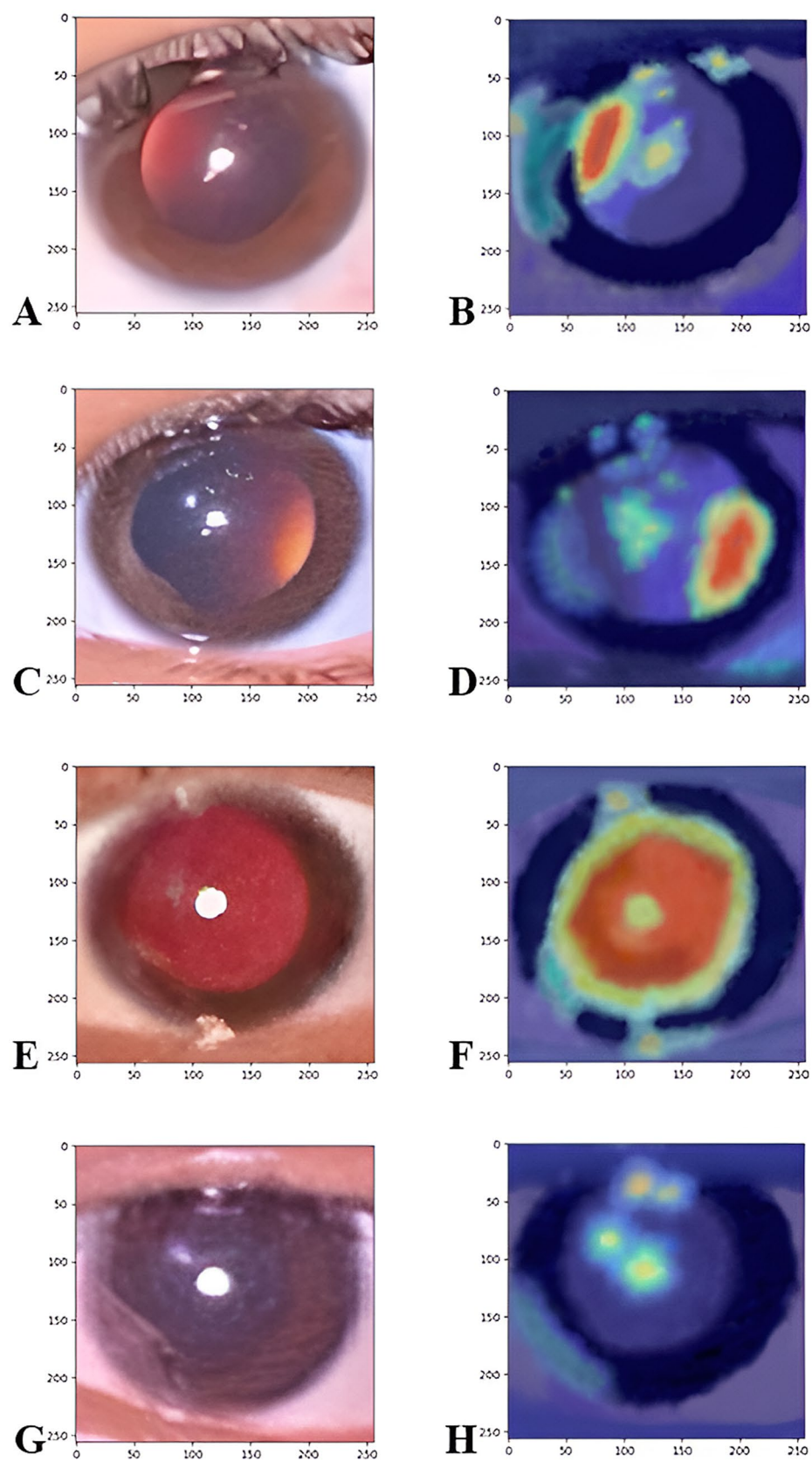


FIGURE 11

The Grad-CAM heatmaps illustrate the corneal image regions that most significantly contributed to the prediction. Figures on the left side refer to the original images of the significant myopia class (A), the significant hypermetropia class (C), insignificant refractive error class (E), and the not applicable class (G). The figure on the right side refers to the heat map of the myopic class (B), the hypermetropic class (D), insignificant refractive error class (F), and not applicable classes (H), which captures important features for classifying refractive error, particularly the crescent region in the pupil.

a smartphone-based setting. This specific multi-branch CNN architecture was chosen after an initial design process in which standard architectures, including CNNs with a single pathway, were tested and offered insufficient performance in terms of learning for distinguishing subtle differences in pupillary red reflex patterns. A single-stream CNN may struggle to learn and fuse these disparate features simultaneously.

Although pre-trained models such as ResNet are retrained to serve as feature extractors, they are not designed with our specific classification task in mind. Refractive error diagnosis requires a holistic assessment of various visual cues, and smartphone-based imaging for diagnostic purposes is inherently prone to noise sources and artifacts (e.g., motion blur, uneven illumination, reflections). A monolithic network forces all features through the same convolutional layers, potentially leading to a suboptimal trade-off. For example, some features may be diluted, or the network may struggle to learn highly discriminative representations for all relevant cues.

Previous studies have developed CNN models for detecting refractive errors in pupillary red reflex images. However, the accuracy of those models is lower than that of our model. A study by [Kriangsakdachai et al. \(2022\)](#) developed a pre-trained CNN model, DenseNet and EfficientNet, for the detection of abnormal red reflex due to refractive error. The model performance has an accuracy of 72.25%, sensitivity of 75.5%, and specificity of 0.69%. The model in this study only classified the class into normal and abnormal cases and did not classify the type of refractive error. The dataset was retrieved from the 322 images captured using an iPhone X or iPhone 7. Another study by [Linde et al. \(2023\)](#) developed a pretrained CNN model, Inception-v3 and EfficientNet, to differentiate myopic and non-myopic eyes. Images of 357 patients were captured with a non-IR fundus camera attached to an Android smartphone. The model achieved 75% accuracy but did not achieve the target as a primary screening tool. A Study in the Korean population by [Chun et al. \(2019\)](#) showed higher accuracy of the pretrained CNN model, ResNet-18, with an overall test accuracy of 81.6%. The model classifies six types of refractive errors based on the dioptric power interval. The dataset in this study is also quite small in size (305 images captured with LGM-X800K smartphone from 164 patients), and the age range of the patients involved is also narrow and limited to a young age (mean age 4.32 years, SD 1.87 years, range 6 months to 8 years). The 3-branch model developed in our study had superior precision (96%) and recall (98%) for critical classes such as “Significant Myopia” compared with the pretrained model, highlighting its ability to handle challenging scenarios where patterns overlap or are subtle. The confusion matrix ([Figure 5, Left](#)) further demonstrates the model’s ability to minimize misclassifications across all classes, solidifying its position as the optimal architecture for this application.

A multi-branch CNN allows dedicated branches to specialize in extracting relevant features from specific regions or visual information types in the input image. Each branch can be optimized using appropriate filter sizes, receptive fields, and feature maps to best capture the nuances of its designated input. The multi-branch approach specifically addresses the need to simultaneously capture features at different scales. This characteristic is particularly important for crescent-shaped red reflex patterns that mainly differ in orientation and position rather than in general appearance. The multi-branch

CNN design facilitates the fusion of these specialized feature representations at a later stage of the network (e.g., concatenation layer followed by fully connected layers). This allows the model to learn complex, non-linear relationships between these different types of information, leading to a more accurate and comprehensive diagnosis than relying on a single, undifferentiated feature set.

Limitations

Despite its strong performance, the proposed model has certain limitations:

1. **Ethnic diversity in dataset:** The training dataset predominantly comprised images from Malay ethnicity, which may limit the model’s generalizability to other populations. Future studies should incorporate images from diverse ethnic groups to broaden the dataset and encompass a broader range of populations. This can be achieved through multi-center research initiatives. Additionally, to improve external validity, stratified sampling during model development and thorough subgroup analyses across different ethnic groups should be implemented during validation. These approaches will contribute to the generalizability of our results.
2. **Resolution constraints:** Although the 72×72 pixel resolution was computationally efficient, it may have led to the loss of finer details in the corneal images. Higher-resolution inputs could further improve the model performance, particularly for challenging classes. We tested the model performance using 224×224 pixel images, which are substantially higher than our original 72×72 resolution. This higher resolution can better capture subtle red reflex morphologies, contributing to our improved test accuracy of 92.91% (vs. the previous 91%). However, despite the improved performance gain, this led to increasing computational demands. The higher resolution also led to longer training times due to the limitations of our computational resources. Efficient training times with lower image resolution are preferred over a slight increase in test accuracy using higher image resolution at the cost of higher computational effort.
3. **Class imbalance:** Although the dataset was balanced by augmentation, real-world applications may still encounter skewed distributions. Dynamic re-weighting techniques during training could address this issue.
4. **Specific device usage:** The specific iPhone 6s used in this study may possess unique image settings, camera characteristics, and light source eccentricity that could influence algorithm performance. The iPhone 6s had a flashlight source position at the left side of the camera lens, with a light source eccentricity (distance of the light source from the edge of the camera aperture) of approximately 6 mm. The images in the dataset were taken in the portrait orientation; thus, the crescent-shaped red reflex was positioned in the horizontal meridian ([Figures 3A,B](#)). In smartphone types where the flashlight is positioned above or below the camera lens (Nokia Lumia 800 or Samsung Galaxy S3), if the image is taken in the portrait orientation, the

crescent-shaped red reflex is positioned in the vertical orientation (Figure 12) (Colicchia et al., 2015). The image should be taken in landscape orientation to produce the crescent-shaped red reflex, which is positioned in the horizontal meridian. Therefore, if the image collection procedure is standardized, the algorithm will continue to produce valid and reliable results and is expected to be more resilient to minor variations in image acquisition, regardless of the type of smartphone.

Recommendations and future directions

The findings emphasize that architectural complexity alone does not guarantee better performance. The integration of diverse and complementary feature extraction pathways, as seen in the 3-branch model, is critical for achieving robust classification in complex datasets. Future studies should focus on refining multi-branch designs to further enhance generalization and stability. Future directions should also focus on expanding the dataset to include images from diverse ethnicities, age groups, and lighting conditions to ensure the fairness of the model in a global population. The use of the model with higher-resolution input images could be beneficial because the influence of subtle variations in the pupillary red reflex is crucial for differentiating more complex classes. Deploying the model on smartphones through lightweight frameworks like TensorFlow Lite will provide a cost-effective tool for real-time predictions and could be a solution to a vision screening at the community level, as it will be much more feasible. Moreover, advanced explainability tools, such as Grad-CAM, should be introduced to generate visual representations of decision-making processes to enhance clinician and patient trust. Future studies may consider hybrid architectures that exploit the benefits of convolutional layers along with attention mechanisms, such as transformers that enhance the capacity of the model to classify overlapping or complex refractive error categories. These improvements will make the proposed solution scalable, reliable, and clinically usable. Finally, future plans to test the model in a real-world environment, particularly in school-aged children, are crucial. The choice for a pilot test in the school-aged children population is due to the high prevalence of uncorrected refractive errors globally, particularly among Indonesian children, which is approximately 12% (Mahayana et al., 2017; Halim et al., 2020). Validating algorithms using external test datasets and in practical environments with diverse demographics is crucial to ensure their generalizability and applicability in clinical practice before implementation.



FIGURE 12

In myopic eyes, the crescent-shaped red reflex is located in the inferior part of the pupil. The images were captured using a Samsung Galaxy S3 smartphone with the light source positioned below the camera lens (Colicchia et al., 2015). Reprinted from Colicchia et al. (2015), with the permission of AIP Publishing.

Conclusion

The integration of CNN models with smartphone-based image acquisition systems offers a scalable and cost-effective refractive error screening solution. The model's performance demonstrates its potential to replace or complement traditional methods that rely on specialized equipment and trained professionals. The model effectively bridges the gap between professional diagnostic tools and accessible, smartphone-based solutions by automating the classification of refractive errors. Therefore, this approach could significantly expand access to eye care in resource-limited settings. By harnessing smartphone technology, this methodology fills the void between the gold standard in diagnostic imaging and currently scalable and low-cost solutions for widespread vision screening.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by The Ethics Committee of the Faculty of Public Health Universitas Indonesia (Ket-541/UN2.F10.D11/PPM.00.02/2024). The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin. Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

Author contributions

MS: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. HP: Software, Methodology, Data curation, Investigation, Writing – review & editing, Conceptualization, Validation, Visualization, Supervision, Writing – original draft, Formal analysis. SH: Writing – review & editing, Conceptualization, Supervision, Validation. KS: Supervision, Writing – review & editing, Conceptualization, Validation. NM: Writing – review & editing, Supervision, Validation, Conceptualization.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This study received funding from the Lembaga Pengelola Dana Pendidikan (LPDP) through a research grant (No. PD2592024110600195).

Acknowledgments

The author would like to thank Lembaga Pengelola Dana Pendidikan (LPDP), Sabang Merauke Eye Center Hospital, and LPPM Universitas Andalas for supporting this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., and Dean, J. (2016). Tensor flow: a system for large-scale machine learning. Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation, 265–83. doi: 10.48550/arXiv.1605.08695
- Bradley, A. P. (1997). Use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn.* 30, 1145–1159. doi: 10.1016/S0031-3203(96)00142-2
- Braverman, R. (2007). Diagnosis and treatment of refractive errors in the pediatric population. *Curr. Opin. Ophthalmol.* 18, 379–383. doi: 10.1097/ICU.0b013e328244dfed
- Cao, H., Cao, X., Cao, Z., Zhang, L., Han, Y., and Guo, C. (2022). The prevalence and causes of pediatric uncorrected refractive error: pooled data from population studies for global burden of disease (GBD) sub-regions. *PLoS One* 17:e0268800. doi: 10.1371/journal.pone.0268800
- Chollet, F. (2017). Deep learning with Python. New York: Manning Publications.
- Chun, J., Kim, Y., Shin, K., Han, S., Oh, S., Chung, T.-Y., et al. (2019). A deep-learning-based prediction of refractive error using smartphone-captured photorefractive images: model development and validation study. *JMIR Med. Inform.* 8, e16225. doi: 10.2196/16225
- Cochrane, G., du Toit, R., and Mesurier, R. (2010). The management of refractive errors. *BMJ* 340:c1711. doi: 10.1136/bmj.c1711
- Colicchia, G., Wiesner, H., and Zollman, D. (2015). Photorefractive of the eye. *Phys. Teach.* 53, 103–105. doi: 10.1119/1.4905810
- Cooper, C. D., Bowling, F. G., Hall, J. E., Colville, D. J., Dortmans, R. J., Munch, J., et al. (1996). Photoscreener instrument evaluation in a childhood population. 1. The Otago photoscreener and Dortmans videophotorefractor. *Aust. N. Z. J. Ophthalmol.* 24, 347–355. doi: 10.1111/j.1442-9071.1996.tb01606.x
- Do, C.-w., Chan, L. Y., Kwok, O.-l., Lai, L. H., Lee, P. H., Ngai, G., et al. (2022). Using a smartphone camera to detect refractive error among schoolchildren: a simple model to address cost-related barriers for refractive error detection. *Invest. Ophthalmol. Vis. Sci.* 63:2819 – A0149.
- Donahue, S. P. (2009). “Screening for pediatric ophthalmologic disorders” in Pediatric ophthalmology: Current thought and a practical guide. eds. M. E. Wilson, R. A. Saunders and R. H. Trivedi (Leipzig, Germany: Springer-Verlag Berlin Heidelberg), 61–70.
- Fu, E., Yang, Z., Leong, H., Ngai, G., Do, C. W., and Chan, L. Exploiting active learning in novel refractive error detection with smartphones (2020). Proceedings of the 28th ACM international conference on multimedia. 2775–2783.
- GBD 2019 Blindness and Vision Impairment Collaborators (2021). Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the right to sight: an analysis for the global burden of disease study. *Lancet Glob. Health* 9, e144–e160. doi: 10.1016/S2214-109X(20)30489-7
- Halim, A., Suganda, R., Sirait, S. N., Memed, F. K., Syumarti Rini, M., et al. (2020). Prevalence and associated factors of uncorrected refractive errors among school children in suburban areas in Bandung, Indonesia. *Power. Med.* 7:1737354. doi: 10.1080/2331205X.2020.1737354
- Hashemi, H., Fotouhi, A., Yekta, A., Pakzad, R., Ostadimoghaddam, H., and Khabazkhoob, M. (2018). Global and regional estimates of refractive error prevalence: systematic review and meta-analysis. *J. Curr. Ophthalmol.* 30, 3–22. doi: 10.1016/j.joco.2017.08.009
- Heaton, J., Goodfellow, I., Bengio, Y., and Courville, A. (2018). Deep learning. Genetic Programming and Evolvable Machines. 19, 305–7. doi: 10.1007/s10710-017-9314-z
- Holden, B. A. (2007). Uncorrected refractive error: the major and most easily avoidable cause of vision loss. *Community Eye Health* 20, 37–39.
- Ik, Z. Q., Lau, S. L., and Chan, J. B., editors. Mobile cataract screening app using a smartphone. 2015 IEEE Conference on e-Learning, e-Management and e-Services (IC3e); (2015) 24–26.
- Ioffe, S., and Szegedy, C.. Batch normalization: accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd international conference on international conference on machine learning-volume 37; Lille, France: JMLR.org; (2015). p. 448–456.
- Keskinbora, K., and Güven, F. (2020). Artificial intelligence and ophthalmology. *Turk. J. Ophthalmol.* 50, 37–43. doi: 10.4274/tjo.galenos.2020.78989
- Kingma, D., and Ba, J.. Adam: a method for stochastic optimization. International conference on learning representations. (2014).
- Kiu, L. Y., Kwok, O. -l., Leung, H. -c., Chan, D. N., Tang, W. -c., Li, H. Y., et al. (2024). Enhancing myopia screening in the community: exploring the potential of smartphone photorefractive. *Invest. Ophthalmol. Vis. Sci.* 65:608.
- Kriangsakdachai, S., Ayudhya, S., Kusakunniran, W., Ayudhya, W., Chantarasagul, C., Manasboonpermpool, R., et al. Anomaly detection in red reflex images using deep learning approaches (2022). 1–6 p.
- Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* 25, 1097–1105. doi: 10.1145/3065386
- Latif, J., Xiao, C., Imran, A., and Tu, S. Medical imaging using machine learning and deep learning algorithms: a review. 2nd International conference on computing, mathematics and engineering technologies (iCoMET).pp. 1–5. IEEE. (2019)
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539
- Lim, J. S., Hong, M., Lam, W. S. T., Zhang, Z., Teo, Z. L., Liu, Y., et al. (2022). Novel technical and privacy-preserving technology for artificial intelligence in ophthalmology. *Curr. Opin. Ophthalmol.* 33, 174–187. doi: 10.1097/ICU.0000000000000846
- Linde, G., Chalakkal, R., Zhou, L., Huang, J. L., O’Keeffe, B., Shah, D., et al. (2023). Automatic refractive error estimation using deep learning-based analysis of red reflex images. *Diagnostics (Basel)* 13:2810. doi: 10.3390/diagnostics13172810
- Ma, S., Guan, Y., Yuan, Y., Tai, Y., and Wang, T. (2020). A one-step and streamlined children vision screening solution based on smartphone imaging for resource-limited areas: design and evaluation. *JMIR Mhealth Uhealth* 8:e18226. doi: 10.2196/18226
- Mahayana, I. T., Indrawati, S. G., and Pawioran, S. (2017). Prevalence of uncorrected refractive error in urban, suburban, exurban and rural primary school children in Indonesian population. *Int. J. Ophthalmol.* 10, 1771–1776. doi: 10.18240/ijo.2017.11.21
- Martinez-Perez, C., Alvarez-Peregrina, C., Villa-Collar, C., and Sánchez-Tena, M. (2022). Application of artificial intelligence to ophthalmology and optometry: a citation network analysis. *J. Optom.* 15, S82–S90. doi: 10.1016/j.optom.2022.06.005
- Ng, W. Y., Zhang, S., Wang, Z., Ong, C. J. T., Gunasekaran, D. V., Lim, G. Y. S., et al. (2021). Updates in deep learning research in ophthalmology. *Clin. Sci. (Lond.)* 135, 2357–2376. doi: 10.1042/CS20210207
- Pan, C.-W., Zheng, Y.-F., Anuar, A. R., Chew, M., Gazzard, G., Aung, T., et al. (2013). Prevalence of refractive errors in a multiethnic Asian population: the Singapore epidemiology of eye disease study. *Invest. Ophthalmol. Vis. Sci.* 54, 2590–2598. doi: 10.1167/iops.13-11725
- Panda, S., and Kumar Dhaka, R. (2022). Application of artificial intelligence in medical imaging. 195–202.
- Powers, D., and Ailab. (2011). Evaluation: from precision, recall, and F-measure to ROC, informedness, markedness, and correlation. *J. Mach. Learn. Technol.* 2, 2229–3981. doi: 10.48550/arXiv.2010.16061
- Prechelt, L. (2000). Early stopping—but when? *Lect. Notes Comput. Sci* 1524, 55–69. doi: 10.1007/3-540-49430-8_3

