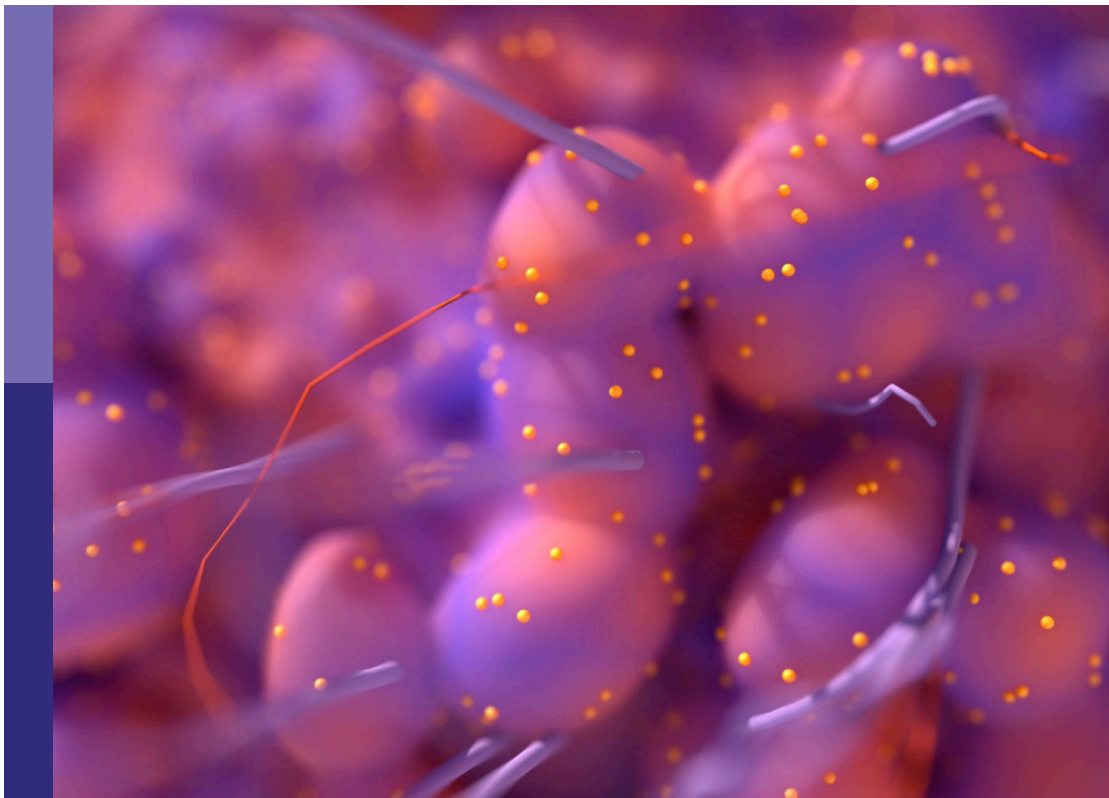
frontiers | Research Topics

Deep learning for medical imaging applications

Edited by
Monica Bianchini, Simone Bonechi, Paolo Andreini
and Sandeep Kumar Mishra

Published in
Frontiers in Oncology
Frontiers in Imaging



ISSN 1664-8714
ISBN 978-2-8325-7374-7
DOI 10.3389/978-2-8325-7374-7

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

Deep learning for medical imaging applications

Topic editors

Monica Bianchini — University of Siena, Italy
Simone Bonechi — University of Siena, Italy
Paolo Andreini — University of Siena, Italy
Sandeep Kumar Mishra — Yale University, United States

Citation

Bianchini, M., Bonechi, S., Andreini, P., Mishra, S. K., eds. (2026). *Deep learning for medical imaging applications*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-7374-7

Table of contents

frontiers | Frontiers in Imaging

TYPE Editorial
PUBLISHED 06 January 2026
DOI 10.3389/fimag.2025.1761718

OPEN ACCESS

EDITED AND REVIEWED BY
Alessandro Piva,
University of Florence, Italy

*CORRESPONDENCE
Simone Bonechi
✉ bonechi@diism.unisi.it
Sandeep Kumar Mishra
✉ sandeep.kumar@yale.edu

RECEIVED 05 December 2025
ACCEPTED 09 December 2025
PUBLISHED 06 January 2026

CITATION
Bonechi S, Bianchini M, Andreini P and Mishra SK (2026) Editorial: Deep learning for medical imaging applications.
Front. Imaging 4:1761718.
doi: 10.3389/fimag.2025.1761718

Editorial: Deep learning for medical imaging applications

Simone Bonechi[1]*, Monica Bianchini[1], Paolo Andreini[1] and Sandeep Kumar Mishra[2,3,4]*

[1]Department of Information Engineering and Mathematics, University of Siena, Siena, Italy, [2]Department of Radiology and Biomedical Imaging, Yale University, New Haven, CT, United States, [3]Yale Biomedical Imaging Institute, Yale University, New Haven, CT, United States, [4]Magnetic Resonance Research Center, Yale University, New Haven, CT, United States

KEYWORDS

artificial intelligence, cancer diagnosis, computed tomography (CT), deep learning, machine learning, magnetic resonance imaging (MRI), medical imaging, ultrasound (US)

Editorial on the Research Topic
Deep learning for medical imaging applications

There is substantial scientific interest in leveraging artificial intelligence (AI), particularly deep learning (DL), for radiological imaging, as these methods are driving significant advancements in disease detection, diagnostic accuracy, and treatment planning (Rubin, 2019). Over the past decade, annual publications on AI in radiology have surged seven-fold, with MRI and CT dominating the field of data acquisition techniques and neuroradiology leading in contributions, followed by musculoskeletal, cardiovascular, breast, urogenital, thoracic, and abdominal subspecialties (Pesapane et al., 2018). AI has evolved into numerous practical tools with significant clinical impact. Modern systems largely depend on artificial neural networks (ANNs) inspired by brain circuitry, including Convolutional Neural Networks (CNNs), recurrent models, and newer transformer architectures. These approaches achieve high performance across MRI, CT, PET, and ultrasound data, uncovering subtle diagnostic features beyond human perception and supporting earlier disease detection and more efficient clinical workflows (Perez-Lopez et al., 2024). As datasets grow and computational frameworks mature, DL continues to reshape the future of precision medicine. Ongoing challenges include model interpretability, generalizability, and unbiased clinical deployment, but the field is rapidly progressing toward robust, trustworthy, and clinically integrated AI systems (Yang et al., 2024). Despite strong research potential on AI, its real-world clinical deployment remains limited, as effective integration into healthcare requires coordinated efforts among stakeholders and careful resolution of ethical challenges (Yang et al., 2024; Saw and Ng, 2022).

Gabriel et al. explored the critical challenge of integrating AI into patient monitoring to support continuous, real-time clinical assessment. Developed by LookDeep Health, the system showed strong performance in object detection and patient-role classification. Their study demonstrated the feasibility of computer vision as a core technology for passive, uninterrupted patient monitoring within operational hospital environments. Performance evaluation showed high accuracy in both object detection and patient-role classification. Using this platform, the investigators compiled a substantial dataset comprising computer-vision, derived predictions from more than 300 high-risk fall patients, totaling over 1,000 monitored patient-days.

Abulajiang et al. explored important insights into the association between age at menopause and the risk of major gynecologic malignancies, including cervical, ovarian, and uterine cancers. Using restricted cubic spline (RCS) regression models, the study rigorously characterized non-linear relationships between menopausal age and subsequent cancer risk. The findings suggest that menopausal age may serve as a meaningful clinical indicator, with potential value in refining individualized cancer risk assessment and informing personalized screening strategies.

Chen, Han et al. conducted a systematic review and meta-analysis evaluating the prognostic significance of growth pattern-based grading in mucinous ovarian carcinoma (MOC). The analysis indicates that expansile MOC is associated with more favorable outcomes, whereas infiltrative MOC correlates with advanced disease and poorer prognosis. The findings further underscore the importance of complete surgical staging for infiltrative MOC, while suggesting that comprehensive staging may be optional in patients with early stage expansile MOC.

Li, Ding et al. investigated radiomic features derived from ultrasound imaging and developed an externally validated predictive model integrating clinical variables with ultrasound-based radiomics to assess residual tumor status in patients with advanced epithelial ovarian cancer. The combined model demonstrated superior performance in preoperatively identifying patients likely to achieve complete resection of all visible disease and exhibited stronger generalizability compared with models based solely on clinical or radiomic features.

Yang et al. presented a comprehensive review of recent advances in the application of AI for the early screening and diagnosis of thyroid diseases. The authors summarized progress across multiple domains, including thyroid pathology and ultrasound-based assessment, and highlight emerging trends in AI-driven clinical decision support. The review further emphasized the potential of integrated AI frameworks that combine ultrasound imaging with clinical data to improve diagnostic accuracy for thyroid cancer and to enable more precise prediction of postoperative survival outcomes.

Chen, Liu et al. introduced a novel multi-class brain tumor classification model, EnSLDe, designed to capture both short-range and long-range dependencies in neuroimaging data. The architecture comprised three key components: a Feature Extraction Module (FExM), a Feature Enhancement Module (FEnM), and a Classification Module. Evaluation on two publicly available datasets demonstrated excellent performance, underscoring the model's ability to effectively integrate multi-scale feature dependencies and thereby enhance brain tumor classification accuracy.

Ma et al. validated a DL signature for non-invasive prediction of spread through air spaces (STAS) in clinical stage I lung adenocarcinoma and compared its performance with a conventional clinical-semantic model. The Swin Transformer-based signature demonstrated superior predictive accuracy, outperforming traditional approaches. This end-to-end DL framework shows strong potential as a reliable tool for estimating STAS preoperatively, providing valuable guidance for surgical planning and supporting more informed decisions regarding adjuvant therapy selection in early-stage disease.

Han et al. developed a radiomics nomogram integrating chest CT features with the ILD-GAP index to improve clinical management of rheumatoid arthritis-associated interstitial lung disease (RA-ILD). CT scans were retrospectively analyzed and staged using ILD-GAP. The model demonstrated strong accuracy in identifying low-risk RA-ILD patients. These findings suggest that this non-invasive, quantitative tool may enhance clinical decision-making by enabling more precise risk stratification and supporting individualized management strategies for RA-ILD. This integrated approach offers added clinical value for patient care.

VanBerlo et al. investigated a self-supervised learning (SSL) approach to address the scarcity of labeled data in medical imaging by leveraging representations learned from unlabeled images. Their findings showed that constructing positive pairs from nearby frames within the same video improves performance compared with pairs derived from the same image, although optimal IVPP hyperparameters vary across downstream tasks. Notably, SimCLR consistently achieved top performance for key B-mode and M-mode lung ultrasound tasks, suggesting that contrastive learning may be better suited than non-contrastive methods for ultrasound imaging applications.

Saavedra et al. developed a novel two-step DL framework to automate the assessment of supraspinatus fatty infiltration in shoulder MRIs. Their method sequentially employs a U-Net architecture to segment the muscle's region of interest, followed by a VGG-19 network to perform binary classification based on Goutallier's scale. Utilizing transfer learning on a dataset of 606 T2-weighted images, the study reported robust performance, achieving a segmentation Dice score of 0.94 and a classification AUROC of 0.99. This approach demonstrates the feasibility of fully automating the diagnostic workflow, significantly reducing the reliance on time-consuming manual segmentation by radiologists.

Li, Chen et al. proposed UnetTransCNN, a hybrid architecture designed for 3D medical image segmentation that effectively integrates CNNs with Transformers. Addressing the limitations of prior sequential fusion models, UnetTransCNN employs a parallel design where a CNN-based module captures local details while a Transformer-based module, enhanced with an Adaptive Fourier Neural Operator, captures global contextual dependencies. The model utilizes adaptive global-local coupling units to dynamically fuse features across multiple scales. Validated on the BTCV and MSD datasets, UnetTransCNN demonstrated state-of-the-art performance, significantly outperforming existing hybrid baselines like TransUNet and CoTr in segmenting both large and small anatomical structures.

Rabah et al. introduced a Vision Transformer (ViT) framework for automated detection of diabetic peripheral neuropathy (DPN) using corneal confocal microscopy (CCM) images. To address the subjectivity and labor-intensive nature of manual assessment, they developed a streamlined ViT model that classifies images as healthy or DPN without requiring pixel-level segmentation. Using a dataset of 692 images, the model achieved state-of-the-art performance (AUC 0.99; F1-score 95%), outperforming CNNs such as ResNet50. Grad-CAM–based interpretability confirmed that the model accurately focused on corneal nerve fiber loss as the key discriminative feature.

Luo et al. introduced a DL-driven data-enhancement framework that sharpens the classification of endometrial lesions in ultrasound imaging. Drawing on 1,875 images from 734 patients across six hospitals, the team couples feature-space anomaly detection for image-quality cleaning with a clustering-based soften-label strategy. After benchmarking multiple CNNs and Vision Transformers, they assembled an ensemble of ResNet50, DenseNet169, DenseNet201, and ViT-B. This model delivers 0.809 accuracy and a 0.911 macro-AUC, markedly outperforming baseline CNNs and demonstrating how targeted data curation can meaningfully elevate diagnostic performance.

Liu et al. investigated the impact of AI-guided MRI instance segmentation on laparoscopic myomectomy, with particular focus on broad-ligament fibroids, which are challenging due to their proximity to critical pelvic anatomy. The DL model segmented fibroids, uterine wall, and uterine cavity on preoperative MRI. In a randomized cohort of 120 patients, AI assistance significantly reduced operative time (118 vs. 140 min), intraoperative blood loss (50 vs. 85 mL), and improved early postoperative recovery. The authors conclude that millimeter-level anatomical mapping can substantially enhance surgical precision in complex gynecologic procedures.

Xiong et al. explored the anticancer actions of 6-gingerol in SKOV3 ovarian carcinoma cells, revealing a targeted apoptotic mechanism. The compound suppressed clonogenic growth and triggers caspase-dependent apoptosis while selectively downregulating the transcription factor Gli3, independent of Bcl-2 family alterations. Notably, 6-gingerol robustly elevated miR-506, typically diminished in ovarian tumors and miR-506 overexpression itself reduces Gli3 and promotes apoptosis. Blocking miR-506 reversed these effects, supporting a model in which 6-gingerol activated a miR-506/Gli3 axis, highlighting its therapeutic promise.

Xie et al. conducted a systematic literature review, spanning the last decade, on the application of machine learning (ML) and DL techniques to psoriasis image analysis. Fifty-three articles were retrieved from major publication repositories (WoS, PubMed, and IEEE Xplore) addressing the problems of lesion localization, lesion recognition, and severity assessment. The authors evaluated commonly used public datasets, summarized prevailing ML/DL architectures and their limitations, and identified persistent challenges, including dataset heterogeneity and limited interpretability. They also outlined emerging trends and proposed future research directions to advance automated psoriasis assessment.

Chen, Shang et al. presented a case study of a patient with recurrent low-grade endometrial stromal sarcoma (LGESS) who refused standard surgery or ablative treatment. After discontinuing chemotherapy due to impaired liver function, the patient was administered high-intensity focused ultrasound (HIFU) together with chemotherapy, which resulted in a significant reduction in tumor volume, inhibition of its progression, and restoration of liver function. This result suggests that HIFU-based combination therapy may represent a valid option for metastatic LGESS or for patients unsuitable for surgery, particularly when integrated with real-time monitoring and precise post-treatment assessment.

Overall, this compilation demonstrates the researchers collectively push forward the development of advanced deep-learning models, reflecting their strong commitment to improving accuracy, reliability, and impact in medical imaging applications.

Author contributions

SB: Writing – original draft, Writing – review & editing. MB: Writing – original draft, Writing – review & editing. PA: Writing – original draft, Writing – review & editing. SKM: Writing – original draft, Writing – review & editing.

Acknowledgments

We are thankful to all the authors of the topics who contributed their research to this Research Topic and to the reviewers for their excellent assessment of all submissions.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) SM, MB, and SB declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declared that generative AI was not 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

References

Perez-Lopez R, Ghaffari Laleh N, Mahmood F, Kather JN. (2024) A guide to artificial intelligence for cancer researchers. *Nat Rev Cancer*. 24, 427–441. doi: 10.1038/s41568-024-00694-7

Pesapane F, Codari M, Sardanelli F. (2018) Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *Eur Radiol Exp*. 2:35. doi: 10.1186/s41747-018-0061-6

Rubin DL. (2019) Artificial intelligence in imaging: the radiologist's role. *J Am Coll Radiol*. 16:1309–17. doi: 10.1016/j.jacr.2019.05.036

Saw S N, and Ng K H. (2022) Current challenges of implementing artificial intelligence in medical imaging. *Physica Medica*. 100:12–7. doi: 10.1016/j.ejmp.2022.06.003

Yang Y, Zhang H, Gichoya J W, Dina Katabi D, Ghassemi M. (2024) The limits of fair medical imaging AI in real-world generalization. *Nat Med*. 30:2838–48. doi: 10.1038/s41591-024-03113-4

frontiers | Frontiers in Imaging

TYPE Original Research
PUBLISHED 10 March 2025
DOI 10.3389/fimag.2025.1547166

OPEN ACCESS

EDITED BY
Sandeep Kumar Mishra,
Yale University, United States

REVIEWED BY
Hina Sultana,
University of North Carolina System, United States
Nazreen Pallikkavaliyaveetil MohammedSheriff,
University of Michigan, United States

*CORRESPONDENCE
Paolo Gabriel
✉ paolo@lookdeep.health

RECEIVED 17 December 2024
ACCEPTED 18 February 2025
PUBLISHED 10 March 2025

CITATION
Gabriel P, Rehani P, Troy T, Wyatt T, Choma M and Singh N (2025) Continuous patient monitoring with AI: real-time analysis of video in hospital care settings.
Front. Imaging 4:1547166.
doi: 10.3389/fimag.2025.1547166

Continuous patient monitoring with AI: real-time analysis of video in hospital care settings

Paolo Gabriel*, Peter Rehani, Tyler Troy, Tiffany Wyatt, Michael Choma and Narinder Singh

Department of R&D, LookDeep Health, Oakland, CA, United States

Introduction: This study introduces an AI-driven platform for continuous and passive patient monitoring in hospital settings, developed by LookDeep Health. Leveraging advanced computer vision, the platform provides real-time insights into patient behavior and interactions through video analysis, securely storing inference results in the cloud for retrospective evaluation.

Methods: The AI system detects key components in hospital rooms, including individuals' presence and roles, furniture location, motion magnitude, and boundary crossings. Inference results are securely stored in the cloud for retrospective evaluation. The dataset, compiled with 11 hospital partners, includes over 300 high-risk fall patients and spans more than 1,000 days of inference. An anonymized subset is publicly available to foster innovation and reproducibility at lookdeep/ai-norms-2024.

Results: Performance evaluation demonstrates strong accuracy in object detection (macro F1-score = 0.92) and patient-role classification (F1-score = 0.98). The system reliably tracks the "patient alone" metric (mean logistic regression accuracy = 0.82 ± 0.15), enabling detection of patient isolation, wandering, and unsupervised movement-key indicators for fall risk and adverse events.

Discussion: This work establishes benchmarks for AI-driven patient monitoring, highlighting the platform's potential to enhance patient safety through continuous, data-driven insights into patient behavior and interactions.

KEYWORDS

artificial intelligence, medical imaging, computer vision, patient monitoring, RGB video, deep learning, healthcare analytics

1 Introduction

In hospitals, direct patient observation is limited–nurses spend only 37% of their shift engaged in patient care (Westbrook et al., 2011), and physicians average just 10 visits per hospital stay (Chae et al., 2021). This limited interaction hinders the ability to fully understand patient behaviors, such as how often patients are left alone, how much they move unsupervised, and how care allocation varies by time or condition. Virtual monitoring systems, which allow remote patient observation via audio-video devices, have improved safety, particularly for high-risk patients (Abbe and O'Keeffe, 2021).

Artificial Intelligence (AI) is transforming healthcare by enhancing diagnostic accuracy, streamlining data processing, and personalizing patient care (Davenport and Kalakota, 2019; Davoudi et al., 2019; Bajwa et al., 2021). While AI has found success in tasks like surgical assistance (Mascagni et al., 2022) and diagnostic imaging (Esteva et al., 2021), patient monitoring represents a critical frontier. Unlike these tasks, continuous patient monitoring involves real-time video analysis over extended periods, requiring AI systems

to process data efficiently and extract actionable insights spanning days, like day-over-day movement (Parker et al., 2022).

Continuous monitoring enhances safety and enables the detection of risks often missed during periodic assessments. For example, trends like delirium fluctuate throughout the day, but infrequent observations make these patterns hard to capture (Wilson et al., 2020). Similarly, patients occasionally leave their beds unattended—a key fall risk—yet monitoring every instance in real-time remains challenging. A robust computer vision-based system can provide immediate, context-aware insights into patient behavior (Chen et al., 2018), caregiver interactions (Avogaro et al., 2023), and room conditions (Haque et al., 2020). Such systems surpass traditional intermittent observation methods by detecting subtle patterns that inform care decisions (Lindroth et al., 2024).

However, achieving scalability, transparency, and adaptability in continuous monitoring systems presents significant challenges. These include efficiently processing video data at higher frame-rates (Posch et al., 2014), ensuring privacy compliance (Watzlaf et al., 2010), and adapting to dynamic hospital settings with varying lighting, camera angles, and patient behaviors. Addressing these technical and operational challenges is critical for AI-driven monitoring systems to gain acceptance and deliver meaningful outcomes, such as reducing falls and other preventable harms.

To bridge these gaps, this research presents a novel AI-driven system for continuous patient monitoring using RGB video (Figure 1), developed collaboratively with industry and healthcare providers. The LookDeep Health platform aims to enhance patient care by providing real-time monitoring and producing computer-vision-based insights into patient behavior, movement, and interactions with healthcare staff.

This study offers several key contributions:

1. **Implementation of advanced computer vision models:** our system utilizes state-of-the-art models for real-time predictions, including localization of people and furniture, monitoring boundary crossings, and calculating motion scores.
2. **Real-world validation:** we rigorously evaluated the system's performance in live hospital settings, illustrating its capability to present care providers with accurate data from continuous monitoring, and laying the foundation for future AI-enabled patient monitoring solutions.
3. **Dataset development:** we developed a comprehensive dataset encompassing over 300 high-risk fall patients tracked across 1,000 collective days and 11 hospitals, creating a valuable resource for studying patient behavior and hospital care patterns. This dataset is publicly available for further research at https://github.com/lookdeep/ai-norms-2024.

2 Methods

2.1 Study design

The LookDeep Health patient monitoring platform was deployed across 11 hospitals in three states within a single healthcare network. The system provides continuous, real-time monitoring of high-risk fall patients. Data collection adhered to institutional guidelines and patient consent procedures (see *Research Ethics*).

2.1.1 Participants

Patients monitored by LookDeep Health were primarily high-risk fall patients identified through mobility assessments as part of standard care protocols. This classification often results in the patient also being categorized as non-ambulatory during the inpatient stay (Capo-Lugo et al., 2023).

Data was organized into three subsets:

1. **Single-frame analysis:** periodic samples from monitoring sessions were used for training and testing object detectors, with over 40,000 frames collected to date. Only patients monitored during the first week of each month were included in the test set, providing 10,000 frames held out for consistent model evaluation.
2. **Observation logging:** ten patients who experienced falls were selected for additional annotation over a twelve month period (Figure 2A).
3. **Public dataset:** over 300 high-risk fall patients were monitored during a six month period, excluding those monitored for less than two days (Figure 2B).

As shown in Figure 3, data collection spanned multiple years, with each subset contributing to the development and validation of the AI system, with some overlap between subsets.

2.1.2 Patient monitoring system overview

The LookDeep Health monitoring system processes video through a computer vision pipeline to detect, classify, and analyze key elements within the patient's room, providing actionable insights to healthcare staff (Figure 4). Key components include:

1. **Video data capture and preprocessing:** video data is captured at 1 frame per second (fps) by LookDeep Video Unit (LVU) devices deployed in patient rooms (Figure 5A). Data is preprocessed to reduce bandwidth and enable efficient analysis.
2. **Object detection and localization:** a custom-trained model detects key objects ("person", "bed", "chair") and localizes them with bounding boxes.
3. **Person-role classification:** detected "person" objects are further classified as "patient", "staff", or "other" using the same object detector model, by augmenting labels with role-specific information.
4. **Motion estimation:** dense optical flow estimates motion between consecutive frames, enabling activity tracking in specific regions (e.g. scene, bed, safety zone).
5. **Logical predictions:** high-level predictions (e.g. "person alone", "patient supervised by staff") are derived by applying rules to detection and motion data, with a 5-second smoothing filter to mitigate detection errors.

Inference results, including object detections, role classifications, motion estimation, and logical predictions, are securely stored in a Google cloud database for further analysis (e.g.

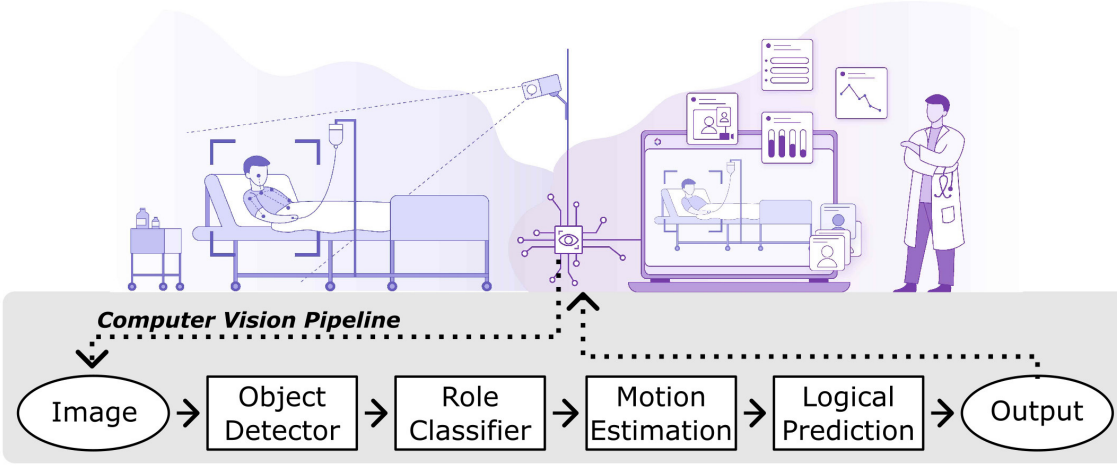


FIGURE 1
Illustrative workflow of the lookdeep health AI-driven patient monitoring platform. The system captures video from a hospital room using mounted cameras and processes each image through a series of computer vision modules. The output is presented as real-time insights for healthcare staff.

trend analysis). Anonymized frames are stored at regular intervals for quality assurance and model improvement.

2.1.3 Data anonymization

To ensure patient privacy in accordance with the Health Insurance Portability and Accountability Act (HIPAA) and institutional guidelines, all video data was processed to remove identifiable information. For training purposes, frames were face-blurred using a two-step procedure to maintain privacy while preserving relevant scene context:

1. **Manual labeling:** faces were manually labeled on fully-blurred images to create bounding boxes without exposing identifiable features.
2. **Local Gaussian blurring:** a strong Gaussian blur was applied to labeled facial regions, preserving scene context while anonymizing identities.

This approach was chosen to ensure privacy while balancing effective model training and validation. Additional obfuscation methods, such as pixelation or complete occlusion of faces, were considered but deemed not necessary for the intended use case. Data handling was conducted under a Business Associate Agreement (BAA) with participating hospitals.

2.2 Data collection

2.2.1 Video patient monitoring

LVU devices capture continuous video in RGB or near-infrared (NIR) mode, depending on ambient lighting. Each device is equipped with a CPU and Neural Processing Unit (NPU), capable of processing data at 1fps to minimize latency and reduce cloud processing requirements. Inference results are uploaded to a secured cloud database (Google BigQuery), with blurred frames stored separately for manual annotation. Camera placement varied based on room layout and clinical workflows (Figure 5B).

2.2.2 Annotations

2.2.2.1 Frame-level labels

A professional labeling team manually annotated over 40,000 images with object bounding boxes, object properties, and scene-level tags (Figure 6). Objects were annotated with 2-d bounding boxes classed as “person”, “bed”, or “chair”, and each “person” bounding box was also assigned a role of “patient”, “staff”, or “other”. Scene level attributes were added for whether the patient was “in bed” or “not in bed”, whether the camera was operating in IR mode, and whether the scene included “exception cases” in comparison to stated norms. Exception cases were applied in any instance of labeler uncertainty (e.g. difficult to see person, patient in street clothes, etc.); in instances of multiple exception cases being applicable, a single “frame exception” catch-all was used. Annotations and quality review were conducted using the Computer Vision Annotation Tool (CVAT, Corporation, 2023), and final QA was conducted using the FiftyOne tool (Moore and Corso, 2024).

2.2.2.2 Observation logs

Blurred video summaries for 10 patients (54 patient-days) were reviewed to log periods when the patient was alone. Logs included timestamps with 1-2 second precision (Figure 6), and underwent secondary quality assurance to provide feedback to labelers and fill out any missing periods.

2.3 Computer vision predictions

The LookDeep Health pipeline processes video data using custom-trained models to detect objects, classify person-role, and estimate motion at 1 fps. Preprocessing compresses frames to JPEG at 80% quality and resizes to a resolution of 1088x612 to reduce bandwidth consumption while still meeting downstream model requirements. Image processing is conducted using OpenCV (Bradski, 2000) and RKNN-toolkit (AI Rockchip, 2024).

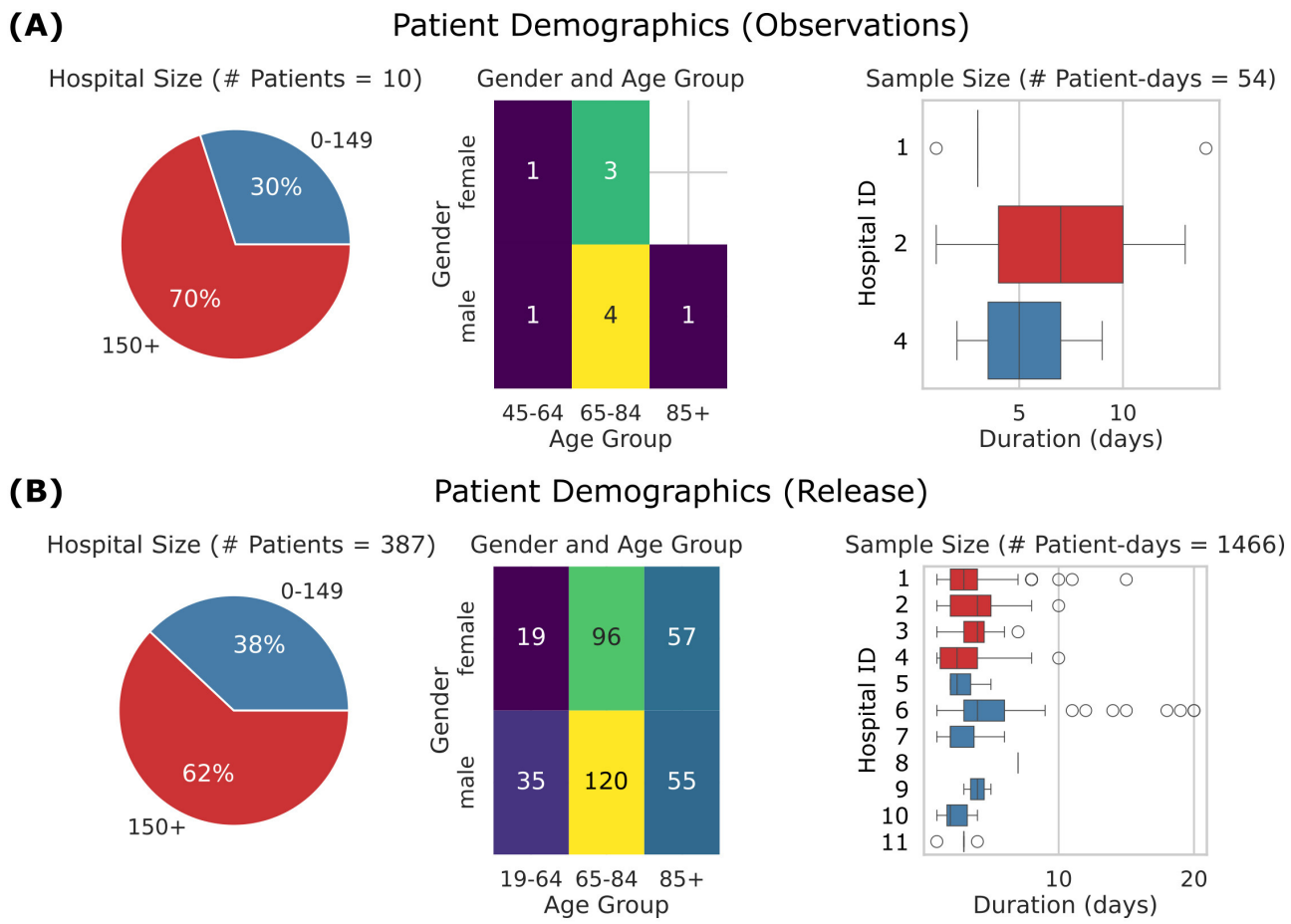


FIGURE 2
Overview of patient demographics. **(A)** Observations subset, comprising 10 patients, 54 patient-days, and 3 hospitals. **(B)** Released dataset, comprising 387 patients, 1,466 patient-days, and 11 hospitals. Left: Pie charts showing the distribution of hospitals by size, where hospitals with an average daily census of 150+ patients are shown in red, and smaller hospitals are shown in blue. Center: Heat maps showing patient age distribution by gender. Right: Box plots showing patient length of monitoring. Central line represents the median, box edges indicate the 25th and 75th percentiles, and whiskers extend to the most extreme data points within 1.5 times the interquartile range. The points represent outliers beyond this range. The y-axis corresponds to hospital IDs, so Hospital 3 is absent from the top-row dataset but included in the bottom-row dataset. The released dataset shows a broader demographic and extended data duration compared to the observations subset.

1. **Object detection (person/bed/chair):** based on the YOLOv4 architecture (Bochkovskiy et al., 2020), the model identifies key objects in each frame, including "person", "bed", and "chair". Training models were initialized using COCO weights (Lin et al., 2014), then fine-tuned on labeled data. Input images were down-sampled to 608 × 608 with OpenCV's cubic interpolation method to fit model requirements. Since the models operate with a smaller fixed input size, increasing the resolution of input images would not significantly improve detection performance unless alternative patch-based approaches were considered. Additionally, the impact of input size on detection accuracy has been well-documented in the original YOLOv4 manuscript, which demonstrated stable performance across various input sizes. Training was conducted on NVIDIA 3070 GPU, and models were subsequently converted for execution on the Rockchip RKNN embedded in the LVU devices.
2. **Person classification (patient/staff/other):** during object detector training, bounding box labels were augmented to classify detected persons by role ("patient", "staff", "other"). Then, at inference time, each "person-" bounding box are re-labeled as "person", with the specific role saved in a separate classification field. Confidence scores for role classifications are derived by taking the highest detection confidence as the primary class and distributing residual scores across remaining classes to indicate potential alternate roles.
3. **Optical flow (motion estimation):** motion between frames was estimated using the Gunnar-Farneback dense optical flow algorithm, which calculates horizontal and vertical displacement for each pixel (Farnebäck, 2003). Optical flow inputs were converted to grayscale and down-sampled to 480x270 to ensure real-time execution. For each region of interest, average motion magnitude was calculated by averaging horizontal and vertical flow vectors, providing an indicator of activity intensity. This estimation does not require training and was implemented using OpenCV with fixed parameters: pyramid scale (pyr_scale = 0.5), number of pyramid levels (levels=3), window size (winsize = 15), number of iterations (iterations=3), size of pixel neighborhood used to find polynomial expansion (poly_n = 5), and the

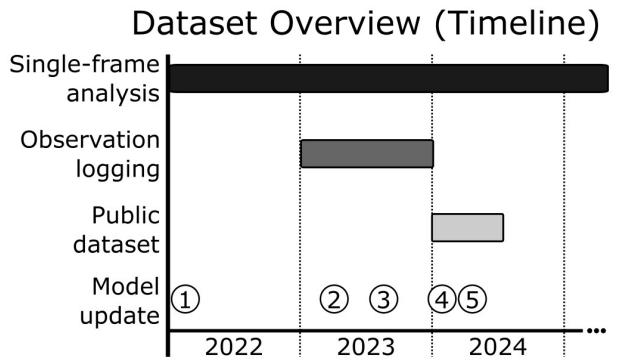


FIGURE 3
Dataset overview and timeline of model updates. Progression of data collection and model updates for the LookDeep Health monitoring system. Single-frame analysis data collection spans a two year period—a broad temporal range for training and validation of object detection and classification tasks. Observation logging data, used for trend validation, was collected over a one year period. The publicly released dataset includes data from a more recent six month period, representing over 1,000 collective patient days. Model updates are indicated by numbered points.

standard deviation of the Gaussian used to smooth derivatives (poly_sigma = 1.2).

2.3.1 Additional components

2.3.1.1 Regions of interest (ROIs)

ROIs, such as "safety zones", provide contextual boundaries for monitoring. They are not predictive outputs themselves, but instead are used to track patient movements and boundary crossings. The "safety zone" was a polygonal region defined by the virtual monitor; its pixel mask is generated by expanding the boundary perimeter by 10% to ensure effective monitoring. Additional ROIs used by the system include the full scene and the detected bed.

2.3.1.2 Logical predictions

Logical predictions summarize patient status and interactions. These predictions were derived from a combination of object detection and role classification results and smoothed with a 5-second filter to mitigate intermittent detection errors.

- **Person alone:** *True* when the average number of detected people in the room is less than two.
- **Patient alone:** *True* when the average number of detected people in the room is less than two, and at least one person is classified as a patient.
- **Supervised by staff:** *True* when the average number of detected people in the room is two or more, and at least one person is classified as healthcare staff.

2.3.1.3 Trend predictions

Trends provide insights into immediate and long-term patient activity, aiding in risk identification and care planning. Hourly trends summarize patient behavior (e.g. "alone" or "moving") based on aggregated logical predictions. For each one-hour interval, predictions were used to calculate the percentage of time the patient spent in key states like "alone," "supervised by staff," or "moving". These percentages were then plotted over time to visualize hourly trends in patient isolation or activity levels throughout the day. These trends provide a high-level overview of patient behavior, aiding in the identification of potential risks and informing care decisions.

2.3.1.3.1 "Assisted" trend predictions

A one-off analysis was conducted to simulate the system's performance when one of the predictions was known. The system's trend predictions based solely on AI inference were compared with those generated using a combination of AI inference and observation logs. For this comparison, "assisted" trends were created by integrating AI-predicted states for "moving" and "supervised by staff" with manually logged periods of "alone" status from the observation logs. This analysis was conducted across the multiple patients and hospitals included in the "Observation Logging" dataset.

2.4 Evaluation

The performance of the AI-driven monitoring system was assessed through two primary methods: **image-level assessment** and **comparison against observation logs**. In the image-level assessment, each frame was analyzed against manual annotations to evaluate the accuracy of the system's object detection, person-role classification, and scene interpretation capabilities. In parallel, observation logs, created from anonymized video summaries of select patients, were compared against predicted trends to assess the system's ability to capture patient behavior patterns.

2.4.1 Frame-level analysis

Each model in the AI system was evaluated independently to assess its performance in object detection and classification tasks. Key performance metrics—precision, recall, and F1-score—were calculated to measure the accuracy and reliability of each model's predictions. Precision assessed the proportion of true positives among all predicted positives, recall measured the ability to identify all true positives, and the F1-score provided a balanced metric between precision and recall.

In addition to these direct object detection and classification tasks, the AI system also generated higher-level, "logical" predictions derived from these outputs. For example, the prediction "is patient alone" was inferred based on a combination of object detection results, such as the absence of healthcare staff within a defined proximity to the patient. These logical predictions were treated as classification tasks themselves, with their accuracy similarly evaluated using precision, recall, and F1-score metrics based on labeled image data. This multi-layered approach allowed us to thoroughly validate both the core object detection functions of each model and the system's ability to interpret and apply these outputs to patient monitoring tasks.

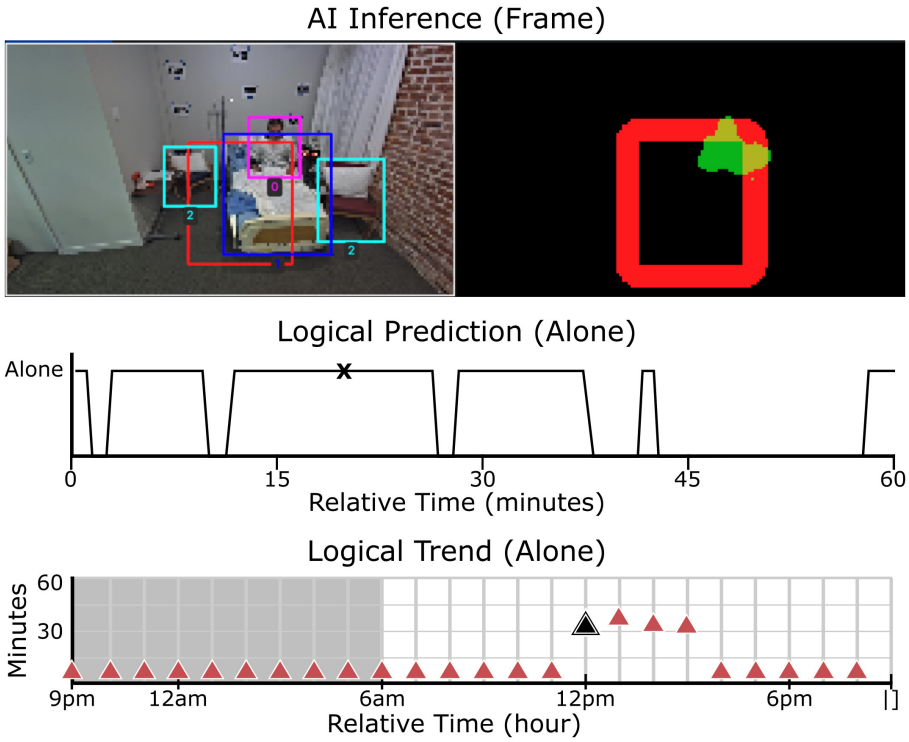


FIGURE 4
Real-time object detection, motion analysis, and patient status monitoring. **Top-left**: Object detection with bounding boxes identifying key elements in the scene. **Top-right**: Segmentation map, where red represents the designated safety zone and green indicates detected motion. **Middle**: The "alone" logical trend over time, showing whether the patient was detected alone in the room for every second within the hour. Peaks indicate periods when the patient was unaccompanied, while lower values indicate caregiver presence. **Bottom**: The "alone" trend over a 24-hour period, aggregated for each hour. This visualization highlights patterns in patient supervision throughout the day. The black markers in the middle and bottom rows correspond to the timestamp of the video frame shown in the top row.

2.4.2 Trend analysis

Trend analysis was conducted by comparing the system's inference-derived metrics to ground truth metrics recorded in observation logs, with both datasets aggregated by patient-day. Unlike the hourly trends shown in Figure 4, analysis was conducted at the per-second level to ensure accurate alignment between AI predictions and observation logs. The primary metric for this analysis was logistic regression accuracy, which assessed the AI system's ability to predict observed behaviors within three time periods: daytime (6 am to 9 pm), nighttime (9 pm–6 am), and the full 24-hour period. In cases where only a single class (e.g. "alone" or "not alone") was present within a specific time period, logistic regression was not feasible. Instead, a manual accuracy score was computed, to allow for consistent accuracy measurements across all time intervals. This score is defined as the proportion of matching values between the AI predictions and ground truth.

Focusing on the "alone" binary behavior trend enables an assessment of the alignment between AI predictions and real-world observations. This analysis validated the AI system's effectiveness in capturing hourly patient behavior trends, underscoring its potential utility in real-time patient monitoring and early detection of deviations from expected patterns.

2.4.3 Camera position meta-analysis

Since cameras were mounted on mobile carts rather than fixed positions, there was variability in camera setup across patients and hospital rooms (Figure 4B). To explore the potential impact of this variability, labeled bed locations were used to estimate each camera's position relative to the hospital bed. Distributions of the labeled bed area and size within each frame, along with the centroid location of the bed relative to the camera's field of view are plotted in Figure 7. These distributions provide an indirect measure of camera position.

This exploratory analysis helped identify patterns and variations in camera setups across different monitoring sessions. However, this information was observational and used only to understand positional variability; no specific adjustments were made during model training or evaluation to account for different camera positions. The results underscore the robustness of our models in handling diverse camera perspectives, as the system maintained consistent detection performance despite these variations.

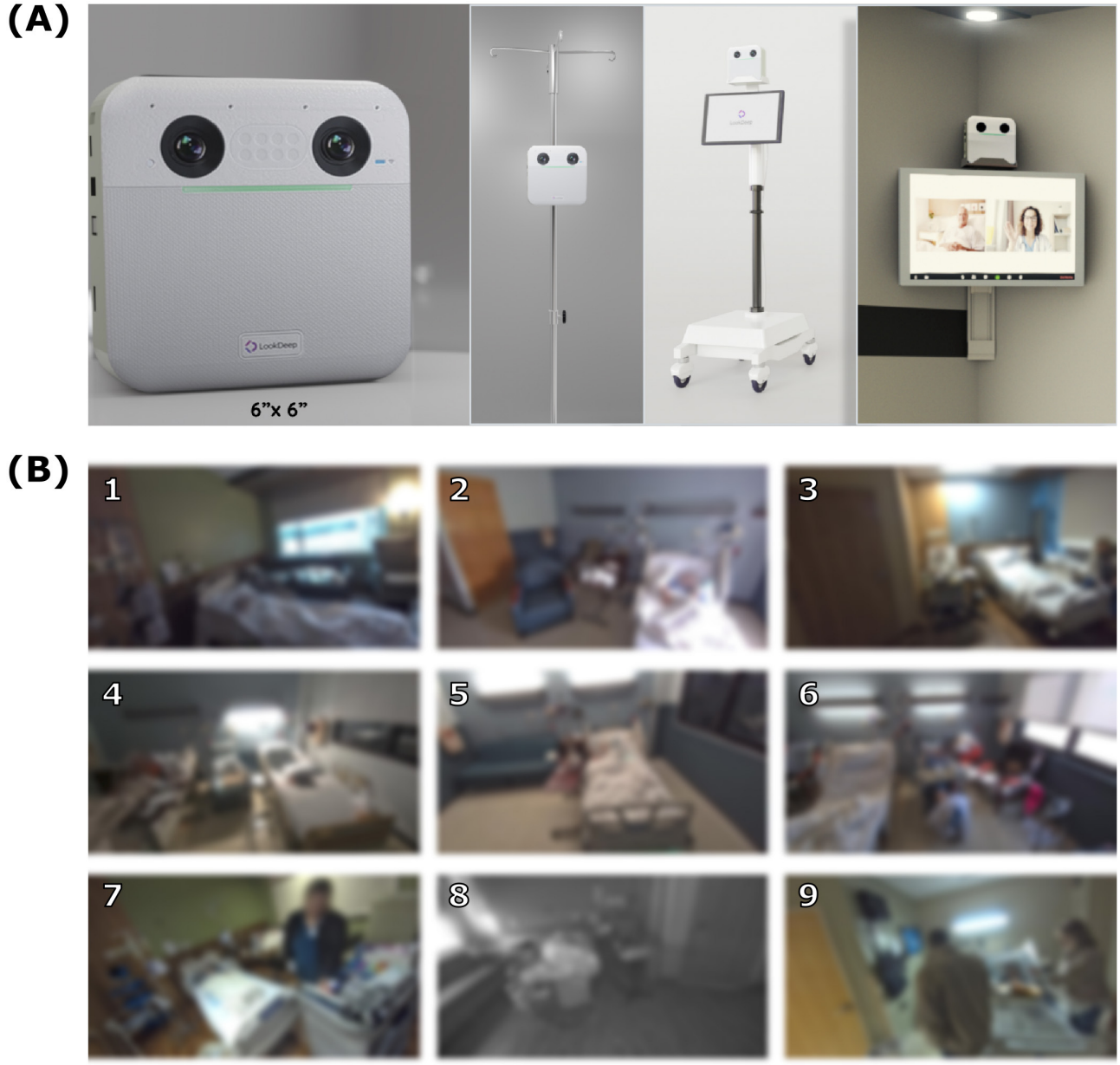


FIGURE 5
Camera setup and example frames. **(A)** LookDeep Video Unit (LVU), a 6" x 6" device, in various mounting configurations. **(B)** A 3 × 3 grid of representative frames captured by the system, showing a diversity of configurations. All images are intentionally blurred to maintain privacy. Each numbered frame provides a unique example that is found in Figure 7.

3 Results

3.1 Frame-level analysis

3.1.1 Object detection, role identification, and patient isolation classification

The evaluations demonstrated that the custom-trained computer vision models perform robustly in real-world hospital settings, achieving high precision across key object detection and classification tasks. We compared five production models alongside a baseline model using an off-the-shelf YOLOv4 configuration (Table 1). Each production model corresponds to a different release, with progressively larger and more refined training datasets incorporated over time (Figure 3). This iterative refinement led to increased model accuracy and adaptability in real-world hospital settings. To ensure consistency, all frame-level analysis was conducted on 10,000 frames collected over a two year period. This representative sample, excluded from model training and validation, highlights the incremental improvements achieved by expanding training datasets across model versions.

As newer models were released, the training set was expanded to include additional annotated data, allowing each successive model to capture more complex and diverse scenarios encountered in hospital environments. The most recent fine-tuned model (v5)

achieved an **F1-score of 0.91** for detecting "person", notably surpassing the baseline YOLOv4 model score of 0.41 (Table 2). Across all object classes–including beds, furniture, and other room elements–the v5 model demonstrated an **F1-score of 0.92**, reflecting a high degree of accuracy and consistency across diverse object types.

In addition to object detection, the system was evaluated on a three-class person-role classification task, distinguishing between patients, healthcare staff, and visitors within the camera's field of view. The v5 model demonstrated particularly strong performance for the "patient" class, achieving an **F1-score of 0.98**, which reflects its high accuracy in identifying patients specifically (Table 2). Accurate person-role classification is essential for monitoring patient interactions and ensuring appropriate caregiving behaviors, as it enables the system to capture not only the presence of individuals but also their roles. Focusing on the "patient" class, the high F1-score underscores the model's robustness in tracking patient activity and interactions, which are critical for effective continuous monitoring in dynamic hospital environments.

The downstream classification task of identifying whether a patient was "alone" in the room showed similarly strong results, with the v5 model achieving an **F1-score of 0.92** (Table 2). This classification task, essential for monitoring patient isolation, consistently improved with each new production release, as more comprehensive training data contributed to better model accuracy. These results confirm the advantage of iterative model refinement

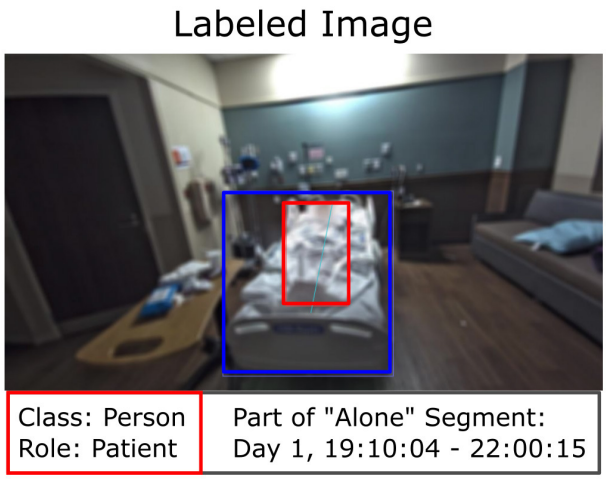


FIGURE 6
Manually labeled image with observation log alignment. The bed is highlighted with a blue bounding box. The patient, identified as a "Person" with the role "Patient", is highlighted with a red bounding box. The associated observation log for "Alone" is shown for illustrative purposes.

TABLE 1 Performance metrics of successive model versions for object detection.

Model version	Fine-tuning data size	Object detection (all)	
		Precision	F1
YOLOv4 (baseline)	n/a	0.84	0.59
Model v1 (2022 Q1)	+700	0.97	0.74
Model v2 (2023 Q2)	+2,474	0.98	0.83
Model v3 (2023 Q3)	+10,133	0.97	0.83
Model v4 (2024 Q1)	+28,914	0.98	0.91
Model v5 (2024 Q2)	+34,239	0.97	0.92

Summary of precision and F1-scores across different versions of the LookDeep Health AI model, highlighting improvements in key tasks as the training data increased. The baseline YOLOv4 model demonstrates initial performance levels, while successive versions (Models v1 to v5) show incremental gains in object detection. With each model iteration, higher precision and F1-scores indicate enhanced detection accuracy and classification reliability, underscoring the impact of additional data and model refinement on real-time patient monitoring capabilities. Evaluation was performed on a fixed dataset containing 10k images.

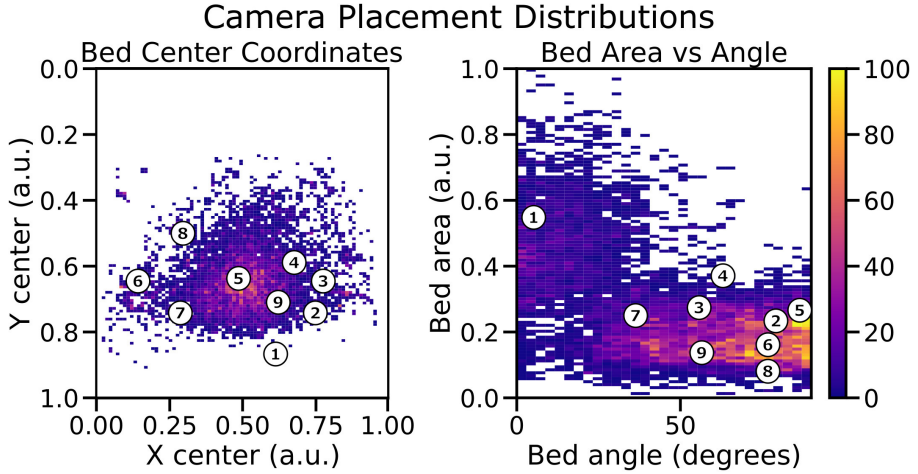


FIGURE 7
Distribution of labeled bed positions relative to the camera. **Left**: Spatial variability of bed center coordinates. **Right**: Distribution of bed area vs. bed angle relative to the camera. Each numbered point is shown in Figure 5B. This highlights the variations in camera perspective and placement across the study.

TABLE 2 Performance metrics of successive model versions for object detection (person), role classification, and "patient alone" classification.

Model version	Object detection (person)		Role classification (patient F1)	"Patient alone" classification (F1)
	Precision	F1		
YOLOv4 (baseline)	0.98	0.41	n/a	0.28
Model v1 (2022 Q1)	0.98	0.85	n/a	0.86
Model v2 (2023 Q2)	0.97	0.89	n/a	0.91
Model v3 (2023 Q3)	0.97	0.86	0.97	0.88
Model v4 (2024 Q1)	0.97	0.91	0.98	0.94
Model v5 (2024 Q2)	0.96	0.91	0.98	0.92

Additional results corresponding to Table 1 are presented here, focusing on object detection of persons, role classification, and "patient alone" classification tasks.

TABLE 3 Performance comparison of models on unblurred vs. face-blurred images across versions.

Model version	Evaluation data size	Unblurred images (F1)	Face-blurred images (F1)	Δ F1
Model v3 (2023 Q3)	2,135	0.81	0.85	+0.04
Model v4 (2024 Q1)	1,809	0.86	0.90	+0.04
Model v5 (2024 Q2)	1,226	0.89	0.91	+0.02

Evaluation of model performance on unblurred and face-blurred images across different versions. The F1-score measures the model's performance, with the "Δ F1" column showing the gap between unblurred and face-blurred images. A Δ value closer to 0 indicates better consistency in model performance between unblurred and face-blurred images.

and dataset expansion, with each production release yielding models that are better adapted to the variability and demands of real-world hospital settings.

3.1.2 Impact of privacy-preserving blurring on model consistency

The performance consistency of the models across unblurred and face-blurred images was evaluated using the Δ metric, which represents the F1-score difference between the two image types (Table 3). Across all model versions, the Δ values were relatively small, indicating that face-blurring–a common privacy-preserving preprocessing step–had minimal impact on model accuracy. For versions v3 and v4, the Δ value was +0.04, while in v5 it decreased to +0.02, suggesting improved robustness to blurring as the training data volume increased.

A smaller Δ value is desirable as it indicates that the model performs consistently regardless of whether the images are unblurred or face-blurred. The reduction in Δ for v5 highlights the value of larger, more diverse training datasets in ensuring that the models generalize well across different image types. This is particularly important in hospital settings, where preserving patient privacy often necessitates the use of face-blurred images. The ability to maintain high accuracy in such scenarios ensures the system's practicality and reliability for real-world deployment.

These results demonstrate that the models not only achieve high accuracy but also exhibit resilience to variations introduced by privacy-preserving preprocessing, a key requirement for scalable applications in healthcare environments.

3.1.3 Object detector performance by IR mode

We analyzed the impact of IR mode on object detection performance by comparing F1-scores across different model versions, broken down into all data, IR-on data, and IR-off data (Figure 8). Results demonstrate a clear trend of increasing F1-scores with newer model versions across all conditions. Notably, the performance gap between IR-on and IR-off scenarios decreases with successive model iterations, indicating improved model robustness to variations in lighting conditions.

At baseline, object detection performance in IR-on scenarios lagged significantly behind IR-off scenarios. However, with the latest model version, this gap narrowed substantially, suggesting that additional training data and model refinements have enhanced the system's ability to generalize across lighting conditions. Despite these improvements, it is worth noting that the test set contains an approximate 25:75 ratio of IR-on to IR-off frames, whereas the population average is closer to 40:60. This imbalance may partially account for residual performance differences and highlights the need for more balanced representation in future datasets.

These findings underscore the importance of accounting for lighting variability in real-world hospital environments and demonstrate the system's potential to adapt to challenging conditions such as low-light monitoring.

3.2 Trend analysis

Inference-derived trends for the "patient alone" metric were compared against observation logs to evaluate the system's ability to accurately capture real-world patterns (Figure 9). This trend analysis utilized data from earlier stages of the project when base models with lower performance were deployed. Specifically, the object detectors used for these inferences had an F1-score of 0.85 for "person" detection, which is lower than the performance of the latest models. Despite this, the analysis showed strong

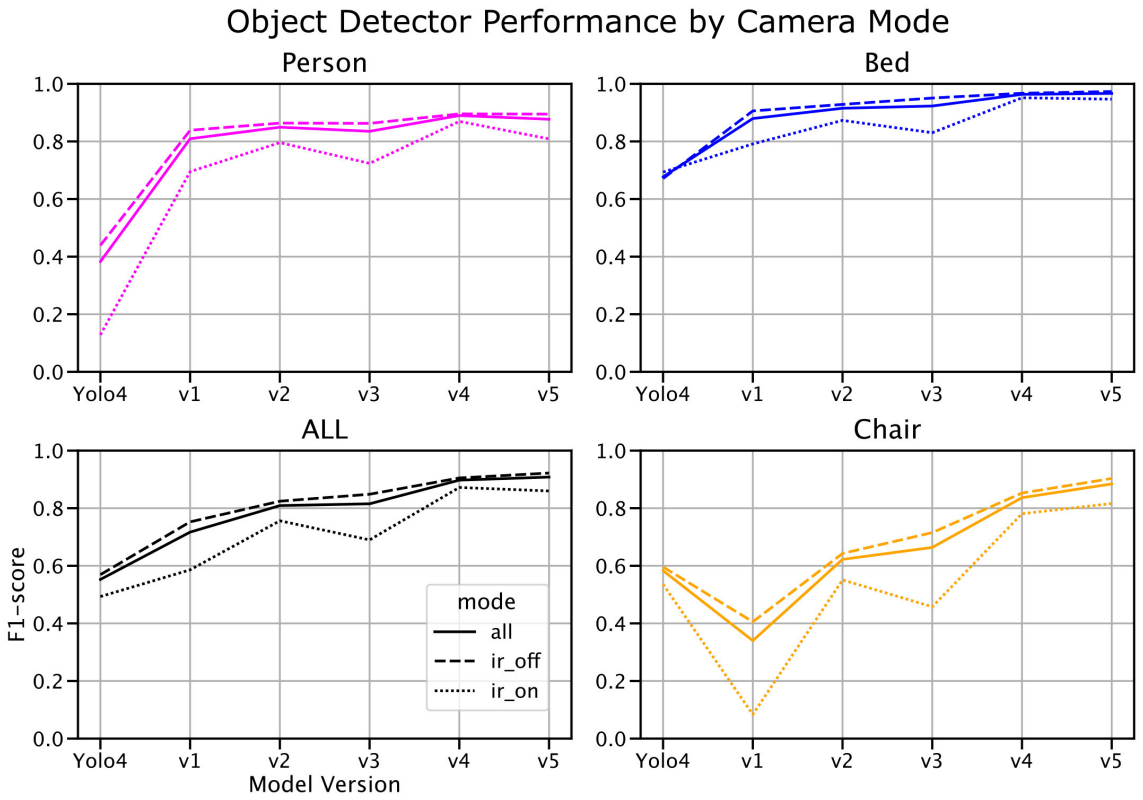


FIGURE 8
Object detection F1-score by model version and infrared (IR) mode. Model performance is shown across successive model versions for all data (solid line), IR-off data (dashed line), and IR-on data (dotted line). The performance gap between IR-on and IR-off modes narrows with more recent model iterations, highlighting increased robustness to varying lighting conditions. Notably, the test set comprises a 25:75 ratio of IR-on to IR-off frames, while the population average is closer to 40:60.

alignment with ground truth data, achieving an **average logistic regression/manual accuracy of 0.84 ± 0.13 during daytime, 0.80 ± 0.16 at nighttime, and 0.82 ± 0.15 across all times**. These results highlight the robustness of the AI system in capturing patient isolation trends, even when using earlier model versions with lower baseline performance.

This accuracy indicates that, even with slightly reduced detection precision in the older models, the system could reliably capture general patterns in patient isolation behavior. The standard deviation (± 0.15) reflects some variability in accuracy across different times of day and patient conditions, possibly influenced by factors such as changing camera angles or environmental conditions. As shown in the normative hourly trends (Figure 10), discrepancies between labeled and AI-inferred "alone" data are more pronounced during nighttime hours, but these differences have minimal impact on the broader trend patterns. For both "Alone and Moving" and "Supervised by Staff" metrics, the AI inferences closely align with label-assisted data, amounting to an **average error of 1–2 min per hour**. This consistency underscores the model's robustness in capturing meaningful patient-alone trends and suggests that any nighttime performance gaps in the "alone" inference do not significantly compromise the overall accuracy. These results highlight the model's potential for improved trend detection as newer, refined models are applied to subsequent data.

4 Discussion

4.1 Implications for clinical practice

The findings of this study underscore the potential for AI-enabled patient monitoring systems to enhance clinical practice through continuous, real-time monitoring. Traditional in-person observations are limited by the time constraints of healthcare staff, who spend limited hours directly interacting with each patient. By providing continuous monitoring, the LookDeep Health platform enables staff to detect patterns that would otherwise go unnoticed, such as extended periods of patient isolation, movement patterns that might indicate a risk of falls, pressure injuries, or irregular interactions with staff. Real-time alerts based on these observations could prompt timely interventions, potentially improving patient safety and outcomes.

Moreover, the data collected by this system can inform trend analysis on a population level, supporting hospital resource allocation and staffing decisions. For instance, identifying times of day when patients are frequently unsupervised could guide adjustments in staffing or the deployment of additional monitoring resources to high-risk patients. Beyond staffing, the system's insights into patient mobility patterns–such as time spent in bed, in a chair, or walking around the room–can help identify markers of successful recovery and readiness for discharge, contributing

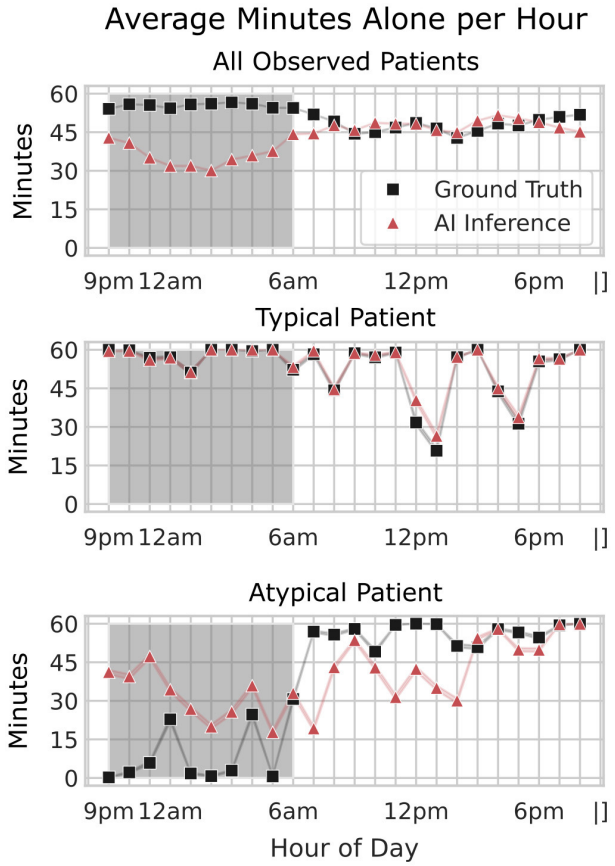


FIGURE 9
Comparison of average minutes alone per hour across all patients, a typical patient, and an atypical patient. The average minutes patients spent alone per hour, comparing ground truth (black squares) and AI inference (red triangles) across three scenarios: all patients **(top)**, a typical patient **(middle)**, and an atypical patient **(bottom)**. The x-axis shows the hour of the day, while the y-axis indicates the average minutes alone per hour. The shaded region represents nighttime hours (9 pm–6 am). For all patients, AI inference closely aligns with ground truth during the day but shows less accuracy at night. An example of a typical and an atypical patient is shown to illustrate the variability in alone time patterns across individual patients. Unlike average trends, this atypical patient exhibits an overprediction of alone time at night, highlighting the need for further model refinement to capture individual patient behaviors accurately.

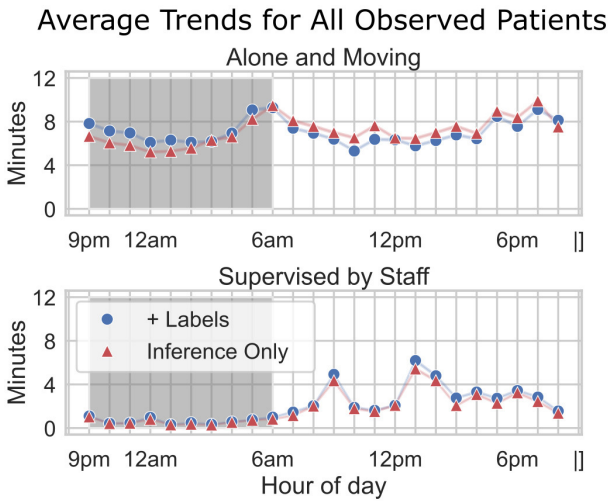


FIGURE 10
Average trends for all observed patients. Hourly trends are compared across two metrics: alone and moving **(top)** and supervised by staff **(bottom)**. Fully AI-inferred data ("Inference Only") is plotted with red triangles. "Assisted" data ("+Labels") is plotted with blue circles—for these data, ground truth "alone" status was used instead of AI-inference. The x-axis indicates the hour of the day, while the y-axis shows the average minutes per hour. The shaded region represents nighttime hours (9 pm–6 am). Although there is a discrepancy between labeled and inferred data for Alone, particularly during nighttime hours, the downstream impact on overall trend accuracy appears minimal (1–2 min per hour).

to improved patient outcomes. These mobility insights could also support the development of best practices for post-procedure mobility, tailored to specific surgeries or treatments, to enhance patient recovery. Altogether, these data-driven insights promote a more efficient, personalized approach to patient care, potentially improving patient satisfaction and clinical outcomes.

4.2 Impact of face-blurring on model performance

While the evaluation of model performance on both unblurred and face-blurred images provides valuable insights, it is important to note that face-blurring is applied only during training and evaluation phases. In real-world deployment, the model will encounter unblurred images as it monitors patients in hospital settings, making this distinction critical to understanding its practical performance. The small Δ values observed across different model versions indicate that the models have been designed to handle face-blurred images without significant degradation in performance. The reduced Δ in the latest version (v5), attributed to increased training data volume, demonstrates improved resilience to face-blurring. However, further studies are needed to assess the model's performance in unblurred scenarios, particularly in environments where face-blurring images for training and evaluation is not an option. This approach ensures privacy during development while maintaining practical deployment fidelity, as real-time monitoring operates on unblurred frames.

4.3 Variation in camera setup

The LookDeep Health patient monitoring platform was deployed in real-world hospital settings with cameras mounted on mobile carts rather than fixed positions, resulting in variation in camera angles, distances, and perspectives across different patient rooms. This variability introduced potential challenges in maintaining consistent object detection and classification accuracy, as model performance can be influenced by changes in camera field of view and positioning relative to the bed. To mitigate these effects, we conducted a camera position meta-analysis using metadata on labeled bed area and centroid location to estimate the approximate

camera placement within each room. Our analysis confirmed that, despite positional differences, the model consistently achieved reliable performance across object detection and classification tasks, demonstrating its robustness to spatial variability. However, this setup presents limitations in controlling for optimal camera positioning, a factor that future studies with standardized camera setups could explore further to minimize variability and enhance model reliability.

4.4 Nuanced differences in time coverage of analyses

A key aspect of this study is the variation in time coverage across different datasets, reflecting the evolving nature of data collection and model validation in real-world hospital settings. The observation logs dataset, which provided ground truth for logical trend validation, was collected exclusively in 2023. In contrast, frame-level annotations for evaluating object detection and person-role classification were gathered over a more extended period from 2022 to 2024. Additionally, the publicly released dataset comprises data collected from a 6 month span across 2024, representing over 1,000 collective patient days across multiple hospitals.

These differences in collection periods introduce nuances in interpretation. For instance, frame-level evaluations benefit from the broader time span, capturing a variety of hospital conditions and patient behaviors across seasons and changing workflows. However, trend analyses were constrained to the observation log time frame, which may limit the ability to generalize trends across the entire study period. Similarly, the released dataset reflects data from the latter phase of the study, aligning with the most refined models but excluding early-stage model iterations.

These variations in time coverage highlight the need to contextualize each analysis within its specific time frame. Future studies could benefit from aligning data collection periods across all evaluation methods, ensuring that models validated on frame-level tasks are continuously validated against trend and behavioral analyses for consistent performance insights over time.

4.5 Challenges and limitations

Several challenges and limitations were encountered in this study. First, the variability in camera setup, as mentioned earlier, introduces potential inconsistencies in model performance due to changing perspectives and distances. While our metadata analysis mitigated this to some extent, a standardized camera setup would likely yield more consistent results.

Second, while the LookDeep Health system demonstrated strong performance in object detection and role classification, real-time video processing presents computational challenges that require balancing accuracy and processing speed. Our use of onboard CPU and NPU on LVU devices provided sufficient processing capabilities for 1 fps inference; however, the scalability of such a setup may be constrained in larger hospital systems requiring higher frame rates for finer details.

Third, the dataset collected in this study primarily consists of high-risk fall patients, which may limit the generalizability of findings to broader patient populations - for example, high-risk patients exhibit limited mobility compared to other patient groups. Additionally, the analysis was conducted on older model versions for some trend analyses, potentially lowering the accuracy of trend detection. Although model refinements are expected to improve results, these differences in model versions should be considered when interpreting the findings.

Lastly, maintaining patient privacy is paramount in continuous video monitoring systems. While the LookDeep Health platform anonymizes all video and stores de-identified data, ongoing attention to data privacy and compliance with healthcare regulations is essential for future deployments in clinical environments.

4.6 Suggestions for future research

While this study provides a foundation for understanding the impact of AI-driven patient monitoring, further research is warranted to explore additional facets of this technology. Future studies could investigate:

- **Enhanced edge case handling:** expanding training datasets to include more examples of diverse scenarios, such as low-light conditions and atypical patient behaviors, could improve model robustness in challenging environments.
- **Advanced deep learning techniques:** integrating more sophisticated deep learning architectures like transformer-based architectures or temporal models could enhance the detection of subtle anomalies, while adaptive pipelines could improve real-time robustness in dynamic hospital environments.
- **Refining architectures and guardrails:** future work could involve refining architectures to detect edge cases more accurately, tracking patterns in prediction errors, and incorporating confidence-based guardrails to prevent catastrophic failures. Such guardrails could include alerts when model confidence is unexpectedly low for consecutive predictions.
- **Higher frame rates and computational scaling:** evaluating the potential for higher frame rates or adaptive frame rate technology to improve real-time responsiveness, particularly in high-activity environments.
- **Standardization of camera placement:** testing standardized, fixed camera setups across patient rooms aims to minimize positional variability and improve model consistency. Although standardization can reduce variability, embracing the inherent diversity of setups may enhance model robustness for real-world applications.
- **Expanded patient cohorts:** extending the analysis to include a wider range of patient demographics and conditions to assess generalizability and adapt the system to diverse populations.
- **Interoperability with hospital systems:** future iterations of the system could integrate more seamlessly with hospital

workflows by automating real-time alerts that directly sync with electronic health record (EHR) systems. For example, patient-specific alerts could be tagged to relevant EHR fields, enabling clinicians to view contextual video data alongside medical records. Additionally, the system could support interoperability with existing hospital tools, such as nurse call systems, to streamline the clinical response to high-risk situations.

These research directions, alongside continued refinement of computer vision models and monitoring systems, will be essential for advancing the practical application of AI in patient monitoring and driving further improvements in healthcare delivery.

5 Conclusion

AI integration in medical imaging is advancing personalized patient treatment but still faces challenges related to effectiveness and scalability. This work demonstrates the potential of computer vision as a foundational technology for continuous and passive patient monitoring in real-world hospital environments.

The contributions of this study are two-fold. First, we introduce the LookDeep Health patient monitoring platform, which leverages computer vision models to monitor patients continuously throughout their hospital stay. This platform scales to support a large number of patients and is designed to handle the complexities of hospital-based data collection. Using this system, we have compiled a unique dataset of computer vision predictions from over 300 high-risk fall patients, spanning 1,000 collective days of monitoring. To encourage further exploration in the field, we released this anonymized dataset publicly at https://github.com/lookdeep/ai-norms-2024.

Second, we rigorously validated the AI system, demonstrating strong performance in image-level object detection and person-role classification tasks. Our analysis also confirms a positive alignment between inference-derived trends and human-observed behaviors on a patient-hour basis, underscoring the reliability of the AI system in capturing patient activity trends. This evaluation can serve as a benchmark for future studies, providing a standard set of criteria for assessing the performance of AI-driven patient monitoring systems.

The extensive dataset and rigorous validation of the LookDeep Health platform highlight the feasibility and impact of continuous patient monitoring through video. By offering real-time insights into patient activity and isolation patterns, continuous monitoring has the potential to reduce fall risks by alerting staff to high-risk situations as they unfold. Beyond improving patient safety, these insights support more efficient staffing and resource allocation, allowing hospitals to adjust care based on real-time patient needs. This predictive capability also aids administrators in managing bed occupancy and optimizing patient flow, particularly during peak times, thus enhancing the responsiveness, efficiency, and scalability of the healthcare system. This work paves the way for future advancements in AI-driven healthcare solutions, promising scalable, data-informed insights to elevate patient care and hospital management.

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.

Ethics statement

The data used in this retrospective study was collected from patients admitted to one of eleven hospital partners across three different states in the USA. The study and handling of data followed the guidelines provided by CHAI standards. Access to this data was granted to the researchers through a Business Associate Agreement (BAA) specifically for monitoring patients at high risk of falls. In compliance with the Health Insurance Portability and Accountability Act (HIPAA), patients provided written informed consent for monitoring as part of their standard inpatient care. To ensure patient privacy, all video data was blurred prior to storage, and no identifiable information is included in this work. Face-blurred frames were used only for training purposes. Faces were manually labeled on fully-blurred images, and the raw images were then treated with a local Gaussian blur in the facial regions, ensuring privacy without compromising model training quality. The outcomes of this analysis did not influence patient care or clinical outcomes.

Author contributions

PG: Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization. PR: Data curation, Investigation, Visualization, Writing – review & editing, Formal analysis. TT: Conceptualization, Investigation, Methodology, Resources, Software, Validation, Writing – review & editing. TW: Investigation, Resources, Writing – review & editing. MC: Writing – review & editing. NS: Funding acquisition, Project administration, Supervision, 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 funded by our hospital system partner as part of a business agreement supporting the development and deployment of AI-driven patient monitoring solutions. The funding provided resources for data collection, system implementation, and analysis within the hospital environment.

Acknowledgments

First and foremost, we extend our gratitude to additional members of the LookDeep Health team, both past and present–Guram Kajaia, James Eitzman, Bill Mers, Mike O'Brien, Jan Marti,

Laura Urbisci, and Tom Hata—for their work in building the patient monitoring platform. We acknowledge the assistance of OpenAI's ChatGPT (version 4, model GPT-4-turbo) in refining the text of this manuscript. This generative AI technology was accessed through OpenAI's platform and used to improve clarity and organization in the presentation of the research. Finally, we thank Aashish Patel, Jacob Hilzinger, Kenny Chen, Tejaswy Pailla, and Quirine van Engen for their valuable feedback on the manuscript.

Conflict of interest

PG, PR, TT, TW, MC, and NS are current or former employees of LookDeep Health, the company that provided the tools used in this study. LookDeep Health was involved in data collection and analysis and reviewed the final manuscript prior to submission. The authors declare no other conflicts of interest related to this work.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. We acknowledge the assistance of OpenAI's ChatGPT (version 4, model GPT-4-turbo) in refining the text of this manuscript. This generative AI technology was accessed through OpenAI's platform and used to improve clarity and organization in the presentation of the research.

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

Abbe, J. R., and O'Keeffe, C. (2021). Continuous video monitoring: implementation strategies for safe patient care and identified best practices. *J. Nurs. Care Qual.* 36, 137–142. doi: 10.1097/NCQ.0000000000000502

AI Rockchip (2024). *Fuzhou Rockchip Electronics Co. RKNN-Toolkit*. Available online at: https://github.com/rockchip-linux/rknn-toolkit (accessed May 26, 2022).

Avogaro, A., Cunico, F., Rosenhahn, B., and Setti, F. (2023). Markerless human pose estimation for biomedical applications: a survey. *Front. Comp. Sci.* 5:1153160. doi: 10.3389/fcomp.2023.1153160

Bajwa, J., Munir, U., Nori, A., and Williams, B. (2021). Artificial intelligence in healthcare: transforming the practice of medicine. *Future Healthc. J.* 8, e188–e194. doi: 10.7861/fhj.2021-0095

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv [Preprint]*. arXiv:2004.10934.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's J. Softw. Tools*. 120, 122–125.

Capo-Lugo, C. E., Young, D. L., Farley, H., Aquino, C., McLaughlin, K., Colantuoni, E., et al. (2023). Revealing the tension: The relationship between high fall risk categorization and low patient mobility. *J. Am. Geriatr. Soc.* 71, 1536–1546. doi: 10.1111/jgs.18221

Chae, W., Choi, D.-W., Park, E.-C., and Jang, S.-I. (2021). Improved inpatient care through greater patient-doctor contact under the hospitalist management approach: a real-time assessment. *Int. J. Environ. Res. Public Health* 18:5718. doi: 10.3390/ijerph18115718

Chen, K., Gabriel, P., Alasfour, A., Gong, C., Doyle, W. K., Devinsky, O., et al. (2018). Patient-specific pose estimation in clinical environments. *IEEE J. Transl. Eng. Health Med.* 6, 1–11. doi: 10.1109/JTEHM.2018.2875464

Corporation, C. (2023). *Computer Vision Annotation Tool (cvat)*. Zenodo. Available online at: https://zenodo.org/records/8416684

Davenport, T., and Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthc. J.* 6, 94–98. doi: 10.7861/futurehosp.6-2-94

Davoudi, A., Malhotra, K. R., Shickel, B., Siegel, S., Williams, S., Ruppert, M., et al. (2019). Intelligent icu for autonomous patient monitoring using pervasive sensing and deep learning. *Sci. Rep.* 9:8020. doi: 10.1038/s41598-019-44004-w

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., et al. (2021). Deep learning-enabled medical computer vision. *NPJ Digit. Med.* 4:5. doi: 10.1038/s41746-020-00376-2

Farnebäck, G. (2003). "Two-frame motion estimation based on polynomial expansion," in *Image Analysis: 13th Scandinavian Conference, SCIA* (Halmstad: Springer), 363–370.

Haque, A., Milstein, A., and Fei-Fei, L. (2020). Illuminating the dark spaces of healthcare with ambient intelligence. *Nature* 585, 193–202. doi: 10.1038/s41586-020-2669-y

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). "Microsoft coco: Common objects in context," in *Computer Vision-ECCV 2014: 13th European Conference* (Zurich: Springer), 740–755.

Lindroth, H., Nalaie, K., Raghu, R., Ayala, I. N., Busch, C., Bhattacharyya, A., et al. (2024). Applied artificial intelligence in healthcare: a review of computer vision technology application in hospital settings. *J. Imag.* 10:81. doi: 10.3390/jimaging10040081

Mascagni, P., Alapatt, D., Sestini, L., Altieri, M. S., Madani, A., Watanabe, Y., et al. (2022). Computer vision in surgery: from potential to clinical value. *NPJ Digit. Med.* 5:163. doi: 10.1038/s41746-022-00707-5

Moore, B. E., and Corso, J. J. (2024). *Fiftyone*. Available online at: https://www.voxel51.com/fiftyone/ and https://github.com/voxel51/fiftyone (accessed April 20, 2022).

Parker, S., Gilstrap, D., Bedoya, A., Lee, P., Deshpande, K., Gabriel, P., et al. (2022). "Continuous artificial intelligence video monitoring of icu patient activity for detecting sedation, delirium and agitation," in *C35. Topics in Critical Care and Respiratory Failure* (Washington DC: American Thoracic Society), A5719–A5719.

Posch, C., Serrano-Gotarredona, T., Linares-Barranco, B., and Delbruck, T. (2014). Retinomorphic event-based vision sensors: bioinspired cameras with spiking output. *Proc. IEEE* 102, 1470–1484. doi: 10.1109/JPROC.2014.2346153

Watzlaf, V. J., Moeini, S., and Firouzan, P. (2010). Voip for telerehabilitation: A risk analysis for privacy, security, and hipaa compliance. *Int. J. Telerehabil.* 2:3. doi: 10.5195/ijt.2010.6056

Westbrook, J. I., Duffield, C., Li, L., and Creswick, N. J. (2011). How much time do nurses have for patients? a longitudinal study quantifying hospital nurses' patterns of task time distribution and interactions with health professionals. *BMC Health Serv. Res.* 11, 1–12. doi: 10.1186/1472-6963-11-319

Wilson, J. E., Mart, M. F., Cunningham, C., Shehabi, Y., Girard, T. D., MacLullich, A. M., et al. (2020). Delirium. *Nat. Rev. Dis. Prim.* 6:90. doi: 10.1038/s41572-020-00223-4

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 11 February 2025
DOI 10.3389/fonc.2025.1541585

OPEN ACCESS

EDITED BY
Sandeep Kumar Mishra,
Yale University, United States

REVIEWED BY
Hina Sultana,
University of North Carolina System, United States
Roberto Altamirano,
University of Chile, Chile

*CORRESPONDENCE
Yumei Wu
wym597118@ccmu.edu.cn

RECEIVED 08 December 2024
ACCEPTED 08 January 2025
PUBLISHED 11 February 2025

CITATION
Abulajiang Y, Liu T, Wang M, Abulai A and Wu Y (2025) The influence of menopause age on gynecologic cancer risk: a comprehensive analysis using NHANES data.
Front. Oncol. 15:1541585.
doi: 10.3389/fonc.2025.1541585

The influence of menopause age on gynecologic cancer risk: a comprehensive analysis using NHANES data

Yiliminuer Abulajiang[1], Tao Liu[2], Ming Wang[1], Abidan Abulai[3] and Yumei Wu[1]*

[1]Beijing Obstetrics and Gynecology Hospital, Capital Medical University, Beijing Maternal and Child Health Care Hospital, Beijing, China, [2]School of Rehabilitation Medicine, Baoding University of Technology, Baoding, China, [3]Department of Endocrinology, The First People's Hospital of Kashi, (The Affiliated Kashi Hospital of Sun Yat-Sen University), Kashi, China

Background: Menopause, a natural transition, affects women's health risks, including gynecologic cancers. Early menopause, linked to lower estrogen, may increase cancer susceptibility. This study analyzed NHANES data from 1999 to 2020 for 8,219 postmenopausal women to explore the relationship between menopausal age and gynecologic cancers. We used regression models and RCS models to assess the risk.

Methods: This study utilized data from the NHANES spanning 1999 to 2020, focusing on 8,219 postmenopausal women selected through stratified sampling. Variables including socioeconomic factors, health behaviors, nutritional status, and medical history were assessed in relation to participants' menopausal age and gynecologic cancer prevalence. We analyzed the relationship between menopausal age and gynecologic cancers (cervical, ovarian, and uterine) using multiple regression models. Additionally, we employed RCS models to evaluate nonlinear relationships between menopausal age and gynecologic cancer risk.

Results: Our findings indicate a significant inverse association between menopausal age and the risk of gynecologic cancers. After controlling for confounding factors such as age, race, BMI, and lifestyle variables, a later age at menopause was associated with a reduced risk of cervical, ovarian, and uterine cancers. The RCS model revealed a non-linear, low-L-shaped relationship, particularly highlighting increased cancer risks at younger menopausal ages. Subgroup analyses demonstrated consistent results across demographic and lifestyle factors, confirming the robustness of the observed associations.

Conclusion: This study reveals the link between menopausal age and gynecologic cancer prevalence. Early menopause is a significant risk factor for cervical, ovarian, and uterine cancers. Our findings support tailored cancer screening based on menopausal age, potentially improving preventive care for postmenopausal women.

KEYWORDS

menopause age, gynecologic cancer risk, personalized cancer screening, NHANES data analysis, risk stratification

Introduction

Gynecological malignancies, including endometrial, cervical, and ovarian cancer, etc., are a leading cause of morbidity and mortality among women, and the yearly number of patients is rising (1, 2). According to statistics, over a million new cases are identified annually, which pose a serious threat to global health. Cervical cancer is the most common gynecological malignancy among women under 40, accounting for 50.4% of cases. Although it can be prevented through vaccination and screening, it remains a leading cause of death, particularly in areas with limited healthcare resources. Following cervical cancer is endometrial cancer, which accounts for 24.2% of cases (3). While its incidence is declining in certain regions, it still contributes significantly to the overall burden of gynecological cancers. Ovarian cancer, at 23%, is the deadliest gynecological malignancy. Due to the lack of effective screening methods, approximately 60% of ovarian cancer cases are diagnosed at an advanced stage, which significantly impacts survival rates (4, 5). Gynecological cancers have a profound impact on women's health and place a significant financial burden on healthcare systems. Their high incidence and mortality rates require management with complex and expensive therapies, which come with several drawbacks, such as treatment-related complications, obesity, social determinants of health, and economic toxicity (6–8).

Menopause is a normal phase of a woman's life, marked by a drop in estrogen levels and usually happening between the ages of 45 and 55. Early menopause refers to menopause that begins between the ages of 40 and 45. The mechanisms behind early and delayed menopause, as well as their relationship with the risk of ovarian cancer, involve a complex interplay between genetic, hormonal, and environmental factors. Genetic predisposition plays a significant role in determining the age of menopause. Hundreds of single nucleotide polymorphisms related to menopausal age have been identified, many of which are associated with immune and mitochondrial functions as well as DNA repair processes. These genetic factors can influence the risk of ovarian cancer (9). Postmenopausal women have persistently high levels of follicle-stimulating hormone, and the changes in hormones are associated with an increase in the expression of inflammatory cytokines and oxidative stress markers, which may lead to malignant transformation of ovarian tissue (10). Delayed menopause is significantly associated with an increased risk of ovarian cancer, which is due to prolonged exposure to estrogen that promotes the development of ovarian cancer (11). Early menopause and primary ovarian insufficiency (POI) are associated with reduced lifetime exposure to estrogen, which may lower the risk of ovarian cancer (9, 12). After menopause, the risk of cervical cancer in women may be reactivated or persist due to human papillomavirus (HPV) infection. Guidelines recommend that screening can stop at age 65 if adequate prior screening has been conducted. However, many women tend to stop screening too early, which increases their risk of developing cervical cancer (13). The prevalence of high-risk HPV infections in postmenopausal women is quite high, with a noticeable increase in infection rates after the age of 65. This suggests that postmenopausal women, particularly those over 65, may benefit from ongoing screening (14). Delayed menopause is associated with a higher risk of endometrial cancer, as long-term exposure to estrogen without the balancing effect of progesterone increases the likelihood of endometrial hyperplasia and cancer. Early menopause shortens the duration of estrogen exposure, thereby reducing the risk of endometrial cancer. This protective effect is due to a shorter reproductive lifespan and decreased cumulative estrogen exposure (15). Obesity and metabolic syndrome are significant risk factors for endometrial cancer. Obesity increases the risk of endometrial cancer in postmenopausal women through various mechanisms, including elevated aromatase activity (16). Insulin resistance and hyperinsulinemia are commonly found in metabolic syndrome, which further increases the risk of endometrial cancer by enhancing the proliferation of endometrial cells (17–19). Furthermore, genetic factors, such as the expression of specific cancer genes like PKD1, have been identified as causes of the progression of endometrial cancer in postmenopausal women. These genetic markers can help predict disease progression and guide targeted therapy (20). Postmenopausal women with endometrial cancer typically exhibit more aggressive disease characteristics, such as higher tumor grades and increased lymphatic metastasis, which are influenced by genetic and hormonal factors (21). The relationship between menopause age and gynecological malignancies is quite complex and influenced by various factors. Racial and cultural factors can affect both the age of menopause and the risk of cancer. A study on Korean women found that menopausal hormone therapy does not increase the risk of melanoma, but certain therapies reduce the risk of non-melanoma skin cancer (22). This indicates that cultural and genetic factors may play a role in cancer risk, which can vary among populations. Lifestyle factors, such as diabetes, may interact with menopausal age. However, one study found no association between menopausal age and microvascular complications in women with diabetes, suggesting that other health factors may obscure. There may be a nonlinear relationship between menopause age and cancer risk. For instance, an earlier onset of menopause is associated with an increased mortality rate, indicating a complex (23, 24).

This study analyzed data from the National Health and Nutrition Examination Survey (NHANES), a cross-sectional survey covering the United States from 1999 to 2020. A total of 8,219 postmenopausal women were selected using stratified sampling methods, ensuring a representative sample. We evaluated various factors, including socioeconomic characteristics, health behaviors, nutritional status, and medical history, and analyzed the relationships between menopausal age and gynecologic cancer prevalence using multivariable logistic regression models. Additionally, restricted cubic spline (RCS) regression models were applied to examine any nonlinear relationships between menopausal age and gynecologic cancer risk, further uncovering potentially complex associations.

Methods

Study design and sample

The NHANES is a nationally representative, cross-sectional survey. The survey used a complex multi-stage probability sample

representing the civilian population in all 50 states and the District of Columbia in the United States (25).

We obtained data from the NHANES, which is representative of the civilian population across all 50 states and the District of Columbia. In the current study, a total of 8,291 NHANES participants were included, representing 21,950,882 postmenopausal women over a span of 14 years (1999-2020). The survey received approval from the Institutional Review Board of the National Center for Health Statistics, and all participants provided informed consent. A flowchart illustrating the process of selecting the study sample is shown in Figure 1. The total population (n=116,876) was screened, resulting in the identification of premenopausal individuals (n=102,652). Among the postmenopausal women, we excluded those with incomplete information regarding menarche and childbirth (n=1,757). From the remaining 12,473 participants, we further excluded individuals with incomplete demographic, disease, dietary, and necessary testing information (n=2,766). Finally, to enhance the scientific validity and reliability of the results, we excluded individuals with premenopausal gynecological diagnoses (n=1,416) from the remaining population, resulting in our ideal study sample of 8,291 individuals.

Sociodemographic characteristics

There are several sociodemographic characteristics, including age, race, and educational level. A non-Hispanic white, a non-Hispanic black, a Mexican-American, another Hispanic, and another race can be excluded from the list. In addition to high school, there are levels of education below high school, high school, and higher education.

Nutritional status

To examine the nutritional quality of the populace, information was gathered from body mass indexes (BMIs) and household property to income ratios (PIRs). Higher PIRs are generally associated with higher levels of physical activity and nutritious intake, as well as higher BMIs compared to low-income populations. The data was divided according to the median, and a cut-off value of 2.3% was chosen as the PIR for households. Screenings were performed on those with BMIs of > 25 kg/m^2 or ≤ 25 kg/m^2.

Habits of behavior

According to how often participants smoked, they were divided into three groups: never smoking, former smoking and now smoking. Alcohol consumption included never drinking, former drinking, mild drinking, moderate drinking, and heavy drinking. In addition to energy intake, behavioral habits are also influenced by population energy intake (kcal).

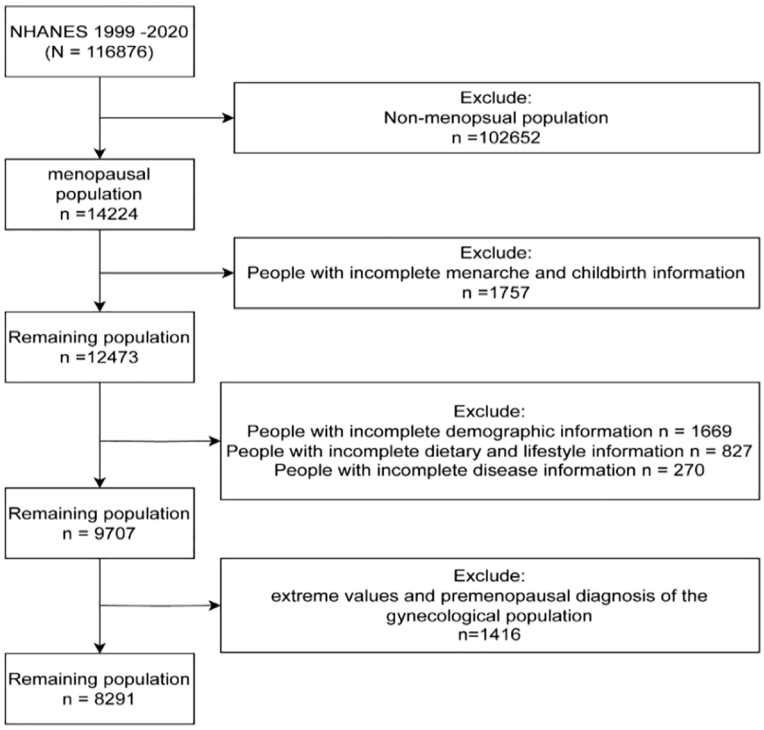


FIGURE 1
Study design and sample.