- Rahi, J. S., Solebo, A. L., and Cumberland, P. M. (2014). Uncorrected refractive error and education. *BMJ* 349:g5991. doi: 10.1136/bmj.g5991
- Resnikoff, S., Pascolini, D., Mariotti, S. P., and Pokharel, G. P. (2008). Global magnitude of visual impairment caused by uncorrected refractive errors in 2004. *Bull. World Health Organ.* 86, 63–70. doi: 10.2471/BLT.07.041210
- Rs, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* 618–626. doi: 10.1007/s11263-019-01228-7
- Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6:60. doi: 10.1186/s40537-019-0197-0
- Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.556 p [Preprint]. Available online at: <https://ui.adsabs.harvard.edu/abs/2014arXiv1409.1556S> (accessed September 01, 2014).
- Syauqie, M., Hastono, S. P., and Siregar, K. N. (2025). Revealing the inherent properties of pupillary red reflex in refractive errors by smartphone imaging: a study in Indonesia's population. *Optics Continuum*, 4, 1519–33. doi: 10.1364/OPTCON.559949
- Varadarajan, A. V., Poplin, R., Blumer, K., Angermueller, C., Ledsam, J., Chopra, R., et al. (2018). Deep learning for predicting refractive error from retinal fundus images. *Invest. Ophthalmol. Vis. Sci.* 59, 2861–2868. doi: 10.1167/iops.18-23887
- Wang, Z., Kempen, J., and Luo, G. (2024). Using smartphones to enhance vision screening in rural areas: pilot study. *JMIR Form. Res.* 8:e55270. doi: 10.2196/55270
- Watts, P., Walker, K., and Beck, L. (1999). Photoscreening for refractive errors in children and young adults with severe learning disabilities using the MTI photoscreener. *Eye (Lond.)* 13, 363–368.
- Yang, Z., Fu, E. Y., Ngai, G., Leong, H. V., Do, C. -w., and Chan, L., editors. Screening for refractive error with low-quality smartphone images. The 18th International Conference on Advances in Mobile Computing & Multimedia; 2020; Chiang Mai, Thailand: Association for Computing Machinery; (2020).



OPEN ACCESS

EDITED BY

Sokratis Makrogiannis,
Delaware State University, United States

REVIEWED BY

Amrutanshu Panigrahi,
Siksha 'O' Anusandhan University, India
Mosiur Rahaman,
Asia University, Taiwan

*CORRESPONDENCE

Indo Intan

✉ indo.intan@undipa.ac.id

RECEIVED 31 January 2025

ACCEPTED 24 June 2025

PUBLISHED 01 September 2025

CORRECTED 05 September 2025

CITATION

Intan I, Karnyoto AS, Harlina S,
Nelwan BJ, Setiawan D, Yamin A and
Puspitasari RE (2025) Heterogeneous
ensemble learning: modified ConvNextTiny
for detecting molecular expression of breast
cancer on standard biomarkers.

Front. Comput. Sci. 7:1569017.

doi: 10.3389/fcomp.2025.1569017

COPYRIGHT

© 2025 Intan, Karnyoto, Harlina, Nelwan,
Setiawan, Yamin and Puspitasari. This is an
open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or reproduction
is permitted which does not comply with
these terms.

Heterogeneous ensemble learning: modified ConvNextTiny for detecting molecular expression of breast cancer on standard biomarkers

Indo Intan^{1*}, Andrea Stevens Karnyoto², Sitti Harlina³,
Berti Julian Nelwan⁴, Devin Setiawan⁵, Amalia Yamin⁶ and
Ririn Endah Puspitasari⁴

¹Department of Informatics Engineering, Concentration of Intelligent Systems, Dipa Makassar University, Makassar, Indonesia, ²Computer Science Department, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia, ³Department of Informatics Engineering, Concentration of Data Science, Dipa Makassar University, Makassar, Indonesia, ⁴Department of Anatomical Pathology, Faculty of Medicine, Hasanuddin University, Makassar, Indonesia, ⁵Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, United States, ⁶Laboratory of Anatomical Pathology, Wahidin Sudirohusodo Hospital, Makassar, Indonesia

Breast cancer is the highest-ranking type of cancer, with 2.3 million new cases diagnosed each year. Immunohistochemistry (IHC) is the gold standard “examination” for determining the expression of cancer malignancies in patients with the ultimate goal of determining prognosis and therapy. Immunohistochemistry refers to the four WHO standard biomarkers: estrogen receptor, progesterone receptor, human epidermal growth factor receptor-2, and Ki-67. These biomarkers are assessed based on the quantity of cell nuclei and the intensity of brown cell membranes. Our study aims to detect the expression of breast cancer malignancy as an initial step in determining prognosis and therapy. We implemented homogeneous and heterogeneous ensemble learning models. The homogeneous ensemble learning model uses the majority vote technique to select the best performance between the Xception, ResNet50V2, InceptionResNet50V2, and ConvNextTiny models. The heterogeneous ensemble learning model takes the ConvNextTiny model as the best model. Feature engineering in ConvNextTiny combines convolution and cell-quantification features as feature fusion. ConvNextTiny, which applies feature fusion, can detect the expression of cancer malignancy. Heterogeneous ensemble learning outperforms homogeneous ensemble learning. The model performs well for accuracy, precision, recall, F_1 -score, and receiver operating characteristic-area under the curve (ROC-AUC) of 0.997, 0.973, 0.991, 0.982, and 0.994, respectively. These results indicate that the model can classify the malignancy expressions of breast cancer well. This model still requires the configuration of the visual laboratory device to test the real-time model capabilities.

KEYWORDS

breast cancer, ConvNextTiny, ensemble learning, Canny, Otsu, IHC, ER/PR/Ki-67, HER-2

1 Introduction

Breast cancer (BC) ranks first in women among all types of cancer in the world (Intan et al., 2024). Approximately 2.3 million cases spread across various countries, and as many as 666,103 deaths (Alismail, 2024). Asian countries have the highest number of 985,817 cases and 315,309 deaths (World Health Organization, 2022). In line with global cases, in Indonesia, breast cancer ranks first in cancer cases in women. Global cancer data in 2022 shows that the total cases in women are 30.1% (66,271 cases) and 19.8% (22,598 cases) (World Health Organization, 2022).

Breast cancer is a type of cancer that occurs when malignant cells grow in breast tissue and is heterogeneous, characterized by various molecular subtypes and genotype profiles (Chen et al., 2024). These cells can form tumors that can be felt on physical examination or detected through mammography. Breast cancer is more common in women but can also occur in men in very small numbers (Kemenkes, 2024). Breast cancer has various presentations with different molecular subtypes, with different biomolecular, pathological, and genetic features, and with different clinical and therapeutic response results, so breast cancer is called a heterogeneous disease. These molecular markers are known to be closely related to oncogenic transformation, cancer cell proliferation, tumor growth, treatment options, and prognosis of breast cancer (Joensuu et al., 2013; Afkari et al., 2021).

An immunohistochemistry (IHC) is an examination to determine the characteristics of breast cancer. The examination involves biomarkers, which are widely used in the process of diagnosis, prognosis, and treatment for patients with breast cancer. Biomarkers are useful for both patients who have recently had breast cancer and those who have experienced a recurrence. There are four types of biomarkers, which are the WHO gold standard that is routinely enforced in the characterization and diagnosis of breast cancer, namely estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor-2 (HER-2), and Ki-67. Especially for ER, PR, and HER-2, implementation is easy because it is effective and inexpensive (Alismail, 2024).

Estrogen receptors (ERs) and progesterone receptors (PRs) are both important for predicting breast cancer pathogenesis and treatment response. Both are hormone receptors in response to estrogen and progesterone; ER and PR contribute to cancer growth in hormone-sensitive breast tissue by facilitating cancer cell proliferation (Alismail, 2024). Estrogen receptors (ERs) have remained the most important biomarker in breast oncology for 60 years after their discovery. ER status is very urgent in clinical decisions and outcome prediction for breast cancer patients, including determining the right therapy for patients. The results can significantly improve clinical outcomes with ER-positive characteristics. As an important predictive biomarker, visualization of its image requires analysis that meets the standard scoring of cell nuclei against stained cell nuclei. The majority of BC are ER-positive. Visually, ER has “positive” and “negative” characteristics. Negative characteristics if ER staining is $\geq 1\%$ of the cell nucleus by IHC (Allison et al., 2021; Reinert et al., 2022; Loggie et al., 2024). Another scoring standard, ER-negative, has a cell nucleus threshold of $\leq 10\%$ (Fei et al., 2021) or ranges from 2 to 7% (Loggie et al., 2024). Progesterone receptor (PR) is a member of the nuclear/steroid hormone receptor (SHR) family of ligand-dependent transcription factors expressed primarily in female reproductive tissues and the central nervous system. In response to the binding of

its related steroid hormone, progesterone, PR regulates the expression of a network of genes to control the development, differentiation, and proliferation of target tissues, as well as pathological processes in endocrine-based cancers (Grimm et al., 2016). PR characteristics include “positive” and “negative.” Visually, PR is “positive” if the score of stained cell nuclei has a cutoff $>1\%$ (Shao et al., 2024) or $>10\%$, conversely if PR is “negative,” then the score of stained cell nuclei is $<10\%$ (2–7%) (Alismail, 2024; Loggie et al., 2024).

Human epidermal growth factor receptor 2 (HER-2) is overexpressed in approximately 15–30% of breast cancer cases. Thus, it is considered an important prognostic and predictive biomarker for breast cancer. Unfortunately, HER-2 overexpression is associated with a more aggressive tumor phenotype characterized by prone metastasis, poor prognosis, and high recurrence rates. This suggests that HER-2-positive breast cancer is often associated with more advanced stages (Alismail, 2024). Routine determination of HER-2 status is performed using techniques such as immunohistochemistry (IHC) and fluorescence *in situ* hybridization (FISH). FISH detects HER-2 gene amplification, while IHC evaluates HER-2 protein expression levels; both techniques determine eligibility for HER-2-targeted therapy (Lv et al., 2016). Accurate determination of HER-2 status influences therapy choice and prognosis, which are critical for the best patient management (Alismail, 2024). HER-2 has a score that is taken from ER and PR. At the cutoff limit of 10%, the threshold $<10\%$ is HER-2 negative, conversely, if the threshold $\geq 10\%$ is HER-2 “positive.” HER-2 is divided into three subtypes based on IHC scores: “negative” (IHC 0/1+), equivocal cases (IHC 2+), and “positive” cases (IHC 3+). Equivocal cases are retested with FISH to verify their HER-2 expression more accurately. Positive cases indicate that patients are eligible for anti-HER-2 therapy (Lv et al., 2016).

Ki-67 is a widely used biomarker to measure and monitor tumor proliferation in breast specimens, although there is poor agreement on the analytical approach to its assessment, assessment methods and cutoffs, data handling, and appropriate clinical utility of the biomarker. Ki-67 appears to be a continuously variable type marker, reflecting tumor biology (Penault-Llorca and Radosevic-Robin, 2017). Testing for Ki-67 is performed using different methods, and cutoffs for defining Ki-67 “positive” and “negative” or “high” and “low” populations are not clear. Consequently, the Tumor Marker Guidelines Committee of the American Society of Clinical Oncology (ASCO) determined that the evidence supporting the clinical utility of Ki-67 is insufficient to recommend routine use of this marker for prognosis in patients with newly diagnosed breast cancer. Standardization of Ki-67 assessment is a global standard set by the WHO to improve its reproducibility. The clinical utility of very low and very high Ki-67 indices is good. The 25% threshold has shown significance for predicting overall survival. Multigene testing can provide useful information to guide the management of patients with ER+/HER-2 breast cancer in the “gray zone” Ki-67 index (between 15 and 25%) (Penault-Llorca and Radosevic-Robin, 2017).

Therefore, the four biomarkers visually indicate the expression of cancer malignancy through the number of cells and the extent of brown color between the stained cells. Determining the characteristics of each biomarker will provide appropriate treatment recommendations for breast cancer patients.

Pathologists have difficulty observing tissues with the naked eye and manually analyzing images based on their knowledge and skills. First, they perform fundamental techniques through microscopic

observation of cell morphological structures (nuclei and cell membranes). The method relies heavily on manual naked-eye observation, so it does not save time because IHC images have complex, uneven cell color distribution between normal and cancerous stained cells, overlapping cells, and uncertain cell sizes. Second, the objectivity of the observation results depends on the experience and accommodation of the pathologist's eyes, so the results sometimes differ. Third, there are so many cases of breast cancer in hospitals that it is tiring if the process only relies on manual techniques. On the other hand, the demand for examination results must be released quickly to patients as part of hospital management.

Therefore, this study proposes a modified ConvNextTiny to detect breast cancer malignancy expression, combining cell quantification and convolution features. The cell quantification feature adapts cell morphology as a fundamental pathologist calculation. The convolution feature is a ConvNextTiny feature that utilizes a convolutional neural network architecture and pre-trained transfer learning weights. The cell quantification calculation is a WHO calculation standard in the ConvNextTiny feature that performs well with transfer weights from “imagenet” (transfer learning weights). Combining both is an advantage in finding unique patterns for each breast cancer image that will improve model performance.

In this study, we experimented with two ensemble transfer learning models, namely homogeneous and heterogeneous ensemble learning. The homogeneous ensemble learning model was constructed using a majority voting scheme among four models: Xception, Resnet50V2, InceptionResnet50v2, and ConvNextTiny. We trained each model individually and performed hyperparameter tuning, with particular focus on evaluating the learning rate. ConvNextTiny was selected as it shows dominant performance within the homogeneous ensemble. Combining the ConvNextTiny and cell quantification model features is the final ensemble model by concatenating their features into the ConvNextTiny neural network classifier. The advantage of the homogeneous ensemble learning model is based on transfer learning from CNN, which has high computational feature extraction and classification capabilities, and ConvNextTiny has a short computational time. Moreover, using cell quantifications involves a practical model and simple computation. Combining these two models using feature fusion balances the complexity of the algorithm and delivers better performance and acceleration. The final model classification result is in the form of ER and PR expressions: “positive” and “negative”; HER-2 is +1 (negative) and +3 (positive), while Ki-67 is low and high. The final results show that our proposed model outperforms all single models and as an ensemble result. Our contributions: (1) built the model using feature fusion that contains 768 ConvNextTiny features and one cell quantification feature from (Canny and Otsu); (2) used four biomarkers (ER, PR, HER, and Ki-67) as input for homogeneous and heterogeneous breast cancer classification.

The study presents a systematic approach to detecting four breast cancer biomarkers—ER, PR, HER-2, and Ki-67. Section 1 discusses the biomarkers and highlights the novelty of the research. Section 2 reviews previous studies and emphasizes the contributions made by this work. Section 3 discusses the methodology in detail, including dataset collection, integration, and the application of ensemble learning algorithms with feature fusion techniques to enhance performance. Section 4 demonstrates that feature fusion significantly improves model accuracy in detecting cell nuclei scores and brown intensity in image objects, leveraging feature extraction and combination in both models. Finally, Section 5 concludes that the

proposed model meets the needs of ensemble learning while aligning with pathologists' practices, achieving superior performance in accuracy, precision, recall, F_1 -score, and ROC-AUC through the combination of cell quantification features and ConvNextTiny features.

2 Related studies

Several previous researchers conducted studies focusing on investigating the status of ER, PR, HER-2, and Ki-67, as well as other biomarkers, using machine learning, deep learning, CNN model adaptation, and stained cell expression scoring.