A medical condition's underlying causes

Diabetes mellitus and hypertension were included because they are two diseases associated with women developing gynecologic cancers and progressing to meet their cancer goals. The following are the clinical diagnostic criteria for diabetes mellitus: (1) The doctor makes the diagnosis; (2) a fasting blood glucose level of 7.0 mmol/L; (3) a glycohemoglobin level of greater than 6.5%; (4) a random blood glucose level of 11.1 mmol/L; (5) a two-hour OGTT level of 11.1 mmol/L; (6) any diabetes medications or insulin already being used.

Statistical analyses

The analyzed data were weighted according to the NCHS. Participants were categorized into two groups based on baseline characteristics according to whether they had gynecologic cancers. Descriptive statistics are used to profile the distribution of participant characteristics, including age, age at menopause, race, education, family PIR, BMI, smoking, alcohol consumption, etc. Data were presented as frequencies with proportions (%), means with standard deviation (SD), or medians with interquartile ranges (IQR). Univariate and multivariate logistic regression analysis between ages at menopause and gynecologic cancers: Crude is an unadjusted model; Model 1 is a model adjusted for age and race; Model 2 is a model adjusted for age, race, first menstruation age, and living birth; Model 3 is a model adjusted for age, race, first menstruation age, living birth, education, BMI, PIR, smoking, alcohol consumption level, energy intake; Model 4 is a model adjusted for age, race, first menstruation age, living birth, education, BMI, PIR, smoking, alcohol consumption level, energy intake, hypertension and diabetes. On this basis, a fully adjusted model was used to assess the association between the age of menopause and major gynecological cancers, including cervical, ovarian, and uterine cancers. To explore the incidence of gynecological cancer in different age groups, participants were further divided into 7 groups based on age at menopause, including ≤30 years, 31-35 years, 36-40 years, 41-45 years, 46-50 years, 51-55 years, and ≥56 years. In the fully adjusted model, a RCS method was used to investigate the non-linear association between age at menopause and gynecologic cancers. In this study, values detected as outliers are treated as missing data and replaced by the result of interpolation. Statistical analyses were conducted using R version 4.4.1 (Posit Software, Boston, MA, USA). A p-value less than 0.05 is considered statistically significant.

Results

Baseline characteristics

The weighted baseline characteristics of the participants, which consisted of 8,219 participants grouped by whether they had gynecological cancer, are shown in Table 1. The results showed statistically significant differences in gynecologic cancer prevalence by family PIR, smoking, menopause, and first menstruation ($P<0.05$). Compared with participants without gynecological cancer, participants with gynecological cancer had lower Family PIR, more smoking, lower age at menopause, and younger age at first menstruation.

Relationship between the age of menopause and the prevalence of gynecological cancers

The results of univariate and multivariate logistic regression analysis between the onset of menopausal age and gynecologic cancers are shown in Table 2. There was an inverse association between age at menopause and the prevalence of gynecological cancer (OR: 0.93, 95% CI: 0.91,0.95), and the difference was statistically significant ($P<0.01$). Model 1 was adjusted for age and race, and the results showed that there was an inverse association between menopausal age and the prevalence of gynecological cancer (OR: 0.92, 95% CI: 0.90-0.94), and the difference was statistically significant ($P<0.01$). Model 2 was adjusted for age, race, first menstruation age, and living birth, and showed an inverse association between age at menopause and gynecologic cancer (OR: 0.92, 95% CI: 0.90-0.94), with statistically significant differences ($P<0.01$). Model 3 was adjusted for age, race, first menstruation age, living birth, education, BMI, PIR, smoking, alcohol consumption level, energy intake, and showed that there was an inverse association between age at menopause and the prevalence of gynecologic cancer (OR: 0.92, 95% CI: 0.91-0.94), and the difference was statistically significant ($P<0.01$). Model 4 was adjusted for age, race, first menstruation age, living birth, education, BMI, PIR, smoking, alcohol consumption level, energy intake, hypertension, and diabetes, and showed that there was an inverse association between age at menopause and the prevalence of gynecologic cancer (OR: 0.92, 95% CI: 0.90-0.94), and the difference was statistically significant ($P<0.01$). Furthermore, to evaluate the effect of specific factors on the gynecological cancers, we performed subgroup analysis and the results are shown in Supplementary Table S1.

Relationship between the age of menopause and the prevalence of major gynecological cancers

To investigate whether menopause is associated with gynecologic age, we performed a regression analysis between menopause and the incidence of three major gynecologic cancers, as shown in Figure 2 for the relationship between age at menopause and the incidence of different gynecologic cancers. After adjusting for age, race, first menstruation age, living birth, education, BMI, PIR, smoking, alcohol consumption level, energy intake, hypertension, and diabetes (model 4), the regression results revealed that age at menopause was inversely associated with the prevalence of gynecologic cancers in patients with cervical (OR:

10.3389/fonc.2025.1541585

TABLE 1 Participant characteristics (N = 8,219) in NHANES 1999–2020.

Characteristic	Non-gynecological cancer (N = 8,017)	Gynecological cancer (N = 274)	*P* value
Age, years	60.00 (53.00, 69.00)	61.00 (49.00, 70.00)	0.430
Race, %			**0.032**
Non-Hispanic White	3,681.00 (73.54)	173.00 (82.70)	
Non-Hispanic Black	1,801.00 (10.90)	36.00 (5.43)	
Mexican American	1,233.00 (4.96)	32.00 (3.52)	
Other Hispanic	741.00 (4.79)	21.00 (3.82)	
Other Race	561.00 (5.81)	12.00 (4.53)	
Education level, %			0.700
Less than high school	2,261.00 (17.49)	82.00 (19.89)	
High school	2,069.00 (28.36)	74.00 (27.81)	
College or above	3,687.00 (54.15)	118.00 (52.30)	
Family PIR	3.03 (1.58, 5.00)	2.48 (1.24, 3.85)	**0.008**
BMI, kg/m^2	28.74 (24.76, 33.46)	30.10 (25.60, 34.99)	0.054
Smoke behavior, %			**0.001**
Never	4,812.00 (57.24)	125.00 (42.91)	
Former	1,940.00 (25.51)	75.00 (29.62)	
Now	1,265.00 (17.24)	74.00 (27.48)	
Alcohol consumption, %			0.068
Never	1,800.00 (16.55)	47.00 (10.00)	
Former	1,606.00 (16.80)	68.00 (20.36)	
Mild	2,539.00 (36.58)	86.00 (32.45)	
Moderate	1,308.00 (20.12)	40.00 (21.83)	
Heavy	764.00 (9.95)	33.00 (15.36)	
Energy intake, kcal	1,627.00 (1,260.00, 2,069.00)	1,602.00 (1,215.00, 2,067.00)	0.700
Hypertension, %	2,503.00 (26.09)	80.00 (22.10)	0.190
Diabetes, %	2,006.00 (19.46)	77.00 (19.91)	0.880
Menopause, years	46.00 (39.00, 51.00)	36.00 (30.00, 46.00)	**<0.001**
First menstruation, years	13.00 (12.00, 14.00)	12.00 (11.00, 13.00)	**0.028**
Living birth	2.00 (2.00, 3.00)	2.00 (2.00, 3.00)	0.950

Bold indicates *P* value < 0.05.

0.88, 95% CI: 0.85-0.90), ovarian (OR: 0.94, 95% CI: 0.91-0.97) and uterine cancer (OR: 0.96, 95% CI: 0.93-0.99). According to different menopause ages, participants were divided into 7 groups, as shown in Figure 3. The percentage of major gynecological cancers occurring at different menopause age groups were observed respectively. The results showed that participants aged 31-35 years had a higher incidence of cervical cancer than other age groups. Cervical cancer incidence decreased gradually among participants aged 31-45 years, with statistical significance (*P* <0.01). The incidence of uterine cancer was higher in participants aged 56 years or older at menopause than in other age groups. Menopausal participants aged 36 to 40 years had a higher incidence of ovarian cancer than the rest of the age group. To further explore more specific associations, subgroup analysis was performed for cervical cancer, ovarian cancer, and uterine cancer, and the results are shown in Supplementary Tables S2–S4.

TABLE 2 Association between age of menopause and gynecologic cancer.

Outcomes	Model	OR (95% CI)	*P* value
Gynecological Cancer	Crude	0.93 (0.91, 0.95)	<0.01
	Model 1	0.92 (0.90, 0.94)	<0.01
	Model 2	0.92 (0.90, 0.94)	<0.01
	Model 3	0.92 (0.91, 0.94)	<0.01
	Model 4	0.92 (0.90, 0.94)	<0.01

OR, odds ratio; CI, confidence interval. Crude is an unadjusted model; model 1 is a model adjusted for age and race; model 2 is a model adjusted for age, race, first menstruation age and living birth; model 3 is a model adjusted for age, race, first menstruation age, living birth, education, BMI, PIR, smoking, alcohol consumption level, energy intake; Model 4 is a model adjusted for age, race, first menstruation age, living birth, education, BMI, PIR, smoking, alcohol consumption level, energy intake, hypertension and diabetes.

Nonlinear relationship between age of menopause and the prevalence of gynecological cancers

By using the RCS models with full adjustment for confounders, the results showed that there was a low L-shaped association between age at menopause and the prevalence of gynecological cancer (Figure 4A). In addition, the results also found that there was a linear association between age at menopause and cervical and uterine cancer (Figures 4B, D), but not with ovarian cancer (Figure 4C).

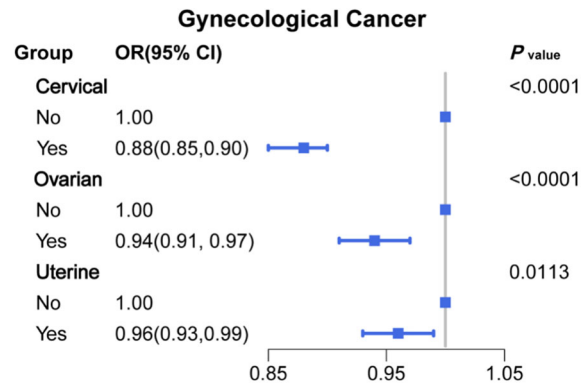


FIGURE 2
Associations between age at menopause and major gynecological cancer (OR (95% CI) and P value).

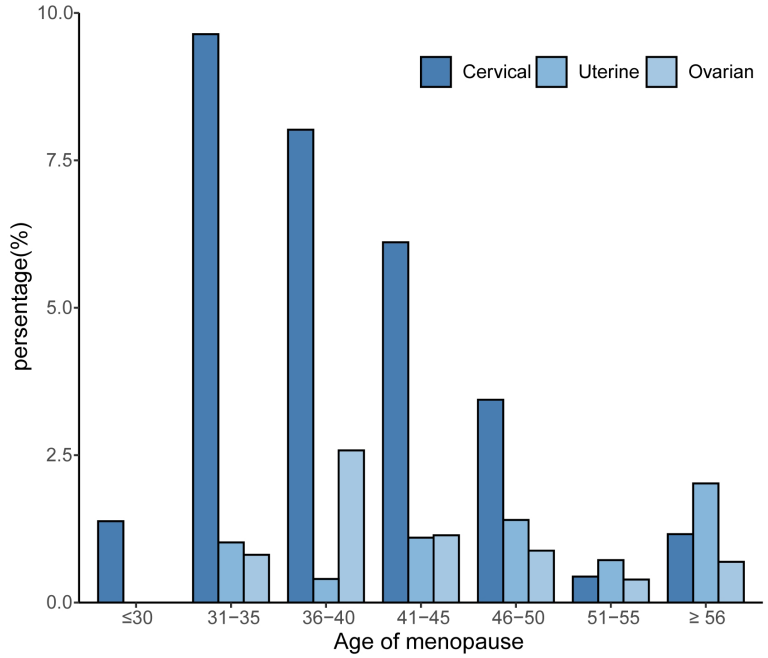


FIGURE 3
Relationship between age at menopause and the prevalence of gynecological cancer by age group.

Discussion

This study revealed that women with an earlier age at menopause face a significantly higher risk of gynecologic cancers (cervical, ovarian, and uterine cancers), supporting the inverse relationship between menopause age and cancer risk. The rapid drop in estrogen associated with early menopause is a key factor in the elevated cancer risk. Previous studies indicate that a quick decline in estrogen may impair tissue repair mechanisms for DNA damage, increase apoptosis, and contribute to chronic inflammation, all of which elevate cancer risks (26–28). A study reported the trends in incidence and mortality rates of cervical cancer in China and analyzed the independent effects of age, period, and cohort on these trends. The results showed that the incidence of cervical cancer has increased among young women under the age of 35 (29). Additionally, a study assessed the incidence, disability-adjusted life years (DALYs), and mortality rates of cervical cancer and found that the incidence has increased among younger age groups, especially among women under the age of 35 (30), which is consistent with the results of the 31-35 age group mentioned in this study.

Moreover, subgroup analysis in this study further refined the risk differences associated with different menopause age groups, showing that women who reached menopause between 36-40 years had a significantly higher risk of ovarian cancer, while women who reached menopause after age 56 had an increased risk of uterine cancer. This finding supports the hypothesis in the literature that late menopause may increase the risk of uterine cancer due to prolonged exposure to high estrogen levels, leading to persistent endometrial stimulation and an elevated risk of uterine cancer (15, 31, 32).

When analyzing the relationship between age at menopause and the risk of gynecological cancer, our study employed RCS and found a low L-shaped relationship between age at menopause and the prevalence of gynecological cancer. This finding is consistent with existing literature, particularly in understanding the impact of changes in estrogen levels on the risk of gynecological cancer.

Firstly, a study based on the NHANES database indicated that a univariate logistic regression analysis of age at menopause and the prevalence of gynecological tumors showed a negative correlation between age at menopause and the prevalence of common gynecological tumors. Particularly for ovarian and cervical cancers, after adjusting for the effects of covariates, a higher risk of gynecological tumors was found, and there were statistically significant differences at earlier ages of menopause. This is in line with our research results, suggesting that before a certain critical point, a lower age at menopause significantly increases the risk of gynecological cancer (33). Furthermore, research has shown that women carrying pathogenic BRCA1/2 gene mutations have up to an 87% risk of developing related cancers. Specifically, multiple breast cancer clusters in BRCA1 and BRCA2 are associated with relatively higher risks of breast cancer and relatively lower risks of ovarian cancer. These findings further emphasize the role of genetic factors in the risk of gynecological cancer and how age at menopause may interact with these genetic risk factors (34). In summary, our research results are consistent with existing literature, highlighting the complex relationship between age at menopause, changes in hormone levels, and the risk of gynecological cancer. These findings provide important scientific evidence for future prevention strategies and intervention measures, especially in identifying high-risk groups and developing

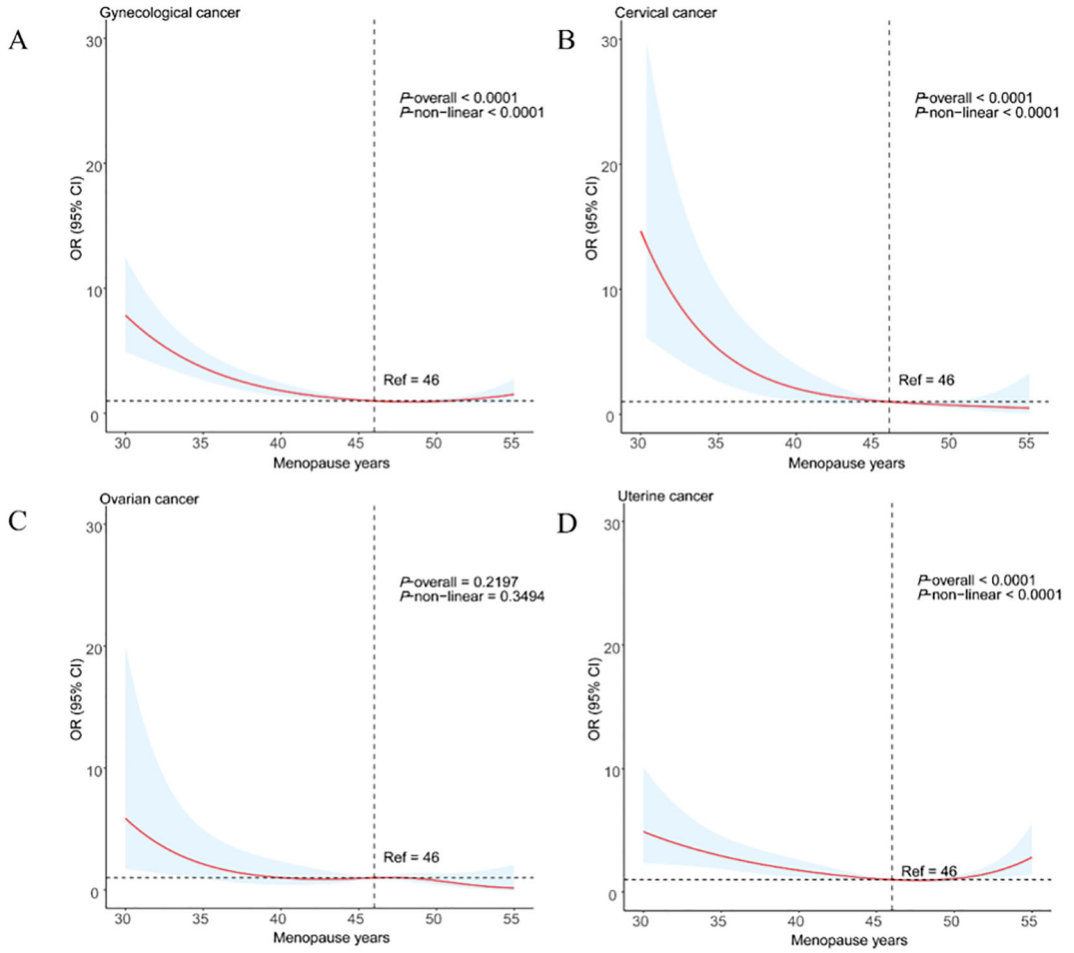


FIGURE 4
Restricted cubic spline analysis of age at menopause and major gynecological cancer **(A)** gynecological cancer; **(B)** cervical cancer; **(C)** ovarian cancer; **(D)** uterine cancer).

personalized prevention plans. We observed a linear inverse relationship between menopause age and the incidence of cervical and uterine cancers, while ovarian cancer showed no significant trend, possibly due to its complex etiology and differing sensitivity to hormones (33, 35, 36).

The chronic inflammatory state post-menopause is also considered a key mechanism in the increase. A 4-year follow-up study that explores the relationship between metabolic health, menopause, and physical activity. The study results indicate that menopause and levels of physical activity have a significant impact on the metabolic health of middle-aged women (37). A literature review based on data from the Study of Women's Health Across the Nation (SWAN), examines the relationship between menopause and metabolic syndrome. The study found that menopause is associated with changes in cardiovascular disease risk factors, which are also related to cancer risk. The study also revealed common genetic signatures associated with metabolic syndrome, type 2 diabetes, cardiovascular diseases, and menopausal status, which are significantly enriched in biological processes, including the positive regulation of binding, the positive regulation of leukocyte cell adhesion, and the regulation of lipid localization (38). Visceral fat accumulation is associated with an increased risk of various cancers, including those of the uterus, cervix, breast, liver, and ovaries. The study also notes that obesity can interfere with therapies and contribute to morbidity from chemotherapy toxicities, thus promoting worse prognosis and mortality (39). The study found that higher levels of insulin resistance are associated with higher breast cancer incidence and higher all-cause mortality after breast cancer (40). Research findings indicate that the link between visceral adipose tissue and cancer risk may involve systemic mechanisms, such as leptin, glucose, insulin, and inflammatory cytokines, which are systemic markers of obesity-related adipose tissue inflammation and may promote tumor development (41). Chronic inflammation may play an important role in the pathogenesis of non-inflammatory diseases such as breast cancer. Activation of innate immunity creates a tissue microenvironment rich in reactive oxygen and nitrogen species that may lead to DNA damage and changes in nearby cells, the study suggests. Inflammation also raises circulating levels of inflammatory cytokines that promote cancer, such as C-reactive protein (CRP) and interleukin-6 (IL-6) (42). There are also studies showing that links between chronic low-grade inflammatory states and multiple chronic diseases are now evident, and controlling this condition may be important to prevent the most common diseases in the general population (43). A case-control study that prospectively

assessed whether plasma levels of inflammatory markers such as CRP, TNF-α, IL-6, leptin, and adiponectin were associated with breast cancer risk showed no significant association between these inflammatory markers and breast cancer risk but found significant interactions between menopausal status and plasma levels. All of these studies support the scientific evidence for a relationship between postmenopausal chronic inflammatory state and cancer risk, and support the idea that postmenopausal chronic inflammatory state may be one of the key mechanisms for increased cancer risk (44). This inflammatory state aligns with our findings, supporting the inclusion of early menopausal women in high-risk cancer screening groups.

In addition to the elevated risk for early menopausal women, late menopausal women also face specific health risks. A study used a meta-analysis to evaluate the relationship between unopposed estrogen or estrogen plus progesterone and endometrial cancer risk. The results showed that women who use estrogen have a higher relative risk than non-users. Risk (RR 2.3) was associated with prolonged use (RR 9.5 for 10 years or more), and the risk of endometrial cancer remained elevated even after 5 years or more of discontinuation of unopposed estrogen therapy (RR 2.3) (45). A systematic review assessed the safety of estrogen plus progestin therapy, particularly considering the impact of treatment regimens and types of progestins on the risk of endometrial cancer. The study found that women who used estrogen alone had an increased risk, while continuous combined therapy was associated with a lower risk compared to sequential combined therapy (46). Hormone replacement therapy should be used with caution in women with a higher risk of endometrial cancer (HR 2.84) in those with a later menopause (age ≥55 years) than in those with the youngest menopause (<45 years)15. These studies underscore the importance of menopause age in gynecologic cancer screening and intervention strategies.

Recently, more studies have viewed age at menopause as an outcome of multiple interacting factors, further highlighting its unique impact on cancer risk. A study points out that lifestyle and dietary factors determine the age of natural menopause. The research indicates that a healthy diet and regular exercise are significant factors affecting the age of menopause, thereby potentially indirectly influencing the cancer risks associated with early menopause (47). A systematic review and meta-analysis studied the impact of psychological interventions on the quality of life of early-stage cancer patients. The study included psychological interventions such as cognitive-behavioral therapy, relaxation training, meditation, stress management, and self-help, which are believed to improve patients' quality of life and may indirectly affect cancer risk (41). There is also a method called "emotional support and case finding" used for the clinical management of cancer patients' emotions. This approach emphasizes the importance of psychological support in cancer treatment and may help reduce cancer risk (48). Psychological interventions for cancer patients include cognitive-behavioral therapy, art therapy, and relaxation therapy, among others. These interventions aim to improve patients' emotional states and quality of life, which may positively impact the reduction of cancer risk (49).

This study makes a significant contribution by providing an in-depth analysis of the relationship between age at menopause and the risk of three major gynecological cancers: cervical, ovarian, and endometrial cancer. The results indicate a notable association: early menopause correlates with an increased risk of cervical and ovarian cancers, whereas late menopause is associated with a higher risk of endometrial cancer. These findings provide a scientific foundation for future personalized screening and health intervention strategies. The inverse association between menopausal age and cancer prevalence suggests that early menopause may be a marker for increased risk, prompting more frequent monitoring and targeted screening for women who experience menopause at younger ages. Additionally, the nonlinear relationship observed highlights the need for personalized risk assessments, taking into account individual factors such as age at menopause, lifestyle, and family history, to optimize prevention and early detection strategies for gynecological cancers. Unlike previous research that broadly examined the link between menopausal age and cancer risk, this study categorizes menopausal age into specific age groups and conducts a subgroup analysis across different cancer types, thereby revealing age-specific cancer risks. Furthermore, by employing a multilevel regression model and adjusting for various confounding variables, the study clarifies the independent effect of menopausal age on cancer risk, enhancing the statistical robustness of the findings. Additionally, the use of the large, representative NHANES database lends strong external validity to the study. NHANES data encompass participants from diverse racial, socioeconomic, and health backgrounds, enhancing the generalizability of the findings. Many previous studies, limited by small sample sizes or specific populations, restricted the applicability of their results. By leveraging NHANES's extensive dataset, this study addresses these limitations and offers a robust reference point for personalized cancer screening in various populations. Another notable achievement of this study is its exploration of a potential nonlinear relationship between menopausal age and gynecological cancer risk. Using RCS regression models, the study is among the first to suggest an L-shaped nonlinear association, indicating that cancer risk may not increase linearly with menopausal age but could be influenced by a combination of factors, with critical risk thresholds for different age groups. This insight provides important theoretical support for age-segmented clinical management strategies.

Despite these valuable insights, the study has several limitations. First, as a retrospective analysis based on cross-sectional data from the NHANES database, it cannot establish causation. Though we have adjusted for multiple confounding factors, the possibility of reverse causation cannot be ruled out. Future longitudinal studies are needed to confirm the causal link between menopausal age and cancer risk, clarifying whether early menopause directly contributes to elevated cancer risk or if other intermediary factors are involved. Second, the study relies on self-reported data, including menopausal age, menarche age, and lifestyle factors, which may introduce recall bias and reporting inaccuracies. Participants may not accurately recall age-related events or health behaviors, particularly over long periods. Future research should incorporate objective biomarkers to reduce self-reporting errors. For example, hormonal and inflammatory biomarkers could more precisely measure physiological changes associated with menopause and their correlation with cancer risk. Moreover, this study does not delve into the variability in the relationship between menopausal age and cancer risk across different demographic groups (e.g., by race, socioeconomic status, and living environment). Both

menopausal age and gynecological cancer incidence may vary significantly across racial and socioeconomic groups, especially in terms of lifestyle factors and healthcare access. Future studies should analyze these differences in greater detail to understand how menopausal age distribution and its impact on cancer risk vary across populations, which would aid in developing more targeted and equitable health management strategies, improving the efficiency of cancer screening and prevention. Finally, the study does not fully explore the biological mechanisms underlying the association between menopausal age and cancer risk. Although hypotheses around estrogen decline and chronic inflammation are proposed, these mechanisms require further verification through experimental and longitudinal studies. Future research could employ animal models or clinical trials to investigate how menopause-induced physiological changes specifically contribute to cancer development, thereby offering a biological basis for prevention and therapeutic strategies.

Conclusion

This study provides significant insights into the association between age at menopause and the risk of developing gynecologic cancers, particularly cervical, ovarian, and uterine cancers. Our findings underscore the role of early menopause as a risk factor for these cancers, while highlighting late menopause as an associated risk for uterine cancer. By employing a large, representative sample and robust analytical methods, our research contributes to the understanding of menopause's impact on cancer risks. These results have potential implications for clinical practice, suggesting that menopausal age could be a critical factor in developing personalized cancer screening strategies. Future studies, ideally longitudinal in design, are essential to further elucidate the causal pathways involved and to explore the biological mechanisms underlying these associations. Such efforts could pave the way for targeted preventive measures and more effective health interventions for women across different menopausal age groups.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.cdc.gov/nchs/nhanes/index.html.

Ethics statement

The studies involving humans were approved by https://www.cdc.gov/nchs/nhanes/index.html. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

YA: Formal analysis, Investigation, Methodology, Resources, Writing – original draft, Writing – review & editing. TL: Data curation, Methodology, Writing – review & editing. MW: Conceptualization, Supervision, Writing – review & editing. AA: Validation, Writing – review & editing. YW: Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by Capital's Funds for Health lmprovement and Research (No. 2024-1-2112).

Acknowledgments

The author thanks the staff and the participants of the NHANES study for their valuable contributions.

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/fonc.2025.1541585/full#supplementary-material

References

1. Li X, Xiao C, Qu K. Aidi injection, a traditional Chinese biomedical preparation for gynecologic tumors: a systematic review and PRISMA-compliant meta-analysis. *Bioscience Rep*. (2021) 41:BSR20204457. doi: 10.1042/BSR20204457

2. Shi P, Zhang X, Lou C, Xue Y, Guo R, Chen S. Hsa_circ_0084927 Regulates Cervical Cancer Advancement via Regulation of the miR-634/TPD52 Axis. *Cancer Manage Res*. (2020) 12:9435–48. doi: 10.2147/CMAR.S272478

3. Han X, Wang Z, Huang D, Deng K, Wang Q, Li C, et al. Analysis of the disease burden trend of Malignant tumors of the female reproductive system in China from 2006 to 2020. *BMC Women's Health*. (2022) 22:504. doi: 10.1186/s12905-022-02104-2

4. Koutras A, Perros P, Prokopakis I, Ntounis T, Fasoulakis Z, Pittokopitou S, et al. Advantages and limitations of ultrasound as a screening test for ovarian cancer. *Diagnostics (Basel Switzerland)*. (2023) 13:2078. doi: 10.3390/diagnostics13122078

5. Liberto JM, Chen SY, Shih IM, Wang TH, Wang TL, Pisanic TR. 2nd. Current and emerging methods for ovarian cancer screening and diagnostics: a comprehensive review. *Cancers*. (2022) 14:2885. doi: 10.3390/cancers14122885

6. Zhang N, Liu C, Di W. Systemic treatment for gynecological cancer patients undergoing hemodialysis. *OncoTargets Ther*. (2023) 16:545–58. doi: 10.2147/OTT.S419445

7. Agnew HJ, Kitson SJ, Crosbie EJ. Gynecological Malignancies and obesity. *Best Pract Res Clin obstetrics gynaecology*. (2023) 88:102337. doi: 10.1016/j.bpobgyn.2023.102337

8. Esselen KM, Baig RA, Gompers A, Stack-Dunnbier H, Hacker MR, Jang JW, et al. Factors associated with increased financial toxicity after the completion of radiation treatment for gynecologic cancer. *Support Care in Cancer*. (2023) 31:388. doi: 10.1007/s00520-023-07849-6

9. Louwers YV, Visser JA. Shared genetics between age at menopause, early menopause, POI and other traits. *Front Genet*. (2021) 12:676546. doi: 10.3389/fgene.2021.676546

10. Ramirez J, Paris E, Basu S, Barua A. Abstract A015: age-associated molecular changes may predispose the ovary to Malignant transformation leading to ovarian cancer (OVCA). *Cancer Res*. (2023) 83:A015–5. doi: 10.1158/1538-7445.AGCA22-A015

11. Moorman PG, Calingaert B, Palmieri RT, Iversen ES, Bentley RC, Halabi S, et al. Hormonal risk factors for ovarian cancer in premenopausal and postmenopausal women. *Am J Epidemiol*. (2008) 167:1059–69. doi: 10.1093/aje/kwn006

12. Kim JM, Yang YS, Lee SH, Jee SH. Association between early menopause, gynecological cancer, and tobacco smoking: a cross-sectional study. *Asian Pacific J Cancer prevention: APJCP*. (2021) 22:3165–70. doi: 10.31557/APJCP.2021.22.10.3165

13. Long ME, Lee YS, Vegunta S. Cervical cancer screening in menopause: when is it safe to exit? *Menopause (New York NY)*. (2023) 30:972–9. doi: 10.1097/GME.0000000000002222

14. Shen Y, Xia J, Li H, Xu Y, Xu S. Human papillomavirus infection rate, distribution characteristics, and risk of age in pre- and postmenopausal women. *BMC Women's Health*. (2021) 21:80. doi: 10.1186/s12905-021-01217-4

15. Katagiri R, Iwasaki M, Abe SK, Islam MR, Rahman MS, Saito E, et al. Reproductive factors and endometrial cancer risk among women. *JAMA Netw Open*. (2023) 6:e2332296. doi: 10.1001/jamanetworkopen.2023.32296

16. Onstad MA, Schmandt RE, Lu KH. Addressing the role of obesity in endometrial cancer risk, prevention, and treatment. *J Clin oncology: Off J Am Soc Clin Oncol*. (2016) 34:4225–30. doi: 10.1200/JCO.2016.69.4638

17. Hernandez AV, Pasupuleti V, Benites-Zapata VA, Thota P, Deshpande A, Perez-Lopez FR. Insulin resistance and endometrial cancer risk: a systematic review and meta-analysis. *Eur J Cancer (Oxford England: 1990)*. (2015) 51:2747–58. doi: 10.1016/j.ejca.2015.08.031

18. Yang X, Wang J. The role of metabolic syndrome in endometrial cancer: a review. *Front Oncol*. (2019) 9:744. doi: 10.3389/fonc.2019.00744

19. Mu N, Zhu Y, Wang Y, Zhang H, Xue F. Insulin resistance: a significant risk factor of endometrial cancer. *Gynecologic Oncol*. (2012) 125:751–7. doi: 10.1016/j.ygyno.2012.03.032

20. Cheng Y, Zhang S, Qiang Y, Dong L, Li Y. Integrated bioinformatics data analysis reveals a risk signature and PKD1 induced progression in endometrial cancer patients with postmenopausal status. *Aging*. (2022) 14:5554–70. doi: 10.18632/aging.204168

21. Lyu YL, Geng L, Wang FX, Yang CL, Rong SJ, Zhou HF, et al. Comparative analysis of pre- and postmenopausal endometrial cancer in 216 patients. *Trans Cancer Res*. (2023) 12:595–604. doi: 10.21037/tcr-22-1616

22. Yuk JS, Lee SK, Uh JA, Seo YS, Kim M, Kim MS, et al. Skin cancer risk of menopausal hormone therapy in a Korean cohort. *Scientific Rep*. (2023) 13:10572. doi: 10.1038/s41598-023-37687-9

23. Xing Z, Alman AC, Kirby RS. Premature menopause and all-cause mortality and life span among women older than 40 years in the NHANES I epidemiologic follow-up study: propensity score matching analysis. *J Women's Health (Larchmont)*. (2023) 32:950–9. doi: 10.1089/jwh.2023.0189

24. Sun S, Du R, Wang S, Guo Y, He H, Wang X, et al. Age at menopause was not associated with microvascular complications in patients with type 2 diabetes mellitus. *Medicine*. (2023) 102:e34066. doi: 10.1097/MD.0000000000034066

25. Fain JA. NHANES. *Diabetes educator*. (2017) 43:151. doi: 10.1177/0145721717698651

26. Jimenéz-Salazar JE, Damian-Ferrara R, Arteaga M, Batina N, Damián-Matsumura P. Non-genomic actions of estrogens on the DNA repair pathways are associated with chemotherapy resistance in breast cancer. *Front Oncol*. (2021) 11:631007. doi: 10.3389/fonc.2021.631007

27. Chimento A, De Luca A, Avena P, De Amicis F, Casaburi I, Sirianni R, et al. Estrogen receptors-mediated apoptosis in hormone-dependent cancers. *Int J Mol Sci*. (2022) 23:1242. doi: 10.3390/ijms23031242

28. Al-Shami K, Awadi S, Khamees A, Alsheikh AM, Al-Sharif S, Ala' Bereshy R, et al. Estrogens and the risk of breast cancer: a narrative review of literature. *Heliyon*. (2023) 9:e20224. doi: 10.1016/j.heliyon.2023.e20224

29. Sun K, Zheng R, Lei L, Zhang S, Zeng H, Wang S, et al. Trends in incidence rates, mortality rates, and age-period-cohort effects of cervical cancer - China, 2003-2017. *China CDC weekly*. (2022) 4:1070–6. doi: 10.46234/ccdcw2022.216

30. Shen X, Cheng Y, Ren F, Shi Z. The burden of cervical cancer in China. *Front Oncol*. (2022) 12:979809. doi: 10.3389/fonc.2022.979809

31. Mahdy H, Casey MJ, Vadakekut ES, Crotzer D. Endometrial cancer. In: *StatPearls*. StatPearls Publishing Copyright B, Treasure Island (FL) (2024). ineligible companies. Disclosure: Murray Casey declares no relevant financial relationships with ineligible companies. Disclosure: Elsa Vadakekut declares no relevant financial relationships with ineligible companies. Disclosure: David Crotzer declares no relevant financial relationships with ineligible companies.

32. Wu Y, Sun W, Liu H, Zhang D. Age at menopause and risk of developing endometrial cancer: a meta-analysis. *BioMed Res Int*. (2019) 2019:8584130. doi: 10.1155/2019/8584130

33. Cheng G, Wang M, Sun H, Lai J, Feng Y, Liu H, et al. Age at menopause is inversely related to the prevalence of common gynecologic cancers: a study based on NHANES. *Front endocrinology*. (2023) 14:1218045. doi: 10.3389/fendo.2023.1218045

34. Petrucelli N, Daly MB, Pal T. BRCA1- and BRCA2-associated hereditary breast and ovarian cancer. In: Adam MP, Feldman J, Mirzaa GM, Pagon RA, Wallace SE, Amemiya A, editors. *GeneReviews(B.)*. University of Washington, Seattle Copyright B, Seattle (WA (1993).

35. Karst AM, Drapkin R. Ovarian cancer pathogenesis: a model in evolution. *J Oncol*. (2010) 2010:932371. doi: 10.1155/2010/932371

36. Kroeger PT Jr., Drapkin R. Pathogenesis and heterogeneity of ovarian cancer. *Curr Opin obstetrics gynecology*. (2017) 29:26–34. doi: 10.1097/GCO.0000000000000340

37. Hyvärinen M, Juppi HK, Taskinen S, Karppinen JE, Karvinen S, Tammelin TH, et al. Metabolic health, menopause, and physical activity-a 4-year follow-up study. *Int J obes (London)*. (2022) 46:544–54. doi: 10.1038/s41366-021-01022-x

38. Janssen I, Powell LH, Crawford S, Lasley B, Sutton-Tyrrell K. Menopause and the metabolic syndrome: the Study of Women's Health Across the Nation. *Arch Internal Med*. (2008) 168:1568–75. doi: 10.1001/archinte.168.14.1568

39. Crudele L, Piccinin E, Moschetta A. Visceral adiposity and cancer: role in pathogenesis and prognosis. *Nutrients*. (2021) 13(6):2101. doi: 10.3390/nu13062101

40. Pan K, Chlebowski RT, Mortimer JE, Gunter MJ, Rohan T, Vitolins MZ, et al. Insulin resistance and breast cancer incidence and mortality in postmenopausal women in the Women's Health Initiative. *Cancer*. (2020) 126:3638–47. doi: 10.1002/cncr.v126.16

41. Chakraborty D, Benham V, Bullard B, Kearney T, Hsia HC, Gibbon D, et al. Fibroblast growth factor receptor is a mechanistic link between visceral adiposity and cancer. *Oncogene*. (2017) 36:6668–79. doi: 10.1038/onc.2017.278

42. Jung SY, Papp JC, Sobel EM, Pellegrini M, Yu H, Zhang ZF. Pro-inflammatory cytokine polymorphisms and interactions with dietary alcohol and estrogen, risk factors for invasive breast cancer using a post genome-wide analysis for gene-gene and gene-lifestyle interaction. *Sci Rep*. (2021) 11:1058. doi: 10.1038/s41598-020-80197-1

43. Masala G, Bendinelli B, Della Bella C, Assedi M, Tapinassi S, Ermini I, et al. Inflammatory marker changes in a 24-month dietary and physical activity randomised intervention trial in postmenopausal women. *Scientific Rep*. (2020) 10:21845. doi: 10.1038/s41598-020-78796-z

44. Agnoli C, Grioni S, Pala V, Allione A, Matullo G, Gaetano CD, et al. Biomarkers of inflammation and breast cancer risk: a case-control study nested in the EPIC-Varese cohort. *Sci Rep*. (2017) 7:12708. doi: 10.1038/s41598-017-12703-x

45. Grady D, Gebretsadik T, Kerlikowske K, Ernster V, Petitti D. Hormone replacement therapy and endometrial cancer risk: a meta-analysis. *Obstetrics gynecology*. (1995) 85:304–13. doi: 10.1016/0029-7844(94)00383-O

46. Sjögren LL, Mørch LS, Løkkegaard E. Hormone replacement therapy and the risk of endometrial cancer: a systematic review. *Maturitas*. (2016) 91:25–35. doi: 10.1016/j.maturitas.2016.05.013

47. Sapre S, Thakur R. Lifestyle and dietary factors determine age at natural menopause. *J Mid-Life Health*. (2014) 5:3–5. doi: 10.4103/0976-7800.127779

48. Dekker J, Karchoud J, Braamse AMJ, Buiting H, Konings IRHM, van Linde ME, et al. Clinical management of emotions in patients with cancer: introducing the approach "emotional support and case finding. *Trans Behav Med*. (2020) 10:1399–405. doi: 10.1093/tbm/ibaa115

49. Semenenko E, Banerjee S, Olver I, Ashinze P. Review of psychological interventions in patients with cancer. *Supportive Care cancer: Off J Multinational Assoc Supportive Care Cancer*. (2023) 31:210. doi: 10.1007/s00520-023-07675-w

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 31 March 2025
DOI 10.3389/fonc.2025.1541572

OPEN ACCESS

EDITED BY
Sandeep Kumar Mishra,
Yale University, United States

REVIEWED BY
Hina Sultana,
University of North Carolina System, United States
Junxiang Huang,
Boston College, United States
Behnaz Jahanbin,
Tehran University of Medical Sciences, Iran
Roberto Altamirano,
University of Chile, Chile

*CORRESPONDENCE
Mengmeng Chen
1277556339@qq.com
Yali Chen
Yalichen182@163.com

RECEIVED 08 December 2024
ACCEPTED 13 March 2025
PUBLISHED 31 March 2025

CITATION
Chen M, Han L, Wang Y, Qiu Q, Chen Y and Zheng A (2025) The prognostic value of growth pattern-based grading for mucinous ovarian carcinoma (MOC): a systematic review and meta-analysis. *Front. Oncol.* 15:1541572. doi: 10.3389/fonc.2025.1541572

The prognostic value of growth pattern-based grading for mucinous ovarian carcinoma (MOC): a systematic review and meta-analysis

Mengmeng Chen[1,2]*, Ling Han[1,2], Yisi Wang[1,2], Qi Qiu[1,2], Yali Chen[1,2]* and Ai Zheng[1,2]

[1]Department of Obstetrics and Gynecology, West China Second University Hospital, Sichuan University, Sichuan, China, [2]Key Laboratory of Birth Defects and Related Diseases of Women and Children (Sichuan University), Ministry of Education, Sichuan, China

Objective: To investigate the prognostic significance of expansile and infiltrative growth patterns in mucinous ovarian carcinoma (MOC).

Methods: A systematic search was conducted in the PubMed, Embase, and Web of Science databases for studies published between January 1, 2010, and September 6, 2024, examining the correlation between expansile and infiltrative tumor growth patterns and prognosis in MOC. Subgroup analyses were performed for mortality, recurrence, and FIGO stage I based on tumor subtype. The Chi-square test was used to evaluate the distribution of expansile and infiltrative tumors across FIGO stages I-IV.

Results: Twelve eligible studies, comprising a total of 1185 patients, were included in this systematic review and meta-analysis. The combined death rate in the expansile and infiltrative MOC was 10.5% (95%CI: 6.2-15.7) and 31.1% (95% CI: 14.1-50.9). The combined recurrence rate in the expansile and infiltrative MOC was 6.9% (95%CI: 3.1-11.9) and 24.5% (95%CI: 14.3-36.2). The combined International Federation of Gynecology and Obstetrics (FIGO) I rate in the expansile and infiltrative MOC was 89.8% (95%CI: 84.9-94.0) and 56.2% (95% CI: 41.5-70.4). A significant association was found between tumor type and FIGO stage (χ^2 (3) = 110.92, $p < 0.00001$).

Conclusion: Expansile MOC predicts better outcomes, while infiltrative MOC is linked to advanced stages and poorer prognosis. Complete surgical staging is crucial for infiltrative MOC but optional for early-stage expansile MOC. Early-stage patients should consider fertility-sparing surgery, timely conception, and close recurrence monitoring.

KEYWORDS

mucinous ovarian carcinoma, pattern-based grading, expansile, infiltrative, prognosis, meta-analysis

1 Introduction

Ovarian cancer is the second most common gynecological malignancy (1). Among its various subtypes, high-grade serous ovarian carcinoma (HGSC) is the most prevalent histological subtype, while mucinous ovarian carcinoma (MOC) is quite rare, constituting approximately 3% to 11% of ovarian cancers (2, 3). MOC is recognized as a distinct entity from other epithelial ovarian cancers (EOCs), exhibiting a unique natural history, molecular profile, chemo-sensitivity, and prognosis compared to HGSC. Notably, MOC is the most common subtype in women under 40 (4), with tobacco smoking identified as the only significant risk factor (5). While most HGSC cases are diagnosed at advanced stages, 80% of MOC cases are identified at stage I (6). Early-stage MOC typically exhibits a better prognosis, however, advanced cases face poorer outcomes, primarily due to a limited response to platinum-based chemotherapy compared to HGSC (7, 8).

Histological grading systems, such as the International Federation of Gynecology and Obstetrics (FIGO) and Silverberg grading systems, have been studied in relation to the ovarian cancer patient prognosis, including MOC (9, 10). As yet, these grading systems alone are insufficient for predicting the clinical course of MOC, unlike their application for other ovarian carcinoma subtypes (11). In 2014, in order to standardize the pathological reporting of gynecological tumors, World Health Organization (WHO) guidelines proposed classifying the mucinous cancers in these two groups based on their growth patterns, calling them expansileand infiltrative-type tumors (12), which was also entered in the newest version CAP protocols (13). However, there is controversy over the treatment of this histological groups using different compasses. Guidelines from the European Society for Medical Oncology and the European Society of Gynecological Oncology (ESMO-ESGO) emphasize the importance of adjuvant chemotherapy for stage IB-IC infiltrative MOC. Even for stage IA, adjuvant chemotherapy may be considered for patients with infiltrative patterns, whereas it is not deemed necessary for stage IA expansile MOC (14, 15). Conversely, the National Comprehensive Cancer Network (NCCN) guidelines do not recommend differentiating histologic subtypes when treating patients with MOC. Instead, they advise administering adjuvant chemotherapy for stage IC or higher MOC, while treatment can be avoided for stage IA-IB, similar to other EOCs (16).

Therefore, we conducted a meta-analysis and systematic review aimed at assessing the prognostic significance of the expansile and infiltrative growth patterns in MOC. This study seeks to provide clearer guidance for the treatment of MOC and improve clinical management and outcomes for patients.

2 Methods

2.1 Protocol registration

This meta-analysis was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (17). Prior to data extraction, the review was registered with the International Prospective Register of Systematic Reviews (PROSPERO) under registration number CRD42024585615.

2.2 Eligibility criteria and exclusion criteria

2.2.1 Eligibility criteria

To be eligible, we aimed for the following inclusion criteria: 1) The study design is a retrospective or prospective study design;2) Included cases need to be classified by expansile or infiltration subtype, and need to be confirmed the diagnosis of MOC;3) Included articles assess at least one of the following parameters: death, recurrence, FIGO I or FIGO stage.

2.2.2 Exclusion criteria

We excluded studies with the following exclusion criteria:1) Reviews, letters, case reports or editorial comments;2) Studies without full text, insufficient data or low-quality scores based on Newcastle- Ottawa Scale (NOS) (18);3) Republished literature or repetitive studies.

2.3 Search strategy

Two researchers (MMC and YSW) conducted a comprehensive search in electronic databases of PubMed, Embase, and Web of Science for relevant researches, published for from January 1, 2010 to September 6, 2024.

The following search terms were used to identify relevant studies on ovarian cancer: “Carcinoma, Ovarian Epithelial”, “Epithelial Carcinoma, Ovarian”, “Ovarian Epithelial Carcinomas”, whereas the following terms were used to identify relevant studies on expansile and infiltrative: “expansile”, “infiltrative”.

Two researchers (LH and YLC) thoroughly reviewed the reference lists of all included articles to identify any potentially missing studies or unpublished data. In cases where multiple studies analyzed overlapping patient populations, the most recent or comprehensive results were selected. Following the initial screening, the full texts of all potential articles were independently reviewed by two researchers (QQ and MMC) for further evaluation. Any disagreements were resolved through discussion with AZ.

2.4 Data extraction

Data were independently extracted by two investigators (QQ and YSW), with any disagreements resolved through discussion with AZ. The extracted data included author, publication date, country, number of cases, growth patterns (expansile and infiltrative), oncological outcomes (death, recurrence), and pathological characteristics (FIGO stage). Attempts were made to obtain missing data by contacting the authors via email; however, no responses were received.

2.4.1 Expansile and infiltrative pattern

In expansile tumor, the tumor consists of a confluent glandular growth pattern with minimal to no stromal invasion. In contrast, infiltrative tumor shows malignant cell clusters with destructive stromal invasion (12).

2.4.2 Oncological outcomes

Death was calculated from the data from surgery to either the last follow-up or the data of death. Recurrence refers as either pathologic proof of cancer or an imaging study showing the regrowth of the tumor, whether it is confined to the pelvic region or outside of it.

2.4.3 Pathological features

For mucinous ovarian carcinoma, Stage I means tumor confined to the ovaries, Stage II means tumor involves one or both ovaries and extends to other pelvic tissues, such as the uterus or fallopian tubes. Stage III means tumor is present in one or both ovaries and has spread to the peritoneum outside the pelvis or to regional lymph nodes. Stage IV means tumor has spread beyond the peritoneum to distant organs, such as the liver or lungs.

2.5 Quality assessment

Two reviewers (MMC and YSW) independently assessed the quality of the included studies, with disagreements resolved through discussion. The quality of each study was evaluated using the Newcastle-Ottawa Scale (NOS), which assesses three categories: case selection, comparability between groups, and outcome assessment. The total NOS score ranges from 0 to 9 points, and studies with a score of ≥6 were considered high-quality and included in our analysis.

2.6 Statistical analysis

Meta-analysis was performed by using STATA 15.0 software. Subgroup analyses were based on expansile and infiltrative pattern, and heterogeneity was determined using orthorhombic test and I^2 statistic. If there was significant heterogeneity (p-value <0.05 or I^2 >50%), a random-effects model was used. Otherwise, a fixed-effect model was used (19). Additionally, a Chi-Square Test was performed to evaluate whether there were statistical differences in the distribution of expansile tumors and infiltrative tumors across stages I, II, III, and IV. Sensitivity analysis to determine the robustness and stability of the results, calculating the herogeneity in each situation in which a single study was removed in turn in noder to evaluate the effect of a single study on the overall outcome. Risk of publication was assessed by visual inspeciton of Begg's funnel plot.

3 Result

3.1 Study selection and characteristics

The initial search retrieved a total of 592 relevant studies from three databases (PubMed = 423, Embase = 132, Web of Science = 37). After removing duplicates and screening titles and abstracts, 27 studies remained. Following a full-text evaluation, 15 studies were excluded. Ultimately, 12 studies, including 1185 patients, met the inclusion criteria and were included in this meta-analysis. A flowchart of the selection process is provided in Figure 1.

All included studies were retrospective and received seven or more stars based on the NOS criteria. The quality assessments of these studies are presented in Table 1, while the general characteristics of the studies included in this meta-analysis are summarized in Table 2.

3.2 Subgroup analysis based on expansile and infiltration tumors.

3.2.1 Death

This meta-analysis of five studies (9, 20, 21, 26, 30) showed that the combined death rate of mucinous ovarian carcinoma was positively correlated with expansile patter (Effect Size=0.105, 95% CI=0.062-0.157, I^2 = 42.001%, n=5), while no significant correlation for infiltrative pattern (Effect Size=0.311, 95%CI=0.141-0.509,I^2 = 78.323%, n=5) Figure 2A. However, the results also indicated high heterogeneity among the studies (I^2 = 80.256%, p<0.05). In order to assess the stability of the model, sensitivity analysis was conducted by excluding each individual study and calculating new effect size. The results showed that the effect size were relatively stable, as illustrated in Figure 2B.

3.2.2 Recurrence

This meta-analysis of eight studies (9, 20, 21, 23–25, 27, 28) showed that the combined recurrence of mucinous ovarian carcinoma was positively correlated with expansile pattern (Effect Size=0.069, 95%CI=0.031-0.119, I^2 = 55.150%, n=8), negatively correlated with infiltrative pattern (Effect Size=0.245, 95% CI=0.143-0.362,I^2 = 79.797%, n=8) Figure 3A. The findings also revealed significant heterogeneity among the studies (I^2 = 80.408%, p<0.05). A sensitivity analysis was performed by omitting each study individually and recalculating the effect size to evaluate model stability. The results indicated that the effect sizes remained fairly stable, as shown in Figure 3B.

3.3.3 FIGO I and FIGO stage

Given that most MOC cases are diagnosed at an early stage, we selected FIGO stage I as one of the key pathological features in our study and found eight studies (Table 3) (9, 21, 22, 24, 25, 27, 29, 30) reported the association between the expansile and infiltrative

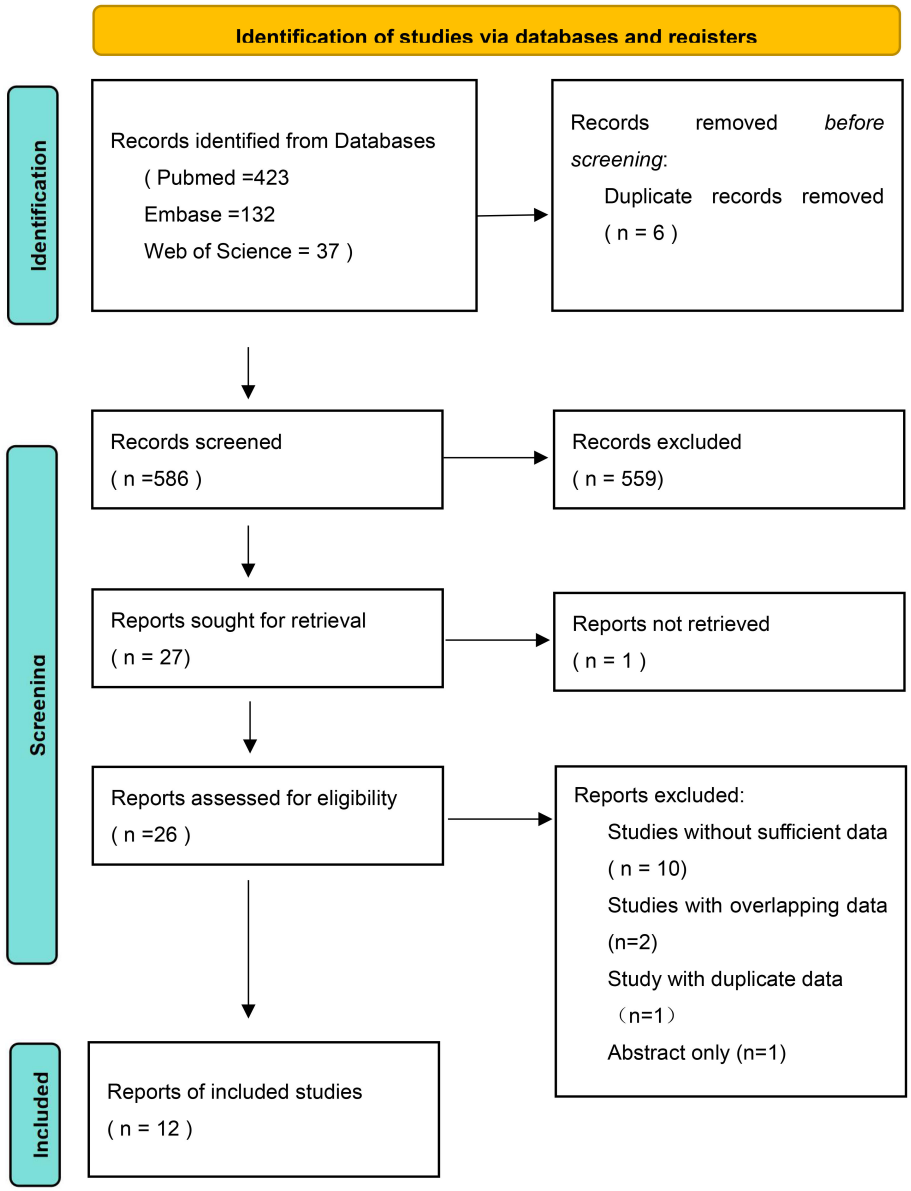


FIGURE 1
Flow diagram of the included studies.

pattern for mucinous ovarian carcinoma and FIGO I stage. The result revealed that the combined FIGO I stage rate of mucinous ovarian carcinoma was positively correlated with expansile pattern (Effect Size=0.898, 95%CI=0.849-0.940, I^2 = 53.137%, n=8), negatively correlated with infiltrative pattern (Effect Size=0.562, 95%CI=0.415-0.704, I^2 = 82.519%, n=8) Figure 4A. Moreover, the results highlighted considerable heterogeneity across the studies (I^2 = 90.752%, p<0.05). To evaluate the robustness of the model, a sensitivity analysis was carried out by removing each study one at a time and recomputing the effect size. The findings suggested that the effect sizes were largely consistent, as depicted in Figure 4B.

Besides, we use the Pearson Chi-Square test to evaluate the distribution of FIGO stages I, II, III, IV among expansile and infiltrative tumors, and found there was a highly significant association between tumor type and FIGO staging (Pearson chi2 (3) = 110.9206, p <0.00001) Figure 4D.

3.3.4 Publication bias

Publication bias was investigated by Begg's funnel plots. Visual inspection of the Begger's funnel plot was almost symmetrical, as depicted in Figures 2C, 3C, 4C, suggesting no obvious evidence of publication bias.

TABLE 1 Quality assessment of included studies.

Study	Selection				Comparability			Outcome		Total
	Representativeness	Selection of non-exposed	Ascertainment of exposure	Outcome not present at start	Comparability on most important factors	Comparability on other risk factors	Assessment of outcome	Long enough follow-up (median>=5 year)	Adequacy (completeness of follow-up)	
Gouy S (20)	✓	✓	✓	✓	✓	×	✓	✓	✓	8
Lim H (21)	✓	✓	✓	✓	✓	×	✓	×	✓	7
Hada T (22)	✓	✓	✓	✓	✓	×	✓	×	✓	7
Tabrizi AD (23)	✓	✓	✓	✓	✓	×	✓	×	✓	7
Sotiropoulou M (24)	✓	✓	✓	✓	✓	×	✓	✓	✓	8
Algera MD (25)	✓	✓	✓	✓	✓	×	✓	×	✓	7
Meagher N (26)	✓	✓	✓	✓	✓	×	✓	×	✓	7
Huin M (27)	✓	✓	✓	✓	✓	✓	✓	✓	✓	8
Muyldermans K (9)	✓	✓	✓	✓	✓	×	✓	✓	✓	8
Hada T (28)	✓	✓	✓	✓	✓	×	✓	×	✓	7
Nistor S (29)	✓	✓	✓	✓	✓	×	✓	×	✓	7
Köbel M (30)	✓	✓	✓	✓	✓	×	✓	×	✓	7

"√" indicates that the criteria are met, while "×" indicates that the criteria are not met.

10.3389/fonc.2025.1541572

TABLE 2 The basic characteristics of included studies.

First author	Publish year	Study period	Region	Study design	Cases	Follow up	Quality
Gouy S (20)	2018	1976-2016	France	R	64	62m	8
Lim H (21)	2023	2003-2021	Korea	R	113	55m	7
Hada T (22)	2022	1984-2019	Japan	R	52	54m	7
Tabrizi AD (23)	2010	1984-2000	Iran	R	31	NM	7
Sotiropoulou M (24)	2013	1998-2008	Greece	R	42	6y	8
Algera MD (25)	2024	2015-2020	Netherlands	R	409	999d	7
Meagher N (26)	2021	NM	Australia	R	133	2y	7
Huin M (27)	2022	2001-2019	France	R	94	5y	8
Muyldermans K (9)	2013	1993-2011	Belgium	R	44	63m	8
Hada T (28)	2021	1984-2018	Japan	R	49	NM	7
Nistor S (29)	2023	2010-2022	UK	R	33	37m	7
Köbel M (30)	2024	NM	Canada	R	121	NM	7

"d" means day, "m" means month and "y" means year. "R" means retrospective. "NM" means not mentioned.

4 Discussion

This meta-analysis revealed that mucinous ovarian carcinoma with expansile-pattern tumors, typically observed in early-stage, tend to have a better oncological prognosis. In contrast, infiltrative-pattern tumors are commonly associated with advanced stages and are linked to poorer outcomes.

Our study indicated that patients with expansile pattern tumors have lower death rate, recurrence rate and a higher proportion of FIGO stage I compared to those with infiltrative tumors. A study conducted by Taira Hada et al. (22) showed that MOC with expansile invasion was a better prognostic factor for progression-free-survival and overall survival than HGSC at all stage. Besides, Taira Hada et al. (31) also conducted a study, and found there was no statistically significant differences in the recurrence rate and prognosis of MOC with expansile and mucinous borderline tumors, it might be possible that expansile MOC biologically behave more like mucinous borderline tumors. These studies suggest that expansile MOC is not an aggressive subtype, leading many researchers to question whether comprehensive staging surgery is necessary for early-stage expansile tumors. Marc D et al. (25) conducted a study of peritoneal staging in clinical early-stage MOC, found limited benefit for routinely performing peritoneal and lymph node staging procedures in patients with expansile tumors, because recurrences, overall survival and recurrence free survival were similar across the different extent of surgical staging groups. In another study (15), researchers concluded that peritoneal metastases are rare in expansile MOC, more than 90% of patients have early-stage disease. Gouy S et al. (32) describes no lymph node involvement in expansile tumors, while one patient upstaged after surgical staging, based on positive peritoneal cytology (3.4%, one out of 29 patients). In conclusion, expansile is a less aggressive pattern. For patients with early-stage expansile MOC, it may be considered safe to forgo additional staging surgery and lymph node sampling following the initial bilateral salpingo-oophorectomy and hysterectomy. Nevertheless, further data is needed to validate this observation and ensure that patient outcomes are not compromised.

In contrast, infiltrative tumors are typically associated with more advanced stages and higher recurrence rates than expansile tumors. Gouy S et al. (20) found lethal recurrences were observed mainly in infiltrative type. Taira Hada et al. (22) reported that univariate analysis showed that MOC with infiltrative invasion at FIGO stages II to IV had worse progression free survival and overall survival than HGSC. Due to the high recurrence rate, it might be considered adjuvant treatment for infiltrative tumor, even in early-stage. According to Lim H et al. (21), one-third of patients who received lymphadenectomy had lymph node involvement. Gouy S et al. (32) investigated 31 infiltrative MOC underwent staging operations and found four patients had nodal involvement. Hence, we suggest lymphadenectomy must be considered during staging operations in patients with infiltrative tumor. Algera MD et al. (15) concluded that upstaging clinical early-stage infiltrative MOC based on positive cytology, peritoneum and omentum metastases occurred in 10.3% of the patients. Besides, Marc D et al. (25) conducted a study of peritoneal staging in clinical early-stage MOC, found that in the infiltrative cohort, overall survival was better for patients undergoing full staging compared with those undergoing fertility sparing or partial staging, patients undergoing fertility-sparing staging for infiltrative tumors experienced significantly more recurrences. In conclusion, patients diagnosed with infiltrative mucinous ovarian carcinoma (MOC) should undergo a thorough surgical staging process. This process should include peritoneal staging, which involves omentectomy, the collection of peritoneal washings, and the acquisition of biopsies, along with pelvic and para-aortic lymph node sampling. Given the potential aggressiveness of this type of cancer, adjuvant treatment should be considered even for tumors identified at an early stage.

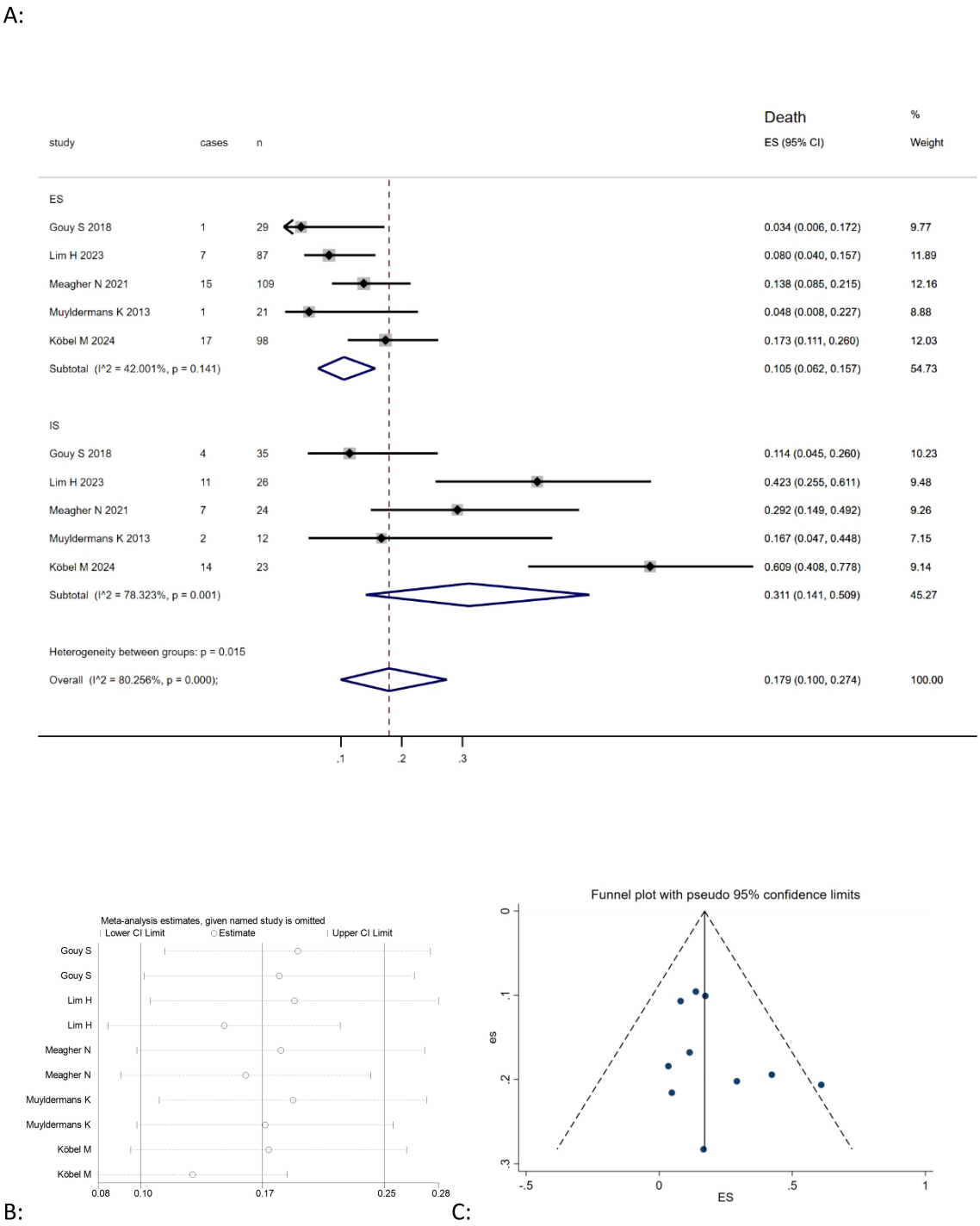


FIGURE 2
(A) Forest plots showing the relationship between infiltrative subtype, expansile subtype, and death rate in MOC; **(B)** sentivity analysis to evaluate robustness and **(C)** funnel plots show publication bias by visual inspection.

In recent years, research on the molecular characteristics of mucinous ovarian cancer (MOC) has increased, providing new insights into its invasion patterns and prognosis. A study found that mucinous ovarian cancer (MOC) with infiltrative invasion was more often positive for CK5/6, CD24, and EGFR, suggesting that these markers may be linked to the aggressive features of this invasion pattern (28). In contrast, expansile invasion showed a higher prevalence of HER2 overexpression/amplification and less frequent HER2 mutation compared to infiltrative MOC, although this difference was not statistically significant (33). Additionally, PAX8 expression was more commonly associated with expansile invasion, but the difference was not statistically significant (75% vs

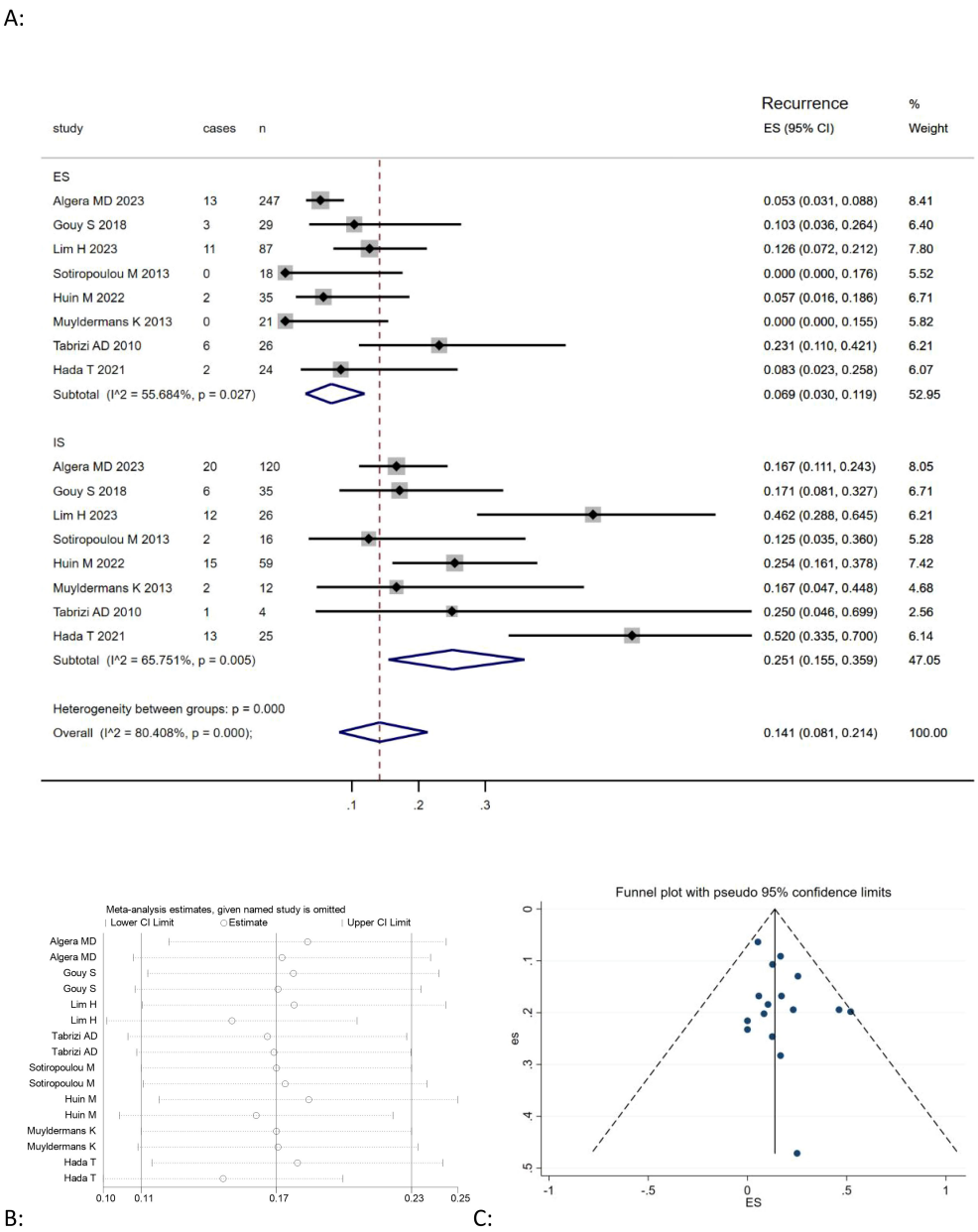
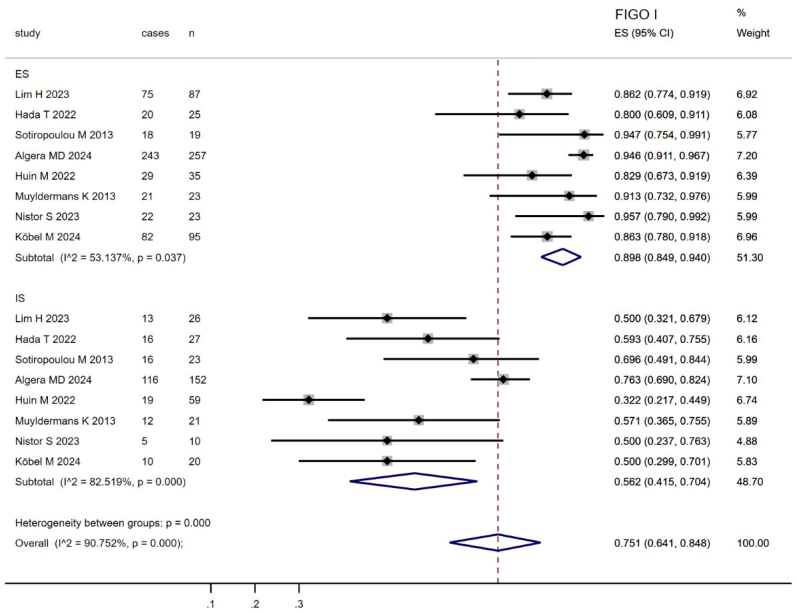


FIGURE 3
(A) Forest plots showing the relationship between death rate and infiltrative subtype, expansile subtype; **(B)** sentivity analysis to evaluate robustness and **(C)** funnel plots show publication bias by visual inspection.

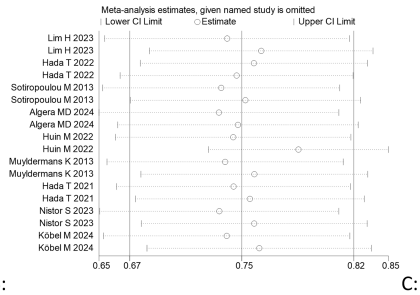
TABLE 3 Distribution of expansile and infiltrative MOC patients across FIGO stages I-IV in various studies.

	Expansile Tumor Stage				Infiltrative Tumor Stage			
	I	II	III	IV	I	II	III	IV
Algera MD (25)	243	6	7	1	116	7	23	2
Lim H (21)	75	3	5	4	13	0	8	5
Hada T (22)	20	2	1	2	16	1	7	3
Huin M (27)	28	1	3	0	19	0	27	9
Nistor S (29)	22	2	0	–	5	3	2	–
Köbel M (30)	82	9	3	1	10	3	6	1

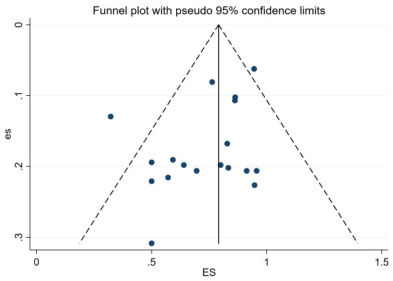
A:



B:



C:



D:

Type	Stage I	II	Total
Expansile tumor	470	23	520
Infiltrative tumor	179	14	286
Total	649	37	806

Type	Stage III	IV	Total
Expansile tumor	19	8	520
Infiltrative tumor	73	20	286
Total	92	28	806

Pearson chi2(3) = 110.9206 Pr = 0.000

FIGURE 4
(A) Forest plots showing the relationship between FIGO I rate and infiltrative subtype, expansile subtype; **(B)** sentivity analysis to evaluate robustness and **(C)** funnel plots show publication bias by visual inspection; **(D)** Cross-tabulation of the distribution of expansile and infiltrative MOC by FIGO stage (I-IV).

37.5%, p=0.99) (29). Overall, the existing data are limited, highlighting the need for further research to integrate molecular data with histological classification for a comprehensive understanding of MOC prognosis.

Fertility-sparing surgery (FSS) is a common topic of discussion because patients diagnosed with MOC are often younger. In recent years, preserving fertility becomes a significant concern in treatment planning, and several studies have focused on the outcomes of fertility-sparing surgery in patients with early-stage MOC. Gouy S et al. (34) conducted a study and emphasized that FFS should be considered for early-stage MOC regardless of its subtype. Similarly, Yoshihara M et al. (35) found patients with stage I MOC underwent uterus preserving surgery was not associated with decreased survival. On the other hand, Hyunji Lim et al. (21)

found infiltrative tumors showed no statistical significance with worse survival, but patients in the infiltrative tumors group who underwent FSS demonstrated a 5-year progression free survival rate of 83.3%, significantly lower than patients without fertility preservation. This suggests that adjuvant chemotherapy should be considered for patients with stage I disease who have undergone FSS, particularly if the histologic subtype is infiltrative. Bentivegna et al. Reported the long-term outcome of 280 MOC patients treated with FFS, the recurrence rate was 6,8% (36). Additional, Wei Lin et al. (37) noted no significant difference in disease-free survival between the FSS and radical surgery groups in patients with stage IA and IC disease, though the FSS group did show a trend toward poorer disease-free survival compared to those who underwent radical surgery. Besides, they found that, among 23 patients diagnosed with early-stage mucinous ovarian carcinoma who underwent fertility-sparing surgery (FSS) and attempted to conceive, 21 (91.3%) successfully achieved 27 pregnancies. These included 26 spontaneous pregnancies and one pregnancy resulting from assisted reproductive technology. However, there is a lack of data on the recurrence rates associated with FSS, highlighting the need for further research in this area. More studies should be conducted to better understand the long-term outcomes and potential risks of recurrence following FSS in patients with mucinous ovarian carcinoma. But we strongly recommend FSS for patients with early-stage MOC, irrespective of the tumor subtype. This approach aims to preserve fertility while effectively treating the cancer. Following treatment, these patients should be encouraged to attempt conception as soon as they are medically cleared and should engage in regular follow-up to monitor for any signs of relapse.

This meta-analysis is the first to evaluate the relationship between growth patterns and prognosis in MOC, but it has limitations. One of the most obvious limitation is the high heterogeneity among the results, although we did sensitivity analysis to explain its robustness, we are currently unable to perform a more thorough investigation into the sources of heterogeneity due to incomplete data. All included studies were retrospective, which may affect the results. Additionally, only English language studies were considered, potentially introducing language bias. The subgroup analysis did not show a significant link between infiltrative patterns and death rate due to limited data. Despite these limitations, the study offers initial insights into the prognostic importance of growth patterns in MOC and suggests areas for future research, calling for more studies, including those with negative findings, to support these conclusions.

5 Conclusion

Our study found that expansile MOC generally has better outcomes, while infiltrative MOC is associated with poorer prognosis and advanced stages. Full surgical staging is recommended for infiltrative MOC, but may be omitted for early-stage expansile MOC. Fertility-sparing surgery is advised for early-stage patients, with early conception and close monitoring.

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/s.

Author contributions

MC: Conceptualization, Writing – original draft. LH: Data curation, Writing – review & editing. YW: Formal Analysis, Investigation, Methodology, Writing – original draft. QQ: Writing – review & editing. YC: Writing – review & editing. AZ: Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research 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 no Generative AI was used in the creation of this manuscript.

Publisher's note

References

1. Bray F, Laversanne M, Bhoo-Pathy N, Ho FDV, Feliciano EJG, Eala MAB. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. (2024) 74:229–63. doi: 10.3322/caac.21834

2. Kobel M, Kalloger SE, Huntsman DG, Santos JL, Swenerton KD, Seidman JD, et al. Cheryl Brown Ovarian Cancer Outcomes Unit of the British Columbia Cancer Agency, Vancouver BC. Differences in tumor type in low-stage versus high-stage ovarian carcinomas. *Int J Gynecol Pathol*. (2010) 29:203–11. doi: 10.1097/PGP.0b013e3181c042b6

3. Matz M, Coleman MP, Sant M, Chirlaque MD, Visser O, Gore M, et al. The histology of ovarian cancer: worldwide distribution and implications for international survival comparisons (CONCORD-2). *Gynecol Oncol*. (2017) 144:405–13. doi: 10.1016/j.ygyno.2016.10.019

4. Yoshikawa N, Kajiyama H, Mizuno M, Shibata K, Kawai M, Nagasaka T, et al. Clinicopathologic features of epithelial ovarian carcinoma in younger vs. older patients: Analysis in Japanese women. *J Gynecol Oncol*. (2017) 25(2):118–23. doi: 10.3802/jgo.2014.25.2.118

5. Gates MA, Rosner BA, Hecht JL, Tworoger SS. Risk factors for epithelial ovarian cancer by histologic subtype. *Am J Epidemiol*. (2017) 171(1):45–53. doi: 10.1093/aje/kwp314

6. Seidman JD, Horkayne-Szakaly I, Haiba M, Boice CR, Kurman RJ, Ronnett BM. The histologic type and stage distribution of ovarian carcinomas of surface epithelial origin. *Int J Gynecol Pathol*. (2004) 23:41–4. doi: 10.1097/01.pgp.0000101080.35393.16

7. Xu W, Rush J, Rickett K, Coward JIG. Mucinous ovarian cancer: A therapeutic review. *Crit Rev Oncol Hematol*. (2016) 102:26–36. doi: 10.1016/j.critrevonc.2016.03.015

8. Morice P, Gouy S, Leary A. Mucinous ovarian carcinoma. *N Engl J Med*. (2019) 380:1256–66. doi: 10.1056/NEJMra1813254

9. Muyldermans K, Moerman P, Amant F, Leunen K, NEven P, Vergote I, et al. Primary invasive mucinous ovarian carcinoma of the intestinal type: Importance of the expansile versus infiltrative type in predicting recurrence and lymph node metastases. *Eur J Cancer*. (2013) 49:1600–08. doi: 10.1016/j.ejca.2012.12.004

10. Busca A, Nofech-Mozes S, Olkhov-Mitsel E, Gien LT, Bassiouny D, Mirkovic J, et al. Histological grading of ovarian mucinous carcinoma - an outcome-based analysis of traditional and novel systems. *Histopathology*. (2020) 77:26–34. doi: 10.1111/his.14039

11. WHO Classification of Tumours Editorial Board. *Female genital tumours. WHO classification of tumours. 5th ed*. Lyon: International Agency for Research on Cancer (2020).

12. Kurman RJ, Carcangiu ML, Herrington CS, et al. *WHO Classification of tumours of female reproductive organs. 4th ed*. Lyon, France: IARC (2014).

13. College of American Pathologists. *CAP Cancer Protocols: Ovarian Mucinous Carcinoma*. (2024).

14. Colombo N, Sessa C, du Bois A, Ledermann J, McCluggage WG, McNeish I, et al. ESMO-ESGO consensus conference recommendations on ovarian cancer: pathology and molecular biology, early and advanced stages, borderline tumours and recurrent disease†. *An Oncol*. (2019) 30:672–705. doi: 10.1093/annonc/mdz062

15. Algera MD, van Driel WJ, van de Vijver KK, Kruitwagen RFPM, Lok CAR. Surgical treatment for clinical early-stage expansile and infiltrative mucinous ovarian cancer: can staging surgeries safely be omitted? *Curr Opin Oncol*. (2022) 34:497–503. doi: 10.1097/CCO.0000000000000862

16. Armstrong DK, Alvarez RD, Backes FJ, et al. NCCN guidelines® Insights: ovarian cancer, version 3.2022. *J Natl Compr Canc Netw*. (2022) 20:972–80. doi: 10.6004/jnccn.2022.0047

17. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. PRISMA-P Group.Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev*. (2015) 4:1. doi: 10.1186/2046-4053-4-1

18. Wells G, Wells G, Shea B, Shea B, O'Connell D, Peterson J, et al. *The Newcastle-Ottawa Scale (NOS) for Assessing the Quality of Nonrandomised Studies in Meta-Analyses*. (2014).

19. DerSimonian R, Laird N. Meta-analysis in clinical trials revisited, Contemp. *Clin Trials*. (2015) 45:139–45.

20. Gouy S, Saidani M, Maulard A, Bach Hamba S, Bentivegna E, Leary A, et al. Characteristics and prognosis of stage I ovarian mucinous tumors according to expansile or infiltrative type. *Int J Gynecol Cancer*. (2018) 28:493–99. doi: 10.1097/IGC.0000000000001202

21. Lim H, Ju Y, Kim SI, Park JH, Kim HS, Chung HH, et al. Clinical implications of histologic subtypes on survival outcomes in primary mucinous ovarian carcinoma. *Gynecol Oncol*. (2023) 177:117–24. doi: 10.1016/j.ygyno.2023.08.013

22. Hada T, Miyamoto M, Ishibashi H, Ishibashi H, Matsurra H, Kakimoto S, et al. Comparison of clinical behavior between mucinous ovarian carcinoma with infiltrative and expansile invasion and high-grade serous ovarian carcinoma: a retrospective analysis. *Diagn Pathol*. (2022) 17:12. doi: 10.1186/s13000-022-01195-7

23. Tabrizi AD, Kalloger SE, Köbel M, Cipollon J, Roskelley CD, Cipollon J, et al. Primary ovarian mucinous carcinoma of intestinal type: Significance of pattern of invasion and immunohistochemical expression profile in a series of 31 cases. *Int J Gynecol Pathol*. (2010) 29:99–107. doi: 10.1097/PGP.0b013e3181bbbcc1

24. Sotiropoulou M, Markoulis P, Thomakos N, Rodolakis A, Koutroumpa I, Zacharakis D, et al. Mucinous carcinoma of the ovary: Significance of prognostic factors in clinical outcome. *Virchows Archiv*. (2013) 463:303. doi: 10.1007/s00428-013-1444-y

25. Algera MD, Van de Vijver KK, van Driel WJ, Slangen BFM, Lof FC, Vander Aa M, et al. Outcomes of patients with early stage mucinous ovarian carcinoma: a Dutch population-based cohort study comparing expansile and infiltrative subtypes. *Int J Gynecol Cancer*. (2024) 34:722–29. doi: 10.1136/ijgc-2023-004955

26. Meagher N, Koebel M, Anderson L, Tan A, Bolithon A, Anglesio M, et al. Pattern of invasion in stage I mucinous ovarian cancer is prognostic within 2-years of diagnosis. *Asia-Pacific J Clin Oncol*. (2021) 17:47–8. doi: 10.1111/ajco.13652

27. Huin M, Lorenzini J, Arbion F, Carcopino X, Touboul C, Dabi Y, et al. Presentation and prognosis of primary expansile and infiltrative mucinous carcinomas of the ovary. *J Clin Med*. (2022) 11(20). doi: 10.3390/jcm11206120

28. Hada T, Miyamoto M, Ishibashi H, Leunen K, Neven P, Vergote I, et al. Survival and biomarker analysis for ovarian mucinous carcinoma according to invasive patterns: retrospective analysis and review literature. *J Ovarian Res*. (2021) 14(7):1600–08. doi: 10.1186/s13048-021-00783-3

29. Nistor S, El-Tawab S, Wong F, Matsuura H, Sakamoto T, Kakimoto S, et al. The clinicopathological characteristics and survival outcomes of primary expansile vs. infiltrative mucinous ovarian adenocarcinoma: a retrospective study sharing the experience of a tertiary centre. *Trans Cancer Res*. (2023) 12:2682–92. doi: 10.21037/tcr-23-863

30. Köbel M, Kang EY, Lee S, Zouridis A, Roux R, Manek S, et al. Infiltrative pattern of invasion is independently associated with shorter survival and desmoplastic stroma markers FAP and THBS2 in mucinous ovarian carcinoma. *Histopathology*. (2024) 84:1095–110. doi: 10.1111/his.15128

31. Hada T, Miyamoto M, Ishibashi H, Matsuura H, Sakamoto T, Kakimoto S, et al. Prognostic similarity between ovarian mucinous carcinoma with expansile invasion and ovarian mucinous borderline tumor: a retrospective analysis. *Med (Baltim)*. (2021) 100:e26895. doi: 10.1097/MD.0000000000026895

32. Gouy S, Saidani M, Maulard A, Terzic T, Karnezis AN, Ghatage P, et al. Staging surgery in early-stage ovarian mucinous tumors according to expansile and infiltrative types. *Gynecol Oncol Rep*. (2017) 22:21–5. doi: 10.1016/j.gore.2017.08.006

33. Dundr P, Bártů M, Bosse T, Kruitwagen RFPM, Lok CAR. Primary mucinous tumors of the ovary: an interobserver reproducibility and detailed molecular study reveals significant overlap between diagnostic categories. *Mod Pathol*. (2023) 36:100040. doi: 10.1016/j.modpat.2022.100040

34. Gouy S, Saidani M, Maulard A, Bach-Hamba S, Bentivegna E, Leary A, et al. Results of fertility-sparing surgery for expansile and infiltrative mucinous ovarian cancers. *Oncologist*. (2018) 23:324–7. doi: 10.1634/theoncologist.2017-0310

35. Yoshihara M, Kajiyama H, Tamauchi S, Iyoshi S, Yokoi A, Suzuki S, et al. Impact of uterus-preserving surgery on stage I primary mucinous epithelial ovarian carcinoma: A multi-institutional study with propensity scoreWeighted analysis. *Int J Gynaecol Obstet*. (2020) 150:177–83. doi: 10.1002/ijgo

36. Rodríguez IM, Prat J. Mucinous tumors of the ovary: A clinicopathologic analysis of 75 borderline tumors(of intestinal type) and carcinomas. *Am J Surg Pathol*. (2002) 26:139–52. doi: 10.1097/00000478-200202000-00001

37. Lin W, Cao D, Shi X, You Y, Yang J, Shen K. Oncological and reproductive outcomes after fertility-sparing surgery for stage I mucinous ovarian carcinoma. *Front Oncol*. (2022) 12:856818. doi: 10.3389/fonc.2022.856818

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 04 February 2025
DOI 10.3389/fonc.2025.1540734

OPEN ACCESS

EDITED BY
Sandeep Kumar Mishra,
Yale University, United States

REVIEWED BY
Junxiang Huang,
Boston College, United States
Tengfei Ke,
Yunnan Cancer Hospital, China

*CORRESPONDENCE
Qingxiu Ai
51331947@qq.com
Bingqing Deng
dbq208@163.com

†These authors have contributed equally to this work

RECEIVED 06 December 2024
ACCEPTED 10 January 2025
PUBLISHED 04 February 2025

CITATION
Li S, Ding Q, Li L, Liu Y, Zou H, Wang Y, Wang X, Deng B and Ai Q (2025) Ultrasonic radiomics-based nomogram for preoperative prediction of residual tumor in advanced epithelial ovarian cancer: a multicenter retrospective study.
Front. Oncol. 15:1540734.
doi: 10.3389/fonc.2025.1540734

Ultrasonic radiomics-based nomogram for preoperative prediction of residual tumor in advanced epithelial ovarian cancer: a multicenter retrospective study

Shanshan Li[1†], Qiuping Ding[2†], Lijuan Li[3†], Yuwei Liu[1], Hanyu Zou[1], Yushuang Wang[1], Xiangyu Wang[4], Bingqing Deng[1*] and Qingxiu Ai[1*]

[1]Department of Medical Ultrasound, The Central Hospital of Enshi Prefecture Tujia and Miao Autonomous Prefecture, Hubei Selenium and Human Health Institute, Enshi, Hubei, China, [2]Reproductive Medicine Center, The Central Hospital of Enshi Prefecture Tujia and Miao Autonomous Prefecture, Enshi, Hubei, China, [3]Department of Medical Ultrasound, The Ethnic Hospital of Enshi Tujia and Miao Autonomous Prefecture, Enshi, Hubei, China, [4]Department of Medical Ultrasound, The Maternal and Child Health and Family Planning Service Center of Enshi Tujia and Miao Autonomous Prefecture, En Shi, Hubei, China

Objectives: To identify radiomic features extracted from ultrasound images and to develop and externally validate a comprehensive model that combines clinical data with ultrasound radiomics features to predict the residual tumor status in patients with advanced epithelial ovarian cancer (OC).

Methods: The study included 112 patients with advanced epithelial OC who underwent preoperative transvaginal ultrasound. Of these, 78 patients were assigned to the development cohort and 34 to the external validation cohort. Tumor contours were manually delineated as regions of interest (ROI) on the ultrasound images, and radiomic features were extracted. The final set of variables was identified using LASSO (least absolute shrinkage and selection operator) regression. Clinical features were also collected and incorporated into the model. A combination model integrating ultrasound radiomics and clinical variables was developed and externally validated. The performance of the predictive models was assessed.

Results: A total of 1,561 radiomic features and 18 clinical features were extracted. The final model included 10 significant ultrasound radiomic variables and 4 clinical features. The comprehensive model outperformed models based on either clinical or radiomic features alone, achieving an accuracy of 0.84, a sensitivity of 0.80, a specificity of 0.75, a precision of 0.88, a positive predictive value of 0.81, a negative predictive value of 0.86, an F1-score of 0.78, and an AUC of 0.82 in the external validation set.

Conclusions: The comprehensive model, which integrated clinical and ultrasound radiomic features, exhibited strong performance and generalizability in preoperatively identifying patients likely to achieve complete resection of all visible disease.

KEYWORDS

ultrasonic radiomics, ovarian cancer, predictive model, nomograms, residual tumor

1 Introduction

Ovarian cancer (OC) ranks among the most prevalent gynecological cancers, holding the position of the third most commonly diagnosed malignancy in the female reproductive system, surpassed only by cervical and endometrial cancers. Moreover, it exhibits the highest mortality rate within this category of cancers, posing a significant threat to women's health (1). Because early symptoms are often nonspecific, the majority of patients are diagnosed at an advanced clinical stage, frequently presenting with localized or widespread pelvic and abdominal metastases. Despite initial treatment, recurrence rates and mortality remain high, with frequent development of drug resistance. As a result, the 5-year survival rate is below 40%, leading to a generally poor prognosis for these patients (2).