Fan et al. (2024) presented an intelligent, holistic breast cancer tumor diagnosis system, including an interpretation module and a subtype module. The interpretation module is used to extract and analyze data based on a CNN-based convolutional neural network from HER-2, ER, PR, and Ki-67 images, followed by classification analysis. The subtype module produces holistic detection results of critical tumor markers with diagnostic suggestions for molecular subtypes validated by three pathologists. The model architecture consists of four convolution layers, four pooling layers, fully connected layers, and one output layer. The used dataset consists of 104 HER-2 cases, 198 ER and PR cases, and 60 Ki-67 cases.

Kildal et al. (2024) proposed a model using Mask R-CNN, YOLOv5, and deep learning to detect nuclear, cytoplasmic, and membranous IHC staining patterns in five image objects, namely colon, two prostate, breast, and endometrial. Image objects of the biomarker Ki-67 for colon, prostate, and breast cancer; PMS2 and MSH6 for colon and endometrial cancer; PTEN, CCNB1, CD44, Flotillin1, Mapre2, and β -catenin for prostate cancer; and ER and PR for breast cancer. The models consist of three, namely the nuclear model, the cytoplasmic model, and the membranous model. The nuclear model consists of 69 whole slide imaging (WSI) from the Ki-67 colon set and 23 WSIs from the PMS2-colon set; the cytoplasmic model consists of 34 WSIs from the PTEN-prostate set; and the membranous model consists of 25 WSIs from the β -catenin prostate. The image size is 800×800 pixels at $40\times$ magnification, as the feature and the labels are “positive” and “negative” for each biomarker.

Zhao et al. (2024) developed a ResNet-18 model based on the framework and an online clinical application platform to predict molecular features and patient prognosis from triple-negative WSI pathology. The framework architecture consists of a serially working part to compare two separate convolutional networks (CNNs). The first is a tissue type classifier developed based on 20 WSIs' pixel-level tissue type annotations connected to the prediction target. The second is a CNN trained based on sample tiles for different targets. The models were trained and validated using the Fudan University Shanghai Cancer Center Triple Negative Breast Cancer (FUSCCTNBC) cohort through three-fold cross-validation. All three models were applied to the The Cancer Genome Atlas Triple Negative Breast Cancer (TCGATNBC) cohort. Each patient received three prediction scores, and the average was used for the final prediction. Performance metrics were then computed for external validation.

Tafavvoghi et al. (2024) performed two scenarios: *first*, classifying tiles in tumor and non-tumor areas for molecular subtypes using InceptionV3, the tile matrix size is 512×512 , then decreased in size (1×1 , 3×3 , 5×5); *second*, using the One-vs-Rest (OvR) strategy to train four binary OvR classifiers and combining the results using the Xtreme Gradient Boosting model. The datasets accessed from The

Cancer Genome Atlas-Breast Cancer Gene (TCGA-BRCA) were 1,175, Breast Cancer Screening System (BRACS) were 129, Clinical Proteomic Tumor Analysis Consortium_Breast Invasive Carcinoma (CPTAC_BRCA) was 382, and HER-2-Warwick was 71.

Solorzano et al. (2024) built a single CNN model and an ensemble model from Inception V3, ResNet50, Inception-ResNet V2, and Xception to determine the presence of invasive carcinoma, IC or not IC. The datasets from Clinseq were 232 WSI and Sos 355 WSI, a total of 2,502,649 tissue tiles of size 598×598 pixels ($271 \times 271 \mu\text{m}$) at $20\times$ magnification. To determine the last decision, voting was applied to the model.

Bychkov et al. (2022) detected mitosis, nuclear pleomorphism, and tubule formation images using ResNet CNN. The biomarkers used were BCSS, ER, and ERBB2. The evaluation technique was applied to the model trained on the FinProg test set, which refers to the internal test set, and the FinHer patient series, which was not used for all training. The average output of the five trained models was used for cross-validation to reduce CNN variance and improve prediction accuracy. Validation between prediction scores (CNN output) and real-time sensor readings was conducted using statistical analysis based on Cox PH multivariate regression.

The advantages of our implemented model are as follows: (1) Accommodating conventional techniques that pathologists use to analyze the cell morphology. The quantification and intensification of the staining of the nucleus and membrane of cancer cells determine the expression of cancer malignancy. (2) Homogeneous ensemble learning improves model performance results during training and testing. The majority vote technique selects the best model among the four models, which is more efficient than the average bagging technique. (3) Heterogeneous ensemble learning, through feature fusion of concatenated different features, significantly improves model performance to be visually and medically representative. Feature engineering uses modified ConvNextTiny as a concatenation of convolution and cell quantification features. Computation time is shorter because it uses one ConvNextTiny classifier.

3 Methods

3.1 Datasets

We used datasets from Hasanuddin University Hospital (HUH) and Wahidin Sudirohusodo Hospital (WSH), consisting of 300 WSI: 200 from WSH, and 100 from HUH, which were then sampled into 23,351 samples. The dataset is closed access, and the owner's consent is required. The immunohistochemistry (IHC) biomarkers used for each patient consisted of estrogen receptor (ER), progesterone receptor (PR), HER-2, and Ki-67. However, the condition of biomarkers in the laboratory is not always complete, so data imbalance is an obstacle. The data composition of each biomarker consists of 9,035 ER image samples, and the total dataset used was 1,499 images. As shown in Table 1, the total dataset used was 23,331 images.

Because the image size during capture varies depending on the image cropping area and device resolution, we need to do data preparation. This step is crucial because, besides being thorough, it is also very time-consuming. Therefore, the input data are standardized to 224×224 and normalized for computing speed needs. In addition, determining the boundaries of cell morphology and membranes

TABLE 1 Datasets of ER, PR, HER-2, and Ki-67.

Images	Train	Validation	Test	Total
ER	4,790	598	506	8,894
PR	3,268	408	410	4,086
HER-2	5,309	663	665	6,637
Ki-67	5,370	671	673	6,714
Total	18,737	2,340	2,254	23,331

requires careful visual analysis to annotate each cell boundary and its cell membrane boundary. Not all WSIs are normal; some have blurriness during WSI pre-processing, so the cell morphology does not match the actual one.

3.2 Pre-processing

We performed physical data acquisition and data preparation that met the required qualification standards. The techniques used included performing the cutting process when capturing the image of each biomarker sample (cropping). The rectangular image dimensions vary according to the ratio of the size and area of the WSI. For instance, an image measuring $1,280 \times 613$ pixels has a size of approximately 224 kb. WSI scans using KBIO, China Scanner, capture the entire cross-section of the WSI image area. Additionally, we read, enlarge, capture, and crop using Slideviewer 3DHISTECH, Budapest, Hungary, with magnifications of $5\times$, $10\times$, $20\times$, and $40\times$. The selection of these magnifications is based on the clarity of the colored cell objects, which adapt to the devices of the two hospitals. Physical data acquisition continues with virtual data acquisition if it is already in the programming framework. Acquisition is done by resizing the original image size into patches (224×224) to ensure uniform data size, facilitating the arithmetic and geometric operations of the input data.

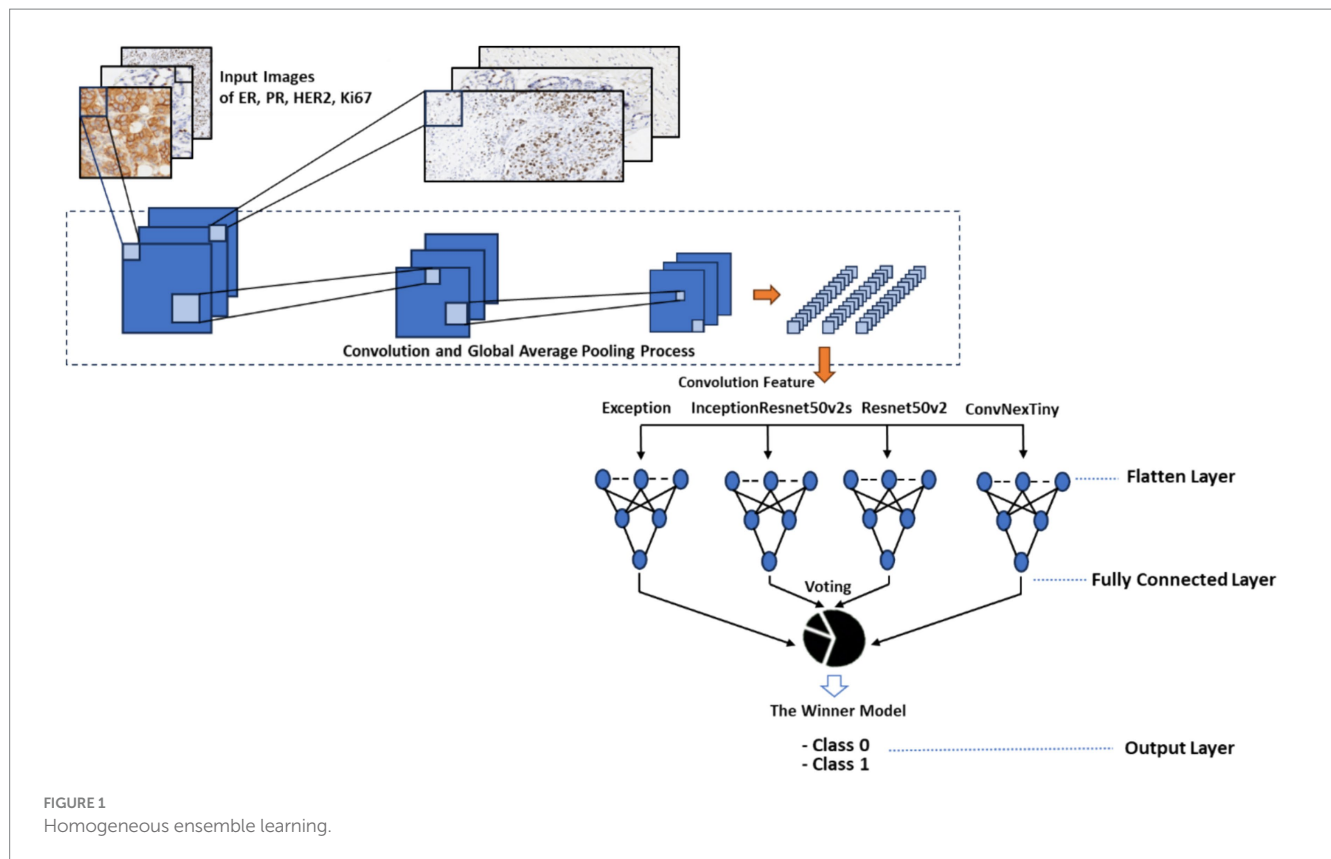
Here, we focus specifically on the preprocessing of handcrafted features, as preprocessing for convolutional neural networks is handled automatically without manual intervention. The preprocessing techniques applied include Gaussian Blur, Otsu thresholding, and mask inversion. An 11×11 Gaussian blur is used to blur the image, remove noise in the image, and make the transition between areas smoother. At the same time, Otsu's method is an automatic technique for determining the optimal threshold to divide grayscale images into two classes: foreground and background. ConvNextTiny was pre-trained using ImageNet weights as transfer learning weights. All handcrafted features use feature standardization for data uniformity.

3.3 Processing

Processing and performing modeling based on the methods used in the feature extractor and classifier.

3.3.1 Feature engineering

Feature engineering consists of two main types: cell quantification feature (handcrafted feature) and convolution feature. The cell quantification feature extracts using Canny edge detection and the Otsu thresholding technique. The Canny edge detection



(Intan et al., 2023) sharpens the edges of cancer cells and colored cell areas while counting the number of cells (Figure 1). Otsu edge thresholding (Chadha et al., 2020) provides a circle boundary on the cell membrane object in a clear circle area through feature extraction (Figure 2) (Aswathy and Jagannath, 2017)—next, automatic convolution feature extraction through the ConvNexTiny model in each of its layers. The third stage of feature extraction is the scored feature. The results of the feature extraction stage are used to extract the number of cells based on their color indications, and then scoring (quantification) is carried out.

In handcrafted image processing, Canny focuses on changes in sharpness between pixels (high gradient) through the following steps: (1) Gaussian blur to reduce noise; (2) Gradient magnitude to detect changes in intensity; (3) Non-maximum suppression to produce thin edges; (4) Hysteresis thresholding to filter strong and weak edges based on pixel intensity; (5) Calculate the average value and standard deviation of the number of cell nuclei. Unlike Canny, Otsu focuses on area segmentation based on pixel intensity. However, Otsu's nature is used to separate between pixel conditions greater than the threshold, where the value is 0, and otherwise, the value is 1. Canny performs edge detection using the hysteresis limit of the image pixel strength. If the pixel is smaller than 30, it is ignored, and if it is above 105, the label is marked as an edge.

The second handcrafted technique uses the Otsu threshold with the following steps: (1) calculate pixel intensity using a binary threshold (8 bits, 0–255), intensity 0–50 is 0, while 100 is 1; (2) calculate the probability of each image intensity (Equation 1); (3) iterate for object and background classes; and (4) calculate the mean and standard deviation of the brown cell membrane area in the HER-2 biomarker.

$$P(i) = \frac{\text{Number of Selected Intensity Pixels}}{\text{Total of Pixels}} \quad (1)$$

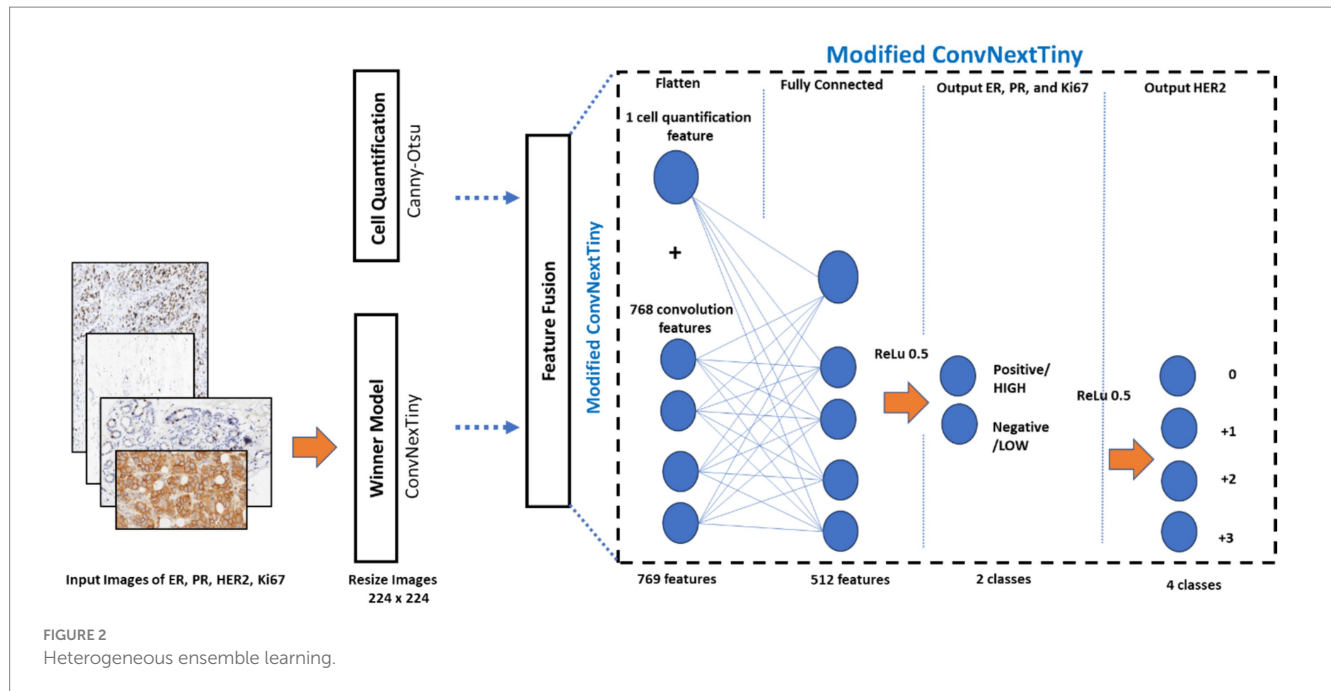
3.3.2 Classifier

Our experimental setup involved two ensemble learning models: homogeneous and heterogeneous. For the homogeneous ensemble, we employed a majority voting technique among four distinct models (Xception, ResNet50V2, InceptionResNet50V2, and ConvNextTiny) to determine the best overall performer. Through this process, ConvNextTiny consistently demonstrated superior performance across various metrics, indicating its strength as an individual classifier. Therefore, for the heterogeneous ensemble learning model, ConvNextTiny was selected as the foundational architecture due to its dominant performance within the homogeneous ensemble. This allowed us to integrate additional handcrafted cell quantification features with ConvNextTiny's convolutional features, creating a robust fusion model.

3.3.2.1 Homogeneous ensemble learning

3.3.2.1.1 Model

The classifier used consists of four models, as follows (Figures 1, 2): *First*, Xception (Li et al., 2023) has a basic CNN architecture; convolution is used to process the main features against the deep convolution separately from the CNN convolution to achieve feature extraction and successful computation, reducing the number of parameters (Sharma and Kumar, 2022; Krishna et al., 2023). *Second*, Resnet50V2 (He et al., 2016; Rahimzadeh and Attar, 2020), Residual



Network 50V2, an architecture that has a stack block with the same connection shape (Residual 3 Unit). This base model has the following advantages: (1) ease of optimization and (2) reducing overfitting, unnormalized signals used as input to the next layer, so that all inputs are normalized. *Third*, InceptionNetResnetV2S (Asif et al., 2022; Talukder et al., 2023), each block is followed by an expansion filter layer used to increase the filter bank dimension before augmentation, according to the input thickness. *Fourth*, ConvNextTiny (Tanvir et al., 2024) is a novel convolutional neural network architecture that leverages standard CNN modules and incorporates optimization techniques inspired by the transformer model. ConvNextTiny has a network structure that shows great development potential through comprehensive experimental demonstrations covering macro and micro designs based on ResNet. This model outperforms the Swin Transformer while maintaining the simplicity and efficiency characteristics of standard CNN architectures (Yang et al., 2022).

In the homogeneous model ensemble (Figure 1), first, all models are applied individually to obtain the best weight hyperparameter tuning results. Starting from Exception, Resnet50v2, InceptionResnet50V2, and ConvNextTiny are based on convolution. We perform the ensemble by majority voting among the four models; the results will be the output of the homogeneous ensemble learning model.

Majority voting is a technique used to determine the final decision based on the largest number of labels in the entire model (Equation 2). Suppose there are $M = 4$ classification models, K possible classes; the m -th model gives a prediction. $\hat{y}$ is the class label with the most votes, and y is an indicator function (1 if true, 0 otherwise).

$$\hat{y} = \arg \max \sum_{m=1}^M \mathbb{I}(y^{(m)} = c_j) \quad (2)$$

Each image will be plotted into an image grid in pixel form using “imagenet” weights with a learning rate of 10^{-3} and 10^{-4} as its training

tuning. The image goes through a feature extraction process, starting from resizing, convolution, and global average pooling 2D to produce 1,024 features in the Exception, Resnet50v2, and InceptionResnet50V2 models and 768 features in ConvNextTiny. This feature is input for a neural network that uses the ReLu activation function and the “softmax” optimizer. The three models have 1,024 features, while ConvNextTiny has 768 features. Those four feature blocks are input to the neural network so that the output produces 512 features and finally produces two classifications, “positive” and “negative,” or low and high. Each single model has the same parameter structure initialization. Similarly, parameter tuning is carried out by taking several learning rate scenarios.

The results of turning parameters with a learning rate of 10^{-3} then become weights for pre-trained. The results of these weights become the initialization when doing the second training to obtain the best weights. Pre-trained has a learning rate of 10^{-4} to load parameters with a learning rate of 10^{-3} , down from before, to obtain a smaller gradient descent, so that the loss decreases and the accuracy improves during training (Equations 6, 7).

3.3.2.2 Heterogeneous ensemble learning model

Medical record data images in ER, PR, Ki-67, and HER-2 images were taken from the examination results released by the pathologist. The pathologist selected the threshold for cell quantification (nucleus and cell membrane) based on WHO standards. The model determines the classification of ER, PR, and Ki-67, focusing on the number of dark brown cell nuclei, while HER-2 focuses on the area of brown cell membranes. The ER and PR use a threshold of 1%. The “negative” class has several cells $\leq 1\%$ in stained cells; conversely, if the number of cells is $> 1\%$, then the class is “positive.” In HER-2, it does not count the number of cells but computes the intensity of the brown color in the image. The HER-2 threshold is at 10%; if the intensity of the dark brown color is greater, then it is “positive,” and if not, then it is “negative.” Unlike the three previous biomarkers, Ki-67 has a higher threshold of 20%. The classification results are the labeling of the

images loaded into the model, and also serve as learning data for the model to recognize classification characteristic patterns: positive, negative, high, and low.

A heterogeneous ensemble model is a concatenation of homogeneous ensemble learning and cell quantification. Cell quantification is handcrafted from Canny edge detection and Otsu segmentation. Cell quantification, as a Canny edge detection model, counts the number of stained cells and Otsu counts segmentation of the area of dark brown cell membranes. Canny (Intan et al., 2023) cell quantification is an edge detection technique to reduce noise, preventing fake edge detection. The image $I(x, y)$ is filtered with a Gaussian Kernel to produce a convolution image, $I_s(x, y)$, as shown in Equations 3, 4.

$$G(x, y) = \frac{1}{2\pi\sigma^2} \cdot e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (3)$$

$$I_s(x, y) = I(x, y) * G(x, y) \quad (4)$$

To determine the feature map of edge boundaries using binary thresholding, 0 and 1. If $T_{\min}$ then 0 (not an edge boundary) and $T_{\max}$ then 1 (edge boundary) (Equation 5). The morphological kernel is rectangular 1×1 and produces the number of contours as the number of cells for classification.

$$g(x, y) = \begin{cases} 1 & \text{if } T_{\min} \leq f(x, y) \leq T_{\max} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Otsu determines the binary threshold using a 3×3 elliptical kernel to segment the brown cell area. The cell quantification feature (Table 2), based on average values and standard deviations, indicates that ER, PR, and Ki-67 show average cell nucleus counts and standard deviations in the training data. These range from 1,600 to 2,400 and 1,250 to 1,650, respectively. On the other hand, HER-2 has an average area percentage of 0.32886 with a standard deviation of 0.17146. These average values are the cell quantification features used to train the model and determine the best final parameters. They are also key parameters for classifying conditions as “positive” or “negative” and “high” or “low.”

The cell quantification model computes the number of cells resulting from extracting Canny features, as explained in Intan et al. (2023). It then computes the average number of cells in the training data and its standard deviation value. The image size, a sample patch for each image, is 224×224 . Patches are partitions of each image into smaller square sizes as two-dimensional images that will be converted into n -dimensional features according to the layer's dimensions. This number of cells only has one feature to be input to the ConvNextTiny classifier. The layer structure of

ConvNextTiny consists of four-layer blocks: the first block has 96 features; the second block has 192 features; the third block has 384 features; and the last block has 768 features and a ReLU activation function. The last block consists of 768 connecting features combined with one cell quantification feature to produce 769 convolution features, also called feature fusion. The process results in a modified ConvNextTiny. Feature fusion will simplify the features of both techniques that were initially separate, aiming to simplify the feature layer while improving the performance of the convolution model from 768 features. The 769 features are input to the neural network to be passed through the RELU activation function, and then 512 features are produced at its dropout (0.5), which are classified into two classes. The ER, PR, and HER-2 produce “negative” (0 and +1) and “positive” (+2 and +3) classes, while Ki-67 produces “low” and “high” classes.

Figure 3 is a general concatenation of ensemble learning. The result of homogeneous ensemble learning in the form of the best model is ConvNextTiny, then the features of the ConvNextTiny head layer consist of 768 convolution features. Cell quantification is a manual feature extraction (handcrafted feature extraction) consisting of one Canny feature and one Otsu feature, each combined into the ConvNextTiny head to form modified ConvNextTiny. Modified ConvNextTiny is the last classifier to determine the final detection and prediction.

3.3.3 Evaluation

Model performance is an indicator of the success of building a new model. A good model will have improved performance with its reference model. It requires parameter tuning to obtain better results if it does not improve. Model classification requires validation of its output; if the probability of correct validation is high, then it is confirmed to be a good model; conversely, if the probability is low, it will reduce model performance. It is also a factor in the feasibility of a model being implemented. This study model uses a confusion matrix (Wang et al., 2025; Chicco and Jurman, 2023), which is an element of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Some of the parameters measured include:

- 1 Loss: Computes the predicted and actual values of the model. y_i is the true label, p_i is the predicted probability for some samples N (Terven et al., 2025).

$$\log \text{loss} = \frac{1}{N} \sum_{i=1}^N y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \quad (6)$$

- 2 Accuracy: Measures the proportion of correct data predictions to the overall model predictions (Wang et al., 2025; Chicco and Jurman, 2023). Higher accuracy values indicate better model performance, indicating that the majority of the data have correct classifications.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (7)$$

TABLE 2 Cell quantification features.

	ER	PR	Ki-67	HER-2
Mean	1610.3472	2343.6925	2048.1196	0.32886
Deviation standard	1454.9679	1610.3472	1252.7431	0.17146

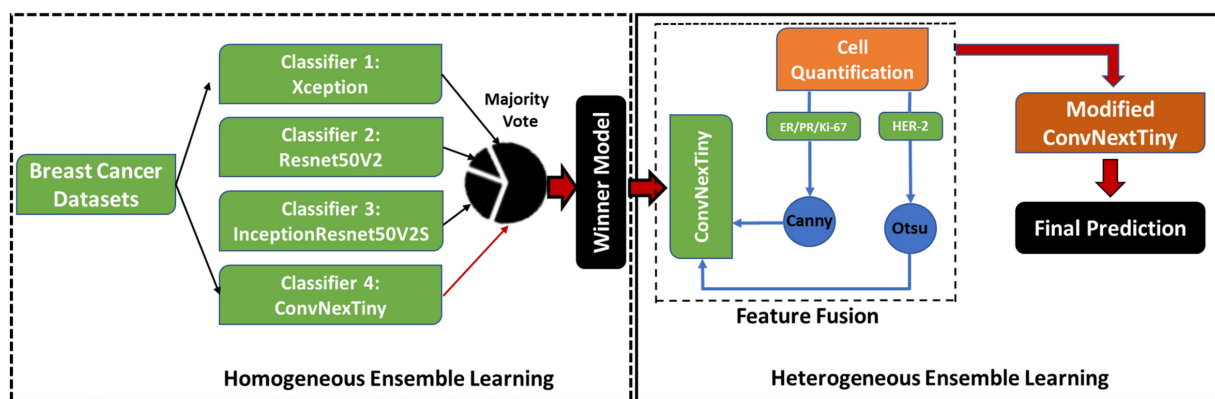


FIGURE 3
The concatenation of ensemble learning.

- 3 Precision: This metric assesses the proportion of true positive predictions to the total positive predictions (Wang et al., 2025; Chicco and Jurman, 2023) (Equation 8). A higher precision value indicates false positive errors of the minor data, focusing on the model's ability to classify samples accurately.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (8)$$

- 4 Recall: This metric measures the proportion of true positive predictions from the total number of positive samples (Wang et al., 2025; Chicco and Jurman, 2023) (Equation 9). A higher recall value indicates fewer false negative errors, reflecting the model's ability to identify positive cases correctly.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (9)$$

- 5 F_1 -score: The harmonic mean of precision and recall measures the model's performance (Wang et al., 2025; Chicco and Jurman, 2023) (Equation 10). A higher F_1 -score indicates better model performance, balancing precision and recall.

$$F_1 - \text{score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

- 6 ROC Curve is a graph that shows the performance of a binary classification model at various threshold values by plotting (Martínez Pérez and Pérez Martín, 2023; Carrington et al., 2023): (1) true positive rate (TPR) on the y-axis (also called sensitivity), and false positive rate (FPR) on the x-axis (which is 1—specificity) (Equations 11, 12).

$$\text{TPR} = \frac{TP}{TP + FN} \quad (11)$$

$$\text{FPR} = \frac{FP}{FP + TN} \quad (12)$$

4 Results and discussion

4.1 Homogeneous ensemble learning

A single model is the original model of each model element combined. Simulation of each model has different characteristics, including shorter computation speed, epoch, accuracy, and loss, which are parameters owned by each model.

In Tables 3–6, all models' performance and classification results show that ConvNexTiny has the highest accuracy and the smallest loss for all types of biomarkers, namely ER, PR, HER-2, and Ki-67. The model achieved the highest accuracy of 0.9933 and the smallest loss of 0.0078. These results indicate that the ConvNexTiny model wins the majority voting results on these metrics. Moreover, the classification results show that its valid data outperforms other models; even its invalid data has the smallest data, so it is very appropriate that ConvNexTiny is the best model among other single models.

Table 2 summarizes the performance of the ConvNexTiny model on the four types of images in detail. The experiments conducted on the training model used an initial learning rate of 10^{-3} , and then fine-tuning was performed at a learning rate of 10^{-4} until the model obtained optimum weights. These optimum weights are used as a reference for testing each biomarker. The lowest learning rate produces higher accuracy and lower loss due to the descent of errors through fine-tuning, which aims to reduce errors in gradient descent (Equation 6).

Table 4 shows the performance capabilities of the four models. ConvNexTiny outperforms the other three models, which can only recognize approximately 925 valid images, while ConvNexTiny can classify as many as 933 valid images and only four invalid images. The same thing is also shown in Table 5; ConvNexTiny outperforms the accuracy of Exception, Resnet50V2, and InceptionResnet50V2 with a value of 0.9974 and the smallest loss

TABLE 3 Model performance of ER and PR.

Models	ER				PR			
	Lr = 10 ⁻⁴		Lr = 10 ⁻³		Lr = 10 ⁻⁴		Lr = 10 ⁻³	
	Test performance							
	Acc	Loss	Acc	Loss	Acc	Loss	Acc	Loss
Exception	0.985	0.0449	0.9658	0.1051	0.9707	0.0577	0.9512	0.1318
Resnet50V2	0.985	0.0538	0.9573	0.1317	0.9585	0.1133	0.9512	0.1096
InceptionResnet50V2	0.979	0.0479	0.9712	0.0973	0.9732	0.0887	0.9512	0.1369
ConvNextTiny	0.9933	0.0078	0.9916	0.0394	0.9800	0.05	0.9636	0.1105

TABLE 4 Data distribution of ensemble model classification of ER, PR, HER-2, and Ki-67.

Models	ER				PR			
	Lr = 10 ⁻⁴		Lr = 10 ⁻³		Lr = 10 ⁻⁴		Lr = 10 ⁻³	
	Numbers of data							
	Valid	Invalid	Valid	Invalid	Valid	Invalid	Invalid	Valid
Exception	925	12	905	32	401	9	390	20
Resnet50V2	925	12	897	40	399	11	390	20
InceptionResnet50V2	925	12	910	27	399	11	390	20
ConvNextTiny	933	4			403	7		

TABLE 5 Model performance of HER-2 and Ki-67.

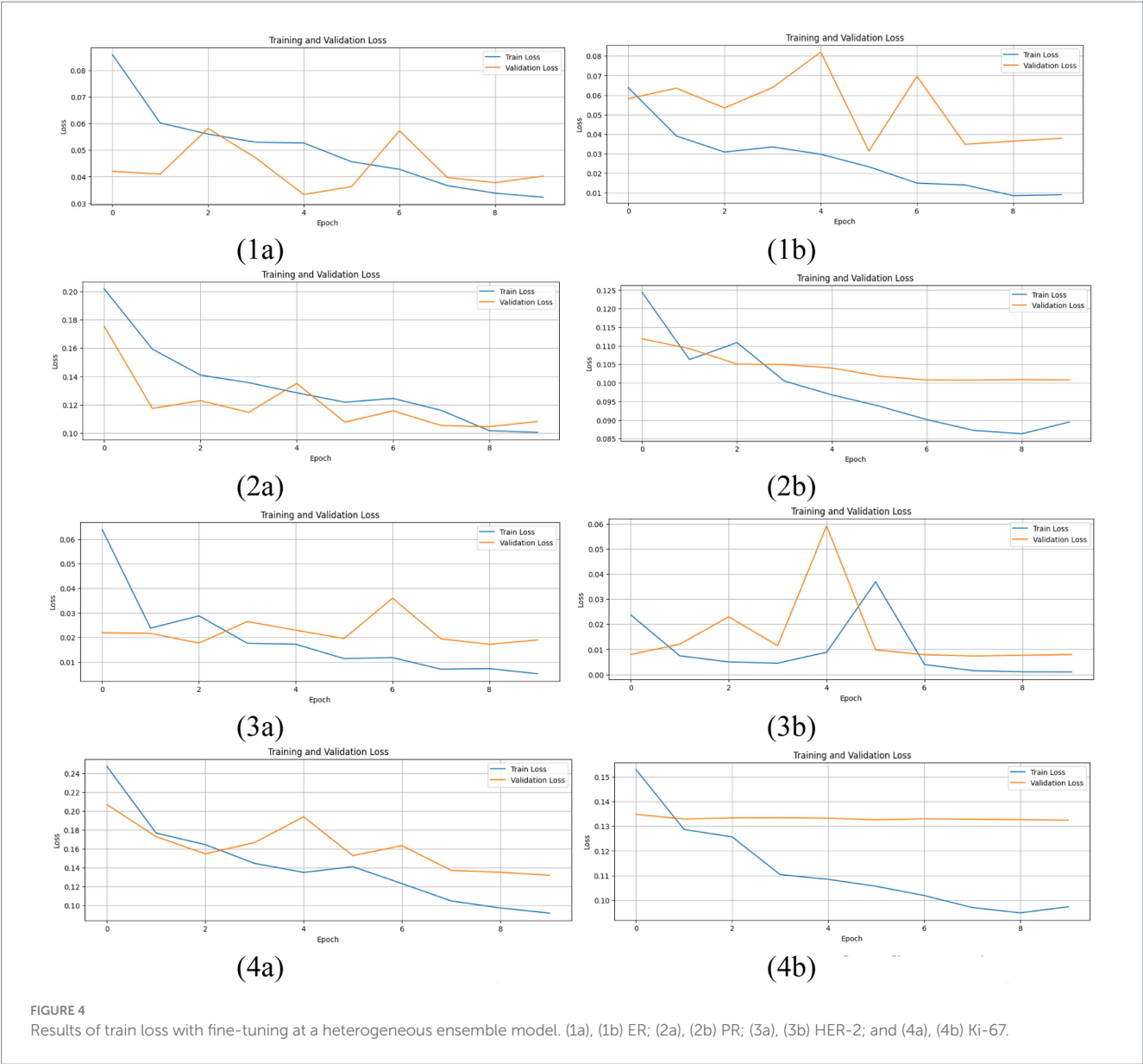
Models	HER-2				Ki-67			
	Lr = 10 ⁻⁴		Lr = 10 ⁻³		Lr = 10 ⁻⁴		Lr = 10 ⁻³	
	Test performance							
	Acc	Loss	Acc	Loss	Acc	Loss	Acc	Loss
Exception	0.9955	0.0185	0.9835	0.0631	0.985	0.0449	0.9658	0.1051
Resnet50V2	0.9925	0.0129	0.9865	0.0439	0.985	0.0538	0.9573	0.1317
InceptionResnet50V2	0.9925	0.0336	0.9835	0.0477	0.979	0.0479	0.9712	0.0973
ConvNextTiny	0.9964	0.0112	0.9960	0.0157	0.990	0.0419	0.9797	0.038

TABLE 6 Model classification of HER-2 and Ki-67.

Models	HER-2				Ki-67			
	Lr = 10 ⁻⁶		Lr = 10 ⁻³		Lr = 10 ⁻⁶		Lr = 10 ⁻³	
	Numbers of data							
	Valid	Invalid	Valid	Invalid	Valid	Invalid	Invalid	Valid
Exception	662	3	654	11	923	14	905	32
Resnet50V2	661	4	656	9	923	14	897	40
InceptionResnet50V2	555	10	654	11	918	19	910	27
ConvNextTiny	663	2			928	9		

value of 0.0112 on HER-2, as well as on Ki-67. The value has an impact on the data that is validated correctly (Table 6), with as many as 663 valid data and only two invalid data on HER-2. Similarly, on Ki-67, there are 928 valid data and nine invalid data. The more valid data depends on the higher accuracy and the lower

loss. Conversely, if the accuracy is lower and the loss is higher, it will affect the number of correctly validated data. Good performance is obtained from the fine-tuning process to obtain the smallest error and high accuracy, even though the training time is longer.