According to the International Federation of Obstetrics and Gynecology (FIGO), there are two main treatment strategies for advanced OC in stages IIIC-IV: (1) primary debulking surgery (PDS) followed by six cycles of postoperative platinum-based chemotherapy, and (2) for patients unlikely to achieve satisfactory tumor reduction, two to three cycles of neoadjuvant chemotherapy can be given before interval debulking surgery (IDS), followed by postoperative adjuvant chemotherapy, a strategy commonly referred to as "sandwich" therapy (3). The primary goal of both treatment approaches is to maximize tumor reduction, ideally leaving a residual tumor (RT) diameter of less than 1 cm, or achieving no visible residual tumor (R0). Maximal cytoreduction stands as a critical prognostic factor in the treatment of advanced OC, showing the most favorable outcomes following adjuvant chemotherapy.

Unfortunately, not all OC patients are suitable candidates for primary debulking surgery (PDS) aimed at achieving an R0 resection (4). For those with a low probability of attaining R0 resection, there is a consensus that surgical intervention should be avoided if incomplete resection (with residual tumor greater than 1 cm) is anticipated, as it has little benefit to patient survival and may lead to a high incidence of perioperative related diseases (3–5). Therefore, assessing the probability of a patient's RT-resection during PDS prior to surgery is advantageous, as it supports the implementation of individualized treatment strategies.

In recent years, the field of imaging has made significant advancements, allowing for a more detailed depiction of tumor heterogeneity and providing valuable prognostic information (6). Various mathematical approaches have been applied to extract a vast array of radiomic features from medical images with high throughput, enabling clinicians to improve diagnostic accuracy and develop personalized, precision treatments (7, 8). Transvaginal ultrasound is a commonly utilized, cost-effective method for the clinical diagnosis of OC, and ultrasound radiomics has been increasingly employed in the study of various malignancies, including thyroid, cervical, liver, and OC (9–11). For example, Chiappa et al. utilized ultrasound radiomics to distinguish between malignant and benign ovarian tumors, highlighting its potential to enhance the preoperative evaluation of patients with ovarian masses and accurately identify those with OC (12). Thus, a comprehensive and unbiased assessment of ultrasound image features is essential (10).

This study seeks to assess the predictive significance of ultrasound radiomics and clinical factors in creating and validating a more reliable and generalizable preoperative model for forecasting RT status in patients with advanced epithelial OC. The goal is to standardize and simplify the process for gynecologists, enabling them to extract critical information from traditional diagnostic imaging more effectively and make informed decisions based on it.

2 Materials and methods

2.1 Study population

The study enrolled 112 patients with histologically confirmed FIGO stage III or IV OC diagnosed between January 2018 and June 2024. Of these, 78 patients from the Central Hospital of Enshi Tujia and Miao Autonomous Prefecture formed the development cohort, while 34 patients, recruited by collaborators at the Ethnic Hospital of Enshi Tujia and Miao Autonomous Prefecture, comprised the external validation cohort. The inclusion and exclusion criteria were consistent across both cohorts. The exclusion criteria included patients currently undergoing neoadjuvant chemotherapy, those lacking essential clinical or surgical data, individuals with poor image quality or significant image artifacts affecting visualization, and patients with a history of repeated biopsies. We established a

standardized protocol to define dataset variables and outcomes, enabling the retrospective collection of data within the same time frame. Patients who met the inclusion criteria were divided into two groups: (1) the RT<1 group, comprising individuals with no visible gross residual tumor (RT) and a maximum tumor diameter of less than 1 cm; and (2) the RT≥1 group, which included patients with a maximum tumor diameter of 1 cm or greater (13). This retrospective study was approved by our institution's ethics review board, with informed consent obtained from all participants.

2.2 Clinical information

Clinical data, including age, body mass index (BMI), parity, presence of hydrothorax, ascites, and ASA score, as well as the metastases in abdomen and pelvis (MAP) score, were collected. Laboratory findings such as perioperative platelet count, perioperative albumin levels, serum cancer antigen-125 (CA125), serum human epididymis protein 4 (HE-4) levels, and the neutrophil-to-lymphocyte ratio (NLR) were also obtained. Additionally, ultrasonic measurement characteristics such as maximum tumor diameter, arterial pulsatility index, resistance index, end diastolic flow rate, peak flow rate, and average flow rate were retrieved from the medical records.

The MAP score was assessed based on preoperative enhanced CT scans of the abdomen and pelvis, with two radiologists, blinded to intraoperative records, scoring and documenting the findings. The score was based on the Zhongshan Hospital rating scale for preoperative OC, which assessed lesions in various regions, including the diaphragmatic peritoneum, liver and kidney recesses, liver capsule, hepato-gastric space, spleen and stomach space, greater omentum (covering both the liver area and splenic curvature), mesentery, peritoneum, intestines, paracolic sulci, uterorectal space, uterine bladder space, and lymph nodes. Each identified lesion contributed 2 points, with the total score being the cumulative sum of all lesions. Any discrepancies in scoring were resolved through consensus.

2.3 Image segmentation

In accordance with the Institutional Review Board's approved protocol, essential clinical data and ultrasound image locations were systematically documented in standardized electronic case report forms (CRFs) and collected within four weeks prior to the primary surgical intervention. The segmentation of images was conducted independently by two seasoned radiologists who were unaware of the patients' tissue pathology. One of the radiologists, possessing around 12 years of experience, utilized the open-source ITK-SNAP software (version 3.8.0; www.itksnap.org) to manually delineate the regions of interest (ROIs) on the image slices. The Kappa consistency analysis was performed to evaluate discrepancies between two radiologists, and a Kappa value ≥ 0.85 was regarded as a good consistency.

2.4 Radiomics feature extraction

PyRadiomics (v.2.0.0; http://www.radiomics.io/pyradiomics.html) software was used to extract features from medical images (14). The process included importing manually delineated ROI images along with the original images into the PyRadiomics platform, where an internal feature analysis program was utilized to extract the relevant features. We adopted nonlinear intensity transformation on image voxels, Gaussian Laplace filter and Eight wavelet transform to obtain high-throughput features. Radiomic features can be categorized into three main groups: (I) geometry, (II) intensity, and (III) texture. Geometric features describe the three-dimensional shape of the tumor, while intensity features reflect the first-order statistical distribution of voxel intensities within the tumor. Texture features, on the other hand, analyze the patterns and the second- and higher-order spatial distributions of these intensities. A total of 1,561 radiomic features were extracted, encompassing first-order features, shape-based features, and a variety of matrix features, including gray level co-occurrence matrix (GLCM) features, gray level dependence matrix (GLDM) features, gray level run length matrix (GLRLM) features, gray level size zone matrix (GLSZM) features, and neighborhood gray-tone difference matrix (NGTDM) features.

2.5 Radiomics feature selection

To eliminate differences in index dimensions, Z-score normalization was applied to account for the varying scales of the manually derived radiomic features. Three methods were employed to select the final variables. Initially, the Mann-Whitney U test was performed to filter all radiomic features, retaining only those with a p-value of less than 0.05. Subsequently, Pearson's rank correlation coefficient was computed to evaluate the correlation between features, and those with an intraclass correlation coefficient (ICC) below 0.9 were discarded to guarantee high repeatability. Finally, the least absolute shrinkage and selection operator (LASSO) regression model was employed to identify the final variables for model construction. Ultimately, the best features were incorporated into the prediction models, which were developed using 10-fold cross-validation.

2.6 Model development and validation

Three models were developed using the development set of 78 patients: model I (the clinical model), model II (the radiomics model), and model III (the clinical-radiomics model). For radiomics models, we tested 15 machine learning algorithms, with the LightGBM model demonstrating the best performance (Appendix 1). However, the clinical-radiomics model was chosen as the nomogram to enhance convenience for clinical application.

The external validation set (34 patients) used to evaluate model performance. The model's performance was assessed through several metrics, including accuracy, sensitivity, specificity,

precision, positive predictive value, negative predictive value, and F1-Score. Additionally, the receiver operating characteristic (ROC) curve was calculated along with the area under the ROC curve (AUC). Calibration was assessed through calibration plots, which depicted the relationship between predicted probabilities and observed proportions. To evaluate the clinical utility and benefits of the predictive model, decision curve analysis (DCA) was conducted.

2.7 Statistical analysis

All statistical analyses were performed using Python packages (version 0.13.2). Group differences were evaluated using either Student's t-test or Mann–Whitney U test for continuous variables, while categorical variables were analyzed using the chi-square test or Fisher's exact test. Multivariate analysis was conducted to select the final variables. Continuous variables that followed a normal distribution are presented as means ± standard deviations (SDs), whereas non-normally distributed variables are reported as medians ± interquartile ranges (IQRs). And odds ratios (ORs), 95% confidence intervals (CIs), HosmerLemeshow (H-L) test were also calculated. And a p value < 0.05 was considered statistically significant.

3 Results

3.1 Clinical and demographic characteristics

The final cohort comprised 112 patients with advanced epithelial OC. This included the development cohort (n=78), which consisted of 55 patients with R0 resection and 23 patients with non-R0 status, and the external validation cohort (n=34), which included 24 patients with R0 resection and 10 patients with non-R0 status. The comparison between the development and external validation cohorts revealed no significant differences between the two groups, nor within each group (p > 0.05), indicating a reasonable classification. Table 1 present the baseline characteristics of patients in each cohort. In the multivariate analysis, age (p = 0.031; OR = 1.011, 95% CI: 1.003-1.018), CA125 level (p = 0.002; OR = 1.001, 95% CI: 1.000-1.001), presence of hydrothorax (p = 0.003; OR = 1.174, 95% CI: 1.078-1.279), and maximum tumor diameter (p = 0.031; OR = 1.002, 95% CI: 1.001-1.004) were identified as independent predictors of RT status (Table 2).

3.2 Radiomics characteristics

A total of 1,561 radiomic features were extracted from ultrasound images, which included 306 first-order features, 14 shape-based features, 374 features from the GLCM, 238 features from the GLDM, 272 features from the GLRLM, 272 features from the GLSZM, and 85 features from the NGTDM. The t-test or Mann-Whitney U test was utilized for the preliminary screening of all features, resulting in the inclusion of 42 features. Subsequently, Pearson correlation analysis was conducted, revealing 25 features that were significantly different between the two groups. Next, LASSO regression was conducted using 10-fold cross-validation with the minimum criterion to determine the optimal λ values. The λ value that resulted in the lowest cross-validation errors is illustrated in Figures 1 and 2. Following this, ten features with nonzero coefficients were used for this task. Finally, ultrasonic radiomic features were established using these 10 features, namely exponential_firstorder_Skewness, exponential_glszm_LargeAreaHighGrayLevelEmphasis, gradient_firstorder_Minimum, lbp_3D_m2_firstorder_90Percentile, logarithm_firstorder_Minimum, squareroot_glcm_Idn, squareroot_glszm_GrayLevelNonUniformityNormalized, squareroot_glszm_SmallAreaEmphasis, wavelet_LHL_ngtdm_Contrast, wavelet_LLL_glcm_Idn (Figures 1, 2).

3.3 Model construction and performance assessment

We developed three models to identify patients suitable for optimal primary debulking surgery. Model 1 (the clinical model) was based solely on clinical characteristics using the LightGBM algorithm. Model 2 (the radiomics model) relied exclusively on ultrasonic radiomics characteristics, also employing the LightGBM algorithm (Appendix 1). Model 3 (the clinical-radiomics model) was an integrative nomogram that combined clinical and radiomics features to enhance clinical application convenience (Figure 3).

The radiomic-clinical nomogram demonstrated superior performance compared to the clinical or radiomics models alone, achieving an accuracy of 0.84, a sensitivity of 0.80, a specificity of 0.75, a precision of 0.88, a positive predictive value of 0.81, a negative predictive value of 0.86, an F1-Score of 0.78, and an AUC of 0.82 in the external validation set (Table 3). Figure 4 illustrates the AUC for both the development and external validation cohorts. The calibration curves for the radiomic-clinical nomogram demonstrated strong agreement between predicted and observed outcomes in both the development and validation cohorts (Figure 4). The Hosmer-Lemeshow (HL) test indicated favorable goodness-of-fit for the data (all p > 0.05). Furthermore, the DCA revealed that the nomogram offers greater clinical benefit (Figure 4), namely, the DCA for the three models indicates that this new diagnostic approach yields a greater net benefit (where a value greater than 0 indicates patient benefit) in predicting the residual tumor status in patients with advanced OC, with the clinical-radiomics model showing a more significant benefit compared to the clinical model or radiomics model.

TABLE 1 Clinical and demographic characteristics of development and validation cohort.

Variables	Development cohort (N=78)			External validation cohort (N=34)		
	R0 (N=55)	Non-R0 (N=23)	*P*	R0 (N=24)	Non-R0 (N=10)	*P*
Age	54.55 ± 9.23	62.13 ± 7.14	<0.01	54.88 ± 9.34	62.10 ± 5.51	0.03
BMI	22.23 ± 3.05	22.64 ± 3.51	0.72	22.08 ± 3.73	24.45 ± 1.89	0.08
NLR	3.07 ± 1.86	3.29 ± 1.82	0.70	3.09 ± 1.88	2.47 ± 1.67	0.46
Perioperative platelet	226.82 ± 82.89	211.04 ± 85.68	0.41	230.00 ± 70.89	175.80 ± 78.42	0.06
Perioperative albumin	45.69 ± 5.48	43.75 ± 4.77	0.09	45.80 ± 5.62	45.37 ± 5.87	0.81
CA125	278.36 ± 163.28	465.04 ± 179.81	<0.01	284.46 ± 136.25	403.10 ± 167.44	0.04
HE-4	285.55 ± 135.99	546.83 ± 183.08	<0.01	318.08 ± 143.04	574.10 ± 184.04	<0.01
MAP score	7.93 ± 2.77	17.83 ± 4.39	<0.01	7.33 ± 2.18	20.80 ± 3.68	<0.01
Maximum tumor diameter	117.25 ± 38.89	141.62 ± 37.31	0.01	119.19 ± 38.81	141.91 ± 23.24	0.10
Arterial pulsatility index	0.31 ± 0.14	0.31 ± 0.12	0.96	0.32 ± 0.14	0.35 ± 0.21	0.63
Resistance index	0.26 ± 0.09	0.28 ± 0.09	0.38	0.27 ± 0.10	0.28 ± 0.13	0.79
End diastolic flow rate	17.09 ± 2.50	16.96 ± 2.09	0.83	16.67 ± 1.88	16.63 ± 2.77	0.97
Peak flow rate	23.07 ± 2.30	23.18 ± 2.02	0.89	22.91 ± 1.97	23.24 ± 2.14	0.67
Average flow rate	19.51 ± 2.22	19.94 ± 1.60	0.40	19.72 ± 1.89	19.48 ± 1.78	0.74
Parity			0.24			0.92
1	4 (7.27)	0		3 (12.50)	1 (10.00)	
2	38 (69.09)	21 (91.30)		17 (70.83)	7 (70.00)	
3	7 (12.73)	2 (8.70)		3 (12.50)	1 (10.00)	
4	5 (9.09)	0		1 (4.17)	1 (10.00)	
5	1 (1.82)	0		0	0	
ASA score			0.14			0.32
1	8 (14.55)	5 (21.74)		8 (33.33)	1 (10.00)	
2	16 (29.09)	1 (4.35)		3 (12.50)	2 (20.00)	
3	10 (18.18)	8 (34.78)		3 (12.50)	2 (20.00)	
4	11 (20.00)	5 (21.74)		8 (33.33)	2 (20.00)	
5	10 (18.18)	4 (17.39)		2 (8.33)	3 (30.00)	
Ascites			0.59			0.13
0	22 (40.00)	3 (13.04)		9 (37.50)	4 (40.00)	
1	19 (34.55)	4 (17.39)		8 (33.33)	3 (30.00)	
2	14 (25.45)	16 (69.57)		7 (29.17)	3 (30.00)	
Hydrothorax			0.01			0.98
0	23 (41.82)	7 (30.43)		4 (16.67)	5 (50.00)	
1	16 (29.09)	9 (39.13)		9 (37.50)	2 (20.00)	
2	16 (29.09)	7 (30.43)		11 (45.83)	3 (30.00)	

A p value < 0.05 was considered statistically significant.
ORs, Odds ratios; CIs, Confidence intervals; BMI, Body mass index; NLR, Neutrophil-to-lymphocyte ratio; CA125, Cancer antigen-125; HE-4, Human epididymis protein 4; MAP score, metastases in abdomen and pelvis score; ASA score, American Society of Anesthesiology score.

TABLE 2 the univariate and multivariate logistic regression analysis of development cohort.

Variables	Univariate logistic regression analysis			Multivariate logistic regression analysis		
	OR	OR 95% CI	*P*	OR	OR 95% CI	*P*
Age	1.02	1.01-1.03	0.001	1.01	1.00-1.02	0.030
BMI	1.00	0.98-1.04	0.609			
NLR	1.01	0.97-1.06	0.639			
Perioperative platelet	1.00	1.00-1.00	0.450			
Perioperative albumin	0.99	0.97-1.00	0.145			
CA125	1.00	1.00-1.00	0.000	1.00	1.00-1.00	0.002
HE-4	0.89	0.85-0.93	0.365			
MAP score	0.95	0.94-0.96	0.210			
Maximum tumor diameter	1.00	1.00-1.00	0.013	1.00	1.00-1.00	0.031
Arterial pulsatility index	0.98	0.51-1.90	0.959			
Resistance index	1.65	0.65-4.24	0.376			
End diastolic flow rate	0.96	0.96-1.03	0.830			
Peak flow rate	1.00	0.97-1.05	0.837			
Average flow rate	1.02	0.98-1.07	0.397			
Parity	0.92	0.81-1.04	0.244			
ASA score	1.01	0.95-1.08	0.757			
Ascites	1.24	1.12-1.36	0.000	1.17	1.08-1.28	0.003
Hydrothorax	1.65	0.65-4.24	0.376			

A p value < 0.05 was considered statistically significant.
ORs, Odds ratios; CIs, Confidence intervals; BMI, Body mass index; NLR, Neutrophil-to-lymphocyte ratio; CA125, Cancer antigen-125; HE-4, Human epididymis protein 4; MAP score, metastases in abdomen and pelvis score; ASA score, American Society of Anesthesiology score.

4 Discussion

In our study, we integrated primary radiomic features, laboratory findings, and clinical factors from patients with advanced epithelial OC to create and validate a radiomics-clinical nomogram. This nomogram is designed for individualized preoperative prediction of treatment response (RT) status. The results demonstrated that the integrated radiomic-clinical nomogram showed enhanced predictive performance compared to using radiomic or clinical signatures individually after external validation. The final model is capable of identification of the RT status prior to surgery. This advancement enhances clinical decision-making, patient communication, and prognosis assessment. For those with a low probability of attaining R0 resection, the surgical intervention should be avoided if incomplete resection. The presence or absence of response to treatment (RT) following PDS or IDS is the most significant factor influencing the prognosis of patients with advanced OC. Notably, a 10% increase in the rate of complete tumor resection can lead to a 5% improvement in overall survival for these patients (15). Research has shown that RT status is an independent and significant prognostic factor for patients with advanced OC. The extent of RT is inversely correlated with patient survival, disease-free survival (DFS), and overall survival (OS) (5, 16). According to Kehoe et al., patients with OC who underwent PDS followed by RT excision experienced the most favorable prognosis (17). High-grade serous ovarian cancer (HGSOC) is the most common and aggressive histological subtype of OC, and complete resection of all visible lesions (RT-resection) in advanced HGSOC patients after PDS is linked to the best outcomes (5, 18). Therefore, it is essential to assess all epithelial OC patients suspected of being at stage IIIC or IV to determine their eligibility for PDS prior to initiating therapy, in line with the clinical practice guidelines set forth by the Society of Gynecologic Oncology and the American Society of Clinical Oncology (19).

For patients in whom achieving satisfactory tumor reduction is challenging, neoadjuvant chemotherapy should be considered prior to PDS. Kevin et al. (21) demonstrated that the mean tumor nuclear area and the major axis length of the stroma are significant factors that can improve risk stratification in patients with HGSOC. For the ultrasonic radiomic characteristics, three methods were employed to select the final variables, resulting in the inclusion of 10 features from a total of 1,561 radiomic features in our model, effectively eliminating invalid variables. Previous studies have demonstrated that all ultrasonic radiomics and clinical features included in our study are relevant to the diagnosis, treatment, and prognosis of ovarian cancer (15, 18, 21, 24).

CA-125 is one of the most commonly used serum biomarkers for OC. Some studies (13, 20) have found that preoperative CA-125 levels can predict gross residual disease at PDS for advanced

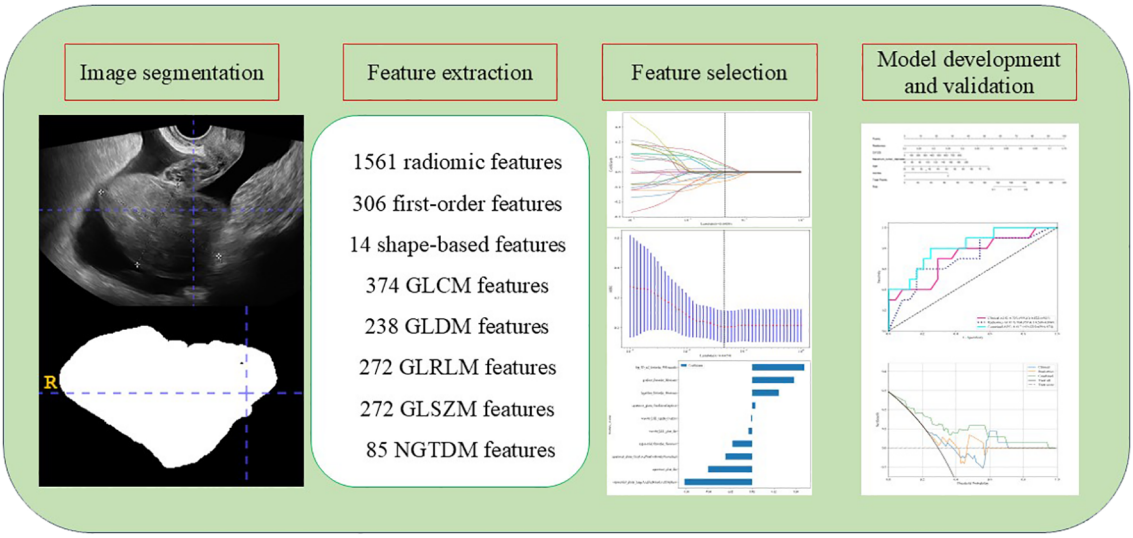


FIGURE 1
Study flowchart of the radiomics analysis.

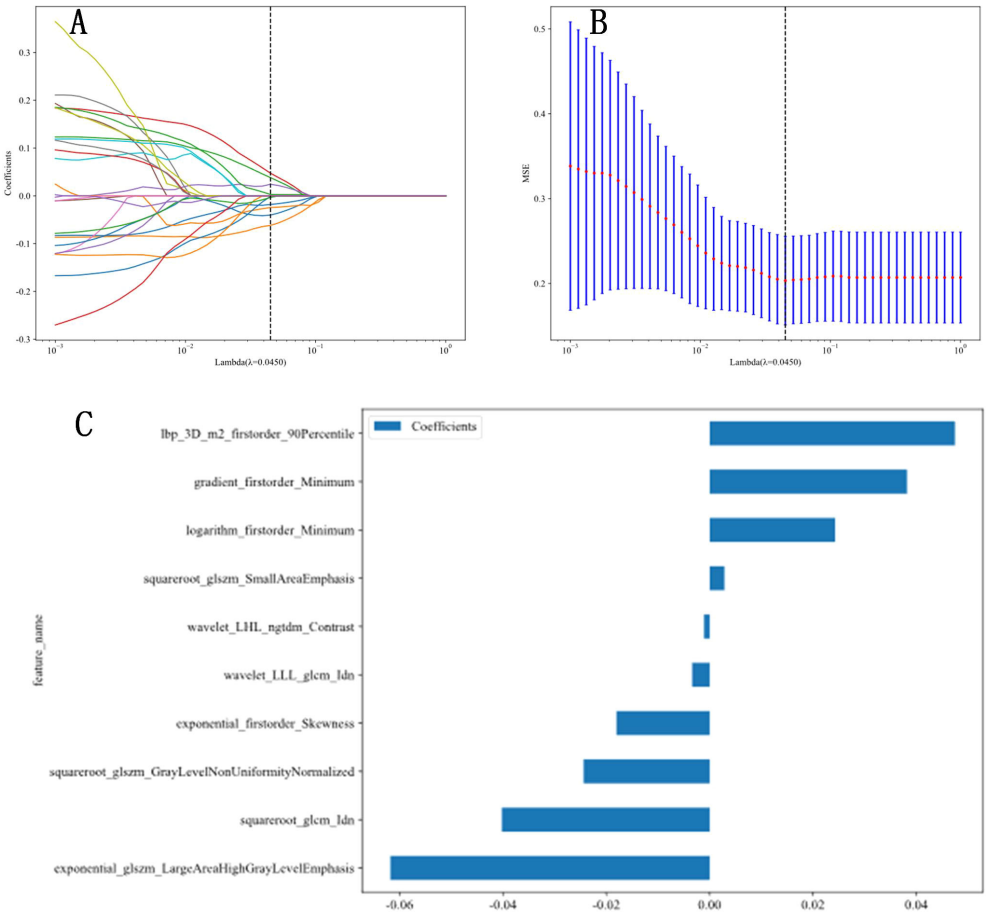


FIGURE 2
Radiomic feature extraction. **(A, B)** Radiomic features extraction using least absolute shrinkage and selection operator (LASSO) algorithm. **(C)** The final features included in our study.

10.3389/fonc.2025.1540734

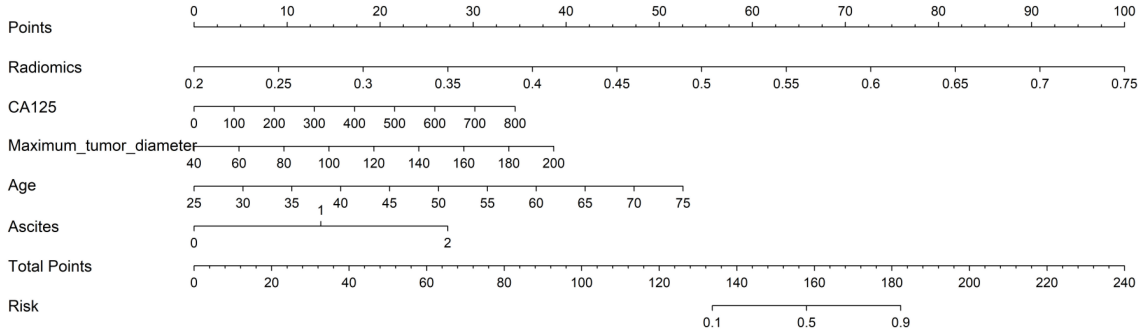


FIGURE 3
A nomogram integrates clinical parameters and radiomics features.

TABLE 3 The performance of clinical model, radiomics model and combined nomogram for predicting RT status.

Model	Cohort	AUC	ACC	Sen	Spe	PPV	NPV	Precision	F1
Clinical	Development	0.883	0.883	0.883	0.883	0.883	0.883	0.883	0.883
	Validation	0.723	0.723	0.723	0.723	0.723	0.723	0.723	0.723
Radiomics	Development	0.878	0.878	0.878	0.878	0.878	0.878	0.878	0.878
	Validation	0.704	0.704	0.704	0.704	0.704	0.704	0.704	0.704
Combined	Development	0.900	0.900	0.900	0.900	0.900	0.900	0.900	0.900
	Validation	0.817	0.817	0.817	0.817	0.817	0.817	0.817	0.817

AUC, Area under the curve; ACC, Accuracy; Sen, Sensitivity; Spe, Specificity; PPV, Positive predictive value; NPV, Negative predictive value; F1, F1-Score.

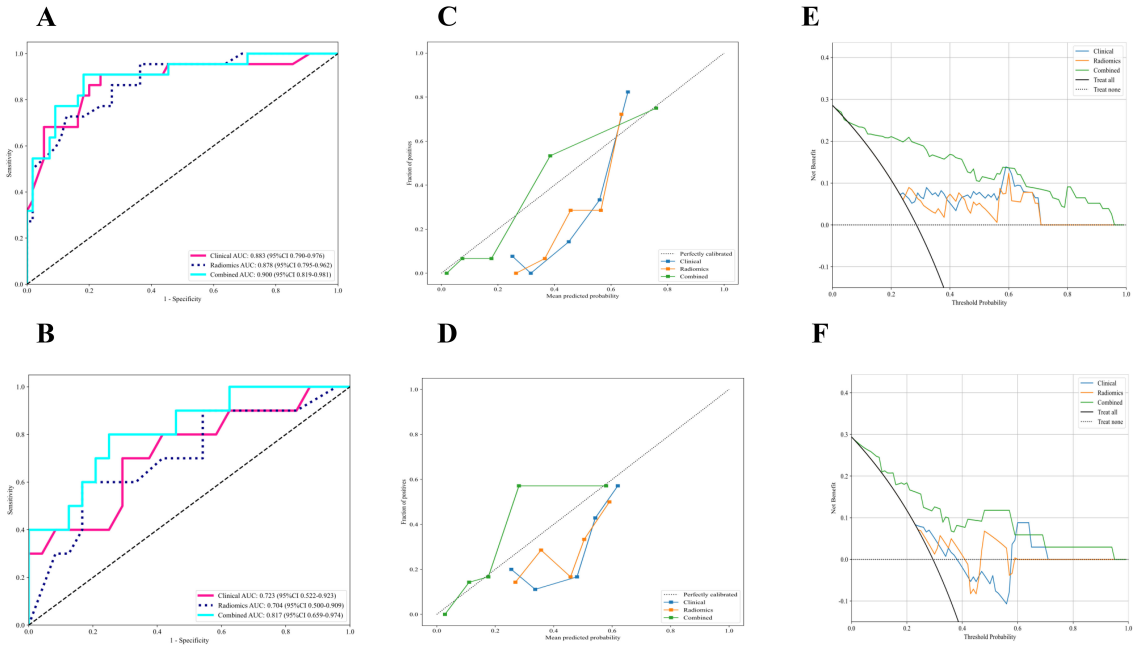


FIGURE 4
The performance of clinical model, radiomics model and combined nomogram with ROC, calibration curves and decision curve analysis. **(A, B)** ROC curves of each model in the **(A)** development and **(B)** external validation cohort for prediction of RT status. **(C, D)** Calibration curves of each model in the **(C)** development and **(D)** external validation cohort for prediction of RT status, and A 45° diagonal line indicates perfect calibration. **(E, F)** Decision curve analysis of each model in the **(E)** development and **(F)** external validation cohort for prediction of RT status, and the colored lines were expected net benefit of per patient. ROC, Receiver operating characteristics curves; RT, Residual tumor.

epithelial OC. Additionally, moderate to severe ascites has been associated with residual disease (13) and may serve as a surrogate indicator of advanced disease across multiple anatomic locations. The maximum tumor diameter is a critical predictor for individualized preoperative assessment of RT status in patients with advanced OC, as reflected in radiomic shape-based features. For patients who are unlikely to achieve satisfactory tumor reduction, neoadjuvant chemotherapy should be considered prior to PDS. Kevin et al. (21) demonstrated that the mean tumor nuclear area and the major axis length of the stroma are important factors for improving risk stratification in patients with HGSOC. In analyzing ultrasonic radiomic characteristics, three methods were utilized to select the final variables, resulting in the inclusion of 10 features from a total of 1,561 radiomic features in our model, effectively eliminating invalid variables.

Ultrasound offers several advantages, including real-time display, convenience, and affordability, making it widely used for screening and preoperative evaluation of OC. Recently, applications of ultrasound-based radiomics have been reported in tumor diagnosis (12), pathology grading (22), vascular invasion assessment, therapeutic evaluation (23), and prognostic prediction (24). However, there are few reports on RT status based on ultrasonics. Meanwhile, several radiomic models for predicting RT status based on computed tomography (CT) and magnetic resonance imaging (MRI) have been developed and validated (25, 26). Lu et al. (26) developed an MRI-based radiomic-clinical nomogram that successfully predicted RT status preoperatively in patients with HGSOC. A multicenter assessment was conducted to evaluate the efficacy of preoperative CT scans and CA-125 levels in predicting gross residual disease following PDS for advanced epithelial OC (25). However, the pelvic CT-based model was primarily developed with a focus on abdominal metastases. These findings support the hypothesis that radiomic features can effectively predict treatment response (RT) status by capturing variations in tumor heterogeneity.

There are several limitations to our study. Firstly, it relies on a small sample size, necessitating larger databases and multicenter studies to confirm the generalizability of this model. Second, future studies should integrate CT or contrast-enhanced CT and MRI or contrast-enhanced MRI into the predictive model to enhance the prediction of RT status in OC. Finally, our study focused exclusively on advanced epithelial OC subtypes, excluding rare variants. Future research should include data from additional OC subtypes to improve the models' universality and clinical applicability.

5 Conclusion

In our study, we confirmed the clinical value of ultrasound-based radiomics for the preoperative prediction of treatment response (RT) status in patients with advanced epithelial OC, and radiomic feature extraction and selection may provide a deeper understanding of ultrasound imaging mechanism. The comprehensive model combined clinical and ultrasonic radiomics features not only had a better performance in preoperative identification of complete resection of all visible diseases but also had a higher generalization ability.

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 authors.

Ethics statement

The studies involving humans were approved by The Central Hospital of Enshi Prefecture Tujia and Miao Autonomous Prefecture. 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.

Author contributions

SL: Data curation, Formal Analysis, Writing – original draft. QD: Data curation, Investigation, Methodology, Writing – review & editing. LJ: Methodology, Resources, Software, Writing – review & editing. YL: Resources, Writing – review & editing. HZ: Data curation, Visualization, Writing – review & editing. YW: Data curation, Visualization, Writing – review & editing. XW: Methodology, Writing – review & editing. BD: Writing – review & editing. QA: Project administration, Writing – review & editing.

Funding

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

Acknowledgments

The authors extend our deepest appreciation to the patients and their families receiving care in the Central Hospital of Enshi Prefecture. The authors also thank team staff who assisted in the data collection and analysis.

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/fonc.2025.1540734/full#supplementary-material

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer J Clin*. (2021) 71:209–49. doi: 10.3322/caac.21660

2. Webb PM, Jordan SJ. Epidemiology of epithelial ovarian cancer. *Best Pract Res Clin Obstetrics Gynaecology*. (2017) 41:3–14. doi: 10.1016/j.bpobgyn.2016.08.006

3. Chi DS, Eisenhauer EL, Lang J, Huh J, Haddad L, Abu-Rustum NR, et al. What is the optimal goal of primary cytoreductive surgery for bulky stage IIIC epithelial ovarian carcinoma (EOC)? *Gynecologic Oncol*. (2006) 103:559–64. doi: 10.1016/j.ygyno.2006.03.051

4. Ghirardi V, Moruzzi MC, Bizzarri N, Vargiu V, DIndinosante M, Garganese G, et al. Minimal residual disease at primary debulking surgery versus complete tumor resection at interval debulking surgery in advanced epithelial ovarian cancer: A survival analysis. *Gynecologic Oncol*. (2020) 157:209–13. doi: 10.1016/j.ygyno.2020.01.010

5. du Bois A, Reuss A, Pujade-Lauraine E, Harter P, Ray-Coquard I, Pfisterer J. Role of surgical outcome as prognostic factor in advanced epithelial ovarian cancer: a combined exploratory analysis of 3 prospectively randomized phase 3 multicenter trials: by the Arbeitsgemeinschaft Gynaekologische Onkologie Studiengruppe Ovarialkarzinom (AGO-OVAR) and the Groupe d'Investigateurs Nationaux Pour les Etudes des Cancers de l'Ovaire (GINECO). *Cancer*. (2009) 115:1234–44. doi: 10.1002/cncr.v115:6

6. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. (2014) 5:4006. doi: 10.1038/ncomms5006

7. Aerts HJ. The potential of radiomic-based phenotyping in precision medicine: A review. *JAMA Oncol*. (2016) 2:1636–42. doi: 10.1001/jamaoncol.2016.2631

8. Kirienko M, Cozzi L, Antunovic L, Lozza L, Fogliata A, Voulaz E, et al. Prediction of disease-free survival by the PET/CT radiomic signature in non-small cell lung cancer patients undergoing surgery. *Eur J Nucl Med Mol Imaging*. (2018) 45:207–17. doi: 10.1007/s00259-017-3837-7

9. Park VY, Han K, Lee E, Kim EK, Moon HJ, Yoon JH, et al. Association between radiomics signature and disease-free survival in conventional papillary thyroid carcinoma. *Sci Rep*. (2019) 9:4501. doi: 10.1038/s41598-018-37748-4

10. Jin X, Ai Y, Zhang J, Zhu H, Jin J, Teng Y, et al. Noninvasive prediction of lymph node status for patients with early-stage cervical cancer based on radiomics features from ultrasound images. *Eur Radiol*. (2020) 30:4117–24. doi: 10.1007/s00330-020-06692-1

11. Hu HT, Wang Z, Huang XW, Chen SL, Zheng X, Ruan SM, et al. Ultrasound-based radiomics score: a potential biomarker for the prediction of microvascular invasion in hepatocellular carcinoma. *Eur Radiol*. (2019) 29:2890–901. doi: 10.1007/s00330-018-5797-0

12. Chiappa V, Bogani G. The Adoption of Radiomics and machine learning improves the diagnostic processes of women with Ovarian MAsses (the AROMA pilot study). *J Ultrasound*. (2021) 24:429–37. doi: 10.1007/s40477-020-00503-5

13. Suidan RS, Ramirez PT, Sarasohn DM, Teitcher JB, Iyer RB, Zhou Q, et al. A multicenter assessment of the ability of preoperative computed tomography scan and CA-125 to predict gross residual disease at primary debulking for advanced epithelial ovarian cancer. *Gynecologic Oncol*. (2017) 145:27–31. doi: 10.1016/j.ygyno.2017.02.020

14. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. (2017) 77:e104–e7. doi: 10.1158/0008-5472.CAN-17-0339

15. Bristow RE, Tomacruz RS, Armstrong DK, Trimble EL, Montz FJ, et al. Survival effect of maximal cytoreductive surgery for advanced ovarian carcinoma during the platinum era: a meta-analysis. *J Clin Oncology: Off J Am Soc Clin Oncol*. (2002) 20:1248–59. doi: 10.1200/JCO.2002.20.5.1248

16. Fagotti A, Vizzielli G, Fanfani F, Costantini B, Ferrandina G, Gallotta V, et al. Introduction of staging laparoscopy in the management of advanced epithelial ovarian, tubal and peritoneal cancer: impact on prognosis in a single institution experience. *Gynecologic Oncol*. (2013) 131:341–6. doi: 10.1016/j.ygyno.2013.08.005

17. Kehoe S, Hook J, Nankivell M, Jayson GC, Kitchener H, Lopes T, et al. Primary chemotherapy versus primary surgery for newly diagnosed advanced ovarian cancer (CHORUS): an open-label, randomised, controlled, non-inferiority trial. *Lancet*. (2015) 386:249–57. doi: 10.1016/S0140-6736(14)62223-6

18. Chang SJ, Bristow RE, Ryu HS. Impact of complete cytoreduction leaving no gross residual disease associated with radical cytoreductive surgical procedures on survival in advanced ovarian cancer. *Ann Surg Oncol*. (2012) 19:4059–67. doi: 10.1245/s10434-012-2446-8

19. Wright AA, Bohlke K, Armstrong DK, Bookman MA, Cliby WA, Coleman RL, et al. Neoadjuvant chemotherapy for newly diagnosed, advanced ovarian cancer: Society of Gynecologic Oncology and American Society of Clinical Oncology Clinical Practice Guideline. *Gynecol Oncol*. (2016) 143:3–15. doi: 10.1016/j.ygyno.2016.05.022

20. Chi DS, Zivanovic O, Palayekar MJ, Eisenhauer EL, Abu-Rustum NR, Sonoda Y, et al. A contemporary analysis of the ability of preoperative serum CA-125 to predict primary cytoreductive outcome in patients with advanced ovarian, tubal and peritoneal carcinoma. *Gynecologic Oncol*. (2009) 112:6–10. doi: 10.1016/j.ygyno.2008.10.010

21. Boehm KM, Aherne EA, Ellenson L, Nikolovski I, Alghamdi M, Vázquez-García I, et al. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat Cancer*. (2022) 3:723–33. doi: 10.1038/s43018-022-00388-9

22. Ren S, Qi Q, Liu S, Duan S, Mao B, Chang Z, et al. Preoperative prediction of pathological grading of hepatocellular carcinoma using machine learning-based ultrasomics: A multicenter study. *Eur J Radiol*. (2021) 143:109891. doi: 10.1016/j.ejrad.2021.109891

23. Wang W, Wu SS, Zhang JC, Xian MF, Huang H, Li W, et al. Preoperative pathological grading of hepatocellular carcinoma using ultrasomics of contrast-enhanced ultrasound. *Acad Radiol*. (2021) 28:1094–101. doi: 10.1016/j.acra.2020.05.033

24. Yao F, Ding J, Hu Z, Cai M, Liu J, Huang X, et al. Ultrasound-based radiomics score: a potential biomarker for the prediction of progression-free survival in ovarian epithelial cancer. *Abdominal Radiol*. (2021) 46:4936–45. doi: 10.1007/s00261-021-03163-z

25. Gu Y, Qin M, Jin Y, Zuo J, Li N, Bian C, et al. A prediction model for optimal primary debulking surgery based on preoperative computed tomography scans and clinical factors in patients with advanced ovarian cancer: A multicenter retrospective cohort study. *Front Oncol*. (2021) 10. doi: 10.3389/fonc.2020.611617

26. Lu J, Cai S, Wang F, Wu PY, Pan X, Qiang J, et al. Development of a prediction model for gross residual in high-grade serous ovarian cancer by combining preoperative assessments of abdominal and pelvic metastases and multiparametric MRI. *Acad Radiol*. (2023) 30:1823–31. doi: 10.1016/j.acra.2022.12.019

frontiers | Frontiers in Oncology

TYPE Review
PUBLISHED 20 February 2025
DOI 10.3389/fonc.2025.1536039

OPEN ACCESS

EDITED BY
Sandeep Kumar Mishra,
Yale University, United States

REVIEWED BY
Tongning Wu,
China Academy of Information and Communications Technology, China
Madhavi Dabbiru,
Dr. Lankapalli Bullayya College of Engineering, India

*CORRESPONDENCE
Kun Cao
461024915@qq.com
Jianjun Yang
qlbsh1@163.com

RECEIVED 28 November 2024
ACCEPTED 31 January 2025
PUBLISHED 20 February 2025

CITATION
Yang L, Wang X, Zhang S, Cao K and Yang J (2025) Research progress on artificial intelligence technology-assisted diagnosis of thyroid diseases.
Front. Oncol. 15:1536039.
doi: 10.3389/fonc.2025.1536039

Research progress on artificial intelligence technology-assisted diagnosis of thyroid diseases

Lina Yang[1], XinYuan Wang[2], Shixia Zhang[1], Kun Cao[1*] and Jianjun Yang[3*]

[1]Development Department of the Wisdom Hospital, Shandong Provincial Third Hospital, Jinan, China, [2]Information Department, Shandong First Rehabilitation Hospital, Linyi, China, [3]General Practice Medicine, Shandong Provincial Third Hospital, Jinan, China

With the rapid development of the "Internet + Medical" model, artificial intelligence technology has been widely used in the analysis of medical images. Among them, the technology of using deep learning algorithms to identify features of ultrasound and pathological images and realize intelligent diagnosis of diseases has entered the clinical verification stage. This study is based on the application research of artificial intelligence technology in medical diagnosis and reviews the early screening and diagnosis of thyroid diseases. The cure rate of thyroid disease is high in the early stage, but once it deteriorates into thyroid cancer, the risk of death and treatment costs of the patient increase. At present, the early diagnosis of the disease still depends on the examination equipment and the clinical experience of doctors, and there is a certain misdiagnosis rate. Based on the above background, it is particularly important to explore a technology that can achieve objective screening of thyroid lesions in the early stages. This paper provides a comprehensive review of recent research on the early diagnosis of thyroid diseases using artificial intelligence technology. It integrates the findings of multiple studies and that traditional machine learning algorithms are widely used as research objects. The convolutional neural network model has a high recognition accuracy for thyroid nodules and thyroid pathological cell lesions. U-Net network model can significantly improve the recognition accuracy of thyroid nodule ultrasound images when used as a segmentation algorithm. This article focuses on reviewing the intelligent recognition technology of thyroid ultrasound images and pathological sections, hoping to provide researchers with research ideas and help clinicians achieve intelligent early screening of thyroid cancer.

KEYWORDS

thyroid disease, machine learning, image recognition, thyroid ultrasound, thyroid pathological slices

1 Introduction

The thyroid gland is a butterfly-shaped gland located in the front of the neck. Its main function is to secrete thyroid hormones. Thyroid hormones play a key role in regulating many physiological processes in the human body, including diabetes management, cardiovascular health, cognitive function, and immune system regulation. Therefore, maintaining normal thyroid hormone levels is essential to maintaining good health (1, 2). When thyroid hormone secretion is disordered, it can lead to abnormal thyroid function or abnormal thyroid structure. Thyroid dysfunction includes hyperthyroidism and hypothyroidism. Thyroid structural abnormalities mainly include thyroid nodules and thyroid cancer. Thyroid nodules refer to solid or cystic masses that appear inside the thyroid gland. Thyroid cancer is a malignant tumor that occurs in thyroid cells and is one of the most common malignant tumors in the endocrine system (3). The causes of thyroid cancer are complex. As a malignant tumor, tumor cells continue to grow and spread, leading to a decline in body function. During the diagnosis and treatment process, it may also cause emotional distress and psychological problems for patients. Studies have shown that cancer patients generally have a higher incidence of mood disorders such as depression and anxiety (4, 5).

Thyroid lesions often have no obvious symptoms in the early stages, but if not discovered and treated in time, they may gradually deteriorate into thyroid cancer, affecting the patient's quality of life and even endangering their life. Therefore, although thyroid cancer has certain hazards, early detection, early diagnosis and early treatment can achieve better treatment results, reduce the surgery rate and mortality rate, improve the cure rate and reduce complications.

In recent years, significant changes in environmental factors, specifically manifested as heavy metal pollutants, persistent organic pollutants (POPs), and increased air pollution (6–8), have adversely affected the normal physiological functions of thyroid hormones. The incidence of thyroid cancer is increasing year by year globally, accounting for approximately 1% to 3% of all new malignant tumors worldwide (9). Currently, the methods for screening thyroid diseases include ultrasound, cell puncture, CT, MRI, etc (10–12). Ultrasound is a common non-invasive and painless examination method (13). Its disadvantage is that it is limited by the doctor's experience and the size, shape, edge, internal echo and other characteristics of the nodule. Therefore, there is a certain misdiagnosis rate when evaluating the benign or malignant nature of thyroid nodules. Thyroid pathology is the gold standard for diagnosis and an important means of determining whether a thyroid nodule is benign or malignant and the type of thyroid tumor. However, pathology is invasive, expensive, and difficult for patients to accept. In order to achieve low-cost, high-accuracy early screening for thyroid disease, researchers have turned their attention to artificial intelligence technology.

The rapid advancement of artificial intelligence in image recognition technology has pushed auxiliary medical care to a highly mature and widely applied stage. In the field of image segmentation, deep learning image segmentation technology can automatically learn the features of images and achieve high-precision image segmentation by training deep neural. Xu (14) proposed an end-to-end FISH-based method (CACNET) for the recognition of genetically abnormal cells (CAC). The CACNET achieves cell nuclear segmentation by an improved Mask region-based convolutional neural network (R-CNN), and the accuracy of circulating CAC recognition using CACNET 94.06%. At the same time, they also developed a deep learning network (FISH-Net) based on 4-color FISH images for CACs, with an accuracy of more than 96% (15). Zhao (16) proposed a breast cancer ultrasound image segmentation method based on the U-Net framework combined with the residual block structure and attention, with a dic of up to 92.1%.

In the field of image classification, it mainly classifies and recognizes objects in images by training deep neural networks. This technology can process large-scale image data and quickly and accurately identify target objects in images. Its advantages include fast recognition speed, high accuracy, and the to handle images of different sizes and resolutions. In 2012, the deep convolutional neural network achieved a significant breakthrough in the ImageNet competition, showing excellent performance of 37.5% top-1 error rate and 17.0% top-5 error rate (17). In addition, Levy (18) proposed an innovative deep convolutional neural network model that cleverly used deep transfer learning technology to successfully achieve high-precision classification of benign and malignant breast tumors with an accuracy rate of up to 92.4%.

Wang (19) developed a mitosis detection method (FMDet) based on breast tissue histopathological images to capture the appearance changes mitotic cells. To achieve more robust feature extraction, the feature extractor was constructed by integrating a channel-level multi-scale attention mechanism into the fully convolutional network structure. The FMDet algorithm has won the first place in the MIDOG 2021 challenge, achieving an accuracy of 74.4%. In 2022, Su (20) used the gene expression data of TCGA to screen characteristic genes by combining WGCNA Lasso algorithms, and used machine learning models to achieve the diagnosis and staging of colorectal cancer. Wang (21) proposed a supervised learning (SSL) scheme of deep learning (DL) framework to address the challenge of high-precision classification seven pulmonary tumor growth patterns in whole slide images (WSIs). This series of technological innovations has undoubtedly injected strong impetus into the field of image segmentation and recognition, and has greatly promoted the application and development of artificial intelligence in early screening of thyroid diseases.

This article analyzes the application of artificial intelligence technology in the early diagnosis of thyroid diseases by comparing a large number of studies, summarizes the current application status of artificial intelligence technology in the early diagnosis of thyroid diseases, and studies the intelligent recognition technology of thyroid ultrasound images and pathological sections respectively. The aim is to explore a technology that can achieve objective screening of thyroid lesions in the early stages. Based on literature research, we explored the application of machine learning and deep learning in thyroid auxiliary diagnosis. We find that for small sample data, SVM and semi-supervised neural networks in deep learning perform better. U-Net has become the benchmark for most

image segmentation tasks, with an accuracy of more than 93%, thanks to its encoder-decoder architecture. Artificial intelligence technology enables auxiliary examination for early screening of thyroid diseases, improving the early cure rate and survival rate of patients, and enhancing the accuracy and of doctors' diagnosis. This study also prospects the future trends of artificial intelligence in the field of thyroid disease research, and constructs a set of artificial intelligence system for the whole process. The development of artificial intelligence in thyroid disease research is no longer limited to thyroid pathology or thyroid ultrasound, but has created an artificial intelligence that integrates thyroid images and clinical data of thyroid cancer, which is used to determine the diagnosis of thyroid cancer and can also accurately predict the postoperative survival period of thyroid cancer patients.

2 Methods

The PubMed database was accessed by computer for retrieval, using "thyroid ultrasound", "thyroid cytopathology" and "machine learning" as search terms. Figure 1 shows the number of publications in the field of thyroid in the past decade. A total of 75 articles were selected for analysis. According to the inclusion and exclusion criteria, 50 articles were finally determined for research and analysis. The inclusion criteria for this review were: (1) Machine learning and deep learning algorithms, such as U-net, K nearest neighbor classification, random forest, support vector machine and artificial neural network. (2) The accuracy of early diagnosis of thyroid disease area under the receiver operating characteristic curve. (3) The time selection is the literature published in 2014 and later in the past 10 years. (4) Except for the GLAS and RITE public datasets, most of them are self-built datasets, which reviewed the data of thyroid patients for years, including thyroid ultrasound images and thyroid pathological slices. The following summary measures were used: machine learning method, sample size, performance measure, and important features. In the early diagnosis of thyroid diseases, the successful application of artificial intelligence technology mainly focuses on two core areas: traditional machine learning methods and deep learning methods.

(1) Traditional Machine Learning: The goal is to train algorithms by analyzing data so that computers can automatically identify and make appropriate decisions (22). Machine learning can be divided into two main types of learning methods: supervised learning and unsupervised learning, which are widely used in many fields such as medical diagnosis, image recognition technology, and sentiment analysis (23). The significant progress made by machine learning in the field of medical image analysis has provided strong technical support for the early screening of thyroid diseases. For example, a study used a dataset from the UCI machine learning library to train a multi-class SVM classifier to classify thyroid diseases (24). The Thy-Wise model uses a random forest algorithm to classify thyroid nodules, showing high accuracy and specificity while reducing the rate of unnecessary biopsies (25).

(2) Deep Learning: Compared with traditional machine learning methods, deep learning has powerful learning capabilities and can better utilize data sets for feature extraction (26). The key technologies of deep learning include convolutional neural network (CNN), recurrent neural network (RNN) and U-Net (27). Deep learning technology has shown great potential and advantages in the classification, detection and segmentation of medical images. For example, the application of U-Net model in biomedical image segmentation (28) and the success of deep residual network in image recognition (29) have demonstrated the effectiveness of deep learning technology in processing complex medical image data.

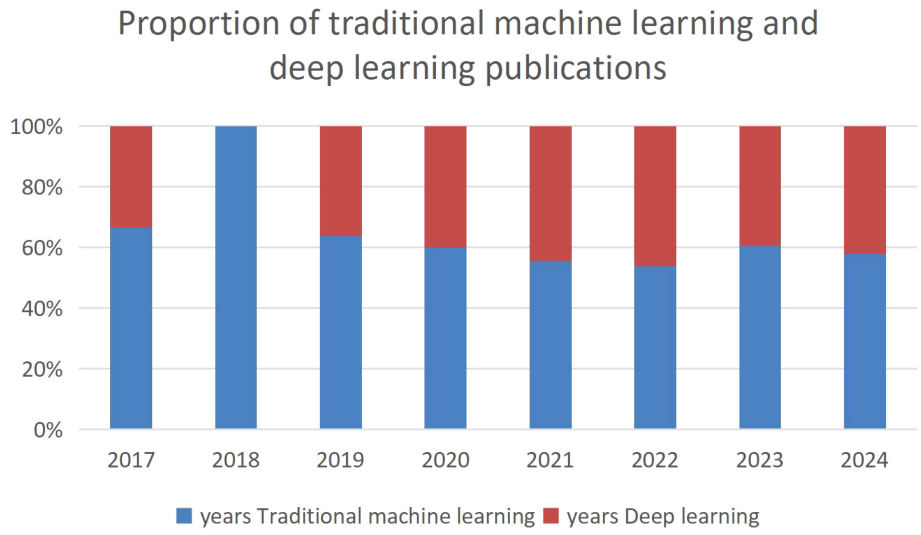


FIGURE 1
Proportion of traditional machine learning and deep learning publications.

3 Results

3.1 Thyroid ultrasound image recognition technology

Thyroid ultrasound diagnosis uses the principle of ultrasonic wave propagation and reflection in human tissues. It transmits ultrasonic waves to thyroid tissues through high-frequency probes, collects the reflected echo signals, and forms ultrasonic images of the thyroid gland. These images can clearly show the size, shape, structure and blood flow of the thyroid gland, providing doctors with rich diagnostic information. Due to its significant advantages of fast imaging, non-invasiveness and no radiation, it has become a widely used and trusted examination method (30–32). Although ultrasound technology has many significant advantages, it also faces some inherent limitations. First, it is unavoidable interference noise and possible artifacts. Second, the shape of thyroid nodules is complex and changeable, blurred, and discontinuous. The boundary characteristics. Third, it is limited by the subjective experience of doctors. These problems have brought certain challenges to accurate diagnosis (33, 34). Therefore, exploring the application of artificial intelligence technology to assist in the diagnosis of thyroid ultrasound has become a research hotspot. Table 1 shows some specific achievements artificial intelligence in the recognition of thyroid ultrasound images.

3.1.1 Traditional machine learning

In previous studies, ultrasound thyroid nodule segmentation methods can be roughly divided into four categories: shape and contour-based (46), region-based (47), machine learning-based (48), and hybrid methods (49).

At the beginning of the introduction of artificial intelligence technology in the medical field, researchers mainly relied on traditional machine learning algorithms. Therefore, the traditional machine algorithm was applied to the diagnosis of thyroid ultrasound images, aiming at improving the diagnostic speed and

TABLE 1 The main results of machine learning algorithms in the study of thyroid nodule ultrasound images.

Published year	Reference	Type of DL	Main Performance	Data	Conclusion
2017	Raghavendra et al. (35)	SVM	ACC: 97.5%, AUC: 94%	242 ultrasound images	spatial gray-level dependence features (SGLDF) and fractal texture.
2017	Ma et al. (36)	CNN	ACC: 91.5%	22123 ultrasound images	A multi-view strategy is used to improve the performance of the CNN based model.
2019	Nguyen et al. (37)	DCNN	Accuracy: 90.88%	237 nodules	cascade classifier
2019	Fu et al. (38)	RF,SVM	RF AUC: 95.4%, SVM AUC: 95.4%	1179 nodules(including 501 benign and 678 malignant)	The performance of RF and SVM is superior to other methods.
2020	Shin et al. (39)	SVM	ACC: 69.0%, Specificity: 79.4%, Sensitivity: 41.7%	348 nodules	GLCM, GLRLM, Gabor, and Haar wavelet
2021	Vadhiraj et al. (40)	MIL	ACC: 96%	99 patients (33 benign, 66 malignant)	GLCM
2021	Peng et al. (41)	ThyNet	AUR: 92.2%	18049 ultrasound images	The proportion of missed malignant thyroid nodules has decreased.
2022	Zhou et al. (42)	MSA-UNet	ACC: 94.6%, Dic: 84.6%	1083 patients	Atrous Spatial Pyramid Pooling.
2023	Li et al. (43)	WSDAC	Dic: 87%	350 ultrasound images	Models can reduce the workload of labeling datasets.
2024	Chen et al. (44)	CNN	CNN AUC: 91%, Inception-ResNet AUC: 94%	11201 ultrasound images	The article conducted substantial, non-substantial, and benign malignant classification studies on ultrasound images. Inception-ResNet, due to the expertise of a senior doctor.
2024	Ma et al. (45)	KNN	ACC: 86.7%	508 ultrasound images	The study considered the impact of different distance weights, k-values, and distance metrics on the classification results.

accuracy of benign and malignant nodules. In 2017, Raghavendra (35) designed a computer-aided diagnosis system (CAD) for the diagnosis of nodules. The system identifies the lesion area by integrating spatial gray-level dependence features (SGLDF) and fractal texture. This feature fusion-based approach achieved an accuracy of 97.5% and an AUC value of 94% for the support SVM using only two features, which is about 3.5% higher than the performance of the SVM proposed by Acharya et al. (50) How to use the right features to improve classification performance has always been a challenge.

Shin I (39) developed an artificial neural network (ANN) based on SVM for the classification model of thyroid tumors in 2020, using 348 preoperative ultrasound images of thyroid nodules as the dataset, and selected 10 important features as the feature input of the model. Then, the effect of the model was compared with the results of manual diagnosis by experienced radiologists. The results showed that the sensitivity, specificity and accuracy of the model were 32.3%, 90.1% and 74.%, respectively, while the sensitivity, specificity and accuracy of the diagnosis by general physicians were 24.0%, 84.0% and 648%. It was proved that the classifier model of machine learning may be helpful in the diagnosis of thyroid cancer.

In 2021, Vadhirajt (40) developed a computer-aided diagnosis system integrating multiple instance learning (MIL) to classify benign and malignant thyroid ultrasound images. Seven ultrasound image features were extracted using the gray-level co-occurrence matrix (GLCM) with an accuracy of 96%. Ma (45) proposed an improved KNN algorithm for automatic classification of thyroid nodules. The paper not only considered the number of class labels of various data categories in KNNs, but also considered the corresponding weights, using the Minkowski distance measurement. Using 508 thyroid nodule hyper images, the improved KNN accuracy was 86.7%. Through summarizing and analyzing the previous studies, we find that different feature selection will have a certain impact on the accuracy of the model.

At the same time, in order to evaluate which algorithm in linear and nonlinear machine learning is better for the benign and malignant classification diagnosis of thyroid nodules, Fu (38) used three linear and five nonlinear machine learning algorithms to evaluate 1039 patients with a total of 1179 nodules. Experimental results have shown that the AUC of machine learning models is higher than that of experienced radiologists. Among them, the AUC of RF and SVM methods in nonlinear machine learning is the highest, both at 95.4%, while the AUC of experienced doctors is only about 83%.

At present, a large number of computer-aided diagnosis systems based on traditional machine learning rely mainly on a variety of texture features and machine learning algorithms differentiating the benign and malignant nature of thyroid nodules, and their accuracy is about 3% higher than that of general doctors. In order to further improve the classification accuracy, the researchers adopted a variety of optimization methods, such as GLCM, SGLDF, to fine-tune the input features and parameters of the machine learning models, making these models show applicability in thyroid diagnosis.

3.1.2 Deep learning

With the continuous advancement of artificial intelligence technology, the application of deep learning in the medical field has become the focus of research. In 2017, Ma (36) first attempted to use a CNN-based model for thyroid nodule segmentation and compared this method with six methods including GA-VBAC, JET, DRLS, SNDRLS, SVM-based method and RBFNN-based method. The study used a total of 22123 thyroid ultrasound images from three hospitals as the dataset. The results show that our proposed CNN-based model has a good performance in the segmentation of thyroid nodules with an accuracy of 91.5%. Peng (41) developed a deep learning model based on ThyNet to distinguish benign and malignant thyroid nodules, and the results showed that the AUC was 92.2%, and the proportion of missed malignant thyroid nodules decreased from 18.9% to 17.0%, reducing fine needle aspiration examinations. In 2024, Chen (44) proposed a convolutional neural network (CNN) model using 11201 images for training, validation and testing. Experiments have shown that the AUC of the model in the classification of benign and malignant thyroid nodules is higher than 91%, among which Inception-ResNet has the highest AUC of 94%, and the performance of the model is better than that of senior physicians.

In artificial intelligence applications, feature selection is key to improving model accuracy. In 2019, Nguyen (37) developed a method for extracting features from thyroid images, using a cascade classifier architecture to improve performance of computer-aided diagnosis systems for thyroid nodule classification. This method combined handcrafted standards and deep learning, achieving a classification accuracy of 90.8%. Gong (51) designed a new multi-task learning framework to simultaneously learn nodule size, glandular location, and nodule position, and proposed an adaptive glandular region feature enhancement module to fully utilize thyroid prior knowledge and use the prior to guide the feature enhancement network to accurately segment thyroid nodules. Different radiomic features were extracted from ultrasound images, including intensity, shape, and texture feature sets.

Although the popularity of deep learning has significantly improved the accuracy of image segmentation, problems with datasets, especially the lack of precisely annotated datasets, can still affect prediction accuracy of models. However, such data is often difficult to obtain in the field of medical image analysis. To solve this problem, Wang (52) proposed an attention-based semi-supervised neural network for thyroid nodule segmentation. The network can complete the thyroid ultrasound image segmentation task using a small amount of fully annotated data and a large amount of weakly annotated data. The article proposes two attention modules, which realize the inhibition or activation of bottom-up and top-down feature channels and image areas through a trainable feed-forward structure, thereby improving network performance. The Jaccard similarity coefficient of the semi-supervised neural network based on attention is 74.91%, which is 4.9% higher than that of the semi-supervised model based on VGG. The accuracy of benign and malignant thyroid tumor classification

was improved from 91.67% to 95.00%, which proved that model had good generalization ability.

Li (43) proposed a weakly supervised deep active contour model for thyroid nodule segmentation, aiming to achieve accurate target segmentation with a small amount of annotation information. The experiment designed three modules: a weakly supervised learning framework, a deep active contour model, and auxiliary edge attention, which can reduce the annotation cost while maintaining a certain segmentation accuracy. The dic value of the model is 87%, which can reduce the workload of dataset annotation.

With the widespread application of deep learning, the U-net algorithm was proposed. U-Net is a convolutional neural network (CNN) structure widely used in deep learning, mainly for image segmentation tasks (53, 54). Ding (55) mainly explored the automatic segmentation technology of thyroid ultrasound images based on U-net. The model embedded an improved residual unit in the jump connection between the encoder and decoder paths and introduced an attention gate mechanism to enhance the weights of feature maps obtained from shallow and deep layers. Experimental results show that the proposed method outperforms other U-shaped models.

In 2020, Zhang (56) proposed two network structures, Cascade U-Net and CH-UNet, for the segmentation and classification of thyroid nodules. Cascade U-Net gradually refines the segmentation results and improves the segmentation accuracy by cascading multiple U-Net modules. CH-UNet combines dilated convolution and hybrid attention mechanism to enhance feature extraction and classification capabilities. Compared with the U-Net proposed by RONNEBERGER (55), the dice of Cascade U-Net in the task of thyroidodule segmentation increased by 2.9%. The dice of the U-Net method by RONNEBERGER (57) was only 80.2%, which fully validated effectiveness of the Cascade U-Net in the segmentation and even classification tasks of thyroid nodules.

In order to accurately detect malignant nodules that are not obvious and have confused boundaries in ultrasound images, and to avoid confusion between tissue and malignant thyroid nod during diagnosis, Yang (58) proposed a deep learning-based thyroid malignant nodule segmentation method of DMU-Net. The method uses the image context information in the U-shaped subnetwork to accurately locate the malignant nodule region, and then captures the fine details of theodule edges in the inverse U-shaped subnetwork. The combination of U-shaped subnetwork and inverse U-shaped subnetwork and the strategy of mutual learning make the dic of DMU-Net on the-built dataset 82.77%, which is 25.86% higher than that of the traditional U-Net network. The research proves that DMU-Net can accurately locate the malignant nodule area by extracting image context information in the U-shaped subnetwork, extract more lesion area features, and help radiologists diagnose thyroid diseases.

In 2022, Zhou (42) proposed an MSA-UNet model with a multi-scale self-attention mechanism for thyroid nodule segmentation. Depth wise separable convolution is used in the Atrous Spatial Pyramid Pooling (ASPP) module, and then in the decoder part, adjacent information of different scales is fused through the channel attention mechanism, allowing the model to learn more important features. The experimental results show that the accuracy of this method is 94.6%, which provides a new research idea for the early detection of thyroid nod. Comparison of accuracy of different U-Net algorithms, as shown in Table 2.

Currently, the research focus of thyroid ultrasound images is mainly on the segmentation and classification tasks of thyroid nodules, but the potential intrinsic connection and mutual influence between nodule characteristics and classification results are often ignored. Thyroid nodule segmentation and classification in ultrasound images are two fundamental but challenging tasks in computer-aided diagnosis of thyroid diseases. Since these two tasks are intrinsically related and share some common features, it is a promising direction to jointly solve these two problems using multi-task learning. However, previous studies have only demonstrated inconsistent predictions between these related tasks. In order to further exploit the effectiveness of the proposed task consistency learning, Kang (61) designed a framework based on multi-task learning (MS-MTL) to improve the performance of thyroid segmentation and classification by improving the consistency between tasks. The first stage of the network performs binary segmentation and classification simultaneously, and the second stage of the network learns multi-class segmentation. The article verifies the feasibility of improving thyroid nodule segmentation and classification performance through multi-task learning and inter-task consistency loss.

The application of deep learning in thyroid ultrasound images has broad significance and value. Various models have been applied to the processing of thyroid ultrasound images, including convolutional neural networks (CNN), U-net etc. By training a large amount of data, these models can learn the key features in ultrasound images for the classification and identification of nodules, thereby reducing misdiagnosis and missed diagnosis caused by human factors and helping to improve the early diagnosis rate. The application of artificial intelligence technology to assist in the early screening of thyroid diseases is not only limited to the diagnosis of thyroid ultrasound pictures, but also shows significant results in the recognition of thyroid pathology icons.

TABLE 2 Comparison of U-Net methods.

Reference	Methods	Recall	Accuracy	Dice
Ronneberger (57)	U-Net	86.1	93.2	80.2
Badrinarayanan (59)	SegNet	88.5	94	81.2
Zhou (60)	UNET++	85.9	93.8	80.8
Zhang et al. (56)	Cascade U-Net	86.6	94.3	83.1
Zhou et al. (42)	MSA-UNet	87	94.6	84.6

3.2 Thyroid pathology section recognition technology

Thyroid pathology examination is a common diagnostic procedure and an important part of the evaluation of thyroid nodules, but there is significant variability in the assessment thyroid cytology specimens by different pathologists and

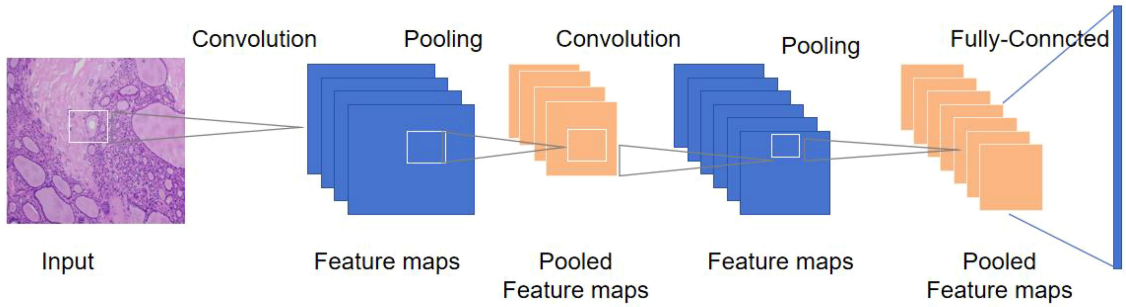


FIGURE 2
Convolutional neural network processing model for thyroid pathological images.

institutions. The sensitivity reported in the literature ranges from 68% to 98%, and the specificity ranges from 56% to 100%. In this case, the use of machine learning can improve accuracy and help standardize the diagnosis of thyroid pathological specimens (62). The process of processing pathological images using convolutional neural networks is shown in Figure 2.

One of the earliest studies on thyroid pathology was conducted by Karakitsos (63), who investigated the ability of a learning vector quantization (LVQ) neural network to distinguish benign from malignant thyroid lesions. The model was trained by measuring 25 features such as size, shape, and texture of approximately 100 nuclei in each case. The results of the study show that the LVQ neural network can distinguish benign from malignant lesions very well, with an accuracy of 90.6%.

In 2011 study also investigated the application of learning vector quantization (LVQ) neural networks in differentiating benign from malignant thyroid lesions using 335 fluid-based cytology, fine needle aspiration (FNA), and Papanicolaou stain specimens. Features extracted by a custom image analysis system were first used to classify each nucleus using an LVQ neural network, and then a second LVQ neural network was used to classify individual lesions. The system was able to distinguish between benign and malignant nuclei and lesions at both the cellular and patient levels (64). Lee (65) developed a machine learning algorithm (MLA) that can classify human thyroid cell clusters by utilizing Papanicolaou staining and intrinsic refractive index (RI) as relevant imaging contrast agents and evaluated the impact of this combination on diagnostic performance. The accuracy of the MLA classifier for 1535 thyroid cell clusters from 124 patients using color images, RI images, and both was 98.0%, 98.0%, and 100%, respectively. The importance of this study lies in the fact that it compares a variety of different diagnostic techniques to improve the accuracy and efficiency of thyroid cancer diagnosis, with MLA classifier achieving the highest accuracy.

Artificial intelligence technology not only achieves precise classification and recognition functions in the processing of thyroid pathological images, but also shows strong prediction capabilities. Improving the of malignant tumor prediction can reduce the incidence of unnecessary surgery. Elliott (66) created a machine learning algorithm (MLA) based on two CNNs to identify follicular cells and predict the malign of the final pathology. The AUC of the model reached 93.2%, which is equivalent to the AUC of 93.1% diagnosed by cell pathologists, demonstrating the effectiveness of the algorithm. Wang (67) developed a prediction system for benign and malignant medullary thyroid cancer and goiter based on SVM and RF algorithms. For the classification of PTC and nodular goiter (NG), the SVM and RF algorithms performed equally well, with 94.2% and 94.4% consistency between the prediction and pathological diagnosis. The system can shorten the diagnosis time of doctors, making the diagnosis time of each sample only 10 minutes, which is very promising for the diagnosis papillary thyroid carcinoma during surgery. This method can also correctly predict the malignancy of a medullary thyroid carcinoma and a follicular thyroid adenoma.

Due to the combined effect of genetic variants, environmental exposure, and immune genetic risk (68, 69), new types of thyroid tumors, as" non-invasive follicular thyroid neoplasm"(NIFTP), have emerged, which has complicated the cytology of thyroid cells, and a lot data have been classified into the category of uncertainty (70).

Hirokawa (71) proposed an artificial intelligence image classification system of EfficientNetV2-L, which proved the efficiency and of artificial intelligence image classification system in identifying thyroid lesions. The research team used 148,395 thyroid pathology smear images from 393 thyroid nodules as the dataset. The researchers reported that the AUC of EfficientNetV2-L exceeded 95%. However, the AUC for poorly differentiated thyroid cancer was only 49%, showing significantly worse performance.

In another study, Yao (72) proposed a digital image analysis method based on feature engineering and supervised machine learning. They focused on cases of poorly differentiated thyroid cancer that were later diagnosed as benign or follicular adenoma in his tissue sections. The method was applied to 40 thyroid pathological slices with high and low power microscopy, and the AUC for the low power image model was 5%, and the AUC for the high power image model was 74%. This method performs better than cellular pathologists in classifying atypical follicular lesions.

The application of artificial intelligence in the field of thyroid pathology image analysis not only significantly enhances the accuracy and timeliness of diagnosis (73), but also relies on its deep learning and image processing technology to realize the analysis of pathological images such as follicular cell morphology and arrangement. Accurate identification of subtle features such as

pattern and abnormal proliferation. These key features are of irreplaceable importance for accurately distinguishing benign and malignant thyroid nodules. Compared with traditional manual diagnostic methods, the integration of artificial intelligence has greatly promoted the early detection, accurate diagnosis and timely treatment of thyroid diseases, bringing patients a higher survival rate and better quality of life.

4 Discussion

In recent years, the research, development and application of artificial intelligence in the field of thyroid diagnosis have achieved significant leaps, providing new horizons and broad possibilities for optimizing the efficiency and accuracy of future diagnostic processes. Especially in the early diagnosis of thyroid cancer, artificial intelligence technology can automatically identify and evaluate complex medical images through machine learning algorithms, thereby improving the accuracy and efficiency of diagnosis.

In the application of thyroid ultrasound images, AI technology has been shown to effectively assist radiologists in the diagnosis of thyroid nodules. For example, one study showed that the performance of an AI system in the diagnosis of thyroid nodules was comparable to that of fine needle aspiration cytology (74). In addition, AI technology also showed high accuracy and efficiency in distinguishing benign from malignant thyroid nodules (75). Based on the previous research, we find that the research methods of thyroid ultrasound images mainly focus on traditional machine learning and deep learning. In traditional machine learning, SVM and RF have high accuracy in thyroid nodule classification due to their superior binary classification performance.

The core concept of SVM lies in the strategy of structural risk minimization, aiming to determine the optimal complexity of the model a limited dataset, thereby enhancing the model's general prediction capability. The model parameters of SVM only depend on the support vectors, which are the data points closest to the decision boundary, and have no direct connection with other points. This means that even with a small number of samples, as long as these support vectors can fully reflect the overall distribution characteristics of the data, SVM can construct an efficient and accurate classification model. Therefore, SVM is particularly suitable for dealing with thyroid datasets with a small sample size.

Compared with machine learning, deep learning has strong learning ability and efficient feature expression ability, which can automatically learn and extract high-level features in images and can more comprehensively capture the details and context information of images, thus improving the accuracy of classification. The deep convolutional neural network (DCNN) model proposed by Krizhevsky (76) achieved breakthrough results in the ImageNet image classification. Therefore, the current research focuses on the classification of thyroid ultrasound and pathological images using deep learning.

Compared with traditional segmentation techniques, the segmentation method based on deep learning does not rely on hand-designed features, and the convolutional neural network (CNN) has shown excellent adaptability in the field of medical image segmentation by virtue of its image hierarchical feature representation capability. ROMÁN (77) reviewed a large number of deep learning-based medical image segmentation methods, among which U-Net is the most typical. The core idea of U-Net is to adopt a symmetric encoder-decoder architecture, which enables deep feature extraction and precise pixel-level segmentation of the input. Liu (78) proposed an automated segmentation algorithm for brain gliomas based on a multi-U-Net network(MU-Net), and conducted experiments on the BRATS2020 dataset. The results show that the Dice coefficients of the MU-Net algorithm for the complete tumor, tumor core, and enhanced tumor are 86.7%, 77.76%, and 76.21%, respectively, which are 2.6%, 2.55%, and 2.41% higher than those of the benchmark model, indicating better segmentation results. The application of these technologies can not only help radiologists diagnose thyroid diseases more accurately and improve diagnostic efficiency, but also reduce their workload.

AI technology also shows great potential in the application of thyroid pathology images. For example, AI technology has been used in cytological analysis of thyroid fine needle aspiration biopsy to distinguish papillary carcinoma from other types of thyroid cancer (79). A hybrid framework combining artificial intelligence was proposed in the study (80), which not only weighted the Thyroid Imaging Reporting and Data System (TIRADS) features, but also used the malignancy score predicted by the convolutional neural network (CNN) to classify and diagnose the malignancy of the nodules.

In summary, artificial intelligence technology has strong clinical significance and application prospects in the application of thyroid ultrasound images and thyroid pathological images. Not only has it improved the accuracy and efficiency of diagnosis, assisted doctors in decision-making, reduced the rate of misdiagnosis, but it can also the allocation of medical resources, reduce unnecessary surgeries and other invasive treatments through artificial intelligence-assisted diagnosis, and reduce the economic burden and pain of patients.

With the continuous advancement of technology and the deepening of clinical applications, artificial intelligence technology has played an increasingly important role in the early diagnosis of thyroid diseases, but the prediction of the postoperative life cycle of thyroid cancer patients is equally important for doctors and patients. This study (81) used artificial neural networks (ANN) to predict the 1-year, 3-year, and 5-year survival of thyroid cancer patients, with accuracy rates of 92.9%, 85.1%, and 86.8%, respectively. Based on our research results, artificial neural networks can effectively represent a survival prediction method for thyroid cancer patients. Liu (9) developed six machine learning models (SVM, XGBoost, LR, DT, RF and KNN) based on the SEER database to predict lung metastasis of thyroid cancer. Although the accuracy of the model is above 90%, prospective studies are still needed to further verify the practicality of the model. And because the genes of thyroid cancer patients may undergo mutation, gene mutation increases the complexity of the data, and the model may have difficulty accurately distinguishing different of diseases. On the other hand, gene mutation may have a complex interaction with other biomarkers or clinical information, which may make a single classification algorithm fail to capture the

information accurately (82), and all these will lead to a bias in the accuracy of the algorithm model.

In the future, we will focus on optimizing the cutting-edge exploration of machine learning algorithm models, integrating patient pathological information, radiology and clinical information, create a more powerful algorithm, aiming to build a set of artificial intelligence system for the whole process. The system will have the ability to deeply analyze massive clinical records and molecular biology data to accurately predict the postoperative survival of thyroid cancer patients, thereby assisting doctors in tailoring more precise treatment strategies for each patient, thereby significantly improving late-stage Prognosis and quality of life in patients with thyroid cancer.

5 Conclusions

This paper reviews the latest application progress of artificial intelligence technology in the field of medical diagnosis, focusing on its potential in the early screening and diagnosis of thyroid. The research hotspot has developed from the initial traditional machine learning to deep learning algorithms, and U-Net has also become the benchmark for most medical image segmentation with the encoder-decoder architecture. Through the previous research, it aims to assist clinicians in achieving intelligent and efficient early identification of thyroid cancer, thereby improving the accuracy of early diagnosis for patients enhancing the efficiency of doctors. Moreover, the article also prospects the future trend of artificial intelligence in the field of thyroid disease research, not only limited to thyroid pathology or thyroid ultrasound but also creating artificial intelligence that integrates thyroid ultrasound images and clinical data of thyroid cancer, which is used to determine the diagnosis of thyroid cancer, and can also accurately predict postoperative survival period of thyroid cancer patients. It aims to provide new research directions for scientific researchers, and bring more personalized treatment plans for doctors and patients through the continuous progress of artificial intelligence technology, treatment strategies, and improve patients' satisfaction and quality of life.

Author contributions

LY: Writing – original draft, Writing – review & editing, Methodology. XW: Funding acquisition, Writing – original draft, Writing – review & editing. SZ: Data curation, Writing – original draft. KC: Supervision, Validation, Writing – review & editing. JY: Funding acquisition, Validation, Writing – review & editing, Data curation.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was supported by the Shandong Medical Science and Technology Project (202325011546) and Natural Science Foundation of Shandong Province (nos. ZR2021MH227).

Conflict of interest

The study was conducted without any commercial or financial relationships where there was no 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

References

1. Mousa S, Hercbergs A, Lin H-Y, Keating K, Davis P. Actions of thyroid hormones on thyroid cancers. *Front Endocrinol.* (2021) 12:36. doi: 10.3389/fendo.2021.691736

2. Ali L, Mohy U Din MT, Gessl A, Lemmens-Gruber R, Kautzky-Willer AGessl A, Lemmens-Gruber R, Kautzky-Willer A, et al. THYROID DISORDERS. *Prof Med J.* (2015). doi: 10.29309/TPMJ/2015.22.10.982

3. Qin J, Shi X. Research progress in the etiology of thyroid cancer. *Chin J Otorhinolaryngol Head Neck Surg.* (2020) 55:711–5. doi: 10.3760/cma.j.cn115330-20191226-00783

4. Liu Z. Emotional disorder and related psychosocial factors in tumor patients. *Chin J Rehabil Theory Pract.* (2017).

5. Wu B-j, Wang M-j. Investigation of psychological health status in Malignant tumor patients. *Int J Nurs.* (2015) 00:1315–7. doi: 10.3760/CMA.J.ISSN.1673-4351.2015.10.008

6. Han X, Meng L, Li Y, Li A, Turyk ME, Yang R, et al. Associations between exposure to persistent organic pollutants and thyroid function in a case-control study of east China. *Environ Sci Technol.* (2019) 53:9866–75. doi: 10.1021/acs.est.9b02810

7. Hybenova M, Hrda P, Procházková J, Stejskal V, Sterzl I, et al. The role of environmental factors in autoimmune thyroiditis. *Neuro Endocrinol Lett.* (2010) 31:283–9.

8. Faustini A, Renzi M, Kirchmayer U, Balducci M, Davoli M, Forastiere F, et al. Short-term exposure to air pollution might exacerbate autoimmune diseases. *Environ Epidemiol.* (2018) 2:e025. doi: 10.1097/EE9.0000000000000025

9. Liu W, Wang S, Ye Z, Xu P, Xia X, Guo M. Prediction of lung metastases in thyroid cancer using machine learning based on SEER database. *Cancer Med.* (2022) 11:2503–15. doi: 10.1002/cam4.v11.12

10. Lee JY, Lee K, Seo BK, Cho KR, Woo OH, Song SE, et al. Radiomic machine learning for predicting prognostic biomarkers and molecular subtypes of breast cancer using tumor heterogeneity and angiogenesis properties on MRI. *Eur Radiol.* (2022) 32:650–60. doi: 10.1007/s00330-021-08146-8

11. Lima LMF, Bogsrud TV, Gharib H, Ryder M, Johnson G, Dursk J. Risk of Malignancy in thyroid nodules with increased 11C-Choline uptake detected incidentally on PET/CT. A diagnostic accuracy study. *Medicine.* (2024) 103:204. doi: 10.1097/MD.0000000000039602

12. Gao L, Wu S, Wongwasuratthakul P, Chen Z, Cai W, Li Q, et al. Label-free surface-enhanced raman spectroscopy with machine learning for the diagnosis of thyroid cancer by using fine-needle aspiration liquid samples. *Biosens (Basel).* (2024) 14:372. doi: 10.3390/bios14080372

13. Haugen BR, Alexander EK, Bible KC, Doherty G M, Mandel SJ, Nikiforov YE, et al. American Thyroid association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American thyroid association guidelines task force on thyroid nodules and differentiated thyroid cancer. *Thyroid*. (2016) 26:1–133. doi: 10.1089/thy.2015.0020

14. Xu X, Li C, Fan X, Lan X, Lu X, Ye X, et al. Attention Mask R-CNN with edge refinement algorithm for identifying circulating genetically abnormal cells. *Cytom A*. (2023) 103:227–39. doi: 10.1002/cyto.a.24682

15. Xu X, Li C, Lan X, Fan X, Lv X, Ye X, et al. A lightweight and robust framework for circulating genetically abnormal cells (CACs) identification using 4-color fluorescence *in situ* hybridization (FISH) image and deep refined learning. *J Digit Imaging*. (2023) 36:1687–700. doi: 10.1007/s10278-023-00843-8

16. Zhao T, Dai H. Breast tumor ultrasound image segmentation method based on improved residual U-Net network. *Comput Intell Neurosci*. (2022) 2022:3905998. doi: 10.1155/2022/3905998

17. Krizhevsky A, Sutskever et al. I. ImageNet classification with deep convolutional neural networks. *Commun ACM*. (2012) 60:84–90. doi: 10.1145/3065386

18. Lévy D, Jain A. Breast mass classification from mammograms using deep convolutional neural networks. *Comput Sci*. (2016). doi: 10.48550/arXiv.1612.00542

19. Wang X, Zhang J, Yang S, Xiang J, Luo F, Wang M, et al. A generalizable and robust deep learning algorithm for mitosis detection in multicenter breast histopathological images. *Med image anal*. (2023) 84:102703. doi: 10.1016/j.media.2022.102703

20. Su Y, Tian X, Gao R, Guo W, Chen C, Chen C, et al. Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis. *Comput Biol Med*. (2022) 145:105409. doi: 10.1016/j.compbiomed.2022.105409

21. Wang Q, Zhang Y, Lu J, Li C, Zhang Y. Semi-supervised lung adenocarcinoma histopathology image classification based on multi-teacher knowledge distillation. *Phys Med Biol*. (2024) 69:7454. doi: 10.1088/1361-6560/ad7454

22. Bernardes R. Machine learning - Basic principles. *Acta ophthalmol*. (2024). doi: 10.1111/aos.v102.s279

23. Chatzilygeroudis K, Hatzilygeroudis I, Parisa E., Andreas K., Mark D, et al. Intelligent computing for interactive system design. *Mach Learn Basics*. (2021). doi: 10.1145/3447404

24. Kumar HHS. (2020). A novel approach of SVM based classification on thyroid disease stage detection, in: *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, (2020). pp. 836–41. doi: 10.1109/ICSSIT48917.2020.9214180

25. Jin Z, Pei S, Ouyang L, Zhang L, Mo X, Chen Q, et al. Thy-Wise: An interpretable machine learning model for the evaluation of thyroid nodules. *Int J Cancer*. (2022) 151:2229–43. doi: 10.1002/ijc.v151.12

26. Du X, Cai Y, Wang S., Zhang L, et al. (2016). Overview of deep learning. In: *Youth Academic Annual Conference of Chinese Association of Automation*, (2016) . pp. 159–64. doi: 10.1109/YAC.2016.7804882

27. Zhang R, Li W, et al. Review of deep learning. *Comput Sci*. (2018).

28. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2015). pp. 234–41.

29. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Comput Vision Pattern Recogn*. (2015) 18:770–8. doi: 10.48550/arXiv.1512.03385

30. Bakkar Y, Bakkar R. The role of ultrasound in thyroid assessment. in: *World Family Med*. (2023) 21:97–106. doi: 10.5742/MEWFM.2023.95256079

31. Chikui T, Okamura K, Tokumori K, Nakamura S, Shimizu M, Koga M, et al. Quantitative analyses of sonographic images of the parotid gland in patients with sjögrens syndrome. *Ultrasound*. (2006) 32:617–22. doi: 10.1016/j.ultrasmedbio.2006.01.013

32. Bojunga J, Trimboli P. Thyroid ultrasound and its ancillary techniques. *Rev Endocr Metab Disord*. (2024) 25:161–73. doi: 10.1007/s11154-023-09841-1

33. Alexander E. Understanding the ability, and inability, of high-resolution ultrasound to guide thyroid nodule evaluation. *Cancer Cytopathol*. (2020) 128:236–7. doi: 10.1002/cncy.v128.4

34. Renshaw A, Gould et al. E. Thyroid FNA: Is cytopathologist review of ultrasound features useful. *Cancer Cytopathol*. (2020) 128:523–7. doi: 10.1002/cncy.v128.8

35. Raghavendra U, Rajendra Acharya U, Gudigar A, Hong Tan J, Fujita H, Hagiwara Y, et al. Fusion of spatial Gray level dependency and fractal texture features for the characterization of thyroid lesions. *Ultrasonics*. (2017) 77:110–20. doi: 10.1016/j.ultras.2017.02.003

36. Ma J, Wu F, Jiang T. Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks. *Int J Comput Assist Radiol Surg*. (2017) 12:1895–910. doi: 10.1007/s11548-017-1649-7

37. Nguyen DT, Pham TD, Batchuluun G, Yoon HS, Park KR. Artificialintelligence-based thyroid nodule classification using information from spatial and frequency domains. *J Clin Med*. (2019) 8:1976. doi: 10.3390/jcm8111976

38. Ouyang F-s, Guo B-l, Ouyang L-z, Liu Z-w, Lin S-j, Meng W. Comparison between linear and nonlinear machine-learning algorithms for the classification of thyroid nodules. *Eur J Radiol*. (2019) 04:251–7. doi: 10.1016/j.ejrad.2019.02.029

39. Shin I, Kim YJ, Han K, Lee E, Kim HJ, Shin JH, et al. Application of machine learning to ultrasound images to differentiate follicular neoplasms of the thyroid gland. *Ultrasonography*. (2020) 39:257–65. doi: 10.14366/usg.19069

40. Vadhiraj VV, Simpkin A, O'Connell J, Singh Ospina N, Maraka S, O'Keeffe DT. Ultrasound image classification of thyroid nodules using machine learning techniques. *Medicina (Kaunas)*. (2021) 57:527. doi: 10.3390/medicina57060527

41. Peng S, Liu Y, Lv W, Liu L, Zhou Q, Yang H, et al. Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study. *Lancet Digit Health*. (2021) 3:250–9. doi: 10.1016/S2589-7500(21)00041-8

42. Zhou X, Zhao J. Multi-scale supervised U-Net ultrasound image segmentation of thyroid nodule. *J Taiyuan Univ Technol*. (2022) 53:1134–42.

43. Li Z, Zhou S, Chang C, Wang Y, Guo Y. A weakly supervised deep active contour model for nodule segmentation in thyroid ultrasound images. *Pattern Recogn Lett*. (2023) 01:128–38. doi: 10.1016/j.patrec.2022.12.015

44. Chen C, Jiang Y, Yao J, Lai M, Liu Y, Jiang X, et al. Deep learning to assist composition classification and thyroid solid nodule diagnosis: a multicenter diagnostic study. *Eur Radiol*. (2024) 34:2323–33. doi: 10.1007/s00330-023-10269-z

45. Ma X, Han X, Zhang L. An improved k-nearest neighbor algorithm for recognition and classification of thyroid nodules. *J Ultrasound Med*. (2024) 43:1025–36. doi: 10.1002/jum.16429

46. Du W, Sang N. (2015). An effective method for ultrasound thyroid nodules segmentation, in: *2015 International Symposium on Bioelectronics and Bioinformatics (ISBB)* (2015) 207–10. doi: 10.1109/ISBB.2015.7344960

47. Alrubaidi WM, Peng B, Yang Y, Chen Q. An interactive segmentation algorithm for thyroid nodules in ultrasound images. In: *International conference on intelligent computing*. Springer (2016) 107–15.

48. Chang CY, Huang HC, Chen SJ. (2009). Thyroid nodule segmentation and component analysis in ultrasound images. In: *Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association*. 2009, 910–7.

49. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. (2016). Learning deep features for discriminative localization, In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016) 2921–9.

50. Acharya UR, Chowriappa P, Fujita H, Bhat S, Dua S, Koh JEW, et al. Thyroid lesion classification in 242 patient population using gabor transform features from high resolution ultrasound images, Knowl. *Based Syst*. (2016) 171:235–45. doi: 10.1016/j.knosys.2016.06.010

51. Gong H, Chen J, Chen G, Li H, Li G, Chen F. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. *Comput Biol Med*. (2023) 155:106389. doi: 10.1016/j.compbiomed.2022.106389

52. Wang J, Zhang R, Wei X, Li X, Yu M, Zhu J. (2019). An attention-based semi-supervised neural network for thyroid nodules segmentation, in: *2019 IEEE International Conference on Bioinformatics an Biomedicine (BIBM)*, (2019) 871–6. doi: 10.1109/BIBM47256.2019.8983288

53. Siddique N, Paheding S, Elkin CP, Devabhaktuni V. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*. (2020) 9:82031–57. doi: 10.1109/ACCESS.2021.3086020

54. Zahra E, Ali B, Siddique W. Medical image segmentation using a U-net type of architecture. *ArXiv abs*. (2005). doi: 10.48550/arXiv.2005.05218

55. Ding J, Huang Z, Shi M, Ning C, et al. Automatic thyroid ultrasound image segmentation based on ushaped network, in: *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. (2019) 1–5. doi: 10.1109/CISP-BMEI48845.2019.8966062

56. Zhang Y, Lai H, Yang W. (2020). Cascade UNet and CH-UNet for thyroid nodule segmentation and benign and Malignant classification, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2020) 129–34.

57. Ronneberger O, Fischer P, Brox T. *U -Net:convolutional networks for biomedical image segmentation*. Springer International Publishing (2015) 234–41.

58. Yang Q, Geng C, Chen R, Pang C, Han R, Lyu L, et al. DMU-Net: Dual-route mirroring U-Net with mutual learning for Malignant thyroid nodule segmentation. *Biomed Signal Process Control*. (2022) 08:103805–18. doi: 10.1016/j.bspc.2022.103805

59. Badrinarayanan V, Kendall A, Cipolla R. SegNet: adeepconvolutionalencoder-decoderarchitectureforimagesegmentation. *IEEE Trans Pattern Anal Mach Intell*. (2017) 39:2481–95. doi: 10.1109/TPAMI.34

60. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: Anested UNet architecture for medical image segmentation. *Deep Learn Med Image Anal Multimodal Learn Clin Decis Support*. (2018) 11045:3–11. doi: 10.1007/978-3-030-00889-5_1.