4.2 Heterogeneous ensemble learning

Figure 4 shows that the fine-tuning results provide better loss conditions. (1a), (2a), (3a), and (4a) show overlapping train and validation at a learning rate of 10^{-3} , while (1b), (2b), (3b), and (4b) show that fine-tuning at a learning rate of 10^{-4} successfully separates the training curve and the validation curve so that overlapping is resolved. The error is getting smaller, indicating that the model’s ability to distinguish between its two classification classes is improving during training and testing. To prove it, Table 7 shows the amount of data and the percentage of model classification on ER, PR, HER-2, and Ki-67.

Heterogeneous ensemble learning is an ensemble model between cell quantification and ConvNextTiny. Cell quantification uses grayscale, binary, blurred image, canny edge detection, dilated image, and contour techniques to determine the radius of the cell circle and compute the number of circles resulting from contours. This circle is a colored cell observed using a microscope display (Figure 5).

TABLE 7 Data distribution of ensemble model classification of ER, PR, HER-2, and Ki-67.

Images	Train	Val	Test	% Valid	% Invalid
ER	4,790	598	506	99.16	0.84
PR	3,268	408	410	97.06	2.94
HER-2	5,309	663	665	99.69	0.30
Ki-67	5,370	671	673	99.25	0.74

In addition to features learned via convolutional neural networks (CNNs), we incorporated a handcrafted feature to quantify cell density or stained area, depending on the biomarker type. We estimated cell quantification using Canny edge detection for ER, PR, and Ki-67 datasets. Images were first converted to grayscale and binarized using Otsu’s thresholding in inverse mode to isolate foreground structures. A Gaussian blur (kernel size: 11×11) was applied to reduce noise, followed by Canny edge detection with thresholds of 30 and 105. The

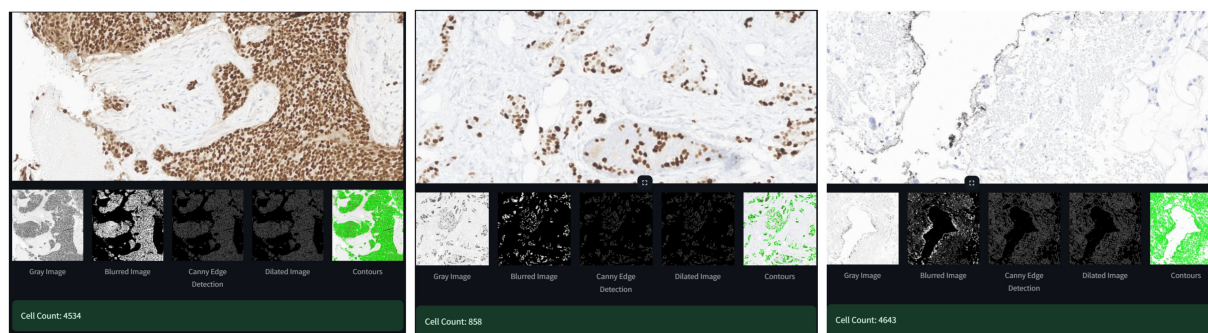


FIGURE 5

Cell quantification as results of feature extraction of Canny edge detection: ER (left), PR (center), and Ki-67 (right).

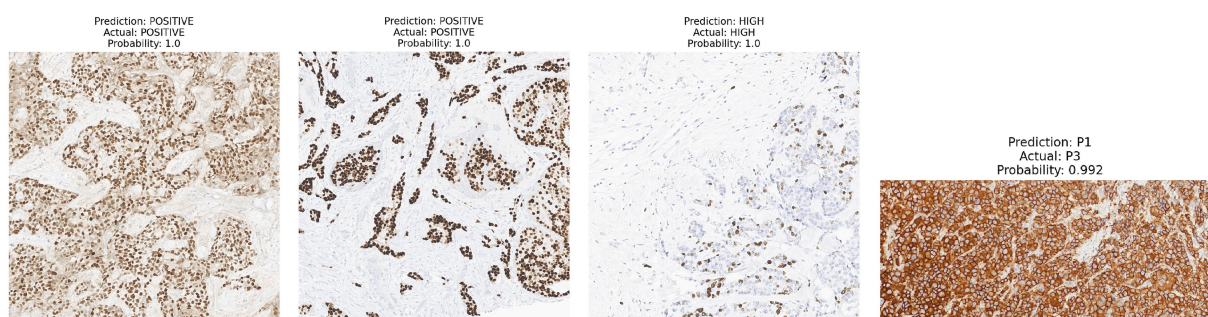


FIGURE 6

Model classification results with labels of ER, PR, Ki-67, and HER-2 (from left to right).

resulting edges were dilated using a 1×1 rectangular kernel for two iterations, and external contours were counted as a proxy for cell nuclei. These threshold values were selected empirically by evaluating multiple samples and identifying the settings that produced the most accurate and consistent contour quantification relative to visual inspection. The extraction of contour cells in ER, PR, and Ki-67 focuses on the brown nucleus circle of cells, totaling 4,534, 858, and 4,643, respectively (Figure 5).

For HER-2 images, we computed the proportion of stained (brown) regions to the total image area. Grayscale images were binarized using Otsu's method and inverted to highlight dark-stained regions. The morphological opening with an elliptical 3×3 kernel removed minor artifacts, and the stained ratio was computed as the fraction of non-zero pixels in the mask. All extracted features were standardized using the mean and standard deviation computed from the training set before being concatenated with CNN outputs for final classification. The model in Figure 6 does not compute the number of cells as circles, but it computes the brown cell area, indicating that the cell membrane is 0.45367 or 45.367%. This area is already within the threshold limit, the area of brown cells for HER-2 (Figure 7).

Table 8 shows that the results of the heterogeneous ensemble model have improved performance beyond the homogeneous ensemble model (Table 2). It shows that the ensemble technique improves model performance by adding one extraction feature (Canny and Otsu). Table 8 shows a fairly significant value of the heterogeneous ensemble model, exceeding the capabilities of the

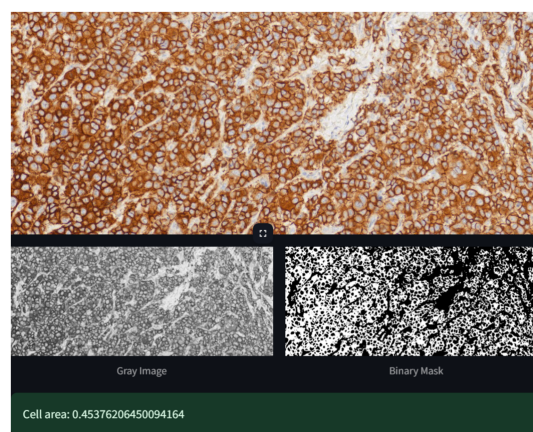


FIGURE 7

Cell quantifications as a result of feature extraction of Otsu edge detection (HER-2).

homogeneous ensemble model compared to the homogeneous values found in Tables 3–5. Technically, the homogeneous ensemble results of the single ConvNextTiny model show quite good performance, but assigning image labels to one technique has not accommodated the needs of pathologists based on their fields of science, which require observation of cell morphology. These observations are based on the

threshold of the number of cells and the brown area on the cell membrane. This is very important to provide confidence in determining the status characteristics of each sample or patch from WSI.

The heterogeneous ensemble model performs classification of ER, PR, Ki-67, and HER-2 with labeling and validation between the prediction and actual. The image shows that the characteristics of the number of cells will be identified based on the brown cycle, as the cell nucleus (the first three images from the left). In contrast, the rightmost image is the HER-2 image, marked as the distribution of cell membranes that dominate the patch area (square images), brown, and focuses on its area rather than on the nucleus. The model will visually compute the number of cell nuclei, which can be computed manually, although it takes an inefficient time. It is unlike the cell membrane, which is observed around the brown area and not the cycle contained in the patches.

The superiority of the heterogeneous ensemble model is also seen in the proportion of correctly recognized image samples. The higher the proportion, the higher the confidence in its classification performance, as shown in Table 8. Indeed, of course, it depends on the train's performance on all types of biomarkers, which greatly determine the performance of the test. Based on the values in Table 8, HER-2 outperforms the other three biomarkers because feature extraction of the brown cell area makes it easy to determine its

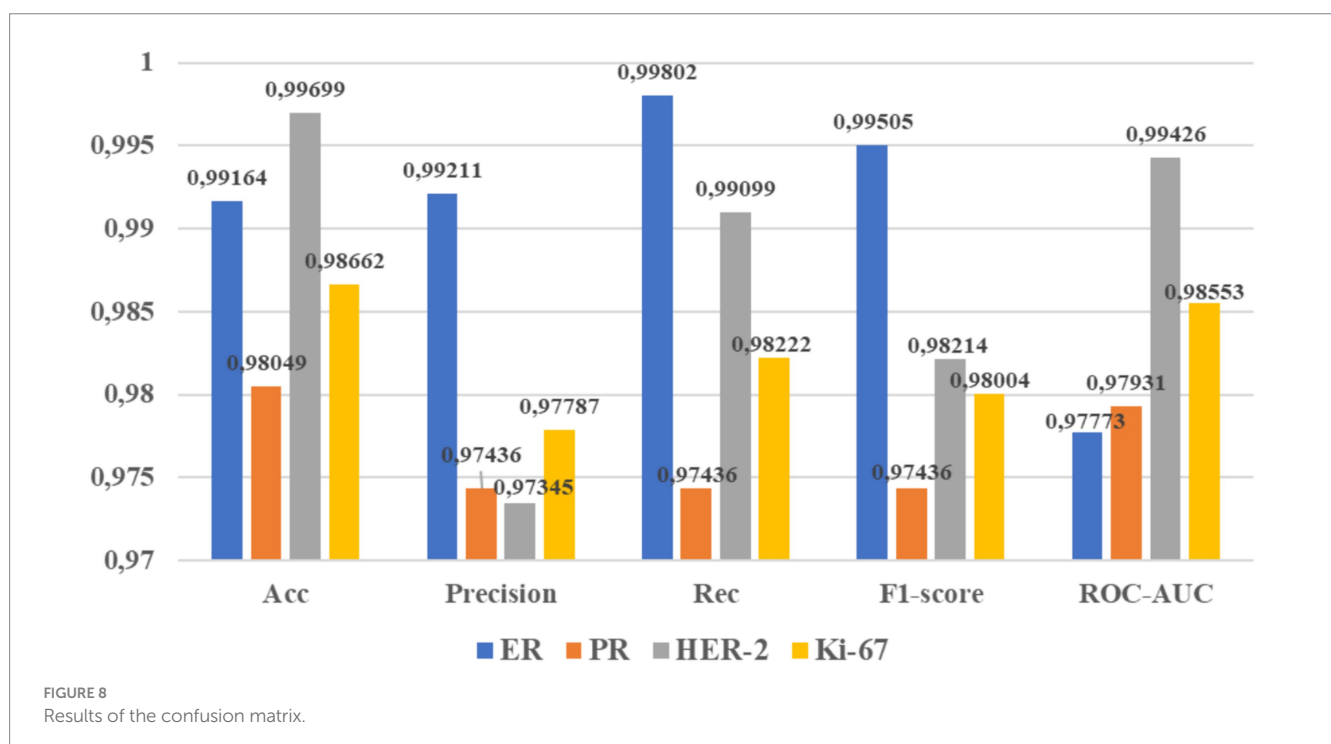
presence with certainty. In comparison, the other three biomarkers, ER, PR, and Ki-67, still need to compute the number of cells, which is not always certain for each image sample. Automated cell quantification makes it more difficult for the model to classify it, even though it already has a threshold number. ER and PR have a threshold of 1%, while HER-2 and Ki-67 have value thresholds of 10 and 20%, respectively. This superiority of HER-2 is indeed unique, even though each type has the same number of image samples. However, HER-2 is still superior for the previous reason.

This study is in line with previous studies using other model ensembles, in that the output of the ensemble model will reinforce the model before the ensemble so that the existence of this technique is significant enough to present a model applied to public service access in the medical field. It is shown in previous studies, including in Table 7.

Figure 8 presents the metrics for evaluating the model using a confusion matrix, including accuracy, precision, recall, F_1 -score, and ROC-AUC. The model achieved its highest accuracy of 0.99699 on HER-2, recall of 0.99802 on ER, F_1 -score of 0.99505 on ER, ROC-AUC of 0.99426 on HER-2, and precision of 0.9921 on ER. HER-2s get superior accuracy when they focus on the cell membrane area, which is influenced by the portion of the brown area in the image. HER-2 also has the highest number of valid predictions. However, accurately counting cancer cells remains challenging due to

TABLE 8 Performance of a heterogeneous ensemble model of ER, PR, HER-2, and Ki-67.

Images	Acc	Precision	Rec	F_1 -score	ROC-AUC
ER	0.99164	0.99211	0.99802	0.99505	0.97773
PR	0.98049	0.97436	0.97436	0.97436	0.97931
HER-2	0.99699	0.97345	0.99099	0.98214	0.99426
Ki-67	0.98662	0.97787	0.98222	0.98004	0.98553



the uncertain distribution of these cells within the image. Despite Ki-67 having the most extensive dataset, in our case, it does not surpass HER-2 for prediction percentage. The model struggles to differentiate between light, regular, and dark brown shades, leading it to rely on edge detection results, which often produce invalid classifications. Additionally, the smallest dataset size for PR contributes to its lower prediction values.

4.3 Research achievement

We conducted multiple scenarios to enhance the model we developed. As expected, these results were also achieved by previous researchers employing different methodologies. Table 9 compares our research findings with those of previous studies.

The performance comparison involves various variables, including different data sources, datasets, feature fusion, ensemble techniques, and models (Table 7). We analyze results across smaller and larger data clusters relative to our dataset, including datasets with restricted access. Our model demonstrates superior performance compared to studies with smaller datasets (Mudeng et al., 2023; Zheng et al., 2023; Abdullakutty et al., 2024; Alam et al., 2024; Islam et al., 2024; Qasrawi et al., 2024; Solorzano et al., 2024; Sreelekshmi and Nair, 2024). It also outperforms those with larger datasets (Khan et al., 2023; Kumari and Ghosh, 2023; Prezja et al., 2024). However, research by Ahmad and Alqurashi (2024) and Rahaman et al. (2024) achieves better results using ensembles of older models. Our approach leverages a new model that integrates feature fusion from handcrafted with ConvNextTiny features in the head block, offering efficient computation and simple edge detection.

Here, our feature fusion model has an average performance above 99%, including previous models (Alam et al., 2024). These features of each class will find their unique patterns and then be combined with the convolution features in the transfer learning model, so that it will produce a unique pattern if only using the convolution model. Feature fusion has provided a unique pattern to the model that produces better performance than without feature fusion, even though homogeneous ensemble learning is carried out. Feature fusion improves model performance on adequate datasets, bagging techniques, and robust models. The selection of models and feature techniques greatly determines the final performance of the model, so that the classification ability becomes a reliable and final result.

This research is an ongoing process of interpreting the molecular expression of breast cancer patients in immunohistochemistry examination. The expression still requires the characteristics of all biomarkers to determine the molecular subtypes that play a role in determining the enforced prognosis and type of therapy. Research by Fan et al. (2024) becomes a reference for the development of this research in the future with various feature engineering and models to obtain robust performance and contribute to efficient architecture in its deployment.

4.4 Limitations

This research endeavors to enhance the model performance utilized by pathologists during immunohistochemistry

examinations, which are concerned with the determination of cell morphology and cell proliferation, explicitly focusing on cell nuclei and membranes. This investigation leverages computer vision to ensure precise and accurate results, thereby facilitating reliable diagnostic outcomes. Nevertheless, various limitations must be acknowledged and addressed in this research context, including:

The fixation of the number of samples for each WSI does not represent the entire area because only three to four samples are taken for each WSI for each magnification. It will compute the number of cells in each WSI accurately and determine the average value and standard deviation of each WSI, not only based on each sample but also focusing on the mean and standard deviation of each WSI. It also overcomes overshooting and overlapping during training, which are still relatively high. It will certainly provide strong confidence in the results released by pathologists for hospitals.

The real-time model integration with the WSI display and scanner has not been configured, so the pointer changes or shifts can provide detection results for each WSI sample and magnification. The real-time configuration between the application and the microscope display hardware makes it easy to determine the final results of the examination, which are released visually.

5 Conclusion

We successfully implemented a heterogeneous ensemble learning model to address the problem of feature classification in anatomical pathology cases. Pathologists typically observe cell morphology to determine the malignant status of cancer, with cell quantification being the gold standard in their practice. This method computes the number of cells and the area of brown-stained cells to identify the cell nucleus and cell membrane. Our study selected the CNN types, ConvNextTiny model based on the majority voting results from four models experimented with cell quantification. Both ensembles demonstrated performance that significantly outperformed the other three single models. The result indicates that our approach, which concatenated a single feature into ConvNextTiny with the simplest structure, achieved performance superior to the more complex structures of the three models with a more significant number of features. Heterogeneous ensemble learning, which is a feature fusion, has significantly better performance on adequate datasets, ensemble techniques, and robust models, so its performance is better than that of homogeneous ensemble learning.

Further development of this study is necessary through more extensive patch exploration for each WSI image sample to achieve a more accurate and precise average. Each WSI sample has a fixed number of samples, and each WSI has its average and standard deviation identified to match the annotation results manually performed by pathologists. Additionally, the WSI shift results have been read, and the WSI status has been detected for each patch shift angle in real-time. This approach will undoubtedly produce an accurate analysis that matches the needs of pathologists and hospitals. The study aims to provide immunohistochemistry examination results that can effectively map the appropriate subtype of malignancy expression of breast cancer.

TABLE 9 Performance of the previous work.