61. Kang Q, Lao Q, Li Y, Jiang Z, Qiu Y, Zhang S, et al. Thyroid nodule segmentation and classification in ultrasound images through intra- and inter-task consistent learning. *Med Image Anal*. (2022) 07:102443. doi: 10.1016/j.media.2022.102443

62. Haugen BR. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: what is new and what has changed. *Cancer*. (2017) 123:372–81. doi: 10.1002/cncr.v123.3

63. Karakitsos P, Cochand-Priollet B, Pouliakis A, Guillausseau PJ, Ioakim-Liossi A. Learning vector quantizer in the investigation of thyroid lesions. *Anal Quant Cytol Histol*. (1999) 21:201–8.

64. Varlatzidou A, Pouliakis A, Stamataki M, Meristoudis C, Margari N, Peros G, et al. Cascaded learning vector quantizer neural networks for the discrimination of thyroid lesions. *Anal quantitative cytol Histol*. (2011) 33:323–34.

656. Lee YK, Ryu D, Kim S, Park J, Park SY, Ryu D, et al. Machine-learning-based diagnosis of thyroid fine-needle aspiration biopsy synergistically by Papanicolaou staining and refractive index distribution. *Sci Rep*. (2023) 13:9847. doi: 10.1038/s41598-023-36951-2

66. Elliott Range DD, Dov D, Kovalsky SZ, Henao R, Carin L, Cohen J. Application of a machine learning algorithm to predict Malignancy in thyroid cytopathology. *Cancer Cytopathol*. (2020) 128:287–95. doi: 10.1002/cncy.v128.4

67. Wang Y, Chen Z, Shima K, Zhong D, Yang L, Wang Q, et al. Rapid diagnosis of papillary thyroid carcinoma with machine learning and probe electrospray ionization mass spectrometry. *Mass Spectorm*. (2022) 57:e4831. doi: 10.1002/jms.v57.6

68. Goldgar DE, Easton DF, Cannon-Albright LA, Skolnick MH. Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. *JNCI: J Natl Cancer Institute*. (1994) 86:1600–8. doi: 10.1093/jnci/86.21.1600

69. Richard MA, Lupo PJ, Morton LM, Yasui YA, Sapkota YA, Arnold MA, et al. Genetic variation in POT1 and risk of thyroid subsequent Malignant neoplasm: A report from the Childhood Cancer Survivor Study. *PloS One*. (2020) 15:e0228887. doi: 10.1371/journal.pone.0228887

70. Chain K, Legesse T, Heath JE, Staats PN. Digital image-assisted quantitative nuclear analysis improves diagnostic accuracy of thyroid fine-needle aspiration cytology. *Cancer Cytopathol*. (2019) 127:501–13. doi: 10.1002/cncy.v127.8

71. Hirokawa M, Niioka H, Suzuki A, Abe M, Arai Y, Nagahara H, et al. Application of deep learning as an ancillary diagnostic tool for thyroid FNA cytology. *Cancer Cytopathol*. (2023) 131:217–25. doi: 10.1002/cncy.v131.4

72. Yao K, Jing X, Cheng J, Balis UGJ, Pantanowitz L, Lew M. A study of thyroid fine needle aspiration of follicular adenoma in the "atypia of undetermined significance" Bethesda category using digital image analysis. *J Pathol Inform*. (2022) 13:100004. doi: 10.1016/j.jpi.2022.100004

73. Girolami I, Marletta S, Pantanowitz L, Torresani E, Ghimenton C, Barbareschi M, et al. Impact of image analysis and artificial intelligence in thyroid pathology, with particular reference to cytological aspects. *Cytopathology*. (2020) 31:432–44. doi: 10.1111/cyt.12828

74. Zhou T, Xu L, Shi J, Zhang Y, Hu T, Xu R, et al. Ultrasonic artificial intelligence shows statistically equivalent performance for thyroid nodule diagnosis to fine needle aspiration cytopathology and BRAFV600E mutation analysis combined. *medRxiv*. (2022). doi: 10.1101/2022.04.28.22274306

75. Arunrukthavon P, Songsaeng D, Keatmanee C, Klabwong S, Ekpanyapong M, Dailey MN, et al. Diagnostic performance of artificial intelligence for interpreting thyroid cancer in ultrasound images. *Int J Knowledge Syst Sci*. (2022) 13:1–13. doi: 10.4018/IJKSS.309431

76. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst*. (2012) 25:1097–105. doi: 10.1145/3065386

77. Liu H, Hu D, Li H, Oguzl I. Medical image segmentation using deep learning. *Deep Learn Healthc*. (2019) 1:17–31. doi: 10.1049/ipr2.12419

78. Liu H, Yu X, Jiang J, Wu S, Xu C, Nai K, et al. Research on brain glioma segmentation algorithm based on multiple U-net networks. *Softw Guide*. (2024) 23:158–65. doi: 10.11907/rjdk.232176

79. Kezlarian BE, Lin O. Artificial intelligence in thyroid fine needle aspiration biopsies. *Acta Cytol*. (2022) 65:324–9. doi: 10.1159/000512097

80. Jia X, Ma Z, Kong D, Li Y, Hu H, Guan L, et al. Novel human artificial intelligence hybrid framework pinpoints thyroid nodule Malignancy and identifies overlooked second-order ultrasonographic features. *Cancers*. (2022) 14:4440. doi: 10.3390/cancers14184440

81. Jajroudi M, Kamkar. L T, Arbabi. F, Sanei M. Prediction of survival in thyroid cancer using data mining technique. *Technol Cancer Res Treat*. (2014) 13:353–9. doi: 10.7785/tcrt.2012.500384

82. Grechanuk PA, Rising ME, Palmer TS. Application of machine learning algorithms to identify problematic nuclear data. *Nucl Sci eng: J Am Nucl Soc*. (2021) 195:1265–78. doi: 10.1080/00295639.2021.1935102

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 11 April 2025
DOI 10.3389/fonc.2025.1512739

OPEN ACCESS

EDITED BY
Sandeep Kumar Mishra,
Yale University, United States

REVIEWED BY
Palash Ghosal,
Sikkim Manipal University, India
Prof. Surjeet Dalal,
Amity University Gurgaon, India
Fatma Taher,
Zayed University, United Arab Emirates

*CORRESPONDENCE
Wenna Chen
chenwenna0408@163.com
Ganqin Du
dgq99@163.com

RECEIVED 19 October 2024
ACCEPTED 21 March 2025
PUBLISHED 11 April 2025

CITATION
Chen W, Liu J, Tan X, Zhang J, Du G, Fu Q and Jiang H (2025) EnSLDe: an enhanced short-range and long-range dependent system for brain tumor classification. *Front. Oncol.* 15:1512739. doi: 10.3389/fonc.2025.1512739

EnSLDe: an enhanced short-range and long-range dependent system for brain tumor classification

Wenna Chen[1]*, Junqiang Liu[2], Xinghua Tan[2], Jincan Zhang[2], Ganqin Du[1]*, Qizhi Fu[1] and Hongwei Jiang[1]

[1]The First Affiliated Hospital, and College of Clinical Medicine of Henan University of Science and Technology, Luoyang, China, [2]College of Information Engineering, Henan University of Science and Technology, Luoyang, China

Introduction: Brain tumors pose significant harm to the functionality of the human nervous system. There are lots of models which can classify brain tumor type. However, the available methods did not pay special attention to long-range information, which limits model accuracy improvement.

Methods: To solve this problem, in this paper, an enhanced short-range and long-range dependent system for brain tumor classification, named as EnSLDe, is proposed. The EnSLDe model consists of three main modules: the Feature Extraction Module (FExM), the Feature Enhancement Module (FEnM), and the Classification Module. Firstly, the FExM is used to extract features and the multi-scale parallel subnetwork is constructed to fuse shallow and deep features. Then, the extracted features are enhanced by the FEnM. The FEnM can capture the important dependencies across a larger sequence range and retain critical information at a local scale. Finally, the fused and enhanced features are input to the classification module for brain tumor classification. The combination of these modules enables the efficient extraction of both local and global contextual information.

Results: In order to validate the model, two public data sets including glioma, meningioma, and pituitary tumor were validated, and good experimental results were obtained, demonstrating the potential of the model EnSLDe in brain tumor classification.

KEYWORDS

brain tumor classification, feature extraction, feature enhancement, long-range dependencies, attention

1 Introduction

The brain is the control center of the body, in addition to maintaining the normal activities of our lives, it also controls our daily senses (hearing, sight, smell, etc.), cognition, memory, thinking, emotions, and many other aspects of our lives (1). Undoubtedly, the brain holds paramount importance in our lives. However, brain tumors stand as one of the

most prevalent afflictions of the nervous system, capable of significantly impairing its functionality. Timely detection of brain tumors is essential for enhancing and prolonging patient survival rates (2, 3). Tumors growing within the skull are generally known as brain tumors, which encompass primary brain tumors originating from brain tissue and secondary tumors that metastasize to the skull from elsewhere in the body (4). The common types of brain tumors include gliomas, meningiomas, and pituitary tumors (5).

Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) are two widely used imaging techniques in medicine that play an important role in labelling abnormalities in the shape, size or location of the brain (6). While CT is limited to cross-sectional imaging, MRI offers the flexibility to image in various orientations, including transverse, sagittal, coronal, and any desired section. Additionally, MRI excels in providing clearer differentiation of soft tissues in three dimensions compared to conventional imaging methods. These advantages have made MRI the most favored method among physicians and have led to increasing interest among researchers. However, the analysis of MRI images by medical professionals to discern the type of tumor is a complex and time-intensive process. The accuracy of their diagnosis can be influenced by the subjective expertise and skills of the physician (7, 8). It is well known that early detection and timely treatment are crucial for the recovery of brain tumor patients (9). If the type of brain tumor can be accurately and early identified, it will greatly increase the patient's valuable treatment time and thus significantly improve the likelihood of recovery.

Traditional Machine Learning (ML) has been widely used for classification problems in Computer-Aided Diagnostic (CAD) systems (10, 11). For example, Singh et al. (12) proposed a new classification method using generalized discriminant analysis and the 1-norm linear programming extreme learning machine. Shahid et al. (13) used a feature selection algorithm to find the effective feature subset, which was then used for classification by an Extreme Learning Machine (ELM) based on hybrid particle swarm optimization. Xie et al. (14) used the combination of Support Vector Machine (SVM) and ELM for feature selection, and the optimal features were used by the classifier to distinguish breast tumor types. Heidari et al. (15) applied stochastic projection algorithm to optimize the constructed SVM model embedded with multiple feature dimensionality reduction methods to improve the classification performance of the model.

Deep learning stands as a cutting-edge innovation in classification and prediction, showcasing outstanding performance in domains necessitating multi-level data processing such as classification, detection, and speech recognition (16). Deep learning has the capability to learn features from extensive image data and extract high-level features from images through layer-by-layer convolution and pooling operations, achieving automatic classification of brain tumors. Compared to traditional image processing methods, deep learning boasts superior feature extraction capability, higher classification accuracy, as well as automation and intelligence. In recent years, many studies have explored the application of deep learning in diagnosing various diseases. For example, Sarki et al. (17) classified mild and multiple diabetic eye diseases by fine-tuning and optimizing the VGG16 model. Jeong et al. (18) used Inception V3 deep learning model to classify the presence or absence of cardiac enlargement, and the classification accuracy reached 96.0%. Chowdhury et al. (19) adopted the improved Xception model to diagnose hair and scalp diseases and achieved a high accuracy rate. Sharifrazi et al. (20) used Convolutional Neural Network (CNN) combined with k-means clustering method to automatically diagnose myocarditis, with an accuracy of 97.41%. The lesion area in brain tumor images constitutes only a small portion of the entire image. Furthermore, when distinguishing between types of brain tumors, both the tumor region and its surrounding area exert a significant impact on the classification results (21). In addition, multi-scale feature fusion has been widely applied to object detection, image segmentation, image classification, and other fields. Multi-scale networks are capable of simultaneously extracting features at different scales in images, thereby more comprehensively capturing the details and overall information of target objects. For example, in object detection tasks, small-scale features can be used to detect small objects, while large-scale features are helpful for detecting large objects. Features at different scales provide different contextual information, and multi-scale networks can effectively integrate this information, offering a more comprehensive and rich visual context. Multi-scale networks can handle input data at different scales, and this characteristic significantly enhances the algorithm's robustness and generalization performance in complex scenarios (22). A common method for multi-scale feature fusion is the pyramid structure. The pyramid structure extracts features at different scales and then fuses these features to obtain a more comprehensive feature representation. Specifically, improved methods based on the Feature Pyramid Network (FPN) architecture achieve deep integration of cross-scale features by constructing multi-level pyramid-like feature representations (23, 24).

However, most previous studies did not pay special attention to the surrounding areas of tumors, i.e., lacking the ability to capture long-range information, which would affect the performance of classification. To overcome the shortcoming, this study proposes a new multi-class brain tumor classification model with enhanced short-range and long-range dependence, named as EnSLDe. The model not only has the ability to capture short-range and long-range dependencies, but also retains local key information. It consists of three main modules: the Feature Extraction Module (FExM), the Feature Enhancement Module (FEnM), and the classification module.Within the FExM, convolutional layers are combined with residual connections to extract features, while incorporating an Effective Multi-scale Attention (EMA) mechanism that simultaneously focuses on channel-wise and spatial information. The FEnM further strengthens feature representation, enabling capture of crucial long-range dependencies while retaining key information within the local range. The classification module adopts a two-layer fully connected structure combined with dropout regularization for brain tumor classification. This approach enhances the model's generalization ability, reducing the risk of overfitting, and further improves the classification performance of the model. We utilized

two datasets to evaluate the model performance: a three-category dataset comprising gliomas, meningiomas, and pituitary tumors, and a four-category dataset including additional healthy categories.

The main contributions of this study are as follows:

- A new model with enhanced short-range and long-range dependence is proposed to classify brain tumor images from MRI.
- FExM is used to extract features from brain tumor images. The EMA module of FExM integrates channel attention and spatial attention to provide a more comprehensive feature representation.
- The FEnM is used to capture important dependencies across larger sequence scales. And it can also cooperate with the global adjustment network to fuse the retained local information with different levels of deep features.
- EnSLDe employs multi-scale parallel subnetworks that integrate shallow and deep features. This architecture enables the model to capture comprehensive contextual information across varying scales, which is critical for distinguishing between diverse tumor types.
- Based on experimental results using two public datasets, the proposed method exhibits excellent performance.

2 Related works

Classification of brain tumors is critical for evaluating tumors and determining treatment options for patients. There are already many CAD systems used in medical industries to help doctors make diagnoses. There have been many methods to classify brain tumors, which can be roughly divided into traditional ML methods, deep learning methods, and hybrid methods.

In the past, traditional ML has been used to classify brain tumors. For example, Bansal and Jindal (25) utilized a combination of grayscale co-occurrence matrix technology and shape-based feature technology to extract mixed features from the tumor area. Subsequently, a hybrid classifier consisting of Random Forest Classifier (RFC), K Nearest Neighbors (KNN) classifier, and Decision Tree (DT) classifier was used to classify brain tumors. 26 performed image segmentation through a marker-based watershed algorithm, then combined features with a sequence-based cascade method, and finally used SVM for classification.

In traditional ML, relevant domain knowledge is needed for feature extraction, while features can be automatically extracted by deep learning. The development of deep learning methods has had a significant impact on the field of medical image analysis applications, especially in disease diagnosis (27). Recently, deep learning has achieved remarkable results in brain tumor classification. For example, Raza et al. (28) proposed a hybrid deep learning model based on the GoogLeNet architecture. The last five layers of GoogLeNet were removed and 15 new layers were added to achieve high accuracy. Díaz-Pernas et al. (29) proposed a multi-scale processing based on CNN architecture design for brain tumor classification. The elastic transformation data expansion method was used to increase the training dataset and prevent over-fitting. Finally, 97.3% classification accuracy was achieved. Ayadi et al. (30) proposed an innovative brain tumor classification model based on CNN architecture, automated processing and minimizing preprocessing requirements. To fully evaluate the accuracy of the model, it was tested on three different brain tumor datasets. Various performance indicators are analyzed in depth. Sreenivasa Reddy and Sathish (31) proposed a brain tumor classification and segmentation scheme based on deep structured architecture. Firstly, adaptive ResUNet3+ with multi-scale convolution was used to process the collected data. Then, the parameters of the deep learning method were optimized and adjusted through the arithmetic optimization algorithm accelerated by the improved mathematical optimizer. Finally, an attention-based ensemble convolutional network was introduced for brain tumor classification. The model demonstrated excellent performance in both segmentation and classification accuracy. P. Ghosal et al. (32) integrated the residual network architecture with the Squeeze and Excitation block to enhance feature extraction and refinement. Islam et al. (33) optimized the EfficientNet series for the purpose of brain tumor classification, with EfficientNetB3 demonstrating superior performance. Aurna et al. (34) utilized multiple MRI datasets and performed feature extraction by combining pre-trained models and newly designed CNN models. Among the extracted features, Principal Component Analysis (PCA) was used to select key features and input them into the classifier. Musallam et al. (35) proposed a three-step preprocessing to improve the quality of MRI images and a new Deep Convolutional Neural Network (DCNN) architecture with 10 convolutional layers. Kumar and Sasikala (36) fused the features extracted from the shallow and deep layers of the pre-trained Resnet18 network, and then adopted a hybrid classifier composed of SVM, KNN, and DT optimized by the Bayesian algorithm perform classification.

In addition, in order to further improve the accuracy and efficiency of brain tumor classification models, optimization algorithms could be used in deep learning. For example, Alshayeji et al. (37) attained a classification accuracy of 97.374% for automatic brain tumor classification by combining the layers of two CNN architectures and fine-tuning the hyperparameters through Bayesian optimization. Irmak (38) used CNN and grid search optimization algorithms to propose three different CNN models to complete three different classification tasks. Almost all hyperparameters in the model were tuned by grid search optimization algorithms. Rammurthy and Mahesh (39) used Whale Harris Hawks Optimization (WHHO), which was a combination of Whale Optimization Algorithm (WOA) and Harris Hawks Optimization (HHO) to optimize the deep convolutional network. Alyami et al. (40) used deep convolutional networks and the slap swarm algorithm to classify brain tumors from brain MRI. To enhance the accuracy of classification, an efficient feature selection technique—the slap swarm algorithm was introduced. This technique helps to identify key features that significantly influence the classification results while excluding

those with minor contributions, thereby ensuring that the classification model achieves optimal accuracy.

It is noteworthy that Transformer models have also been employed in brain tumor classification tasks. Sudhakar Tummala et al. (41) investigated the capability of pretrained and fine-tuned Vision Transformer (ViT) models for brain tumor classification using MRI images. GAZI JANNATUL FERDOUS et al. (42) proposed a novel Linear Complexity Data-efficient Image Transformer (LCDEiT). The LCDEiT adopts a teacher-student strategy, where the teacher model is a customized gated pooling convolutional neural network (CNN) responsible for transferring knowledge to the transformer-based student model. The student model achieves linear computational complexity through an external attention mechanism. Asiri et al. (43) employed Swin Transformer for multi-class brain tumor classification. Tapas Kumar Dutta et al. (44) developed GT-Net for brain tumor classification tasks. The core component of this model is the Global Transformation Module (GTM), which contains multiple Generalized Self-Attention Blocks (GSB) designed to explore long-range global feature relationships between lesion regions.

These studies, whether based on traditional ML methods, deep learning approaches, or hybrid methodologies, have achieved notable success in brain tumor classification. Many deep learning models (e.g., CNNs) automatically extract features but typically focus on local or global information rather than both. For instance, architectures like Inception-v3, ResNet, and DenseNet demonstrate strong performance yet generally emphasize localized details or global context without comprehensive integration. Hybrid approaches combining traditional machine learning and deep learning techniques may still fail to fully exploit multi-scale feature fusion or advanced attention mechanisms. While some models employ attention mechanisms, they often prioritize either channel-wise or spatial attention. This paper proposes a novel model named EnSLDe (Enhanced Short- and Long-range Dependency Extractor), designed to strengthen both short-term and long-range dependencies while preserving essential local information. EnSLDe uniquely integrates short- and long-range dependencies through its FExM and FEnM. This dual processing proves critical for concurrently capturing localized tumor details and global contextual patterns in brain MRI images.

3 Proposed method

This section introduces our proposed brain tumor classification framework, which is shown in Figure 1. The training and testing phase of the proposed system works as follows:

1. The brain MRI dataset is divided into two disjoint sets: a training set and a test set.
2. Data augmentation techniques such as random rotation, random horizontal and vertical flipping are applied to the training dataset to mitigate overfitting issues.
3. The proposed network is trained by selecting appropriate hyperparameters and specifying the cross-entropy loss function.
4. Once training is completed, the trained model is saved.
5. The model is validated on a randomly partitioned test dataset, and the performance of the model is evaluated.

3.1 Proposed brain tumor classification model

The EnSLDe consists of three main modules, namely feature extraction module, feature enhancement module and classification module, which is shown in Figure 2. Since both local and long-range dependent features play a crucial role in effectively classifying brain tumors from MRI images, the EnSLDe employs FExM and FEnM to extract and enhance these features. The classification module comprises two fully connected layers integrated with Dropout regularization, which enhances the model's generalization ability. Moreover, the stacked utilization of two fully connected layers can

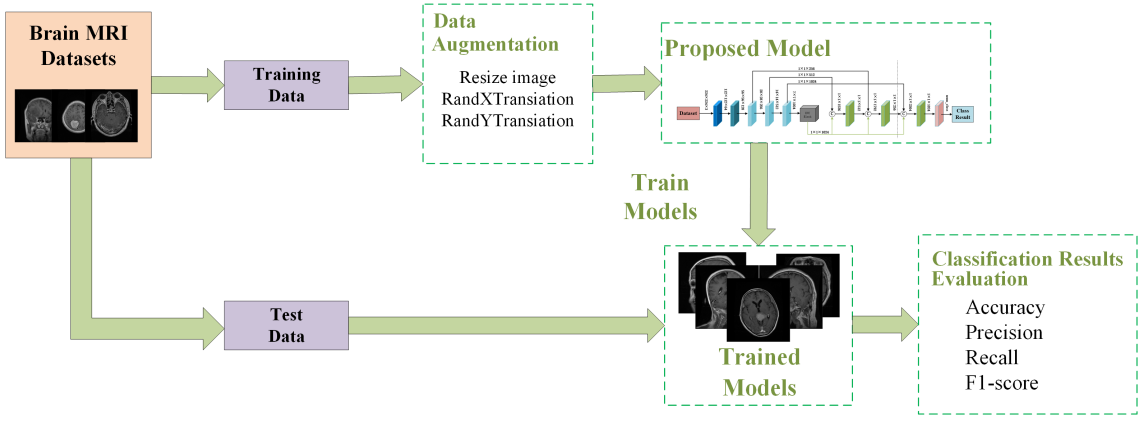


FIGURE 1
The proposed framework for brain tumor classification system.

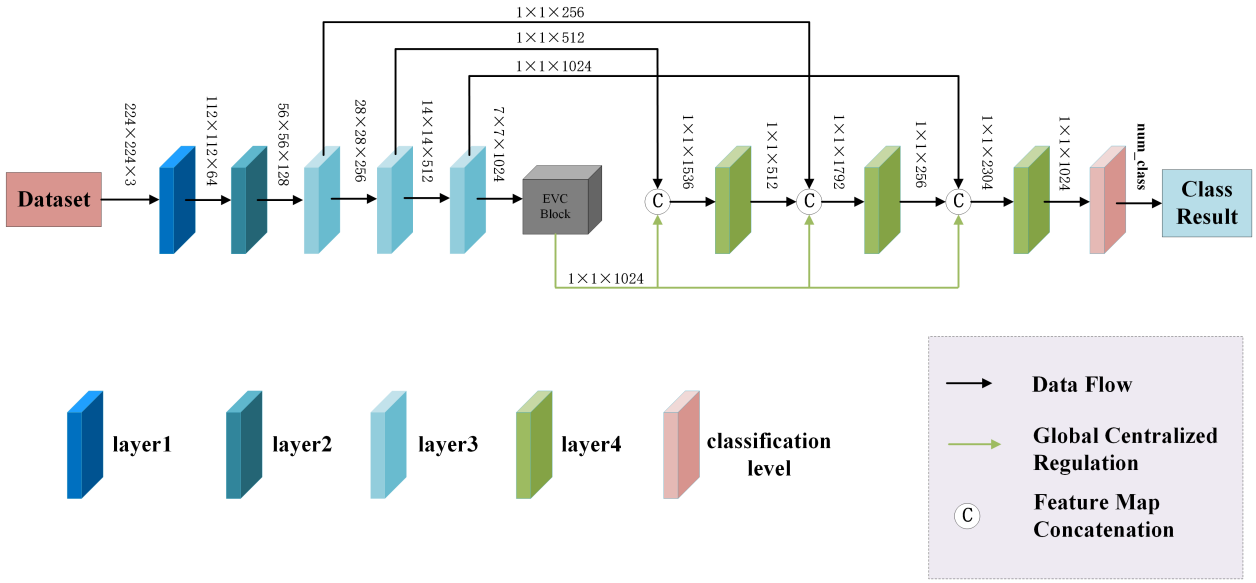


FIGURE 2
The proposed model.

amalgamate and transform features, thereby capturing more information and optimizing the representation capabilities of features to enhance model performance.

3.1.1 The feature extraction module

The feature extraction module consists of layer1, layer2, layer3-1, layer3-2, layer3-3, layer4-1, layer4-2, and layer4-3, and is used to extract multiple depth-level features from brain tumor images. The Feature Extraction Module (FExM) was designed to extract features from multiple intermediate layers to simultaneously capture short-range and long-range dependencies. This multi-scale parallel sub-network fuses shallow features (which retain fine-grained details) with deep features (encoding abstract, high-level contextual information). The selection of feature extraction layers was guided by empirical validation through ablation studies, which demonstrated that combining multiple layers achieved higher classification accuracy compared to those obtained using a single layer of features. Inspired by the C3 module in YOLOv5 and integrating the Effective Multi-scale Attention (EMA) proposed by (Ouyang et al. (45), we have developed a novel Conv and Depthwise_conv with EMA (CDE) module, as illustrated in Figure 3. The CDE module consists of a residual network and EMA. The structure of the residual network involves adding skip connections on top of the serial connection of two convolutional layers and a depthwise separable convolutional layer. This allows for the direct addition of input and output. Subsequently, the output features of the entire residual network are processed by EMA. Incorporating the residual network into the CDE module effectively alleviates the issues of gradient explosion or vanishing, making the model training process more stable and easier to optimize.

Additionally, depthwise separable convolution is used by CDE module, which significantly reduces computational costs while maintaining powerful feature extraction capabilities, thus achieving a good balance between efficiency and performance. The inclusion of EMA allows the CDE module to form multi-scale parallel subnetwork while extracting features, which fuses shallow and deep features. This further enhances feature extraction and strengthens short-range and long-range dependencies. Moreover, it reshapes part of the channel dimensions into batch dimensions, effectively avoiding potential information loss caused by dimensionality reduction through conventional convolution. This improvement not only reduces computational overhead but also allows the model to focus more on extracting key features while retaining information from each channel. Layer1 consists of two convolutional layers and is mainly used to extract shallow image features. Layer2 consists of the residual network in the CDE module. layer3-1, layer3-2, and layer3-3 are all composed of CDE modules. Layer4-1, layer4-2, and layer4-3 are all composed of convolutional layers with a convolution kernel size of 1×1, which are used for channel dimensionality reduction after feature fusion.

The EMA divides the channel dimension of input feature maps into multiple sub-features and redistributes spatial-semantic features within each feature group. Specifically, EMA avoids traditional channel dimensionality reduction operations by reshaping the channel dimension into the batch dimension. This design enables EMA to model inter-channel dependencies through standard convolution operations without losing channel information. The EMA employs three parallel branches to extract attention weights:

1. 1×1 Branch: Encodes channel attention along horizontal and vertical directions using two 1D global average pooling operations, thereby capturing long-range spatial dependencies while preserving precise positional information.

2. 3×3 Branch: Captures multi-scale feature representations through a 3×3 convolution kernel to expand the feature space.
3. Cross-Space Interaction: Fuses output feature maps from the two parallel branches via matrix dot product operations to capture pixel-level pairwise relationships and highlight global contextual information.

For an input feature$X \in \mathbb{R}^{C\times H\times W}$, it is first partitioned into G sub-features, each with a shape of (C/G) × H×W. In the 1×1 branch, two 1D feature vectors Z_H and Z_W are obtained by encoding channel attention through 1D global average pooling along horizontal and vertical directions, respectively. Z_H and Z_W can be calculated by Equation 1:

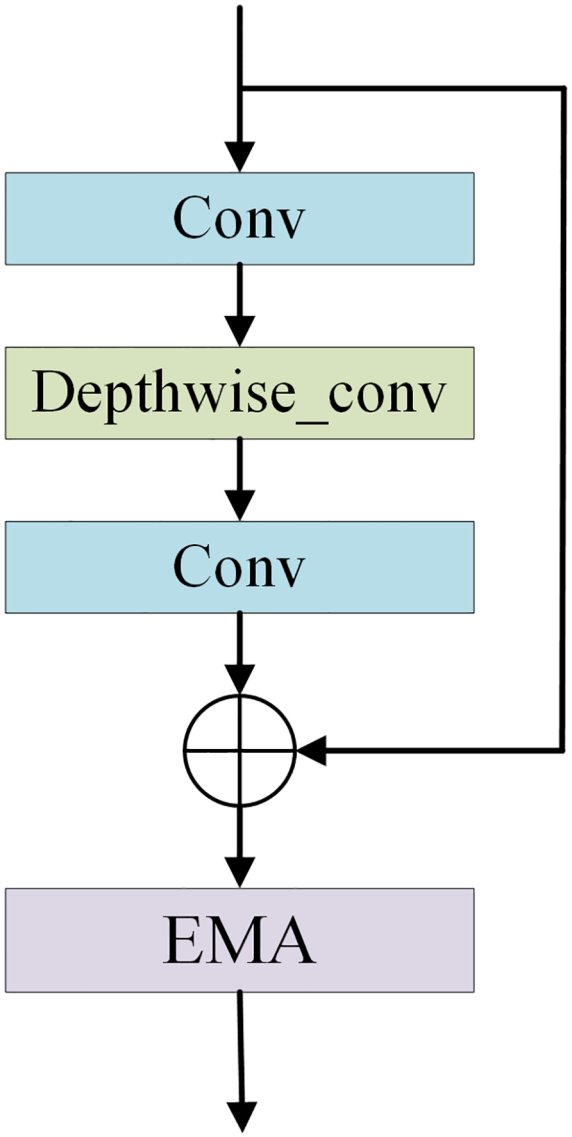


FIGURE 3
The structural diagram of the CDE module.

$$Z_H = \sum_{j=1}^{H} \chi_{c,j} \tag{1}$$

$$Z_W = \sum_{j=1}^{H} \chi_{c,j}$$

where, $x_{c,i}$ and $x_{c,j}$ denote the eigenvalues of the c channel in the horizontal and vertical directions, respectively. The vectors Z_H and Z_W are processed through 1×1 convolutions and the Sigmoid function to generate the channel attention maps A_H and A_W, can be calculated by Equation 2:

$$A_H = \sigma(conv(Z_W)) \tag{2}$$

$$A_W = \sigma(conv(Z_W))$$

Where, σ denotes the Sigmoid function. In the 3×3 branch, multi-scale feature representation $F_{3\times3}$ is captured by the 3×3 convolution operation as shown in Equation 3:

$$F_{3\times3} = Conv_{3\times3}(X) \tag{3}$$

The final output feature map Y is obtained by fusing A_H and A_W matrix dot product is performed by $F_{3\times3}$, and the calculation formula is shown in Equation 4:

$$Y = \sigma(A_H \cdot A_W \cdot F_{3\times3}) \tag{4}$$

3.1.2 The feature enhancement module

The Explicit Visual Center (EVC) method (46) is used to enhance the features extracted by the model. The EVC can effectively extract global long-range dependencies from images while preserving crucial local information. The EVC combines a Multi-Layer Perceptron (MLP) based on top-level features with a Learnable Visual Center (LVC) mechanism, both of which operate in parallel to complement each other. The MLP is responsible for capturing the global long-range dependencies of the image, effectively addressing complex long-range dependency issues, and enhancing the model's perception of global information. Meanwhile, the LVC operates along the path of the MLP, focusing on preserving the crucial local information of the image to ensure that the model does not lose important local details while attending to the global context. For input F_{in}, the equation is calculated as follows (Equation 5):

$$F = Cat(MLP(F_{in}), LVC(F_{in})) \tag{5}$$

in the LVC model, the input (X) is mapped to a set of (C)-dimensional features, ($\{X_{in} = x_1, x_2, \ldots, x_n\}$), where (N=H×W) represents the total number of input features. Subsequently, LVC computes an intrinsic codebook (B = {b1, b2, …, b_k}), which includes (K) codewords (or visual centers) along with a set of smoothing factors (S = {s_1, s_2, …, s_k}). The feature encoding is achieved through a series of convolutional layers. The encoded features are then matched against each codeword in the codebook. The discrepancies between the features and the codewords are computed, and learnable weights are derived from these

differences. The ultimate output is a (C)-dimensional vector (e) (Equation 6).

$$e_k = \sum_{i=1}^{n} \frac{e^{-S_k\|x_i - b_k\|^2}}{\sum_{j=1}^{K} e^{-S_k\|x_i - b_k\|^2}} (x_i - b_k) \tag{6}$$

The output of LVC is obtained by summing the features vector (X_{in}) and the local features (Z) for each channel, as shown in Equation 7.

$$X_{out} = X_{in} \oplus Z \tag{7}$$

here, the local feature (Z) is derived by applying a Fully Connected (FC) layer that maps the feature (e) to an influence factor of dimensions C×1×1. Subsequently, a channel-wise multiplication operation is conducted with (X_{in}). The output following the Feature Enhancement Module is then obtained as follows (Equation 8):

$$F = Cat(X_{EVC}, X_d) \tag{8}$$

where, F represents the fusion feature, X_{EVC} denotes the feature output from the EVC, and X_d signifies the depth feature derived from various levels.

3.2 Loss function

The loss function we used during model training is the cross-entropy loss function (47). One can assume there are n classes, where the true label is represented by a K-dimensional vector y (with only one element being 1 and others being 0), and the model output probability is represented by a K-dimensional vector y' (with each element ranging from 0 to 1 and summing up to 1). The formula for multi-class cross-entropy loss function is defined as shown in Equation 9.

$$Loss = -\sum_{i=1}^{n} y_i \log y_i' \tag{9}$$

where, n is the number of categories, y_i is the i-th element of the true label vector y, and y_i' is the i-th element of the model output probability vector y_i.

The cross-entropy loss function is an efficient loss function in classification problems as it accurately measures the similarity between the true label distribution and the model's predicted label distribution. Specifically, a smaller cross-entropy value indicates a closer resemblance between these two probability distributions, implying more accurate predictions by the model. When there is a significant disparity between the true and predicted distributions, the cross-entropy loss function yields a large loss value. This characteristic enables the model to update parameters more quickly during training, thus accelerating the learning process. The amplifying effect of the cross-entropy loss function makes the model more sensitive to prediction errors during training, facilitating more effective adjustment of model parameters and reducing the likelihood of erroneous predictions. Therefore, the cross-entropy loss function is well-suited as a loss function for classification models, particularly excelling in handling multi-class classification problems.

4 Results and discussion

This study was conducted on a computer equipped with RTX3080 graphics card of 10 GB video memory and 64 GB of RAM.

4.1 Brain tumor dataset and preprocessing

In this paper, two publicly available brain tumor MRI datasets are applied for the brain tumor multi-classification task. Details of these two datasets are provided in Table 1. Both Cheng dataset and BT-large-4c dataset contain different views of brain anatomy: axial, coronal and sagittal views. Additionally, both datasets contain different numbers of brain tumor categories obtained from different patients with differences in tumor grade, race, and age. The Cheng dataset contains 3 types of brain tumors, namely glioma, meningioma and pituitary tumor. Among them, there are 1426 glioma images, 708 meningioma images and 930 pituitary tumor images, for a total of 3064 grayscale brain Magnetic Resonance (MR) images (48). The BT-large-4c dataset consists of 3264 brain MR images, including 926 glioma, 940 meningioma and 901 pituitary tumor images, and the remaining 497 normal images (49). These two datasets are split into 80% for training and 20% for testing.

During the dataset preprocessing phase, we implemented an efficient and streamlined data preprocessing protocol. To ensure

TABLE 1 Details of the datasets used in this study.

NO.	Dataset name	Classes	Number of Each class	Total number of images
1	Cheng	Glioma	1426	3064
		Meningioma	708	
		Pituitary	930	
2	BT-large-4c	Glioma	926	3264
		Meningioma	940	
		Pituitary	901	
		No tumor	497	

image content integrity and feature stability in experimental settings, all images were uniformly resized to dimensions of 224×224×3 pixels. This standardized resizing not only preserves the spatial structure and informational completeness of images but also significantly reduces computational overhead during network training, thereby enhancing training efficiency. Additionally, a standardization procedure was applied—a conventional preprocessing technique in deep learning—to mitigate variations in illumination, contrast, and other attributes across images, enabling the model to focus on learning intrinsic features. Considering that deep neural networks typically require large-scale datasets for training while our study employed a relatively limited dataset, data augmentation strategies were systematically deployed to alleviate overfitting. Specifically, techniques including random rotation, cropping, and horizontal flipping were implemented. These operations effectively enhanced dataset diversity without introducing additional noise, thereby strengthening the model's generalization capabilities.

4.2 System implementation and evaluation metrics

During the model training process, we will fine-tune hyperparameters such as batch size, optimizer type, learning rate, epochs, and loss function based on experience and actual requirements. The objective of this process is to identify the optimal combination of hyperparameters to enhance the model's performance and achieve the desired training outcomes. In this model, we employ the Adam optimizer with an initial learning rate of 0.001, 150 epochs, and a mini-batch size of 16 samples.

In this study, the performance of the proposed method is given by accuracy, recall, precision, and F1 -score (Cohen's) were used for evaluation Kappa(κ), Matthews Correlation Coefficient (MCC) are given by this is given by Equations 10–15 (50):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{10}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

$$\text{F1} - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{14}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{15}$$

where, True Positives (TP) are the number of actual and predicted positives. True Negatives (TN) are the number of negatives that are both actual and predicted. False Positives (FP) are the number of actual negatives that are predicted to be positive. False Negatives (FN) are the number of actual positives that are predicted to be negative. p_o is the proportion of inter-observers who actually agree. p_e is the proportion of agreement expected based on a random assignment.

4.3 Experimental results

The proposed method is applied to the Cheng dataset and the BT-large-4c dataset for classification, and the corresponding confusion matrix is generated, as shown in Figures 4A, B. In these matrices, the label "G" represents glioma, "M" represents meningioma, "P" represents pituitary tumor, and "N" represents no tumor. The confusion matrices vividly illustrate the classification performance of the model for each category. Additionally, the detailed values of model metrics obtained on the Cheng and BT-large-4c datasets are shown in Table 2. These metrics offer a quantitative basis for comparison, facilitating the evaluation of the model's performance and comparison with other methods. It is noteworthy that on the Cheng dataset, our model demonstrated

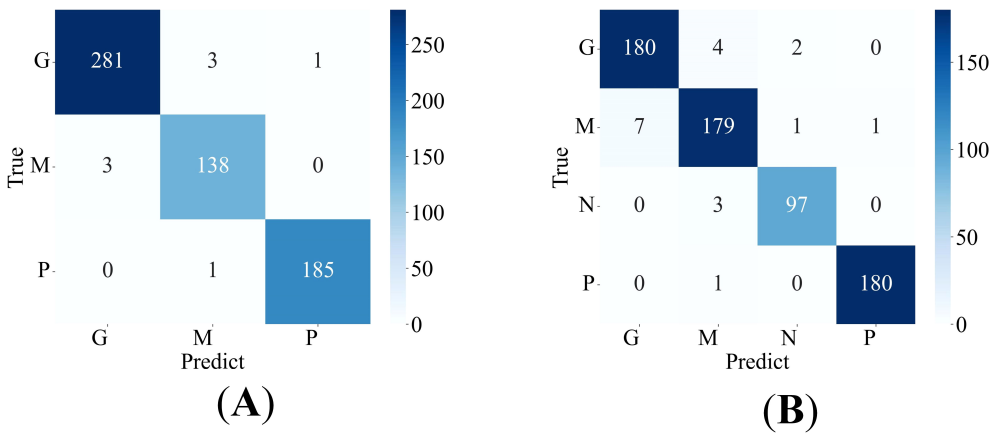


FIGURE 4
Confusion matrix of the proposed model **(A)** on the Cheng dataset, **(B)** on the BT-large-4c dataset.

TABLE 2 Detailed metric values of the proposed model on Cheng and BT-large-4c datasets.

Dataset	Tumor type	Precision	Recall	F1-score	Accuracy	κ	Mcc
Cheng	Glioma	0.9894	0.9860	0.9877	0.9869	0.9795	0.9795
	Meningioma	0.9718	0.9787	0.9753			
	Pituitary	0.9946	0.9946	0.9946			
	Average	**0.9853**	**0.9864**	**0.9859**			
BT-large-4c	Glioma	0.9626	0.9677	0.9651	0.9710	0.9607	0.9607
	Meningioma	0.9572	0.9521	0.9547			
	No tumor	0.9700	0.9700	0.9700			
	Pituitary	0.9945	0.9945	0.9945			
	Average	**0.9711**	**0.9711**	**0.9711**			

exceptionally high classification performance, achieving an accuracy of 98.69%. Similarly, on the BT-large-4c dataset, the model achieved a classification accuracy of 97.10%. The total number of parameters in the EnSLDe model is 87 million (87M). The total memory size required for the model during operation (including training and inference) is 2792.73MB. The memory size required for one forward and backward propagation process in the model is 2459.25MB.

The Receiver Operating Characteristic (ROC) curve is a graphical tool used to represent the performance of a classification model. It effectively evaluates the performance of the model under different classification thresholds by taking the False Positive Rate (FPR) and True Positive Rate (TPR) as the horizontal and vertical coordinates. The Area Under the Curve (AUC) quantitatively assesses the quality of the classification model. Higher AUC values indicate better model performance, with values closer to 1 indicating more ideal classification performance. Specifically, the ROC curves of our proposed model on the Cheng dataset and BT-large-4c dataset are depicted in Figures 5A, B, respectively. On the Cheng dataset, the AUC values for glioma, meningioma, and pituitary tumor in our proposed model are 0.9982, 0.9991, and 1.0000, respectively. On the BT-large-4c dataset, the AUC values for glioma, meningioma, pituitary tumor, and no tumor in our proposed model are 0.9941, 0.9921, 0.9999, and 0.9967, respectively. These results indicate that our proposed model exhibits excellent classification performance on both the Cheng dataset and BT-large-4c dataset.

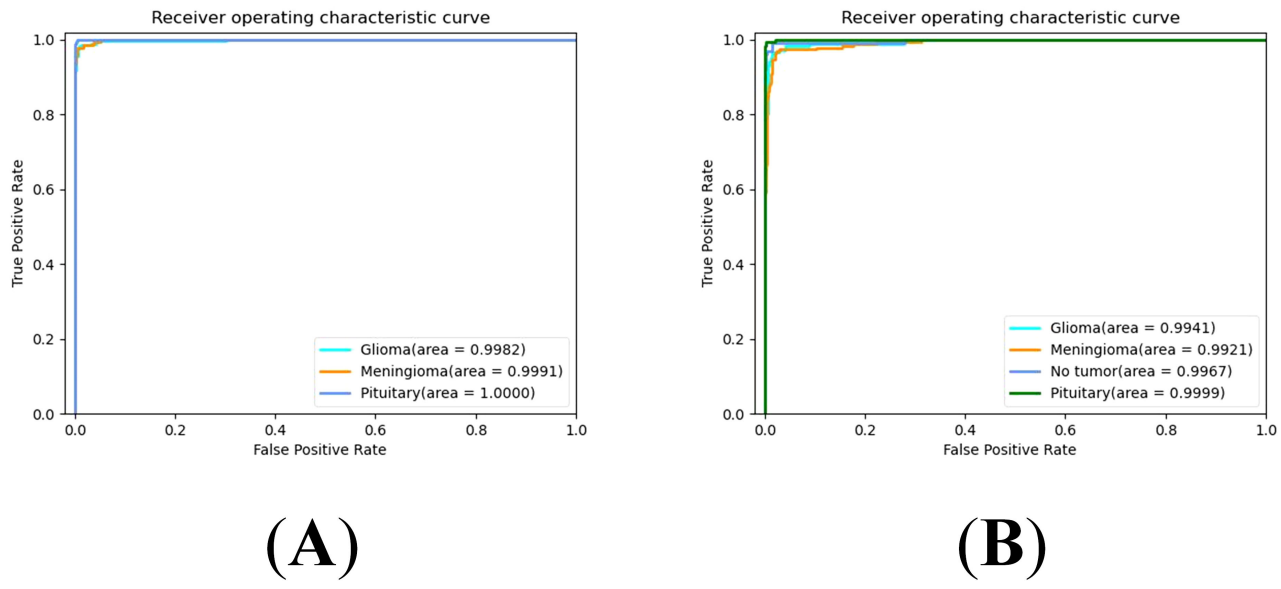


FIGURE 5
ROC curve for EnSLDe **(A)** on the Cheng dataset, **(B)** on the BT-large-4c dataset.

4.4 Ablation experiment

This ablation experiment aims to comprehensively evaluate the impact of attention module, FEnM and data enhancement on model performance. The following three subsections will demonstrate in detail the contribution and importance of these three key components to model performance.

4.4.1 The impact of the attention module on the model

In this section, the influence of various attention modules on our proposed model is investigated. The new models reconstructed from these attention modules and our proposed model include: Squeeze-and-Excitation(SE) (51) instead of EMA in EnSLDe named as EnSLDe-SE, Coordinate Attention (CA) (52) instead of EMA in EnSLDe named as EnSLDe-CA, Convolutional Block Attention Module (CBAM) (53) instead of EMA in EnSLDe named as EnSLDe-CBAM and the one removing EMA from EnSLDe named as EnSLDe-NoEMA. These models are used for classification prediction on the Cheng dataset, and the results are shown in Figure 6.

From Figure 6, it is evident that the EnSLDe-SE does not perform well in these models, with an accuracy of only 96.41%. Conversely, the EnSLDe exhibits exceptional performance in these models, achieving an accuracy of 98.69% and demonstrating excellent performance across other evaluation metrics. Specifically, the EnSLDe attains 98.53%, 98.64%, and 98.59% in precision, recall, and F1-score parameters, respectively. Moreover, when the EMA module is removed, the model's accuracy significantly drops to 97.06%. This comparison underscores the crucial role of the EMA module in enhancing the performance of the proposed model. The inclusion of the EMA module not only boosts the classification accuracy of the model but also achieves balanced optimization across multiple evaluation metrics, thereby enabling the model to maintain high performance levels.

4.4.2 The impact of the FExM on the model

FExM is the cornerstone of the EnSLDe architecture, designed to hierarchically extract multi-scale contextual features through the combination of convolutional layers, residual connections, and the EMA mechanism. To rigorously evaluate its contribution, we conducted a comparative analysis of the model's performance with and without the FExM module. When the FExM was not used, the model's performance metrics—Precision, Recall, F1-score, and Accuracy—were 0.9656, 0.9722, 0.9683, and 0.9706, respectively, which were consistently lower than those of the model with FExM. It is worth noting that the precision dropped by 1.63%, highlighting the crucial importance of FExM to the overall model performance. Furthermore, in the ablation study, the p-value for the paired t-test of accuracy was 0.0013 (below the significance level, $\alpha = 0.05$), with a confidence interval ranging from [0.0442, 0.1815].

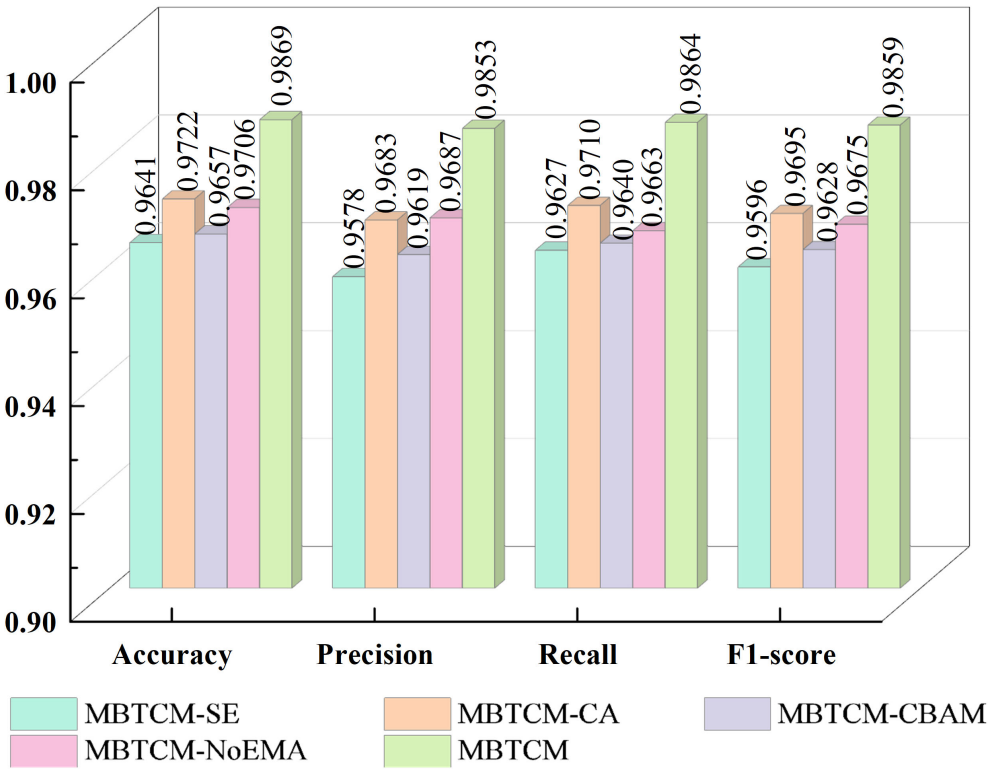


FIGURE 6
Impact of each attention module.

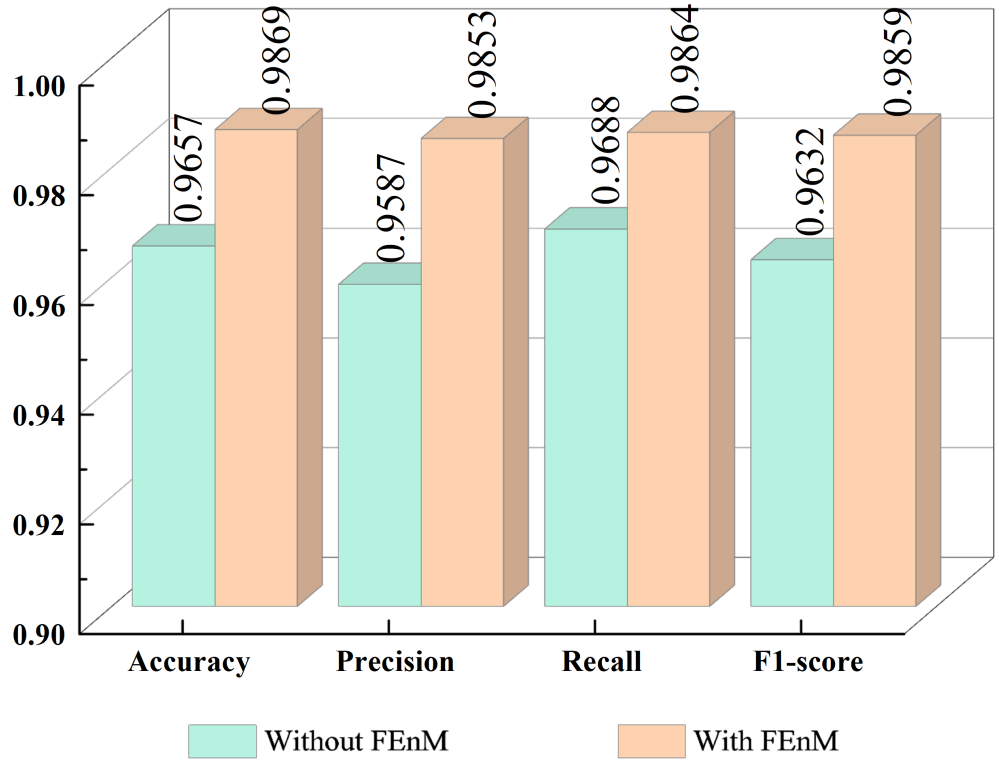


FIGURE 7
Impact of FEnM.

4.4.3 The impact of the FEnM on the model

This section primarily examines the impact of the FEnM on the proposed model, with specific results depicted in Figure 7. The figure clearly illustrates that introducing FEnM significantly enhances the classification performance of the model on the Cheng dataset. Specifically, the accuracy, precision, recall, and F1-score of the model have increased by 2.12%, 2.66%, 1.76%, and 2.27%, respectively. The p-value of the paired t-test for accuracy with and without FEnM was 0.0094 (which is below the significance level, $\alpha = 0.05$), and the confidence interval range was [0.0228, 0.1595]. The notable performance improvement can be attributed to the effective role of the FEnM. The FEnM not only substantially enhances the extracted features but also excels in capturing important long-range dependencies. Moreover, the FEnM can integrate the retained local key information with different levels of deep features, thereby enriching the expressive capabilities of features. Through this feature enhancement method, the model can more accurately identify brain tumors in classification tasks.

4.4.4 The impact of data augmentation on models

This experiment utilizes two datasets: the Cheng dataset and the BT-large-4c dataset. Through the application of data augmentation techniques, the classification performance of the proposed model on these datasets is significantly enhanced. The impact of data augmentation on the model is illustrated in Figure 8. Specifically, for the Cheng dataset, the accuracy is improved by 3.92%, and for the BT-large-4c dataset, the accuracy is improved by 3.51%. These results highlight the crucial role of data augmentation techniques in enhancing model performance. In particular, by incorporating data augmentation with random horizontal or vertical flipping of images, the model becomes adept at learning tumor characteristics from various orientations and locations. This implies that the model can effectively identify and classify tumors even when their orientation or location varies in real-world applications.

4.4.5 Ablation studies on layer selection

To further validate the selection of feature extraction layers, we conducted an ablation study, the results of which are summarized in Table 3. When features were extracted from a single layer (shallow or deep), classification accuracy was consistently lower than that achieved via a multilayer fusion approach. To assess whether the observed differences in performance were statistically significant, paired t-tests were conducted. The tests compared classification accuracies of deep layers (which demonstrated superior performance to shallow layers) and multilayer fusion, positing the null hypothesis that there was no significant difference in performance. The paired t-test produced a p-value of 0.03 (below the significance level, $\alpha = 0.05$), indicating a statistically significant difference in performance. By combining features from shallow and deep layers, the model captured a more holistic representation of the input data. The confidence interval for the difference in accuracy (which ranged from [-0.013, -0.0019]) excluded zero, confirming

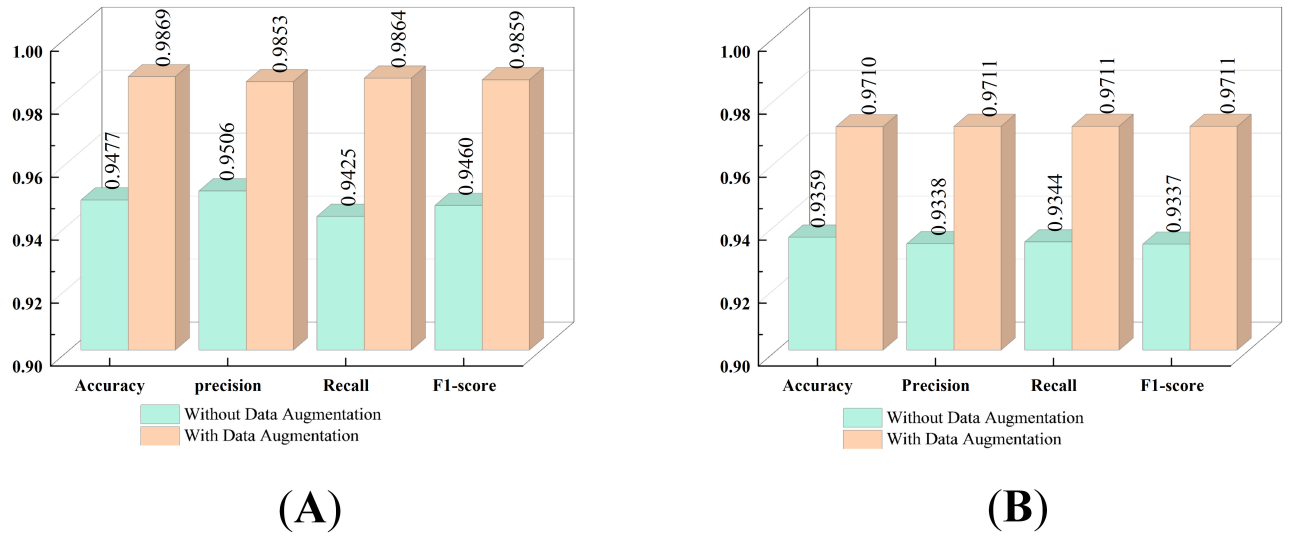


FIGURE 8
Impact of Data Augmentation **(A)** on the Cheng dataset, **(B)** on the BT-large-4c dataset.

that the multilayer fusion approach surpassed single-layer extraction. The shallow layer provided detailed local information, whereas the deep layer captured global contextual features. This combination enhanced the model's ability to discern complex patterns in brain tumor images.

4.4.6 Impact of hyperparameter selection on model performance

Hyperparameters are an important aspect that affects model performance, and different hyperparameters can lead to different experimental results. In this section, the impact of the hyperparameters batch size, lr, and optimizer on model performance will be verified. Table 4 presents the experimental results. By comparing Tables 2, 4, it can be found that the hyperparameter values selected in this paper are quite good.

4.5 Cross-dataset validation

To comprehensively validate the model, cross-validation was employed. The BT-large-4c dataset, comprising glioma, pituitary tumor, and meningioma data, was used to evaluate the model trained on the Cheng dataset. The cross-validation results for accuracy, precision, recall, and F1-score were 92.98%, 93.2%, 93.02%, and 93.01%, respectively. These outcomes indicate that the proposed model exhibits significant robustness.

4.6 Discussion

To further quantify the performance of the proposed model. The classification results obtained by our proposed model are compared with those obtained by previous state-of-the-art models using the same dataset, as shown in Table 5. Noreen et al. (54) proposed a method integrating deep learning with machine learning models, employing deep learning for feature extraction, including the Inception-v3 and Xception models. Additionally, the classification of brain tumors through deep learning and machine learning algorithms such as softmax, RF, SVM, KNN, and ensemble techniques were explored. Bodapati et al. (55) developed a dual-channel deep neural network architecture for brain tumor classification using pre-trained InceptionResNetV2 and Xception models, incorporating attention mechanisms to enhance accuracy and generalization capabilities in brain tumor recognition. Shaik and Cherukuri (56) designed and implemented a multi-level attention network (MANet). The proposed MANet includes spatial and channel-wise attention mechanisms, prioritizing tumor regions while maintaining the inter-channel temporal dependencies in the semantic feature sequences obtained from the abnormal areas. Öksüz et al. (57) utilized pre-trained AlexNet, ResNet-18, GoogLeNet, and ShuffleNet networks to extract deep features from images, and designed a shallow network for extracting shallow features, fusing these features and classifying them with SVM and KNN. Jaspin and Selvan (58) proposed a multi-class

TABLE 3 Layer selection of experimental results in dataset Chen.

Method	Precision	Recall	F1-score	Accuracy
Shallow layer	0.9554	0.9539	0.9546	0.9592
Deep layer	0.9683	0.9645	0.9664	0.9690
Multilayer fusion (ours)	0.9853	0.9864	0.9859	0.9869

TABLE 4 Experimental results for different hyper-paramete.

Hyper-parameter	Value	Precision	Recall	F1-score	Accuracy
Batch	8	0.9739	0.9770	0.9754	0.9771
Lr	0.0001	0.9835	0.9811	0.9823	0.9837
Optimizer	SGD	0.9800	0.9757	0.9778	0.9804

convolutional neural network (MCCNN) model for identifying tumors in brain MRI images. This network, consisting of an 11-layer structure including three convolutional layers, three max-pooling layers, one flattening layer followed by three dense layers, and an output layer, achieved classification performance on par with pre-trained models. Md. S. I. Khan et al. (59) designed a 23-layer convolutional neural network for brain tumor classification. Satyanarayana et al. (60) introduced a density convolutional neural network model based on mass correlation mapping (DCNN-MCM) for brain tumor classification. This model leverages the average mass elimination algorithm (AMEA) and mass correlation analysis (MCA) for the extraction and training of significant features of brain tumors, using a CNN model for efficient classification. Kibriya et al. (61) developed a 13-layer CNN specifically for brain tumor classification. Dutta et al. (62) introduced an attention-based residual multi-scale CNN, termed ARM-Net. This model includes a lightweight residual multi-scale CNN architecture known as RM-Net and introduces a lightweight global attention module (LGAM) to selectively learn more discriminative features. S. U. R. Khan et al. (63) employed the DenseNet169 model for feature extraction and fed the extracted features into three multi-class machine learning classifiers: RF, SVM, and gradient-boosting decision trees (XGBoost). Brain tumor classification was performed through the integration of these classifiers using a majority voting strategy. Demir and Akbulut (64) used a new multi-level feature selection algorithm to select the 100 deep features with the highest significance and adopted the SVM algorithm with Gaussian kernel for classification and achieved better performance. Senan et al. (65) employed both AlexNet and ResNet18 in conjunction with SVM for brain tumor classification and diagnosis. Initially, deep learning techniques were used to extract robust and significant deep features through deep convolutional layers, followed by classification using SVM. Ravinder et al. (66) proposed a graph convolutional neural network (GCN) model. This model integrates graph neural networks (GNN) with traditional CNNs. Our EnSLDe achieves superior performance compared to other methods. This depends on its ability to enhance short-range and long-range dependencies. EnSLDe yields experimental results for the Chen dataset. On the BT-large-4c dataset, EnSLDe underperforms AlexNet+SVM by a margin of 0.0139 in terms of precision. Nonetheless, it excels in other performance indicators. The EnSLDe model demonstrates exceptional performance on the Cheng and BT-large-4c datasets, achieving high accuracy rates of 98.69% and 97.10%, respectively. These results highlight the

TABLE 5 Comparison of our proposed model with previous models.

Reference	Dataset	Method	Precision	Recall	F1-score	Accuracy
Noreen et al. (54)	Cheng	Inception-v3+Ensemble	–	–	–	0.9434
Bodapati et al. (55)		Two-Channel DNN	–	–	0.9779	0.9523
Shaik and Cherukuri (56)		MANet	0.9614	0.9599	0.9603	0.9651
Öksüz et al. (57)		ResNet18+ShallowNet+SVM	0.9525	0.9527	0.9526	0.9725
Jaspin and Selvan (58)		MCCNN	0.95	0.95	0.96	0.9517
Md. S. I. Khan et al. (59)		23-layer CNN	0.965	0.964	0.964	0.978
Satyanarayana et al. (60)		DCNN-MCN	–	–	–	0.94
Kibriya et al. (61)		13-layer CNN	0.97	0.96	0.965	0.972
Dutta et al. (62)		ARM-Net	0.9646	0.9609	0.9620	0.9664
S. U. R. Khan et al. (63)		Hybrid-NET	0.95	0.94	0.94	0.951
Dutta et al. (44)		GT–Net	–	–	96.39	97.11
The Proposed Method		EnSLDe	0.9853	0.9864	0.9859	0.9869
Demir and Akbulut (64)	BT-large-4c	R-CNN+SVM	0.964	0.9645	0.964	0.966
Senan et al. (65)		AlexNet+SVM	0.985	–	–	0.951
Ravinder et al. (66)		GCNN	0.9525	0.965	0.9587	0.9501
The Proposed Method		EnSLDe	0.9711	0.9711	0.9711	0.971

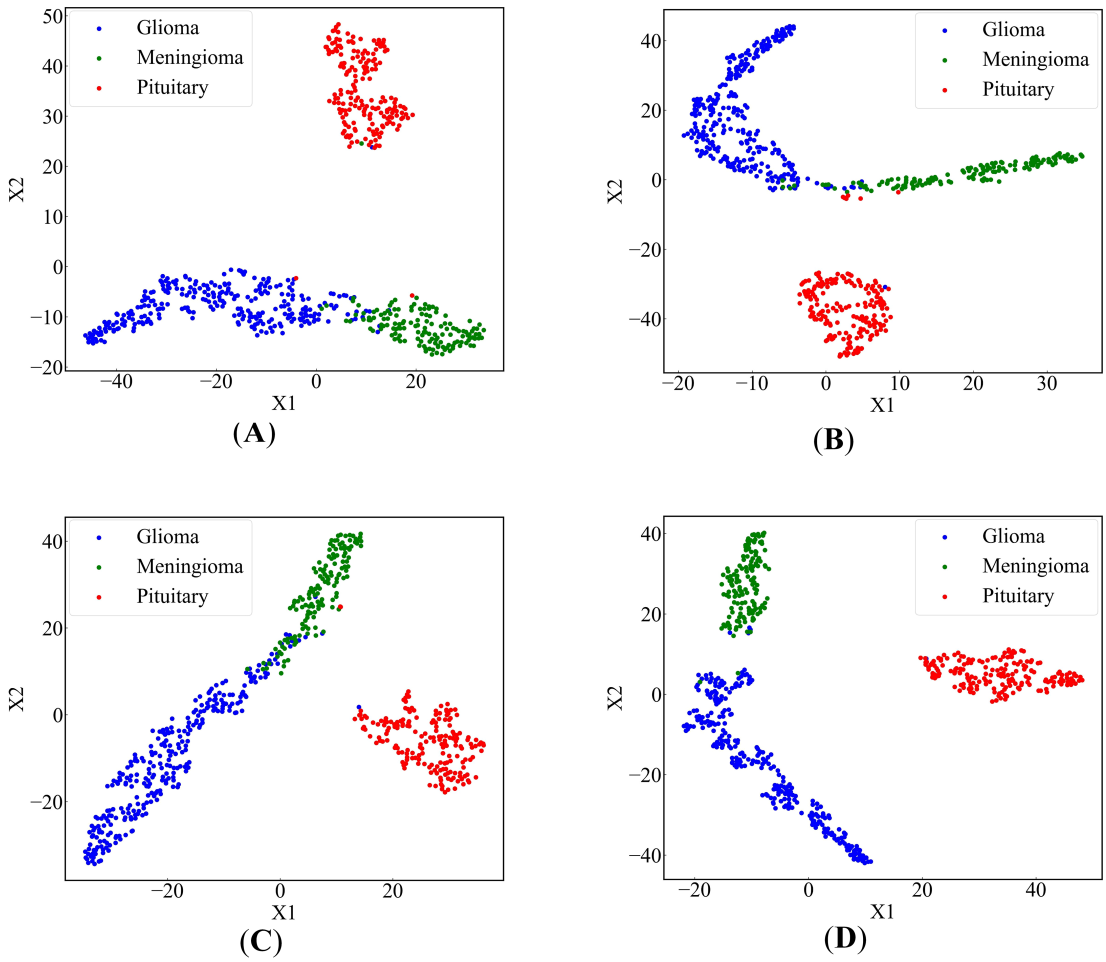


FIGURE 9
2-dimensional scatter plots of deep feature sets **(A)** EnSLDe without FEnM, **(B)** EnSLDe without EMA, **(C)** EnSLDe without Data Augmentation, **(D)** EnSLDe.

model's ability to effectively capture both short-range and long-range dependencies in brain tumor images, leading to improved classification accuracy. And multi-scale parallel subnetworks fuse shallow and deep features to capture comprehensive information. However, it is important to note that the performance of any model, including EnSLDe, can vary depending on the specific characteristics of the data it is applied to. While EnSLDe outperforms several state-of-the-art models on these datasets, its generalizability to real-world applications requires further validation.

In order to more intuitively display the effect of our proposed method, we used the t-SNE (67) algorithm to reduce the dimensionality of high-dimensional feature data and drew a scatter plot on a 2-dimensional plane. Figures 9A–C depict scatter plots obtained by removing FEnM, EMA, and Data Augmentation, respectively. There are instances where the glioma class and the meningioma class are interconnected and nested. However, in Figure 9D, obtained by EnSLDe, the sample points of each class are closely clustered together, with clear separation between different categories. This intuitively underscores the significance of FEnM, EMA, and Data Augmentation for the model. The ability of the model to distinguish features effectively is enhanced by them.

As shown in Figure 4 and Table 2, the EnSLDe model achieves superior classification performance for pituitary tumors (precision: 0.9946, recall: 0.9946) compared to gliomas (precision: 0.9894, recall: 0.9946) and meningiomas (precision: 0.9718, recall: 0.9787), the latter of which exhibits the lowest performance metrics. A comparison of Figures 9A–D illustrates that EnSLDe employs effective strategies to differentiate gliomas from meningiomas. However, persistent feature overlap hinders the model's ability to achieve optimal classification accuracy.

The EnSLDe model is designed to capture both short- and long-range dependencies within images, demonstrating considerable potential for generalization beyond the classification of brain tumors. Its architecture, which incorporates a multi-scale parallel subnetwork and feature enhancement modules, is well-suited for a wide range of medical imaging tasks. Additionally, the model is adaptable to the classification of tumors in various organs, such as lung, breast, and liver tumors. The model's ability to effectively capture contextual information makes it suitable for the

identification of different lesion types and the detection of abnormalities across a diverse array of medical conditions.

Adapting the EnSLDe model to a new task necessitates several adjustments. First, the model requires retraining on a task-specific dataset, including modifying the number of output categories and fine-tuning the classification module. Furthermore, the feature extraction module may require modification to account for variations in imaging characteristics, such as resolution and contrast. Despite its design efficiency, the EnSLDe model exhibits limited scalability, particularly in resource-constrained environments. Training the model demands substantial computational resources, particularly for large-scale datasets. However, incorporating efficient convolutional layers and depthwise separable convolutions mitigates these computational demands. To address scalability challenges, several strategies may be implemented. For instance, model compression techniques (e.g., pruning and quantization) can substantially reduce computational complexity while maintaining competitive performance.

To further understand the decision-making process of the proposed EnSLDe model and validate its ability to focus on relevant regions in brain tumor classification, we visualized the feature maps using the Grad-CAM++ method. The results are shown in Figure 10. Grad-CAM++ is a widely used technique for visualizing the regions of interest in image classification tasks, providing insights into the model's attention mechanism. As shown in Figure 10, the feature maps generated by the EnSLDe model effectively highlight brain tumor regions, demonstrating the model's ability to distinguish between brain tumor and non-tumor regions. This visualization confirms that the model focuses on tumor regions, which is critical for accurate classification. However, it is also clear that the model focused on other non-tumor regions. This observation suggests that the model effectively captures key brain tumor features while incorporating additional contextual information from surrounding brain regions, which may contribute to its high classification accuracy. While the EnSLDe model demonstrated strong performance in focusing on relevant regions, the visualization results also highlighted areas for potential improvement. Specifically, the model's focus on non-tumor regions suggests that there may be opportunities to refine the feature extraction and enhancement modules to emphasize the most critical features further. Future work could explore advanced attention mechanisms or additional regularization techniques to ensure that the model focuses more precisely on tumor regions, potentially leading to higher classification accuracy.

5 Conclusion

A new multi-class brain tumor classification model, named EnSLDe, has been proposed. This model is primarily composed of three modules: FExM (Feature Extraction Module), FEnM (Feature Enhancement Module), and the classification module. FExM efficiently extracts features using convolutional layers and residual networks and combines EMA (Efficient Multi-Attention) to simultaneously focus on both channel and spatial information of the features. This effectively preserves the information of each channel, preventing the loss of important features during the compression of the channel dimension. The design of FEnM aims to deeply integrate shallow and deep features, facilitating a more comprehensive understanding of the features and the extraction of advanced and important features. Additionally, the model's ability to capture short-range and long-range dependencies has been enhanced. The feature enhancement module further strengthens the features by effectively capturing important dependencies over a large sequence range while preserving local key information. The double-layer fully connected structure is adopted as the core of the classification module and combined with dropout regularization technology, which further improves the model classification

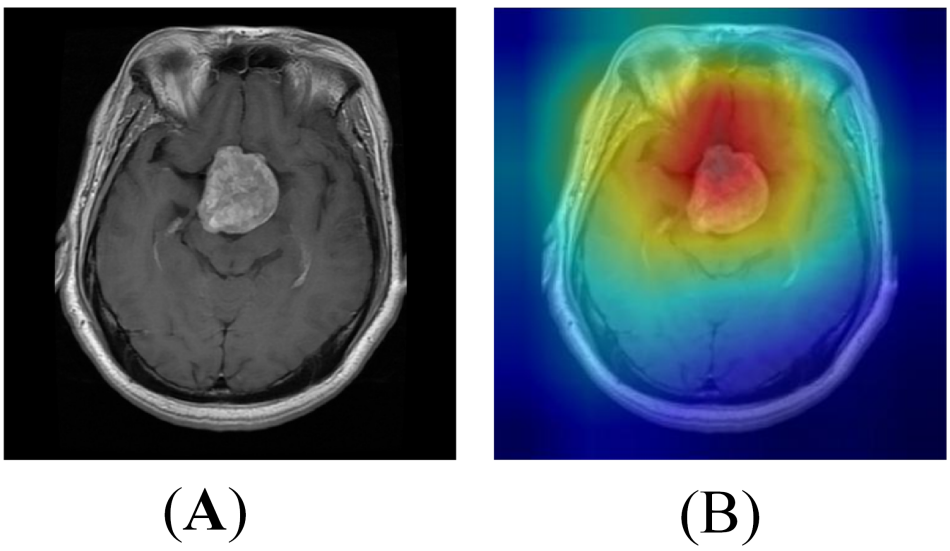


FIGURE 10
Heat map visualization of the model **(A)** Original image **(B)** Heat map.

performance. Experimental evaluations conducted on the challenging Cheng dataset and BT-large-4c dataset demonstrate the excellent performance of our model in brain tumor classification tasks. On the Cheng dataset, the model achieves accuracy, recall, precision, and F1-score of 98.69%, 98.53%, 98.64%, and 98.59%, respectively. Similarly, on the BT-large-4c dataset, the model attains accuracy, recall, precision, and F1-score of 97.10%, 97.11%, 97.11%, and 97.11%, respectively. Indeed, the differentiation between glioma and meningioma remains suboptimal. Further refinement is required to enhance the model's ability to distinguish accurately between these two tumor types. Future studies should augment the dataset to include a broader range of brain disorders, thereby enriching the model's training corpus and enhancing its capacity to differentiate among diverse neurological pathologies. Additionally, strategic modifications to the model's architecture, training protocols, and loss functions could be implemented to optimize its discriminative performance in distinguishing gliomas from meningiomas. And the model was deployed, and the clinical capabilities of the model were verified by combining the doctors commanded by experience.

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://figshare.com/articles/dataset/brain_tumor_dataset/1512427 https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri.

Author contributions

WC: Conceptualization, Investigation, Project administration, Writing – original draft. JL: Formal Analysis, Software, Validation, Visualization, Writing – review & editing. XT: Formal Analysis, Software, Writing – original draft. JZ: Conceptualization, Project administration, Software, Writing – review & editing. GD: Project administration, Supervision, Writing – review & editing. QF: Validation, Writing – review & editing. HJ: Project administration, Validation, 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 Foundation of He'nan Educational Committee (Grant No. 24A320004), Major Science and Technology Projects of Henan Province (Grant No. 221100210500), the Medical and Health Research Project in Luoyang (Grant No. 2001027A), and the Construction Project of Improving Medical Service Capacity of Provincial Medical Institutions in Henan Province (Grant No. 2017-51).

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

References

1. Asif S, Yi W, Ain QU, Hou J, Yi T, Si J. Improving effectiveness of different deep transfer learning-based models for detecting brain tumors from MR images. *IEEE Access*. (2022) 10:34716–30. doi: 10.1109/ACCESS.2022.3153306

2. Bouhafra S, El Bahi H. Deep learning approaches for brain tumor detection and classification using MRI images, (2020 to 2024): A systematic review. *J Of Imaging Inf In Med*. (2024). doi: 10.1007/s10278-024-01283-8

3. Ghosal P, Reddy S, Sai C, Pandey V, Chakraborty J, Nandi D. A deep adaptive convolutional network for brain tumor segmentation from multimodal MR images. *TENCON 2019 - 2019 IEEE Region 10 Conf (TENCON)*. (2019), 1065–70. doi: 10.1109/TENCON.2019.8929402

4. Yu Z, Li X, Li J, Chen W, Tang Z, Geng D. HSA-net with a novel CAD pipeline boosts both clinical brain tumor MR image classification and segmentation. *Comput Biol Med*. (2024) 170:108039. doi: 10.1016/j.compbiomed.2024.108039

5. Almalki YE, Ali MU, Kallu KD, Masud M, Zafar A, Alduraibi SK, et al. Isolated convolutional-neural-network-based deep-feature extraction for brain tumor classification using shallow classifier. *Diagnostics*. (2022) 12:1793. doi: 10.3390/diagnostics12081793

6. Mehnatkesh H, Jalali SMJ, Khosravi A, Nahavandi S. An intelligent driven deep residual learning framework for brain tumor classification using MRI images. *Expert Syst Appl*. (2023) 213:119087. doi: 10.1016/j.eswa.2022.119087

7. Mondal A, Shrivastava VK. A novel Parametric Flatten-p Mish activation function based deep CNN model for brain tumor classification. *Comput Biol Med*. (2022) 150:106183. doi: 10.1016/j.compbiomed.2022.106183

8. ZainEldin H, Gamel SA, El-Kenawy E-SM, Alharbi AH, Khafaga DS, Ibrahim A, et al. Brain tumor detection and classification using deep learning and sine-cosine fitness grey wolf optimization. *Bioengineering*. (2022) 10:18. doi: 10.3390/bioengineering10010018

9. Louis DN, Perry A, Reifenberger G, Von Deimling A, Figarella-Branger D, Cavenee WK, et al. The 2016 world health organization classification of tumors of the central nervous system: A summary. *Acta Neuropathologica*. (2016) 131:803–20. doi: 10.1007/s00401-016-1545-1

10. Amin J, Sharif M, Raza M, Saba T, Rehman A. Brain tumor classification: feature fusion. In: *2019 International Conference on Computer and Information Sciences*

(ICCIS) (2019) (Sakaka, Saudi Arabia: Institute of Electrical and Electronics Engineers (IEEE)). p. 1–6. doi: 10.1109/ICCISci.2019.8716449

11. Njeh I, Sallemi L, Ben Slima M, Ben Hamida A, Lehericy S, Galanaud D. A Computer Aided Diagnosis `CAD' for Brain Glioma Exploration. In: *2014 1st International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)* (2014) (Sousse: Institute of Electrical and Electronics Engineers (IEEE)). p. 243–8. doi: 10.1109/ATSIP.2014.6834615

12. Singh RS, Saini BS, Sunkaria RK. Classification of cardiac heart disease using reduced chaos features and 1-norm linear programming extreme learning machine. *Int J For Multiscale Comput Eng*. (2018) 16:465–86. doi: 10.1615/IntJMultCompEng.2018026587

13. Shahid AH, Singh MP, Roy B, Aadarsh A. Coronary artery disease diagnosis using feature selection based hybrid extreme learning machine. In: *2020 3rd International Conference On Information And Computer Technologies (Icict 2020)* (2020) (San Jose, CA, USA: Institute of Electrical and Electronics Engineers (IEEE)). p. 341–6. doi: 10.1109/ICICT50521.2020.00060

14. Xie W, Li Y, Ma Y. Breast mass classification in digital mammography based on extreme learning machine. *Neurocomputing*. (2016) 173:930–41. doi: 10.1016/j.neucom.2015.08.048

15. Heidari M, Lakshmivarahan S, Mirniaharikandehei S, Danala G, Maryada SKR, Liu H, et al. Applying a random projection algorithm to optimize machine learning model for breast lesion classification. *IEEE Trans On Biomed Eng*. (2021) 68:2764–75. doi: 10.1109/TBME.2021.3054248

16. Pyrkov TV, Slipensky K, Barg M, Kondrashin A, Zhurov B, Zenin A, et al. Extracting biological age from biomedical data via deep learning: Too much of a good thing? *Sci Rep*. (2018) 8:5210. doi: 10.1038/s41598-018-23534-9

17. Sarki R, Ahmed K, Wang H, Zhang Y. Automated detection of mild and multi-class diabetic eye diseases using deep learning. *Health Inf Sci And Syst*. (2020) 8:32. doi: 10.1007/s13755-020-00125-5

18. Jeong W-Y, Kim J-H, Park J-E, Kim M-J , Jong-Min L. Evaluation of classification performance of inception V3 algorithm for chest X-ray images of patients with cardiomegaly. 한국방사선학회논문지. (2021) 15:455–61. doi: 10.7742/jksr.2021.15.4.455

19. Chowdhury MS, Sultan T, Jahan N, Mridha MF, Safran M, Alfarhood S, et al. Leveraging deep neural networks to uncover unprecedented levels of precision in the diagnosis of hair and scalp disorders. *Skin Res And Technol*. (2024) 30:e13660. doi: 10.1111/srt.13660

20. Sharifrazi D, Alizadehsani R, Joloudari JH, Band SS, Hussain S, Sani ZA, et al. CNN-KCL: Automatic myocarditis diagnosis using convolutional neural network combined with k-means clustering. *Math Biosci And Eng*. (2022) 19:2381–402. doi: 10.3934/mbe.2022110

21. Shahin AI, Aly W, Aly S. MBTFCN: A novel modular fully convolutional network for MRI brain tumor multi-classification. *Expert Syst Appl*. (2023) 212:118776. doi: 10.1016/j.eswa.2022.118776

22. Xue Z, Chen W, Li J. Enhancement and fusion of multi-scale feature maps for small object detection. In: *2020 39th Chinese Control Conference (CCC)* (2020) (Shenyang, China: Institute of Electrical and Electronics Engineers (IEEE)). p. 7212–7. doi: 10.23919/CCC50068.2020.9189352

23. Chen Y, Zhang C, Chen B, Huang Y, Sun Y, Wang C, et al. Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases. *Comput Biol Med*. (2024) 170:107917. doi: 10.1016/j.compbiomed.2024.107917

24. Lin T-Y, Dollar P, Girshick R, He K, Hariharan B, Belongie S. (2017). Feature pyramid networks for object detection, in: *The conference title is Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2117–25.

25. Bansal T, Jindal N. An improved hybrid classification of brain tumor MRI images based on conglomeration feature extraction techniques. *Neural Computing Appl*. (2022) 34:9069–86. doi: 10.1007/s00521-022-06929-8

26. Khan MA, Lali IU, Rehman A, Ishaq M, Sharif M, Saba T, et al. Brain tumor detection and classification: A framework of marker-based watershed algorithm and multilevel priority features selection. *Microscopy Res Technique*. (2019) 82:909–22. doi: 10.1002/jemt.23238

27. Mehmood A, Yang S, Feng Z, Wang M, Ahmad AS, Khan R, et al. A transfer learning approach for early diagnosis of Alzheimer's disease on MRI images. *Neuroscience*. (2021) 460:43–52. doi: 10.1016/j.neuroscience.2021.01.002

28. Raza A, Ayub H, Khan JA, Ahmad I, S. Salama A, Daradkeh YI, et al. A hybrid deep learning-based approach for brain tumor classification. *Electronics*. (2022) 11:1146. doi: 10.3390/electronics11071146

29. Díaz-Pernas FJ, Martínez-Zarzuela M, Antón-Rodríguez M, González-Ortega D. A deep learning approach for brain tumor classification and segmentation using a multiscale convolutional neural network. *Healthcare*. (2021) 9:153. doi: 10.3390/healthcare9020153

30. Ayadi W, Elhamzi W, Charfi I, Atri M. Deep CNN for brain tumor classification. *Neural Process Lett*. (2021) 53:671–700. doi: 10.1007/s11063-020-10398-2

31. Sreenivasa Reddy B, Sathish A. A Multiscale Atrous Convolution-based Adaptive ResUNet3 + with Attention-based ensemble convolution networks for brain tumour segmentation and classification using heuristic improvement. *Biomed Signal Process Control*. (2024) 91:105900. doi: 10.1016/j.bspc.2023.105900

32. Ghosal P, Nandanwar L, Kanchan S, Bhadra A, Chakraborty J, Nandi D. Brain tumor classification using resNet-101 based squeeze and excitation deep neural network. In: *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)* (2019) (Gangtok, India: Institute of Electrical and Electronics Engineers (IEEE)). p. 1–6. doi: 10.1109/ICACCP.2019.8882973

33. Islam M, Talukder M, Uddin M, Akhter A, Khalid M. BrainNet: precision brain tumor classification with optimized efficientNet architecture. *Int J Of Intelligent Syst*. (2024) 2024. doi: 10.1155/2024/3583612

34. Aurna NF, Abu Yousuf M, Abu Taher K, Azad AKM, Moni MA. A classification of MRI brain tumor based on two stage feature level ensemble of deep CNN models. *Comput In Biol And Med*. (2022) 146. doi: 10.1016/j.compbiomed.2022.105539

35. Musallam AS, Sherif AS, Hussein MK. A new convolutional neural network architecture for automatic detection of brain tumors in magnetic resonance imaging images. *IEEE Access*. (2022) 10:2775–82. doi: 10.1109/ACCESS.2022.3140289

36. Kumar SA, Sasikala S. Automated brain tumour detection and classification using deep features and Bayesian optimised classifiers. *Curr Med Imaging*. (2023) 20. doi: 10.2174/1573405620666230328092218

37. Alshayeji M, Al-Buloushi J, Ashkanani A, Abed S. Enhanced brain tumor classification using an optimized multi-layered convolutional neural network architecture. *Multimedia Tools Appl*. (2021) 80:28897–917. doi: 10.1007/s11042-021-10927-8

38. Irmak E. Multi-classification of brain tumor MRI images using deep convolutional neural network with fully optimized framework. *Iranian J Sci Technology Trans Electrical Eng*. (2021) 45:1015–36. doi: 10.1007/s40998-021-00426-9

39. Rammurthy D, Mahesh PK. Whale Harris hawks optimization based deep learning classifier for brain tumor detection using MRI images. *J Of King Saud University-Computer And Inf Sci*. (2022) 34:3259–72. doi: 10.1016/j.jksuci.2020.08.006

40. Alyami J, Rehman A, Almutairi F, Fayyaz AM, Roy S, Saba T, et al. Tumor localization and classification from MRI of brain using deep convolution neural network and salp swarm algorithm. *Cogn Comput*. (2023) 16:2036–46. doi: 10.1007/s12559-022-10096-2

41. Tummala S, Kadry S, Bukhari SAC, Rauf HT. Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. *Curr Oncol*. (2022) 29:7498–511. doi: 10.3390/curroncol29100590

42. Ferdous GJ, Sathi KA, Md. A, Hoque MM, Dewan MAA. LCDEiT: A linear complexity data-efficient image transformer for MRI brain tumor classification. *IEEE Access*. (2023) 11:20337–50. doi: 10.1109/ACCESS.2023.3244228

43. Asiri AA, Shaf A, Ali T, Pasha MA, Khan A, Irfan M, et al. Advancing brain tumor detection: Harnessing the Swin Transformer's power for accurate classification and performance analysis. *Peerj Comput Sci*. (2024) 10. doi: 10.7717/peerj-cs.1867

44. Dutta TK, Nayak DR, Pachori RB. GT-Net: Global transformer network for multiclass brain tumor classification using MR images. *Biomed Eng Lett*. (2024) 14:1069–77. doi: 10.1007/s13534-024-00393-0

45. Ouyang D, He S, Zhang G, Luo M, Guo H, Zhan J, et al. Efficient Multi-Scale Attention Module with Cross-Spatial Learning. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (2023), 1–5. doi: 10.1109/ICASSP49357.2023.10096516

46. Quan Y, Zhang D, Zhang L, Tang J. Centralized feature pyramid for object detection. *IEEE Trans Image Process*. (2023) 32:4341–54. doi: 10.1109/TIP.2023.3297408

47. Zhang Z, Sabuncu M. Generalized cross entropy loss for training deep neural networks with noisy labels. *Adv Neural Inf Process Syst*. (2018) 31.

48. Cheng J, Huang W, Cao S, Yang R, Yang W, Yun Z, et al. Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PloS One*. (2015) 10:e0140381. doi: 10.1371/journal.pone.0140381

49. Bhuvaji S, Kadam A, Bhumkar P, Dedge S, Kanchan S. Brain tumor classification (MRI). *Kaggle*. (2020) 10.

50. Alsaggaf W, Cömert Z, Nour M, Polat K, Brdesee H, Toğaçar M. Predicting fetal hypoxia using common spatial pattern and machine learning from cardiotocography signals. *Appl Acoustics*. (2020) 167:107429. doi: 10.1016/j.apacoust.2020.107429

51. Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (2018) (Salt Lake City, USA: Institute of Electrical and Electronics Engineers (IEEE)). pp. 7132–41. pp. 7132–41.

52. Hou Q, Zhou D, Feng J. Coordinate Attention for Efficient Mobile Network Design. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. (2021) (Nashville, TN, USA: Institute of Electrical and Electronics Engineers (IEEE)). pp. 13713–22. pp. 13713–22.

53. Woo S, Park J, Lee J-Y, Kweon IS. CBAM: Convolutional Block Attention Module. *Proceedings of the European Conference on Computer Vision (ECCV)*. (2018) (Munich, Germany: Springer). pp. 3–19.

54. Noreen N, Palaniappan S, Qayyum A, Ahmad I, O. Alassafi M. Brain tumor classification based on fine-tuned models and the ensemble method. *Computers Materials Continua*. (2021) 67:3967–82. doi: 10.32604/cmc.2021.014158

55. Bodapati JD, Shaik NS, Naralasetti V, Mundukur NB. Joint training of two-channel deep neural network for brain tumor classification. *Signal Image Video Process*. (2021) 15:753–60. doi: 10.1007/s11760-020-01793-2

56. Shaik NS, Cherukuri TK. Multi-level attention network: Application to brain tumor classification. *Signal Image Video Process*. (2022) 16:817–24. doi: 10.1007/s11760-021-02022-0

57. Öksüz C, Urhan O, Güllü MK. Brain tumor classification using the fused features extracted from expanded tumor region. *Biomed Signal Process Control*. (2022) 72:103356. doi: 10.1016/j.bspc.2021.103356

58. Jaspin K, Selvan S. Multiclass convolutional neural network based classification for the diagnosis of brain MRI images. *Biomed Signal Process Control*. (2023) 82:104542. doi: 10.1016/j.bspc.2022.104542

59. Khan Md.SI, Rahman A, Debnath T, Karim M, Nasir MK, Band SS, et al. Accurate brain tumor detection using deep convolutional neural network. *Comput Struct Biotechnol J*. (2022) 20:4733–45. doi: 10.1016/j.csbj.2022.08.039

60. Satyanarayana G, Appala Naidu P, Subbaiah Desanamukula V, Satish kumar K, Chinna Rao B. A mass correlation based deep learning approach using deep Convolutional neural network to classify the brain tumor. *Biomed Signal Process Control*. (2023) 81:104395. doi: 10.1016/j.bspc.2022.104395

61. Kibriya H, Masood M, Nawaz M, Nazir T. Multiclass classification of brain tumors using a novel CNN architecture. *Multimedia Tools Appl*. (2022) 81:29847–63. doi: 10.1007/s11042-022-12977-y

62. Dutta TK, Nayak DR, Zhang Y-D. ARM-Net: Attention-guided residual multiscale CNN for multiclass brain tumor classification using MR images. *Biomed Signal Process Control*. (2024) 87:105421. doi: 10.1016/j.bspc.2023.105421

63. Khan SUR, Zhao M, Asif S, Chen X. Hybrid-NET: A fusion of DenseNet169 and advanced machine learning classifiers for enhanced brain tumor diagnosis. *Int J Imaging Syst Technol*. (2024) 34:e22975. doi: 10.1002/ima.22975

64. Demir F, Akbulut Y. A new deep technique using R-CNN model and L1NSR feature selection for brain MRI classification. *Biomed Signal Process Control*. (2022) 75:103625. doi: 10.1016/j.bspc.2022.103625

65. Senan EM, Jadhav ME, Rassem TH, Aljaloud AS, Mohammed BA, Al-Mekhlafi ZG. Early diagnosis of brain tumour MRI images using hybrid techniques between deep and machine learning. *Comput Math Methods Med*. (2022) 2022:1–17. doi: 10.1155/2022/8330833

66. Ravinder M, Saluja G, Allabun S, Alqahtani MS, Abbas M, Othman M, et al. Enhanced brain tumor classification using graph convolutional neural network architecture. *Sci Rep*. (2023) 13:14938. doi: 10.1038/s41598-023-41407-8

67. Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. (2008) 9:2579–605.

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 08 January 2025
DOI 10.3389/fonc.2024.1482965

OPEN ACCESS

EDITED BY
Sandeep Kumar Mishra,
Yale University, United States

REVIEWED BY
Xiaoliang Shao,
The Third Affiliated Hospital of Soochow University, China
Xiaonan Shao,
Third Affiliated Hospital of Soochow University, China
Nazreen Pallikkavaliyaveetil MohammedSheriff,
University of Michigan, United States

*CORRESPONDENCE
Liming Xia
xialiming2017@outlook.com
Dazhi Chen
chdazhi@sina.com

RECEIVED 19 August 2024
ACCEPTED 17 December 2024
PUBLISHED 08 January 2025

CITATION
Ma X, He W, Chen C, Tan F, Chen J, Yang L, Chen D and Xia L (2025) A CT-based deep learning model for preoperative prediction of spread through air spaces in clinical stage I lung adenocarcinoma.
Front. Oncol. 14:1482965.
doi: 10.3389/fonc.2024.1482965

A CT-based deep learning model for preoperative prediction of spread through air spaces in clinical stage I lung adenocarcinoma

Xiaoling Ma[1], Weiheng He[1], Chong Chen[2], Fengmei Tan[3], Jun Chen[4], Lili Yang[1], Dazhi Chen[1]* and Liming Xia[2]*

[1]Medical imaging center, People's Hospital of Ningxia Hui Autonomous Region, Yinchuan, China, [2]Department of Radiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [3]Department of Pathology, People's Hospital of Ningxia Hui Autonomous Region, Yinchuan, China, [4]Department of Radiology, Bayer Healthcare, Wuhan, China

Objective: To develop and validate a deep learning signature for noninvasive prediction of spread through air spaces (STAS) in clinical stage I lung adenocarcinoma and compare its predictive performance with conventional clinical-semantic model.