Author	Dataset source	Number of data	Methods	Results
Mudeng et al. (2023)	BreakHis DataBiox	7,909 922	InceptionResNetV2; InceptionV3; NASNet-Large; ResNet50; ResNet101; VGG19; and Xception as single model Majority voting	Accuracy: 97.67% F_1 -score: 97.60%
Zheng et al. (2023)	P&D Laboratory	7,909	VGG16 + Xception + ResNet50 + DenseNet201 Weighted voting strategy	Accuracy: 98.90%
Khan et al. (2023)	ITMP; University of Bern RUMC	53,814	U-Net + ViT Segmentation and classifier	F_1 -score: 97.4% Sensitivity: 99.5% Specificity: 96.7%
Analysis: concatenate				
Kumari and Ghosh (2023)	IDC BreakHis	277,524 7,909	VGG-16 + Xception + DenseNet201	Accuracy: 94.2%
Sreelekshmi and Nair, 2024	MIAS	332	U-Net + Auto Encoder	Accuracy: 75.3%
Prezja et al. (2024)	NCT UMM	100,000	EfficientNet + Vision Transformer + Random Forest	Accuracy: 96.74%
Abdullakutty et al. (2024)	Electronic Medical Record	3,764	PCA + auto encoder; VGG-16; ViT; and ResNet-50	Accuracy: 78.84%
Parshionikar and Bhattacharyya (2024)	BreakHis IR Thermal Image Dataset	9,713 1,279	Inception + CapsNet	Accuracy: 99.74%
Rahaman et al. (2024)	In-House	12,156	EfficientnetB3 + ResNet50 + SCL	Accuracy: 99.92% Precision: 99.88% Recall: 99.90% F_1 -score: 99.89%
Ahmad and Alqurashi (2024)	American Oncology Institute at Shrimann Hospital	1,935	ResNet50 + InceptionV3	Accuracy: 99.80% F_1 -score: 99% Sensitivity: 99% Specificity: 99%
Solorzano et al. (2024)	Clinseq Sos	355 284; total 2,502,649 tiles	Majority vote of Inc. Xception V3, ResNet50, Inception-ResNet V2 and Xception	Accuracy: 91.2% Dice: 86.2% Specificity: 85.9% Precision: 83.7%
Qasrawi et al. (2024)	HMUH	20,000	YOLO; VGG-16, DenseNet121	Accuracy: 88.9% Precision: 88.9% Recall: 88.7% F_1 -score: 88.8% AUC: 89.4%
Karuppasamy et al. (2024)	SQUH BreakHis	158 7,909	AlexNet + VggNet	AUC: 95%
Islam et al. (2024)	BUSI UDAIT	780 163	MobileNet + Xception	Accuracy: 87.82% Precision: 87.33% Recall: 85.33% F_1 -score: 86.00%
Alam et al. (2024)	BUSI	1,312	GAN + SVM + U-Net + VGG-19	Accuracy: 99.48% Sensitivity: 99.40% Specificity: 99.55%
Proposed model	Hasanuddin University Hospital Wahidin Sudirohusodo Hospital	23,154	Majority voting: Exception, ResNet50V2; InceptionResnet50V2; ConvNextTiny Ensemble (feature fusion): Canny/Otsu + ConvNextTiny for every Biomarker	Accuracy: 99.7% Precision: 97.35% Recall: 99.1% F_1 -score: 98.21% ROC-AUC: 99.43%

Data availability statement

Datasets from Hasanuddin University Hospital, Dr. Wahidin Sudirohusodo Hospital, and <https://www.kaggle.com/datasets/akbarnejad1991/ihc4bc-compressed/> were used for the study. The datasets from the two hospitals are private, while the third dataset is publicly available.

Ethics statement

The studies involving humans were approved by Prof. Dr. Muh Nasrum Massi, Ph.D., Sp.MK(K) Komite Etik Penelitian Kesehatan RSPTN Universitas Hasanuddin RSUP Dr. Wahidin Sudirohusodo Makassar. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

II: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. AK: Formal analysis, Project administration, Software, Visualization, Writing – review & editing. SH: Data curation, Formal analysis, Investigation, Project administration, Visualization, Writing – review & editing. BN: Conceptualization, Data curation, Formal analysis, Project administration, Supervision, Writing – review & editing. DS: Conceptualization, Formal analysis, Software, Validation, Writing – review & editing. AY: Data curation, Formal analysis, Supervision, Project administration, Writing – review & editing. RP: Data curation, Formal analysis, Validation, Visualization, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the Key Research and Development Project, Ministry of Higher Education, Science and Technology of the Republic of Indonesia (111/E5/PG.02.00.PL/2024), Special Funds at the Fundamental Research Scheme to Dipa Makassar University (103/UNDIPA/G.4/VI/2024).

References

- Abdulkutty, F., Akbari, Y., Al-Maadeed, S., Bouridane, A., Talaat, I. M., and Hamoudi, R. (2024). Towards improved breast cancer detection via multi-modal fusion and dimensionality adjustment. *Comput. Struct. Biotechnol. J.* 1:100019. doi: 10.1016/j.csbr.2024.100019
- Afkari, H., Makrufardi, F., Hidayat, B., Budiawan, H., and Sundawa Kartamihardja, A. H. (2021). Correlation between ER, PR, HER-2, and Ki-67 with the risk of bone metastases detected by bone scintigraphy in breast cancer patients: a cross sectional study. *Ann. Med. Surg.* 67:102532. doi: 10.1016/j.amsu.2021.102532
- Ahmad, I., and Alqurashi, F. (2024). Early cancer detection using deep learning and medical imaging: a survey. *Crit. Rev. Oncol. Hematol.* 204:104528. doi: 10.1016/j.critrevonc.2024.104528
- Alam, M. N. A., Mohi Uddin, K. M., Rahman, M. M., Manu, M. M. R., and Nasir, M. K. (2024). A novel automated system to detect breast cancer from ultrasound images using deep fused features with super resolution. *Intell. Based Med.* 10:100149. doi: 10.1016/j.ibmed.2024.100149
- Alismail, H. (2024). Review: merging from traditional to potential novel breast cancer biomarkers. *J. King Saud Univ. Sci.* 36:103551. doi: 10.1016/j.jksus.2024.103551
- Allison, K. H., Hammond, M. E. H., Dowsett, M., McKernin, S. E., Carey, L. A., Fitzgibbons, P. L., et al. (2021). Estrogen and progesterone receptor testing in breast cancer: ASCO/CAP guideline update special articles abstract. *J. Clin. Oncol.* 38, 1346–1366. doi: 10.1200/JCO.19.02309

Acknowledgments

The authors would like to thank the Ministry of Higher Education, Science and Technology of the Republic of Indonesia for funding, the Rector and Head of the Research and Development Agency of Dipa Makassar University for their support in this research until it could be published, and also to Hasanuddin University Hospital and Dr. Wahidin Sudirohusodo Hospital for their data support and expert consultation.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Correction note

This article has been corrected with minor changes. These changes do not impact the scientific content of the article.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fcomp.2025.1569017/full#supplementary-material>

- Asif, S., Yi, W., Ain, Q. U., Hou, J., Yi, T., and Si, J. (2022). Improving effectiveness of different deep transfer learning-based models for detecting brain tumors from MR images. *IEEE Access* 10, 34716–34730. doi: 10.1109/ACCESS.2022.3153306
- Aswathy, M. A., and Jagannath, M. (2017). Detection of breast cancer on digital histopathology images: present status and future possibilities. *Inform. Med. Unlocked* 8, 74–79. doi: 10.1016/j.imu.2016.11.001
- Bychkov, D., Joensuu, H., Nordling, S., Tiulpin, A., Kückel, H., Lundin, M., et al. (2022). Outcome and biomarker supervised deep learning for survival prediction in two multicenter breast cancer series. *J. Pathol. Inform.* 13:100171. doi: 10.4103/jpi.jpi_29_21
- Carrington, A. M., Manuel, D. G., Fieguth, P. W., Ramsay, T., Osmari, V., Wernly, B., et al. (2023). Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45, 329–341. doi: 10.1109/TPAMI.2022.3145392
- Chadha, G. K., Srivastava, A., Singh, A., Gupta, R., and Singla, D. (2020). An automated method for counting red blood cells using image processing. *Procedia Comput. Sci.* 167, 769–778. doi: 10.1016/j.procs.2020.03.408
- Chen, H., Gui, X., Zhou, Z., Su, F., Gong, C., Li, S., et al. (2024). Distinct ER and PR expression patterns significantly affect the clinical outcomes of early HER2-positive breast cancer: a real-world analysis of 871 patients treated with neoadjuvant therapy. *Breast* 75:103733. doi: 10.1016/j.breast.2024.103733
- Chicco, D., and Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, 16, 1–23. doi: 10.1186/s13040-023-00322-4
- Fan, L., Liu, J., Ju, B., Lou, D., and Tian, Y. (2024). A deep learning based holistic diagnosis system for immunohistochemistry interpretation and molecular subtyping. *Neoplasia* 50:100976. doi: 10.1016/j.neo.2024.100976
- Fei, F., Siegal, G. P., and Wei, S. (2021). Characterization of estrogen receptor-low-positive breast cancer. *Breast Cancer Res. Treat.* 188, 225–235. doi: 10.1007/s10549-021-06148-0
- Grimm, S. L., Hartig, S. M., and Edwards, D. P. (2016). Progesterone receptor signaling mechanisms. *J. Mol. Biol.* 428, 3831–3849. doi: 10.1016/j.jmb.2016.06.020
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity mappings in deep residual networks, European Conference on Computer Vision. Springer: Cham. 630–645.
- Intan, I., Nelwan, B. J., Henry, M. M., Karnyoto, A. S., Puspitasari, R. E., and Pardamean, B. (2024). Deep ensemble transfer learning for detecting breast cancer in histopathological images. *Commun. Math. Biol. Neurosci.* 2024:101. doi: 10.28919/cmbn/8823
- Intan, I., Nurdin, and Pangerang, F. (2023). Facial recognition using multi edge detection and distance measure. *IAES Int. J. Artif. Intell.* 12, 1330–1342. doi: 10.11591/ijai.v12.i3.pp1330-1342
- Islam, M. R., Rahman, M. M., Ali, M. S., Nafi, A. A. N., Alam, M. S., Godder, T. K., et al. (2024). Enhancing breast cancer segmentation and classification: an ensemble deep convolutional neural network and U-Net approach on ultrasound images. *Mach. Learn. Appl.* 16:100555. doi: 10.1016/j.mlwa.2024.100555
- Joensuu, K., Leidenius, M., Kero, M., Andersson, L. C., Horwitz, K. B., and Heikkilä, P. (2013). ER, PR, HER2, Ki-67 and CK5 in early and late relapsing breast cancer-reduced CK5 expression in metastases. *Breast Cancer* 7, 23–34. doi: 10.4137/BCBCR.S10701
- Karuppasamy, A. D., Abdesselam, A., zidoum, H., Hedjam, R., and al-Bahri, M. (2024). Combining a forward supervised filter learning with a sparse NMF for breast cancer histopathological image classification. *Intell. Based Med.* 10:100174. doi: 10.1016/j.ibmed.2024.100174
- Kemenkes. (2024). Kanker Payudara, Kemenkes. Available online at: <https://ayosehat.kemkes.go.id/topik-penyakit/neoplasma/kanker-payudara>. (Accessed December 22, 2024)
- Khan, A., Brouwer, N., Blank, A., Müller, F., Soldini, D., Noske, A., et al. (2023). Computer-assisted diagnosis of lymph node metastases in colorectal cancers using transfer learning with an ensemble model. *Mod. Pathol.* 36:100118. doi: 10.1016/j.modpat.2023.100118
- Kildal, W., Cyll, K., Kalsnes, J., Islam, R., Julbø, F. M., Pradhan, M., et al. (2024). Deep learning for automated scoring of immunohistochemically stained tumour tissue sections—validation across tumour types based on patient outcomes. *Heliyon* 10:e32529. doi: 10.1016/j.heliyon.2024.e32529
- Krishna, S., Suganthi, S. S., Bhavsar, A., Yesodharan, J., and Krishnamoorthy, S. (2023). An interpretable decision-support model for breast cancer diagnosis using histopathology images. *J. Pathol. Inform.* 14:100319. doi: 10.1016/j.jpi.2023.100319
- Kumari, V., and Ghosh, R. (2023). A magnification-independent method for breast cancer classification using transfer learning. *Healthc. Anal.* 3:100207. doi: 10.1016/j.health.2023.100207
- Li, L., Yang, Z., Yang, X., Li, J., and Zhou, Q. (2023). PV resource evaluation based on Xception and VGG19 two-layer network algorithm. *Heliyon* 9:e21450. doi: 10.1016/j.heliyon.2023.e21450
- Loggie, J., Barnes, P. J., Carter, M. D., Rayson, D., and Bethune, G. C. (2024). Is Oncotype DX testing informative for breast cancers with low ER expression? A retrospective review from a biomarker testing referral center. *Breast* 75:103715. doi: 10.1016/j.breast.2024.103715
- Lv, Q., Meng, Z., Yu, Y., Jiang, F., Guan, D., Liang, C., et al. (2016). Molecular mechanisms and translational therapies for human epidermal receptor 2 positive breast cancer. *Int. J. Mol. Sci.* 17:2095. doi: 10.3390/ijms17122095
- Martínez Pérez, J. A., and Pérez Martín, P. S. (2023). La curva ROC. *Medicina de Familia. SEMERGEN*, 49:101821. doi: 10.1016/j.semerg.2022.101821
- Mudeng, V., Farid, M. N., Ayana, G., and Choe, S. W. (2023). Domain and histopathology adaptations-based classification for malignancy grading system. *Am. J. Pathol.* 193, 2080–2098. doi: 10.1016/j.ajpath.2023.07.007
- Parshionikar, S., and Bhattacharyya, D. (2024). An enhanced multi-scale deep convolutional orchard capsule neural network for multi-modal breast cancer detection. *Healthc. Anal.* 5:100298. doi: 10.1016/j.health.2023.100298
- Penault-Llorca, F., and Radosevic-Robin, N. (2017). Ki67 assessment in breast cancer: an update. *Pathology* 49, 166–171. doi: 10.1016/j.pathol.2016.11.006
- Prezja, F., Annala, L., Kiiskinen, S., Lahtinen, S., Ojala, T., Ruusuvaari, P., et al. (2024). Improving performance in colorectal cancer histology decomposition using deep and ensemble machine learning. *Heliyon* 10:e37561. doi: 10.1016/j.heliyon.2024.e37561
- Qasrawi, R., Daraghme, O., Qdaih, I., Thwib, S., Vicuna Polo, S., Owienah, H., et al. (2024). Hybrid ensemble deep learning model for advancing breast cancer detection and classification in clinical applications. *Heliyon* 10:e38374. doi: 10.1016/j.heliyon.2024.e38374
- Rahaman, M. M., Millar, E. K. A., and Meijering, E. (2024). Generalized deep learning for histopathology image classification using supervised contrastive learning. *J. Adv. Res.* 1–16. doi: 10.1016/j.jare.2024.11.013
- Rahimzadeh, M., and Attar, A. (2020). A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2. *Inform. Med. Unlocked* 19:100360. doi: 10.1016/j.imu.2020.100360
- Reinert, T., Cascell, F., Resende, C. A. A., Gonçalves, A. C., Godo, V. S. P., and Barrios, C. H. (2022). Clinical implication of low estrogen receptor (ER-low) expression in breast cancer. *Front. Endocrinol.* 13:1015388. doi: 10.3389/fendo.2022.1015388
- Shao, Y., Guan, H., Luo, Z., Yu, Y., He, Y., Chen, Q., et al. (2024). Clinicopathological characteristics and value of HER2-low expression evolution in breast cancer receiving neoadjuvant chemotherapy. *Breast* 73:103666. doi: 10.1016/j.breast.2023.103666
- Sharma, S., and Kumar, S. (2022). The Xception model: a potential feature extractor in breast cancer histology images classification. *ICT Express* 8, 101–108. doi: 10.1016/j.icte.2021.11.010
- Solorzano, L., Robertson, S., Acs, B., Hartman, J., and Rantalainen, M. (2024). Ensemble-based deep learning improves detection of invasive breast cancer in routine histopathology images. *Heliyon* 10:e32892. doi: 10.1016/j.heliyon.2024.e32892
- Sreelekshmi, V., and Nair, J. J. (2024). Variational auto encoders for improved breast cancer classification. *Procedia Comput. Sci.* 233, 801–811. doi: 10.1016/j.procs.2024.03.269
- Tafavvoghi, M., Sildnes, A., Rakaee, M., Shvetsov, N., Bongo, L. A., Busund, L. R., et al. (2024). Deep learning-based classification of breast cancer molecular subtypes from H & E whole-slide images. *J. Pathol. Inform.* 16:100410. doi: 10.1016/j.jpi.2024.100410
- Talukder, M. A., Islam, M. M., Uddin, M. A., Akhter, A., Pramanik, M. A. J., Aryal, S., et al. (2023). An efficient deep learning model to categorize brain tumor using reconstruction and fine-tuning. *Expert Syst. Appl.* 230:120534. doi: 10.1016/j.eswa.2023.120534
- Tanvir, J., Mehedi, S. T., Paul, B. K., and Morshed, M. (2024). TrashNeXt: classification of recyclable water pollutants using deep transfer learning method. *Case Stud. Chem. Environ. Eng.* 11:101073. doi: 10.1016/j.csee.2024.101073
- Terven, J., Cordova-Esparza, D. M., Romero-González, J. A., Ramírez-Pedraza, A., and Chávez-Urbiola, E. A. (2025). A comprehensive survey of loss functions and metrics in deep learning. *Artificial Intelligence Review*, 58. doi: 10.1007/s10462-025-11198-7
- Wang, Z., Yin, Y., Xu, W., Mo, Y. K., Yang, H., Xiong, J., et al. (2025). Illuminating breast cancer malignancies: Lightweight histopathology computer vision classifier for precise breast cancer screening. *Eng. Medicine*, 2:100053. doi: 10.1016/j.engmed.2024.100053
- World Health Organization. (2022). Cancer today. Available online at: https://gco.iarc.who.int/today/en/dataviz/bars?mode=cancer&key=total&group_populations=1&types=1&sort_by=value0&populations=900&multiple_populations=0&values_position=out&cancers_h=39&sexes=2 (Accessed February, 2024).
- Yang, X., Zhao, J., Zhang, H., Dai, C., Zhao, L., Ji, Z., et al. (2022). Remote sensing image detection based on YOLOv4 improvements. *IEEE Access* 10, 95527–95538. doi: 10.1109/ACCESS.2022.3204053
- Zhao, S., Yan, C. Y., Lv, H., Yang, J. C., You, C., Li, Z. A., et al. (2024). Deep learning framework for comprehensive molecular and prognostic stratifications of triple-negative breast cancer. *Fundam. Res.* 4, 678–689. doi: 10.1016/j.fmr.2022.06.008
- Zheng, Y., Li, C., Zhou, X., Chen, H., Xu, H., Li, Y., et al. (2023). Application of transfer learning and ensemble learning in image-level classification for breast histopathology. *Intell. Med.* 3, 115–128. doi: 10.1016/j.imed.2022.05.004



OPEN ACCESS

EDITED BY

Sokratis Makrogiannis,
Delaware State University, United States

REVIEWED BY

Imran Iqbal,
Helmholtz Association of German Research
Centres (HZ), Germany
Abdus Sattar Mollah,
Military Institute of Science and Technology
(MIST), Bangladesh

*CORRESPONDENCE

Feng Tian,
✉ tf_0145@163.com

[†]These authors have contributed equally
to this work

RECEIVED 25 March 2025

ACCEPTED 22 August 2025

PUBLISHED 26 September 2025

CITATION

Zhang C, Zhang D, Tang L and Tian F (2025)
Research on a machine learning method for
predicting discharge time of thyroid cancer
patients receiving ¹³¹I treatment: a
retrospective study.
Front. Physiol. 16:1599657.
doi: 10.3389/fphys.2025.1599657

COPYRIGHT

© 2025 Zhang, Zhang, Tang and Tian. This is
an open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with
these terms.