Methods: A total of 513 patients with pathologically-confirmed stage I lung adenocarcinoma were retrospectively enrolled and were divided into training cohort (n = 386) and independent validation cohort (n = 127) according to different center. Clinicopathological data were collected and CT semantic features were evaluated. Multivariate logistic regression analyses were conducted to construct a clinical-semantic model predictive of STAS. The Swin Transformer architecture was adopted to develop a deep learning signature predictive of STAS. Model performance was assessed using area under the receiver operating characteristic curve (AUC), sensitivity, specificity, positive and negative predictive value, and calibration curve. AUC comparisons were performed by the DeLong test.

Results: The proposed deep learning signature achieved an AUC of 0.869 (95% CI: 0.831, 0.901) in training cohort and 0.837 (95% CI: 0.831, 0.901) in validation cohort, surpassing clinical-semantic model both in training and validation cohort (all $P<0.01$). Calibration curves demonstrated good agreement between STAS predicted probabilities using deep learning signature and actual observed probabilities in both cohorts. The inclusion of all clinical-semantic risk predictors failed to show an incremental value with respect to deep learning signature.

Conclusions: The proposed deep learning signature based on Swin Transformer achieved a promising performance in predicting STAS in clinical stage I lung adenocarcinoma, thereby offering information in directing surgical strategy and facilitating adjuvant therapeutic scheduling.

KEYWORDS

deep learning, lung adenocarcinoma, spread though air space, computer tomography, prediction

Introduction

Lung cancer remains the leading lethal malignancy, responsible for 12.4% of all newly-diagnosed cases worldwide in 2022 (1). As the predominant cause of lung cancer-related mortality, lung adenocarcinoma exhibits distinctive histological growth pattern and molecular genotyping (2). Spread through air spaces (STAS) is a unique invasion pattern separate from lymphatic-vascular and visceral pleural invasion, with a predisposition in lung adenocarcinoma. Initially introduced by Kadota et al. and explicitly defined in the World Health Organization Classification of Lung Cancer in 2015, STAS refers to the dissemination of tumor cells as solid nests, micropapillary clusters or single cells into the peritumoral alveolar airspaces (3). Multiply studies have consistently demonstrated that STAS serves as a well-established prognosticator for lung adenocarcinoma undergoing sublobectomy, indicating an increased risk of postoperative relapse and worse prognosis (4–6). STAS is recognized as a pathological indicator for identifying the beneficiaries of adjuvant chemotherapy among stage IB patients (7). Therefore, STAS is of great significance in identifying high-risk patients and guiding personalized therapeutic strategies.

However, intraoperative pathological assessment for STAS through rapid frozen sections has been proved to be of limited sensitivity and reproducibility (8). The shifting of tumor cells to the peritumoral alveolar airspaces caused by manual operations such as extrusion, blade cutting and tissue dysfixation were hardly distinguished from STAS cell clusters, thereby hindering the reliable application of this approach. Several scholars exploited CT semantic indicators for STAS by visual inspection or manual measurement, such as tumor diameter, ground-glass opacity (GGO) components, and pleural retraction (9, 10). Nevertheless, these indicators rely on subjective judgement and professional skills, making them unsuitable for widespread clinical practice due to inconsistent interpretation criteria. Several studies developed CT-based radiomics signature predictive of STAS, but the radiomics approach involves several sequential processing steps such as tumor delineation, dimension reduction and model building (11, 12). The efficiency of radiomics modeling is highly influenced by interobserver heterogeneity and handling quality at each step.

Deep learning is an end-to-end network architecture, characterized by the ingestion of data from the input end and the generation of prediction results from the output end. The error between prediction result and actual observation is iteratively propagated through each layer, facilitating model adjustment and convergence. On account of the advantages of automatically learning and extracting representative information, deep learning has achieved remarkable efficacy in distinguishing histological subtypes, evaluating treatment response, and predicting survival (13–15). In this study, we employed Swin Transformer, a deep learning framework exploited by Microsoft Research Asia, to construct and validate a CT-based deep learning predictive model for STAS in lung adenocarcinoma. This study also sought to investigate the incremental value of clinical characteristics and conventional CT semantic features over the deep learning signature.

Methods

Patients

This study was approved by the Ethics Committee and the requirement for informed consent was waived due to its retrospective nature. The patients who underwent radical resection at the main campus of Tongji Hospital (Center 1) from October 2021 to June 2022 were systematically reviewed. Inclusion criteria were: (1) invasive lung adenocarcinoma confirmed by pathology; (2) maximum tumor diameter on CT images ≤ 4 cm; (3) no radiological signs of locoregional lymph node invasion or distant metastasis; (4) no preoperative radiotherapy, chemotherapy or targeted therapy; (5) interval time of preoperative CT examination and operation within two weeks. The exclusion criteria were: (a) rare histological variants; (b) simultaneous or metachronous tumors; (c) unavailable thin-section CT images or obvious image artifacts; (d) insufficient peritumoral parenchyma reserved for STAS assessment; (e) subjected to other cancers. Tumor staging was based on the eighth edition of the TNM staging system. Following the same criteria, patients undergoing radical surgical resection at the Sino-Germany Guanggu Campus of

Tongji Hospital (Center 2) from January 2022 to June 2022 were retrospectively enrolled. Clinical information including gender, age, smoking history, pack-year and serum CEA level were acquired from clinical electronic records. The recruitment workflow is illustrated in Figure 1.

Histological assessment

Pathological characteristics including histological subtype, Ki-67 labeling index (LI), visceral pleural invasion, lymphatic-vascular invasion, pathological TNM staging and STAS were documented. The excision specimen was fixed in 10% formalin and embedded in paraffin before sectioned. Hematoxylin-eosin staining, immunohistochemistry staining and elastic fiber staining were performed accordingly. Two pathologists with experiences of 5 years and 11 years independently interpreted STAS on the sections. Initially, tumor smooth interfaces were recognized by naked eyes and at low-magnification (×10). Subsequently, three areas with the most abundant STAS were selected for interpretation at high-magnification (×200). If any of the following forms of tumor cells are observed within peritumoral alveolar airspaces, it is judged to be STAS-positive: (1) micropapillary clusters without a central fibrovascular core; (2) solid tumor nests; (3) discrete single tumor cells. Ki-67 LI is determined by the percentage of cells with stained-brown nuclei among 1000 tumor cells via immunohistochemical staining. Invasive lung adenocarcinoma is categorized into five histological subtypes based on growth architecture: lepidic, acinar, papillary, micropapillary and solid predominant adenocarcinoma.

CT scanning protocol and semantic feature interpretation

The patients were examined using multi-slice spiral CT scanners including GE Discovery 750 HD, TOSHIBA Aquilion One TSX-301A, Philips Brilliance ICT 256 and GE Optima CT 660. The acquisition parameters were detailed in Supplementary Data Sheet 1. CT semantic features were independently evaluated by two radiologists with 12 and 7 years of experience, respectively, blinded to the clinicopathological information. The lung window (width: 1600 HU; level: -600 HU) and mediastinal window (width: 400 HU; level: 40 HU) were fixed, respectively. CT semantic features included affiliated lobe, location, attenuation type, tumor total diameter, tumor consolidation diameter, consolidation-to-tumor ratio (CTR), shape, boundary, lobulation, spiculation, cavity, vacuole, air bronchogram, and plural attachment. CTR is quantified by the ratio of tumor consolidation diameter and total diameter. The definitions of CT semantic features were elucidated in Supplementary Data Sheet 1. The interobserver agreement for categorical and continuous variables was evaluated using Cohen 's kappa coefficient and intraclass correlation coefficient (ICC), respectively. The average measured by two radiologists was taken as the final value for continuous variables. Consensus on divergent categorical variables was reached through discussion involving a third radiologist.

Tumor segmentation and deep learning signature development

The automatic virtual adversarial training segmentation algorithm, based on a three-dimensional U-shape convolutional neural network known as 3D U-Net, was employed to achieve tumor segmentation. The topology of U-net was showed in Supplementary Data Sheet 1. For modeling, we proposed a deep learning framework called Swin Transformer to develop a signature predictive of STAS. The overall architecture consists of four transformer stages comprising Patch Embedding/Merging and Swin Transformer Blocks in each stage as revealed in Figure 2 and Supplementary Data Sheet 1. To mitigate overfitting due to limited amounts of data, the model was pretrained in CT images of lung cancer from the Cancer Imaging Archive followed by fine-tuned in 13510 CT images of lung adenocarcinoma in the training cohort. Furthermore, to compare the efficacy of different deep learning methods in predicting STAS, we applied ResNet-50, EfficientNet and ConvNeXt for modeling denoted as $Model_{ResNet-50}$, $Model_{EfficientNet}$ and $Model_{ConvNeXt}$. The original code for implementing Swin Transformer can be acquired at https://github.com/microsoft/Swin-Transformer. We implemented the neural network using PyTorch 1.4.1 library in Python 3.7.0 (https://pytorch.org).

Clinical-semantic model construction

Univariate analysis was initially performed to identify statistically significant clinical characteristics and CT semantic features between STAS positive and negative subgroups ($P < 0.05$) in the training cohort. Afterwards, features with Spearman correlation coefficient > 0.7 were removed in view of multicollinearity inference. The remaining features as candidate variables were included in multivariate logistic regression analysis to determine the features independently associated with STAS. The features were combined linearly weighted by their corresponding regression coefficients to construct clinical-semantic model. Given that the inherent design of preoperative prediction, pathological indicators were not included in logistic regression analysis, but compared across different STAS subgroups.

Statistical analysis

Statistical analysis was performed using MATLAB (MathWorksInc., Natick, MA) and SPSS (IBM, ver.26.0). Shapiro-Wilk test and Levene test were used to analyze the normality and

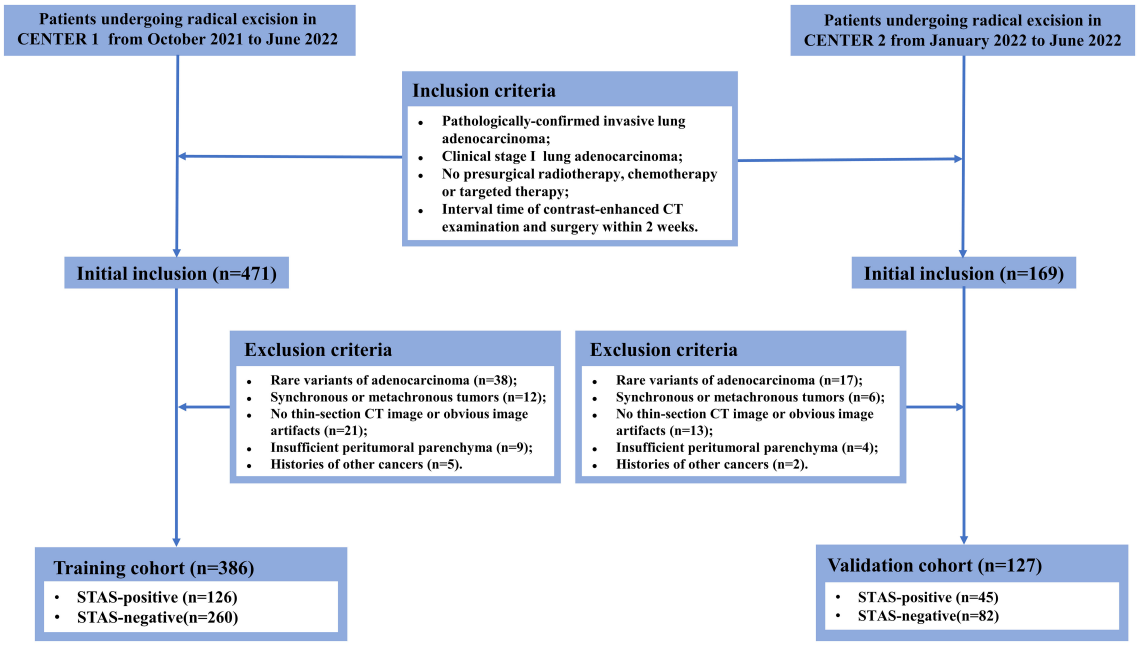


FIGURE 1
The workflow diagram of patient recruitment.

homogeneity of variance for continuous variables. The continuous variables were compared using the Student's t-test and Mann-Whitney U test, as appropriate. The comparisons of categorical variables were conducted by Chi-square test or Fisher exact test. Pearson correlation analysis was used to evaluate the correlation between features. The area under receiver operating characteristic curve (AUC), sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were used to quantify model performance. The calibration curve and Hosmer-Lemeshow test were employed to evaluate the consistency between predicted probabilities by deep learning signature and actual observations. A double-tailed $P<0.05$ indicated statistical significance.

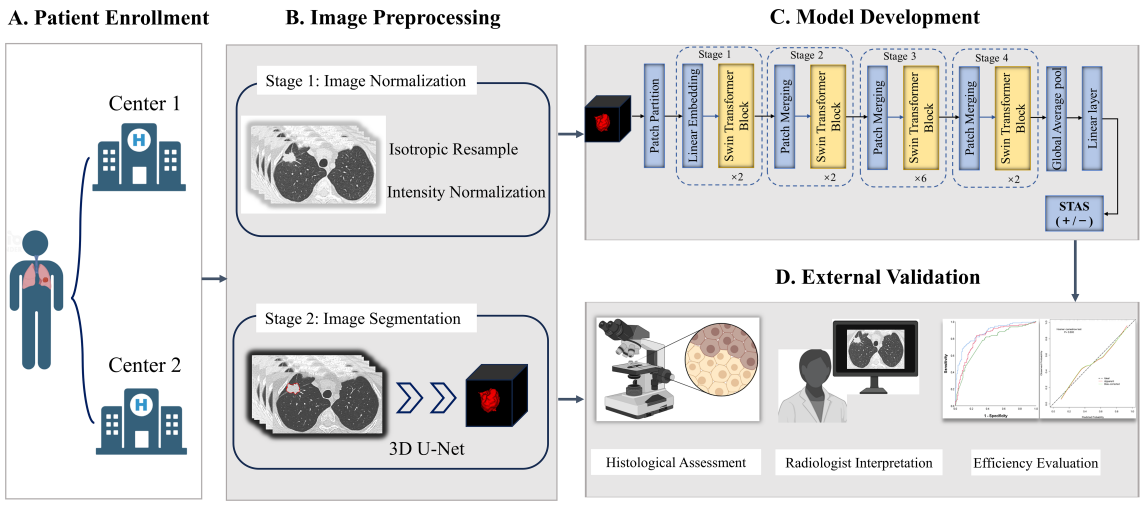


FIGURE 2
The overall framework of STAS prediction model development and validation. **(A)** Patients with lung adenocarcinoma were respectively enrolled from Center 1 and Center 2. **(B)** Imaging preprocessing included isotropic resample, intensity normalization and tumor automatic segmentation. **(C)** Deep learning signature predictive of STAS was developed based on Swin Transformer. **(D)** Histological assessment and radiologist interpretation were conducted for all patients in Center 1 and Center 2, and then model performance comparisons were performed. STAS, spread through air spaces.

Results

Baseline characteristics

In total, 126 eligible STAS-positive and 260 STAS-negative patients from Center 1 were enrolled to construct a training cohort (n=386). Accordingly, a total of 45 STAS-positive and 82 STAS-negative patients from Center 2 constituted an independent validation cohort (n=127). As revealed in Table 1, all clinicopathological characteristics and CT semantic features exhibited a balanced distribution between the training cohort and validation cohort. Of 513 patients, 239 (46.6%) were male [median age (interquartile): 59.0 (53.0, 65.0)] and 274 (53.4%) were female [median age (interquartile): 61.0 (54.0, 68.0)]. Totally, there were 171 (33.3%) and 342 (66.7%) patients pathologically-confirmed to be STAS-positive and STAS-negative, respectively.

The interobserver consistency assessment for CT semantic features

As shown in Table 2, ICC for tumor total diameter, tumor consolidation diameter and CTR were 0.988 (95% CI: 0.985, 0.990), 0.991 (95% CI: 0.990, 0.993) and 0.982 (95% CI: 0.979, 0.985), respectively. Cohen 's kappa coefficients for the categorical variables ranged from 0.808 to 0.992, indicative of satisfactory interobserver agreement in interpreting CT semantic features. The discrepant numbers (frequency) of categorical variables between two radiologists were also documented as revealed in Table 2.

The association of clinicopathological characteristics with STAS

As shown in Table 3, STAS was more likely occurred in patients with pack-year > 40 (P=0.002) and CEA > 5 ug/L (P<0.001), but it had no significant association with gender, age and smoking history. STAS was more frequently observed in micropapillary and solid predominant adenocarcinoma, but rarely occurred in lepidic predominant adenocarcinomas (P<0.001). Furthermore, STAS was closely related with visceral pleural invasion and lymphatic-vascular invasion (P<0.001 and P<0.001). Ki-67 LI in STAS-positive subgroup significantly exceeded that of STAS-negative subgroup (P<0.001). Additionally, lung adenocarcinoma with higher pathological T and N stages showed a higher prevalence of STAS (P<0.001 for both).

The association of CT semantic features with STAS

Tumor total diameter, tumor consolidation diameter and CTR in STAS-positive subgroup were significantly higher than those in STAS-negative subgroup (all P<0.001; Figures 3 and 4). Solid tumors, obscure boundary, spiculation, vacuole and pleural attachment were more frequent in STAS, but air bronchogram was less common in STAS (all P< 0.05). The tumor consolidation diameter and attenuation subtype were excluded from logistic regression analysis considering a strong correlation with CTR (r=0.839 and 0.913, P< 0.001). Finally, CEA (odds ratio [OR]: 2.022; 95% CI: 1.080, 3.784; P=0.028), vacuole (OR: 3.509; 95% CI: 1.488, 8.278; P=0.004), obscure boundary (OR: 2.716; 95% CI: 1.628, 4.529; P<0.001) and CTR (OR: 1.023; 95% CI: 1.014, 1.033; P<0.001) were included to construct the clinical-semantic model as the independent risk indicators for predicting STAS.

Model construction and efficacy evaluation

As shown in the Table 4 and Figure 5, the AUC for Swin Transformer based deep learning signature in the training cohort and validation cohort was 0.869 (95% CI: 0.831, 0.901) and 0.837 (95% CI: 0.761, 0.896), respectively. Encouragingly, Swin Transformer based deep learning signature achieved significantly higher AUC than $Model_{ResNet\text{-}50}$, $Model_{EfficientNet}$ and $Model_{ConvNeXt}$ in training cohort (0.869 vs. 0.800, 0.797 and 0.783; all P < 0.001), as well as than $Model_{EfficientNet}$ and $Model_{ConvNeXt}$ in validation cohort (0.837 vs. 0.775 and 0.795; P = 0.025 and 0.027), as shown in Supplementary Table E2. Deep learning signature showed an improvement in predictive performance than $Model_{ResNet\text{-}50}$ in validation cohort, but it did not reach statistical significance (0.837 vs. 0.799, P = 0.087).

Meanwhile, The AUC for CTR alone and clinical-semantic model was 0.709 (95% CI: 0.660, 0.754) and 0.764 (95% CI: 0.719, 0.806) in training cohort, as well as 0.734 (95% CI: 0.648, 0.808) and 0.714 (95% CI: 0.627, 0.790) in validation cohort, respectively. In the training cohort, deep learning signature performed far superior to CTR (0.869 vs. 0.709, P < 0.001) and clinical-semantic model (0.869 vs.0.764, P < 0.001), with a statistically significant difference. Notably, deep learning signature yielded significantly higher AUC than both CTR (0.837 vs. 0.734, P=0.006) and clinical-semantics model (0.837 vs. 0.714, P=0.002) in validation cohort. The sensitivity, specificity, PPV and NPV of deep learning signature in predicting STAS ranged from 0.578 to 0.706, 0.892 to 0.951, 0.761 to 0.867 and 0.804 to 0.862 across two cohorts, respectively. According to the Hosmer-Lemeshow test and calibration curve, the predicted STAS probabilities by deep learning signature revealed good agreement with the actual observations both in training cohort and validation cohort (P=0.600 and 0.082, respectively). Furthermore, when deep learning signature was incorporated into clinical-semantic model, all CT semantic risk predictors were eliminated from multivariate regression analysis, with merely deep learning signature remained. Pearson correlation analysis revealed a strong correlation between CTR and deep learning signature (r = 0.789, P < 0.001).

Discussion

This study revealed that CEA, tumor boundary, vacuolation and CTR are the independent clinical-semantic features associated with STAS in lung adenocarcinoma. The proposed deep learning model predictive of STAS based on Swin Transformer yielded an AUC of 0.869 (95% CI: 0.821, 0.908) and 0.837 (95% CI: 0.742, 0.908) in the training cohort and independent validation cohort, superior to

10.3389/fonc.2024.1482965

TABLE 1 The distribution of clinicopathological characteristics in training cohort and validation cohort.

Characteristic	All patients (n=513)	Training cohort (n=386)	Validation cohort (n=127)	*P* value
A. Clinical characteristics				
Gender				0.108
Female	274 (53.4%)	214 (55.4%)	60 (47.2%)	
Male	239 (46.6%)	172 (44.6%)	67 (52.8%)	
Age* (year)	60.0 (54.0, 66.0)	59.0 (54.0, 67.0)	62.0 (53.0, 66.0)	0.320
Smoking history				0.728
Nonsmoker	369 (72.0%)	280 (72.5%)	89 (70.1%)	
Former smoker	70 (13.6%)	50 (13.0%)	20 (15.7%)	
Current smoker	74 (14.4%)	56 (14.5%)	18 (14.2%)	
Pack-year				0.907
≤ 3	372 (72.5%)	280 (72.5%)	92 (72.4%)	
4-40	89 (17.3%)	68 (17.6%)	21 (16.5%)	
> 40	52 (10.2%)	38 (9.9%)	14 (11.1%)	
CEA (ug/L)				0.930
≤ 5	435 (84.8%)	327 (84.7%)	108 (85.0%)	
> 5	78 (15.2%)	59 (15.3%)	19 (15.0%)	
Surgical modalities				0.367
Wedge resection	14 (2.7%)	12 (3.1%)	2 (1.6%)	
Sublobectomy	25 (4.9%)	21 (5.4%)	4 (3.1%)	
Lobectomy	474 (92.4%)	353 (91.5%)	121 (95.3%)	
B. Histopathological characteristics				
Histological subtype				0.352
Lepidic	88 (17.2%)	63 (16.3%)	25 (19.7%)	
Acinar	240 (46.8%)	185 (47.9%)	55 (43.3%)	
Papillary	103 (20.0%)	78 (20.2%)	25 (19.7%)	
Micropapillary	43 (8.4%)	28 (7.3%)	15 (11.8%)	
Solid	39 (7.6%)	32 (8.3%)	7 (5.5%)	
Ki-67 LI* (%)	10 (3.5, 20.0)	9 (5, 20)	10 (3, 20)	0.171
Ki-67 LI				0.817
< 10%	255 (49.7%)	193 (50.0%)	62 (48.8%)	
≥ 10%	258 (50.3%)	193 (50.0%)	65 (51.2%)	
Visceral pleural invasion				0.786
Present	93 (18.1%)	71 (18.4%)	22 (17.3%)	
Absent	420 (81.9%)	315 (81.6%)	105 (82.7%)	
Lymph-vascular invasion				0.112
Present	70 (13.6%)	58 (15.0%)	12 (9.4%)	
Absent	443 (86.4%)	328 (85.0%)	115 (90.6%)	

(Continued)

TABLE 1 Continued

Characteristic	All patients (n=513)	Training cohort (n=386)	Validation cohort (n=127)	*P* value
B. Histopathological characteristics				
Pathological T stage				0.176
T1a	64 (12.5%)	48 (12.4%)	16 (12.6%)	
T1b	238 (46.4%)	170 (44.1%)	68 (53.5%)	
T1c	103 (20.1%)	85 (22.0%)	18 (14.2%)	
T2	108 (21.0%)	83 (21.5%)	25 (19.7%)	
Pathological N stage				0.402
N0	437 (85.1%)	328 (85.0%)	109 (85.8%)	
N1	27 (5.3%)	23 (6.0%)	4 (3.1%)	
N2	49 (9.6%)	35 (9.0%)	14 (11.1%)	
STAS				0.563
Positive	171 (33.3%)	126 (32.6%)	45 (35.4%)	
Negative	342 (66.7%)	260 (67.4%)	82 (64.6%)	

Unless otherwise stated, data were presented as numbers (percentages) and compared using the Chi-square test or Fisher's exact test.
*Data were presented as medians (inter-quartiles) and compared using the Mann-Whitney U test.
CEA, carcinoembryonic antigen; CTR, consolidation-to-tumor ratio; LI, labeling index; T, tumor; N, node; STAS, spread through air space.

TABLE 2 The interobserver agreement of CT semantic features for lung adenocarcinoma.

CT semantic feature	Disagreement	Kappa value/ICC	95% CI
Affiliated lobe‡	2 (0.4%)	0.992	0.980, 1.000
Location‡	20 (4.9%)	0.808	0.728, 0.888
Tumor total diameter§	NA	0.988	0.985, 0.990
Tumor consolidation diameter§	NA	0.991	0.990, 0.993
CTR§	NA	0.982	0.979, 0.985
Shape‡	17 (3.3%)	0.883	0.826, 0.940
Boundary‡	29 (5.7%)	0.844	0.789, 0.890
Lobulation‡	9 (1.8%)	0.871	0.787, 0.955
Spiculation‡	12 (2.3%)	0.953	0.928, 0.978
Cavity‡	8 (1.6%)	0.941	0.900, 0.982
Vacuole‡	11 (2.1%)	0.859	0.777, 0.941
Air bronchogram‡	37 (7.2%)	0.856	0.811, 0.901
Pleural attachment‡	13 (2.5%)	0.945	0.914, 0.976

§ICC was calculated for the continuous variables.
‡Cohen's kappa coefficient was calculated for the categorical variables.
Disagreement was presented as numbers (percentages).
ICC, intraclass correlation coefficient; CTR, consolidation-to-tumor ratio; CI, interval confidence.
NA, not applicable.

conventional CTR and clinical-semantic model. Furthermore, neither CTR nor clinical-semantic model exhibited an incremental value over deep learning signature, further confirming its superior predictive value.

For early-stage patients, sublobectomy can preserve more pulmonary function, reduce surgical complications, and shorten hospitalization time, particularly with an equivalent therapeutic effect to lobectomy (16). However, sublobectomy is not appropriate for STAS-positive patients due to a higher risk of locoregional relapse and distant metastasis compared with lobectomy. Another study proved that STAS had negligible adverse effects on prognosis if surgical margin distance exceeded 2 cm in limited resection (17). Thus, anatomic lobectomy and sufficient surgical margin should be recommended for STAS-positive patients to prevent recurrence caused by STAS. Dai et al. also demonstrated that recurrence-free survival rates and overall survival rates of stage IA STAS-positive patients were comparable to those of stage IB patients (18). Furthermore, stage IB patients with STAS-positive can benefit from adjuvant chemotherapy (7, 19). Consequently, STAS serves as a pathological indicator for T upstaging and risk stratification, as well as an effective biomarker for identifying the beneficiaries of adjuvant chemotherapy in early-stage patients.

Currently, there is limited research on leveraging deep learning technique to predict STAS, and the predictive capacity remains modest. Tao et al. applied 3D convolutional neural network to predict STAS in NSCLC, yielding an AUC of 0.790 in validation cohort (20). Wang et al. presented SE-Resnet50 for risk estimation of STAS in solid or part-solid lung adenocarcinoma, resulting a

TABLE 3 The relationships of clinicopathological characteristics and CT semantic features with STAS in training cohort.

Characteristics	Training cohort (n=386)	STAS positive (n=126)	STAS negative (n=260)	*P* value
A. Clinical characteristics				
Gender				0.201
Female	214 (55.4%)	64 (50.8%)	150 (57.7%)	
Male	172 (44.6%)	62 (49.2%)	110 (42.3%)	
Age* (year)	59.0 (54.0, 67.0)	59.0 (53.8, 68.3)	59.0 (54.0, 65.0)	0.422
Smoking history				0.051
Nonsmoker	280 (72.5%)	84 (66.7%)	186 (75.4%)	
Former smoker	50 (13.0%)	24 (19.0%)	26 (10.0%)	
Current smoker	56 (14.5%)	18 (14.3%)	38 (14.6%)	
Pack-year				0.002
≤ 3	280 (72.5%)	84 (66.7%)	196 (75.4%)	
4-40	68 (17.6%)	20 (15.9%)	48 (18.5%)	
> 40	38 (9.9%)	22 (17.4%)	16 (6.1%)	
CEA				< 0.001
≤ 5 ug/L	327 (84.7%)	95 (75.4%)	232 (89.2%)	
> 5 ug/L	59 (15.3%)	31 (24.6%)	28 (10.8%)	
Surgical modalities				0.147
Wedge resection	12 (3.1%)	3 (2.4%)	9 (3.5%)	
Sublobectomy	21 (5.4%)	3 (2.4%)	18 (6.9%)	
Lobectomy	353 (91.5%)	120 (95.2%)	233 (89.6%)	
B. Histopathological characteristics				
Histological subtype				< 0.001
Lepidic	63 (16.3%)	5 (4.0%)	58 (22.3%)	
Acinar	185 (47.9%)	51 (40.5%)	134 (51.5%)	
Papillary	78 (20.2%)	24 (19.0%)	54 (20.8%)	
Micropapillary	28 (7.3%)	26 (20.6%)	2 (0.8%)	
Solid	32 (8.3%)	20 (15.9%)	12 (4.6%)	
Ki-67 LI* (%)	10.0 (5.0, 20.0)	10.8 (7.4, 30.0)	5.0 (3.0, 10.0)	< 0.001
Ki-67 LI				< 0.001
< 10%	193 (50.0%)	35 (27.8%)	158 (60.8%)	
≥ 10%	193 (50.0%)	91 (72.2%)	102 (39.2%)	
Visceral pleural invasion				< 0.001
Present	71 (18.4%)	36 (28.6%)	35 (13.5%)	
Absent	315 (81.6%)	90 (71.4%)	225 (86.5%)	
Lymph-vascular invasion				< 0.001
Present	58 (15.0%)	47 (37.3%)	11 (4.2%)	
Absent	328 (85.0%)	79 (62.7%)	249 (95.8%)	

(Continued)

TABLE 3 Continued

Characteristics	Training cohort (n=386)	STAS positive (n=126)	STAS negative (n=260)	*P* value
B. Histopathological characteristics				
Pathological T stage				< 0.001
T1a	48 (12.4%)	10 (7.9%)	38 (14.6%)	
T1b	170 (44.1%)	42 (33.3%)	128 (49.3%)	
T1c	85 (22.0%)	29 (23.1%)	56 (21.5%)	
T2	83 (21.5%)	45 (35.7%)	38 (14.6%)	
Pathological N stage				< 0.001
N0	328 (85.0%)	82 (65.1%)	246 (94.6%)	
N1	23 (6.0%)	17 (13.5%)	6 (2.3%)	
N2	35 (9.0%)	27 (21.4%)	8 (3.1%)	
C. CT Semantic characteristics				
Affiliated lobe				0.044
Upper lobe	236 (61.1%)	68 (54.0%)	168 (64.6%)	
Middle/lower lobe	150 (38.9%)	58 (46.0%)	92 (35.4%)	
Location				0.060
Central	47 (12.2%)	21 (16.7%)	26 (10.0%)	
Peripheral	339 (87.8%)	105 (83.3%)	234 (90.0%)	
Attenuation type				< 0.001
GGO	26 (6.7%)	4 (3.2%)	22 (8.5%)	
Sub-solid	208 (53.9%)	46 (36.5%)	162 (62.3%)	
Solid	152 (39.4%)	76 (60.3%)	76 (29.2%)	
Tumor total diameter (mm)*	22.0 (17.0, 27.0)	25.0 (19.0, 31.0)	21.0 (16.0, 26.0)	< 0.001
Tumor consolidation diameter (mm)*	15.5 (10.0, 23.0)	21.0 (15.8, 28.3)	13.0 (8.0, 20.0)	< 0.001
CTR* (%)	78.6 (46.5, 100.0)	100.0 (78.5,100.0)	64.1 (38.1, 100.0)	< 0.001
Shape				0.065
Round or oval	324 (83.9%)	112 (88.9%)	212 (81.5%)	
Irregular	62 (16.1%)	14 (11.1%)	48 (18.5%)	
Presence of obscure boundary	101 (26.2%)	53 (42.1%)	48 (18.5%)	< 0.001
Presence of lobulation	359 (93.0%)	121 (96.0%)	238 (91.5%)	0.105
Presence of spiculation	188 (48.7%)	76 (60.3%)	112 (43.1%)	0.001
Presence of cavity	54 (14.0%)	18 (14.3%)	36 (13.8%)	0.907
Presence of vacuole	29 (7.5%)	19 (15.1%)	10 (3.8%)	< 0.001
Presence of air bronchogram	200 (51.8%)	56 (44.4%)	144 (55.4%)	0.044
Presence of pleural attachment	118 (30.6%)	48 (38.1%)	70 (26.9%)	0.025

Unless otherwise stated, data were presented as numbers (percentages) and compared using the Chi-square test or Fisher's exact test.
*Data were presented as medians (inter-quartiles) and compared using the Mann-Whitney U test.
CEA, carcinoembryonic antigen; LI, labeling index; T, tumor; N, node; GGO, ground-glass opacity; CTR, consolidation-to-tumor ratio; STAS, spread through air space.

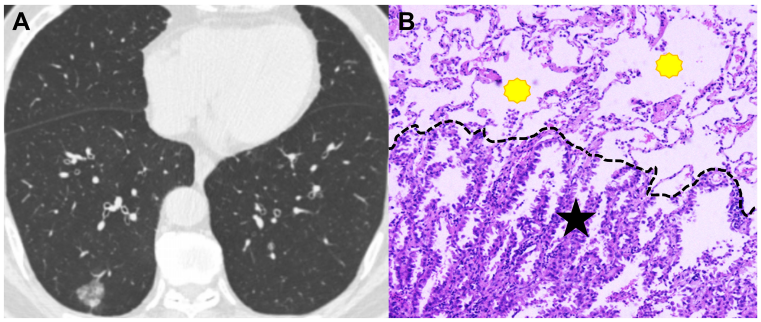


FIGURE 3
CT image and pathological image obtained from a 65-year-old man with spread though air spaces negative lung adenocarcinoma. **(A)** Th axial CT image (width, 1600 HU; level, -600 HU) shows a sub-solid nodule in the right lower lobe. **(B)** The photomicrograph of hematoxylin-eosin-stained histological section (magnification × 200) shows clean alveolar spaces (yellow polygon) beyond the boundary (dashed line) of the tumor (black star).

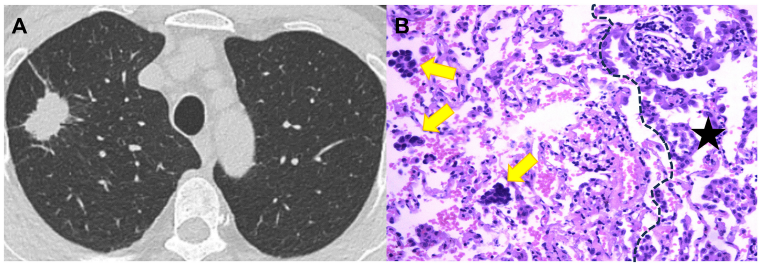


FIGURE 4
CT image and pathological image obtained from a 59-year-old woman with spread though air spaces positive lung adenocarcinoma. **(A)** Th axial CT image (width, 1600 HU; level, -600 HU) shows a solid nodule in the right upper lobe. **(B)** The photomicrograph of hematoxylin-eosin-stained histological section (magnification × 200) shows several solid nests of tumor cell (yellow arrow) beyond the boundary (dashed line) of the tumor (black star).

TABLE 4 The model performances in the training cohort and validation cohort.

Model	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
Training cohort					
CTR	0.709 (0.660,0.754)	0.706 (0.619, 0.784)	0.692 (0.632, 0.748)	0.527 (0.449, 0.604)	0.829 (0.773, 0.877)
Clinical-semantic model	0.764 (0.719,0.806)	0.778 (0.695, 0.847)	0.669 (0.608, 0.726)	0.533 (0.458, 0.606)	0.861 (0.806, 0.906)
Deep learning signature	0.869 (0.831,0.901)	0.706 (0.619, 0.784)	0.892 (0.848, 0.927)	0.761 (0.673, 0.835)	0.862 (0.815, 0.901)
Validation cohort					
CTR	0.734 (0.648,0.808)	0.689 (0.534, 0.818)	0.744 (0.636, 0.834)	0.596 (0.450, 0.731)	0.813 (0.707, 0.894)
Clinical-semantic model	0.714 (0.627,0.790)	0.778 (0.629, 0.888)	0.671 (0.558, 0.771)	0.565 (0.431, 0.691)	0.846 (0.735, 0.924)
Deep learning signature	0.837 (0.761,0.896)	0.578 (0.422, 0.723)	0.951 (0.880, 0.987)	0.867 (0.693, 0.962)	0.804 (0.711, 0.878)

CTR, consolidation-to-tumor ratio; AUC, area under the receiver operating characteristic curve; CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value.

highest AUC of 0.933 achieved so far in training cohort. Nevertheless, their model exhibited a substantial performance reduction in validation cohorts (AUC=0.783-0.806), which approximated the performance of our developed $Model_{ResNet\text{-}50}$ in training and validation cohorts (AUC=0.799-0.800)This unfavorable generalization may attribute to model overfitting by reason of complicated architecture (21). Lin et al. enrolled 581 patients with tumor smaller than 3 cm and CTR less than 0.5 from two institutions. They extracted the deep learning features from solid components and the entire tumors respectively, thereby developing deep learning models with and without solid component gate (SCG). The results revealed deep learning model with SCG achieved higher AUCs than deep learning model without SCG (22). Thus, further investigation is required to develop deep

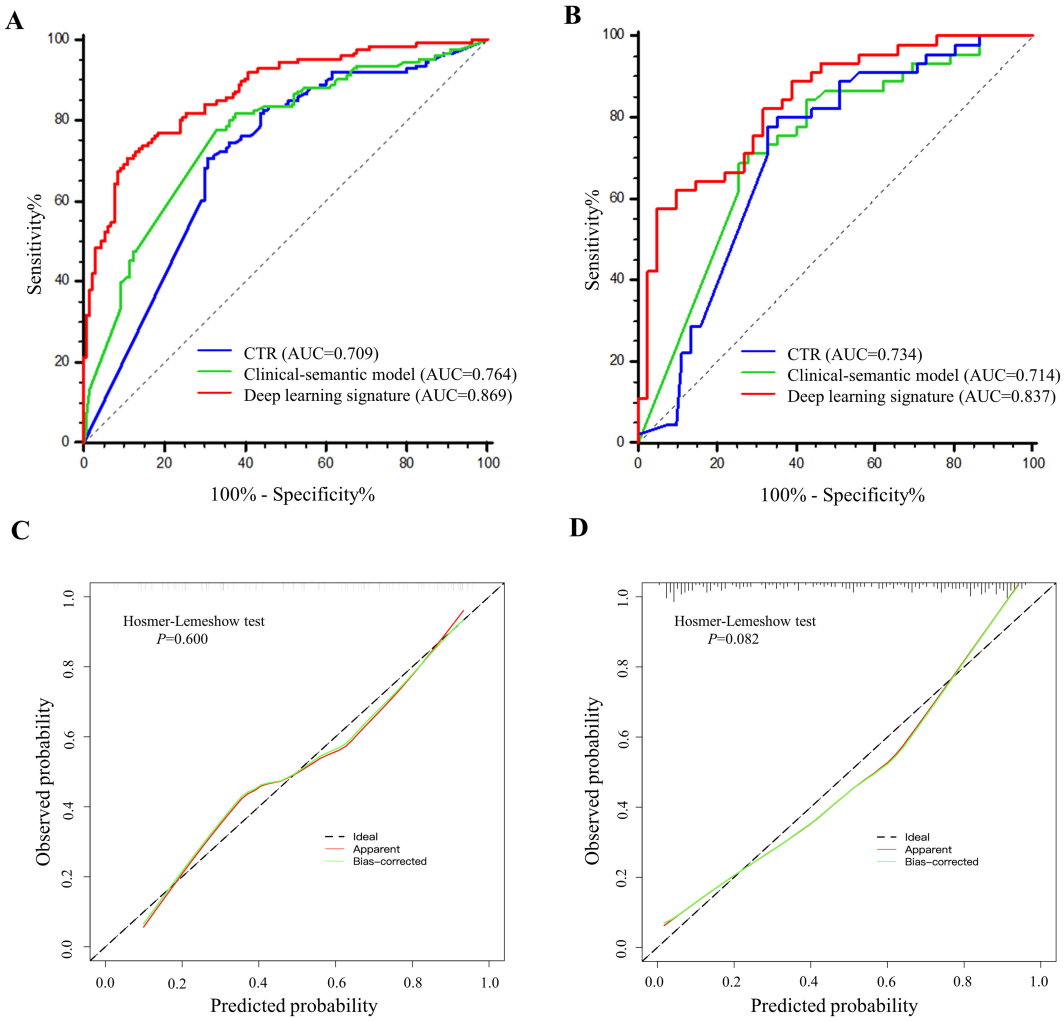


FIGURE 5
The performance comparisons of deep learning signature, CTR and clinical-semantic model in predicting STAS. **(A, B)** The receiver operating characteristic curves of CTR, clinical-semantic model and deep learning signature in training cohort **(A)** and validation cohort **(B)**. Number in parenthesis is the area under receiver operating characteristic curve. **(C, D)** The calibration curves depicted the good agreements between predicted probabilities by deep learning signature and actual observed probabilities of STAS in training cohort **(C)** and validation cohort **(D)**.

learning signature with SCG based on Swin Transformer, in expectation to further improve the prediction efficacy. In this study, Swin Transformer was adopted as the backbone architecture in modeling, achieving a satisfactory and comparable performance in training and validation cohorts with AUC ranging from 0.837 to 0.869, which was superior to $Model_{ResNet-50}$, $Model_{EfficientNet}$ and $Model_{ConvNeXt}$. This finding lent support to the potential of our proposed Swin Transformer in predicting STAS in lung adenocarcinoma. The state-of-the-art Swin Transformer is regarded as the new backbone of machine vision. With two key strengths of non-overlapping shifted windows and hierarchical structures, Swin Transformer can flexibly process images at various scales and reduce computational complexity from the exponential level to the linear level. Growing evidence validated the efficient processing capabilities of Swin Transformer in handling multitasking such as image classification and density detection (23–25). Our previously published study has affirmed the remarkable efficiency of Swin Transformer in predicting lymph node metastasis in lung adenocarcinoma (26). Aside from that, automatic tumor segmentation was conducted in this study using a 3D U-shape convolutional neural network. This deep learning architecture serves as a highly effective tool for accurate, robust, and efficient segmentation. It surpasses the time-consuming and labor-intensive manual delineation or semi-automated segmentation, as evidenced by the Dice similarity coefficients across multiple institutions (27, 28).

Further exploring the relationship between STAS and histopathological factors, micropapillary and solid predominant adenocarcinoma were more commonly observed in STAS. Our findings demonstrated a significant association between STAS and visceral pleural invasion, lympho-vascular invasion and higher pathological T stage, consistent with previous literature (29). Additionally, lymph node invasion was more frequently found in STAS-positive subgroup (34.9% vs 5.4%). In line with our results, Vaghjiani et al. also reported that STAS was an independent predictor of occult lymph node metastasis in clinical stage IA

lung adenocarcinoma (30). Although the underlying mechanism of STAS remains unclear till now, it has been found that epithelial-mesenchymal transition (EMT) prominently promotes the occurrence of STAS (31). EMT is widely recognized as a biological process wherein polarized epithelial cells transform into loosely connected interstitial cells; this process is regarded as the key driver of tumor genesis, invasion and metastasis. This may account for the strong association between STAS and the aforementioned invasive histopathological factors.

In clinical-semantic model, tumor boundary, vacuolation and CTR were the independent CT semantic features in predicting STAS. As a reflection of tumor aggressiveness, CTR weighted heavily in regression analysis with a 1.25-fold increased risk of STAS for every 10% increase. In accordance with our finding, Ding et al. and Chen et al. confirmed that CTR was independently associated with STAS (32, 33). Unexpectedly, the inclusion of all clinical-semantic risk predictors failed to show an incremental value with respect to deep learning signature. We found a strong correlation between CTR and deep learning signature ($r = 0.789$, $P < 0.001$), which might account for this result. These findings also lead to speculation on whether deep learning signature contains biological information regarding tumor boundary and vacuolation, which should be explored by future in-depth research. We also found CTR and clinical-semantic model showed equivalent NPV and sensitivity to deep learning signature. Notably, in both training and validation cohorts, the deep learning signature exhibited far superior AUC, specificity, and PPV compared to CTR and the clinical-semantic model, which lent support to its predominant efficacy in predicting STAS.

There are some limitations to this study. First, data were retrospectively collected from different CT equipment, so heterogeneity in acquisition parameters and reconstruction protocols might be inevitable. The class-imbalance in sample should be addressed using resample techniques in the future. Second, it is necessary to expand sample size and enroll multi-institutional data to further affirm the repeatability and generalization of deep learning signature. Besides, long-term follow up and survival data should be warranted to affirm the prognostic value of STAS, as well as the relationship of deep learning signature with prognosis. Further investigation is required to enhance the biological interpretability of deep learning, which inherently possesses a black box nature, thereby facilitating its application in clinical practice. Common approaches involve employing the Grad-CAM algorithm for generating visualizations of deep learning and incorporating attentional mechanisms into deep learning networks to achieve the significance weight of diagnosis and decision-making based on attention regions. Additionally, exploring the associations between deep learning and genomics or proteomics can further improve the biological interpretability of deep learning. Last, given that biological behavior varies in different histological subtypes of lung cancer, future research needs to supplement the predictive value of the deep learning signature for STAS in other histological subtypes.

In conclusion, the proposed deep learning signature based on Swin Transformer offers a promising predictive performance for STAS in clinical stage I lung adenocarcinoma, surpassing the conventional clinical-semantic model. The end-to-end deep learning approach harbors the potential as a well-established tool for noninvasive estimation of STAS, directing surgical strategy and facilitating adjuvant therapeutic scheduling.

Data availability statement

I'm regretful that the generated dataset is inappropriate to be disclosed so far, because some work based on parts of this dataset have not yet published. Requests to access the datasets should be directed to maxiaoling0417@163.com.

Ethics statement

The studies involving humans were approved by Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology. The studies were conducted in accordance with the local legislation and institutional requirements. The ethics committee/institutional review board waived the requirement of written informed consent for participation from the participants or the participants' legal guardians/next of kin because Written informed consent was waived because of retrospective study.

Author contributions

XM: Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing, Conceptualization. WH: Conceptualization, Data curation, Methodology, Writing – review & editing, Validation, Visualization. CC: Conceptualization, Writing – review & editing, Data curation, Resources, Software, Visualization. FT: Conceptualization, Methodology, Writing – review & editing, Formal analysis, Investigation. JC: Conceptualization, Software, Visualization, Writing – review & editing, Methodology, Validation. LY: Formal analysis, Investigation, Methodology, Writing – review & editing. DC: Conceptualization, Methodology, Validation, Writing – review & editing, Formal analysis, Funding acquisition, Project administration, Supervision. LX: Formal Analysis, Investigation, Methodology, Writing – review & editing, Conceptualization, Data curation, Project administration, Resources, Supervision.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study has received funding from the Ningxia Natural Science Foundation (2024AAC03483 and 2024AAC03457), and the Pilot Experiment Project of the National Natural Science Foundation of China (2025GZRYSY003).

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/fonc.2024.1482965/full#supplementary-material

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. (2024) 74:229–63. doi: 10.3322/caac.21834

2. Yan R, Fan X, Xiao Z, Liu H, Huang X, Liu J, et al. Inhibition of DCLK1 sensitizes resistant lung adenocarcinomas to EGFR-TKI through suppression of Wnt/β-Catenin activity and cancer stemness. *Cancer Lett*. (2022) 531:83–97. doi: 10.1016/j.canlet.2022.01.030

3. Kadota K, Nitadori JI, Sima CS, Ujiie H, Rizk NP, Jones DR, et al. Tumor Spread through Air Spaces is an Important Pattern of Invasion and Impacts the Frequency and Location of Recurrences after Limited Resection for Small Stage I Lung Adenocarcinomas. *J Thorac Oncol*. (2015) 10:806–14. doi: 10.1097/jto.0000000000000486

4. Laville D, Désage AL, Fournel P, Bayle-Bleuez S, Neifer C, Picot T, et al. Spread through air spaces in stage I to III resected lung adenocarcinomas: should the presence of spread through air spaces lead to an upstaging? *Am J Surg Pathol*. (2024) 48:596–604. doi: 10.1097/pas.0000000000002188

5. Si H, Xu L, Zhao Y, Su H, Dai C, Xie H, et al. Spread through air spaces in residual tumor classification for clinical IA lung adenocarcinoma. *Ann Thorac Surg*. (2024) 118:825–33. doi: 10.1016/j.athoracsur.2024.03.007

6. Chen X, Zhou H, Wu M, Xu M, Li T, Wang J, et al. Prognostic impact of spread through air spaces in patients with ≤2 cm stage IA lung adenocarcinoma. *J Thorac Dis*. (2024) 16:2432–42. doi: 10.21037/jtd-24-444

7. Lv Y, Li S, Liu Z, Ren Z, Zhao J, Tao G, et al. Impact of surgery and adjuvant chemotherapy on the survival of stage I lung adenocarcinoma patients with tumor spread through air spaces. *Lung Cancer*. (2023) 177:51–8. doi: 10.1016/j.lungcan.2023.01.009

8. Villalba JA, Shih AR, Sayo TMS, Kunitoki K, Hung YP, Ly A, et al. Accuracy and reproducibility of intraoperative assessment on tumor spread through air spaces in stage 1 lung adenocarcinomas. *J Thorac Oncol*. (2021) 16:619–29. doi: 10.1016/j.jtho.2020.12.005

9. Suh JW, Jeong YH, Cho A, Kim DJ, Chung KY, Shim HS, et al. Stepwise flowchart for decision making on sublobar resection through the estimation of spread through air space in early stage lung cancer(1). *Lung Cancer*. (2020) 142:28–33. doi: 10.1016/j.lungcan.2020.02.001

10. Qin L, Sun Y, Zhu R, Hu B, Wu J. Clinicopathological and CT features of tumor spread through air space in invasive lung adenocarcinoma. *Front Oncol*. (2022) 12:959113. doi: 10.3389/fonc.2022.959113

11. Liao G, Huang L, Wu S, Zhang P, Xie D, Yao L, et al. Preoperative CT-based peritumoral and tumoral radiomic features prediction for tumor spread through air spaces in clinical stage I lung adenocarcinoma. *Lung Cancer*. (2022) 163:87–95. doi: 10.1016/j.lungcan.2021.11.017

12. Chen LW, Lin MW, Hsieh MS, Yang SM, Wang HJ, Chen YC, et al. Radiomic values from high-grade subtypes to predict spread through air spaces in lung adenocarcinoma. *Ann Thorac Surg*. (2022) 114:999–1006. doi: 10.1016/j.athoracsur.2021.07.075

13. Wulaningsih W, Villamaria C, Akram A, Benemile J, Croce F, Watkins J. Deep learning models for predicting Malignancy risk in CT-detected pulmonary nodules: A systematic review and meta-analysis. *Lung*. (2024) 202:625–36. doi: 10.1007/s00408-024-00706-1

14. Peng J, Xie B, Ma H, Wang R, Hu X, Huang Z. Deep learning based on computed tomography predicts response to chemoimmunotherapy in lung squamous cell carcinoma. *Aging Dis*. (2024). doi: 10.14336/ad.2024.0169

15. Na KJ, Kim YT, Goo JM, Kim H. Clinical utility of a CT-based AI prognostic model for segmentectomy in non-small cell lung cancer. *Radiology*. (2024) 311:e231793. doi: 10.1148/radiol.231793

16. Zhang B, Liu R, Ren D, Li X, Wang Y, Huo H, et al. Comparison of lobectomy and sublobar resection for stage IA elderly NSCLC patients (≥70 years): A population-based propensity score matching's study. *Front Oncol*. (2021) 11:610638. doi: 10.3389/fonc.2021.610638

17. Masai K, Sakurai H, Sukeda A, Suzuki S, Asakura K, Nakagawa K, et al. Prognostic impact of margin distance and tumor spread through air spaces in limited resection for primary lung cancer. *J Thorac Oncol*. (2017) 12:1788–97. doi: 10.1016/j.jtho.2017.08.015

18. Dai C, Xie H, Su H, She Y, Zhu E, Fan Z, et al. Tumor Spread through Air Spaces Affects the Recurrence and Overall Survival in Patients with Lung Adenocarcinoma >2 to 3 cm. *J Thorac Oncol*. (2017) 12:1052–60. doi: 10.1016/j.jtho.2017.03.020

19. Chen D, Wang X, Zhang F, Han R, Ding Q, Xu X, et al. Could tumor spread through air spaces benefit from adjuvant chemotherapy in stage I lung adenocarcinoma? A multi-institutional study. *Ther Adv Med Oncol*. (2020) 12:1758835920978147. doi: 10.1177/1758835920978147

20. Tao J, Liang C, Yin K, Fang J, Chen B, Wang Z, et al. 3D convolutional neural network model from contrast-enhanced CT to predict spread through air spaces in non-small cell lung cancer. *Diagn Interv Imaging*. (2022) 103:535–44. doi: 10.1016/j.diii.2022.06.002

21. Wang S, Liu X, Jiang C, Kang W, Pan Y, Tang X, et al. CT-based super-resolution deep learning models with attention mechanisms for predicting spread through air spaces of solid or part-solid lung adenocarcinoma. *Acad Radiol*. (2024) 31:2601–9. doi: 10.1016/j.acra.2023.12.034

22. Lin MW, Chen LW, Yang SM, Hsieh MS, Ou DX, Lee YH, et al. CT-based deep-learning model for spread-through-air-spaces prediction in ground glass-predominant lung adenocarcinoma. *Ann Surg Oncol*. (2024) 31:1536–45. doi: 10.1245/s10434-023-14565-2

23. Li L, Mei Z, Li Y, Yu Y, Liu M. A dual data stream hybrid neural network for classifying pathological images of lung adenocarcinoma. *Comput Biol Med*. (2024) 175:108519. doi: 10.1016/j.compbiomed.2024.108519

24. Lo CM, Wang CC, Hung PH. Interactive content-based image retrieval with deep learning for CT abdominal organ recognition. *Phys Med Biol*. (2024) 69. doi: 10.1088/1361-6560/ad1f86

25. Li C, Bagher-Ebadian H, Sultan R, Elshaikh M, Movsas B, Zhu D, et al. A new architecture combining convolutional and transformer-based networks for automatic 3D multi-organ segmentation on CT images. *Med Phys*. (2023) 50:6990–7002. doi: 10.1002/mp.16750

26. Ma X, Xia L, Chen J, Wan W, Zhou W. Development and validation of a deep learning signature for predicting lymph node metastasis in lung adenocarcinoma: comparison with radiomics signature and clinical-semantic model. *Eur Radiol*. (2023) 33:1949–62. doi: 10.1007/s00330-022-09153-z

27. Carles M, Kuhn D, Fechter T, Baltas D, Mix M, Nestle U, et al. Development and evaluation of two open-source nnU-Net models for automatic segmentation of lung tumors on PET and CT images with and without respiratory motion compensation. *Eur Radiol*. (2024) 34:6701–11. doi: 10.1007/s00330-024-10751-2

28. Park J, Kang SK, Hwang D, Choi H, Ha S, Seo JM, et al. Automatic lung cancer segmentation in [(18)F]FDG PET/CT using a two-stage deep learning approach. *Nucl Med Mol Imaging*. (2023) 57:86–93. doi: 10.1007/s13139-022-00745-7

29. Shih AR, Mino-Kenudson M. Updates on spread through air spaces (STAS) in lung cancer. *Histopathology*. (2020) 77:173–80. doi: 10.1111/his.14062

30. Vaghjiani RG, Takahashi Y, Eguchi T, Lu S, Kameda K, Tano Z, et al. Tumor spread through air spaces is a predictor of occult lymph node metastasis in clinical stage IA lung adenocarcinoma. *J Thorac Oncol.* (2020) 15:792–802. doi: 10.1016/j.jtho.2020.01.008

31. Niu Y, Han X, Zeng Y, Nanding A, Bai Q, Guo S, et al. The significance of spread through air spaces in the prognostic assessment model of stage I lung adenocarcinoma and the exploration of its invasion mechanism. *J Cancer Res Clin Oncol.* (2023) 149:7125–38. doi: 10.1007/s00432-023-04619-z

32. Ding Y, Chen Y, Wen H, Li J, Chen J, Xu M, et al. Pretreatment prediction of tumour spread through air spaces in clinical stage I non-small-cell lung cancer. *Eur J Cardiothorac Surg.* (2022) 62:ezac248. doi: 10.1093/ejcts/ezac248

33. Chen Y, Jiang C, Kang W, Gong J, Luo D, You S, et al. Development and validation of a CT-based nomogram to predict spread through air space (STAS) in peripheral stage IA lung adenocarcinoma. *Jpn J Radiol.* (2022) 40:586–94. doi: 10.1007/s11604-021-01240-3

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 01 August 2024
DOI 10.3389/fimmu.2024.1417156

OPEN ACCESS

EDITED BY
Sandeep Kumar Mishra,
Yale University, United States

REVIEWED BY
Mayra Mejia,
Instituto Nacional de Enfermedades Respiratorias, Mexico
Alarico Ariani,
University Hospital of Parma, Italy

*CORRESPONDENCE
Xiaoli Gu
Hu906@163.com
Lin Jin
jinlin205@163.com

†These authors have contributed equally to this work

RECEIVED 14 April 2024
ACCEPTED 22 July 2024
PUBLISHED 01 August 2024

CITATION
Han N, Guo Z, Zhu D, Zhang Y, Qin Y, Li G, Gu X and Jin L (2024) A nomogram model combining computed tomography-based radiomics and Krebs von den Lungen-6 for identifying low-risk rheumatoid arthritis-associated interstitial lung disease. *Front. Immunol.* 15:1417156. doi: 10.3389/fimmu.2024.1417156

A nomogram model combining computed tomography-based radiomics and Krebs von den Lungen-6 for identifying low-risk rheumatoid arthritis-associated interstitial lung disease

Nie Han[1†], Zhinan Guo[1†], Diru Zhu[2†], Yu Zhang[2], Yayi Qin[3], Guanheng Li[1], Xiaoli Gu[2*] and Lin Jin[1*]

[1]Department of Ultrasound, Guanghua Hospital Affiliated to Shanghai University of Traditional Chinese Medicine, Shanghai, China, [2]Department of Radiology, Guanghua Hospital Affiliated to Shanghai University of Traditional Chinese Medicine, Shanghai, China, [3]Department of Pulmonary Function, Guanghua Hospital Affiliated to Shanghai University of Traditional Chinese Medicine, Shanghai, China

Objectives: Quantitatively assess the severity and predict the mortality of interstitial lung disease (ILD) associated with Rheumatoid arthritis (RA) was a challenge for clinicians. This study aimed to construct a radiomics nomogram based on chest computed tomography (CT) imaging by using the ILD-GAP (gender, age, and pulmonary physiology) index system for clinical management.

Methods: Chest CT images of patients with RA-ILD were retrospectively analyzed and staged using the ILD-GAP index system. The balanced dataset was then divided into training and testing cohorts at a 7:3 ratio. A clinical factor model was created using demographic and serum analysis data, and a radiomics signature was developed from radiomics features extracted from the CT images. Combined with the radiomics signature and independent clinical factors, a nomogram model was established based on the Rad-score and clinical factors. The model capabilities were measured by operating characteristic curves, calibration curves and decision curves analyses.

Results: A total of 177 patients were divided into two groups (Group I, n = 107; Group II, n = 63). Krebs von den Lungen-6, and nineteen radiomics features were used to build the nomogram, which showed favorable calibration and discrimination in the training cohort [AUC, 0.948 (95% CI: 0.910–0.986)] and the testing validation cohort [AUC, 0.923 (95% CI: 0.853–0.993)]. Decision curve analysis demonstrated that the nomogram performed well in terms of clinical usefulness.

Conclusion: The CT-based radiomics nomogram model achieved favorable efficacy in predicting low-risk RA-ILD patients.

KEYWORDS

computed tomography, radiomics, KL-6, rheumatoid arthritis, interstitial lung disease

1 Introduction

Rheumatoid arthritis (RA) is one of the most immune-mediated diseases that affects 0.5–1% of the global population. It is primarily characterized by joint swelling and tenderness, leading to the destruction of synovial joints (1). Beyond the joints, RA is associated with systemic inflammation that can result in multiple coexisting conditions and extra-articular manifestations (2). Pulmonary involvement is recognized as the most prevalent extra-articular complication in RA, encompassing a broad range of disorders such as airway diseases, pleural effusions, and rheumatoid nodules (3–5). Among these pulmonary complications, interstitial lung disease (ILD) has the highest prevalence (6). Importantly, RA-ILD is a significant cause of mortality among RA patients and contributes to considerable morbidity (7). Consequently, accurately assessing mortality risk associated with RA-ILD is of great clinical significance.

The ILD-GAP (gender, age, and pulmonary physiology) index, initially proposed by Ley et al. in 2012 (8), is a simple scoring system designed to predict mortality risk in patients with idiopathic pulmonary fibrosis. Utilizing variables such as gender, age, predicted forced vital capacity (FVC), and diffusion capacity for carbon monoxide (DL_{CO}), which has been refined and validated for various types of ILD (9). Its accuracy in predicting outcomes for RA-ILD has been confirmed by multiple studies (10–12). However, pulmonary function tests (PFTs) necessitate active participation from patients, such as performing deep breaths or forceful exhalations (13). This can be particularly challenging for special populations, including those with cognitive impairments or concurrent pulmonary conditions, potentially compromising the precision of the test results. To our knowledge, there is an absence of universal, quantitative, non-invasive techniques for the staging of RA-ILD.

The current primary method for diagnosing RA-ILD remains Computed Tomography (CT) scan, owing to its noninvasive and sensitive nature in detecting lung involvement (14, 15). However, there are many features to determine the presence of ILD and inter-reader variability, especially in unexperienced readers, is an issue (16). Visual analysis of ILDs on CT images faces difficulties in providing prognosis information, as various stages of RA-ILD exhibit overlapping imaging features, making the diagnosis and assessment of severity challenging with conventional imaging modalities (17, 18). Radiomics technology, capable of extracting numerous high-dimensional features from CT images, emerges as a solution to address the limitations of visual assessment. Although radiomics has predominantly been explored in the context of various tumors (19, 20), its potential has been demonstrated in identifying the GAP staging of connective tissue disease-associated interstitial lung disease (CTD-ILD) (21, 22). However, ILD associated with different CTDs can be characterized by distinct clinical manifestations, imaging, and pathological features, indicating their unique developmental and regression patterns. In the context of RA-ILD, evidence from a small cohort study suggested that radiomics may hold the potential for predicting mortality (23). However, limited studies are focusing on the application of radiomics in the staging of RA-ILD. Therefore, it is still necessary to explore the discriminative value of radiomics in various stages of RA-ILD.

In this retrospective study, we aimed to establish a novel CT-based radiomics nomogram to differentiate between the different stages of RA-ILD.

2 Materials and methods

2.1 Patients

The study included patients clinically diagnosed with RA-ILD between April 2020 and December 2023 at Guanghua Hospital Affiliated with Shanghai University of Traditional Chinese Medicine. Inclusion criteria comprised patients meeting all of the following conditions: 1) diagnosed with RA according to the 2010 American College of Rheumatology criteria for RA (24); 2) diagnosed with ILD according to the American Thoracic Society, European Respiratory Society, Japanese Respiratory Society, and Latin American Thoracic Society (ATS/ERS/JRS/ALAT) criteria for ILD (25); 3) underwent a CT scan showing signs of ILD within 3 months after clinical diagnosis; and 4) underwent pulmonary function tests and laboratory examination within 30 days before or after the CT scan. Exclusion criteria were applied for patients meeting any of the following conditions: 1) those with pulmonary edema, infection, drug toxicity, allergy tumor, or heart disease; 2) diagnosed with a combination of other types of CTD; 3) incomplete demographic or clinical data. The flowchart of the study subjects is shown in Figure 1.

2.2 Pulmonary function test

The routine PFTs were conducted using the Master Screen Diffusion Pulmonary Function Instrument (Eric Jaeger, Germany). The following indicators were assessed: the percentage predicted values (% predicted) of forced expiratory volume in 1 s (FEV1), FVC, total lung capacity (TLC), and DL_{CO}. The ILD-GAP index was calculated in accordance with the method proposed by Ryerson et al. (9). Subsequently, patients were categorized into two groups: Group I comprised patients with an ILD-GAP index ≤1, while Group II included patients with an ILD-GAP index >1.

2.3 CT image acquisition and evaluation

All enrolled patients underwent nonenhanced chest CT examinations using one of two multidetector CT systems: SOMATOM Definition Flash (Siemens Healthcare, Tokyo, Japan) or Access CT (Philips Healthcare, Andover, Massachusetts, USA). The parameters used for CT scanning were as follows: tube voltage of 120 kVp and tube current-time product of 60-220 mAs with automatic dose modulation; detector collimation of 64 × 0.6 mm; rotation time of 1.0 second; and matrix size of 512 × 512. All CT scans were reconstructed with a 1-mm slice thickness and lung convolution kernels. The semiquantitative CT (SQCT) assessment

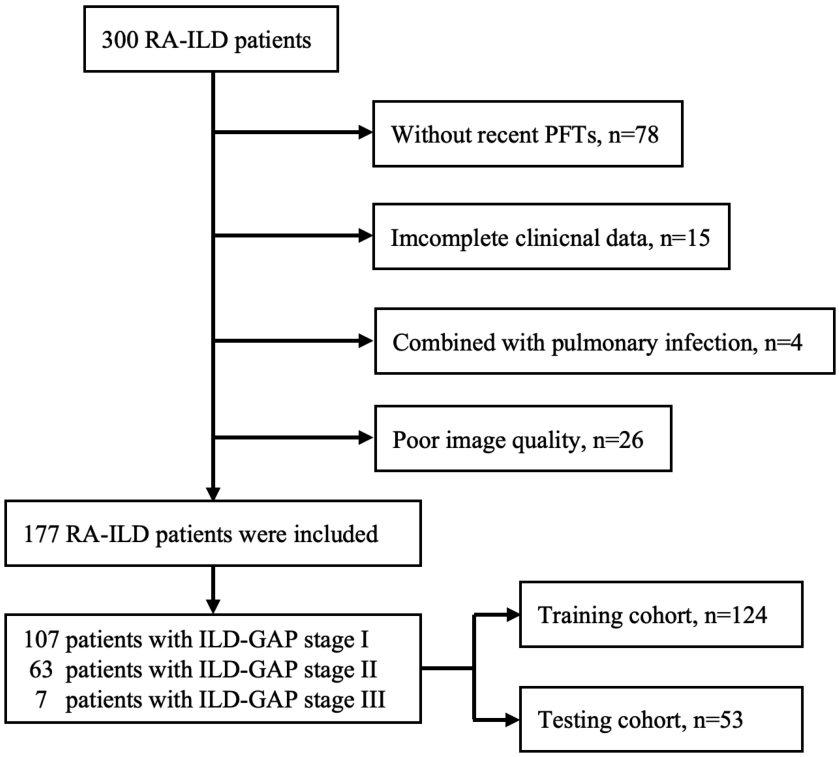


FIGURE 1
Flowchart of the patient cohort.

was carried out to calculate Goh score for each CT scan (26). RA-ILD findings from HRCT were classified as UIP or non-UIP patterns following recent IPF guidelines (25).

2.4 Three-dimensional lung segmentation

All image segmentation was executed using 3D Slicer software (version 5.6.1, www.slicer.org). The preprocessing steps were carried out as follows: 1) All CT images were reprocessed using the "Resample Scalar Volume" module by resampling them into 1-mm thick slices and normalizing the intensity values within the range of [−1, 1]. 2) Using the "Radiomics" module, the voxel intensity values were discretized with a fixed bin width of 25 HU to reduce noise and standardize intensity across the images. 3) Z-score normalization was performed on the image gray values to reduce the impact of inconsistent imaging parameters on the variability of radiomics features. 4) The region of interest (ROI) of the bilateral lungs was automatically segmented, encompassing blood vessels and the trachea in the lung lobes (window width = 1,250; window level = -875). A threshold-based region growing method was utilized. The seeding strategy involved the placement of a total of 13 seed points across different anatomical planes. On the axial plane, three seed points were positioned in the peripheral regions of the left and right lungs, respectively. A similar approach was adopted on the coronal plane. Additionally, one seed point was positioned at the location of the main bronchus. Subsequently, the segmentation results underwent manual correction by a radiologist with 5 years of experience in imaging diagnosis of chest diseases, and confirmation was obtained from another radiologist with 8 years of experience in imaging diagnosis of chest diseases.

Interclass and intraclass correlation coefficients (ICCs) were employed in the following manner: A total of 20 cases were randomly selected for region of interest (ROI) segmentation by Radiologist 1. Radiologist 2 then replicated the segmentation for these 20 cases. Subsequently, Radiologist 1 repeated the segmentation after a one-month interval. The segmentation was deemed well-matched in terms of interobserver reliability and intraobserver reproducibility when the ICC value surpassed 0.75.

2.5 Radiomics feature extraction and model establishment

Figure 2 shows the workflow of radiomics analysis in this study. The patient cohort was randomly split into training and test cohorts at a ratio of 7:3. Feature extraction was performed utilizing the open-source Pyradiomics software package (http://pypi.org/project/pyradiomics/). This package facilitates the extraction of a comprehensive suite of radiomics features, categorized into seven distinct classes: Gray Level Dependence Matrix (GLDM), Gray Level Co-occurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM), Neighboring Gray Tone Difference Matrix (NGTDM), First Order Statistics, and Shape-based features (3D). A detailed description of the extracted features is accessible via the Pyradiomics

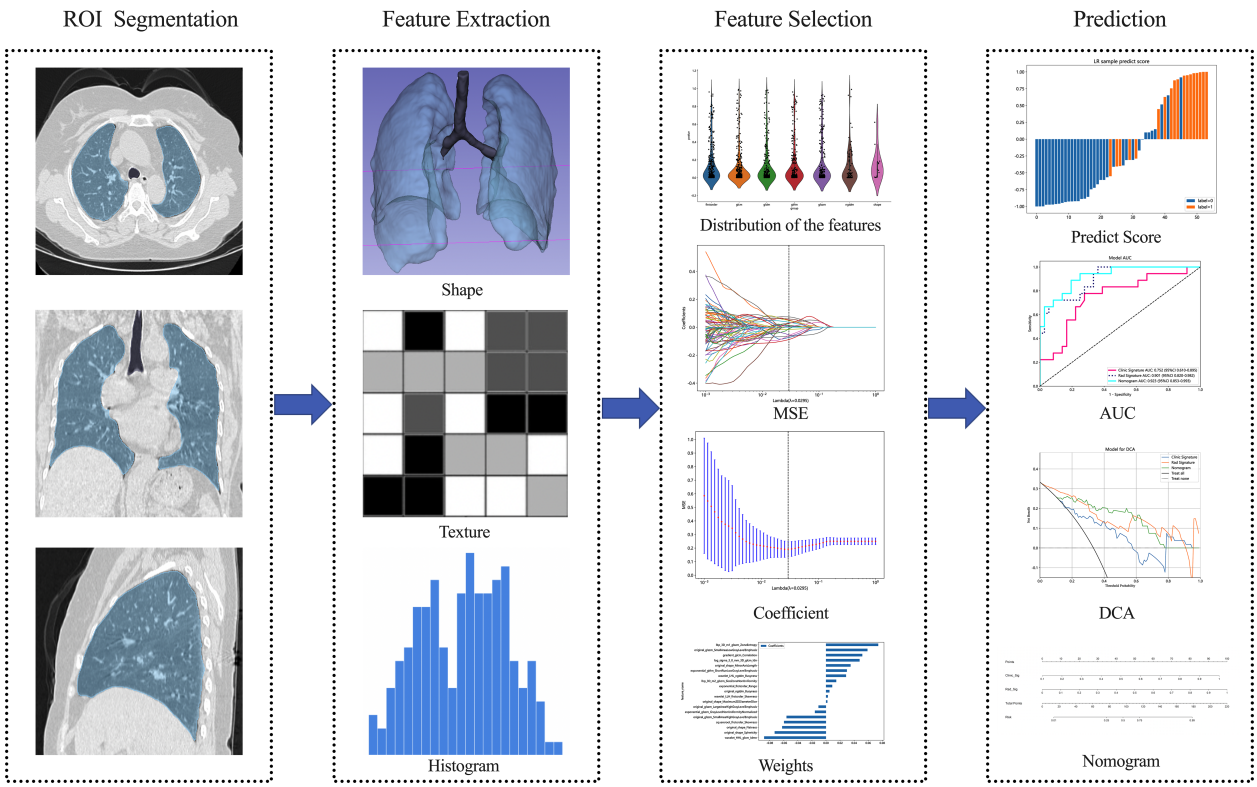


FIGURE 2
Overview workflow of radiomics analysis. Semi-automatic segmentation of the whole lung was performed on CT images, followed by manual adjustment of the confirmed dissection range, with the region of interest delineated in blue. Imaging-derived histologic features, including shape and texture characteristics, were extracted from CT images of both lungs. Feature selection was conducted using inter- and intra-observer reliability assessment as well as the LASSO method. The performance and clinical utility of predictive model were evaluated using ROC, DCA, and nomogram analysis. MSE Mean standard error, ROC Receiver operating characteristic, DCA Decision curve analysis.

documentation (http://pyradiomics.readthedocs.io). A total of 1,834 radiomics features were extracted from the ROIs. Statistical analysis involved the Student's t-test for normally distributed features and the Mann-Whitney U test for others. Features with a p-value ≤ 0.05 were retained, resulting in 1,171 features. Spearman's rank correlation coefficient was then applied to identify robustly repeatable features, retaining one feature from pairs with a correlation coefficient > 0.75. A recursive elimination strategy further refined the features to a subset of 102. The dataset's signature was constructed using the least absolute shrinkage and selection operator (LASSO) regression model. The optimal λ value was determined via tenfold cross-validation. Features with non-zero coefficients formed the Radiomics Signature, combining linearly to compute the radiomics score for each patient. Scikit-learn in Python was employed for LASSO regression, and logistic regression was used for model formulation after 10-fold cross-validation to verify model adequacy.

2.6 Construction of the clinical model

The clinical factor model incorporated variables that were significantly different ($p < 0.05$) as determined by univariate logistic regression analysis. These variables included clinical data and laboratory examinations from the training cohort. Odds ratios (ORs) with 95% confidence intervals (CIs) were calculated for the significantly correlated variables. To mitigate the risk of data leakage within the models, gender, age, and PFT parameters were excluded.

2.7 The building of the clinical-radiomics nomogram

A multivariate logistic regression analysis, combining both the clinical signature and radiomics signature, was employed in a backward step-down selection procedure to develop the final integrated radiomics-clinical prediction model.

2.8 Statistical analysis

Statistical analyses were performed using SPSS (version 26.0; IBM Corp.). Statistical significance was defined as a two- sided p-value ≤ 0.05. Normally distributed data were analyzed using independent T-tests, and non-normally distributed data were presented as medians (interquartile range) using Mann-Whitney U tests. Categorical variables were analyzed using chi-square tests.

The predictive performance of the three models was evaluated using receiver operating characteristic (ROC) curves, with the area under the ROC curve (AUC) calculated. Model performance was tested in both the training and test cohorts. The Delong test was applied to compare AUCs among the three models. Calibration efficiency of the nomogram was assessed through calibration curves, and the Hosmer–Lemeshow analytical fit was used to evaluate calibration ability. Decision curve analysis (DCA) was employed to evaluate the clinical utility of the radiomics-clinical model.

3 Results

3.1 Patient characteristics

A total of 177 patients with RA-ILD were enrolled in this study. Among these patients, 107, 63, and 7 were allocated to ILD-GAP stage I, II, and III, respectively. To prevent excessive data bias, the patients in ILD-GAP stage II and III were combined into a single group. Table 1 listed the baseline patient characteristics in group I and group II. Age, gender, FVC, FEV1, TLC, DL_{CO}, and serum Krebs von den Lungen-6 (KL-6) level showed significant differences ($p < 0.05$) between the two groups, while the differences in smoking history, ACPA, RF-IgM, RF-IgA, and RF-IgG were not significant ($p > 0.05$). In addition, there was no significant statistical difference between the two groups in terms of ESR, CRP, TNFα, IFNγ, IFNα, as well as disease activity score ($p > 0.05$).

3.2 Development of the clinical model

Univariate logistic regression was performed to analyze the clinical data and laboratory examinations (Table 2). To ensure the reliability of the model construction, factors such as gender, age, and PFT parameters were excluded. Then, KL-6 (ORs = 1.007; 95% CI, 1.004-1.010; $p < 0.001$) was selected as independent clinical risk factors.

3.3 Development of the radiomics model

A total of 1,834 radiomics features were extracted from the CT images, with 1,171 exhibiting promising interobserver and intraobserver agreement (intraclass correlation coefficient > 0.75). Through LASSO logistic regression analysis, 102 significantly different ($p < 0.05$) radiomics features were selected to identify optimally related features. Ultimately, 19 features were included in the construction of the radiomics model. Figures 3A,B show the coefficients and mean standard error (MSE) for the 10-fold validation, while Figure 3C presents the coefficient values for the final selection of non-zero features Rad score is shown as follows: Rad-score= 0.4227 + 0.0088 × exponential_firstorder_Range +0.0296 × exponential_glrlm_ShortRunLowGrayLevelEmphasis -0.0157 × exponential_glszm_GrayLevelNonUniformity Normalized +0.0516 × gradient_glcm_Correlation +0.0743 × lbp_3D_m1_glszm_ZoneEntropy +0.0146 × lbp_3D_m2_glszm_

TABLE 1 Patient characteristics.

Variables	Group I (n=107)	Group II (n=70)	*p* value
Female (%)	91(85.05%)	40(57.14%)	<0.001
Age, years	58.8 ± 8.9	71.5 ± 5.5	<0.001
RA duration, years	10.00 [4.00-9.25]	11.00 [4.00-20.00]	0.084
Smoking (%)	7(6.54%)	5(7.14%)	0.876
Lung function			
FVC%	86.5 ± 18.1	66.3 ± 17.4	<0.001
FEV1%	86.2 ± 18.0	67.2 ± 17.0	<0.001
TLC%	83.6 ± 15.7	56.8 ± 14.6	<0.001
DL_{CO}%	61.5 ± 17.7	32.2 ± 13.4	<0.001
Laboratory Examinations			
ACPA, RU/ml	653.90 [240.30-1249.50]	582.10 [138.75-1364.88]	0.782
RF-IgA, U/ml	32.77 [8.22-300.00]	28.18 [6.43-146.40]	0.496
RF-IgG, U/ml	30.01 [6.11-96.76]	40.50 [4.23-136.63]	0.957
RF-IgM, U/ml	127.00 [33.90-369.00]	135.00 [40.25-574.00]	0.590
TNFα, pg/ml	2.56 [1.68-2.67]	2.00 [1.36-2.56]	0.075
IFNγ, pg/ml	2.46 [2.27-5.65]	2.46 [1.82-5.05]	0.745
IFNα, pg/ml	1.36 [0.95-2.09]	1.50 [0.96-1.88]	0.830
ESR, mm/h	37.50 [23.75-65.25]	40.00 [18.00-69.00]	0.682
CRP, mg/l	12.35 [2.06-32.95]	7.14 [0.80-22.98]	0.197
KL-6, U/ml	216.58 [137.09-297.30]	376.84 [261.07-539.88]	<0.001
Disease activity			
DAS-28-ESR	3.51 ± 1.56	3.33 ± 1.37	0.489
DAS-28-CRP	4.25 ± 1.54	4.12 ± 1.46	0.611
CT images			
ILD pattern (UIP/ non-UIP)	50 (46.7%)	64.3(64.3%)	0.022
Goh score, %	12 [8-15]	19 [13-27]	<0.001
Treatment for RA			
Methotrexate	75 (72.8%)	45 (66.2%)	0.353
Methylprednisolone	47 (46.5%)	37 (57.8%)	0.158
Hydroxychloroquine	18 (18.2%)	11 (16.4%)	0.769
Leflunomide	20 (19.8%)	18 (26.9%)	0.284
Biological agent	69 (67.0%)	30 (44.8%)	0.004

Categorical variables are presented as n (%). Continuous variables are listed as median (interquartile range, IQR) or as mean ± standard deviation. n, number of patients; FVC, Forced vital capacity; FEV1, Forced expiratory volume in 1 s; TLC, Total lung capacity; DLCO, Diffusion capacity for carbon monoxide; ESR, erythrocyte sedimentation rate; RF, rheumatoid factor; CRP, C-reactive protein; APLA, anti-phospholipid antibodies; KL-6, Krebs von den Lungen-6; TNFα, tumor necrosis factor alpha; IFNγ, interferon gamma; IFNα, interferon alpha; DAS, disease activity score; UIP, usual interstitial pneumonia.

TABLE 2 Independent risk factors in training cohort.

Variables	Odds ratio (95% CI)	*p* value
Age	1.27(1.17-1.38)	<0.001
Gender	0.30(0.13-0.69)	0.005
RA duration	1.03(1.00-1.07)	0.068
FVC%	0.91(0.88-0.94)	<0.001
FEV1%	0.94(0.92-0.97)	<0.001
TLC%	0.89(0.85-0.92)	<0.001
DL_{CO}%	0.88(0.84-0.92)	<0.001
ACPA	1.000(0.996-1.004)	0.936
RFIgM	1.000(0.998-1.001)	0.678
RFIgG	1.000(0.996-1.004)	0.936
RFIgA	0.998(0.995-1.002)	0.347
KL-6	1.007(1.004-1.010)	<0.001
TNFα	1.02(0.97-1.07)	0.457
IFNα	1.04(0.94-1.15)	0.419
IFNγ	0.99(0.91-1.07)	0.771
CRP	0.99(0.97-1.00)	0.183
ESR	0.99(0.97-1.00)	0.168
DAS-28-CRP	0.84(0.64-1.11)	0.219
DAS-28-ESR	0.85(0.65-1.11)	0.227

CI, confidence-interval; ORs, Odds ratio; FVC, Forced vital capacity; FEV1, Forced expiratory volume in 1 s; TLC, Total lung capacity; DLCO, Diffusion capacity for carbon monoxide; ESR, erythrocyte sedimentation rate; RF, rheumatoid factor; CRP, C-reactive protein; APLA, anti-phospholipid antibodies; KL-6, Krebs von den Lungen-6, TNFα, tumor necrosis factor alpha; IFNγ, interferon gamma; IFNα, interferon alpha; DAS, disease activity score.

SizeZoneNonUniformity +0.0477 × log_sigma_3_0_mm_3D_glcm_Idn -0.0107 × original_glszm_LargeAreaHighGrayLevelEmphasis -0.0561 × original_glszm_SmallAreaHighGrayLevelEmphasis +0.0590 × original_glszm_SmallAreaLowGrayLevelEmphasis +0.0049 × original_ngtdm_Busyness -0.0623 × original_shape_Flatness +0.0020 × original_shape_Maximum2DDiameterSlice +0.0349 × original_shape_MinorAxisLength -0.0730 × original_shape_Sphericity -0.0597 × squareroot_firstorder_Skewness -0.0879 × wavelet_HHL_glcm_Idmn +0.0285 × wavelet_LHL_ngtdm_Busyness +0.0026 × wavelet_LLH_firstorder_Skewness.

3.4 Comparison of clinical, radiomics, and nomogram models

As shown in Figure 4, for the AUC, the clinical features [0.736, 95%CI = 0.642–0.830) and the radiomics features (0.939, 95%CI = 0.892–0.985) were perfectly fitted for the training cohort. In the testing cohort, the clinical characteristics (0.752, 95%CI = 0.610–0.894) and the radiomics signature remained well-fitted (0.901, 95% CI = 0.820–0.982). As shown in Figure 5, The nomogram using the LR algorithm, combining clinical features and radiomics features, showed the best performance in the training (0.948, 95%CI = 0.910–0.987) and testing cohort (0.923, 95%CI = 0.853–0.993), respectively. The detailed diagnostic efficiency capability for each model is presented in Supplementary Table S1.

To compare the clinical signature, radiomics signature, and nomogram, the Delong test was utilized (Supplementary Table 2). In the testing cohort, the results indicated that the AUC comparison between the nomogram and the clinical signature achieved 0.021, suggesting that the nomogram outperformed the clinical model in discriminating the GAP staging of RA-ILD. The AUC comparison between the nomogram and radiomics signature was 0.219, indicating that both models performed well in differentiating the GAP staging of RA-ILD.

3.5 Comparison of visual assessment, radiomics, and nomogram models

In the testing cohort, the Goh score achieved an AUC of 0.820 (95%CI=0.700-0.941; Supplementary Figure 1). Comparatively, both the radiomics model (0.901, 95% CI: 0.820-0.982) and the combined radiomics-KL-6 nomogram model (0.923, 95% CI: 0.853-0.993) showed superior AUC values relative to the Goh score.

3.6 Calibration curve and DCA of the models

The calibration curves for the training and testing cohorts were shown in Figure 6. The p-values from the Hosmer-Lemeshow test for clinical features, radiologic features, and nomograms were 0.557, 0.171, 0.305, and 0.193, 0.072, 0.160 in the training and test cohorts, respectively. These p-values suggest a perfect agreement for each model (Supplementary Table 3).

As shown in Figure 7, the DCA for clinical features, radiographic features, and nomograms, covering predictive probabilities from 0.12 to 0.41, 0.02 to 0.91, and 0.1 to 0.78. The nomogram achieves the largest net benefit compared to other models when the threshold probability ranges from 0.23 to 0.58.

4 Discussion

In our study, the radiomics model based on chest CT has great performance to distinguish different ILD-GAP stage patients with an AUC of 0.901 in validation cohort. The nomogram model, combining the radiomics model and serum KL-6, further enhanced the prediction efficiency of GAP staging with an AUC of 0.948 and 0.923 in the training and validation cohort, respectively.

Among the serological markers, anti-citrullinated protein antibodies (ACPA) have been implicated in the extra-articular manifestations of RA, including ILD (27–29). Correia et al. reported a correlation between ACPA titers and the risk of developing ILD (30). On the contrary, many studies have shown no association between ACPA and ILD, as well as related RF factors. Similarly, our study revealed no significant differences

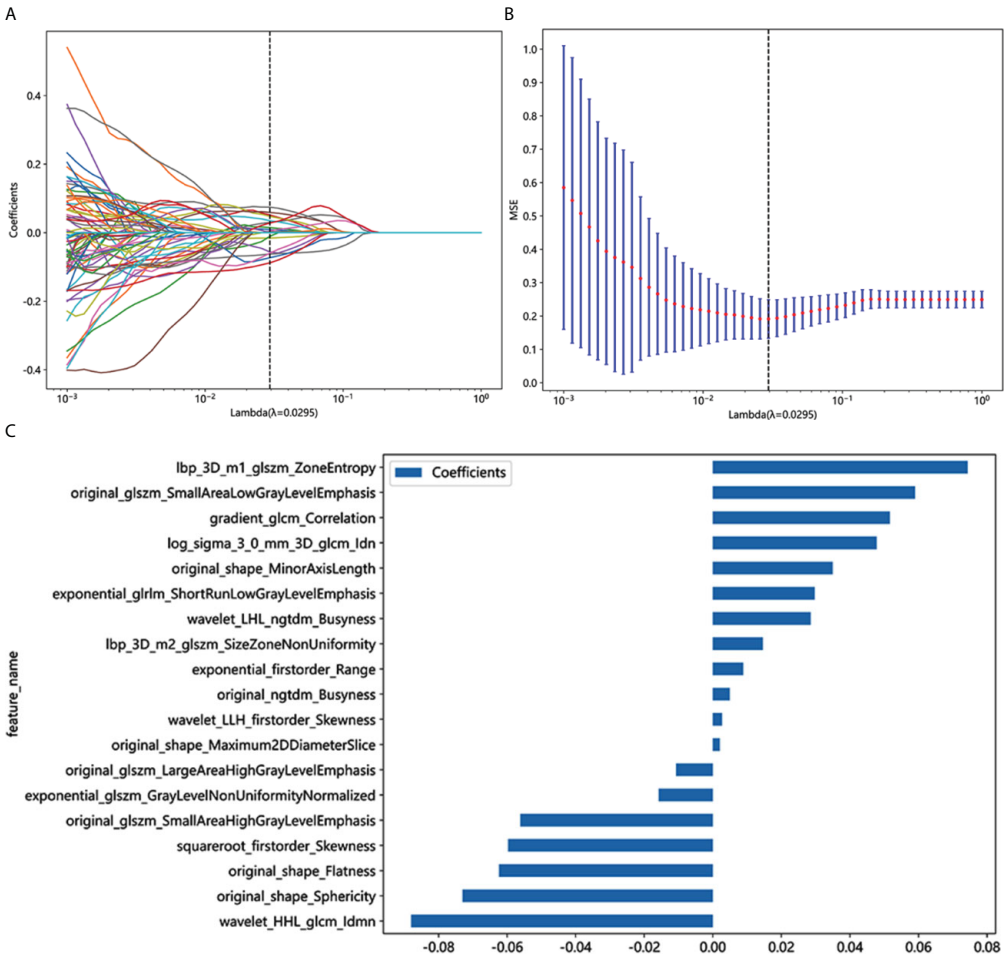


FIGURE 3
Radiomics feature selection based on LASSO algorithm and Rad score establishment. **(A)** LASSO coefficient profile plot with different log (λ)was shown. **(B)** Ten-fold cross-validated coefficients and 10-fold cross-validated MSE. **(C)** The histogram of the Rad score based on the selected features.

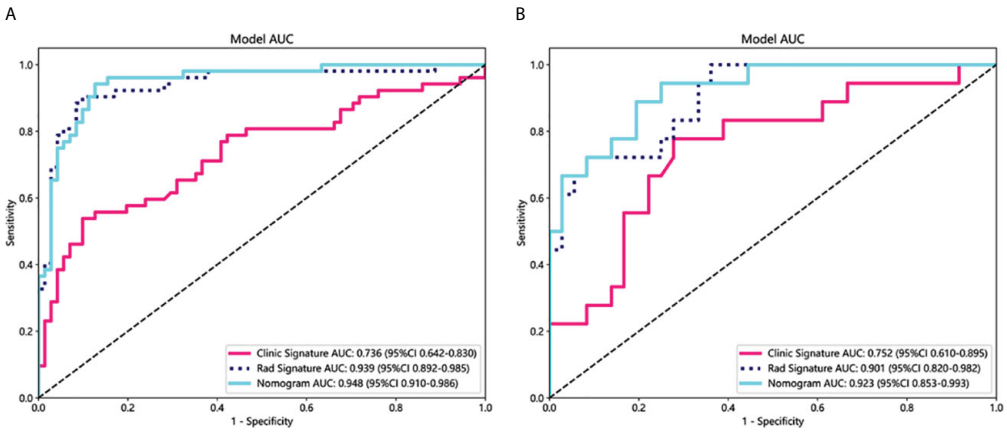


FIGURE 4
Comparison of receiver operating characteristic (ROC) curves for the clinical, radiomics, and nomogram models in the training **(A)** and testing **(B)** cohorts. The combined nomogram performed optimally in both the training and testing cohorts.

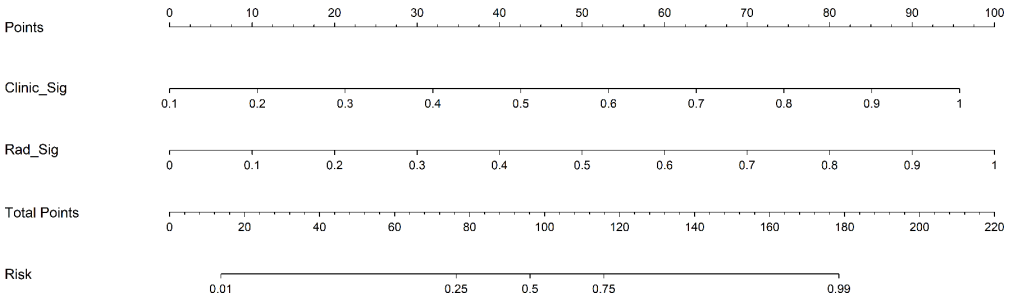


FIGURE 5
A constructed nomogram for predicting the GAP staging of RA-ILD.

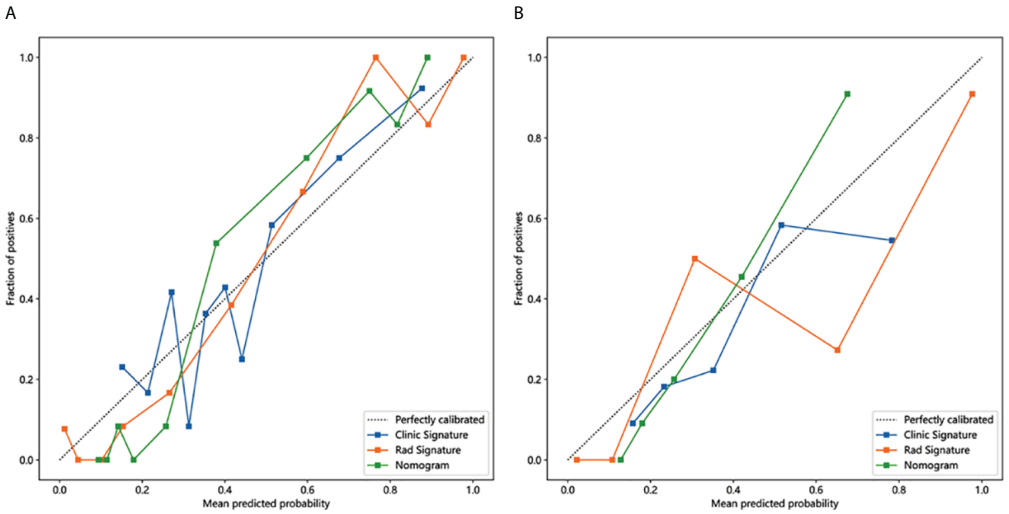


FIGURE 6
Calibration curves in the training and testing cohorts showing that the nomogram fits perfectly well in both the training **(A)** and testing cohorts **(B)**.