Research on a machine learning method for predicting discharge time of thyroid cancer patients receiving ¹³¹I treatment: a retrospective study

Chen Zhang^{1†}, Dandan Zhang^{1,2,3†}, Lijun Tang¹ and Feng Tian^{1*}

¹Department of Nuclear Medicine, Jiangsu Province Hospital, The First Affiliated Hospital with Nanjing Medical University, Nanjing, China, ²Department of Operations Management, Jiangsu Province Hospital, The First Affiliated Hospital with Nanjing Medical University, Nanjing, China, ³China Hospital Reform and Development Research Institute of Nanjing University, Nanjing Drum Tower Hospital, Nanjing, China

Radioactive iodine-¹³¹I based internal irradiation therapy has become one of the main methods for treating thyroid cancer, but patient usually need to be hospitalized after taking ¹³¹I until the residual activity meets the discharge criteria. However, the complex metabolism of ¹³¹I drug in individualized patient may make it difficult to assess when patients would meet discharge criteria, thereby increasing the hospital stay. In this study, some basic data of 1,044 thyroid cancer patients received ¹³¹I treatment at the First Affiliated Hospital with Nanjing Medical University from January 2022 to January 2024 are collected. Numerical analysis methods are used to analyze the absorption and metabolism of ¹³¹I drug in different patients and support vector machine (SVM) model is used to predict the discharge time of different patients. Results show that the effective half-life of ¹³¹I in both male and female patients are 10.35 h and 9.64 h, whose residual activity less than 400 MBq after 48 h of taking ¹³¹I. While the effective half-life of ¹³¹I in both male and female patients are 14.07 and 13.47 h for that the residual activity are greater than 400 MBq after 48 h of taking ¹³¹I. Furthermore, a discharge time prediction method based on SVM has been developed and the accuracy and precision of this method in predicting whether a patient could be discharged from the hospital after 48 h of taking the ¹³¹I drug are 88.04% and 96.89%. These results show that the discharge time prediction method could be expected to improve the rotation efficiency of nuclear medicine wards and provide timely treatment for more thyroid cancer patients receiving ¹³¹I treatment in the future.

KEYWORDS

¹³¹I, thyroid cancer, effective half-life, support vector machine, discharge time

Introduction

According to the GLOBOCAN 2022 database of cancer incidence and mortality by the WHO International Agency for Research on Cancer, thyroid cancer is ranking

in seventh place for incidence (Bray et al., 2024). Thyroid cancer can be treated through surgical removal of the thyroid gland (Nguyen et al., 2015). However, some effects would have an impact on the choice of surgery, such as the size and grading of tumors et al. Studies have shown that the thyroid cancer may recur when surgical resection is incomplete (CSJGs, 2015; Sippel and Chen, 2009). Moreover, the patient's quality of life after surgery of thyroid gland may also be affected (Dogan et al., 2017; Singer et al., 2012). Recent years, due to the high uptake of iodine by the thyroid gland, the utilization of radioactive iodine-131 drugs (i.e., [¹³¹I]NaI) has become one of the main methods for the treatment of thyroid cancer (Verburg et al., 2020; Giovannella et al., 2023).

Since the maximum energy of the β^- ray released by ¹³¹I is only 606.5 keV, ¹³¹I could only deposit energy in its accumulated tissue, ultimately achieving the killing of tumor cells (Al-jubeh et al., 2012; Kim et al., 2019). Considering the physical half-life of ¹³¹I (i.e., 8.02 d) and the dose deposited in tumor areas, the activity typically used to treat thyroid cancer ranges from 100 to 200 mCi in clinical (Gao et al., 2024; Tran et al., 2022; Nguyen et al., 2024). As a result of the thyroid cancer patients received the ¹³¹I treatment may have a radiation dose impact on the public, and different countries and regions have set some discharge criteria to protect the public (Venencia et al., 2002; Tunçel et al., 2016). For example, in Argentina, the Nuclear Regulatory Authority has stipulated that the patient received ¹³¹I treatment needed to be hospitalized in special wards for 2 or 3 d if patients receive a dose >1 GBq (30 mCi), or if the emitting radiation dose rate is >50 μ Sv/h at 1 m. The US Nuclear Regulatory Commission regulatory guide (No. 8.39) allows the release of differentiated thyroid carcinomas patients based on a measured dose rate of 7 mR/h at 1 m (Shahhosseini et al., 2004). According to the GB 18871-2002 and GBZ 120-2020 formulated by relevant departments in China, patients can be discharged from the hospital when the residual activity less than 400 MBq (Zhou et al., 2019; Jin et al., 2018).

When a thyroid cancer patient is admitted to the nuclear medicine departments, doctors would mainly analyze the patient's condition to determine the administration activities of ¹³¹I. However, it is difficult to estimate when the patient would meet the discharge criteria after taking the ¹³¹I. Zahra, et al. studied the radiation exposure rate of 100 patients who were treated with 3.7, 5.5 or 7.4 GBq of ¹³¹I, the exposure rates after each of the three first days of hospitalization were 30, 50 and 70 μ Sv/h at 1 m, and all patients had an acceptable dose rate on days 2 and 3 that allowed their hospital discharge (Azizmohammadi et al., 2013). Sometimes, patients could meet the discharge criteria after 48 h of taking ¹³¹I drug, but due to the complex absorption or metabolic abilities of ¹³¹I drug in patients, residual activity of ¹³¹I drug in some patients may exceed the discharge criteria after 48 h. In order to ensure the radiation safety, the hospitalization duration for different thyroid cancer patients is often fixed, resulting in some patients being unable to leave the hospital even after meeting the discharge criteria. Such behavior not only wastes medical resources and leads to some patients being unable to receive timely treatment to some extent, but also imposes unnecessary financial burdens on them. Considering

the reality that the incidence of thyroid cancer in China has been increasing year by year in recent years, more and more patients would receive ¹³¹I treatment. However, the inability to accurately assess patient's discharge time in advance has led to a waste of medical resources, making it challenging to ensure patients' safety.

If the discharge time for patients receiving ¹³¹I treatment could be accurately assessed before or during their hospitalization, it would not only curtail their hospital stay and waiting period but also enhance the rotation efficiency of nuclear medicine wards. To achieve this goal, the new methods and technologies should be established. In recent years, artificial intelligence has made great progress in all fields of medicine. Related researches on disease diagnosis using patient's data combined with artificial intelligence has provided methods and ideas for solving the problem of predicting patient discharge time in the real world (Iqbal et al., 2021; Iqbal et al., 2022). In this study, the metabolism of [¹³¹I] NaI in different patients are analyzed, and a new technology for accurately predicting the discharge time of patient receiving ¹³¹I treatment based on the basic data of patients, the metabolism of ¹³¹I drug and machine learning algorithm is established and the performance are evaluated accordingly.

Materials and methods

Basic information of thyroid cancer patients selected in this study

In this study, some data of 1,044 thyroid cancer patients who received ¹³¹I treatment at the department of nuclear medicine of the First Affiliated Hospital with Nanjing Medical University from January 2022 to January 2024 are collected. The basic information of these patients could be found in Table 1, which consisted of number of cases, gender, age, ¹³¹I drug activity used.

Method and equipment for monitoring the residual activity of patients

After taking a sufficient amount activity of ¹³¹I drug orally at one time, the patient returned to the ward to wait for the ¹³¹I drug to be metabolized in the body until the residual activity in the body met the discharge criteria. During hospitalization, patients undergo residual activity monitoring every 24 h after taking ¹³¹I drug in a specific dose monitoring room. A real-time dosimeter is placed on the wall of the monitoring room, which produced by Shanghai Juyin Technology Co., Ltd. The basic information of the radiation monitoring system are shown in Table 2. Through calibration, the dosimeter can obtain the radiation activity of ¹³¹I through the dose rate measured. The patient is required to stand on the identification line for 1 min which is 1 m away from the dosimeter. The real-time reading of the dosimeter will be displayed on the remote monitoring device of the nurse station, and the staff will record the reading of the dosimeter when it is stable. By analyzing the residual activity of ¹³¹I drug in different patients at a specific time after taking the medicine, the metabolisms of ¹³¹I drug in different patients can be revealed.

Abbreviations: ¹³¹I, Radioactive iodine-131; SVM, Support vector machine; MC, Meet discharge conditions; NMC, Not meet discharge conditions.

TABLE 1 Basic information of thyroid cancer patients selected in this work.

Gender	Average age (years)	Median age (years)	Number of patients taking 100 mCi ¹³¹ I	Number of patients taking 150 mCi ¹³¹ I	Number of patients taking 200 mCi ¹³¹ I
Male	41.67 ± 12.72	39	47	345	16
Female	42.55 ± 12.83	42	106	507	23

TABLE 2 The basic information of radiation monitoring system.

Parameter	Value
Dose range	0.1 μSv/h ~ 10 mSv/h
Energy range	33 keV ~ 3 MeV
Sensitivity	~2.2 cps/μSv/h(@137Cs)
Dose rate linearity	≤ ±10%(0.1 μSv/h ~ 10 mSv/h)

Machine learning algorithms and evaluation parameters used in this study

In order to achieve more efficient rotation and utilization of nuclear medicine wards and ensure timely discharge of patients, machine learning algorithms could be used. Basing the previous comparison of the performance of different machine learning algorithms, a support vector machine (SVM) algorithm is used to analyze the relationship between when the patient can discharge and collected patients' data which consisted of gender, age, ¹³¹I drug activity used, and the metabolism of ¹³¹I drug (Suthaharan and SJMlm, 2016; WSJNb, 2006). The regularization parameter of SVM is set to 1. Patients with residual activity less than 400 MBq at 48 h after taking ¹³¹I drug are classified into one category (MC), while patients with residual activity greater than 400 MBq are classified into another category (NMC). 80% of patients are selected randomly as training data and rest 20% of patients as testing data. SMOTE is used to train and optimize the SVM model, and stratified sampling is adopted during train-test splitting to preserve the original class proportions in both training and testing sets. All the training and optimization involved in this study were done basing Python 3.11.3 (Chao et al., 2014; Markowetz et al., 2003). And quantitative parameters such as accuracy, precision, sensitivity, specificity, F1-score, AUC value are used to analyze the test results (Abjijoml, 2013).

Results

¹³¹I activity changes in thyroid cancer patients

For all patients selected in this study, NMC patient accounts for 16.08% of the total patients. NMC male patient accounts for 8.51% of the total patients, and NMC male patient accounts for 22.34% of all male patients. NMC female patient accounts for 7.57% of the

total patients, while accounts for 12.23% of all female patients. The residual activity changes in all thyroid cancer patients are shown in Figures 1, 2. The value of residual activity measured are fitted using the e-exponential function to obtain the effective half-life of ¹³¹I in different patients. The fitting results of male MC, male NMC, female MC and female NMC are shown in Formula 1–4. And the fitted values and measured values of the residual activity of ¹³¹I in patients after taking the ¹³¹I drug for a specific time are shown in Table 3. From these results, it can be seen that the residual activity attenuation of ¹³¹I drug in all patients follows the exponential attenuation law. The effective and biological half-life of ¹³¹I for different patients are calculated according to the measured results are shown in Table 4.

$$\text{Male MC} \quad \text{Activity} = 4761.04 \cdot e^{-\frac{t}{14.93}} + 38.23 (R^2 = 99.987\%) \quad (1)$$

$$\text{Male NMC} \quad \text{Activity} = 5665.07 \cdot e^{-\frac{t}{20.30}} + 30.77 (R^2 = 99.955\%) \quad (2)$$

$$\text{FemaleMC} \quad \text{Activity} = 5033.50 \cdot e^{-\frac{t}{13.91}} + 37.74 (R^2 = 99.999\%) \quad (3)$$

$$\text{FemaleNMC} \quad \text{Activity} = 5987.29 \cdot e^{-\frac{t}{19.44}} + 109.73 (R^2 = 99.995\%) \quad (4)$$

Iodine uptake rate in different thyroid cancer patients

In addition, 70 of all thyroid cancer patients would take iodine uptake rates test after 2, 6, and 24 h of taking ¹³¹I drug (male/NMC: 26/9, female/NMC: 44/5). The results are shown in Figures 3, 4. It can be seen that for MC patients, their iodine uptake rate shows a downward trend, while for NMC patients, it shows an upward trend.

Discharge time prediction of thyroid cancer patients based on support vector machine

There are significant individualized differences in the absorption and metabolism of ¹³¹I drug among thyroid cancer patients, which also would affect the rotation and utilization of nuclear medicine ward to some extent. Based on the patient information collected, and the metabolisms of ¹³¹I drug, the impact weights of patient information on the classification of MC and NMC are analyzed through feature engineering, and the results are shown in Figure 5. From the results, it can be seen that for the patient information

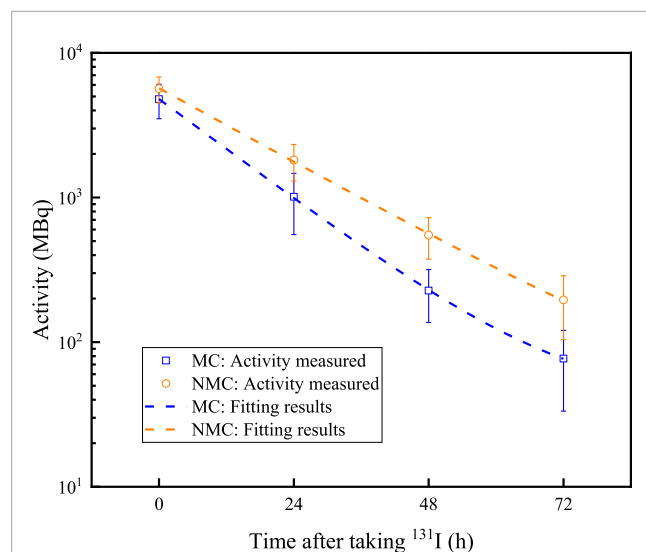


FIGURE 1
Residual activity decreases over time in male thyroid cancer patients. Dots represent the measured data, and the dashed lines represent the fitting results.

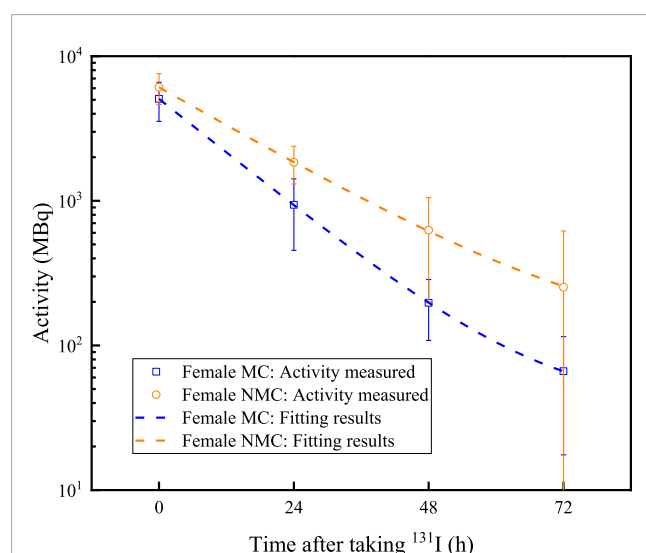


FIGURE 2
Residual activity decreases over time in female thyroid cancer patients. Dots represent the measured data, and the dashed lines represent the fitting results.

collected in this study, the patient's gender and the activity after 24 h of taking ^{131}I drug are key factors affecting whether the patient could meet the discharge criteria after 48 h of taking ^{131}I drug. Finally, based on the data collected, the SVM model is used to predict whether patients could meet the discharge criteria after 48 h of taking the ^{131}I drug. The ROC curve and the value of quantitative parameters are shown in [Figure 6](#) and [Table 5](#), respectively. It can be seen that the ROC curve with an AUC of 0.92, which demonstrates that the SVM classifier has good discrimination ability. Combined with the accuracy of 88.04%, precision of 96.89%, sensitivity of 88.64%, and an F1-score of 0.93, the SVM model could effectively

identify true positives while maintaining an acceptable false positive rate. These results show that based on the SVM algorithm and the metabolism of ^{131}I drug of individual patient collected, whether the patient could be discharged from the hospital after 48 h of taking ^{131}I can be accurately predicted.

Discussion

In recent years, the incidence of thyroid cancer has continued to increase, partly due to the increased incidence of diseases and also partly due to the improvement of diagnostic techniques. ^{131}I drug has become an important technology for the treatment of thyroid cancer. However, due to the presence of residual radioactivity in patients after taking the ^{131}I drug, patients generally required hospitalization until their residual activity meet the discharge criteria. Considering the large number of thyroid cancer patients and the shortage of medical resources, in order to ensure treatment efficiency and the radiation safety of the staff and public, it is necessary to clarify the patient's discharge status in advance to ensure the rotation of the nuclear medicine wards.