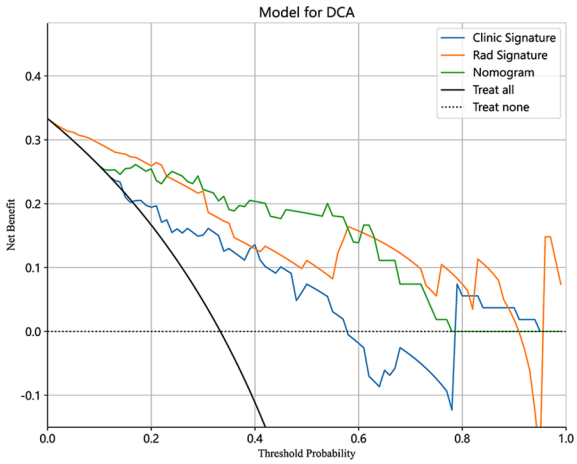


FIGURE 7
Decision curve analysis (DCA) of the clinical, radiomics, and nomogram models in the testing cohort.

between ACPA and RF factors in different stages of RA-ILD. However, these different results may be attributed to the heterogeneity of ACPA specificity and search methods (5, 31). It is worth noting that treatment strategies may play a crucial role in the development and progression of RA-ILD. A higher proportion of biological agent use was revealed in the low-risk group by our analysis. This suggests that patients using biological agents may represent a cohort receiving early and aggressive treatment. The use of biological agents may interrupt the inflammatory cascade leading to ILD, thereby reducing the risk of developing severe ILD in later stages (32, 33).

In addition, older age and male sex have been strongly associated with RA-ILD (34). We excluded gender, age, and PFTs parameters from the clinical model to prevent data leakage, despite their status as independent risk factors. Eventually, univariate logistic regression analysis revealed that KL-6 was an independent predictor in our present study. KL-6 is a mucin-like glycoprotein which stimulates fibrosis and inhibits apoptosis of pulmonary fibroblasts (35, 36). Elevated serum KL-6 levels have been observed in RA patients with lung involvement, suggesting its potential utility in early detection of ILD. In a cohort of 50 RA patients, KL-6 levels positively correlated with the high-resolution computed tomography fibrosis score, indicating that high KL-6 levels are a significant biomarker for ILD and may serve as a predictor for ILD severity in RA patients (37). Moreover, a study suggests that high KL-6 levels might be an independent risk factor and useful for the prognosis in patients with RA-ILD (38). So far, the utility of serum KL-6 has been evaluated in several forms of ILD and its sensitivity and specificity for RA-ILD ranged from 67%-85% and 60%-90%, respectively, depending on the cutoff value (36, 37, 39). In our study, a clinical factor model to classify RA-ILD stages was developed based on KL-6, and then achieved an AUC of 0.752 in the testing cohorts.

Radiomics is an objective technique offering a reliable and comprehensive quantitative assessment of images, unaffected by inter-reader variability (40). Feature extraction involves mathematical operations on digital images to generate numerical descriptors of texture, shape, and other distinct characteristics. These descriptors can be computationally analyzed to explore potential associations with clinical parameters (41). Particularly useful for diseases challenging to describe through simple visual features, high-dimensional abstract features extracted from wavelet-transformed images can provide diverse perspectives in capturing hidden information not easily observed visually. Radiomics features have indeed proven their potential for severity estimation in Systemic sclerosis-ILD and guiding treatment decisions (42). At present, the literature on the application of radiomics is limited. Venerito et al. (23) retrieved the HRCTs of 30 RA-ILD patients and suggested that radiomics analysis could predict patient mortality. This finding suggests that HRCT could serve as a digital biomarker for RA-ILD, offering prognostic value that is independent of the clinical characteristics of the disease. Recently, some scholars have developed radiomics models based on CT images to differentiate GAP staging in CTD patients. Qin et al. (21) manually segmented the right lung of CTD-ILD patients and constructed a radiomics model from the 9 extracted texture features. The AUC of their radiomics models in the validation cohort was 0.787 and 0.718 in the internal and external test cohort, respectively. A similar study utilized a semi-automatic segmentation method to segment bilateral lungs, obtaining a total of 4 features (22). Their developed radiomics model demonstrated an AUC of 0.801 in the test cohort. Instead of focusing on all types of CTDs, we concentrated on patients with RA. In our work, totally1,834 radiomics features obtained from the CT images, 19 higher-order texture features extracted from wavelet transformed images were acquired as remarkable elements to build the radiomics model, resulting in an AUC of 0.939, and 0.901 in the training and testing cohorts, respectively. It is speculated that by targeted with ILD specifically caused by RA, to some extent excluded the imbalance of training data arising from the heterogeneous imaging characteristics of various CTD-ILD subtypes (43), which eventually screened out more features. In the current study, we constructed a nomogram model that integrates the radscore with serum KL-6 levels to further enhance the accuracy of predicting low-risk RA-ILD. In contrast to the GAP index, the nomogram model can predict GAP staging in patients with RA-ILD even when precise lung function parameters are challenging to obtain. This radiomics-based approach may serve as a supportive tool for assessing the severity of RA-ILD. Moreover, the proposed model can be readily implemented in clinical practice, as it leverages routinely acquired chest CT imaging and serum biomarker data to automate the computational process, thereby minimizing the operational burden on clinicians.

There are certain limitations in our study. Firstly, the single-center design with a relatively small overall sample size, especially the limited representation of more severe ILD-GAP stage III patients, may restrict the model ability. Future studies based on larger datasets from other centers are needed to evaluate model generalizability. Secondly, the exact mortality of the retrospective study verified by the GAP index system may less precise than actual mortality of patient. Nevertheless, as an available method to predict mortality, the GAP index system has been validated in RA-ILD. The precise assessment of mortality risk will be conducted in our further research. In addition, our study serves as a foundational exploration, offering valuable insights for selecting valuable imaging biomarkers in RA-ILD.

In conclusion, a novel nomogram model combining CT-based radiomics and serum KL-6 was developed in our study. It shows good prediction accuracy in predicting low-risk RA-ILD patients, which implies that this noninvasive and quantitative method may impact the clinical decision-making process, offering a more precise management strategy for patients with RA-ILD.

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 authors.

Ethics statement

The study protocol was approved by the Ethics Committee of Guanghua Hospital Affiliated to Shanghai University of Traditional Chinese Medicine (2023-K-46). 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.

Author contributions

NH: Writing – original draft. ZG: Writing – original draft. DZ: Writing – original draft. YZ: Writing – original draft. YQ: Writing – original draft. GL: Writing – original draft. XG: Conceptualization, Supervision, Writing – review & editing. LJ: Conceptualization, Supervision, 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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2024.1417156/full#supplementary-material

References

1. Em G, Gs F. Rheumatoid arthritis - common origins, divergent mechanisms. *New Engl J Med*. (2023) 388:529–42. doi: 10.1056/NEJMra2103726

2. Conforti A, Di Cola I, Pavlych V, Ruscitti P, Berardicurti O, Ursini F, et al. Beyond the joints, the extra-articular manifestations in rheumatoid arthritis. *Autoimmun Rev*. (2021) 20:102735. doi: 10.1016/j.autrev.2020.102735

3. Hyldgaard C, Hilberg O, Pedersen AB, Ulrichsen SP, Løkke A, Bendstrup E, et al. A population-based cohort study of rheumatoid arthritis-associated interstitial lung disease: comorbidity and mortality. *Ann Rheum Dis*. (2017) 76:1700–6. doi: 10.1136/annrheumdis-2017-211138

4. Olson AL, Swigris JJ, Sprunger DB, Fischer A, Fernandez-Perez ER, Solomon J, et al. Rheumatoid arthritis-interstitial lung disease-associated mortality. *Am J Respir Crit Care Med*. (2011) 183:372–8. doi: 10.1164/rccm.201004-0622OC

5. Koduri G, Norton S, Young A, Cox N, Davies P, Devlin J, et al. Interstitial lung disease has a poor prognosis in rheumatoid arthritis: results from an inception cohort. *Rheumatol (Oxford)*. (2010) 49:1483–9. doi: 10.1093/rheumatology/keq035

6. Yunt ZX, Solomon JJ. Lung disease in rheumatoid arthritis. *Rheum Dis Clin North Am*. (2015) 41:225–36. doi: 10.1016/j.rdc.2014.12.004

7. Kadura S, Raghu G. Rheumatoid arthritis-interstitial lung disease: manifestations and current concepts in pathogenesis and management. *Eur Respir Rev*. (2021) 30:210011. doi: 10.1183/16000617.0011-2021

8. Ley B, Ryerson CJ, Vittinghoff E, Ryu JH, Tomassetti S, Lee JS, et al. A multidimensional index and staging system for idiopathic pulmonary fibrosis. *Ann Intern Med*. (2012) 156:684–91. doi: 10.7326/0003-4819-156-10-201205150-00004

9. Ryerson CJ, Vittinghoff E, Ley B, Lee JS, Mooney JJ, Jones KD, et al. Predicting survival across chronic interstitial lung disease. *Chest*. (2014) 145:723–8. doi: 10.1378/chest.13-1474

10. Nurmi HM, Purokivi MK, Kärkkäinen MS, Kettunen H-P, Selander TA, Kaarteenaho RL, et al. Are risk predicting models useful for estimating survival of patients with rheumatoid arthritis-associated interstitial lung disease? *BMC Pulm Med*. (2017) 17:16. doi: 10.1186/s12890-016-0358-2

11. Zamora-Legoff JA, Krause ML, Crowson CS, Ryu JH, Matteson EL. Patterns of interstitial lung disease and mortality in rheumatoid arthritis. *Rheumatol (Oxford)*. (2017) 56:344–50. doi: 10.1093/rheumatology/kew391

12. Morisset J, Vittinghoff E, Lee BY, Tonelli R, Hu X, Elicker BM, et al. The performance of the GAP model in patients with rheumatoid arthritis associated interstitial lung disease. *Respir Med*. (2017) 127:51–6. doi: 10.1016/j.rmed.2017.04.012

13. Graham BL, Steenbruggen I, Miller MR, Barjaktarevic IZ, Cooper BG, Hall GL, et al. Standardization of spirometry 2019 update. An official american thoracic society and european respiratory society technical statement. *Am J Respir Crit Care Med*. (2019) 200:e70–88. doi: 10.1164/rccm.201908-1590ST

14. Paschalaki KE, Jacob J, Wells AU. Monitoring of lung involvement in rheumatologic disease. *Respiration*. (2016) 91:89–98. doi: 10.1159/000442890

15. Spagnolo P, Lee JS, Sverzellati N, Rossi G, Cottin V. The lung in rheumatoid arthritis: focus on interstitial lung disease. *Arthritis Rheumatol*. (2018) 70:1544–54. doi: 10.1002/art.40574

16. Walsh SLF, Calandriello L, Sverzellati N, Wells AU, Hansell DMUIP Observer Consort. Interobserver agreement for the ATS/ERS/JRS/ALAT criteria for a UIP pattern on CT. *Thorax*. (2016) 71:45–51. doi: 10.1136/thoraxjnl-2015-207252

17. Gruden JF. CT in idiopathic pulmonary fibrosis: diagnosis and beyond. *AJR Am J Roentgenol*. (2016) 206:495–507. doi: 10.2214/AJR.15.15674

18. Tominaga J, Sakai F, Johkoh T, Noma S, Akira M, Fujimoto K, et al. Diagnostic certainty of idiopathic pulmonary fibrosis/usual interstitial pneumonia: The effect of the integrated clinico-radiological assessment. *Eur J Radiol*. (2015) 84:2640–5. doi: 10.1016/j.ejrad.2015.08.016

19. Chen M, Copley SJ, Viola P, Lu H, Aboagye EO. Radiomics and artificial intelligence for precision medicine in lung cancer treatment. *Semin Cancer Biol*. (2023) 93:97–113. doi: 10.1016/j.semcancer.2023.05.004

20. Bera K, Braman N, Gupta A, Velcheti V, Madabhushi A. Predicting cancer outcomes with radiomics and artificial intelligence in radiology. *Nat Rev Clin Oncol*. (2022) 19:132–46. doi: 10.1038/s41571-021-00560-7

21. Qin S, Jiao B, Kang B, Li H, Liu H, Ji C, et al. Non-contrast computed tomography-based radiomics for staging of connective tissue disease-associated interstitial lung disease. *Front Immunol*. (2023) 14:1213008. doi: 10.3389/fimmu.2023.1213008

22. Jiang X, Su N, Quan S, Linning E, Li R. Computed tomography radiomics-based prediction model for gender–age–physiology staging of connective tissue disease-associated interstitial lung disease. *Acad Radiol*. (2023) 30:2598–605. doi: 10.1016/j.acra.2023.01.038

23. Venerito V, Manfredi A, Lopalco G, Lavista M, Cassone G, Scardapane A, et al. Radiomics to predict the mortality of patients with rheumatoid arthritis-associated interstitial lung disease: A proof-of-concept study. *Front Med (Lausanne)*. (2022) 9:1069486. doi: 10.3389/fmed.2022.1069486

24. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO, et al. 2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Arthritis Rheum*. (2010) 62:2569–81. doi: 10.1002/art.27584

25. Raghu G, Remy-Jardin M, Myers JL, Richeldi L, Ryerson CJ, Lederer DJ, et al. Diagnosis of idiopathic pulmonary fibrosis. An official ATS/ERS/JRS/ALAT clinical practice guideline. *Am J Respir Crit Care Med*. (2018) 198:e44–68. doi: 10.1164/rccm.201807-1255ST

26. Goh NSL, Desai SR, Veeraraghavan S, Hansell DM, Copley SJ, Maher TM, et al. Interstitial lung disease in systemic sclerosis: a simple staging system. *Am J Respir Crit Care Med*. (2008) 177:1248–54. doi: 10.1164/rccm.200706-877OC

27. Kelly CA, Saravanan V, Nisar M, Arthanari S, Woodhead FA, Price-Forbes AN, et al. Rheumatoid arthritis-related interstitial lung disease: associations, prognostic factors and physiological and radiological characteristics–a large multicentre UK study. *Rheumatol (Oxford)*. (2014) 53:1676–82. doi: 10.1093/rheumatology/keu165

28. Zhang Y, Li H, Wu N, Dong X, Zheng Y. Retrospective study of the clinical characteristics and risk factors of rheumatoid arthritis-associated interstitial lung disease. *Clin Rheumatol*. (2017) 36:817–23. doi: 10.1007/s10067-017-3561-5

29. Yin Y, Liang D, Zhao L, Li Y, Liu W, Ren Y, et al. Anti-cyclic citrullinated Peptide antibody is associated with interstitial lung disease in patients with rheumatoid arthritis. *PloS One*. (2014) 9:e92449. doi: 10.1371/journal.pone.0092449

30. Correia CS, Briones MR, Guo R, Ostrowski RA. Elevated anti-cyclic citrullinated peptide antibody titer is associated with increased risk for interstitial lung disease. *Clin Rheumatol*. (2019) 38:1201–6. doi: 10.1007/s10067-018-04421-0

31. Sebastiani M, Manfredi A, Cerri S, Della Casa G, Luppi F, Ferri C. Radiologic classification of usual interstitial pneumonia in rheumatoid arthritis-related interstitial lung disease: correlations with clinical, serological and demographic features of disease. *Clin Exp Rheumatol*. (2016) 34:564–5.

32. Yunt ZX, Chung JH, Hobbs S, Fernandez-Perez ER, Olson AL, Huie TJ, et al. High resolution computed tomography pattern of usual interstitial pneumonia in rheumatoid arthritis-associated interstitial lung disease: Relationship to survival. *Respir Med*. (2017) 126:100–4. doi: 10.1016/j.rmed.2017.03.027

33. Kawano-Dourado L, Doyle TJ, Bonfiglioli K, Sawamura MVY, Nakagawa RH, Arimura FE, et al. Baseline characteristics and progression of a spectrum of interstitial lung abnormalities and disease in rheumatoid arthritis. *Chest*. (2020) 158:1546–54. doi: 10.1016/j.chest.2020.04.061

34. Doyle TJ, Patel AS, Hatabu H, Nishino M, Wu G, Osorio JC, et al. Detection of rheumatoid arthritis-interstitial lung disease is enhanced by serum biomarkers. *Am J Respir Crit Care Med*. (2015) 191:1403–12. doi: 10.1164/rccm.201411-1950OC

35. Kohno N, Akiyama M, Kyoizumi S, Hakoda M, Kobuke K, Yamakido M. Detection of soluble tumor-associated antigens in sera and effusions using novel monoclonal antibodies, KL-3 and KL-6, against lung adenocarcinoma. *Jpn J Clin Oncol*. (1988) 18:203–16.

36. Satoh H, Kurishima K, Ishikawa H, Ohtsuka M. Increased levels of KL-6 and subsequent mortality in patients with interstitial lung diseases. *J Intern Med*. (2006) 260:429–34. doi: 10.1111/j.1365-2796.2006.01704.x

37. Zheng M, Lou A, Zhang H, Zhu S, Yang M, Lai W. Serum KL-6, CA19-9, CA125 and CEA are diagnostic biomarkers for rheumatoid arthritis-associated interstitial lung disease in the chinese population. *Rheumatol Ther*. (2021) 8:517–27. doi: 10.1007/s40744-021-00288-x

38. Kim HC, Choi KH, Jacob J, Song JW. Prognostic role of blood KL-6 in rheumatoid arthritis-associated interstitial lung disease. *PloS One*. (2020) 15:e0229997. doi: 10.1371/journal.pone.0229997

39. Avouac J, Cauvet A, Steelandt A, Shirai Y, Elhai M, Kuwana M, et al. Improving risk-stratification of rheumatoid arthritis patients for interstitial lung disease. *PloS One*. (2020) 15:e0232978. doi: 10.1371/journal.pone.0232978

40. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. (2012) 48:441–6. doi: 10.1016/j.ejca.2011.11.036

41. Barnes H, Humphries SM, George PM, Assayag D, Glaspole I, Mackintosh JA, et al. Machine learning in radiology: the new frontier in interstitial lung diseases. *Lancet Digit Health*. (2023) 5:e41–50. doi: 10.1016/S2589-7500(22)00230-8

42. Martini K, Baessler B, Bogowicz M, Blüthgen C, Mannil M, Tanadini-Lang S, et al. Applicability of radiomics in interstitial lung disease associated with systemic sclerosis: proof of concept. *Eur Radiol*. (2021) 31:1987–98. doi: 10.1007/s00330-020-07293-8

43. Travis WD, Costabel U, Hansell DM, King TE, Lynch DA, Nicholson AG, et al. An official American Thoracic Society/European Respiratory Society statement: Update of the international multidisciplinary classification of the idiopathic interstitial pneumonias. *Am J Respir Crit Care Med*. (2013) 188:733–48. doi: 10.1164/rccm.201308-1483ST

frontiers | Frontiers in Imaging

TYPE Original Research
PUBLISHED 20 June 2024
DOI 10.3389/fimag.2024.1416114

OPEN ACCESS

EDITED BY
Sandeep Kumar Mishra,
Yale University, United States

REVIEWED BY
Aili Wang,
Harbin University of Science and Technology, China
Junxiang Huang,
Boston College, United States

*CORRESPONDENCE
Blake VanBerlo
✉ bvanberl@uwaterloo.ca

RECEIVED 11 April 2024
ACCEPTED 04 June 2024
PUBLISHED 20 June 2024

CITATION
VanBerlo B, Wong A, Hoey J and Arntfield R (2024) Intra-video positive pairs in self-supervised learning for ultrasound. *Front. Imaging.* 3:1416114. doi: 10.3389/fimag.2024.1416114

Intra-video positive pairs in self-supervised learning for ultrasound

Blake VanBerlo[1]*, Alexander Wong[2], Jesse Hoey[1] and Robert Arntfield[3]

[1]Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada, [2]Department of Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada, [3]Schulich School of Medicine and Dentistry, Western University, London, ON, Canada

Introduction: Self-supervised learning (SSL) is a strategy for addressing the paucity of labelled data in medical imaging by learning representations from unlabelled images. Contrastive and non-contrastive SSL methods produce learned representations that are similar for pairs of related images. Such pairs are commonly constructed by randomly distorting the same image twice. The videographic nature of ultrasound offers flexibility for defining the similarity relationship between pairs of images.

Methods: We investigated the effect of utilizing proximal, distinct images from the same B-mode ultrasound video as pairs for SSL. Additionally, we introduced a sample weighting scheme that increases the weight of closer image pairs and demonstrated how it can be integrated into SSL objectives.

Results: Named *Intra-Video Positive Pairs* (IVPP), the method surpassed previous ultrasound-specific contrastive learning methods' average test accuracy on COVID-19 classification with the POCUS dataset by $\geq$ 1.3%. Detailed investigations of IVPP's hyperparameters revealed that some combinations of IVPP hyperparameters can lead to improved or worsened performance, depending on the downstream task.

Discussion: Guidelines for practitioners were synthesized based on the results, such as the merit of IVPP with task-specific hyperparameters, and the improved performance of contrastive methods for ultrasound compared to non-contrastive counterparts.

KEYWORDS

self-supervised learning, ultrasound, contrastive learning, non-contrastive learning, representation learning

1 Introduction

Medical ultrasound (US) is a modality of imaging that uses the amplitude of ultrasonic reflections from tissues to compose a pixel map. With the advent of point-of-care ultrasound devices, ultrasound has been increasingly applied in a variety of diagnostic clinical settings, such as emergency care, intensive care, oncology, and sports medicine (Yim and Corrado, 2012; Whitson and Mayo, 2016; Sood et al., 2019; Soni et al., 2020; Lau and See, 2022). It possesses several qualities that distinguish it from other radiological modalities, including its portability, lack of ionizing radiation, and affordability. Despite morphological distortion of the anatomy, ultrasound has been shown to be comparable to radiological alternatives, such as chest X-ray and CT, for several diagnostic tasks (van Randen et al., 2011; Alrajhi et al., 2012; Nazerian et al., 2015).

Deep learning has been extensively studied as a means to automate diagnostic tasks in ultrasound. As with most medical imaging tasks, the lack of open access to large datasets is a barrier to the development of such systems, since large training sets are required for deep computer vision models. Organizations that have privileged access to large datasets are also faced with the problem of labeling ultrasound data. Indeed, many point-of-care ultrasound examinations in acute care settings are not archived or documented (Hall et al., 2016; Kessler et al., 2016).

When unlabeled examinations are abundant, researchers turn to unsupervised representation learning to produce pretrained deep learning models that can be fine-tuned using labeled data. Self-supervised learning (SSL) is a broad category of methods that has been explored for problems in diagnostic ultrasound imaging. Broadly, SSL refers to the supervised pretraining of a machine learning model for a task that does not require labels for the task of interest. The pretraining task (i.e., *pretext task*) is a supervised learning task where the target is a quantity that is computed from unlabeled data. After optimizing the model's performance on the pretext task, the weights are recast as initial weights for a new model that is trained to solve the task of interest (referred to as the *downstream task*). If the pretrained model has learned to produce representations of salient information in ultrasound images, then it is likely that it can be fine-tuned to perform the downstream task more proficiently than had it been randomly initialized. Contrastive learning is a type of pretext task in SSL that involves predicting whether two inputs are related (i.e., positive pairs) or unrelated (i.e., negative pairs). In computer vision, a common way to define positive pairs is to apply two randomly defined transformations to an image, producing two distorted views of the image with similar content. Positive pairs satisfy a *pairwise relationship* that indicate semantic similarity. All other pairs of images are regarded as negative pairs. Non-contrastive methods disregard negative pairs, focusing only on reducing the differences between representations of positive pairs.

Unlike other forms of medical imaging, US is a dynamic modality acquired as a stream of frames, resulting in a video. Despite this, there are several US interpretation tasks that can be performed by assessing a still US image. Previous studies exploring SSL in US have exploited the temporal nature of US by defining contrastive learning tasks with *intra-video positive pairs* – positive pairs comprised of images derived from the same video (Chen et al., 2021; Basu et al., 2022). Recent theoretical results indicate that the pairwise relationship must align with the labels of the downstream task in order to guarantee that self-supervised pretraining leads to non-inferior performance on the downstream task (Balestriero and LeCun, 2022). For classification tasks, this means that positive pairs must have the same class label. Due to the dynamic nature of US, one cannot assume that all frames in a US video possess the same label for all downstream US interpretation tasks. As a result, it may be problematic to indiscriminately designate any pair of images originating from the same US video as a positive pair. Moreover, since US videos are often taken sequentially as a part of the same examination or from follow-up studies of the same patient, different US videos may bear a striking resemblance to each other. It follows that designating images from different US videos as negative pairs may result in negative pairs that closely resemble positive pairs.

In this study, we aimed to examine the effect of proximity and sample weighting of intra-video positive pairs for common SSL methods. We also intended to determine if non-contrastive methods are more suitable for classification tasks in ultrasound. Since non-contrastive methods do not require the specification of negative pairs, we conjectured that non-contrastive methods would alleviate the issue of cross-video similarity and yield stronger representations for downstream tasks. Our contributions and results are summarized as follows:

- A method for sampling intra-video positive pairs for joint embedding SSL with ultrasound.
- A sample weighting scheme for joint embedding SSL methods that weighs positive pairs according to the temporal or spatial distance between them in their video of origin.
- A comprehensive assessment of intra-video positive pairs integrated with SSL pretraining methods, as measured by downstream performance in B-mode and M-mode lung US classification tasks. We found that, with proper downstream task-specific hyperparameters, intra-video positive pairs can improve performance compared to the standard practice of producing two distortions of the same image.
- An comparison of contrastive and non-contrastive learning for multiple lung US classification tasks. Contrary to our initial belief, a contrastive method outperformed multiple non-contrastive methods on multiple lung US downstream tasks.

Figure 1 encapsulates the novel methods proposed in this study. To the authors' knowledge, there are no preceding studies that systematically investigate the effect of sampling multiple images from the same US video in non-contrastive learning. More generally, we believe that this study is the first to integrate sample weights into non-contrastive objectives.

2 Background

2.1 Joint embedding self-supervised learning

Having gained popularity in recent years in multiple imaging modalities, joint embedding SSL refers to a family of methods where the pretext task is to produce output vectors (i.e., *embeddings*) that are close for examples satisfying a similarity pairwise relationship. Pairs of images satisfying this relationship are known as *positive pairs*, and they assumed to share semantic content with respect to the downstream task. For example, positive pairs could belong to the same class in a downstream supervised learning classification task. On the other hand, *negative pairs* are pairs of images that do not satisfy the pairwise relationship. In the label-free context of SSL, positive pairs are often constructed by sampling distorted versions of a single image (Chen et al., 2020; Grill et al., 2020; Zbontar et al., 2021; Bardes et al., 2022). The distortions are sampled from a distribution of sequentially applied

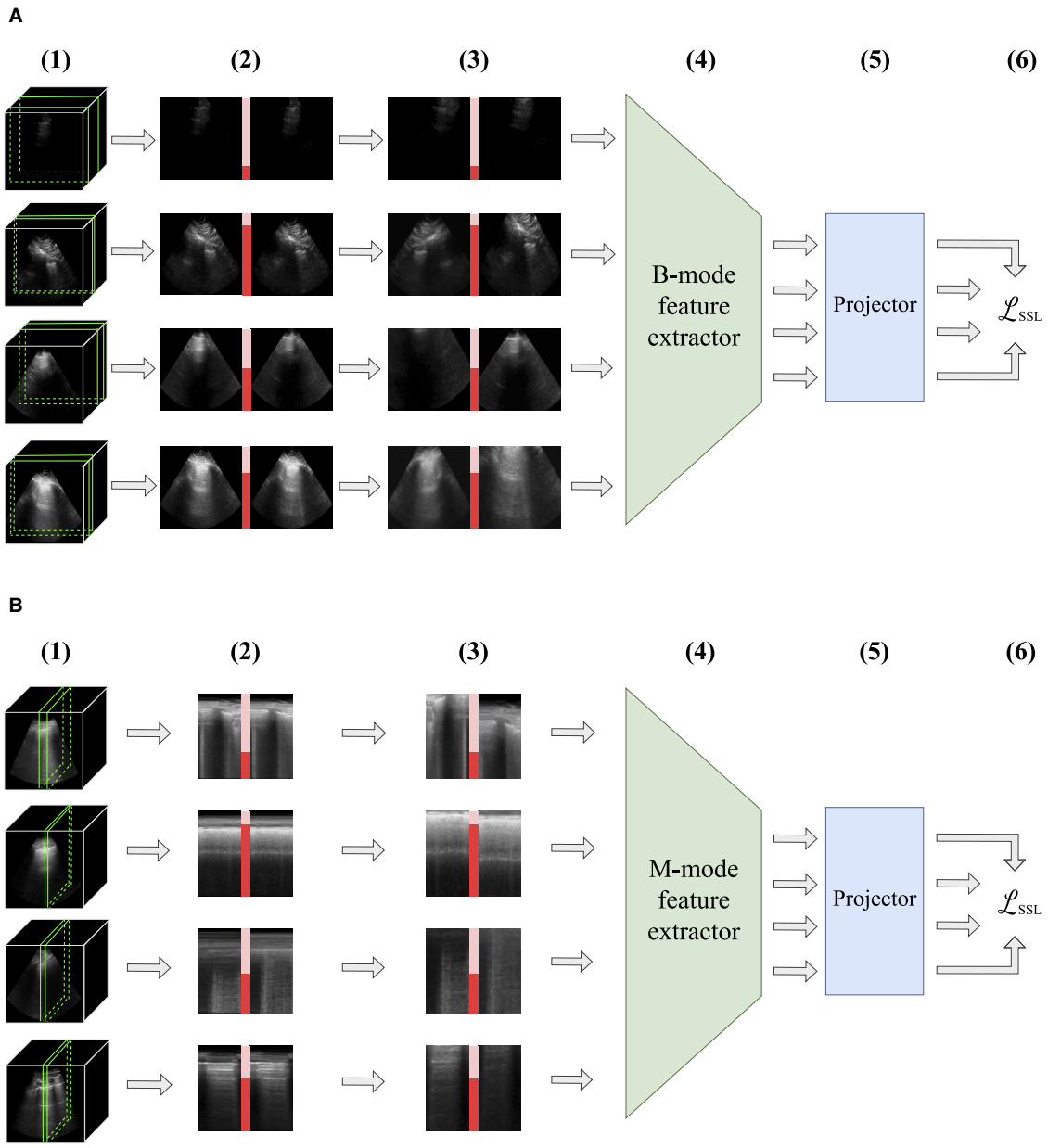


FIGURE 1
An overview of the methods introduced in this study. Positive pairs of images separated by no more than a threshold are sampled from the same B-mode video (1). Sample weights inversely proportional to the separation between each image (red bars) are calculated for each pair (2). Random transformations are applied to each image (3). Images are sent to a neural network consisting of a feature extractor (4) and a projector (5) connected in series. The outputs are used to calculate the objective $\mathcal{L}_{SSL}$ (6). The trained feature extractor is retained for downstream supervised learning tasks. **(A)** For B-mode ultrasound, positive pairs are temporally separated images from the same video. **(B)** For M-mode ultrasound, positive pairs are spatially separated images from the same video.

transformations that are designed to preserve the semantic content of the image. Horizontal reflection is a common example of a transformation that meets this criterion in many forms of imaging.

The architecture of joint embedding models commonly consists of two modules connected in series: a feature extractor and a projector. The feature extractor is typically a convolutional neural network (CNN) or a variant of a vision transformer, while the projector is a multi-layer perceptron. After pretraining, the projector is discarded and the feature extractor is retained for weight initialization for the downstream task.

Contrastive learning and *non-contrastive learning* are two major categories of joint embedding methods. Contrastive methods rely on objectives that explicitly attract positive pairs and repel negative pairs in embedding space. Many of these methods adopt the InfoNCE objective (Oh Song et al., 2016), which may be viewed as cross-entropy for predicting which combination of embeddings in a batch correspond to a positive pair. In most contrastive methods, positive pairs and negative pairs are distorted versions of the same image and different images, respectively. MoCo is a contrastive method that computes pairs of embeddings using two

feature extractors: a "query" encoder and a "key" encoder (He et al., 2020). The key encoder, which is an exponentially moving average of the query encoder, operates on negative examples Its output embeddings are queued to avoid recomputation of negative embeddings. SimCLR (Chen et al., 2020) is a widely used contrastive method that employs a variant of the InfoNCE objective that does not include the embedding of the positive pair in the demoninator (Oh Song et al., 2016). It does not queue negative embeddings, relying instead on large batches of negative examples.

Non-contrastive methods dispense with negative pairs altogether, limiting their focus to reducing the difference between embeddings of positive pairs. By design, they address the information collapse problem – a degenerate solution wherein all examples map to a null representation vector. Self-distillation non-contrastive methods use architectural and asymmetrical training strategies to avoid collapse [e.g., BYOL (Grill et al., 2020)]. Information maximization non-contrastive methods address collapse by employing objectives that maximize the information content of the embedding dimensions. For instance, the Barlow Twins method is a composite objective that contains a term for penalizing dimensional redundancy for batches of embeddings, in addition to a term for the distances between embeddings of individual positive pairs (Zbontar et al., 2021). VICReg introduced an additional term that explicitly maximizes variance across dimensions for batches of embeddings (Bardes et al., 2022). Despite a common belief that contrastive methods need much larger batch sizes than non-contrastive methods, recent evidence showed that hyperparameter tuning can boost the former's performance with smaller batch sizes (Bordes et al., 2023). Non-contrastive methods have been criticized for requiring embeddings with greater dimensionality than the representations outputted by the feature extractor; however, a recent study suggested that the difference may be alleviated through hyperparameter and design choices (Garrido et al., 2022).

Theoretical works have attempted to unify contrastive and non-contrastive methods. Balestriero and LeCun (2022) found that SimCLR, VICReg, and Barlow Twins are all manifestations of spectral embedding methods. Based on their results, they recommended that practitioners define a pairwise relationship that aligns with the downstream task. For example, if the downstream task is classification, then positive pairs should have the same class. Garrido et al. (2022) challenged the widely held assumptions that non-contrastive methods perform better than contrastive methods and that non-contrastive methods rely on large embedding dimensions. They showed that the methods perform comparatively on benchmark tasks after hyperparameter tuning and that VICReg can be modified to reduce the dependence on large embeddings (Garrido et al., 2022).

2.2 Joint embedding methods for B-mode lung ultrasound

Ultrasound is a dynamic imaging modality that is typically captured as a sequence of images and stored as a video. As such, images originating from the same video are highly correlated and are likely to share semantic content. Accordingly, recent works have developed US-specific contrastive learning methods that construct positive pairs from the same video. The Ultrasound Contrastive Learning (USCL) method (Chen et al., 2021) is a derivative of SimCLR in which positive pairs are weighted sums of random images within the same video [i.e., the mixup operation (Zhang et al., 2017)], while negative pairs are images from different videos. They reported an improvement on the downstream task of COVID-19 classification with the POCUS dataset (Born et al., 2020). Improving on USCL, Meta-USCL concurrently trains a separate network that learns to weigh positive pairs (Chen et al., 2022). The work was inspired by the observation that the intra-video positive pairs may exhibit a wide range of semantic similarity or dissimilarity. Basu et al. (2022) proposed a MoCo-inspired solution where positive pairs are images that are temporally close within a video, while negative pairs consist of either pairs from different videos or pairs from the same video that are separated temporally by a no less than a gradually decreasing threshold. Lastly, the HiCo method's objective is the sum of a softened InfoNCE loss calculated for the feature maps outputted by various model blocks (Zhang et al., 2022). The authors reported greatly improved performance with respect to USCL.

Standard non-contrastive methods have been applied for various tasks in US imaging. In addition to assessing contrastive methods, Anand et al. (2022) conducted pretraining with two self-distillation non-contrastive methods [BYOL (Grill et al., 2020) and DINO (Caron et al., 2021)] on a large dataset of echocardiograms. BYOL pretraining has also been applied in anatomical tracking tasks (Liang et al., 2023). Information maximization methods have been investigated for artifact detection tasks in M-mode and B-mode lung ultrasound (VanBerlo et al., 2023a,b). To our best knowledge, no studies have trialed non-contrastive learning methods for B-mode ultrasound with intra-video positive pairs. The present study seeks to address this gap in the literature by investigating the effect of sampling positive pairs from the same video on the efficacy of non-contrastive pretraining for tasks in ultrasound.

3 Methods

3.1 Joint embedding methods for ultrasound with intra-video positive pairs

3.1.1 Setup

We consider the standard joint embedding scenario where unlabeled data are provided and the goal is to maximize the similarity between embeddings of positive pairs. In contrastive learning, the goal is augmented by maximizing the dissimilarity of negative pairs. Let x_1 and x_2 denote a positive pair of US images. Self-supervised pretraining results in a feature extractor $f(x)$ that outputs representation vector h. The goal of SSL is to produce a feature extractor that is a better starting point for learning the downstream task than random initialization.

In this study, we propose a simple method for sampling and weighing positive pairs in the joint embedding setting that can be adopted for any joint embedding SSL method. We adopt SimCLR (Chen et al., 2020), Barlow Twins (Zbontar et al., 2021) and VICReg (Bardes et al., 2022) for our experiments. In these methods, a MLP projector is appended to the feature extractor during pretraining. $z = g(h) = g(f(x))$ is the embedding vector

outputted by the projector. The SSL objective is then computed in embedding space.

3.1.2 Intra-video positive pairs: (IVPP)

Recall that positive pairs are images that are semantically related. Previous work in contrastive SSL for US has explored the use of intra-video positive pairs (Chen et al., 2021, 2022; Basu et al., 2022; Zhang et al., 2022). A problem with naively sampling intra-video positive pairs is that it rests on the assumption that all images in the video are sufficiently similar. However, clinically relevant signs commonly surface and disappear within the same US video as the US probe and/or the patient move. For example, B-lines are an artifact in lung US that signify diseased lung parenchyma (Soni et al., 2020). B-lines may disappear and reappear as the patient breathes or as the sonographer moves the probe. The A-line artifact appears in the absence of B-lines, indicating normal lung parenchyma. In the absence of patient context, an image containing A-lines and an image containing B-lines from the same video convey very different impressions. While most previous methods only considered inter-video images to be negative pairs, Basu et al. (2022) argued that that temporally distant intra-video pairs of US images are more likely to be dissimilar, which inspired their method that treats such instances as negative pairs. Despite this, we argue that distant intra-video images may sometimes exhibit similar content. For example, the patient and probe may remain stationary throughout the video, or the probe may return to its original position and/or orientation. Moreover, periodic physiological processes such as the respiratory cycle may result in temporally distant yet semantically similar images. Without further knowledge of the US examinations in a dataset, we conjectured that it may be safest to only assume that positive pairs are intra-video images that are close to each other. Closer pairs are likely to contain similar semantic content, yet they harbor different noise samples that models should be invariant to. In summary, this method distinguishes itself from prior work by only considering proximal frames to be positive pairs and treating distant pairs as neither positive nor negative pairs.

For B-mode US videos, we define positive pairs as intra-video images x_1 and x_2 that are temporally separated by no more than $\delta_{\max}$ seconds. To accomplish this, x_1 is randomly drawn from the video's images, and x_2 is randomly drawn from the set of images that are within δ_t seconds of x_1. The frame rate of the videos must be known in order to determine which images are sufficiently close to x_1. Note that videos with higher frame rates will provide more candidates for positive pairs, potentially increasing the diversity of pairs with respect to naturally occurring noise.

A similar sampling scheme is applied for M-mode US images. Like previous studies, we define M-mode images as vertical slices through time of a B-mode video, taken at a specific x-coordinate in the video (Jasčůr et al., 2021; VanBerlo et al., 2022b, 2023b). The x-axis of an M-mode image is time, and its y-axis is the vertical dimension of the B-mode video. We define positive pairs to be M-mode images whose x-coordinates differ by no more than δ_x pixels. To avoid resolution differences, all B-mode videos are resized to the same width and height prior to sampling M-mode images. The positive pair sampling process for B-mode and M-mode images is depicted in Figure 2.

As is customary in joint embedding methods, stochastic data augmentation is applied to each image, encouraging the feature extractor to become invariant to semantically insignificant differences. Any data augmentation pipeline may be adopted for this formulation of intra-video positive pairs; however, we recommend careful selection of transformations and the distributions of their parameters to ensure that the pairwise relationship continues to be consistent with the downstream US task.

3.1.3 Sample weights

The chance that intra-video images are semantically related decreases as temporal or spatial separation increases. To temper the effect of unrelated positive pairs, we apply sample weights to positive pairs in the SSL objective according to their temporal or spatial distance. Distant pairs are weighed less than closer pairs. For a positive pair of B-mode images occurring at times t_1 and t_2 or M-mode images occurring at positions x_1 and x_2, the sample weight is calculated using Equation 1:

$$w = \frac{\delta_t - |t_2 - t_1| + 1}{\delta_t + 1} \qquad w = \frac{\delta_x - |x_2 - x_1| + 1}{\delta_x + 1} \tag{1}$$

Sample weights were incorporated into each SSL objective trialed in this study. Accordingly, we modified the objective functions for SimCLR, Barlow Twins, and VICReg in order to weigh the contribution to the loss differently based on sample weights. Appendix 1 describes the revised objective functions. To the authors' knowledge, this study is the first to propose sample weighting schemes for the aforementioned self-supervised learning methods.

3.2 Ultrasound classification tasks

3.2.1 COVID-19 classification (COVID)

As was done in previous studies on on US-specific joint embedding methods (Chen et al., 2021, 2022; Basu et al., 2022; Zhang et al., 2022), we evaluate IVPP on the public POCUS lung US dataset (Born et al., 2020). This dataset contains 140 publicly sourced US videos (2116 images) labeled for three classes: COVID-19 pneumonia, non-COVID-19 pneumonia, and normal lung.[1] When evaluating on POCUS, we pretrain on the public Butterfly dataset, which contains 22 unlabeled lung ultrasound videos (Butterfly Network, 2020).[2]

3.2.2 A-line vs. B-line classification (AB)

A-lines and B-lines are two cardinal artifact in B-mode lung US that can provide quick information on the status of a patient's

1 See dataset details at the public POCUS repository (Born et al., 2020): https://github.com/jannisborn/covid19_ultrasound.

2 Accessed via a URL available at the public USCL repository (Chen et al., 2021): https://github.com/983632847/USCL.

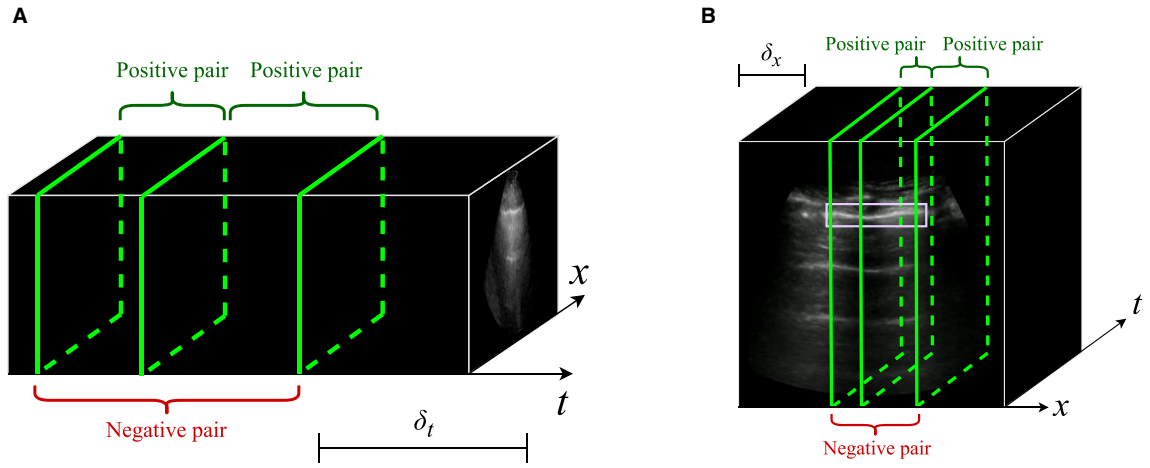


FIGURE 2
Illustration of intra-video positive pairs. Positive pairs are considered images that are no more than a threshold apart from each other within the same ultrasound video. **(A)** For B-mode ultrasound, positive pairs are frames in the same video that are within δ_t seconds of each other. **(B)** For M-mode ultrasound, positive pairs are M-mode images originating from the same B-video that are located within δ_x pixels from each other. In the context of lung ultrasound, M-mode images should intersect the pleural line (outlined in mauve).

lung tissue. A-lines are reverberation artifacts that are indicative of normal, clear lung parenchyma (Soni et al., 2020). On lung US, they as horizontal lines deep to the pleural line. Conversely, B-lines are indicative of diseased lung tissue (Soni et al., 2020). Generally, the two are mutually exclusive. We evaluate on the binary classification task of A-lines versus B-lines on lung US, as was done in previous work benchmarking joint embedding SSL methods for lung US tasks (VanBerlo et al., 2023a).

We use a private dataset of 25917 parenchymal lung US videos (5.9e6 images), hereafter referred to as *ParenchymalLUS*. It is a subset of a larger database of de-identified lung US videos that was partially labeled for previous work (Arntfield et al., 2021; VanBerlo et al., 2022b). Access to this database was permitted via ethical approval by Western University (REB 116838). Before experimentation, we split the labeled portion of ParenchymalLUS by anonymous patient identifier into training, validation, and test sets. The unlabeled portion of ParenchymalLUS was assembled by gathering 20000 videos from the unlabeled pool of videos in the database that were predicted to contain a parenchymal view of the lungs by a previously trained lung US view classifier (VanBerlo et al., 2022a). All videos from the same patient were in either the labeled or the unlabeled subset. Table 1 provides further information on the membership of ParenchymalLUS.

3.2.3 Lung sliding classification (LS)

Lung sliding is a dynamic artifact that, when observed on a parenchymal lung US view, rules out the possibility of a pneumothorax at the site of the probe (Lichtenstein and Menu, 1995). The absence of lung sliding is suggestive of pneumothorax, warranting further investigation. On B-mode US, lung sliding manifests as a shimmering of the pleural line (Lichtenstein and Menu, 1995). The presence or absence of lung sliding is also appreciable on M-mode lung US images that intersect the pleural line (Lichtenstein et al., 2005; Lichtenstein, 2010). We evaluate on the binary lung sliding classification task, where positive pairs are M-mode images originating from the same B-mode video.

ParenchymalLUS is adopted for the lung sliding classification task. We use the same train/validation/test partition as described above. Following prior studies, we estimate the horizontal bounds of the pleural line using a previously trained object detection model (VanBerlo et al., 2022b) and use the top half of qualifying M-mode images, in decreasing order of total pixel intensity (VanBerlo et al., 2023b).

4 Results

4.1 Training protocols

Unless otherwise stated, all feature extractors are initialized with ImageNet-pretrained weights. Similar studies concentrating on medical imaging have observed that this practice improves downstream performance when compared to random initialization (Azizi et al., 2021; VanBerlo et al., 2023b). Moreover, we designate fully supervised classifiers initialized with ImageNet-pretrained weights as a baseline against which to compare models pretrained with SSL.

Evaluation on POCUS follows a similar protocol employed in prior works (Chen et al., 2021; Basu et al., 2022). Feature extractors with the ResNet18 architecture (He et al., 2016) are pretrained on the Butterfly dataset. Prior to training on the POCUS dataset, a 3-node fully connected layer with softmax activation was appended to the pretrained feature extractor. Five-fold cross validation is conducted with POCUS by fine-tuning the final three layers of the pretrained feature extractor. Unlike prior works, we adopt the average across-folds validation accuracy, instead of taking the accuracy of the combined set of validation set predictions across folds. Presenting the results in this manner revealed the high

TABLE 1 Breakdown of ParenchymalLUS at the video and image level.

		Unlabeled	Labeled		
			Train	Validation	Test
Total	Patients	5, 204	1, 540	330	329
	Videos	20,000	4123	858	936
	Images	4,611,063	927,889	191,437	208,648
A/B line labels	Videos	–	2, 100 / 998	441 / 197	512 / 213
	Images	–	484,287 / 216,505	99,132 / 40,608	116,648 / 42,122
Lung sliding labels	Videos	–	3, 169 / 477	631 /103	707 / 96
	Images	–	727,205 / 96,771	146,322 / 23,218	166,753 / 21,911

x/y indicates the number of negative and positive labeled examples available for each task, respectively. Video labels apply to each image within the video. Note that some videos were not labeled for both tasks.

variance of model performance across folds, which may be due to the benchmark dataset's small video sample size.

All experiments with ParenchymalLUS utilize the MobileNetV3-Small architecture as the feature extractor, which outputs a 576-dimensional representation vector (Howard et al., 2019). Feature extractors are pretrained on the union of the unlabeled videos and labeled training set videos in ParenchymalLUS. Performance is assessed via test set classification metrics. Prior to training on the downstream task, a single-node fully connected layer with sigmoid activation was appended to the pretrained feature extractor. We report the performance of linear classifiers trained on the frozen feature extractor's representations, along with classifiers that are fine-tuned end-to-end.

For each joint embedding method, the projectors were multilayer perceptrons with two 768-node layers, outputting 768-dimensional embeddings. Pretraining is conducted for 500 epochs using the LARS optimizer (You et al., 2019) with a batch size of 384 and a learning rate schedule with warmup and cosine decay as in Bardes et al. (2022).

The pretraining and training data augmentation pipelines consist of random transformations, including random cropping, horizontal reflection, brightness jitter, contrast jitter, and Gaussian blurring. Additional data preprocessing details are available in Appendix 2.

Source code will be made available upon publication.[3]

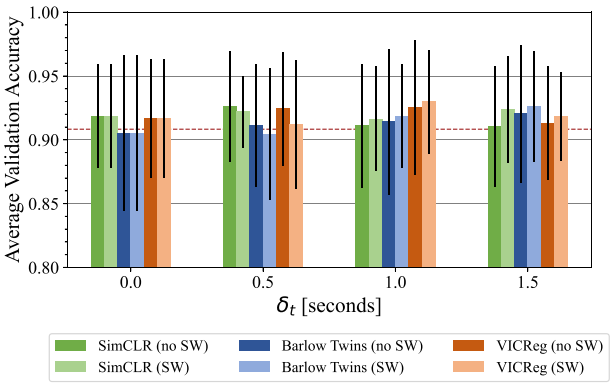


FIGURE 3
Average test accuracy across 5-fold cross validation on the POCUS dataset. Models were pretrained with a variety of intra-video positive pair thresholds with and without sample weights. Error bars indicate the standard deviation across folds. The dashed line indicates initialization with ImageNet-pretrained weights.

4.2 Performance

The two main proposed features of IVPP are intra-video positive pairs and distance-based sample weights. Accordingly, we assess the performance of IVPP across multiple assignments of the maximum image separation. Separate trials were conducted for SimCLR, Barlow Twins, and VICReg pretraining. For the COVID and AB tasks, we explored the values $\delta_t \in \{0, 0.5, 1, 1.5\}$ seconds. The LS task is defined for M-mode US, and so we explored $\delta_x \in \{0, 5, 10, 15\}$ pixels. The standardized width of B-mode US videos should be considered when determining an appropriate range for δ_x. Note that when $\delta = 0$, sample weights are all 1 and therefore do not modify any of the SSL objectives investigated in this study.

Figure 3 summarizes the performance of IVPP on the public POCUS dataset after pretraining on the Butterfly dataset, which is measured by average test accuracy in 5-fold cross validation. In most cases, pretrained models obtained equal or greater average accuracy than the ImageNet-pretrained baseline, with the exception of Barlow Twins with $\delta_t = 0$ and $\delta_t = 0.5$. The performance of models pretrained with SimCLR, Barlow Twins, and VICReg peaked at different nonzero values of δ_t (0.5, 1, and 1.5 respectively), indicating a possible benefit of selecting temporally close yet distinct intra-video positive pairs. It was also observed across all three pretraining methods that the inclusion of sample weights resulted in worsened test AUC when $\delta = 0.5$, but improved test AUC when $\delta = 1.0$ and $\delta = 1.5$.

Similar experiments were conducted with ParenchymalLUS for the AB task and LS task, using B-mode and M-mode images respectively as input. ParenchymalLUS represents a scenario where there is an abundance of unlabeled data, which differs greatly from the preceding evaluation on public, yet small,

3 https://github.com/bvanberl/IVPP

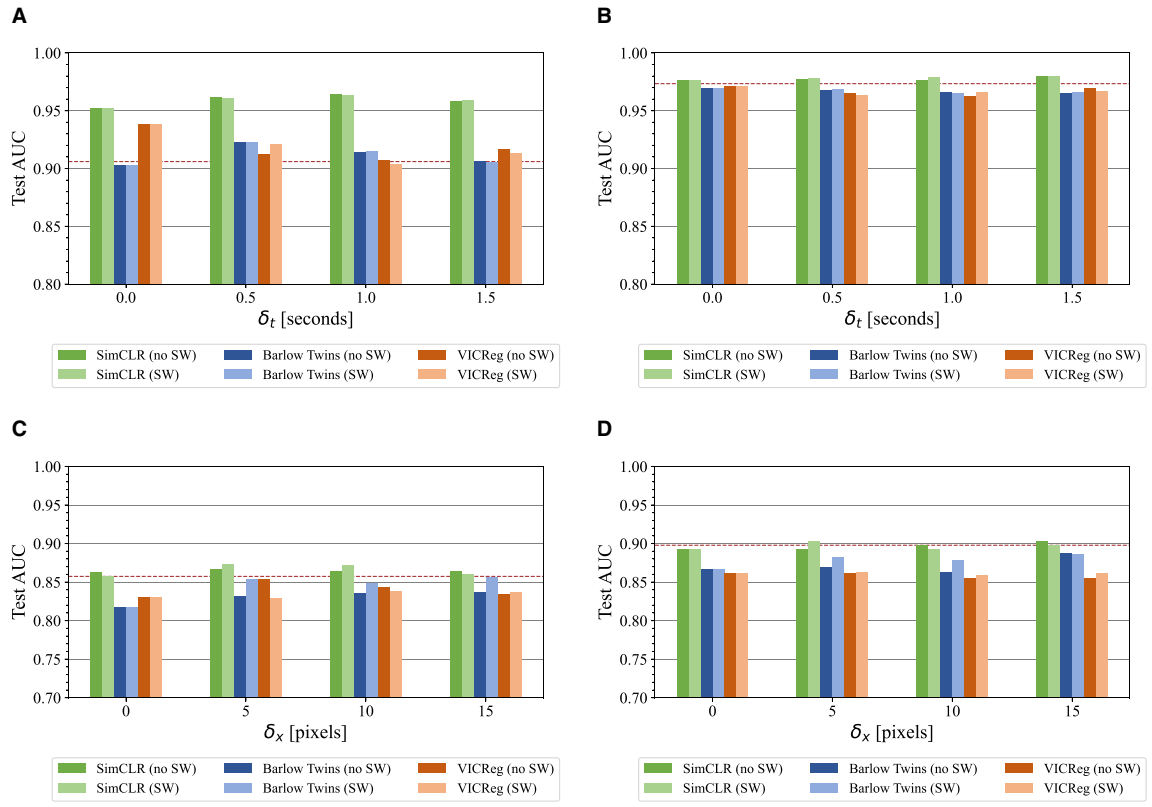


FIGURE 4
ParenchymalLUS test set AUC for the AB and LS binary classification tasks, calculated for models pretrained with a selection of contrastive and non-contrastive learning methods and employing a variety of intra-video positive pair thresholds with and without sample weights (SW). The dashed line indicates initialization with ImageNet-pretrained weights. **(A)** Linear classifiers for the AB task. **(B)** Fine-tuned classifiers for the AB task. **(C)** Linear classifiers for the LS task. **(D)** Fine-tuned classifiers for the LS task.

datasets. The unlabeled and labeled portions of ParenchymalLUS contained at least an order of magnitude more videos than either the public Butterfly and POCUS datasets. B-mode and M-mode feature extractors were pretrained on the union of the unlabeled and training portions of ParenchymalLUS—one for each value of δ, with and without sample weights. For these evaluations, we use all training examples that have been assigned a label for the downstream task. Figure 4 provides a visual comparison of the test AUC obtained by linear feature representation classifiers and fine-tuned models for the AB and LS tasks. An immediate trend across both tasks and evaluation types is that SimCLR consistently outperformed Barlow Twins and VICReg, which are both non-contrastive methods. Furthermore, pretraining with non-contrastive methods often resulted in worse test AUC compared to initialization with ImageNet-pretrained weights. Another observation across all experiments was that there was no discernible trend for the effect of sample weights that was consistent for any task, pretraining method, δ_t, or δ_x.

Focusing on AB, linear classifiers achieved the greatest performance when $\delta_t > 0$, with the exception of VICReg (Figure 4A). The use of SimCLR compared to the other pretraining methods appeared to be responsible for the greatest difference in test performance. As shown in Figure 4A, SimCLR-pretrained models outperformed non-contrastive methods, and were the only models that outperformed ImageNet-pretrained weights. The use of a nonzero δ_t resulted in slight improvement in combination with SimCLR pretraining, but degraded performance of non-contrastive methods.

Similar results were observed for the LS M-mode classification task. Models pretrained with SimCLR were the only ones that matched or surpassed fully supervised models. Nonzero δ_x generally improved the performance of linear classifiers, with $\delta_x = 5$ pixels corresponding to the greatest test AUC for SimCLR and VICReg, and $\delta_x = 15$ for Barlow Twins. Inclusion of sample weights appreciably improved the performance of Barlow Twins-pretrained models. Fine-tuned models pretrained with SimCLR performed similarly to fully supervised models, while non-contrastive methods resulted in degradation of test AUC.

Table 2 compares the top-performing IVPP-pretrained models for each SSL method with two prior US-specific contrastive learning methods— USCL (Chen et al., 2021) and US UCL (Basu et al., 2022). Of note is that all three self-supervised methods pretrained with IVPP outperformed ImageNet-pretrained initialization for POCUS, a task where very little pretraining *and* training data were utilized. For the B-mode and M-mode tasks assessed with ParenchymalLUS, a contrastive method (including the baseline) outperformed non-contrastive methods. Appendix 4 provides additional results that exhibit a similar trend with different pretraining batch sizes. Overall, the most salient result from the above experiments is that SimCLR, a contrastive method,

TABLE 2 Performance of fine-tuned models pretrained using IVPP compared to US-specific contrastive learning methods, USCL, and UCL, and to baseline random and ImageNet initializations.

Dataset	POCUS	ParenchymalLUS	
Pretraining method	Mean (std) test accuracy	A/B Test AUC	LS Test AUC
Random initialization	0.881 (0.050)	0.954	0.790
ImageNet initialization	0.908 (0.043)	0.973	0.898
USCL (Chen et al., 2021)	0.905 (0.044)	0.979	0.874
US UCL (Basu et al., 2022)	0.901 (0.054)	0.967	0.809
IVPP [SimCLR]	0.926 (0.043)	0.980	0.903
IVPP [Barlow Twins]	0.921 (0.054)	0.969	0.887
IVPP [VICReg]	0.930 (0.046)	0.971	0.862

outperformed both non-contrastive methods when unlabeled data is abundant.

4.3 Label efficiency

ParenchymalLUS is much larger than public ultrasound datasets for machine learning. Although the majority of its videos are unlabeled, it contains a large number of labeled examples. To simulate a scenario where the fraction of examples that are labeled is much smaller, we investigated the downstream performance of models that were pretrained on all the unlabeled and training ParenchymalLUS examples and then fine-tuned on a very small subset of the training set.

Label efficiency investigations are typically conducted by fitting a model for the downstream task using progressively smaller fractions of training data to gauge how well self-supervised models fare in low-label scenarios. The results of these experiments may be unique to the particular training subset that is randomly selected. We designed an experiment to determine if the choice of δ_t, δ_x, or the introduction of sample weights influenced downstream performance in low-label settings. To reduce the chance of biased training subset sampling, we divided the training set into 20 subsets and repeatedly performed fine-tuning experiments on each subset for each pretraining method and δ value, with and without sample weights. To ensure independence among the subsets, we split the subsets by patient. Inspection of the central moments and boxplots from each distribution indicated that the normality and equal variance assumptions for ANOVA were not violated. For each pretraining method, a two-way repeated-measures analysis of variance (ANOVA) was performed to determine whether the mean test AUC scores across values of δ and sample weight usage were different. The independent variables were δ and the presence of sample weights, while the dependent variable was test AUC. Whenever the null hypothesis of the ANOVA was rejected, *post-hoc* paired *t*-tests were performed to compare the following:

- Pretraining with nonzero δ against standard positive pair selection ($\delta = 0$).
- For the same nonzero δ value, sample weights against no sample weights.

For each group of *post-hoc* tests, the Bonferroni correction was applied to establish a family-wise error rate of $\alpha = 0.05$. To ensure that each training subset was independent, we split the dataset by anonymous patient identifier. This was a necessary step because intra-video images are highly correlated, along with videos from the same patient. As a result, the task became substantially more difficult than naively sampling 5% of training images because the volume *and* heterogeneity of training examples was reduced by training on a small fraction of examples from a small set of patients.

The fine-tuning procedure was identical to that described in Section 4.1, with the exception that the model's weights at the end of training were retained for evaluation, instead of restoring the best-performing weights on the validation set. Figure 5 provides boxplots for all trials that indicate the distributions of test AUC under the varying conditions for both the AB and LS tasks. Again, SimCLR performance appeared to be substantially higher than both non-contrastive methods.

Table 3 gives the mean and standard deviation of each set of trials, for each hyperparameter combination. For each task and each pretraining method, the ANOVA revealed significant interaction effects ($p \leq 0.05$). Accordingly, all intended *post-hoc* *t*-tests were performed to ascertain (1) which combinations of hyperparameters were different from the baseline setting of augmenting the same frame twice ($\delta = 0$) and (2) values of δ where the addition of sample weights changes the outcome. First, we note that SimCLR was the only pretraining method that consistently outperformed full supervision with ImageNet-pretrained weights. Barlow Twins and VICReg pretraining – both non-contrastive methods – resulted in worse performance.

For the AB task, no combination of intra-video positive pairs or sample weights resulted in statistically significant improvements compared to dual distortion of the same image ($\delta_t = 0$). For Barlow Twins and VICReg, several nonzero δ_t resulted in significantly worse mean test AUC. Sample weights consistently made a difference in Barlow Twins across δ_t values, but only improved mean test AUC for $\delta_t = 1$ and $\delta_t = 1.5$.

Different trends were observed for the LS task. SimCLR with $\delta_x = 5$ and no sample weights improved mean test AUC compared to the baseline where $\delta_x = 0$. No other combination of hyperparameters resulted in a significant improvement. For Barlow Twins, multiple IVPP hyparameter combinations resulted in improved mean test AUC over the baseline. No

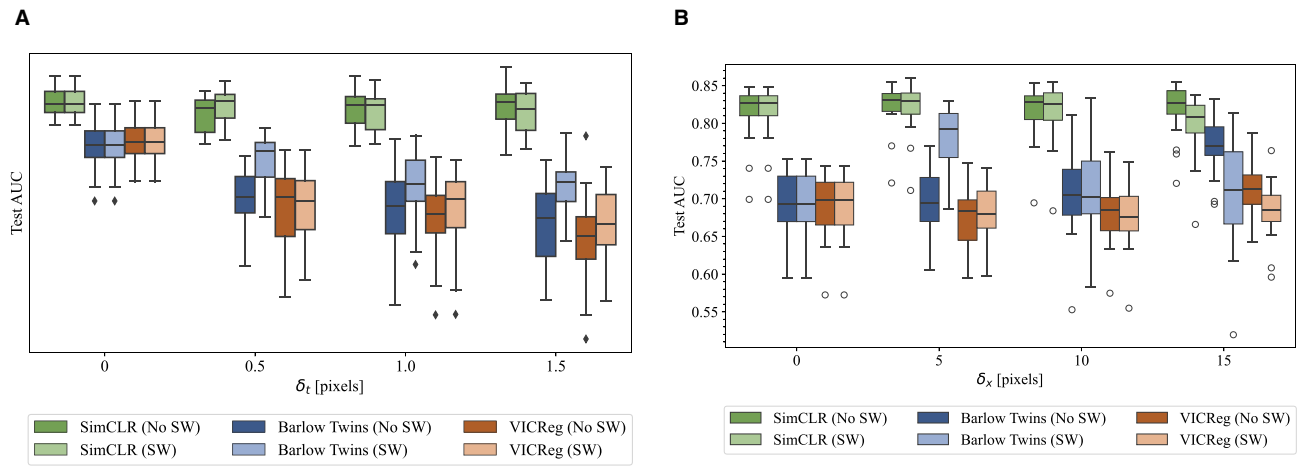


FIGURE 5
Boxplots conveying the quartile ranges of test AUC distributions for each pretraining method and assignment to δ, with and without sample weights. Each experiment is repeated 20 times on disjoint subsets of the training set, each containing all images from a group of patients. **(A)** AB task. **(B)** LS task.

TABLE 3 ParenchymalLUS test AUC for the the AB and LS tasks when trained using examples from 5% of the patients in the training set.

	AB			LS		
Pretrain method	δ_t	SW	Mean (std) test AUC	δ_x	SW	Mean (std) test AUC
SimCLR	0	✗	0.938 (0.007)	0	✗	0.812 (0.037)
	0.5	✗	0.931 (0.010)*	5	✗	0.824 (0.030)*
	0.5	✓	0.936 (0.007)†	5	✓	0.820 (0.033)
	1	✗	0.934 (0.011)	10	✗	0.815 (0.035)
	1	✓	0.933 (0.011)	10	✓	0.816 (0.037)
	1.5	✗	0.936 (0.013)	15	✗	0.819 (0.034)
	1.5	✓	0.932 (0.012)	15	✓	0.798 (0.039)*†
Barlow Twins	0	✗	0.914 (0.014)	0	✗	0.693 (0.044)
	0.5	✗	0.914 (0.010)*	5	✗	0.694 (0.040)
	0.5	✓	0.883 (0.017)*†	5	✓	0.780 (0.040)*†
	1	✗	0.877 (0.022)*	10	✗	0.705 (0.051)
	1	✓	0.891 (0.018)*†	10	✓	0.706 (0.066)
	1.5	✗	0.870 (0.024)*	15	✗	0.769 (0.037)*
	1.5	✓	0.892 (0.015)*†	15	✓	0.707 (0.071)†
VICReg	0	✗	0.917 (0.011)	0	✗	0.690 (0.042)
	0.5	✗	0.879 (0.024)*	5	✗	0.675 (0.036)
	0.5	✓	0.879 (0.021)*	5	✓	0.679 (0.038)
	1	✗	0.872 (0.023)*	10	✗	0.680 (0.039)
	1	✓	0.876 (0.024)*	10	✓	0.675 (0.040)
	1.5	✗	0.860 (0.026)*	15	✗	0.710 (0.036)
	1.5	✓	0.870 (0.021)*†	15	✓	0.685 (0.039)†
None (ImageNet-pretrained)			0.896 (0.017)			0.783 (0.028)
None (random initialization)			0.774 (0.051)			0.507 (0.022)

Twenty trials were performed for each pretraining method, value of δ, with and without sample weights (SW). Mean and standard deviation of the test AUC across trials are reported for each condition. *Significantly different ($p < 0.05$) than baseline for the pretraining method where $\delta = 0$. †Significantly different ($p < 0.05$) for particular δ when sample weights are applied, compared to no sample weight.

IVPP hyperparameter combinations significantly improved the performance of VICReg.

5 Discussion

5.1 Guidelines for practitioners

Insights were derived to guide practitioners working with deep learning for ultrasound interpretation. First, SimCLR was observed to achieve the greatest performance consistently across multiple tasks. With the exception of the data-scarce COVID-19 classification task, SimCLR decisively outperformed Barlow Twins and VICReg on the A/B and LS tasks. The results provide evidence toward favoring contrastive learning over non-contrastive learning for problems in ultrasound. It could be that the non-contrastive methods studied may be less effective for lung ultrasound examinations. We suspect that the lack of diversity in parenchymal lung ultrasound and the fine-grained nature of the classification tasks is problematic for non-contrastive methods, as the objectives are attractive and focus on maximizing embedding information. Perhaps explicit samples of negative pairs may be needed to learn a meaningful embedding manifold for fine-grained downstream tasks. Future work assessing non-contrastive methods for tasks in different ultrasound examinations or alternative imaging modalities altogether would shed light on the utility of non-contrastive methods outside the typical evaluation setting of photographic images.

While the experimental results do not support the existence of overarching trends for hyperparameter assignments for intra-video positive pairs across pretraining methods, it was observed that some combinations improved performance on particular downstream tasks. For example, each pretraining method's downstream performance on COVID-19 classification was improved by a nonzero value of δ_t. Overall, the results indicated that the optimal assignment for IVPP hyperparameters may be problem-specific. Clinically, IVPP may improve performance on downstream ultrasound interpretation tasks; however, practitioners are advised to include a range of values of δ with and without sample weights in their hyperparameter search.

5.2 Limitations

The methods and experiments conducted in this study were not without limitations. As is common in medical imaging datasets, the ParenchymalLUS dataset was imbalanced. The image-wise representation for the positive class was 30.0% for the AB task and 11.7% for the lung sliding task. Although some evidence exists in support for self-supervised pretraining for alleviating the ill effects of class imbalance in photographic images (Yang and Xu, 2020; Liu et al., 2021), computed tomography, and fundoscopy images (Zhang et al., 2023), we found no such evidence for tasks in medical ultrasound.

As outlined in the background, the pretraining objectives employed in this study have been shown to improve downstream performance when the pairwise relationship aligns with the downstream task (Balestriero and LeCun, 2022). These guarantees compare to the baseline case of random weight initialization. While it was observed that all pretraining methods outperformed full supervision with randomly initialized weights, ImageNet-pretrained weights outperformed non-contrastive methods in several of the experiments. ImageNet-pretrained weights are a strong and meaningful baseline for medical imaging tasks, as they have been shown to boost performance in several supervised learning tasks across medical imaging modalities (Azizi et al., 2021). It is possible that some extreme data augmentation transformations and intra-positive pairs could jeopardize the class agreement of positive pairs (as is likely in most pragmatic cases); however, near-consistent alignment was achieved through data augmentation design and small ranges of δ. Although there exists evidence that VICReg and SimCLR can achieve similar performance on ImageNet with judicious selection of hyperparameters (e.g., temperature, loss term weights, learning rate) (Garrido et al., 2022), we used default hyperparameters. Due to limited computational resources, we avoided expansion of the hyperparameter space by only studying IVPP hyperparameters.

Lastly, M-mode images were designated by selecting x-coordinates in B-mode videos that intersect a pleural line region of interest, as predicted by an object detection model utilized in previous work (VanBerlo et al., 2022b, 2023b). LUS M-mode images must intersect the pleural line in order to appreciate the lung sliding artifact. While we mitigated potential inaccuracies in localization by limiting training and evaluation data to the brightest half of eligible x-coordinates, it is possible that a small fraction of M-mode images were utilized that did not intersect the pleural line.

5.3 Conclusion

Intra-video positive pairs have been proposed as a means of improving the downsteam performance of ultrasound classifiers pretrained with joint embedding self supervised learning. In this study, we suggested a scheme for integrating such positive pairs into common contrastive and non-contrastive SSL methods. Applicable to both B-mode and M-mode ultrasound, the proposed method (IVPP) consists of sampling positive pairs that are separated temporally or spatially by no more than a threshold, optionally applying sample weights to each pair in the objective depending on the distance. Investigations revealed that using nearby images from the same video for positive pairs can lead to improved performance when compared to composing positive pairs from the same image, but that IVPP hyperparameter assignments yielding benefits may vary by the downstream task. Another salient result was the persistent top performance of SimCLR for key tasks in B-mode and M-mode lung ultrasound, indicating that contrastive learning may be more suitable than non-contrastive learning methods for ultrasound imaging.

Future work could investigate IVPP for other types of medical ultrasound exams. IVPP could also be integrated into other SSL objectives. The sample weights formulation proposed in this study could also be applied to SSL for non-US videos. Given the high performance of SimCLR, subsequent work should perform a comprehensive comparison contrastive and non-contrastive

SSL methods for tasks in medical US. Lastly, future work could evaluate US-specific data augmentation transformations that preserve semantic content. As a natural source of differences between positive pairs, IVPP could be studied in tandem with US-specific augmentations.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/983632847/USCL. Some of the datasets generated and/or analyzed during the current study are available in via online repositories. The Butterfly dataset and the 5-fold splits of the POCUS dataset can be found in the above USCL repository.

Ethics statement

The studies involving humans were approved by Lawson Research Institute, Western University. 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

BV: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Visualization, Writing original draft, Writing – review & editing. AW: Methodology, Supervision, Writing – review & editing. JH: Methodology, Supervision, Writing – review & editing. RA: Data curation, Resources, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This study was funded by the Natural Sciences and Engineering Research Council of Canada, as BV was a Vanier Scholar (FRN 186945). The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Acknowledgments

Computational resource support was also provided by Compute Ontario (computeontario.ca) and the Digital Research Alliance of Canada (alliance.can.ca).

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/fimag.2024.1416114/full#supplementary-material

References

Alrajhi, K., Woo, M. Y., and Vaillancourt, C. (2012). Test characteristics of ultrasonography for the detection of pneumothorax: a systematic review and meta-analysis. *Chest* 141, 703–708. doi: 10.1378/chest.11-0131

Anand, D., Annangi, P., and Sudhakar, P. (2022). "Benchmarking self-supervised representation learning from a million cardiac ultrasound images," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (Glasgow: IEEE), 529–532.

Arntfield, R., Wu, D., Tschirhart, J., VanBerlo, B., Ford, A., Ho, J., et al. (2021). Automation of lung ultrasound interpretation via deep learning for the classification of normal versus abnormal lung parenchyma: a multicenter study. *Diagnostics* 11:2049. doi: 10.3390/diagnostics11112049

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., et al. (2021). "Big self-supervised models advance medical image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3478–3488.

Balestriero, R., and LeCun, Y. (2022). "Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods," in *Advances in Neural Information Processing Systems*, eds. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (New York: Curran Associates, Inc), 26671–26685.

Bardes, A., Ponce, J., and LeCun, Y. (2022). "VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning," in *International Conference on Learning Representations*.

Basu, S., Singla, S., Gupta, M., Rana, P., Gupta, P., and Arora, C. (2022). "Unsupervised contrastive learning of image representations from ultrasound videos with hard negative mining," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer), 423–433.

Bordes, F., Balestriero, R., and Vincent, P. (2023). Towards democratizing joint-embedding self-supervised learning. *arXiv[preprint] arXiv*:2303.01986. doi: 10.48550/arXiv.2303.01986

Born, J., Brändle, G., Cossio, M., Disdier, M., Goulet, J., Roulin, J., et al. (2020). POCOVID-net: automatic detection of COVID-19 from a new lung ultrasound imaging dataset (POCUS). *arXiv[preprint] arXiv*:2004.12084. doi: 10.48550/arXiv.2004.12084

Butterfly Network (2020). *Covid-19 Ultrasound Gallery*. Available online at: https://www.butterflynetwork.com/covid19/covid-19-ultrasound-gallery (accessed September 20, 2020).

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., et al. (2021). "Emerging properties in self-supervised vision transformers," in *Proceedings*

of the IEEE/CVF International Conference on Computer Vision (Montreal, QC: IEEE), 9650–9660.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning* (New York: PMLR), 1597–1607.

Chen, Y., Zhang, C., Ding, C. H., and Liu, L. (2022). Generating and weighting semantically consistent sample pairs for ultrasound contrastive learning. *IEEE Trans. Med. Imag.* 42, 1388–1400. doi: 10.1109/TMI.2022.3228254

Chen, Y., Zhang, C., Liu, L., Feng, C., Dong, C., Luo, Y., et al. (2021). "USCL: pretraining deep ultrasound image diagnosis model through video contrastive representation learning," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference* (Strasbourg: Springer), 627–637.

Garrido, Q., Chen, Y., Bardes, A., Najman, L., and Lecun, Y. (2022). On the duality between contrastive and non-contrastive self-supervised learning. *arXiv[preprint] arXiv*:2206.02574. doi: 10.48550/arXiv.2206.02574

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inform. Proc. Syst.* 33, 21271–21284. doi: 10.5555/3495724.3497510

Hall, M. K., Hall, J., Gross, C. P., Harish, N. J., Liu, R., Maroongroge, S., et al. (2016). Use of point-of-care ultrasound in the emergency department: insights from the 2012 medicare national payment data set. *J. Ultrasound Med.* 35, 2467–2474. doi: 10.7863/ultra.16.01041

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 9729–9738.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., et al. (2019). "Searching for MobileNetV3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1314–1324. doi: 10.1109/ICCV.2019.00140

Jasčůr, M., Bundzel, M., Malík, M., Dzian, A., Ferenčík, N., and Babič, F. (2021). Detecting the absence of lung sliding in lung ultrasounds using deep learning. *Appl. Sci.* 11:6976. doi: 10.3390/app11156976

Kessler, R., Stowell, J. R., Vogel, J. A., Liao, M. M., and Kendall, J. L. (2016). Effect of interventional program on the utilization of pacs in point-of-care ultrasound. *J. Digi. Imag.* 29, 701–705. doi: 10.1007/s10278-016-9893-x

Lau, Y. H., and See, K. C. (2022). Point-of-care ultrasound for critically-ill patients: a mini-review of key diagnostic features and protocols. *World J. Criti. Care Med.* 11:70. doi: 10.5492/wjccm.v11.i2.70

Liang, H., Ning, G., Zhang, X., and Liao, H. (2023). "Semi-supervised anatomy tracking with contrastive representation learning in ultrasound sequences," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*.

Lichtenstein, D. A. (2010). *Whole Body Ultrasonography in the Critically Ill*. Cham: Springer Science & Business Media.

Lichtenstein, D. A., and Menu, Y. (1995). A bedside ultrasound sign ruling out pneumothorax in the critically iii: lung sliding. *Chest* 108, 1345–1348. doi: 10.1378/chest.108.5.1345

Lichtenstein, D. A., Mezière, G., Lascols, N., Biderman, P., Courret, J.-P., Gepner, A., et al. (2005). Ultrasound diagnosis of occult pneumothorax. *Criti. Care Med.* 33, 1231–1238. doi: 10.1097/01.CCM.0000164542.86954.B4

Liu, H., HaoChen, J. Z., Gaidon, A., and Ma, T. (2021). Self-supervised learning is more robust to dataset imbalance. *arXiv[preprint] arXiv*:2110.05025. doi: 10.48550/arXiv.2110.05025

Nazerian, P., Volpicelli, G., Vanni, S., Gigli, C., Betti, L., Bartolucci, M., et al. (2015). Accuracy of lung ultrasound for the diagnosis of consolidations when compared to chest computed tomography. *Am. J. Emerg. Med.* 33, 620–625. doi: 10.1016/j.ajem.2015.01.035

Oh Song, H., Xiang, Y., Jegelka, S., and Savarese, S. (2016). "Deep metric learning via lifted structured feature embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4004–4012.

Soni, N. J., Arntfield, R., and Kory, P. (2020). *Point-of-Care Ultrasound*. Philadelphia: Elsevier.

Sood, R., Rositch, A. F., Shakoor, D., Ambinder, E., Pool, K.-L., Pollack, E., et al. (2019). Ultrasound for breast cancer detection globally: a systematic review and meta-analysis. *J. Global Oncol.* 5, 1–17. doi: 10.1200/JGO.19.00127

van Randen, A., Laméris, W., van Es, H. W., van Heesewijk, H. P., van Ramshorst, B., Ten Hove, W., et al. (2011). A comparison of the accuracy of ultrasound and computed tomography in common diagnoses causing acute abdominal pain. *Eur. Radiol.* 21, 1535–1545. doi: 10.1007/s00330-011-2087-5

VanBerlo, B., Li, B., Hoey, J., and Wong, A. (2023a). Self-supervised pretraining improves performance and inference efficiency in multiple lung ultrasound interpretation tasks. *arXiv[preprint] arXiv*:2309.02596. doi: 10.1109/ACCESS.2023.3337398

VanBerlo, B., Li, B., Wong, A., Hoey, J., and Arntfield, R. (2023b). "Exploring the utility of self-supervised pretraining strategies for the detection of absent lung sliding in m-mode lung ultrasound," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Vancouver, BC: IEEE), 3076–3085.

VanBerlo, B., Smith, D., Tschirhart, J., VanBerlo, B., Wu, D., Ford, A., et al. (2022a). Enhancing annotation efficiency with machine learning: Automated partitioning of a lung ultrasound dataset by view. *Diagnostics* 12:2351. doi: 10.3390/diagnostics12102351

VanBerlo, B., Wu, D., Li, B., Rahman, M. A., Hogg, G., VanBerlo, B., et al. (2022b). Accurate assessment of the lung sliding artefact on lung ultrasonography using a deep learning approach. *Comp. Biol. Med.* 148:105953. doi: 10.1016/j.compbiomed.2022.105953

Whitson, M. R., and Mayo, P. H. (2016). Ultrasonography in the emergency department. *Crit. Care* 20:1–8. doi: 10.1186/s13054-016-1399-x

Yang, Y., and Xu, Z. (2020). Rethinking the value of labels for improving class-imbalanced learning. *Adv. Neural Informat. Proc. Syst.* 33, 19290–19301. doi: 10.5555/3495724.3497342

Yim, E. S., and Corrado, G. (2012). Ultrasound in sports medicine: relevance of emerging techniques to clinical care of athletes. *Sports Med.* 42, 665–680. doi: 10.1007/BF03262287

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., et al. (2019). Large batch optimization for deep learning: training bert in 76 minutes. *arXiv[preprint] arXiv*:1904.00962. doi: 10.48550/arXiv.1904.00962

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021). "Barlow twins: self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*, 12310–12320.

Zhang, C., Chen, Y., Liu, L., Liu, Q., and Zhou, X. (2022). "Hico: hierarchical contrastive learning for ultrasound video model pretraining," in *Proceedings of the Asian Conference on Computer Vision*, 229–246.

Zhang, C., Zheng, H., and Gu, Y. (2023). Dive into the details of self-supervised learning for medical image analysis. *Med. Image Analy.* 89:102879. doi: 10.1016/j.media.2023.102879

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv[preprint] arXiv*:1710.09412. doi: 10.48550/arXiv.1710.09412

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 03 September 2024
DOI 10.3389/fmed.2024.1416169

OPEN ACCESS

EDITED BY
Simone Bonechi,
University of Siena, Italy

REVIEWED BY
Surjeet Dalal,
Amity University Gurgaon, India
Fred Nicolls,
University of Cape Town, South Africa

*CORRESPONDENCE
Felipe Feijoo
✉ felipe.feijoo@pucv.cl

RECEIVED 11 April 2024
ACCEPTED 20 August 2024
PUBLISHED 03 September 2024

CITATION
Saavedra JP, Droppelmann G, Jorquera C and Feijoo F (2024) Automated segmentation and classification of supraspinatus fatty infiltration in shoulder magnetic resonance image using a convolutional neural network.
Front. Med. 11:1416169.
doi: 10.3389/fmed.2024.1416169

Automated segmentation and classification of supraspinatus fatty infiltration in shoulder magnetic resonance image using a convolutional neural network

Juan Pablo Saavedra[1], Guillermo Droppelmann[2,3], Carlos Jorquera[4] and Felipe Feijoo[1]*

[1]School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile, [2]Clínica MEDS, Santiago, Chile, [3]Harvard T.H. Chan School of Public Health, Boston, MA, United States, [4]Facultad de Ciencias, Escuela de Nutrición y Dietética, Universidad Mayor, Santiago, Chile

Background: Goutallier's fatty infiltration of the supraspinatus muscle is a critical condition in degenerative shoulder disorders. Deep learning research primarily uses manual segmentation and labeling to detect this condition. Employing unsupervised training with a hybrid framework of segmentation and classification could offer an efficient solution.

Aim: To develop and assess a two-step deep learning model for detecting the region of interest and categorizing the magnetic resonance image (MRI) supraspinatus muscle fatty infiltration according to Goutallier's scale.

Materials and methods: A retrospective study was performed from January 1, 2019 to September 20, 2020, using 900 MRI T2-weighted images with supraspinatus muscle fatty infiltration diagnoses. A model with two sequential neural networks was implemented and trained. The first sub-model automatically detects the region of interest using a U-Net model. The second sub-model performs a binary classification using the VGG-19 architecture. The model's performance was computed as the average of five-fold cross-validation processes. Loss, accuracy, Dice coefficient (CI. 95%), AU-ROC, sensitivity, and specificity (CI. 95%) were reported.

Results: Six hundred and six shoulders MRIs were analyzed. The Goutallier distribution was presented as follows: 0 (66.50%); 1 (18.81%); 2 (8.42%); 3 (3.96%); 4 (2.31%). Segmentation results demonstrate high levels of accuracy (0.9977 ± 0.0002) and Dice score (0.9441 ± 0.0031), while the classification model also results in high levels of accuracy (0.9731 ± 0.0230); sensitivity (0.9000 ± 0.0980); specificity (0.9788 ± 0.0257); and AUROC (0.9903 ± 0.0092).

Conclusion: The two-step training method proposed using a deep learning model demonstrated strong performance in segmentation and classification tasks.

KEYWORDS

classification, deep learning, fatty infiltration, MRI, supraspinatus

Introduction

Rotator cuff tears (RCTs) are a prevalent musculoskeletal shoulder condition that affects millions of people worldwide, regardless of sex (1, 2). This degenerative and progressive condition becomes increasingly common with age in the general population (3), leading to significant economic consequences for patients and healthcare systems alike (4, 5). The magnitude of tear size, muscle atrophy, and fatty infiltration are important variables in predicting the prognosis of patients (6, 7). Specifically, low levels of fatty infiltration have been shown to have significantly better outcomes than those with more severe conditions, as they are less likely to experience re-tears (7, 8). Therefore, identifying specific stages of fatty infiltration and the supraspinatus muscle is crucial in accurately predicting patients' prognoses, particularly for those that are to be exposed to a major surgery or in population of high risk with such as older patients. For this purpose, magnetic resonance image (MRI) is one of the most commonly used medical imaging techniques available for the detection of RCT and fatty infiltration, owing to its high diagnostic accuracy (9). However, patient access to MRI results may take several days due to the large number of exams and the time specialists can dedicate to this task. Therefore, developing tools that can speed-up this process, while having a high accuracy in identifying fatty infiltration, can help reduce waiting times suffered by patients and the burden faced by medical experts.

Goutallier et al. (10) proposed one of the most widely used qualitative scales for identifying supraspinatus fatty infiltration, consisting of five stages ranging from 0 (normal muscle) to 4 (severe fat accumulation). Although Goutallier's scale was originally developed based on CT scan analysis, it has been adapted for use with MRI. Fuchs et al. (11) proposed a new scale by combining the previously defined stages in Goutallier's work. Specifically, levels zero and one were merged to create the normal stage, level two was redefined as moderate, and level three or four were considered to represent severe fatty infiltration. However, there has been some controversy over the adaptation of the original scale for use with MRI (12). Furthermore, reducing inter-observer variability when assessing rotator cuff quality from MRI remains a major challenge in diagnostic imaging (13).

On the other hand, deep learning algorithms, especially convolutional neural networks (CNNs), have rapidly become the preferred methodology for analyzing medical images (14–16). Some of the most commonly used deep learning architectures for computer vision tasks include Inception-v3, ResNet50, VGG19, and U-Net (17–20). However, due to complexity of medical image datasets and smaller size compared to other sources of data, transfer learning has become a suitable approach for building and training deep learning models in clinical research. With transfer learning, most of the proposed models for medical diagnosis are based on pre-trained models from the ImageNet dataset and trained using transfer learning techniques (21). This technique involves using a well-trained model from a non-medical source dataset, such as ImageNet, and re-training it in a target dataset, such as medical images, including MRIs (22–24).

Most of the existing deep learning applications are based on supervised training, a commonly used technique for classification using medical images. However, supervised training requires labeled images for the models to learn from their structure. Additionally, in supervised learning, in order to improve the model's performance, researchers manually select the region of interest (manual segmentation). However, manual segmentation is a time-consuming task, and manual labelling from medical experts is not always available (25). Therefore, to address these limitations, unsupervised training for segmenting the region of interest could be a viable solution. In the context of identifying shoulder fatty infiltration, four recent and highly important articles addressing this problem or closely related have been published. Three of these studies focused on magnetic resonance images (22, 23, 26) while only one utilized CT scans (27). However, all these studies relied on annotated data, which means that each image was manually labeled by an expert to create an image and corresponding infiltration level pairs, or each image was manually segmented to generate a corresponding segmentation mask for that specific image.

In order to address the gap in the literature, the objective of this research is to develop and assess a two-step deep learning framework. The first step performs and automated detection the region of interest (segmentation of the region of interest), while the second step uses the information from the segmentation model to classify the region of interest into one of the Goutallier's fatty infiltration levels using MRI images, hence, fully automating the process of identifying the Goutallier's fatty infiltration levels via the usage of deep learning techniques (segmentation and classification hybrid framework).

Materials and methods

Study design

This research was designed as a retrospective, single-site study, following the guidelines outlined in the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE). Patient records were exclusively obtained from MRI examinations conducted at the MEDS Clinic in Santiago, Región Metropolitana, Chile. The study started on September 25th, 2020.

Learning approach

An end-to-end deep learning model was developed to classify the patient risk based on the fatty infiltration of the supraspinatus muscle. The training process was performed in a two-step fashion. In the first step, we trained a segmentation model to extract the region of interest from the image. In the second step, we trained a classification model to determine if there was a risk or not for further surgery based on the level of fatty infiltration in the region of interest detected in the first step. Both models (segmentation and classification) are trained independently and non-recursively. However, segmented images from the first step (segmentation model) are used to train the classification model. Therefore, the training process of the classification model, as well as the testing phase, are performed using results from the segmentation model (segmented images). The training process and workflow of the proposed two-step model is described in Figure 1 as well as in Figure 2.

Dataset characteristics

The medical institution provided all the data, consisting of 900 DICOM files corresponding to unique exams. Each file corresponds

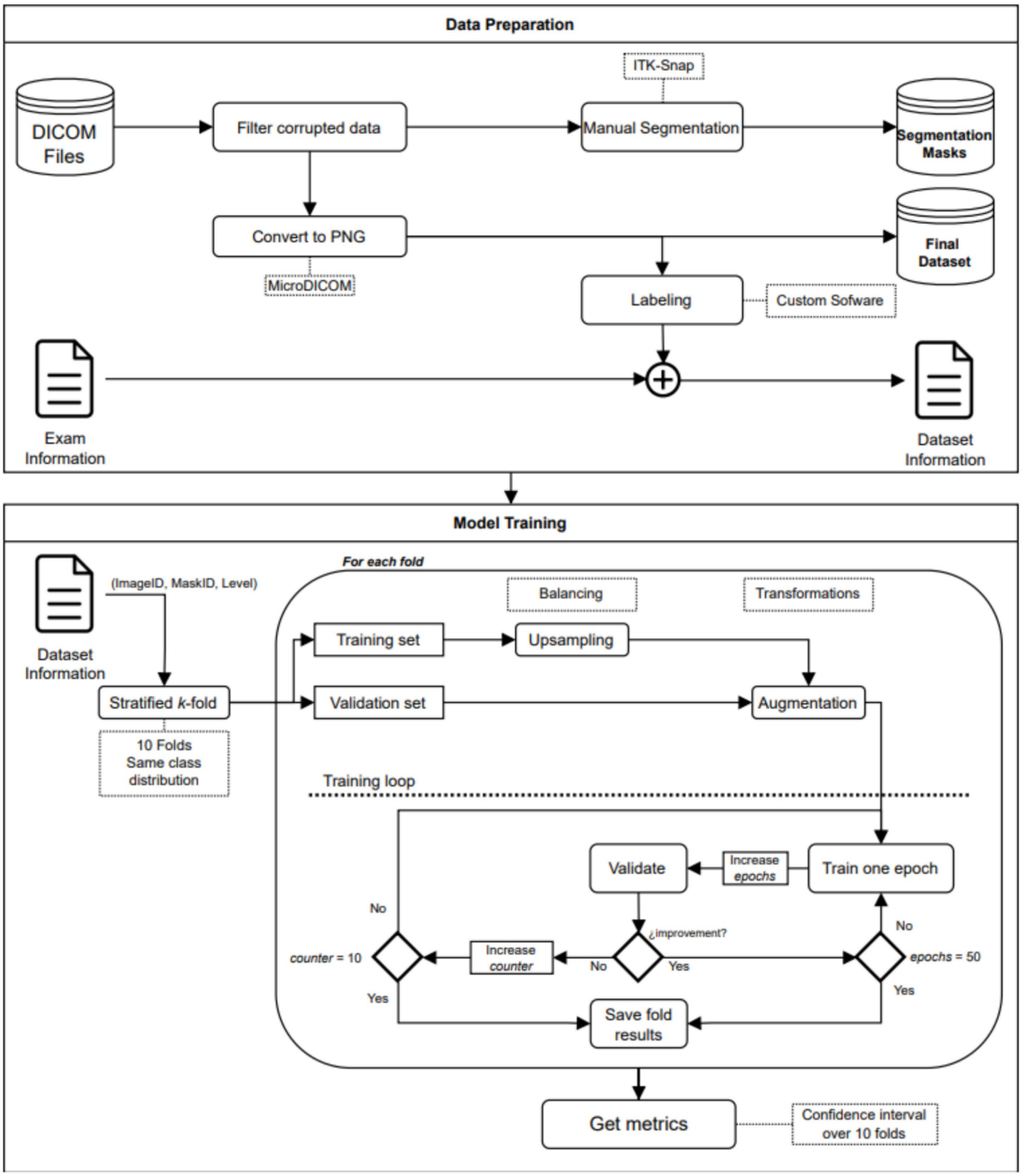


FIGURE 1
Workflow diagram.