Some data of 1,044 thyroid cancer patients received ^{131}I treatment at the department of nuclear medicine of the First Affiliated Hospital with Nanjing Medical University from January 2022 to January 2024 are collected. Firstly, the metabolisms of ^{131}I drug of different patients are analyzed. The data shows that approximately 83.92% of patients met the discharge criteria, which the residual activity less than 400 MBq after 48 h of taking ^{131}I . The discharge rate of female patients is 87.77%, while that of male patients is only 77.66%. This may be due to the number of male patients collected in this study is much smaller than female patients, resulting in significant statistical errors. And it is also possible that male thyroid cancer patients have poorer metabolic capacity for ^{131}I than female patients. The residual activity in patient decreases exponentially with the time after taking ^{131}I drug. By fitting the activity profiles of MC and NMC patients, the effective half-life of ^{131}I in both male and female MC patients are 10.35 h and 9.64 h, while the effective half-life of male and female NMC patients are 14.07 h and 13.47 h, respectively. The longer biological half-life would be mainly due to delayed renal excretion of ^{131}I . In addition, these data further indicate that the male patients have a longer effective half-life of ^{131}I than female patients both of MC and NMC patients. In addition, there are also differences in the changes in iodine uptake rate in different patients. For NMC patients, the iodine uptake rate in 2 h, 6 h and 24 h after taking ^{131}I drug showed an upward trend, while in MC patients showed a downward trend. Iodine uptake rate can also provide a reference for assessing when a patient can be discharged to some extent.

In order to accurately determine whether patients could be discharged after 48 h of taking ^{131}I , a machine learning prediction method based on SVM has been developed with patient data collected and the metabolisms of ^{131}I drug. The results show that the accuracy and precision of this method to determine whether a patient is MC or NMC are 88.04% and 96.89%, respectively. Considering that 83.92% of all patients are MC, these results indicate that the discharge time prediction method based on the metabolisms of ^{131}I drug and SVM established in this work can improve the accuracy of predicting patient discharge time, thereby providing

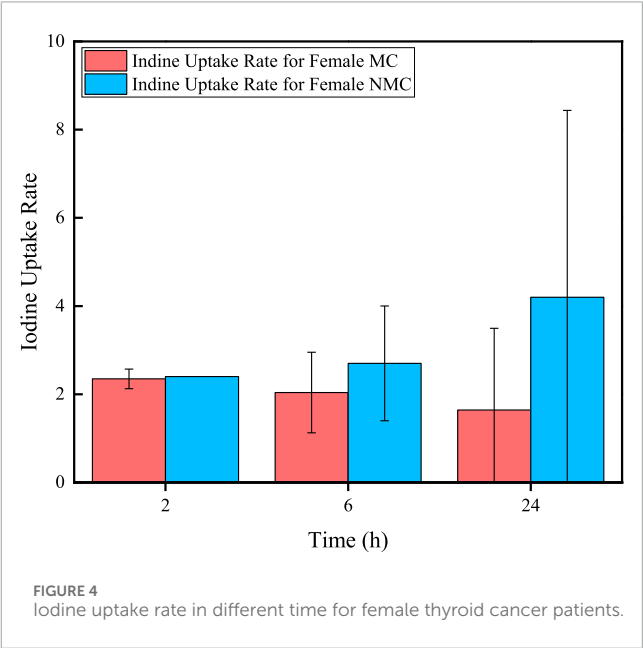
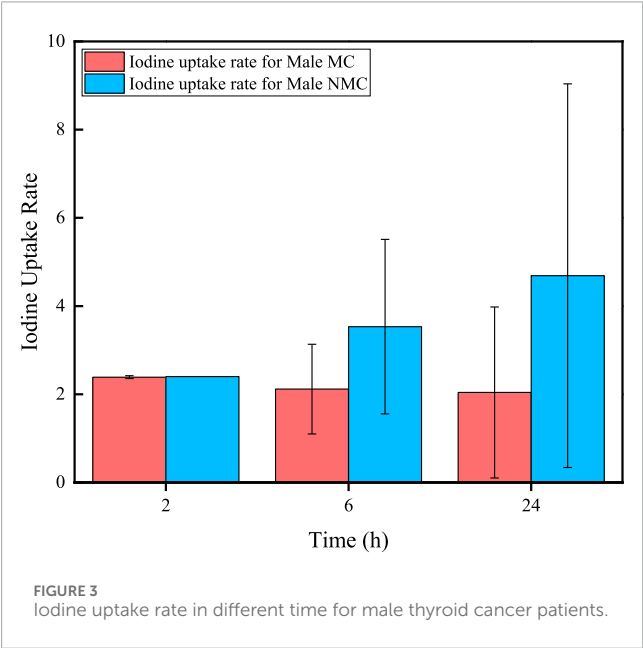
TABLE 3 The residual activity values in patient at different time points.

Category		0 h	24 h	48 h	72 h
Male MC	Measured value	4,780.65 (4,630.59, 4,930.70)	1,012.00 (958.39, 1,065.61)	227.37 (216.76, 237.97)	77.13 (72.03, 82.22)
	Fitted value	4,490.82	992.28	229.41	76.54
Female MC	Measured value	5,058.57 (4,926.63, 5,190.51)	937.24 (895.42, 979.06)	196.63 (188.97, 204.30)	66.14 (61.92, 70.36)
	Fitted value	4,722.08	934.24	197.41	66.18
Male NMC	Measured value	5,663.81 (5,409.41, 5,918.21)	1814.37 (1701.86, 1926.88)	550.36 (511.75, 588.97)	196.07 (175.87, 216.26)
	Fitted value	5,423.53	1767.59	563.25	194.02
Female NMC	Measured value	6,009.06 (5,661.10, 6,357.01)	1833.54 (1703.37, 1963.71)	619.63 (514.98, 724.29)	247.11 (157.21, 337.02)
	Fitted value	5,796.82	1851.79	616.61	257.21

(): The 95% confidence interval of the data.

TABLE 4 The effective and biological half-life of ¹³¹I for different patients.

Category	Average age (years)	Effective half-life (h)	Biological half-life (h)
Male MC	41.84 ± 14.51	10.35	10.94
Female MC	43.33 ± 16.42	9.64	10.15
Male NMC	42.17 ± 13.88	14.07	15.18
Female NMC	52.10 ± 33.59	13.47	14.49



technical support for the rotation and utilization of nuclear medicine wards to some extent. However, the patient information currently collected in this work is relatively limited, and more information of

patient, such as tumor staging, would be collected, and combined with the residual activity at more time points after taking the ¹³¹I drug, so as to achieve a rough assessment of the patient's discharge

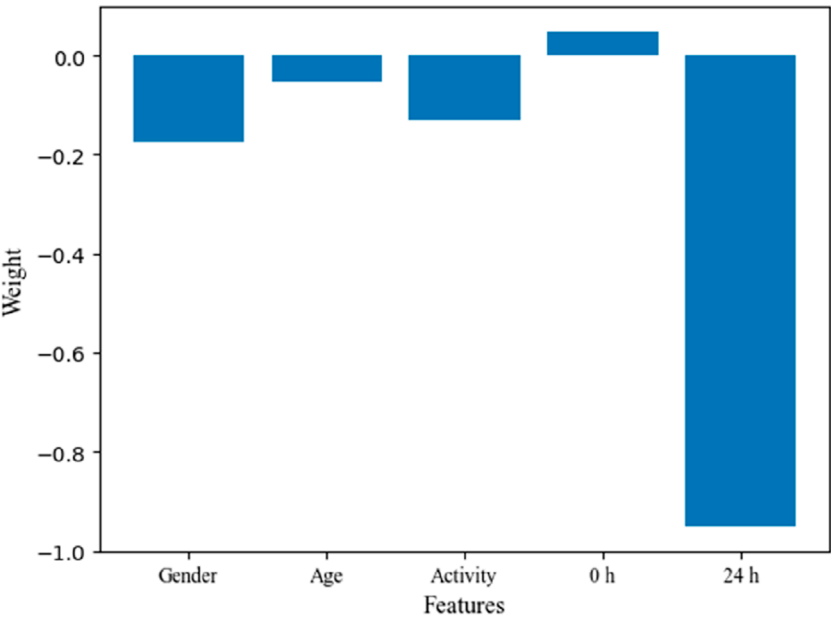


FIGURE 5
The weight of different features of patient about discharge judgment.

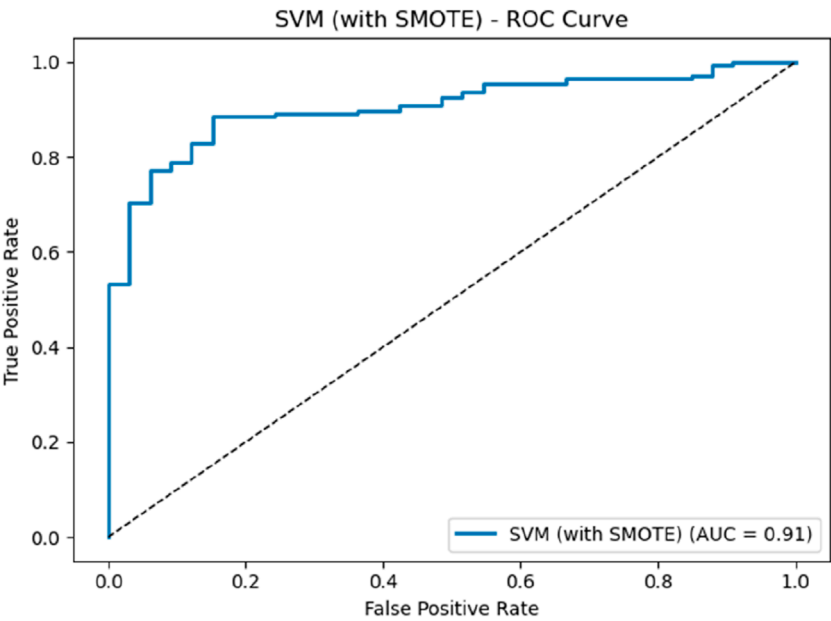


FIGURE 6
ROC curve for predicting whether patient can be discharged after 48 h of taking ¹³¹I.

TABLE 5 The value of quantitative evaluation parameters.

Parameter	Accuracy	Precision	Sensitivity	Specificity	F1-score	AUC
Value	88.04%	96.89%	88.64%	84.85%	0.93	0.91

time before the patient is hospitalized in the future. In addition, this study is currently a single-center study, and multi-center studies would be conducted in the future to explore the impact of more factors in the real-world on the performance of discharge time prediction method, so as to optimize the method to help more patients and more centers.

Conclusion

The number of new cases of thyroid cancer has been continuously increasing year by year, and ^{131}I drug has become an important technology for the treatment of thyroid cancer. However, due to individual differences in the absorption of ^{131}I drug by patients, there are also individual differences in when patients meet the discharge criteria after taking the ^{131}I drug. A discharge time prediction method basing the metabolisms of ^{131}I drug in different patients and the SVM algorithm is established in this study, and the results show that this method could be expected to improve the rotation efficiency of nuclear medicine wards and provide timely treatment for more patients in the future.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

CZ: Formal Analysis, Writing – original draft, Data curation. DZ: Software, Writing – review and editing. LT: Resources, Writing – review and editing, Investigation. FT: Funding acquisition, Writing – review and editing, Data curation, Methodology, Writing – original draft.

References

- Abjijoml, M. (2013). Comparative study on classification performance between support vector machine and logistic regression. *Int. J. Mach. Learn. Cybern.* 4, 13–24. doi:10.1007/s13042-012-0068-x
- Al-jubeh, W., Shaheen, A., and Zalloum, O. (2012). *Radioiodine I-131 for diagnosing and treatment of thyroid diseases*.
- Azizmohammadi, Z., Tabei, F., Shafiei, B., Babaei, A. A., Jukandan, S., Naghshine, R., et al. (2013). A study of the time of hospital discharge of differentiated thyroid cancer patients after receiving iodine-131 for thyroid remnant ablation treatment. *Hell. J. Nucl. Med.* 16 (2), 103–106. doi:10.1967/s002449910081
- Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., et al. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca. Cancer J. Clin.* 74 (3), 229–263. doi:10.3322/caac.21834
- Chao, C.-M., Yu, Y.-W., Cheng, B.-W., and Kuo, Y.-L. (2014). Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree. *J. Med. Syst.* 38, 106–107. doi:10.1007/s10916-014-0106-1
- Csjgs, G. (2015). Recurrence of papillary thyroid cancer after optimized surgery. *Gland. Surg.* 4 (1), 52–62. doi:10.3978/j.issn.2227-684X.2014.12.06
- Dogan, S., Sahbaz, N., Aksakal, N., Tural, F., Torun, B., Yildirim, N., et al. (2017). Quality of life after thyroid surgery. *J. Endocrinol. Invest.* 40, 1085–1090. doi:10.1007/s40618-017-0635-9
- Gao, X., Wu, B., Zhou, Q., Liu, Y., and Wang, R. J. B. (2024). Could effective iodine-131 half-life be extended by lithium carbonate in graves' disease patients: results from a retrospective analysis. *Biomol. Biomed.* 24 (6), 1711–1716. doi:10.17305/bb.2024.10659
- Giovannella, L., Garo, M. L., Campenni, A., Petranović Ovcariček, P., and Görges, R. J. C. (2023). Thyroid hormone withdrawal versus recombinant human TSH as preparation for I-131 therapy in patients with metastatic thyroid cancer: a systematic review and meta-analysis. *Cancers (Basel)*. 15 (9), 2510. doi:10.3390/cancers15092510
- Iqbal, I., Younus, M., Walayat, K., Kakar, M. U., and Ma, J. (2021). Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images. *Comput. Med. Imaging Graph.* 88, 101843. doi:10.1016/j.compmedimag.2020.101843
- Iqbal, I., Walayat, K., Kakar, M. U., and Ma, J. (2022). Automated identification of human gastrointestinal tract abnormalities based on deep convolutional neural network with endoscopic images. *Intell. Syst. Appl.* 16, 200149. doi:10.1016/j.iswa.2022.200149
- Jin, P., Feng, H., Ouyang, W., Wu, J., Chen, P., Wang, J., et al. (2018). Radiation dose rates of differentiated thyroid cancer patients after 131 I therapy. *Radiat. Environ. Biophys.* 57, 169–177. doi:10.1007/s00411-018-0736-7
- Kim, M., Bae, J. K., Hong, B. H., Kim, K. M., and Lee, W. (2019). Study on high density scintillators and multi-energy windows for improving I-131 gamma image quality: monte carlo simulation approach. 74, 305–311.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the National Natural Science Foundation of China (Grant No.12405387), the Natural Science Foundation of Jiangsu Province (Grant No.BK20241107), the Jiangsu Preventive Medicine Project (Grant Ym2023101), and Aid project of Jiangsu Ningai Medical Development and Medical Aid Foundation (Grant NDYG2024003).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Markowetz, F., Edler, L., and MjbjjoMMiB, V. (2003). Support vector machines for protein fold class prediction. *45*(3):377–389.
- Nguyen, Q. T., Lee, E. J., Huang, M. G., Park, Y. I., Khullar, A., Rajah, P., et al. (2015). Diagnosis and treatment of patients with thyroid cancer. *cancer* 8 (1), 30–40.
- Nguyen, N. C., Anigati, E. M., Desai, N. B., and Öz, O. K. (2024). “Radioactive iodine therapy in differentiated thyroid cancer: an update on dose recommendations and risk of secondary primary malignancies,” *Seminars in nuclear medicine* (Elsevier).
- Shahhosseini, S., Beiki, D., Dadashzadeh, S., Eftekhari, M., HjhjoNM, T., and Moosazadeh Rashti, G. (2004). Radiation dose rate and urinary activity in patients with differentiated thyroid carcinoma treated with radioiodine-131; a survey in Iranian population. *Hell. J. Nucl. Med.* 7 (3), 192–194.
- Singer, S., Lincke, T., Gamper, E., Bhaskaran, K., Schreiber, S., Hinz, A., et al. (2012). Quality of life in patients with thyroid cancer compared with the general population. *Thyroid* 22 (2), 117–124. doi:10.1089/thy.2011.0139
- Sippel, R. S., and Chen, H. J. T. (2009). Controversies in the surgical management of newly diagnosed and recurrent/residual thyroid cancer. *cancer* 19 (12), 1373–1380. doi:10.1089/thy.2009.1606
- Suthaharan, S., and Sjmml, S. (2016). *Learning afbdctwefe. Support vector machine*, 207–235.
- Tran, T.-V.-T., Rubino, C., Allodji, R., Andruccioli, M., Bardet, S., Diallo, I., et al. (2022). Breast cancer risk among thyroid cancer survivors and the role of I-131 treatment. *Br. J. Cancer* 127 (12), 2118–2124. doi:10.1038/s41416-022-01982-5
- Tunçel, N., Karayalçın, B., Özkan, G., EjjoR, B., and Therapy, R. (2016). The environmental dose measurements of high dose iodine-131 treated thyroid cancer patients during hospitalization period. 1 (2), 38–41. doi:10.15406/ijrrt.2016.01.00009
- Venencia, C. D., Germanier, A. G., Bustos, S. R., Giovannini, A. A., and EpjjoNM, W. (2002). Hospital discharge of patients with thyroid carcinoma treated with 131I. *J. Nucl. Med.* 43 (1), 61–65.
- Verburg, F. A., Flux, G., Giovanella, L., van Nostrand, D., Muylle, K., Mjejonm, L., et al. (2020). Differentiated thyroid cancer patients potentially benefitting from postoperative I-131 therapy: a review of the literature of the past decade. *Eur. J. Nucl. Med. Mol. Imaging* 47, 78–83. doi:10.1007/s00259-019-04479-1
- Wsjnb, N. (2006). What is a support vector machine? *Nat. Biotechnol.* 24 (12), 1565–1567. doi:10.1038/nbt1206-1565
- Zhou, Y., Yu, N., Ma, J., Yang, X., Chen, W., Shi, X., et al. (2019). Investigation and analysis on level of medical exposure in radiodiagnosis and radiotherapy in Jiangsu Province in 2016. *J. Public Health Emerg.* 3, 8. doi:10.21037/jpbe.2019.04.01

Frontiers in Computer Science

Explores fundamental and applied computer science to advance our understanding of the digital era

An innovative journal that fosters interdisciplinary research within computational sciences and explores the application of computer science in other research domains.

Discover the latest Research Topics

[See more →](#)

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact