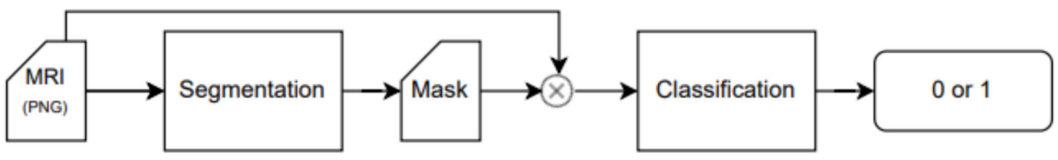


FIGURE 2
Diagram of the end-to-end model sequence.

to a T2-weighted Y-view MR sequence of the shoulder. Furthermore, we extracted all 900 medical reports associated to each of the DICOM files. The medical reports were, authored by three different radiologists. These reports used various scales or standards to document the fatty infiltration or degeneration stage. To ensure accurate labeling, we enlisted the expertise of an experienced radiologist who manually labeled the dataset. Moreover, images with diagnostic uncertainties underwent manual segmentation under the supervision of another radiologist, ensuring detailed and reliable annotations.

According to Figure 3, the labeling process resulted in 666 registered images, with one being marked as inconclusive and two remaining unregistered. Additionally, there were 60 images for which segmentation masks could not be created due to a file error. Consequently, our ground truth dataset comprises 606 labeled images along with their corresponding segmentation masks. Table 1 provides an overview of the image label counts, indicating 403, 114, 51, 24, and 14 for Goutallier 0, 1, 2, 3, and 4, respectively. More than 82% of the images fall into grades 0 or 1, indicating a significant imbalance towards lower fatty infiltration grades. The female group exhibited a greater number of samples in the higher grades compared to the male group. Furthermore, except for the observed mean age in the Goutallier 0 group ($p < 0.05$), there were no significant differences between the female and male groups across Goutallier levels in terms of proportions or mean age.

Dataset preparation

The DICOM file format is extensively adopted as a standard for medical images in clinical settings. A DICOM data object consists of multiple attributes, including fields such as name, ID, and more. It also incorporates a distinct attribute that contains the image pixel data. In order to enhance the efficiency of image processing during model ingestion, we extracted the pixel data from every DICOM file and converted it to PNG format. This extraction process was facilitated by MicroDICOM, a freely available software for viewing DICOM files.

The ITK-Snap3 software was utilized to generate the segmentation masks. In this case, separate masks were created for the supraspinous fossa area and the supraspinatus muscle area. Considering the specific evaluation of the fatty infiltration grade of the muscle based on the muscle area alone by physiologists, the focus was directed towards the supraspinatus muscle area mask for the subsequent steps. The final outcome of the segmentation process is visualized in Figure 4.

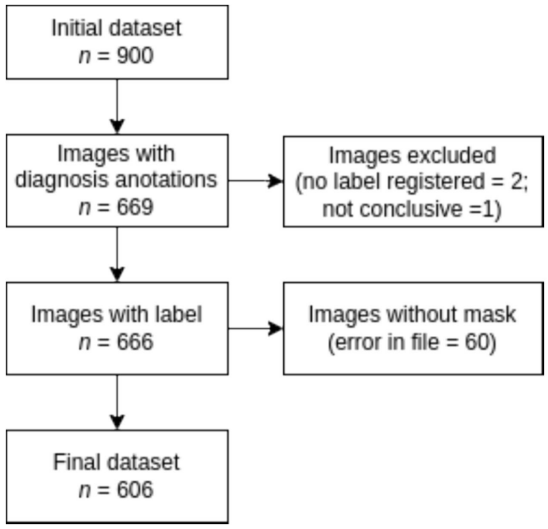


FIGURE 3
Flowchart for dataset selection.

The data preparation process resulted in multiple images in PNG file format, each accompanied by its corresponding segmentation mask and label. Figure 1 provides a visual representation of the workflow involved in the data preparation.

Criteria for fatty infiltration

The criteria were based on Goutallier's fatty infiltration definitions. According to the original paper, five levels of fatty infiltration were proposed, ranging from zero to four, to signify the qualitative presence of fat in the muscle. A level zero indicates the absence of fat in the muscle, while higher levels correspond to increasing fatty infiltration. Goutallier's scale assigns higher values as the fatty infiltration intensifies. A level four indicates a higher amount of fat than muscle present.

As mentioned earlier, the objective is to assist clinicians in determining the risk associated with performing surgery based on the quality of the supraspinatus muscle. From a classifier perspective, this task can be viewed as a binary classification. In this study, Goutallier's fatty infiltration levels zero or one were classified as "not risky," while levels three or four were categorized as "risky." Samples labeled as Goutallier level two were excluded from the analysis. This choice is based on previous research [see Saavedra et al. (20)] where it is shown that including Goutallier's level 2 into a binary classification task does not significantly impact the performance of a classification model. Also, clinical relevance falls in correctly those cases where there is high or low level of fatty infiltration [see references (10) and (11)].

Proposed model

The proposed model is composed of two sequential neural network models that serve distinct purposes. Model A is designed to narrow down the region of interest in the MRI image by leveraging both the image and the segmentation mask as inputs. The U-Net model is proposed for this task (see next). Its primary objective is to predict the supraspinatus muscle area. The hypothesis is that this approach effectively eliminates irrelevant information from the image, thereby enhancing the performance of the second network. Following Model A (segmentation), Model B (classification task) takes the supraspinatus muscle area of the image as input and predicts the fatty infiltration level based on the Goutallier's fatty infiltration level scale. An overview of the workflow is provided in Figure 2, while the subsequent subsections offer a detailed explanation.

Cross validation (k-fold) was performed during the training process. The total of 606 Y-view MRI shoulder images were grouped into five non overlapping folds. Each time, four folds were used as the training set and one as the validation set. Every fold was used four times as part of the training set and one time as part of the validation set. Fold composition was the same for both models (Model A and Model B). Model performance was computed as the average of those five training processes and 95% confidence intervals (CI) were obtained. In every training process the model with the lowest loss function value was considered the best model.

TABLE 1 Patient data distribution Goutallier's level by sex.

Goutallier level	*N* (%)	Female		Male		*p*-value	
		N (%)	Age mean (SD)	*N* (%)	Age mean (SD)	*N*	Age
0	403 (66.50)	140 (35)	53.06 (10.55)	263 (65)	49.24 (13.13)	0.477	***
1	114 (18.81)	74 (65)	61.50 (10.37)	40 (35)	63.58 (8.17)	0.465	0.371
2	51 (8.42)	31 (61)	66.65 (9.53)	20 (39)	66.40 (10.13)	0.447	0.992
3	24 (3.96)	16 (67)	68.88 (7.74)	8 (33)	64.25 (7.59)	0.424	0.230
4	14 (2.31)	13 (93)	67.31 (7.33)	1 (7)	N.A.	0.354	0.8
Total	606 (100)	274 (45)	58.47 (11.67)	332 (55)	52.42 (13.81)	0.483	

Mann–Whitney or *t*-test were used to compute the significance (alpha 0.05).

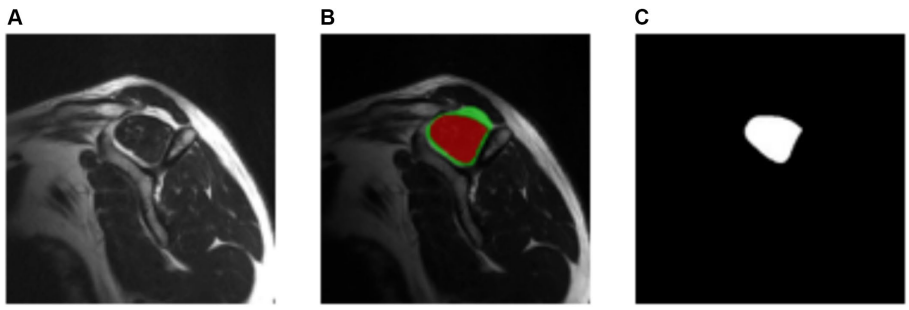


FIGURE 4
Manually segmentation process. **(A)** Original image. **(B)** Resulting segmentation masks. Supraspinous fossa in green, supraspinatus muscle in red. **(C)** Supraspinatus muscle mask.

Model development and training

The proposed model was built using two sequenced architectures: U-Net (28) (Model A) and VGG-19 (29) (Model B). The first sub-model created the segmentation mask of the input image, and the second, performed the fatty infiltration classification for that same image. The selection of the VGG-19 model for the classification task is supported by previous research [see reference (20)] where it is shown that the VGG-19 is among the best CNN for fatty infiltration (among the tested models). Although the proposed framework follows sequential stages, the training process was performed in two steps. In the first step, we trained the segmentation model using every image and the corresponding segmentation mask as input.

The objective was for the model A to learn to predict the corresponding segmentation mask for an image that had not been seen previously. In the second step, a classification model was trained using the region of interest of the image and its corresponding label. Before feeding the classification model, automatic cropping of the image was performed, and only the region of interest was used as input for the classification model.

A repeated stratified *k*-fold cross-validation was performed in both steps. This method allowed us to use the entire dataset in the training process and minimize the influence of data selection, as occurs when using random train/validation/test splitting. The *k* value was set equal to 5 and, therefore, 5 non-intersecting groups were created at random. The proportion of every class in the original dataset was replicated in every group. Each time, four groups were used to create the training set and one was used to create the validation set.

The model performance was computed as the average of 5 training processes, and the corresponding confidence intervals were reported. Confidence intervals obtained from the cross-validation training process was used to assess robustness of the trained models. Due to the high imbalance of the dataset, the minority class was up sampled. In every training process, the smaller class was replicated until the proportion between classes was close to 1:1. The added images were copies of their originals but with slight differences in terms of rotation (±35°), horizontal flipping, and center cropping. The up-sampling process was carried out for the training data only. Figure 1 shows the workflow of the model training process.

Step 1: Training the segmentation model. For the segmentation task, a "U"-shaped neural network was built as described in Khouy et al. (28). The only difference is that (1, 1) padding was used in every convolutional layer to allow the network to utilize the entire image during the training process. The model was training for a maximum of 50 epochs and feeding the network with batches of five images at a time. We used binary cross-entropy loss, implemented in the PyTorch framework. The optimization algorithm used was Adam optimizer with its standard configuration. The learning rate was set to 10-5.

The segmentation process was performed using the U-Net model. The training hyperparameters were as follows: batch size = 8, maximum epochs = 50, input size = 224 × 224 (px), learning rate = 10^{-3}, optimizer = Adam (standard configuration). The loss function used was the Dice loss, which was defined as:

$$\text{Dice score} = 2 \times p \times t / \left(p^2 + t^2\right) \quad (1)$$

$$\text{Dice loss} = 1 - \text{Dice score} \quad (2)$$

In Equation 1, "p" represents predicted values from the output, and "t" represents true values from the input. Basically, the Dice score (see Equation 2) measures the ratio of the intersection over the union for the resulting segmentation mask (30). The better the performance of the segmentation model, the higher the Dice score value. On the other hand, the Dice loss is the function to be minimized. The higher the value of the Dice score, the lower the value of the loss function.

Step 2. Training the classification model: The VGG-19 architecture was used for the classification task. We kept the convolutional layers of the model as the original and only the last layer of the fully connected layers was changed. Originally, the output of the VGG-19 architecture was 1,000 neurons. In our case we use only one output unit. That way, the model was able to perform the binary classification of the inputs.

To train the model, we used transfer learning. This means that all the weights of the original models trained on the ImageNet dataset were utilized. These weights were not optimized during the training process, and only the classifier layers were optimized. We employed the same maximum number of epochs, batch size, loss function, and optimizer as in the segmentation training process. A termination function was implemented to stop the training process if there was no improvement in the last 10 epochs. The best performance was saved and recorded. The only hyper-parameter that was optimized was the learning rate, and the best performance was achieved at 10^{-5}. In the following section, we will present the output of both models, including the segmentation mask and a detailed explanation of the obtained metric values.

Statistical analysis

Normality tests were conducted, and the analysis of statistical differences between groups utilized either the Mann–Whitney U test or t-test. A significance level of $p < 0.05$ was employed to establish statistical significance. Descriptive analysis of patient ages was performed, presenting the mean and standard deviation ($m \pm sd$). Categorical data were expressed as percentages and frequencies.

The performance of the models was evaluated and compared based on accuracy, sensitivity, specificity, and area under the receiver operator curve (AU-ROC). A binary classifier produces either 0 or 1 for a given input, corresponding to the actual expected output. True positive (TP) was defined as the model correctly predicting the positive class. False positive (FP) refers to the model incorrectly predicting the positive class when it is actually negative. False negative (FN) occurs when the model incorrectly predicts the negative class when it is actually positive. True negative (TN) is when the model correctly predicts the negative class. Sensitivity, specificity and accuracy (Equations 3–5), were computed as follows:

$$\text{Sensitivity}\left(\text{true positive rate}\right): \text{TP} / \left(\text{TP} + \text{FN}\right) \tag{3}$$

$$\text{Specificity}: \text{TN} / \left(\text{TN} + \text{FP}\right) \tag{4}$$

$$\text{Accuracy}: \left(\text{TN} + \text{TP}\right) / \left(\text{TN} + \text{FP} + \text{FN} + \text{TP}\right) \tag{5}$$

The AU-ROC measures the classifier's performance regardless of the threshold used to convert probability scores into class decisions. The horizontal axis represents recall (sensitivity), while the vertical axis corresponds to precision, calculated as TP/(TP + FP). As both axes range from 0 to 1, the maximum value of the area under the curve inside the square is 1, indicating better classifier performance. A random classifier would have an AU-ROC equal to 0.5.

For metrics such as accuracy, sensitivity, specificity, and AU-ROC, 95% confidence intervals over the mean were calculated to assess model performance. All statistical analyses were conducted using the Python programming language.

Results

Sociodemographic characteristics

Male subjects presented 333 images, representing 55% of the sample. The patient's average age was 55.1 ± 13.2 years. The data showed the presence of various types of Goutallier levels in MRI exams. An asymmetrical distribution of Goutallier grades was identified. A significant majority, exceeding 82% of the images, fell into grades 0 and 1, indicating a notable prevalence of low fatty infiltration: Goutallier 0 (66.50%), Goutallier 1 (18.81%), Goutallier 2 (8.42%), Goutallier 3 (3.96%), and Goutallier 4 (2.31%). Furthermore, the female group exhibited a higher frequency of samples in higher grades compared to the male group, although this disparity did not reach statistical significance. For more information, refer to Table 1.

Step 1. Segmentation

At the outset of the training process, the loss value was recorded at 0.8498 ± 0.0102, serving as an initial baseline for assessing the model's performance. As training progressed through successive epochs, a consistent reduction in the loss value was observed. Ultimately, post-training, the loss value significantly decreased to 0.0623 ± 0.0050. The training loss value (and other performance metrics) can be observed in Figure 5.

The substantial decline in the loss value reflects a considerable improvement in the model's predictive accuracy. The reduction over the epochs suggests that the model became increasingly proficient at minimizing errors and refining its predictions. The tight standard deviations associated with the initial and final loss values underscore the reliability and consistency of the observed improvements.

These results imply that the deep learning model underwent effective training, optimizing its ability to generalize patterns and make accurate segmentation tasks. The detailed evolution of the loss value throughout the epochs provides a quantitative measure of the model's learning process and its enhanced performance at the training's conclusion.

The segmentation task performed by the model can be observed in Figure 6. The original input mask is highlighted in red, and the model's output mask is highlighted in green. The background of each case displays the original image. Before making modifications, the images were rotated before being fed into the segmentation model. This rotation aims to prevent the model from memorizing specific patterns and, instead, encourages it to learn more generalized concepts from the data.

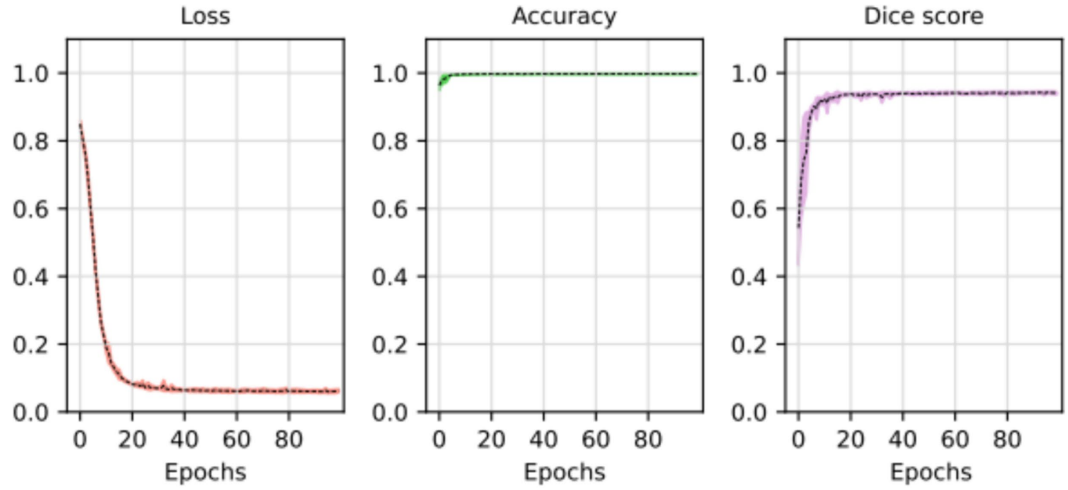


FIGURE 5
Loss, accuracy, and Dice score for the segmentation model. The average of the five training processes is shown in segment line. The color shadow shows the confidence interval (C.I. 95%).

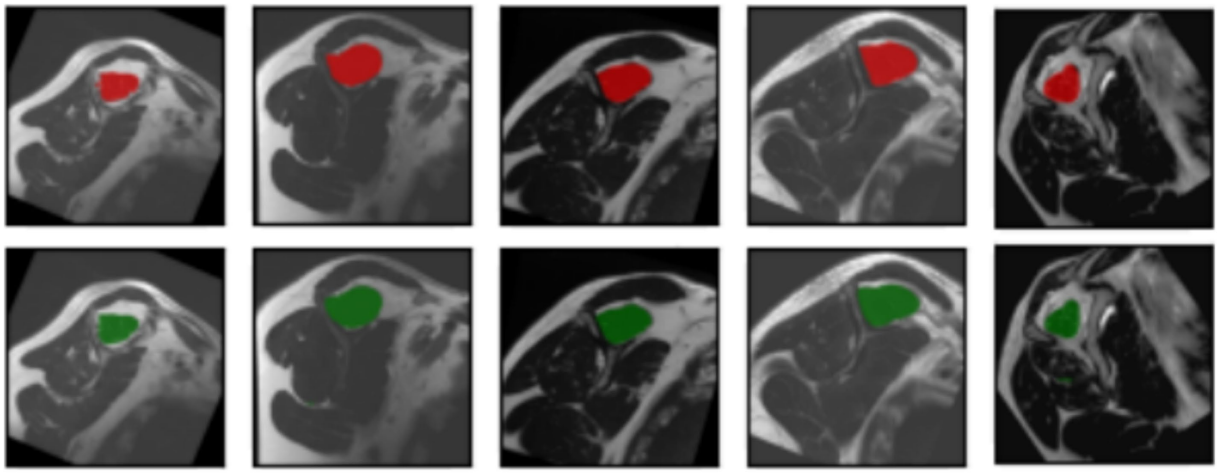


FIGURE 6
Input masks and the respectively, output masks obtained from the U-Net model. The original masks are shown in red; the resulting masks are shown in green. For each input image showed in every column of the first or the third row, the corresponding output mask from the U-Net is showed on the same column in the second and fourth rows, respectively.

In most cases, the resulting segmentation mask (in green) closely resembles the original input segmentation mask (in red). This suggests that the model effectively learned to perform the segmentation task without memorizing specific samples from the training dataset. The similarity between the masks indicates that the model has generalized correctly and can apply its knowledge to new images effectively. In this sense, the model efficiently minimized errors during the training process, as indicated by the computed average loss value of 0.0587. This low loss value is crucial because it signifies the model's ability to consistently converge toward accurate predictions. The small standard deviation of 0.0048 further emphasizes the precision and stability of the model's training, reinforcing its reliability in capturing intricate patterns within the data. At the same time, the model shows its proficiency in correctly classifying instances with an average accuracy of 0.9977. With a minimal standard deviation of 0.003, the model also shows consistent accuracy across various data points. These findings highlight the robustness of the model in performing precise segmentation tasks. Finally, the model achieved an average Dice score of 0.9441, indicative of its efficacy in capturing the spatial agreement between predicted and ground truth segmentations. A small standard deviation of 0.0035 shows the model's stability in consistently achieving high Dice scores. These results affirm the model's performance in image segmentation tasks. For more details, please refer to Table 2.

Step 2. Classification

Figure 7 shows the original image (A) and the segmentation mask obtained from the U-Net model (B). Then using that segmentation mask, the region of interest was cropped (automated process) from the original image (C). Finally, a resizing function was applied to the

image, resulting in (D). This pre-processing allowed the model to decide considering only the supraspinatus muscle, similarly as how the clinicians do.

During the training process, the loss function value for the validation set was monitored. At the beginning of the training process the loss value was 0.6645 ± 0.0228, decreasing to 0.01178 ± 0.0037 after the training process was concluded. The accuracy, sensitivity, specificity and AUROC were computed as the average of the model performance over the validation set in each of the five training processes of the *k*-fold. Table 3 shows the results for those metrics in terms of the confidence interval ($\alpha = 0.05$). As shown, every metric value is above 0.9 (on average), hence showing a good binary classification performance of fatty infiltration of the supraspinatus muscle based on Goutallier's fatty infiltration scale. In particular, the accuracy reached a level of 97.3% with a 0.023 95% CI, showing high precision (low variability). Even though the results show a higher value of specificity compared to sensitivity, the difference could increase if no oversampling (or other data-balancing technique) was used. In this case, sensitivity reached a level of 90% with 0.98 95% CI, while the sensitivity showed a high level of 97.9% with a low 95% CI of 0.02. Finally, the balancing of these two metrics was computed by the AU-ROC, which has an average level of 99% with a low 95% CI of 0.009, indicating a high level of capability to differentiate risky from non-risky levels of fatty infiltration based on automated segmented images from the U-Net model (see Figures 8, 9).

Results of the proposed automated two steps training model shows that the segmentation model could first learn how to find the region of interest (supraspinatus muscle). Then, the classification model could learn how to classify the input, based on that region of interest, as risky or not risky. Cropping the region of interest before feeding the classifier, allowed the model to learn as clinicians do. However, the two step process proposed here shows a small reduction in classification performance (sensitivity, specificity, accuracy and AU-ROC) when compared to different CNN trained on the same data but considering manual segmentation of the ROI [see Saavedra et al. (20) for details]. Table 4 shows the comparison of the two step proposed model (U-NET + VGG-19) with VGG-19, ResNET-50 and Inception-v3 models. As noted, given that manual segmentation done by professional clinicians and medical expert is more accurate that segmentation performed by U-NET, errors from the U-NET model are passed on to the VGG-19 classification model, resulting a slightly lower performance. However, the (almost insignificant) reduction of performance is valid as the proposed model completely automates the process of identifying the level of fatty infiltration, reducing hence the need for lengthily process of manual segmentation of the ROI of the supraspinatus muscle.

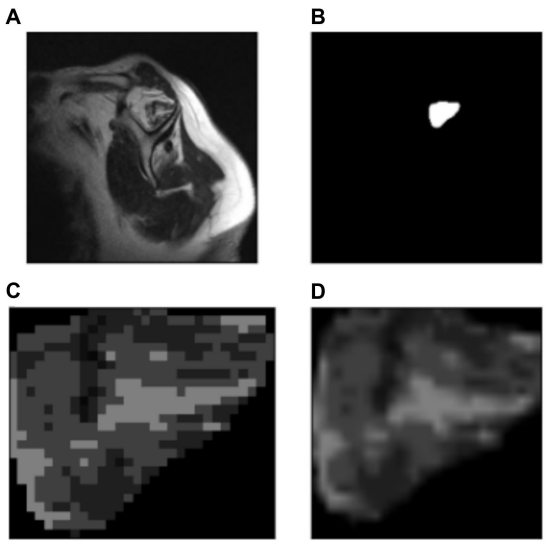


FIGURE 7
Automatic cropping process. **(A)** Original image. **(B)** Output mask from the U-Net model. **(C)** Cropped region of interest from the original image (ROI). **(D)** Resized region of interest (224 × 224 px).

TABLE 2 Segmentation results.

	Loss	Accuracy	Dice score
Average	0.0587	0.9977	0.9441
S.D	0.0048	0.003	0.0035
CI. (95%)	0.0586 ± 0.0042	0.9977 ± 0.0002	0.9441 ± 0.0031

Loss, accuracy, and Dice score were computed as the average of five training processes. Confidence interval calculated at $\alpha = 0.05$.

Discussion

This article introduces a novel deep-learning framework for assessing the degree of fatty infiltration in the supraspinatus muscle. The framework performs two main tasks: segmenting the region of interest and classifying the level of fatty infiltration on a five-level scale proposed by Goutallier et al. (10) based on the automated segmentation process. To achieve this, we developed two sub-models: the first based on the U-Net architecture for segmentation, and the second based on the VGG-19 architecture with modified classifier layers for binary classification. We first trained the segmentation sub-model using segmentation masks and then trained the classification sub-model using the labels associated with the fatty infiltration diagnosis. We used transfer-learning weights to train both sub-models. The binary output of the model (0 or 1) was interpreted as "not risky" or "risky," respectively, with higher levels of fatty infiltration indicating a greater risk of re-tear or poor surgical outcomes.

TABLE 3 Classification model results.

	Loss	Accuracy	Sensitivity	Specificity	AUROC
Average	0.1065	0.9731	0.9000	0.9788	0.9903
S.D	0.0584	0.0263	0.1118	0.0293	0.0105
CI. (95%)	0.1065 ± 0.0512	0.9731 ± 0.0230	0.9000 ± 0.0980	0.9788 ± 0.0257	0.9903 ± 0.0092

Confidence interval computed from the validation set of the five training processes at $\alpha = 0.05$. The loss, accuracy, sensitivity, specificity, and area under the ROC curve (AUROC) are shown.

Our model achieved strong performance thanks to the implementation of transfer learning and k-fold cross-validation techniques. By leveraging these approaches, we were able to reduce the number of parameters requiring optimization and utilize the full dataset for both training and validation purposes, effectively guarding against overfitting issues given our relatively small dataset of slightly more than 600 samples. However, some research has made efforts to optimize the process of hyperparameter optimization (31). Still, it's worth noting that relying on transfer learning from a pre-trained model on the ImageNet dataset may not always represent the most ideal solution. This can be seen as a possible limitation of the relatively small sample of images obtained for this study. Future research should focus on evaluating the effect of the proposed training process. This is needed to understand if the high accuracy levels obtained in this research are driven by transfer learning and data augmentation techniques or to identify if the task or segmenting and classifying fatty infiltration in the supraspinatus muscle is a simpler task compared to more complex images (such as X-rays or ultrasounds of different body or biological structures).

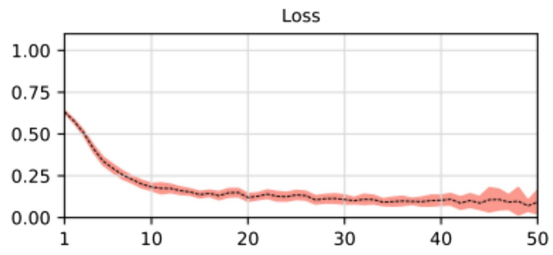


FIGURE 8
Average and confidence interval ($\alpha = 0.05$) for the classification validation loss over five-folds cross-validation training processes. Average is shown in segmented line, and confidence interval is shown in shadow.

In the medical domain, obtaining labeled data or segmentation masks for images can be challenging. Meanwhile, radiological reports are abundant and readily available. Manual labeling or segmentation is a labor-intensive process, but leveraging the valuable information contained in reports can facilitate model training without significant human effort. Another approach worth considering is unsupervised learning, which can enable the model to learn without relying on fully labeled or segmented data. Additionally, using transfer learning with a pre-trained model in a related domain, such as shoulder MRI images or MRI images more broadly, has the potential to enhance the model's performance.

Deep learning models have been increasingly applied in radiology, with the U-Net (28) being a particularly popular choice for segmentation tasks. One example of this is Taghizadeh et al. (27), who employed the U-Net model to assess muscle degeneration levels in CT scans. Through a supervised training approach with annotated data, they successfully segmented the structures and characterized the pre-morbid state based on clinical information. By comparing these two states, they were able to quantify the degree of muscle degeneration.

Medina et al. (22) proposed two sequential models trained in a supervised manner via transfer learning from a model pre-trained on the ImageNet dataset. Both models had all their weights initially frozen except for the classifier layers, which were optimized by training the network on a shoulder MRI dataset. Model A aimed to identify the best image in a series depicting the rotator cuff muscles, while Model B focused on segmenting the four rotator cuff muscles. Model A was constructed using the Inception-v3 architecture, while Model B was based on the VGG19 architecture.

Kim et al. (26) proposed a unique approach for assessing muscle atrophy in the supraspinous fossa by measuring the occupation ratio

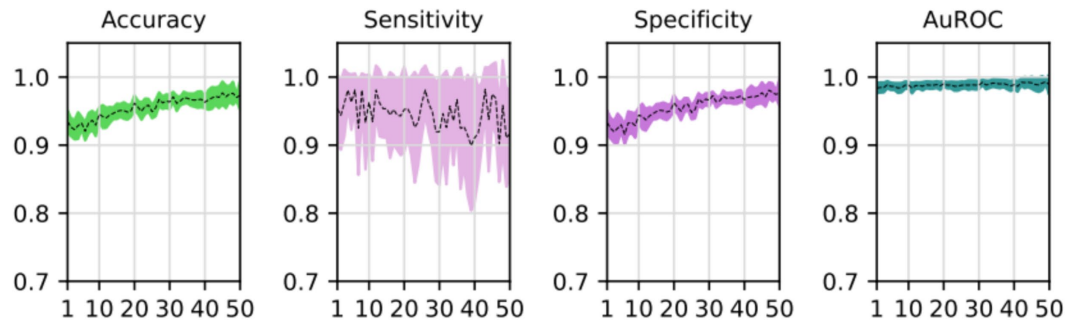


FIGURE 9
Average and confidence interval ($\alpha = 0.05$) for the classification validation accuracy, sensitivity, specificity, and area under the ROC curve (AUROC), over five-folds cross-validation training processes. Average is shown in segmented line, and confidence interval is shown in shadow.

TABLE 4 Classification model comparison with literature.

	Loss	Accuracy	Sensitivity	Specificity	AUROC
Proposed model	0.106 ± 0.051	0.973 ± 0.023	0.900 ± 0.098	0.978 ± 0.025	0.990 ± 0.009
VGG-19	0.096 ± 0.010	0.973 ± 0.006	0.947 ± 0.039	0.975 ± 0.006	0.991 ± 0.003
ResNet-50	0.123 ± 0.011	0.976 ± 0.006	0.925 ± 0.053	0.980 ± 0.006	0.992 ± 0.003
Inception-v3	0.102 ± 0.009	0.974 ± 0.007	0.869 ± 0.085	0.981 ± 0.006	0.991 ± 0.004

Confidence interval computed from the validation set of the five training processes at $\alpha = 0.05$. The loss, accuracy, sensitivity, specificity, and area under the ROC curve (AUROC) are shown. Models used for comparison are obtained from Saavedra et al. (20).

(O.R.) of the supraspinatus muscle. They used a VGG19-like network to segment the region of interest with annotated data, but gaps in the muscle area obtained from the model required filling with a post-processing algorithm. The authors then determined the stage of muscle atrophy based on the O.R. (stage I: O.R. ≥ 0.6; stage 2: $0.4 \leq$ O.R. ≤ 0.6; stage 3: O.R. > 0.4). Although this method did not assess the fatty infiltration grade precisely, it was still a valuable contribution.

Ro et al. (23) also utilized the VGG19 model to perform a segmentation task for identifying the region of interest. To convert the grayscale image into a binary representation, they applied Otsu's thresholding (32), a technique commonly used to separate the foreground (fat) from the background (muscle) in the image. However, as in other studies, post-processing was required, and the results were not directly applicable to a fatty infiltration scale like Goutallier's.

This study has some limitations that must be considered. Firstly, a domain bias might have been introduced to the prediction because the MRI images and natural images used in the training process came from very different dataset. While we used the cross-validation technique to overcome the over-fitting problem, we were unable to test our data on an external dataset, which could limit the model's generalizability if it is intended to be used in a production environment. To address this issue, future studies could focus on training the model on a larger set of MRI images to improve both the model's performance and the clinician's reliance on an artificial intelligence-driven solution. Also, it is important to consider that in order to bring these new models and technologies to production environment (deployment), computational resources must be considered as the models must be retrained as new data comes in. This also helps improving and refining the deployed models. To properly do this, deployment environments (hospitals or clinics) must be equipped with appropriate computational tools (servers or computers) to efficiently manage the update of models, which also increase in complexity and computational resources needed as more data becomes available. Additionally, the manual labeling task was performed by only one trained radiologist, which might limit the reliability of the ground truth. To improve the accuracy and consistency of the labeling process, future studies could consider involving multiple trained radiologists in the task and comparing the model's performance with that of the professionals. Finally, further efforts should be pursued to evaluate the feedback-loops during the training process of the proposed two-stage algorithm. This research did not focus on the possible improvements of the segmentation and classification models when feeding their results and predictive errors, similar to what boosting or sequential machine learning algorithms do.

In summary, this study analyzed a dataset of MRI images to assess fatty infiltration levels in the supraspinatus muscle among patients with rotator cuff conditions. We proposed a two-step training method using deep learning models, which demonstrated strong performance in segmentation and classification tasks. These findings indicate the potential of these models for accurate and reliable evaluation of musculoskeletal conditions in similar clinical settings.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: data might be requested to authors and it will be sent if authorized by corresponding authorities as they are images of patients. Requests to access these datasets should be directed to guillermo.droppelmann@meds.cl.

Ethics statement

The studies involving humans were approved by Comité de Ética Científico Adultos, Servicio de Salud Metropolitano Oriente, Santiago, Chile. 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

JS: Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. GD: Writing – original draft, Writing – review & editing. CJ: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. FF: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The publication was financially supported by the Universidad Mayor.

Acknowledgments

The authors are grateful for the kind collaboration and assistance of the Sports Medicine Data Science Center MEDS-PUCV.

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

References

1. Sabo MT, LeBlanc J, Hildebrand KA. Patient gender and rotator cuff surgery: are there differences in outcome? *BMC Musculoskelet Disord*. (2021) 22:838. doi: 10.1186/s12891-021-04701-y

2. Yamamoto A, Takagishi K, Osawa T, Yanagawa T, Nakajima D, Shitara H, et al. Prevalence and risk factors of a rotator cuff tear in the general population. *J Shoulder Elb Surg*. (2010) 19:116–20. doi: 10.1016/j.jse.2009.04.006

3. Lazarides AL, Alentorn-Geli E, Choi JHJ, Stuart JJ, Lo IKY, Garrigues GE, et al. Rotator cuff tears in young patients: a different disease than rotator cuff tears in elderly patients. *J Shoulder Elb Surg*. (2015) 24:1834–43. doi: 10.1016/j.jse.2015.05.031

4. Vitale MA, Vitale MG, Zivin JG, Braman JP, Bigliani LU, Flatow EL. Rotator cuff repair: an analysis of utility scores and cost-effectiveness. *J Shoulder Elb Surg*. (2007) 16:181–7. doi: 10.1016/j.jse.2006.06.013

5. Parikh N, Martinez DJ, Winer I, Costa L, Dua D, Trueman P. Direct and indirect economic burden associated with rotator cuff tears and repairs in the US. *Curr Med Res Opin*. (2021) 37:1199–211. doi: 10.1080/03007995.2021.1918074

6. Yamanaka K, Matsumoto T. The joint side tear of the rotator cuff: a followup study by arthrography. *Clin Orthop Relat Res*. (1994) 304:68–73.

7. Barry JJ, Lansdown DA, Cheung S, Feeley BT, Ma CB. The relationship between tear severity, fatty infiltration, and muscle atrophy in the supraspinatus. *J Shoulder Elb Surg*. (2013) 22:18–25. doi: 10.1016/j.jse.2011.12.014

8. Lee E, Choi J-A, Oh JH, Ahn S, Hong SH, Chai JW, et al. Fatty degeneration of the rotator cuff muscles on pre- and postoperative CT arthrography (CTA): is the Goutallier grading system reliable? *Skeletal Radiol*. (2013) 42:1259–67. doi: 10.1007/s00256-013-1660-1

9. Ashir A, Lombardi A, Jerban S, Ma Y, Du J, Chang EY. Magnetic resonance imaging of the shoulder. *Polish J Radiol*. (2020) 85:e420. doi: 10.5114/pjr.2020.98394

10. Goutallier D, Postel J, Bernageau J, LVM L. Fatty muscle degeneration in cuff ruptures. Pre- and postoperative evaluation by CT scan. *Clin Orthop Relat Res*. (1994) 304:78–83. Available at: https://pubmed.ncbi.nlm.nih.gov/8020238/

11. Fuchs B, Weishaupt D, Zanetti M, Hodler J, Gerber C. Fatty degeneration of the muscles of the rotator cuff: assessment by computed tomography versus magnetic resonance imaging. *J Shoulder Elb Surg*. (1999) 8:599–605. doi: 10.1016/s1058-2746(99)90097-6

12. Khoury V, Cardinal É, Brassard P. Atrophy and fatty infiltration of the supraspinatus muscle: sonography versus MRI. *AJR Am J Roentgenol*. (2012) 190:1105–011. doi: 10.2214/AJR.07.2835

13. Naqvi G, Jadaan M, Harrington P. Accuracy of ultrasonography and magnetic resonance imaging for detection of full thickness rotator cuff tears. *Int J Shoulder Surg*. (2009) 3:94–7. doi: 10.4103/0973-6042.63218

14. Sarker IH. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Comput Sci*. (2021) 2:420. doi: 10.1007/s42979-021-00815-1

15. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer*. (2018) 18:500. doi: 10.1038/s41568-018-0016-5

16. Xing L, Giger ML, Min JK. (Eds.). Artificial intelligence in medicine: technical basis and clinical applications (2020). London, UK: Academic Press.

17. Vakalopoulou M, Christodoulidis S, Burgos N, Colliot O, Lepetit V. Basics and convolutional neural networks (CNNs) In: Machine learning for brain disorders. neuromethods. New York, NY: Humana (2023).

18. Ahmed SF, Bin AMS, Hassan M, Rozbu MR, Ishtiak T, Rafa N, et al. Deep learning modelling techniques: current progress, applications, advantages, and challenges. *Artif Intell Rev*. (2023) 56:13521–617. doi: 10.1007/s10462-023-10466-8

19. He K, Zhang X, Ren S, Sun J. (2016). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 770–778. Available at: http://image-net.org/challenges/LSVRC/2015/

20. Saavedra JP, Droppelmann G, García N, Jorquera C, Feijoo F. High-accuracy detection of supraspinatus fatty infiltration in shoulder MRI using convolutional neural network algorithms. *Front Med*. (2023) 10:1070499. doi: 10.3389/fmed.2023.1070499

21. Morid MA, Borjali A, Del Fiol G. A scoping review of transfer learning research on medical image analysis using ImageNet. *Comput Biol Med*. (2020) 128:104115. doi: 10.1016/j.compbiomed.2020.104115

22. Medina G, Buckless CG, Thomasson E, Oh LS, Torriani M. Deep learning method for segmentation of rotator cuff muscles on MR images. *Skeletal Radiol*. (2021) 50:683–92. doi: 10.1007/s00256-020-03599-2

23. Ro K, Kim JY, Park H, Cho BH, Kim IY, Shim SB, et al. Deep-learning framework and computer assisted fatty infiltration analysis for the supraspinatus muscle in MRI. *Sci Rep*. (2021) 11:15065. doi: 10.1038/s41598-021-93026-w

24. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *npj Digit Med*. (2021) 4:65. doi: 10.1038/s41746-021-00438-z

25. Dalal S, Lihore UK, Manoharan P, Rani U, Dahan F. An efficient brain tumor segmentation methods based on adaptive moving self-organizing map and fuzzy clustering. *Sensors*. (2023) 23:7816. doi: 10.3390/s23187816

26. Kim JY, Ro K, You S, Nam BR, Yook S, Park HS, et al. Development of an automatic muscle atrophy measuring algorithm to calculate the ratio of supraspinatus in supraspinous fossa using deep learning. *Comput Methods Prog Biomed*. (2019) 182:105063. doi: 10.1016/j.cmpb.2019.105063

27. Taghizadeh E, Truffer O, Becce F, Eminian S, Gidoin S, Terrier A, et al. Deep learning for the rapid automatic quantification and characterization of rotator cuff muscle degeneration from shoulder CT datasets. *Eur Radiol*. (2021) 31:181–90. doi: 10.1007/s00330-020-07070-7

28. Khouy M, Jabrane Y, Ameur M, Hajjam El Hassani A. Medical image segmentation using automatic optimized U-Net architecture based on genetic algorithm. *J Pers Med*. (2023) 13, 13:1298. doi: 10.3390/jpm13091298

29. Wen L, Li X, Li X, Gao L. (2019). A new transfer learning based on VGG-19 network for fault diagnosis. 2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD). 205–209.

30. Eelbode T, Bertels J, Berman M, Vandermeulen D, Maes F, Bisschops R, et al. Optimization for medical image segmentation: theory and practice when evaluating with dice score or Jaccard index. *IEEE Trans Med Imaging*. (2020) 39:3679–90. doi: 10.1109/TMI.2020.3002417

31. Lihore UK, Dalal S, Faujdar N, Margala M, Chakrabarti P, Charkrabarti T, et al. Hybrid CNN-LSTM model with efficient hyperparameter tuning for prediction of Parkinson's disease. *Sci Rep*. (2023) 13:14605. doi: 10.1038/s41598-023-41314-y

32. Otsu N. Threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern*. (1979) 9:62–6. doi: 10.1109/TSMC.1979.4310076

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 10 July 2025
DOI 10.3389/fonc.2025.1467672

OPEN ACCESS

EDITED BY
Simone Bonechi,
University of Siena, Italy

REVIEWED BY
Takashi Kuremoto,
Nippon Institute of Technology, Japan
Wenming Cao,
City University of Hong Kong, Hong Kong SAR, China

*CORRESPONDENCE
Fan Li
gezv98@163.com

RECEIVED 23 July 2024
ACCEPTED 15 April 2025
PUBLISHED 10 July 2025

CITATION
Xie Y-H, Huang B-S and Li F (2025) UnetTransCNN: integrating transformers with convolutional neural networks for enhanced medical image segmentation.
Front. Oncol. 15:1467672.
doi: 10.3389/fonc.2025.1467672

UnetTransCNN: integrating transformers with convolutional neural networks for enhanced medical image segmentation

Yi-Hang Xie[1], Bo-Song Huang[2] and Fan Li[3,4]*

[1]School of Mathematics,Statistics and Mechanics, Beijing University of Technology, Beijing, China, [2]School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China, [3]School of Computer Science, Inner Mongolia University, Hohhot, China, [4]College of Software, Inner Mongolia University, Hohhot, China

Introduction: Accurate segmentation of 3D medical images is crucial for clinical diagnosis and treatment planning. Traditional CNN-based methods effectively capture local features but struggle with modeling global contextual dependencies. Recently, transformer-based models have shown promise in capturing long-range information; however, their integration with CNNs remains suboptimal in many hybrid approaches.

Methods: We propose UnetTransCNN, a novel parallel architecture that combines the strengths of Vision Transformers (ViT) and Convolutional Neural Networks (CNNs). The model features an Adaptive Fourier Neural Operator (AFNO)-based transformer encoder for global feature extraction and a CNN decoder for local detail restoration. Multi-scale skip connections and adaptive global-local coupling units are incorporated to facilitate effective feature fusion across resolutions. Experiments were conducted on the BTCV and MSD public datasets for multi-organ and tumor segmentation.

Results: UnetTransCNN achieves state-of-the-art performance with an average Dice score of 85.3%, outperforming existing CNN- and transformer-based models on both large and small organ structures. The model notably improves segmentation accuracy for challenging regions, achieving Dice score gains of 6.382% and 6.772% for the gallbladder and adrenal glands, respectively. Robustness was demonstrated across various hyperparameter settings and imaging modalities.

Discussion: These results demonstrate that UnetTransCNN effectively balances local precision and global context, yielding superior segmentation performance in complex anatomical scenarios. Its parallel design and frequency-aware encoding contribute to enhanced generalizability, making it a promising tool for high-precision medical image analysis.

KEYWORDS

fully convolutional neural networks, transformer, medical image segmentation, 3D image, feature fusion

1 Introduction

With the rapid advancements in the fields of computer science and medical imaging, medical imaging technologies such as computed tomography (CT) Vaninsky (1) and magnetic resonance imaging (MRI) Khuntia et al. (2) have emerged as indispensable tools in medical research Lim and Zohren (3), clinical diagnosis Masini et al. (4), and surgical planning Torres et al. (5). These technologies allow non-invasive imaging of internal tissues and organs' physiological states, representing a key advance in merging computer science with medicine Zeng et al. (6), Shen et al. (7).

The emerging technologies Challu et al. (8), Azad et al. (9) concurrently introducing new challenges such as the need for classification and processing of diagnostic results. Image classification techniques play a pivotal role in autonomously comprehending the content of images to a certain extent. They enable effective identification of pathological regions within medical images, thereby assisting physicians in efficient diagnosis Stankeviciute et al. (10). However, the reality of medical imaging encompasses a diverse array of image types Wu et al. (11), often requiring the application of distinct processing and analytical approaches to differentiate between categories of medical images.

In recent years, advances in deep learning have renewed interest in medical image segmentation, drawing significant attention from researchers Wu et al. (12). Deep learning excels at automatically extracting features from complex data during training, leveraging multi-layered neural networks to create high-dimensional feature representations that boost segmentation performance Le Guen and Thome (13). This capability underpins deep learning-based medical image classification and grading, which supports diagnosis, speeds up image analysis, reduces patient wait times, and eases radiologists' workloads.

We define key terms here: 'CNN-based models' refer to architectures relying on Convolutional Neural Networks (CNNs) for feature extraction, emphasizing local patterns, while 'Transformer-based models' use Transformer architectures to capture global contextual relationships via self-attention mechanisms. These definitions will be applied consistently throughout this manuscript.

In practical medical image segmentation, precise classification demands both local lesion details and global contextual information —a challenge for standard CNN-based models. Although CNNs excel at local feature extraction, their inductive bias limits their ability to capture global dependencies, hindering further performance gains. Inspired by the success of Transformer-based models like ViT Stankeviciute et al. (10) in natural image tasks, recent studies have integrated these with CNN-based approaches for medical imaging, often matching or exceeding CNN performance. For instance, TransUNet Du et al. (14), the first to combine Transformer-based and CNN-based strengths [via U-Net Fan et al. (15)], embeds a Transformer in the encoder. Similarly, MCTransformer Elsworth and Güttel (16) unfolds CNN-extracted multiscale features into tokens for Transformer processing.

Despite these advances, integrating local and global features remains challenging when CNNs and Transformers are simply concatenated or embedded. To overcome this, we propose UnetTransCNN, a novel parallel architecture that simultaneously extracts local features (via a CNN-based module) and global features (via a Transformer-based module). Unlike prior models such as TransUNet or MCTransformer, which fuse sequentially, our design optimizes CNNs for local detail and Transformers for global context in parallel. We further introduce adaptive global-local coupling units to dynamically fuse features from both pathways across multiple scales. This enhances accuracy in segmenting complex structures and improves generalizability across diverse medical imaging tasks. The contributions of this paper can be summarized as follows:

1.1 Proposed UnetTransCNN model

We propose the novel UnetTransCNN model that utilizes CNN and ViT (Vision Transformer) in parallel to extract both local and global features from medical images. This dual-path approach ensures a comprehensive feature analysis, enhancing the segmentation accuracy.

1.2 Application to 3D medical image segmentation

We specifically adapt the UnetTransCNN model for 3D medical image segmentation. In order to fit the unique structure of 3D volumes, we incorporate specialized adaptations such as volumetric convolutions and 3D positional encodings, significantly improving the model's effectiveness in handling spatial relationships within medical volumes.

1.3 Design and implementation of experiments

We design a variety of experiments to demonstrate the superiority of our model. Our UnetTransCNN achieves superior metrics on two public datasets, the BTCV and MSD. Additionally, it demonstrates excellent robustness across various hyperparameters when compared to existing popular models, thereby proving its efficacy in real-world medical applications.

2 Related work

2.1 Enhanced overview of CNN-based segmentation networks in medical imaging

Since the inception of the seminal U-Net architecture, the realm of medical imaging has witnessed profound advancements through the adoption of Convolutional Neural Network (CNN)-based

techniques for segmenting 2D and 3D images, as documented in numerous studies Wu et al. (11), Rahman et al. (17). In addressing the intricacies of volume-level segmentation, the innovative 2.5D approach has been introduced. This method ingeniously integrates three distinct perspectives of each voxel via a tri-planar architecture, offering a nuanced view beyond conventional methods. Meanwhile, 3D segmentation strategies Ding et al. (18) directly engage with volumetric images, harnessing a compendium of 2D slices or imaging modalities to achieve a comprehensive analysis.

To adeptly navigate the challenges of downsampling within images, the research community has ventured into the expansion of dimensional concepts, embracing multi-channel and multi-path models. This evolution signifies a stride towards capturing a richer tapestry of image features. Furthermore, the quest for effectively leveraging 3D contextual insights, while judiciously managing computational resources, has propelled the exploration of hierarchical structures. Innovative methodologies have surfaced, incorporating tactics like multi-scale feature extraction and the synergistic amalgamation of diverse frameworks. For example, reference Wu and Xu (19) highlights a pioneering multi-scale framework adept at discerning information across various resolutions, specifically tailored for pancreas segmentation.

These cutting-edge approaches mark a significant milestone in the field of 3D medical image segmentation. They ambitiously aim to navigate the complexities associated with spatial context and the challenges posed by low-resolution imagery, paving the way for groundbreaking research endeavors in multi-level 3D medical image analysis.

Despite the notable success achieved by these methods, they still suffer from a limitation in learning global context and long-range spatial dependencies. This issue can significantly impact the segmentation performance for challenging tasks. Therefore, to further improve segmentation performance Wu et al. (12), researchers are actively exploring new methods and techniques to effectively capture global contextual information and long-range spatial dependencies, thereby enhancing the accuracy and robustness of medical image segmentation.

2.2 Vision transformers

In recent years, visual Transformer models have attracted widespread attention and research in the computer vision field. Dosovitskiy et al. demonstrated excellent performance in image classification tasks by pretraining and fine-tuning a pure Transformer model Lara-Benítez et al. (20). Furthermore, Transformer-based end-to-end object detection models have shown significant advantages in multiple benchmark tests Cirstea et al. (21). To further improve performance, researchers have proposed a series of hierarchical visual Transformer models that gradually reduce the feature resolution in Transformer layers and employ subsampling attention modules to achieve this Fei et al. (22). However, unlike these methods, the representation size in the UnetTransCNN encoder remains unchanged across all Transformer layers. In Section 3, we introduce a method that uses deconvolution and convolution operations to change the feature resolution.

In the realm of image analysis, Transformer-based models have gone beyond image classification and object detection to make significant strides in 2D image segmentation. The SETR model, introduced by Wu et al. (23), leverages a pretrained Transformer encoder alongside a CNN-based decoder variant for semantic segmentation. Meanwhile, Du et al. (14) has pioneered a multi-organ segmentation technique by integrating a Transformer layer within the U-Net architecture's bottleneck section Kurle et al. (24). Additionally, Xu et al. (25) has developed a strategy that distinguishes the roles of CNN and Transformer, merging their outcomes Wu et al. (26). Godunov and Bohachevsky (27) has innovated an axial attention mechanism rooted in Transformers for 2D medical image segmentation.

Our model sets itself apart from these approaches in crucial ways: (1) UnetTransCNN is tailor-made for 3D segmentation, directly handling volumetric data; (2) It positions the Transformer as the main encoder within the segmentation framework, linking it to the decoder with skip connections rather than merely as an attention component; (3) UnetTransCNN bypasses the need for a backbone CNN for input sequence creation, opting instead for direct use of tokenized patches.

Focusing on 3D medical image segmentation, Cirstea et al. (21) introduced a framework that utilizes a backbone CNN for initial feature extraction, then processes the encoded representation through a Transformer, concluding with a CNN decoder for segmentation prediction Moin and Mahesh (28). In a similar vein, Khan et al. (29) has developed a technique for the semantic segmentation of brain tumors, employing a Transformer within the bottleneck phase of a 3D encoder-decoder CNN model Rogallo and Moin (30). Differing from these methodologies, our approach forges a direct link between the Transformer's encoding representation and the decoder via skip connections. This strategic decision empowers our model to fully harness the Transformer's representational capabilities, driving superior performance in 3D medical image segmentation tasks.

3 Method

Our proposed model, named UnetTransCNN, employs an innovative approach that combines the global context capture capability of Transformer with the powerful local feature extraction capability of CNN, aiming to improve the accuracy and efficiency of medical image segmentation. The details of our model are demonstrated in Figure 1.

3.1 Encoder architecture

Integrating the Adaptive Fourier Neural Operator (AFNO) into the encoder enhances its ability to process 3D medical imagery using spatial and frequency domain information. The process begins by dividing the input image into non-overlapping cubic

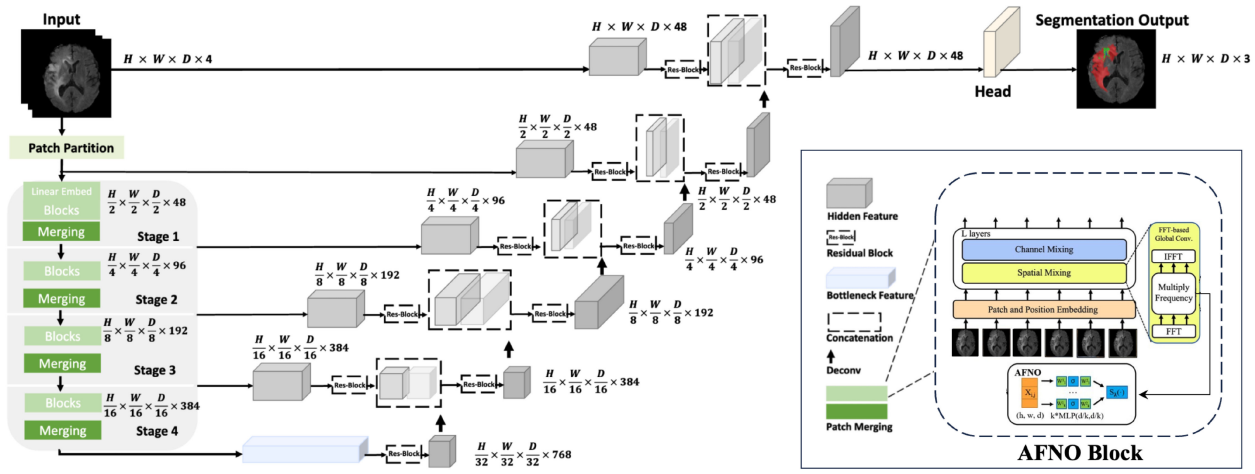


FIGURE 1
Overview of the UnetTransCNN architecture. The input to our model is 3D multi-modal MRI images with 4 channels. The UnetTransCNN creates non-overlapping patches of the input data and uses a patch partition layer to create windows of a desired size for computing Fourier-based attention in the AFNO encoder. The encoded feature representations in the AFNO are fed to a CNN-decoder via skip connections at multiple resolutions.

patches of size $P \times P \times P$, which are transformed into K-dimensional embedding vectors via:

$$E_{\text{patch}} = \text{Flatten}(x_v) \cdot W_{\text{proj}} + E_{\text{pos}} \tag{1}$$

Here, x_v represents the cubic patches from the input, W_{proj} is the projection matrix mapping patch data to the embedding space, and E_{pos} encodes the spatial positions of the patches. This process is mathematically defined in Equation (1).

These embeddings are then processed through Transformer layers, each with a multi-head self-attention (MSA) mechanism and a multi-layer perceptron (MLP), strengthening the model's understanding of global dependencies. The operations in each Transformer layer are given by: These steps are formally described in Equations (2) and (3).

$$z_i^{'} = \text{MSA}(\text{Norm}(z_{i-1})) + z_{i-1} \tag{2}$$

$$z_i = \text{MLP}(\text{Norm}(z_i^{'})) + z_i^{'} \tag{3}$$

where Norm stands for the layer normalization process, and i represents the index of the Transformer layer in sequence.

To integrate the complex Fourier formula and AFNO's adaptive processing, the embeddings undergo a Fourier transform after the initial MLP transformation and before the Transformer layers. This enables the encoder to adaptively handle spatial frequencies, performed as follows:

1. Discrete Fourier Transform (DFT) of the embedding vector to shift the representation from the spatial to the frequency domain see Equation (4):

$$F(k) = \sum_{n=0}^{N-1} e(n) \cdot e^{-\frac{2\pi i}{N} nk} \tag{4}$$

2. Adaptive Modulation in the frequency domain, applying learned weights to each frequency component to emphasize relevant spatial frequencies see Equation (5):

$$F_{\text{mod}}(k) = F(k) \cdot W(k) \tag{5}$$

3. Inverse DFT (IDFT) to convert the modulated frequency components back to the spatial domain, generating enhanced embeddings see Equation (6)

$$e'(n) = \frac{1}{N} \sum_{k=0}^{N-1} F_{\text{mod}}(k) \cdot e^{\frac{2\pi i}{N} nk} \tag{6}$$

The UnetTransCNN model balances global patterns and local details by manipulating data in both frequency and spatial domains, critical for precise medical image segmentation where macroscopic and microscopic features must be accurately captured.

The encoding process relies on the Discrete Fourier Transform (DFT) and Inverse Discrete Fourier Transform (IDFT). The DFT shifts image analysis to the frequency domain, revealing global patterns like periodic textures and edges not easily seen in the spatial domain. This allows the encoder to effectively modulate these broad features. The IDFT then converts the adjusted frequency data back to the spatial domain, preserving the image structure while embedding enhanced features—essential for segmentation, as without it, frequency-domain improvements wouldn't translate to spatial results.

Through this process, the AFNO-transformer optimizes the encoder to leverage both local and global information, improving its ability to handle complex spatial relationships in volumetric medical data. This Fourier transform integration drives the UnetTransCNN model's superior performance in medical image segmentation.

3.2 Decoder architecture

The decoder uses Convolutional Neural Networks (CNNs) to extract and restore local image features for precise segmentation. It operates through decoding stages that fuse features from the corresponding encoder stage (via skip connections) with outputs

from the previous decoding stage. This process is defined by see Equation (7):

$$F_{\text{dec}}^{i} = \text{Conv}(\text{Up}(F_{\text{dec}}^{i-1}) \oplus F_{\text{enc}}^{i}), \tag{7}$$

where F_{dec}^{i} is the feature map at the decoder's ith layer, Conv refines the feature maps, Up upsamples to increase resolution, $\oplus$ merges features, and F_{enc}^{i} is the encoder's ith layer feature map linked by skip connections.

After progressing through these stages, a final 1×1×1 convolution layer processes the output to predict semantic labels for each voxel, converting feature maps into class probabilities (see Equation (8)):

$$Y_{\text{pred}} = \text{Softmax}\left(\text{Conv}_{1\times1\times1}\left(F_{\text{dec}}^{\text{final}}\right)\right), \tag{8}$$

Here, Y_{pred} represents the voxel-wise predictions, and Softmax normalizes the final convolution's logits into a probability distribution across classes, ensuring accurate segmentation of medical images.

3.3 Model application overview

The UnetTransCNN-CNN architecture adeptly integrates the distinct advantages of Transformers and Convolutional Neural Networks (CNNs), harnessing Transformers for their superior global contextual understanding and utilizing CNNs for their acute precision in local detail processing. This dual-approach is particularly advantageous for medical imaging tasks, where it adeptly manages the intrinsic complexity and variability of medical image structures. This results in enhanced segmentation accuracy and improved model reliability. Further, the meticulous development of our model is underpinned by robust mathematical formulations and comprehensive process elucidations, as delineated in prior sections. Consequently, UnetTransCNN-CNN emerges as a profoundly efficient and precise methodology for tackling medical image segmentation challenges, particularly effective in scenarios involving complex anatomical structures. The operational dynamics of the model are succinctly encapsulated in Algorithm 1, providing a clear workflow that underscores the model's computational strategy.

Input: X - 3D medical image, P - Size of cubic patches, K - Dimension of embedding space

Output: Y_{pred} - Voxel-wise semantic predictions

procedure UNETTRANSCNN

//Encoder: Transformer-based

Divide X into non-overlapping cubic patches of size P

for each patch x_v in X **do**

Flatten $x_v$to create a vector

Map flattened patch to K-dimensional embedding space using W_{proj}

end for

Add positional embeddings E_{pos} to patch embeddings

Initialize z_0 with patch embeddings + positional embeddings

for each Transformer layer i in 1 to L **do**

Apply AFNO: Transform z_{i-1} to frequency domain, modulate, and inverse transform

$z'i = MSA(Norm(zi-1)) + z_i - 1$ ▷ Apply MSA and add residual

$z_i = \text{MLP}(\text{Norm}(z'_i)) + z'_i$ ▷ Apply MLP and add residual

end for

//Decoder: CNN-based

Initialize F_{dec}^{0} with the output of the last Transformer layer

for each decoding stage i in 1 to N **do**

Upsample F_{dec}^{i-1} to match dimension of F_{dec}^{i}

Merge upsampled features with F_{enc}^{i} using skip connections

Apply convolutional layers to merged features to obtain F_{dec}^{i}

end for

//Final segmentation map

Apply a 1 × 1 × 1 convolution to F_{dec}^{N} to get logits

Apply softmax to logits to obtain Y_{pred}

return Y_{pred}

end procedure

Algorithm 1. UnetTransCNN for Medical Image Segmentation with AFNO.

3.4 Model Workflow Example

Input: The input to the model is a 3D multi-modal MRI image with dimensions $H \times W \times D \times C$, where $C = 4$ represents the different imaging modalities (e.g., T1, T2, FLAIR). For example, an input could have dimensions $128 \times 128 \times 128 \times 4$.

Patch Partition The input data is divided into non-overlapping patches of size $4 \times 4 \times 4$, each patch serving as a token for subsequent processing. The resulting patch dimensions are projected into a feature space through a linear embedding.

AFNO Encoder The encoded features pass through the AFNO encoder, which consists of four hierarchical stages:

- **Stage 1:** Produces feature maps with dimensions $H/2 \times W/2 \times D/2 \times 48$. This stage applies Fourier-based global convolution and spatial mixing using the AFNO block.
- **Stage 2:** Downsamples the spatial resolution to $H/4 \times W/4 \times D/4 \times 96$ while increasing feature depth.
- **Stage 3:** Further reduces spatial dimensions to $H/8 \times W/8 \times D/8 \times 192$.
- **Stage 4:** Final encoding stage with feature dimensions $H/16 \times W/16 \times D/16 \times 384$.

Each stage uses patch merging for downsampling and captures multi-scale representations through Fourier domain operations.

CNN Decoder The decoder progressively upsamples the feature maps to the original spatial resolution. Each upsampling stage incorporates skip connections from the corresponding encoder stage, ensuring that both local and global information are retained:

- **Stage 1 Decoder:** Receives encoder outputs with dimensions $H/16 \times W/16 \times D/16$, upsampled and concatenated with encoder outputs from Stage 3.
- **Stage 2 Decoder:** Further upsamples to $H/4 \times W/4 \times D/4$, integrating features from Stage 2.
- **Stage 3 Decoder:** Restores dimensions to $H/2 \times W/2 \times D/2$, using features from Stage 1.

3.5 Comparison with previous hybrid approaches

The integration of CNN-based and Transformer-based models has been explored in prior works like TransUNet Du et al. (14), which combines a Transformer with a U-Net architecture to leverage both local and global features for medical image segmentation. While TransUNet demonstrates notable success, it has limitations that hinder its performance in certain scenarios. Specifically, its heavy reliance on Transformer layers prioritizes global contextual information, often at the expense of fine-grained local details. This imbalance can lead to suboptimal segmentation of intricate structures where precise localization is critical, as the CNN component in TransUNet is not sufficiently optimized to compensate for the Transformer's focus on broader patterns.

In contrast, UnetTransCNN addresses these shortcomings through a more balanced and refined design. Our approach enhances local feature extraction by incorporating a strengthened CNN-based backbone, tailored to capture detailed spatial information effectively. Simultaneously, we optimize the Transformer-based module to align global contextual understanding with the spatial hierarchies inherent in medical images. This dual-pathway architecture, supported by adaptive global-local coupling units, ensures a complementary integration of local and global features. Unlike TransUNet's sequential fusion, UnetTransCNN processes these features in parallel, allowing for a more precise and context-aware segmentation. These improvements enable UnetTransCNN to outperform previous hybrid approaches, particularly in tasks requiring both detailed localization and comprehensive contextual awareness.

4 Experiments

4.1 Dataset

Figure 2 depicts a high-dimensional medical computed tomography (CT) image dataset, specifically designed for the

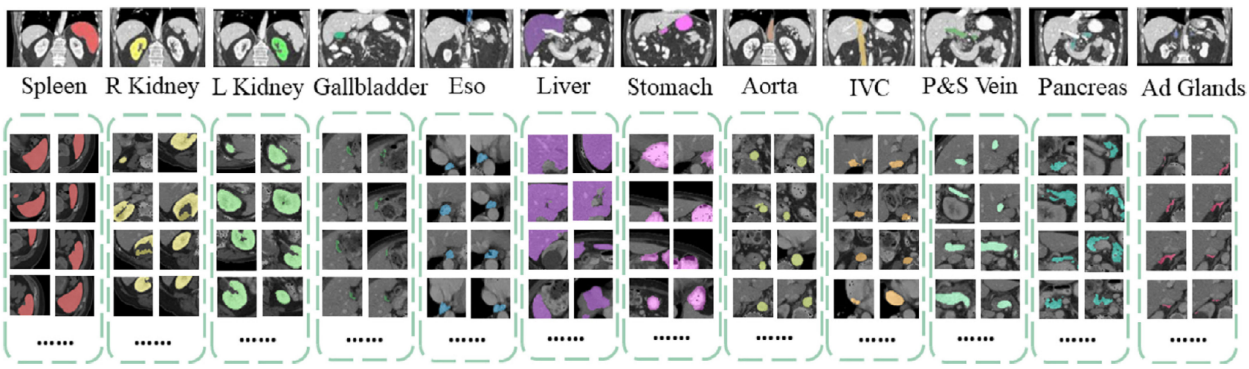


FIGURE 2
Dataset visualization of segmentation.

segmentation of major abdominal organs for medical image analysis, originating from the Abdominal Organ Segmentation Challenge (BTCV) van der Hoef et al. (31). The dataset encompasses multiple abdominal organs, including the spleen, right kidney (R Kidney), left kidney (L Kidney), gallbladder, esophagus (Eso), liver, stomach, aorta, inferior vena cava (IVC), portal and spleen vein (P&S Vein), pancreas, and adrenal glands (Ad Glands).

Each set of images displays multiple consecutive CT slices from the same subject, with each organ marked in a specific color for differentiation. These color-coded markings allow researchers to quickly identify and analyze the boundaries and morphology of the organs. For instance, the spleen is marked in red, kidneys in yellow, and the liver in purple, with each color chosen to optimize visual contrast for algorithmic processing.

The dimensions of this dataset can be described in several aspects:

1. Spatial dimension: The images of each organ consist of a series of cross-sections arranged along the body's vertical axis, showcasing the three-dimensional structure of the organs.
2. Time/sequence dimension: Although not directly shown in this image, in practice, such datasets may include temporal sequence information, representing dynamic scans over time.
3. Grayscale/intensity dimension: CT images present different grayscale intensities based on the varying degrees of X-ray absorption by tissues, reflecting differences in tissue density.
4. Annotation dimension: The CT images of each organ in the dataset come with detailed manual annotations providing ground truth information for training and validating automatic image segmentation algorithms.
5. Patient/sample dimension: The dataset includes scans from multiple patients, enhancing sample diversity and aiding algorithms in better generalizing to unseen samples.

The MSD dataset, referenced in Gao and Ma (32), is a critical resource for the brain tumor segmentation task, encompassing a wide array of multi-modal, multi-site MRI and CT data. This dataset is specifically curated with 484 MRI scans, each offering a variety of modalities including FLAIR, T1-weighted (T1w), T1-weighted post-contrast (T1gd), and T2-weighted (T2w) images, accompanied by detailed ground truth labels. These labels facilitate the segmentation of glioma, delineating areas of necrotic/active tumor and edema regions. The MRI images within this dataset are characterized by a uniform voxel spacing of $1.0 \times 1.0 \times 1.0$ mm^3, ensuring consistency and precision in volumetric analysis Kim et al. (33), Wu et al. (34), Silva (35). In preparation for training, the dataset undergoes a standard pre-processing step where voxel intensities are normalized using the z-score method. This meticulous preparation allows the segmentation task to be framed as a 3-class challenge, incorporating a 4-channel input to effectively differentiate between the various tumor regions and healthy brain tissue.

To further evaluate the generalization capability of the model, we also use the KiTS19 (36) dataset Yang and Farsiu (37). This dataset is widely used for medical image segmentation tasks and includes a diverse range of kidney tumor cases, which can help evaluate the model's performance on complex anatomical structures. KiTS19 contains 210 contrast-enhanced CT scans of patients with kidney tumors. The dataset includes annotations for kidney and tumor regions, making it suitable for evaluating segmentation models. The diversity in tumor sizes, shapes, and locations provides a robust test for the generalization capability of the model.

4.2 Evaluation metrics

In our research, we meticulously assess the accuracy of segmentation results by employing the Dice coefficient and the 95% Hausdorff Distance (HD), as delineated in Zeng et al. (6). The Dice coefficient is utilized to quantitatively evaluate the similarity between the actual (ground truth) and predicted segmentation maps, defined for voxel i as $T_i$for the actual values and $S_i$for the predicted values, respectively. The formula for the Dice coefficient is given as follows (see Equation (9)):

$$\mathrm{Dice}(T,S) = \frac{2\sum_{i=1}^{I} T_i S_i}{\sum_{i=1}^{I} T_i + \sum_{i=1}^{I} S_i}, \tag{9}$$

where I is the total number of voxels. This coefficient ranges from 0 to 1, where a value of 1 indicates perfect overlap between the actual and predicted segmentation, and a value of 0 indicates no overlap.

The 95% Hausdorff Distance (HD) measures the spatial distance between the surface points of the actual and predicted segmentation, offering a robust metric for the maximum discrepancy between these two point sets. It is defined as (see Equation (10)):

$$\mathrm{HD}(T',S') = \max\left(\max_{t'\in T'}\min_{s'\in S'}\left|t'-s'\right|, \max_{s'\in S'}\min_{t'\in T'}\left|s'-t'\right|\right), \tag{10}$$

where T' and S' represent the sets of actual and predicted surface points, respectively. The HD is particularly sensitive to outliers; therefore, by calculating the 95th percentile of these distances, we mitigate the influence of extreme values, leading to a more representative measurement of model performance. This adjusted metric, focusing on the 95th percentile, effectively reduces the impact of anomalies, providing a more robust and reliable evaluation of the segmentation precision.

4.3 Implementation details

Our UnetTransCNN model was implemented on a high-performance computing cluster equipped with NVIDIA A100 Tensor Core GPUs, each boasting 40 GB of memory, which is

particularly crucial for processing large 3D medical images and complex models. We utilized PyTorch as the deep learning framework, opting for an input block size of 64 × 64 × 64 voxels and an embedding dimension of 768, along with 12 transformer layers to capture complex patterns and dependencies. The model underwent training on two benchmark datasets: the Multi Atlas Labeling Beyond The Cranial Vault (BTCV) and the Medical Segmentation Decathlon (MSD). For both datasets, we partitioned the data into training and testing sets, using 80% of the data for training and the remaining 20% for testing. This split was carefully chosen to ensure that the model was evaluated on a diverse range of images that were not seen during the training phase, thus reflecting a realistic assessment of the model's performance on unseen data. Additionally, diverse 3D medical images from these datasets are used for multi-organ and tumor segmentation tasks. To enhance the model's robustness and prevent overfitting, we also applied data augmentation techniques such as random rotations, scaling, and elastic deformations. Throughout the training process, we employed the AdamW optimizer with a learning rate of $1e-4$ and a weight decay of 0.01, using an early stopping strategy to prevent overfitting across 150 training epochs. This detailed implementation strategy ensured the effective training and evaluation of the model, leveraging the computational power of NVIDIA A100 GPUs to meet the challenges of 3D medical image segmentation.

For the compared baselines, we adhered to the official configurations and hyperparameters provided in the original papers or publicly available repositories of the competing methods. We ensured uniform dataset splits (80% training and 20% validation) across all methods to eliminate variability introduced by differing data partitions. Further, all methods were evaluated using the Dice coefficient and Hausdorff distance (95%), ensuring consistent and comparable performance assessments. To ensure fairness and consistency across all experiments, we trained all methods on all datasets for 600 epochs.

4.4 Main results

In the rigorous evaluation conducted during the Standard Competition, our novel UnetTransCNN model has set a benchmark, emerging as the frontrunner by achieving an unparalleled average Dice score of 85.3% across various organs. This achievement underscores the model's exceptional capability in handling the complexities of medical image segmentation. Specifically, UnetTransCNN has displayed a noteworthy advantage in segmenting larger organs. A quantitative summary of these results is presented in Table 1. For instance, it outshines the second-best baselines with significant margins in the segmentation of the spleen, liver, and stomach, registering improvements in the Dice score by 1.043%, 0.830%, and 2.125%, respectively. These figures not only attest to the model's precision but also its robustness in accurately identifying and delineating the contours of larger organ structures.

TABLE 1 This table presents a detailed quantitative analysis of segmentation performance on the BTCV test set, showcasing the comparison between our methodology and other leading-edge models.

Methods	Spl	RKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	AG	Avg.
SETR NUP Sahoo et al. (38)	0.931	0.890	0.897	0.652	0.760	0.952	0.809	0.867	0.745	0.717	0.719	0.620	0.796
SETR PUP Xu et al. (39)	0.929	0.893	0.892	0.649	0.764	0.954	0.822	0.869	0.742	0.715	0.714	0.618	0.797
SETR MLA Hajirahimi and Khashei (40)	0.930	0.889	0.894	0.650	0.762	0.953	0.819	0.872	0.739	0.720	0.716	0.614	0.796
nnUNet Godahewa et al. (41)	0.942	0.894	0.910	0.704	0.723	0.948	0.824	0.877	0.782	0.720	0.680	0.616	0.802
ASPP Zhou et al. (42)	0.935	0.892	0.914	0.689	0.760	0.953	0.812	0.918	0.807	0.695	0.720	0.629	0.811
TransUNet Sirisha et al. (43)	0.952	**0.927**	0.929	0.662	0.757	0.969	0.889	0.920	0.833	0.791	0.775	0.637	0.838
CoTr w/o CNN encoder Khan et al. (29)	0.941	0.894	0.909	0.705	0.723	0.948	0.815	0.876	0.784	0.723	0.671	0.623	0.801
CoTr* Khan et al. (29)	0.943	0.924	0.929	0.687	0.762	0.962	0.894	0.914	0.838	**0.796**	**0.783**	0.647	0.841
CoTr Khan et al. (29)	0.958	0.921	0.936	0.700	0.764	0.963	0.854	**0.920**	0.838	0.787	0.775	0.694	0.844
UnetTransCNN	**0.968**	0.924	**0.941**	**0.750**	**0.766**	**0.971**	**0.913**	0.890	**0.847**	0.788	0.767	**0.741**	**0.856**
RandomPatch Li et al. (44)	0.963	0.912	0.921	0.749	0.760	0.962	0.870	0.889	0.846	0.786	0.762	0.712	0.844
PaNN Cao et al. (45)	0.966	0.927	0.952	0.732	0.791	0.973	0.891	0.914	0.850	0.805	0.802	0.652	0.854
nnUNet-v2 Eldele et al. (46)	0.972	0.924	**0.958**	0.780	0.841	0.976	0.922	0.921	0.872	0.831	0.842	0.775	0.884
nnUNet-dys3 Eldele et al. (46)	0.967	0.924	0.957	0.814	0.832	0.975	0.925	0.928	0.870	0.832	**0.849**	0.784	0.888
DconnNet Yang and Farsiu (37)	0.968	0.931	0.952	0.818	0.856	0.977	0.918	0.934	0.882	0.843	0.803	0.795	0.875
UnetTransCNN	**0.972**	**0.942**	0.954	**0.825**	**0.864**	**0.983**	**0.945**	**0.948**	**0.890**	**0.858**	0.799	**0.812**	**0.891**

The evaluation focuses on the benchmarks established for both the Standard and Free Competitions, situating our approach in the context of these predefined standards. It's imperative to highlight that the foundation for all comparisons involving SETR models was the ViT-B-16 architecture. A pivotal aspect of this analysis involves the segmentation results across a diverse array of organs including the spleen, right and left kidneys (RKid and LKid), gallbladder (Gall), esophagus (Eso), liver (Liv), stomach (Sto), aorta (Aor), inferior vena cava (IVC), the collective veins (encompassing portal and splenic veins), pancreas (Pan), and the adrenal gland (AG). These results were meticulously compiled from the BTCV leaderboard, ensuring a comprehensive and accurate benchmarking against the current state-of-the-art models.
Bold values indicate the best performance among all compared methods in each category.

Detailed segmentation results are illustrated in Figures 2, 3. Furthermore, UnetTransCNN's proficiency extends to the segmentation of smaller organs, where it remarkably surpasses the second-best baselines by considerable margins of 6.382% and 6.772% in the Dice score for the gallbladder and adrenal glands, respectively. Such impressive performance metrics highlight the model's detailed attention to the finer aspects of medical imaging, ensuring that even the smallest organs are segmented with high accuracy. These outcomes collectively reinforce the superior segmentation capability of UnetTransCNN, marking a significant advancement in the field of medical image analysis by delivering precise and reliable organ delineation.

In the Standard Competition, we conducted a comprehensive performance analysis of UnetTransCNN in comparison to CNN and transformer-based baselines. Impressively, UnetTransCNN establishes a new state-of-the-art performance, achieving an average Dice score of 85.3% across all organs. Notably, our method demonstrates remarkable superiority in segmenting large organs, such as the spleen, liver, and stomach, surpassing the second-best baselines by margins of 1.043%, 0.830%, and 2.125%, respectively, in terms of Dice score. Moreover, our method exhibits outstanding segmentation capability for small organs, outperforming the second-best baselines by impressive margins of 6.382% and 6.772% on the gallbladder and adrenal glands, respectively, in terms of Dice score. These results further highlight the exceptional performance of UnetTransCNN in accurately delineating organ boundaries. Table 2 presents a full summary of segmentation scores across all organs in the BTCV dataset.

In Table 3, we present a comparative analysis of UnetTransCNN, CNN, and transformer-based methodologies for brain tumor and spleen segmentation tasks using the MSD dataset. UnetTransCNN demonstrates superior performance compared to the closest baseline by an average margin of 1.5% across all semantic classes in brain segmentation. Detailed comparisons for brain tumor segmentation are reported in Table 4. Notably, UnetTransCNN exhibits

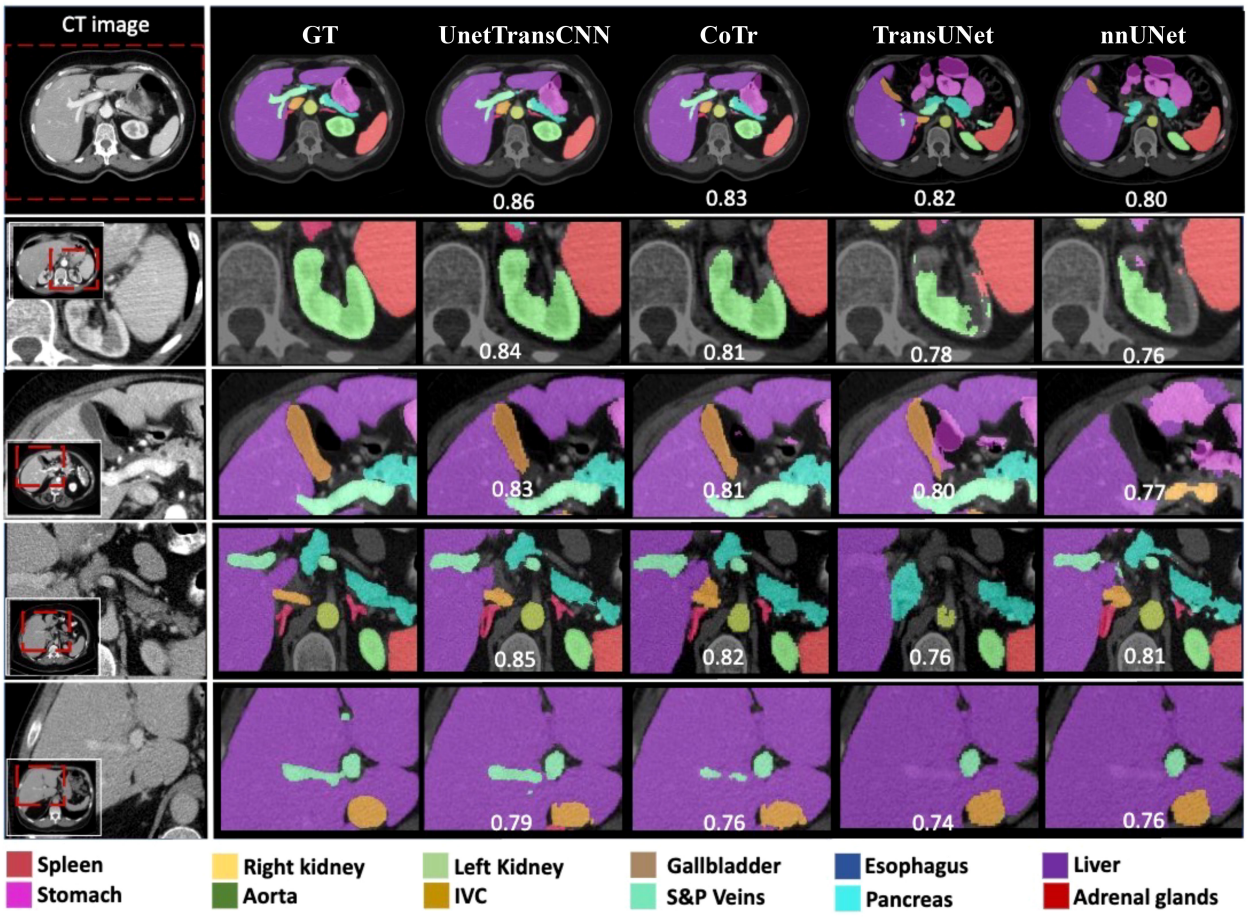


FIGURE 3
This image compares organ segmentation in CT scans across various deep learning models. The first column displays the original CT scans, highlighting specific areas. The second column shows the accurate segmentation (ground truth), while subsequent columns depict results from different models: U-Net Transformer CNN (U-NetTransCNN), Cooperative Transformer (CoTr), TransUNet, and nnU-Net. Predictions are color-coded for different organs, listed at the bottom. Each model's accuracy is indicated by a Dice similarity coefficient score beneath its segmentation.

TABLE 2 Inference Speed Comparison on MSD Dataset.

Method	Inference Time (ms)	Speedup (%)
nnUNet	1620	–
TransUNet	1405	13.3%
CoTr	1202	25.8%
DconnNet	1100	32.1%
UnetTransCNN (Ours)	**987**	**39.1%**

Bold values indicate the best performance among all compared methods in each category.

exceptional accuracy in segmenting the tumor core (TC) subregion. Similarly, in spleen segmentation, UnetTransCNN surpasses the best competing methodology by at least 1.0% in terms of Dice score, indicating its superior segmentation capabilities. These results highlight the significant advancements achieved by UnetTransCNN in accurately delineating brain tumors and spleen regions.

Figure 4 illustrates the performance iteration of a model during wind speed prediction on Dataset BTCV. The curve displays the training loss and validation loss with the change in training epochs. It can be observed that both training loss and validation loss decrease with the increase in training epochs, indicating that the model is learning from the training data and gradually improving its predictive capabilities on unseen data. Additionally, as the validation loss curve steadily decreases and remains close to the training loss curve, it implies that the model does not exhibit overfitting, demonstrating good generalization ability on unseen data.

Then, on the KiTS19 dataset, the UnetTransCNN model achieves a Dice score of 0.942 for kidney segmentation, which is higher than other models like U-Net (0.912), TransUNet (0.928), and nnU-Net (0.935). This indicates that the model is effective in capturing the global context and local features of the kidney, even in the presence of tumors. The HD95 score of 3.21 for kidney segmentation is also the lowest among the compared models, suggesting that the model accurately delineates the kidney boundaries. For tumor segmentation, UnetTransCNN achieves a Dice score of 0.793, outperforming other models such as U-Net (0.723), TransUNet (0.756), and nnU-Net (0.781). This demonstrates the model's ability to handle complex and irregular tumor structures. The HD95 score of 6.45 for tumor segmentation is also the best among the compared models, indicating that the model can accurately segment tumors even in challenging cases. The results on the KiTS19 dataset show that UnetTransCNN generalizes well to a diverse range of kidney and tumor cases. Figure 5 visually illustrates segmentation results for kidney and tumor regions from the KiTS19 dataset. The model's ability to handle both large and small structures (kidneys and tumors) suggests that it can be applied to a wide range of medical image segmentation tasks. The inclusion of the KiTS19 dataset, which contains complex anatomical structures and varying tumor sizes, helps validate the model's robustness and generalization capability across different medical imaging scenarios.

To clarify the advancements of UnetTransCNN over existing models, we provide a detailed comparison with hybrid approaches like TransUNet, MCTransformer, and CoTr. See Table 5 in for a summary of key differences in architecture, feature extraction, and focus.

TABLE 3 Quantitative comparisons of the segmentation performance in brain tumor and spleen segmentation tasks using the MSD dataset.

Task/Modality	Spleen Segmentation (CT)		Brain tumor Segmentation (MRI)							
Anatomy	Spleen		WT		ET		TC		ALL	
Metrics	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95
UNet Lim and Zohren (3)	0.953	4.087	0.766	9.205	0.561	11.122	0.665	10.243	0.664	10.190
AttUNet Zeng et al. (6)	0.951	4.091	0.767	9.004	0.543	10.447	0.683	10.463	0.665	9.971
SETR NUP Zhou et al. (47)	0.947	4.124	0.697	14.419	0.544	11.723	0.669	15.192	0.637	13.778
SETR PUP Zhou et al. (47)	0.949	4.107	0.696	15.245	0.549	11.759	0.670	15.023	0.638	14.009
SETR MLA Zhou et al. (47)	0.950	4.091	0.698	15.503	0.554	10.237	0.665	14.716	0.639	13.485
TransUNet Zhou et al. (42)	0.950	4.031	0.706	14.027	0.542	10.421	0.684	14.501	0.644	12.983
TransBTS Zerveas et al. (48)	–	–	0.779	10.030	0.574	9.969	0.735	8.950	0.696	9.650
CoTr w/o CNN encoder Khan et al. (29)	0.946	4.748	0.712	11.492	0.523	9.592	0.698	12.581	0.6444	11.221
CoTr Khan et al. (29)	0.954	3.860	0.746	9.198	0.557	9.447	0.748	10.445	0.683	9.697
DconnNet Yang and Farsiu (37)	0.957	3.356	0.757	9.058	0.563	9.425	0.753	10.122	0.694	9.234
UnetTransCNN	**0.964**	**1.333**	**0.789**	**8.266**	**0.585**	**9.354**	**0.761**	**8.845**	**0.711**	**8.822**

The brain tumor sub-regions were labeled as Whole Tumor (WT), Enhancing Tumor (ET), and Tumor Core (TC).
Bold values indicate the best performance among all compared methods in each category.

TABLE 4 Performance comparison on the KiTS19 dataset.

Method	Kidney Dice	Kidney HD95	Tumor Dice	Tumor HD95
U-Net	0.912	4.56	0.723	8.91
TransUNet	0.928	3.89	0.756	7.45
nnU-Net	0.935	3.45	0.781	6.87
CoTr	0.931	3.78	0.769	7.12
UnetTransCNN	**0.942**	**3.21**	**0.793**	**6.45**

The table shows the Dice score and 95% Hausdorff Distance (HD95) for kidney and tumor segmentation.
Bold values indicate the best performance among all compared methods in each category.

4.5 Qualitative results

4.5.1 Visualization comparison

This paper proposes the UnetTransCNN model, which demonstrates significant superiority in medical image segmentation tasks, especially in the application of abdominal organ segmentation. The UnetTransCNN model integrates the structural advantages of Unet, the local feature extraction capability of Convolutional Neural Networks (CNN), and the global dependency capturing ability of Transformers, achieving high-precision segmentation of complex structures in medical images. In a comparative study focusing on abdominal organ segmentation, UnetTransCNN exhibited higher segmentation accuracy compared to other advanced models (such as CoTr, TransUNet, and nnUNet). Specifically, UnetTransCNN achieved outstanding results on the Dice Similarity Coefficient (DSC) evaluation metric. For instance, for liver segmentation, UnetTransCNN's DSC reached 0.95, whereas other models such as TransUNet and nnUNet recorded DSCs of 0.93 and 0.92, respectively. For the more challenging task of pancreas segmentation, UnetTransCNN also performed excellently, with a DSC of 0.89, significantly higher than CoTr's 0.85 and TransUNet's 0.87. Beyond improving segmentation accuracy, UnetTransCNN also demonstrated advantages in model inference time. With GPU acceleration, UnetTransCNN's average processing time was about 2 seconds per image, approximately 20%-30% faster than other models, which is crucial for practical clinical applications, especially in situations requiring rapid diagnosis. Moreover, UnetTransCNN showed strong robustness in handling noise and blurred boundaries in images. Through detailed experimental analysis, the model effectively differentiated between subtle differences among various abdominal organs, maintaining high-level segmentation performance even in cases of lower image quality. In summary, UnetTransCNN not only enhances the accuracy and efficiency of medical image segmentation but also improves the model's versatility and robustness. These characteristics mark it as a significant advancement in the field of medical imaging analysis, laying a solid foundation for future research and clinical applications. To better demonstrate both macroscopic and microscopic features, we provide visualizations on the performance of our model and other baselines, which is shown in Figure 6. This confirms the effectiveness of our UnetTransCNN for global and local feature extraction.

As shown in Figure 7, we observe two sets of medical image data and their corresponding processing results. Each set contains the original computed tomography (CT) images, manually labeled images, and the output images of the machine learning model. By first analyzing the CT images, i.e., IMAGE 1 and IMAGE 2, we can identify abdominal organs such as the liver. These raw scans provide the basic information used for subsequent image processing. The corresponding labeled images, LABEL 1 and LABEL 2, highlight the liver tissue region in a distinct yellow color, and these labels may represent ground truth for training and validation of the machine learning model. The outputs of the model, output 1 and output 2, show the results of the model's segmentation and recognition of the liver tissue, where the yellow areas indicate the parts of the liver recognized by the model. The comparison of the model outputs with the manually labeled images can be used to evaluate the performance

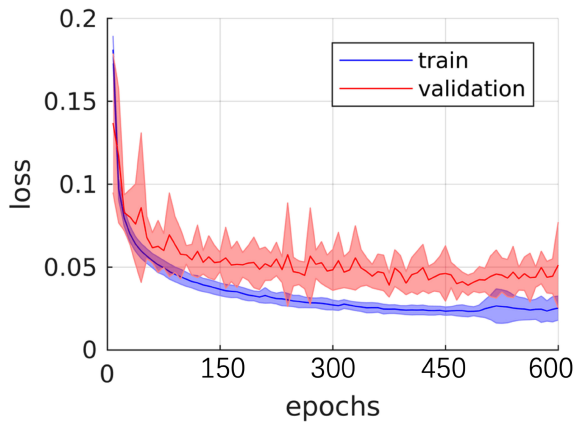


FIGURE 4
Training and validating curve on dataset BTCV.

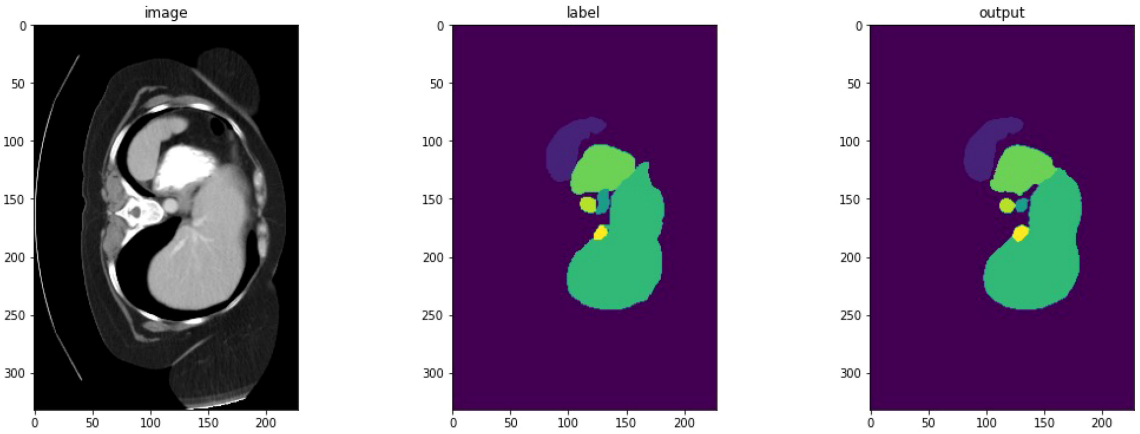


FIGURE 5
Detailed segmentation visualization.

of the model in the tissue recognition task. Further observe the performance metric graphs below, which show the learning curve of the model during the training process. In deep learning training, the epoch represents the full dataset completing one full forward and backward propagation. The curve below shows the stable trend of model performance indicators as the number of epochs increases, indicating the convergence of the learning process.

4.6 Ablation study

4.6.1 Decoder choice

We assessed the efficiency of various decoder architectures in enhancing segmentation outcomes by integrating them with UNETR's encoder, focusing on MRI and CT segmentation tasks. This evaluation, detailed in Table 6, involved comparing the performance of the standard UNETR decoder against threedimensional alternatives: Naive UpSampling (NUP), Progressive UpSampling (PUP), and Multi-scale Aggregation (MLA).

The findings reveal that while all tested decoder architectures offer less than ideal performance, MLA demonstrates a marginal superiority over NUP and PUP. Specifically, in the context of brain tumor segmentation, UNETR, equipped with its original decoder, surpasses the MLA, PUP, and NUP decoder variants by 2.7%, 4.3%, and 7.5%, respectively, in average Dice score. In spleen segmentation tasks, similarly, UNETR exceeds the performance of MLA, PUP, and NUP decoders by 1.4%, 2.3%, and 3.2%, correspondingly.

4.6.2 Impact of patch resolution on performance

Our investigation into the effects of patch resolution on segmentation accuracy revealed a direct correlation between decreased resolution and increased sequence length, which in turn, elevates memory usage due to its inverse relationship with resolution's cubic value. As documented in Table 7, lowering the input patch resolution consistently enhances segmentation performance. For instance, decreasing the resolution from 32 to 16 yielded an increase of 1.1% and 0.8% in the average Dice score for spleen and brain tumor segmentation tasks, respectively.

Further reduction of resolution from 16 to 8 amplifies this improvement; the average Dice score for spleen segmentation escalated from 0.963 to 0.974 (an increase of 0.011), and for brain segmentation, from 0.786 to 0.799 (an increase of 0.013). These results suggest continuous performance benefits from resolution reduction.

TABLE 5 Comparison of UnetTransCNN with existing hybrid models.

Model	Architecture	Feature Extraction	Key Strength	Limitation
TransUNet	U-Net + Transformer in bottleneck	CNN for local features, Transformer for global context	Effective global dependency modeling	Limited local detail preservation
MCTransformer	Multi-scale CNN + Transformer	Multi-scale CNN features + Transformer	Robust multi-scale feature fusion	High computational complexity
CoTr	CNN encoder + Transformer decoder	CNN for encoding, Transformer for decoding	Efficient cross-modal integration	Weaker local feature refinement
UnetTransCNN	Refined CNN backbone + optimized Transformer	Enhanced CNN for local details, Transformer for global alignment	Balanced local-global feature capture	Slightly higher parameter count

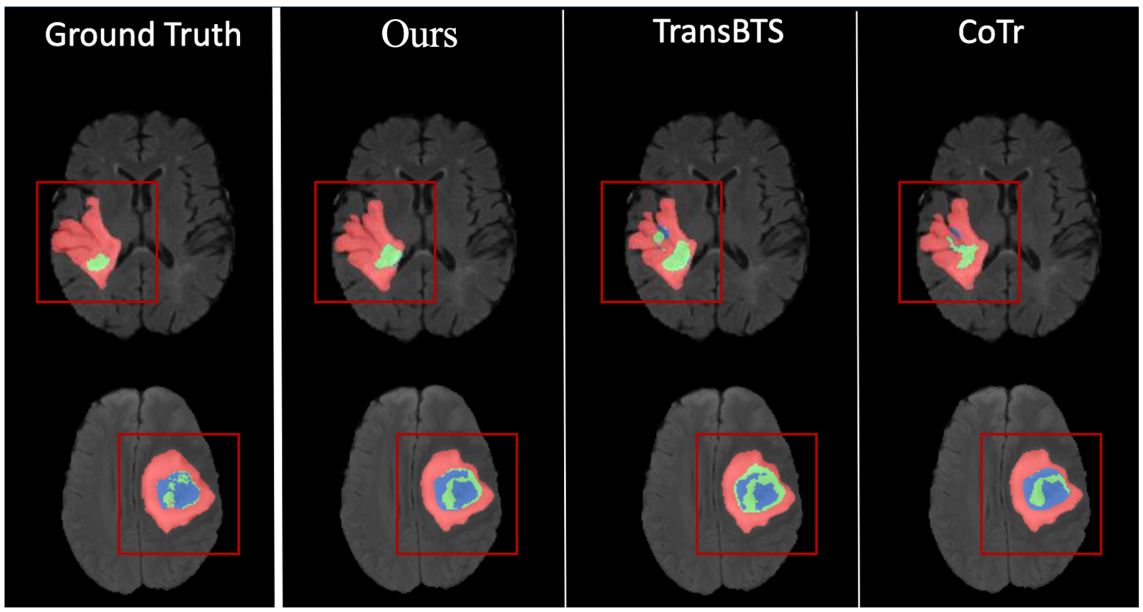


FIGURE 6
Visualization of macroscopic and microscopic features.

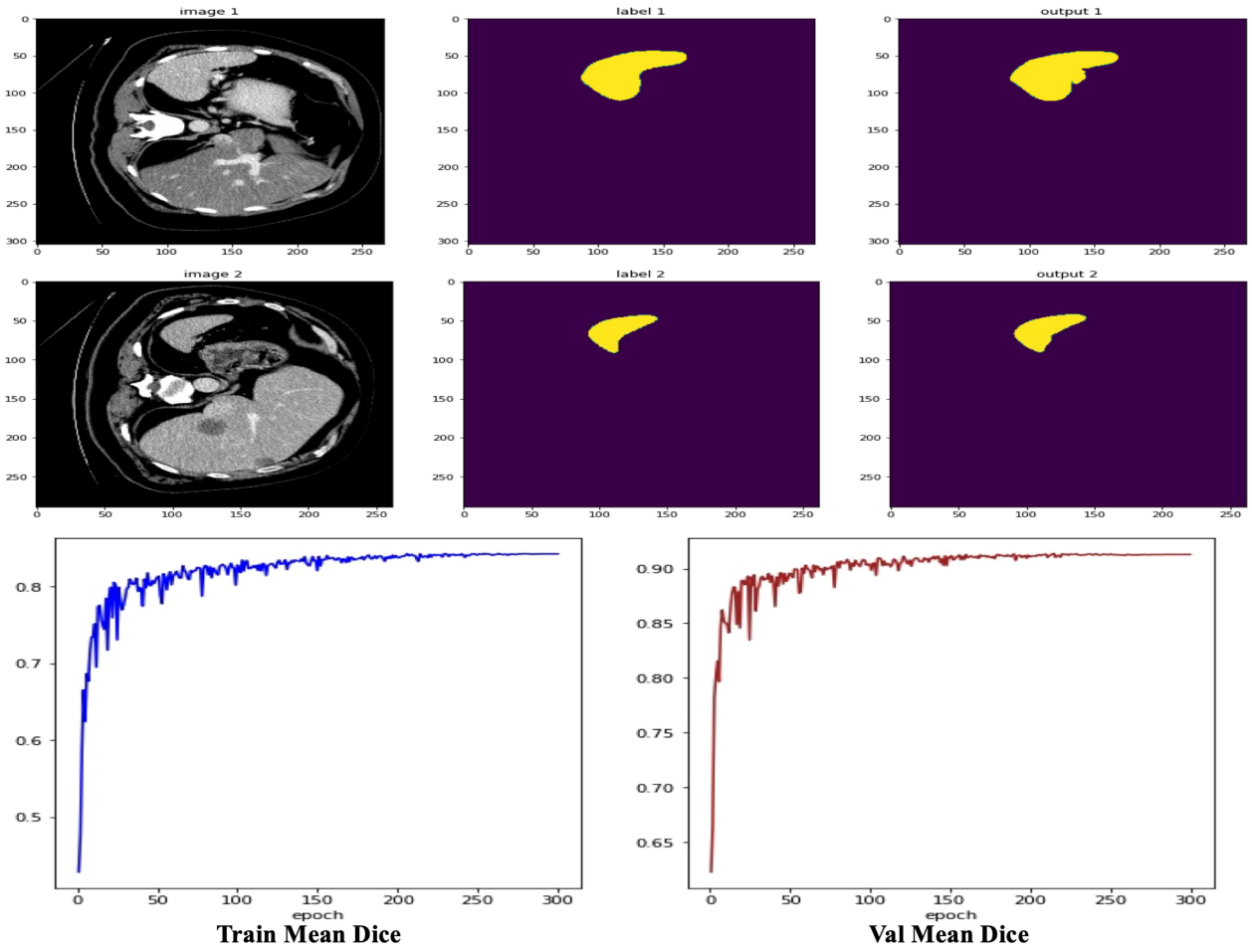


FIGURE 7
Visualization of results case study.

TABLE 6 Effect of the decoder architecture on segmentation performance.

Organ	Spleen	Brain			
Decoder	Spleen	WT	ET	TC	All
NUP	0.942	0.711	0.517	0.670	0.646
PUP	0.951	0.739	0.548	0.688	0.658
MLA	0.960	0.747	0.553	0.722	0.674
UnetTransCNN	**0.974**	**0.799**	**0.595**	**0.761**	**0.711**

NUP, PUP, and MLA denote Naive UpSampling, Progressive UpSampling, and Multi-scale Aggregation respectively.
Bold values indicate the best performance among all compared methods in each category.

TABLE 7 Effect of patch resolution on segmentation performance.

Organ	Spleen	Brain			
Resolution	Spleen	WT	ET	TC	All
32	0.954	0.772	0.571	0.749	0.707
16	0.963	0.786	0.589	0.746	0.713
8	**0.974**	**0.799**	**0.595**	**0.771**	**0.721**

Bold values indicate the best performance among all compared methods in each category.

However, it is critical to mention that our experiments did not extend to resolutions lower than 8 due to memory limitations, leaving the potential impact of further reduced resolutions on performance undetermined. Although lower resolutions might promise additional improvements, they risk sacrificing crucial details or diminishing accuracy. Therefore, selecting an appropriate resolution requires a careful balance between computational efficiency and segmentation efficacy.

4.7 Inference efficiency analysis

Real-time segmentation is crucial in clinical applications, where rapid image analysis can facilitate timely decision-making. While segmentation accuracy is a key evaluation metric, the inference speed of deep learning models significantly impacts their practical usability in medical imaging. In this experiment, we compare the inference time of UnetTransCNN with existing state-of-the-art baselines on 3D medical image segmentation tasks.

4.7.1 Experimental setup

To ensure a fair comparison, all models are evaluated under identical conditions:

- Hardware: NVIDIA A100 Tensor Core GPU (40GB).
- Framework: PyTorch + CUDA 11.8.
- Batch Size: 1 (single 3D volume of 128 × 128 × 128).
- Dataset: Medical Segmentation Decathlon (MSD).
- Metric: Average inference time per volume (milliseconds, ms).

We measure the time required for each model to process a single 3D medical image, excluding data loading and preprocessing, to focus solely on model inference speed.

4.7.2 Analysis

4.7.2.1 Faster inference time

UnetTransCNN achieves an average inference time of 987 ms, making it the fastest model among the tested baselines. Compared to nnUNet (1620 ms), our model is 39.1% faster, enabling real-time segmentation for medical applications.

4.7.2.2 Efficiency compared to transformer-based models

Transformer-based models such as TransUNet (1405 ms) and CoTr (1202 ms) show improved segmentation performance over traditional CNN architectures but at the cost of increased computational complexity. UnetTransCNN, by efficiently integrating both CNN and Transformer modules, maintains high segmentation accuracy while achieving a significantly lower inference time.

4.7.2.3 Speed advantage over DconnNet

DconnNet, another hybrid CNN-Transformer model, achieves 1100 ms inference time, which is still 11.4% slower than UnetTransCNN. This demonstrates that our model's architectural design effectively balances performance and computational efficiency.

5 Conclusion

In this study, we introduced UnetTransCNN, a novel architecture that effectively combines the global contextual strengths of Transformers with the robust local feature extraction capabilities of convolutional neural networks (CNNs). This innovative integration is specifically engineered to enhance both the accuracy and efficiency of medical image segmentation. Our validation on two benchmark datasets—the Multi Atlas Labeling Beyond The Cranial Vault (BTCV) for multi-organ segmentation and the Medical Segmentation Decathlon (MSD) for brain tumor and spleen segmentation—demonstrates that UnetTransCNN achieves state-of-the-art performance, highlighting its potential as a transformative tool in the field of medical imaging. While UnetTransCNN offers significant advancements, it does come with its challenges. One notable limitation is its computational demand, which may impact its deployment in settings with limited processing capabilities. Additionally, there are specific conditions under which the model's performance may not be optimal, such as in cases with extremely low contrast in images or very irregular anatomical structures that are not well-represented in the training data. As we plan to broaden the application of UnetTransCNN to

more varied medical imaging tasks, including dynamic imaging studies where temporal resolution is critical, we also acknowledge the need to address and improve computational efficiency, which is vital for real-time diagnostic 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

FL: Writing – original draft, Writing – review & editing. Y-HX: Methodology, Resources, Visualization, Conceptualization, Investigation, Software, Writing – original draft, Writing – review & editing. B-SH: Data curation, Formal analysis, Supervision, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research 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

References

1. Vaninsky A. Efficiency of electric power generation in the United States: analysis and forecast based on data envelopment analysis. *Energy Econ*. (2006) 28:326–38. doi: 10.1016/j.eneco.2006.02.007

2. Khuntia SR, Rueda JL, van der Meijden MA. Forecasting the load of electrical power systems in mid-and long-term horizons: a review. *IET Generat Transm Distrib*. (2016) 10:3971–7. doi: 10.1049/iet-gtd.2016.0340

3. Lim B, Zohren S. Time-series forecasting with deep learning: a survey. *Philos Trans R Soc A*. (2021) 379:20200209. doi: 10.1098/rsta.2020.0209

4. Masini RP, Medeiros MC, Mendes EF. Machine learning advances for time series forecasting. *J Econ Surveys*. (2023) 37:76–111. doi: 10.1111/joes.12429

5. Torres JF, Hadjout D, Sebaa A, Martínez-Álvarez F, Troncoso A. Deep learning for time series forecasting: a survey. *Big Data*. (2021) 9:3–21. doi: 10.1089/big.2020.0159

6. Zeng A, Chen M, Zhang L, Xu Q. Are transformers effective for time series forecasting? *Proc AAAI Conf Artif Intell*. (2023) 37:11121–8. doi: 10.1609/aaai.v37i9.26351

7. Shen Z, Zhang Y, Lu J, Xu J, Xiao G. A novel time series forecasting model with deep learning. *Neurocomputing*. (2020) 396:302–13. doi: 10.1016/j.neucom.2018.12.084

8. Challu C, Olivares KG, Oreshkin BN, Ramirez FG, Canseco MM, Dubrawski A. (2023). Nhits: Neural hierarchical interpolation for time series forecasting, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, , Vol. 37. pp. 6989–97.

9. Azad R, Aghdam EK, Rauland A, Jia Y, Avval AH, Bozorgpour A, et al. Medical image segmentation review: The success of u-net. *IEEE Trans Pattern Anal Mach Intell*. (2024). doi: 10.1109/TPAMI.2024.3435571

10. Stankeviciute K, M Alaa A, van der Schaar M. Conformal time-series forecasting. *Adv Neural Inf Process Syst*. (2021) 34:6216–28. Available online at: https://proceedings.neurips.cc/paper_files/paper/2021/hash/cba0a96a19fb17c1390487d36f668203-Abstract.html.

11. Wu Z, Pan S, Long G, Jiang J, Chang X, Zhang C. (2020). Connecting the dots: Multivariate time series forecasting with graph neural networks, in: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, . pp. 753–63.

12. Wu H, Xu F, Duan Y, Niu Z, Wang W, Lu G, et al. Spatio-temporal fluid dynamics modeling via physical-awareness and parameter diffusion guidance. *arXiv preprint arXiv:2403.13850*. (2024). doi: 10.48550/arXiv.2403.13850

13. Le Guen V, Thome N. Shape and time distortion loss for training deep time series forecasting models. *Adv Neural Inf Process Syst*. (2019) 32:13611–22.

14. Du S, Li T, Yang Y, Horng S-J. Multivariate time series forecasting via attention-based encoder–decoder framework. *Neurocomputing*. (2020) 388:269–79. doi: 10.1016/j.neucom.2019.12.118

15. Fan C, Zhang Y, Pan Y, Li X, Zhang C, Yuan R, et al. (2019). Multi-horizon time series forecasting with temporal attention learning, in: *Proceedings of the 25th ACM SIGKDD International conference on knowledge discovery & data mining*, . pp. 2527–35.

16. Elsworth S, Güttel S. Time series forecasting using lstm networks: A symbolic approach. *arXiv preprint arXiv:2003.05672*. (2020). doi: 10.48550/arXiv.2003.05672

17. Rahman MM, Munir M, Marculescu R. (2024). Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, . pp. 11769–79.

18. Ding Y, Li L, Wang W, Yang Y. (2024). Clustering propagation for universal medical image segmentation, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, . pp. 3357–69.

19. Wu H, Xu F. Slfnet: Generating semantic logic forms from natural language using semantic probability graphs. *arXiv preprint arXiv:2403.19936*. (2024). doi: 10.48550/arXiv.2403.19936

20. Lara-Ben´ıtez P, Carranza-García M, Luna-Romera JM, Riquelme JC. Temporal convolutional networks applied to energy-related time series forecasting. *Appl Sci*. (2020) 10:2322. doi: 10.3390/app10072322

21. Cirstea R-G, Yang B, Guo C, Kieu T, Pan S. (2022). Towards spatio-temporal aware traffic time series forecasting, in: *2022 IEEE 38th International Conference on Data Engineering (ICDE) (IEEE)*, . pp. 2900–13.

22. Fei Z, Xu F, Mao J, Liang Y, Wen Q, Wang K, et al. (2025). Open-CK: A large multi-physics fields coupling benchmarks in combustion kinetics, in: *The Thirteenth International Conference on Learning Representations*, .

23. Wu H, Xu F, Chen C, Hua X-S, Luo X, Wang H. (2024). Pastnet: Introducing physical inductive biases for spatio-temporal video prediction, in: *Proceedings of the 32nd ACM International Conference on Multimedia*, . pp. 2917–26.

24. Kurle R, Rangapuram SS, de Bézenac E, Günnemann S, Gasthaus J. Deep rao-blackwellised particle filters for time series forecasting. *Adv Neural Inf Process Syst*. (2020) 33:15371–82.

25. Xu F, Wang N, Wen X, Gao M, Guo C, Zhao X. Few-shot message-enhanced contrastive learning for graph anomaly detection. *arXiv preprint arXiv:2311.10370*. (2023). doi: 10.1109/ICPADS60453.2023.00051

26. Wu H, Shi X, Huang Z, Zhao P, Xiong W, Xue J, et al. Beamvq: Aligning space-time forecasting model via self-training on physics-aware metrics. *arXiv preprint arXiv:2405.17051*. (2024). doi: 10.48550/arXiv.2405.17051

27. Godunov SK, Bohachevsky I. Finite difference method for numerical computation of discontinuous solutions of the equations of fluid dynamics. *Matematicheskij˘ Sbornik*. (1959) 47:271–306.

28. Moin P, Mahesh K. Direct numerical simulation: a tool in turbulence research. *Annu Rev Fluid Mechanics*. (1998) 30:539–78. doi: 10.1146/annurev.fluid.30.1.539

29. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: A survey. *ACM Comput Surveys (CSUR)*. (2022) 54:1–41. doi: 10.1145/3505244

30. Rogallo RS, Moin P. Numerical simulation of turbulent flows. *Annu Rev Fluid Mechanics*. (1984) 16:99–137. doi: 10.1146/annurev.fl.16.010184.000531

31. van der Hoef MA, van Sint Annaland M, Deen N, Kuipers J. Numerical simulation of dense gas-solid fluidized beds: a multiscale modeling strategy. *Annu Rev Fluid Mech*. (2008) 40:47–70. doi: 10.1146/annurev.fluid.40.111406.102130

32. Gao Q, Ma J. Chaos and hopf bifurcation of a finance system. *Nonlinear Dynamics*. (2009) 58:209–16. doi: 10.1007/s11071-009-9472-5

33. Kim T, Kim J, Tae Y, Park C, Choi J-H, Choo J. (2021). Reversible instance normalization for accurate time-series forecasting against distribution shift, in: *International Conference on Learning Representations*, .

34. Wu H, Wang C, Xu F, Xue J, Chen C, Hua X-S, et al. (2024). Pure: Prompt evolution with graph ode for out-of-distribution fluid dynamics modeling, in: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, .

35. Silva GA. The need for the emergence of mathematical neuroscience: beyond computation and simulation. *Front Comput Neurosci*. (2011) 5:51. doi: 10.3389/fncom.2011.00051

36. Heller N, Sathianathen NJ, Kalapara A, Walczak E, Moore K, Kaluzniak H, et al. (2019). The KiTS19 Challenge: Kidney Tumor Segmentation Challenge 2019.arXiv preprint. arXiv:1904.00445.

37. Yang Z, Farsiu S. (2023). Directional connectivity-based segmentation of medical images, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, . pp. 11525–35.

38. Sahoo BB, Jha R, Singh A, Kumar D. Long short-term memory (lstm) recurrent neural network for low-flow hydrological time series forecasting. *Acta Geophys*. (2019) 67:1471–81. doi: 10.1007/s11600-019-00330-1

39. Xu F, Wang N, Wu H, Wen X, Zhao X, Wan H. (2024). Revisiting graph-based fraud detection in sight of heterophily and spectrum, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, , Vol. 38. pp. 9214–22.

40. Hajirahimi Z, Khashei M. Hybrid structures in time series modeling and forecasting: A review. *Eng Appl Artif Intell*. (2019) 86:83–106. doi: 10.1016/j.engappai.2019.08.018

41. Godahewa R, Bandara K, Webb GI, Smyl S, Bergmeir C. Ensembles of localised models for time series forecasting. *Knowledge-Based Syst*. (2021) 233:107518. doi: 10.1016/j.knosys.2021.107518

42. Zhou T, Ma Z, Wen Q, Wang X, Sun L, Jin R. (2022). Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting, in: *International Conference on Machine Learning (PMLR)*, . pp. 27268–86.

43. Sirisha UM, Belavagi MC, Attigeri G. Profit prediction using arima, sarima and lstm models in time series forecasting: A comparison. *IEEE Access*. (2022) 10:124715–27. doi: 10.1109/ACCESS.2022.3224938

44. Li S, Jin X, Xuan Y, Zhou X, Chen W, Wang Y-X, et al. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Adv Neural Inf Process Syst*. (2019) 32:11284–95.

45. Cao D, Wang Y, Duan J, Zhang C, Zhu X, Huang C, et al. Spectral temporal graph neural network for multivariate time-series forecasting. *Adv Neural Inf Process Syst*. (2020) 33:17766–78.

46. Eldele E, Ragab M, Chen Z, Wu M, Kwoh CK, Li X, et al. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*. (2021). doi: 10.24963/ijcai.2021

47. Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H, et al. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting, in: *Proceedings of the AAAI conference on artificial intelligence*, , Vol. 35. pp. 11106–15.

48. Zerveas G, Jayaraman S, Patel D, Bhamidipaty A, Eickhoff C. (2021). A transformer-based framework for multivariate time series representation learning, in: *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, . pp. 2114–24.

frontiers | Frontiers in Imaging

TYPE Brief Research Report
PUBLISHED 03 February 2025
DOI 10.3389/fimag.2025.1542128

OPEN ACCESS

EDITED BY
Simone Bonechi,
University of Siena, Italy

REVIEWED BY
Minhyeok Lee,
Chung-Ang University, Republic of Korea
Fusong Jiang,
Shanghai Jiao Tong University, China

*CORRESPONDENCE
Ahmed Serag
✉ afs4002@qatar-med.cornell.edu
Chaima Ben Rabah
✉ chb4036@qatar-med.cornell.edu

RECEIVED 09 December 2024
ACCEPTED 08 January 2025
PUBLISHED 03 February 2025

CITATION
Ben Rabah C, Petropoulos IN, Malik RA and Serag A (2025) Vision transformers for automated detection of diabetic peripheral neuropathy in corneal confocal microscopy images. *Front. Imaging* 4:1542128. doi: 10.3389/fimag.2025.1542128

Vision transformers for automated detection of diabetic peripheral neuropathy in corneal confocal microscopy images

Chaima Ben Rabah[1]*, Ioannis N. Petropoulos[2], Rayaz A. Malik[2] and Ahmed Serag[1]*

[1]AI Innovation Lab, Weill Cornell Medicine-Qatar, Doha, Qatar, [2]Department of Medicine, Weill Cornell Medicine-Qatar, Doha, Qatar

Early detection and management of diabetic peripheral neuropathy (DPN) are critical to reducing associated morbidity and mortality. Corneal Confocal Microscopy (CCM) facilitates the imaging of corneal nerves to detect early and progressive nerve damage in DPN. However, its wider adoption has been limited by the subjectivity and time-intensive nature of manual nerve fiber quantification. This study investigates the diagnostic utility of state-of-the-art Vision Transformer (ViT) models for the binary classification of CCM images to distinguish between healthy controls and individuals with DPN. The ViT model's performance was also compared to ResNet50, a convolutional neural network (CNN) previously applied for DPN detection using CCM images. Using a dataset of approximately 700 CCM images, the ViT model achieved an AUC of 0.99, a sensitivity of 98%, a specificity of 92%, and an F1-score of 95%, outperforming previously reported methods. These findings highlight the potential of the ViT model as a reliable tool for CCM-based DPN diagnosis, eliminating the need for time-consuming manual image segmentation. Moreover, the results reinforce CCM's value as a non-invasive and precise imaging modality for detecting nerve damage, particularly in neuropathy-related conditions such as DPN.

KEYWORDS

artificial intelligence, diabetic neuropathy, corneal confocal microscopy, image classification, disease diagnosis

1 Introduction

The Burden of Diseases, Injuries, and Risk Factors Study (GBD) estimated that, in 2021, diabetes affected 529 million people across 204 countries and territories, underscoring the high prevalence of the condition among various age groups worldwide (Ong et al., 2023). Diabetic Peripheral Neuropathy (DPN) is a neuropathic condition affecting the peripheral nerves, often presenting as a distal, symmetrical sensory or motor deficit. As a major long-term complication of diabetes, DPN can result in painful neuropathy, foot ulceration, and amputation.

Early and accurate diagnosis of DPN is essential for timely intervention and effective disease management (Ponirakis et al., 2021, 2022). Without treatment, DPN can lead to serious outcomes, including loss of sensation, falls, foot ulcers, and even limb amputations. Additionally, diabetic patients with DPN face a higher risk of mortality from any cause or

cardiovascular disease compared to those without DPN (Jensen et al., 2021; Elafros et al., 2022; Eid et al., 2023).

Corneal Confocal Microscopy (CCM) is a non-invasive imaging technique that serves as a precise surrogate biomarker for small fiber neuropathy. The corneal nerves, accessible through CCM, are frequently impacted in the early stages of DPN, enabling clinicians to detect nerve damage before more severe symptoms develop. Manual analysis of CCM images is labor-intensive, subjective, and requires significant expertise, with interobserver variability that can limit diagnostic accuracy for DPN. Using Deep Learning (DL), Salahouddin et al. (2021) employed a U-Net-based model to automate the segmentation and quantification of corneal nerves in CCM images, achieving discrimination between patients with and without DPN, with an average area under the curve (AUC) of 0.93. Moving toward eliminating the need for pixel-wise annotations, Preston et al. (2022) utilized a ResNet model to diagnose peripheral neuropathy, reporting an average sensitivity of 84% in correctly identifying DPN patients on a test set of 40 images.

Following recent advancements in automated DPN diagnostics, we evaluated a state-of-the-art Vision Transformer (ViT) model for classifying DPN patients using CCM images, comparing its performance to the established ResNet architecture. Our approach, which eliminates the need for pixel-wise annotations, is the first to apply ViTs for DPN classification on CCM images, demonstrating high accuracy on a relatively large dataset. Additionally, we employed Grad-CAM to generate heatmaps, visually highlighting regions that contribute most to the classification decision and confirming a focus on corneal nerves. Figure 1 shows an overview of the transformer-based model architecture for corneal nerve classification.

2 Method

2.1 Dataset

The experiment was carried out on a database of 692 CCM images (358 healthy controls and 334 DPN cases) collected from 106 subjects (29 patients with DPN and 77 healthy controls), captured using the Heidelberg HRTIII corneal confocal microscope. This is a sub-analysis of the LANDMark study (Pritchard et al., 2014)—a multi-center study conducted at the University of Manchester, UK and Queensland University of Technology, Australia in 2009–2014. The LANDMark study adhered to the tenets of the Declaration of Helsinki and was approved by the relevant institutional review boards. Informed, written consent was obtained from all subjects prior to participation.

The images have a size of 384 × 384 pixels, 8-bit gray levels, and are stored in BMP format. To mitigate potential biases arising from the relatively small sample size and the varying number of images per subject, we employed a rigorous data splitting strategy. The dataset was divided into training (60%), validation (20%), and testing sets (20%). To ensure balanced representation across sets, we performed stratified splitting based on subject-level allocation. This ensured that no images from the same subject were included in more than one set, preventing potential bias arising from inter-subject variability.

2.2 Vision transformer

Introduced by Dosovitskiy et al. (2020), Vision Transformers (ViTs) have quickly gained prominence in classification tasks, often outperforming traditional methods (Bazi et al., 2021; Ding et al., 2023; Long et al., 2024). The ViT model includes an embedding layer, a transformer encoder, and an MLP head. The input image is divided into non-overlapping patches, each treated as a token, with position embeddings added to retain spatial information. These embeddings are processed by the encoder, which consists of stacked layers with multiheaded self-attention (capturing relationships across image regions), an MLP block (refining extracted information), and a normalization layer (ensuring data stability). Finally, the MLP head translates encoded information into the predicted class.

Our work leverages the capabilities of ViTs to construct a robust and scalable system, while addressing technical complexities associated with data preprocessing and model development. To enhance efficiency and potentially improve performance, we optimized the original ViT architecture (Dosovitskiy et al., 2020) by reducing the number of Transformer layers, thereby streamlining the model and overfitting. Furthermore, we decreased the MLP size, leading to a substantial decrease in model parameters and computational cost. We modified the input patch size. This trade-off increases the effective sequence length for the Transformer while simultaneously reducing computational complexity, as the number of patches decreases quadratically with the increase in patch size. These modifications resulted in a dramatic reduction in model parameters from 86M to 6M, making our model significantly more compact and potentially easier to deploy on resource-constrained devices.

To enhance model performance and stability, we incorporated a batch normalization layer after the Transformer block. Unlike the original model's layer normalization, which normalized across all features within a sample, our batch normalization normalizes each feature independently across the mini-batch. This modification aims to improve training stability and potentially enhance generalization. To further mitigate overfitting, we integrated Dropout throughout the model architecture. Dropout randomly deactivates a fraction of neurons during training, preventing excessive reliance on specific features and encouraging weight sharing across the network, ultimately leading to more robust and generalizable models.

2.3 Model training

We trained our ViT model using Python 3.7.10 and TensorFlow with Keras on a GPU P100 for 150 epochs. Images were resized to 256 × 256 pixels and divided by the ViT into 144 patches of 20 × 20 pixels each. During training, we applied a combination of feature normalization and data augmentation techniques on each patch, including horizontal flipping, zooming (height and width factor 0.2), and slight rotation (factor 0.02), to enhance model robustness. Optimizing ViT's complex structure is challenging, so we used the AdamW optimizer with Decoupled Weight Decay Regularization, with specific parameters listed in Table 1, carefully selected for a balance of accuracy and efficiency (https://github.com/serag-ai/ViT-CCM).

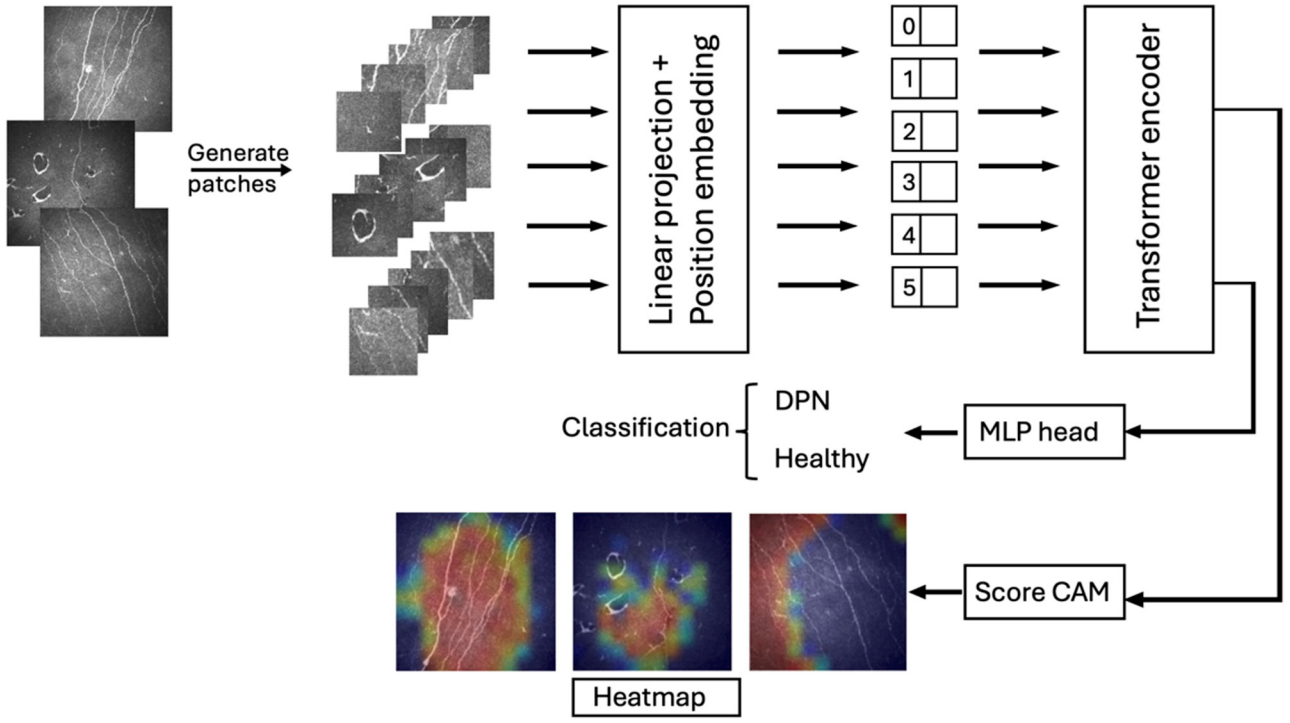


FIGURE 1
Overview of the transformer-based model architecture for corneal nerve classification. The input image is divided into patches, which are linearly projected and positionally embedded before being fed into the transformer encoder. The output representations are used for classification between DPN and healthy cases through an MLP head. Additionally, Score-CAM generates heatmaps highlighting relevant regions in the patches, aiding interpretability in classification.

TABLE 1 Parameters of the trained ViT.

Parameters	Values
Learning rate	0.0001
Weight decay	0.0001
Patch size	20
Batch size	20
number of heads	6
Projection dimension	128
Number of training epochs	150

3 Evaluation metrics

We assessed our model's performance using several key metrics: Area Under the Receiver Operating Characteristic Curve (AUC), Specificity, Sensitivity, and F1-score.

AUC is a threshold-independent metric that evaluates the performance of a classification model. It represents the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance. The AUC ranges from 0 to 1, where a value closer to 1 indicates superior discriminative ability. An AUC of 0.5 suggests no discriminative power, equivalent to random guessing.

Sensitivity, also known as recall, measures the proportion of true positives (*TP*) correctly identified out of all actual positives. It is calculated as:

$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FN} \tag{1}$$

Specificity measures the proportion of true negatives (*TN*) correctly identified out of all actual negatives. It is calculated as:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{2}$$

The F1-score is a harmonic mean of Precision (*Pre*) and Recall (*Rec*), combining them into a single metric. It is calculated as:

$$F_1 = 2 \times \frac{Pre \times Rec}{Pre + Rec} \tag{3}$$

Recall (*Rec*) is defined as in Equation (1), while Precision (*Pre*) is defined as the proportion of true positives out of all positive predictions:

$$Pre = \frac{TP}{TP + FP} \tag{4}$$

In these formulas, *TP* (True Positives) refers to instances correctly classified as positive, while *FP* (False Positives) denotes negative instances that are incorrectly classified as positive. Similarly, *FN* (False Negatives) represents positive instances that are incorrectly classified as negative, and *TN* (True Negatives) refers to instances correctly classified as negative.

TABLE 2 Comparison of AUC, sensitivity, specificity, and F1-score between the ViT model and other methods for the binary classification task.

	AUC	Specificity	Sensitivity	F1-score
EfficientNetB7	0.96	91.35%	94.82%	91.66%
MobileNet	0.98	95.06%	96.55%	94.91%
ResNet50	0.98	96%	98%	96%
ViT	**0.99**	92%	98%	95%

The value in bold shows the highest AUC value.

3.1 Statistical analysis

We also performed a statistical analysis to test the differences between classification results. A *t*-test was conducted, and a *P*-value > 0.05 was interpreted as indicating insufficient evidence to conclude a significant difference between the classification results.

4 Results

4.1 Model performance

The trained ViT model demonstrated outstanding performance in this binary classification task, achieving an AUC of 0.99, which underscores the effectiveness of ViT architectures in extracting discriminative features from CCM images. Specifically, the model correctly classified 75 out of 81 healthy controls, with only one misclassification among DPN cases, resulting in a sensitivity of 98%, specificity of 92%, and a high F1-score of 95%.

4.2 Comparison against other methods

We further compared our method against ResNet50 pretrained on ImageNet (Deng et al., 2009), which has previously been used for detecting DPN in CCM images (Preston et al., 2022; Meng et al., 2023). Table 2 presents the AUC, sensitivity, specificity, and F1-scores for both methods. Our proposed ViT model outperformed ResNet50, achieving a higher AUC compared to ResNet50. Although ResNet50 exhibited a slightly higher F1-score than the ViT model, the difference was not significant ($P = 0.397$).

Besides ResNet50, we have compared our results to well-known DL models including the EfficientNetB7 (Tan and Le, 2019), and MobileNet (Howard, 2017), chosen for their exceptional performance in tasks such as feature extraction and image classification, particularly their capability to detect anomalies within images. In Table 2, we reported a remakrbale AUC for MobileNet of 0.98. However, our ViT beats all these models in term of AUC and F1-score.

To enhance the interpretability of our model's predictions on test images and provide clinicians with greater insight, we employed Grad-CAM (Selvaraju et al., 2017). This attribution method uses the gradients flowing into the final convolutional layer to generate a coarse "attribution map," visually highlighting the regions of the image with the strongest influence on the classification outcome. In essence, the map reveals which parts of the image were most significant in the model's decision-making process. Figure 2 illustrates original and Grad-CAM images from healthy controls (Figure 2A) and patients with DPN (Figure 2B). This clearly identifies areas where corneal nerves are located as providing the most influence to identify DPN.

5 Discussion

In our research, e investigated the potential of the Vision Transformer (ViT) model for classifying corneal confocal microscopy (CCM) images. By splitting images into patches and processing them within a transformer-based architecture, the ViT model effectively captures both local and global features, making it particularly well-suited for tasks requiring a comprehensive view of image content. To our knowledge, this is the first study to apply a ViT model for analyzing and classifying CCM images, achieving a high AUC of 0.99, which surpasses results reported in previous studies (Silva et al., 2015; Salahouddin et al., 2021; Alam et al., 2022; Preston et al., 2022; Meng et al., 2023). These classification results underscore the effectiveness of ViT in distinguishing between healthy controls and individuals with DPN in this context.

To enhance model interpretability and provide clinicians with insights into the ViT model's predictions, we employed Grad-CAM as an explainability tool. Recognizing that Grad-CAM is traditionally designed for CNNs with their hierarchical convolutional layers, we adapted this technique for our ViT architecture. Instead of relying on convolutional feature maps, we leveraged the attention maps generated by the Transformer encoder. By analyzing the attention weights assigned to different image patches, we effectively identified the regions within the CCM images that most significantly influenced the model's predictions. The generated heatmaps, qualitatively validated for their effectiveness, highlighted regions within images that are clinically relevant for diagnosing DPN, such as corneal nerves. This approach not only provides valuable insights into the model's decision-making process but also enhances clinician trust and confidence in its predictions, thereby facilitating potential adoption in clinical settings. To address this, one of the co-authors (RAM), a pioneer of corneal nerve analysis undertook visual inspection of the Grad-CAM heatmaps and confirmed that the highlighted regions were identifying corneal nerve fiber loss, a hallmark of DPN.

While ResNet, a widely adopted CNN architecture, demonstrated competitive results, it has certain limitations. ResNet requires a fixed input size, which can be restrictive when working with images of varying dimensions (Salehi et al., 2023), and it struggles to capture long-range dependencies, which are often essential for identifying complex patterns. In contrast, ViT models, in principle, can process images of different dimensions due to their inherent self-attention mechanisms and the fact of processing images with patches. Practical implementations often necessitate training with a specific input resolution for computational efficiency. In our case, the input to our ViT model consists of CCM images with their original size of 384 × 384 pixels. However, during the internal image augmentation process within the model, these images are resized to 256 × 256 pixels. This choice was made to optimize training efficiency by enabling efficient batch processing and optimized memory usage, leading to faster training times. This approach, while introducing a degree of constraint,

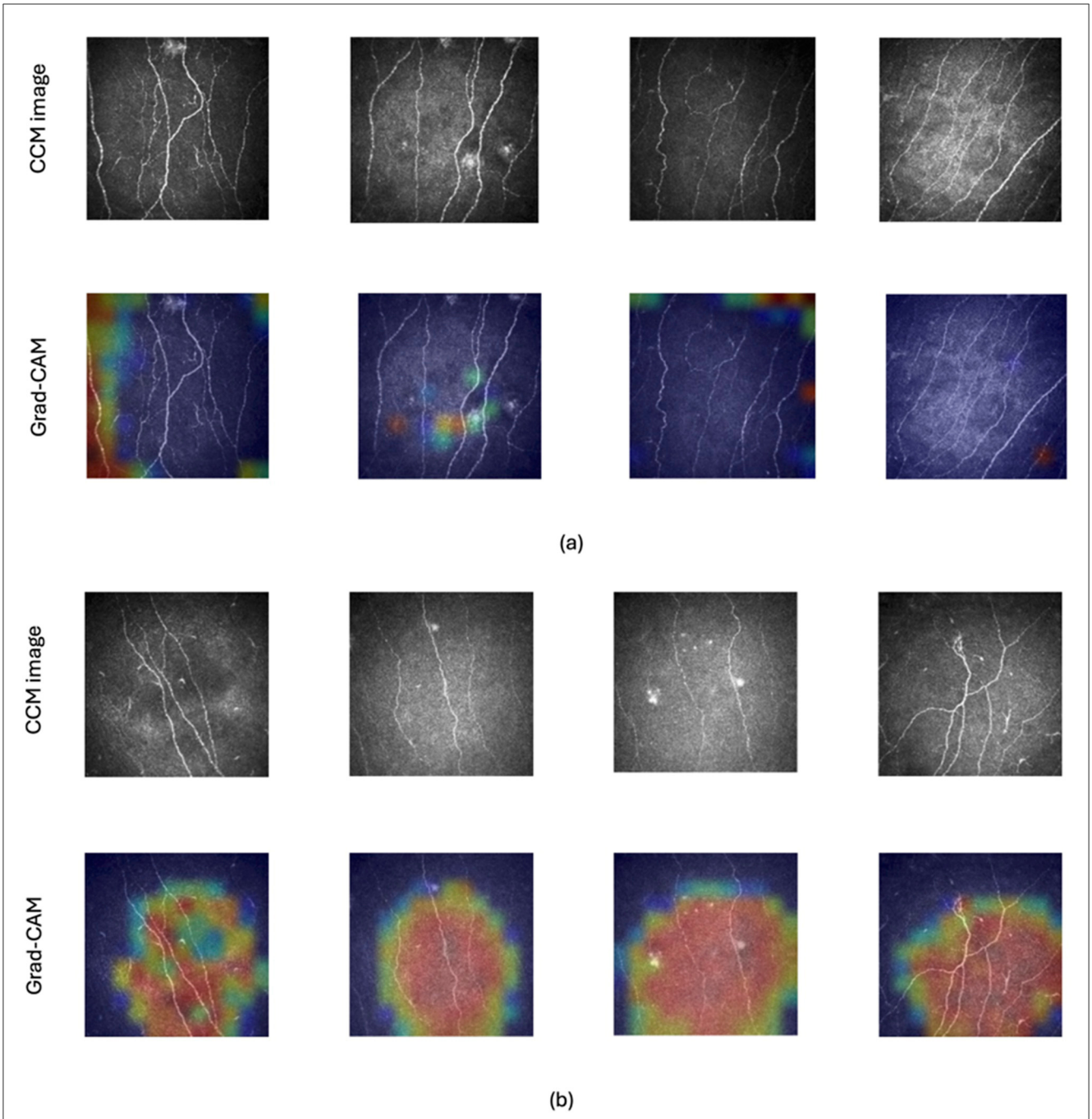


FIGURE 2
Example of CCM images of healthy controls and patients with and without DPN along with their corresponding Grad-CAM images. Grad-CAM creates a heatmap where hotter colors (red) indicate the image regions that had the strongest impact on the model's classification decision. Cooler colors (orange, yellow, and green) represent progressively less influential areas, with shades of blue highlighting the regions with the weakest influence. The top two rows **(A)** display images of healthy control subjects and the bottom two rows **(B)** present images of patients with DPN.

does not inherently limit the model's generalizability to images of different dimensions. ViT's architecture, with its self-attention mechanisms, allows it to flexibly handle varying input sizes while capturing long-range dependencies, making it a more adaptable and powerful choice for tasks that demand a deep understanding of image-wide context. In real-world applications, this approach, combined with the inherent flexibility of the ViT architecture, allows for a degree of adaptability to varying input dimensions.

Furthermore, ViTs are renowned for their scalability (Pan et al., 2021; Chen et al., 2022; Dehghani et al., 2023), as their performance typically improves with larger datasets and increased model complexity. This scalability is particularly advantageous for medical applications, where large datasets and robust models are often essential for achieving high diagnostic accuracy. Building on this scalability, our research demonstrates that ViT models can effectively detect DPN using CCM images without requiring

complex pre-processing steps, segmentation, or adaptive feature extraction techniques.

This study, while demonstrating promising results, acknowledges several limitations. Firstly, the relatively small sample size (692 images) may limit the generalizability of the findings.

Secondly, the integration of this AI model into clinical practice presents several challenges. The computational demands of ViT models, while mitigated through optimizations employed in this study, may still pose challenges in resource-limited clinical settings.

Furthermore, the use of AI in healthcare raises important ethical considerations, including data privacy, algorithmic bias, and the potential for unintended consequences. Ensuring responsible and equitable AI development and deployment is paramount. To safeguard patient privacy while advancing AI models in healthcare, two promising approaches are federated learning and synthetic data generation. Federated learning enables model evaluation and refinement without transferring sensitive patient data, while synthetic data generation creates artificial data that mimics real data without containing any actual patient information. These innovative solutions offer a balance between model improvement and robust privacy protection.

These findings suggest that ViT models may offer a more efficient and accurate approach to DPN diagnosis compared to traditional methods. To fully harness the potential of ViTs, future research should focus on developing training sets encompassing a broader range of normal and abnormal pathologies, exploring the practical implementation of this algorithm in clinical workflows, and comparing its performance to existing diabetic neuropathy screening techniques. This will be crucial for translating this technology into real-world healthcare solutions.

This study serves as a foundation for future research that will address the identified shortcomings. Further research with larger, more diverse cohorts is warranted to confirm these initial observations. Moreover, incorporating other medical image modalities can be used to assess the robustness of the model in peripheral neuropathies classification.

In conclusion, this study presents a novel application of AI for the automated classification of CCM images, enabling rapid and objective detection of DPN. Our vision transformer-based model demonstrated remarkable accuracy in distinguishing patients with DPN from healthy controls. By eliminating the subjectivity and time-intensive processes of manual image segmentation and interpretation, this approach offers a faster and more consistent analysis. The integration of this AI-driven tool into clinical workflows has the potential to revolutionize DPN diagnosis by enabling quicker decision-making, facilitating timely interventions, and ultimately improving patient outcomes. While the results are promising, further research is needed to refine the model and extend its applicability. Future studies should utilize larger datasets, including diabetic patients with diverse comorbidities, to enhance model interpretability and provide clinicians with more actionable insights. This research highlights the transformative potential of AI in medical diagnostics. By automating complex tasks and improving diagnostic accuracy, AI-driven solutions can advance patient care and contribute to the effective management of diabetes-related neuropathies.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: the dataset contains de-identified patient data and is subject to ethical and confidentiality requirements. Access is limited to authorized researchers for academic and non-commercial purposes only. Approval from the institutional review board is required before access is granted. All requests for data access should be submitted to RM. along with a detailed research proposal and signed data use agreement. Requests to access these datasets should be directed to ram2045@qatar-med.cornell.edu.

Ethics statement

The study was approved by the North Manchester Research Ethics Committee. Informed consent: All study participants provided written informed consent. Registry and the registration no. of the study/trial: (Ethical approval number: #09/H1006/38). 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

CBR: Conceptualization, Methodology, Project administration, Writing – original draft. INP: Data curation, Writing – review & editing. RAM: Data curation, Validation, Visualization, Writing – review & editing. AS: Conceptualization, Methodology, Supervision, Validation, Writing – review & editing.

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 all volunteers for their participation.

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.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

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

Publisher's note

References

Alam, U., Anson, M., Meng, Y., Preston, F., Kirthi, V., Jackson, T. L., et al. (2022). Artificial intelligence and corneal confocal microscopy: the start of a beautiful relationship. *J. Clin. Med.* 11:6199. doi: 10.3390/jcm11206199

Bazi, Y., Bashmal, L., Rahhal, M. M. A., Dayil, R. A., and Ajlan, N. A. (2021). Vision transformers for remote sensing image classification. *Rem. Sens.* 13:516. doi: 10.3390/rs13030516

Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., et al. (2022). "Adaptformer: adapting vision transformers for scalable visual recognition," in *Advances in Neural Information Processing Systems*, 16664–16678.

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., et al. (2023). "Scaling vision transformers to 22 billion parameters," in *International Conference on Machine Learning* (PMLR), 7480–7512.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "ImageNet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE), 248–255. doi: 10.1109/CVPR.2009.5206848

Ding, M., Qu, A., Zhong, H., Lai, Z., Xiao, S., and He, P. (2023). An enhanced vision transformer with wavelet position embedding for histopathological image classification. *Patt. Recognit.* 140:109532. doi: 10.1016/j.patcog.2023.109532

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Eid, S. A., Rumora, A. E., Beirowski, B., Bennett, D. L., Hur, J., Savelieff, M. G., et al. (2023). New perspectives in diabetic neuropathy. *Neuron* 111, 2623–2641. doi: 10.1016/j.neuron.2023.05.003

Elafros, M. A., Andersen, H., Bennett, D. L., Savelieff, M. G., Viswanathan, V., Callaghan, B. C., et al. (2022). Towards prevention of diabetic peripheral neuropathy: clinical presentation, pathogenesis, and new treatments. *Lancet Neurol.* 21, 922–936. doi: 10.1016/S1474-4422(22)00188-0

Howard, A. G. (2017). MobileNets: efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Jensen, T. S., Karlsson, P., Gylfadottir, S. S., Andersen, S. T., Bennett, D. L., Tankisi, H., et al. (2021). Painful and non-painful diabetic neuropathy, diagnostic challenges and implications for future management. *Brain* 144, 1632–1645. doi: 10.1093/brain/awab079

Long, Z., McCreadie, R., and Imran, M. (2024). CrisisViT: a robust vision transformer for crisis image classification. *arXiv preprint arXiv:2401.02838*.

Meng, Y., Preston, F. G., Ferdousi, M., Azmi, S., Petropoulos, I. N., Kaye, S., et al. (2023). Artificial intelligence based analysis of corneal confocal microscopy images for diagnosing peripheral neuropathy: a binary classification model. *J. Clin. Med.* 12:1284. doi: 10.3390/jcm12041284

Ong, K. L., Stafford, L. K., McLaughlin, S. A., Boyko, E. J., Vollset, S. E., Smith, A. E., et al. (2023). Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the global burden of disease study 2021. *Lancet* 402, 203–234. doi: 10.1016/S0140-6736(23)01301-6

Pan, Z., Zhuang, B., Liu, J., He, H., and Cai, J. (2021). "Scalable vision transformers with hierarchical pooling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 377–386. doi: 10.1109/ICCV48922.2021.00043

Ponirakis, G., Elhadd, T., Al Ozairi, E., Brema, I., Chinnaiyan, S., Taghadom, E., et al. (2022). Prevalence and risk factors for diabetic peripheral neuropathy, neuropathic pain and foot ulceration in the Arabian gulf region. *J. Diab. Investig.* 13, 1551–1559. doi: 10.1111/jdi.13815

Ponirakis, G., Elhadd, T., Chinnaiyan, S., Hamza, A. H., Sheik, S., Kalathingal, M. A., et al. (2021). Prevalence and risk factors for diabetic neuropathy and painful diabetic neuropathy in primary and secondary healthcare in Qatar. *J. Diab. Investig.* 12, 592–600. doi: 10.1111/jdi.13388

Preston, F. G., Meng, Y., Burgess, J., Ferdousi, M., Azmi, S., Petropoulos, I. N., et al. (2022). Artificial intelligence utilising corneal confocal microscopy for the diagnosis of peripheral neuropathy in diabetes mellitus and prediabetes. *Diabetologia* 65, 457–466. doi: 10.1007/s00125-021-05617-x

Pritchard, N., Edwards, K., Dehghani, C., Fadavi, H., Jeziorska, M., Marshall, A., et al. (2014). Longitudinal assessment of neuropathy in type 1 diabetes using novel ophthalmic markers (landmark): study design and baseline characteristics. *Diabetes Res. Clin. Pract.* 104, 248–256. doi: 10.1016/j.diabres.2014.02.011

Salahouddin, T., Petropoulos, I. N., Ferdousi, M., Ponirakis, G., Asghar, O., Alam, U., et al. (2021). Artificial intelligence-based classification of diabetic peripheral neuropathy from corneal confocal microscopy images. *Diabetes Care* 44:e151. doi: 10.2337/dc20-2012

Salehi, A. W., Khan, S., Gupta, G., Alabduallah, B. I., Almjally, A., Alsolai, H., et al. (2023). A study of CNN and transfer learning in medical imaging: advantages, challenges, future scope. *Sustainability* 15:5930. doi: 10.3390/su15075930

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 618–626. doi: 10.1109/ICCV.2017.74

Silva, S. F., Gouveia, S., Gomes, L., Negrão, L., Quadrado, M. J., Domingues, J. P., et al. (2015). "Diabetic peripheral neuropathy assessment through texture based analysis of corneal nerve images," in *Journal of Physics: Conference Series* (IOP Publishing), 012002. doi: 10.1088/1742-6596/616/1/012002

Tan, M., and Le, Q. (2019). "EfficientNet: rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning* (PMLR), 6105–6114.

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 15 October 2024
DOI 10.3389/fonc.2024.1440881

OPEN ACCESS

EDITED BY
Paolo Andreini,
University of Siena, Italy

REVIEWED BY
Diego Raimondo,
University of Bologna, Italy
Kui Sun,
Shandong Provincial Hospital, China

*CORRESPONDENCE
Yuwang Zhou
1447129042@qq.com
Yun Fang
494853510@qq.com

†These authors have contributed equally to this work

RECEIVED 05 June 2024
ACCEPTED 23 September 2024
PUBLISHED 15 October 2024

CITATION
Luo Y, Yang M, Liu X, Qin L, Yu Z, Gao Y, Xu X, Zha G, Zhu X, Chen G, Wang X, Cao L, Zhou Y and Fang Y (2024) Achieving enhanced diagnostic precision in endometrial lesion analysis through a data enhancement framework. *Front. Oncol.* 14:1440881. doi: 10.3389/fonc.2024.1440881

Achieving enhanced diagnostic precision in endometrial lesion analysis through a data enhancement framework

Yi Luo[1,2†], Meiyi Yang[3†], Xiaoying Liu[4], Liufeng Qin[4], Zhengjun Yu[5], Yunxia Gao[6], Xia Xu[7], Guofen Zha[8], Xuehua Zhu[9], Gang Chen[5], Xue Wang[4], Lulu Cao[10], Yuwang Zhou[4*] and Yun Fang[4*]

[1]Medical Engineering Cross Innovation Consortium, Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, Zhejiang, China, [2]Medical Engineering Cross Innovation Consortium, The Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou People's Hospital, Quzhou, Zhejiang, China, [3]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, China, [4]Department of Ultrasound, The Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou People's Hospital, Quzhou, Zhejiang, China, [5]Department of Ultrasound, Kaihua County People's Hospital, Quzhou, Zhejiang, China, [6]Department of Ultrasound, The Second People's Hospital of Quzhou, Quzhou, Zhejiang, China, [7]Department of Ultrasound, Changshan County People's Hospital, Quzhou, Zhejiang, China, [8]Department of Ultrasound, People's Hospital of Quzhou Kecheng, Quzhou, Zhejiang, China, [9]Department of Ultrasound, Quzhou Maternal and Child Health Care Hospital, Quzhou, Zhejiang, China, [10]Department of Pathology, The Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou People's Hospital, Quzhou, Zhejiang, China

Objective: The aim of this study was to enhance the precision of categorization of endometrial lesions in ultrasound images via a data enhancement framework based on deep learning (DL), through addressing diagnostic accuracy challenges, contributing to future research.

Materials and methods: Ultrasound image datasets from 734 patients across six hospitals were collected. A data enhancement framework, including image features cleaning and soften label, was devised and validated across multiple DL models, including ResNet50, DenseNet169, DenseNet201, and ViT-B. A hybrid model, integrating convolutional neural network and transformer architectures for optimal performance, to predict lesion types was developed.

Results: Implementation of our novel strategies resulted in a substantial enhancement in model accuracy. The ensemble model achieved accuracy and macro-area under the receiver operating characteristic curve values of 0.809 of 0.911, respectively, underscoring the potential for use of DL in endometrial lesion ultrasound image classification.

Conclusion: We successfully developed a data enhancement framework to accurately classify endometrial lesions in ultrasound images. Integration of

anomaly detection, data cleaning, and soften label strategies enhanced the comprehension of lesion image features by the model, thereby boosting its classification capacity. Our research offers valuable insights for future studies and lays the foundation for creation of more precise diagnostic tools.

KEYWORDS

deep learning, data enhancement framework, endometrial cancer, ultrasonography, diagnosis

1 Introduction

Patients with endometrial cancer, otherwise referred to as cancer of the uterine body, have a highly variable prognosis; crucially, the survival rate can be significantly improved through early detection and diagnosis (1, 2). In clinical practice, patients with postmenopausal bleeding are generally diagnosed through various means, including imaging, pathological examination, and serum tumor markers (3, 4). Magnetic resonance imaging (MRI) and computed tomography (CT) are relatively accurate imaging methods, but are expensive and CT poses significant radiation hazards. Further, although curettage and hysteroscopy are key steps in the diagnostic process, they are somewhat invasive for patients. In contrast, ultrasound examination is convenient, non-invasive, inexpensive, and repeatable, and is often used as a first-line diagnostic tool for endometrial lesions (5, 6). Ultrasonography is also an important means of large-scale asymptomatic population screening, where early detection of endometrial cancer by large-scale screening can significantly improve patient prognosis (7). Nevertheless, since physical condition and disease state vary in each patient, there is currently no universal diagnostic indicator for endometrial cancer (4). Additionally, the accuracy of ultrasound examination is affected by factors including the technical ability of medical personnel and environmental noise. Reznak et al. found that the success rate of ultrasound examination in predicting polyps is 65.1%, and that it has limited predictive value when used alone (8). Therefore, there is an urgent need for an auxiliary screening method that can effectively improve the accuracy of ultrasound examination in diagnosing endometrial cancer.

In recent years, artificial intelligence, particularly deep learning (DL), has made significant progress in medical image recognition (9–11). For instance, numerous developmental directions have emerged in the application of deep learning for the diagnosis of endometrial lesions. Based on MRI images, DL models can automatically locate, segment, and measure the degree of muscle infiltration of endometrial cancer (12–15); however, DL research based on ultrasound images is relatively scarce. Hu et al. (16) and Liu et al. (17) each proposed endometrial thickness measurement models based on transvaginal ultrasound (TVUS) images; however, these models cannot be directly applied to endometrial lesion classification. Other features in ultrasound images, such as uniformity of endometrial echo and blood flow signals, are also crucial for distinguishing benign and malignant endometrial lesions (18, 19). Further, DL also performs poorly in the task of ultrasound image classification. Raimondo et al. (20) used a DL model to diagnose adenomyosis based on TVUS images, and the results indicated that the diagnostic accuracy of the DL model was lower than that of general ultrasound doctors, although it had higher specificity in identifying healthy uteruses and reducing overdiagnosis. Therefore, we sought to improve model learning and utilization of various ultrasound image features using DL methods to enhance endometrial lesion classification accuracy.

In this study, we developed a DL model for automatic identification of endometrial lesions using an innovative combination comprising multi-stage anomaly detection, a data cleaning process, and a soft label strategy, to improve model understanding of lesion image features and enhance its classification ability. Our experiments explored the relationships among lesion features, models, and different degrees of softening (τ). Final accuracy was also enhanced through integration of several different models.

2 Materials and methods

2.1 Patients

This multicenter retrospective diagnostic study was conducted in line with the principles of the Declaration of Helsinki. This study was approved by the Ethics Committee of the People's Hospital of Quzhou City (No. 2022-148). Ultrasound examination images were collected from March 2014 to March 2023 at six hospitals: The Quzhou Affiliated Hospital of Wenzhou Medical University, Changshan County People's Hospital, Kaihua County People's Hospital, People's Hospital of Quzhou Kecheng, The Second People's Hospital of Quzhou, and Quzhou Maternal And Child Health Care Hospital. Inclusion criteria: 1. Non-pregnant women who have had sexual intercourse and consent to transvaginal ultrasound examinations. 2. Patients with confirmed pathological diagnoses via hysteroscopy or endometrial biopsies. Exclusion criteria: 1) Patients who have not had sexual intercourse and are thus ineligible for transvaginal ultrasound examinations. 2) Patients are allergic to condoms and thus unsuitable for ultrasound examinations. 3) Patients with severe reproductive system abnormalities or acute inflammation who are contraindicated

for transvaginal ultrasound examinations. 4) Patients with severe psychological disorders who are unsuitable for transvaginal ultrasound examinations. 5) Each patient's endometrial ultrasound images are collected in two views: all longitudinal images and all transverse images for each case. 6) Image blurring due to significant visual losses and damages during the collection process, along with interferences like gas and artifacts. All images were collected by professional radiologists, and saved in DICOM format. Then, the ultrasound images are further screened, as shown in Figure 1; 734 patients were ultimately included in the study.

2.2 Data processing

After collection, all ultrasound data were converted from DICOM into JPG files using Python for research. Since data were derived from multiple different hospitals, some preprocessing measures were performed on all images for experiments, including manual cropping to retain only the part captured by the instrument and scaling to 224 × 224. Finally, to improve model robustness and generalization ability, data augmentation techniques, including random-cropping, random-flipping, and TrivialAugment (21) were also used during the training phase. In the testing phase, only size adjustment and normalization of the original images were conducted.

2.3 Data enhancement framework

An innovative data augmentation framework, primarily encompassing data cleaning and label softening procedures, was developed in this study. The processing of training set data using this framework is summarized in Figure 2. Following a feature extraction process, image feature cleaning, and soften label implementation, the training set was utilized to generate a softened set for training purposes.

2.3.1 Image feature cleaning

Medical data are often intricate, encompassing numerous variables and factors, and the diverse types of noise they contain represents a substantial challenge (22). For example, data for the present research was sourced from multiple hospitals, where the process of ultrasound image acquisition is influenced by objective factors, such as equipment performance, environmental noise, and patient size and positioning, which can lead to the presence of abnormal images and noise within the dataset, with potential to impair model performance. To mitigate this possibility, a rigorous data cleaning process was initiated following division of the original data into training and testing sets.

As illustrated in Figure 2, five-fold cross-validation was first applied to partition the training set into five subsets, four of which were used to train an independent DL model. These models were primarily tasked with predicting the results from the remaining subsets and generating corresponding image feature vectors. In this study, ResNet34 was used as the backbone network of the framework. Finally, five sets of experimental results were connected to form a complete training set of image features.

Subsequently, anomaly detection methods, such as Isolation Forests (23), were introduced to analyze the feature vectors of the generated training set and exclude potential anomalous data. The training sets selected by three methods were then merged to form a new, cleaned training set. In this study, we selected Isolation Forest, Local Outlier Factor, and One-Class SVM. The selection of methods is

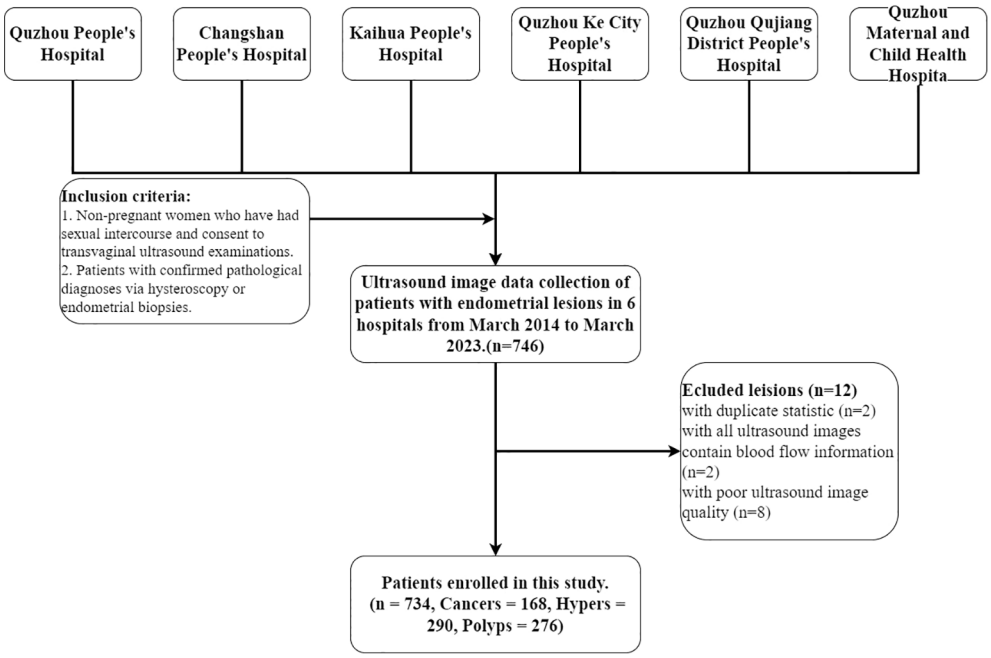


FIGURE 1
Patient selection workflow. A total of 746 patients with endometrial lesions were collected, of which 734 were ultimately included in the analysis. Cancers, Hypers, and Polyps indicate patients with these types of lesions.

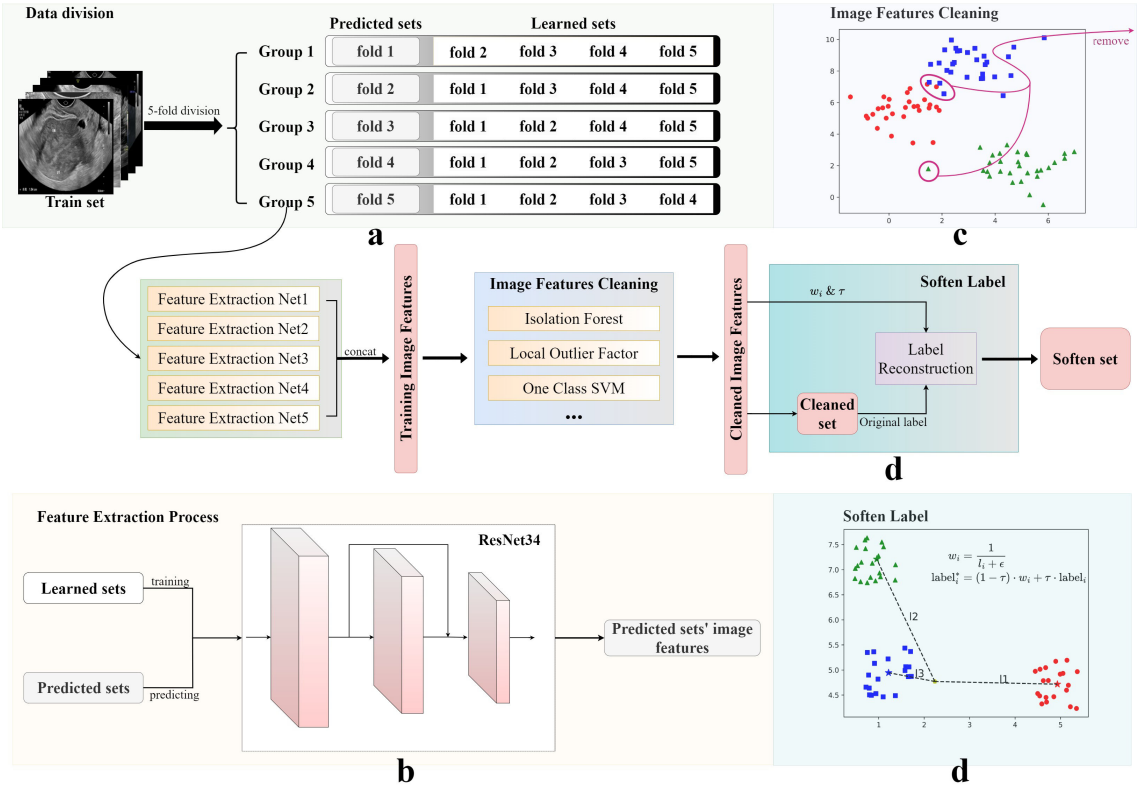


FIGURE 2
Image features cleaning and soften label processes. The original training set was obtained using four steps: **(A)** data division, **(B)** image features cleaning, **(C)** feature extraction, and **(D)** soften label, to obtain the final soften set. The Soften Label subfigure shows the calculation formula used for softening labels.

contingent upon the data and the specifics of the research. This innovative approach to data cleaning ensured the robustness of the developed model, despite the diverse and potentially noisy data sources.

2.3.2 Soften label

To enhance generalization ability of the model and alleviate overfitting, a label smoothing strategy was implemented, based on the inverse proportion of image-to-cluster center distance. As shown in Figure 2, Soften Label included the following processes: first, dimension reduction and clustering were performed on the new processed training set; then, the center of each category cluster and the distance of each image to each cluster center were calculated; finally, new labels were formed, according to the distance ratio. In addition, an adjustable temperature, τ, was introduced, to control the smoothness of the label. The new label for training was obtained by calculating the inverse distance ratio multiplied by τ, plus the hard label value. Datasets were named at different processing stages as the cleaned set and the softened set.

2.4 Model architecture and training strategy

In this study, a hybrid model to predict patient lesion types, based on convolutional neural network (CNN) and Transformer architectures, is proposed, with the aim of maximizing prediction accuracy. As shown in Figure 3, the proposed model combines three classic CNN models (ResNet50, DenseNet169, and DenseNet201) and ViT-B, leveraging the complementary strengths of these different models to enhance the accuracy of endometrial ultrasound image classification.

The multilayer perceptron layer of the original model was tailored to suit this classification task. Each preprocessed image was fed into the model for automatic processing, outputting a three-dimensional array. After Log-Softmax function processing, the prediction probability for each image was obtained. During model integration, the prediction probabilities from all sub-models were weighted to yield the final result. In the testing phase, the average prediction probability for all images from a single patient was calculated, to determine the prediction result.

The experiment comprised three stages. Initially, unmodified ResNet50 was employed as the base model and the impact of different data processing methods on model performance assessed. Subsequently, the applicability of the proposed method was explored by training various CNN and visual transformer models, and the results statistically analyzed after setting the τ value. Finally, high-performing models from the second stage were integrated to test the performance of the optimal model. During the training process, CrossEntropyLoss was used as the objective function, and AdamW was used as the optimizer for end-to-end training. Additionally, the transformer architecture network was loaded with pretraining parameters.

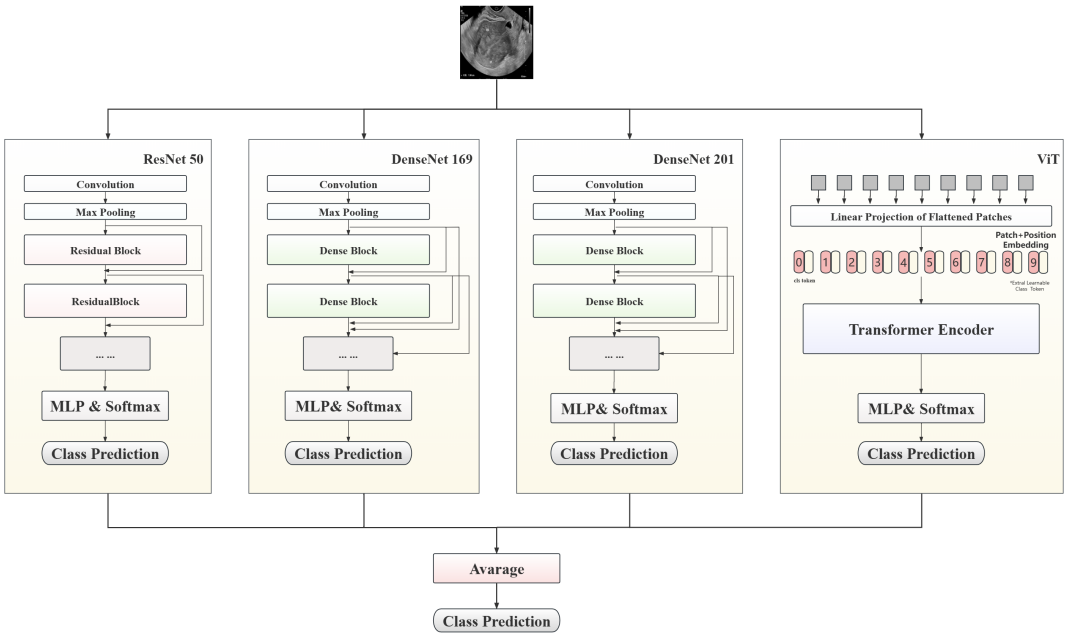


FIGURE 3
Architecture of model integrating ResNet50, DenseNet169, DenseNet201, and ViT-B.

In this manuscript, suffixes have been added to indicate different models; for example, ResNet50_b represents the baseline model, while ResNet50_c represents the model trained using the cleaned set data. Similarly, models trained using softened set data have the suffix "_s".

2.5 Devices and software

This was a multicenter study, and different hospitals used various devices for the data collection process, including Samsung WS80A, GE Volkswagen E10, GE Volkswagen E8, PhilipPsQ5, PhilipPsQ7, and Mindray Resona 6s. All equipment met the experimental requirements. The protocols for each scanning instrument are shown in Table 1.

2.6 Statistical analysis

Statistical analyses were performed during the testing phase, with individual cases serving as the smallest unit of measurement. Models were validated on a test set, followed by statistical evaluation of the confusion matrix derived from the validation outcomes. Additionally, receiver operating characteristic (ROC) curves were plotted. Primary indicators for comparing model performance were accuracy and area under the ROC curve (AUC); sensitivity and specificity were also considered as indicators of the classification capabilities of models. Two visualization techniques, Grad-CAM (24) and t-SNE (25), were employed to elucidate the operational mechanism of the model.

3 Results

3.1 Case inclusion and grouping

Among 1875 high-quality images from 734 patients, we randomly extracted 30% of cases as a test set. The remaining images were used as the original training set for data augmentation and model training. The detailed dataset partitions used in this study are presented in Table 2. All experiments were trained and tested using the same data-division. Our final model achieved the best performance, with accuracy and macro-AUC values of 0.809 and 0.911, respectively.

3.2 Impact of innovative strategies

In the methods testing phase, we chose ResNet50 as the baseline model. Model performance was significantly improved through feature cleaning and soften label processing. As shown in Table 3, when the original training set was used for training, the accuracy of the test set was only 0.691. This provided us with a comparison baseline; the baseline was determined in the same way for each model in subsequent multi-model comparisons. We noticed that abnormal images in the training set could affect model training; therefore, we used feature cleaning to reprocess the training set. After obtaining relatively clean data, the accuracy of the model on the test set increased to 0.741. In subsequent experiments, we used a label-softening method to reconstruct the labels in the new dataset. Under the same data augmentation and image preprocessing, the

TABLE 1 Scanning Instrument Protocol.

	GE Voluson E8	GE Voluson E6	mindray Resona 7s	mindray Resona 6s	PHILIPS EPIQ-7	PHILIPS EPIQ-5
Intracavitary probe	RIC5-9	IC5-9-D	V11-3HU	DE10-3U	3D9-3V	C10-3V
Probe frequency	5-9MHz	5-9MHz	3-11MHz	3-10MHz	3-9MHz	3-10MHz
Bandwidth	4.5-9.8MHz	4.5-9.8MHz	2.5-12.2MHz	2.8-11.8MHz	2.7-9.2MHz	2.8-1.2MHz
TIS	0.4	0.4	0.3	0.3	0.3	0.4
Depth	6.0cm	7.0cm	7.0cm	8.0cm	7.0cm	6.0cm
Magnification	1.2	1.5	1.1	1.1	1.1	1.1
Maximum fan angle	180°	180°	180°	180°	180°	180°
Frame rate	40HZ	41HZ	42HZ	42HZ	49HZ	47HZ
Gain	40%-80%	40-70%	40-70%	40-70%	40-70%	40-70%
Dynamic range	50-120	50-120	50-120	50-120	50-120	50-120

TABLE 2 Partition details of the endometrial lesion classification dataset.

Category	Datasets		Training set		Testing set	
	Patients	Images	Patients	Images	Patients	Images
Cancer	168	460	118	323	50	137
Hyper	290	661	203	470	87	191
Polyp	276	746	193	506	83	240
Total	734	1867	514	1299	220	568

accuracy of the model increased to 0.764. The independence and invariance of the test set were ensured in each training batch.

Label smoothness was controlled using the parameter, τ, which is similar to the smoothing coefficient in Label Smooth (26). In this experiment, we introduced a variety of different τ values, to generate different soften-labels. ResNet50 showed different classification capabilities under different values of τ. As shown in Table 4, ResNet50 performed best when τ was 0.7. To further study the impact of τ on model training, we introduced five other models, including DenseNet169, DenseNet201, EfficientNetB4, VGG16-bn and ViT-B. As shown in Table 4, our framework effectively improved the representation learning of various models, indicating that the improvement in the performance of ResNet50 was not isolated. Further, the best performance of each model corresponded to different values of τ. Among individual models, DenseNet201 achieved the best accuracy when τ was 0.9. When τ was 0.7, the performances of ResNet50, DenseNet169, and VGG16-bn were better than those achieved with other softening coefficients. These conditions may indicate that the optimal value of τ may vary depending on the characteristics of the dataset, model, and study.

3.3 Prediction model performance

As shown in Figure 4, the confusion matrixes for each model effectively reflected their classification performance. In terms of overall accuracy, the DenseNet201_s model exhibited outstanding performance, achieving a best score of 0.786, particularly in recognition of polyp class images, for which it had the best single-category recall rate. We also plotted ROC curves for DenseNet169_s and DenseNet201_s, to evaluate and compare their performances by measuring AUC values (Figure 4). We

TABLE 3 Impact of different data processing approaches on model performance.

Dataset	Model	ACC	AUC	F1	Recall	Precision
Base	ResNet50	0.691	0.811	0.680	0.665	0.697
Cleaned set		0.741	0.845	0.736	0.728	0.744
Soften label		**0.764**	**0.873**	**0.752**	**0.745**	**0.759**

Boldface numerals are utilized to underscore the optimal results in this group's trial.

TABLE 4 Model performance comparison (Accuracy).

Model	Base	Soften label (τ)			
		0.6	0.7	0.8	0.9
ResNet50	0.691	0.727	0.764	0.714	0.718
DenseNet169	0.727	0.755	0.782	0.736	0.736
DenseNet201	0.731	0.764	0.745	0.75	**0.786**
EfficientNetB4	0.672	0.7	0.69	0.714	0.745
VGG16-bn	0.682	0.732	0.745	0.695	0.727
ViT-B	0.736	0.782	0.723	0.759	0.75

Boldface numerals are utilized to underscore the optimal results in this group's trial.

found that DenseNet-201_s was the single model with the best comprehensive classification performance in this study.

In the final phase of our experiment, we implemented an ensemble model approach to enhance the performance of our model. The ensemble models were constructed based on the performance ranking of models as indicated in Table 4. As demonstrated in Table 5, the Ensemble Model2, which is comprised of ResNet50_s, DenseNet169_s, DenseNet201_s, and ViT-B models, yielded the most superior test results, achieving an accuracy of 0.809 and a macro-AUC of 0.911. As illustrated in Figure 4, the Ensemble Model2 outperforms DenseNet201_s in the classification of cancer and hyperplasia. The macro-AUC value of the Ensemble Model2 has significantly improved, and the ROC curve is also more reasonable.

3.4 Model visualization

The operation process of DL models is often viewed as a 'black box' prediction; however, we applied the Grad-CAM and t-SNE visualization methods to explain the working mechanism used by our DL model.

In Grad-CAM, we used hook functions to generate the gradient of the last dense module of the model and stacked these gradients onto the original image to generate heat maps. As shown in Figure 5, the areas of interest for the model can be distinguished by depth of color. From these images, it can be observed that the model accurately focused on lesion areas in the endometrium; more attention was paid to these areas, and these local features deeply affected model prediction.

We also intuitively observed the training effect of the model using the t-SNE method to count the feature vectors extracted by the model. In the high-dimensional space of feature vectors, we calculated the

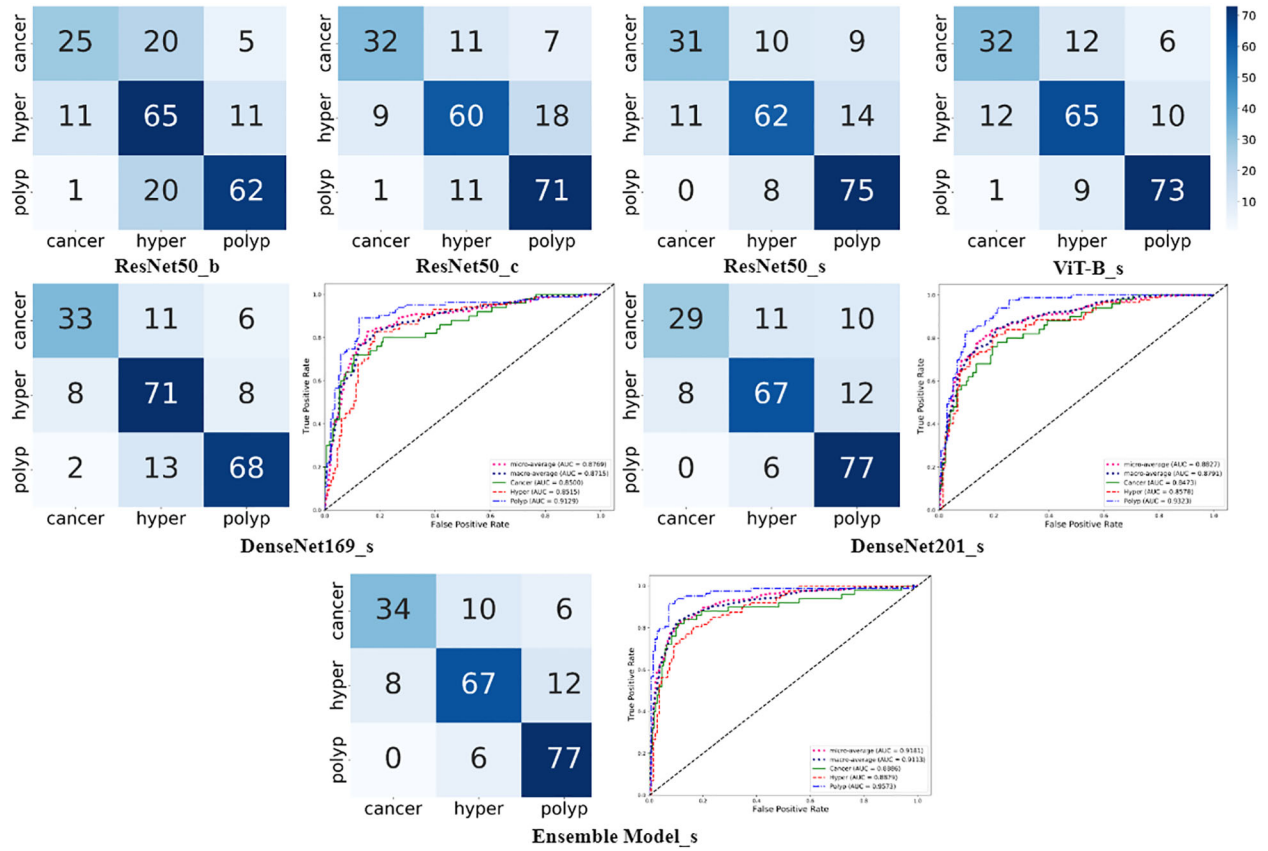


FIGURE 4
Charts summarizing statistical analysis of results from seven different models. Matrix diagrams represent confusion matrices, while the line plots are ROC curves.

TABLE 5 Performance of ensemble models with different compositions.

Model	Model Composition	ACC	AUC
Ensemble model1	**DenseNet169+DenseNet201 +ViT-B**	0.777	0.898
Ensemble model2	**Ensemble model1+ ResNet50**	**0.809**	**0.911**
Ensemble model3	**Ensemble model2+ EfficientNetB4**	0.805	0.908
Ensemble model4	**Ensemble model2+ VGG16-bn**	0.791	0.906
Ensemble model5	**Ensemble model3+ VGG16-bn**	0.782	0.912

Boldface numerals are utilized to underscore the optimal results in this group's trial.

similarities between each data point and mapped these data to low-dimensional space for visualization, and compared the clustering diagrams before and after model training (Figure 6). As illustrated in Figure 6, most images were mapped in their fixed areas through training, but there was overlap among certain categories. Further, the distance between different category cluster centers reflected the intrinsic relationship of their key image features to a certain extent. We proposed a soften-label method based on this principle.

4 Discussion

In the burgeoning field of DL, our study represents a pioneering effort to accurately classify endometrial lesions in ultrasound images using DL models. We achieved an automatic classification with a final accuracy of 0.809 and a macro-AUC value of 0.911.

To maximize DL model effectiveness, we established an innovative data augmentation framework. In this study, collection of datasets from multiple centers ensured inclusion of diverse endometrial lesion ultrasound data. Although this diversity ensured the generalization performance of the model, it also introduced additional noise, which is an inherent challenge commonly present in medical datasets. Within our data augmentation framework, we implemented a scalable data cleaning process, including selection of appropriate feature extraction networks and anomaly detection methods, which significantly improved the accuracy of ResNet50 on our test set, from 0.70 at baseline to 0.741. Another challenge arose from the low signal-to-noise ratio of ultrasound images and the similarity of lesion image features. To address this, we incorporated a label softening strategy, based on clustering and inverse distance, into the

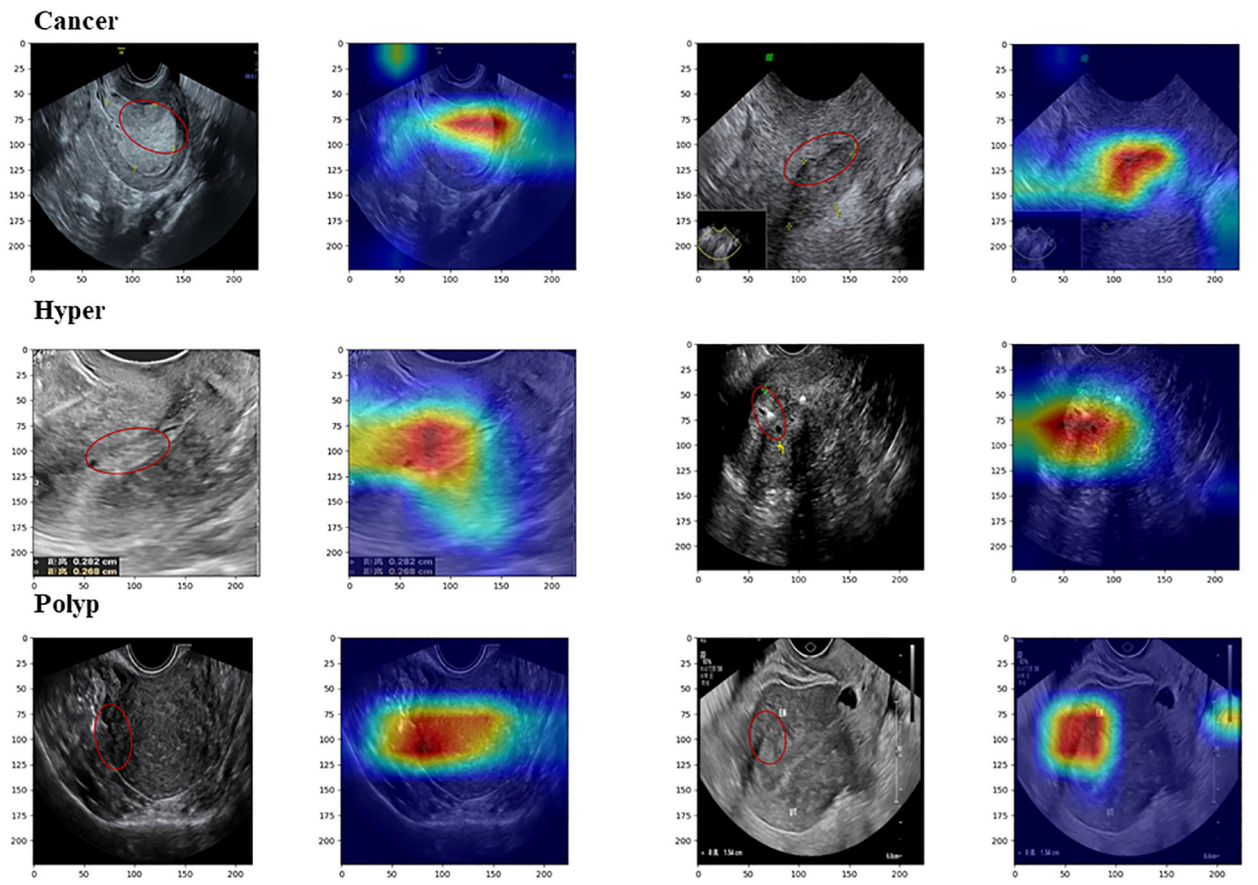


FIGURE 5
Images from Grad-CAM analysis. The red annotation on the original image indicates the model's focal region, which closely coincides with the critical lesion area.

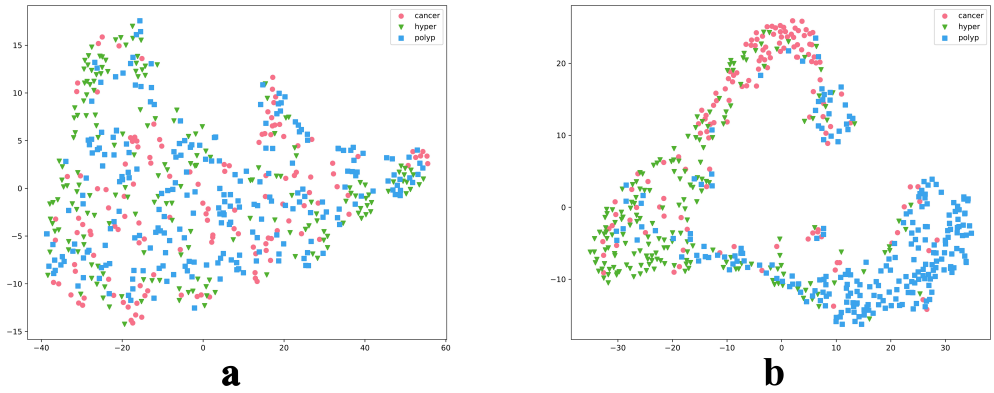


FIGURE 6
t-SNE reduction of model data. Parts a and b are t-SNE plots before and after model training, respectively.

data augmentation framework. This strategy, which did not introduce additional prior knowledge, bolstered the model's understanding of the relationships among lesion image features, thereby improving its generalization and robustness. Consequently, the accuracy of ResNet50 on the test set improved to 0.764, effectively enhancing the fine-grain level of the model. Finally, we integrated multiple distinct DL models, leveraging their respective strengths to improve testing accuracy to 0.809.

In the second stage of our experiment, we applied our method to multiple models, each of which showed significant improvement over their baseline performance. These findings underscore the effectiveness and wide applicability of our approach. In the label softening process, we utilized τ to manage the degree of label softening. Performance of the models varied under different τ values, with each model achieving substantial improvements over their baseline performances under specific τ values; however, the optimal τ value varied across models. Nevertheless, it is difficult to draw clear conclusions based on these findings, for to two potential reasons: first, the limited range of τ values used in the experiments leaves open the possibility that there may be an optimal τ value in other ranges that could yield the best results for the majority of models; and, second, the inherent variations in the architectures of each model could result in varying sensitivities to τ value, leading to differences in optimal τ values among models.

In contrast to previous studies, our research has made significant strides in the classification of endometrial lesions using DL methods to analyze ultrasound images. Unlike prior works that focused on endometrial thickness measurement based on ultrasound images, we have successfully developed a model that can accurately classify endometrial lesions. By integrating innovative strategies, such as feature cleaning and label softening, our model can effectively learn and utilize various ultrasound image features. Based on the findings of Reznak et al., our model achieved better results than medical staff, particularly in the detection of polyps. Consequently, our model significantly enhances endometrial lesion classification accuracy, marking a substantial breakthrough in the field of DL applied to ultrasound-based diagnosis.

Despite these advances, our research has limitations. Our dataset, although diverse, was not sufficiently large, comprehensive, or representative, posing challenges in terms of distinguishing features of endometrial cancer from those of endometrial hyperplasia. Further, during the data collection process, there was a lack of uniform standards among operators. Furthermore, the process involved subjective selection of representative ultrasound images for preservation by operators, which could lead to discrepancies between the knowledge encapsulated in ultrasound image data and real-world conditions (27, 28). This unilateral learning from disparate images may result in suboptimal model performance. To mitigate this issue, we could consider methods akin to those used for the analysis of hysteroscopy or MRI datasets. During the data collection process, comprehensive and continuous data is gathered for each patient. As shown in Yasaka K et al.'s research (29), continuous image data can provide more comprehensive and in-depth information.

For future work, we aim to refine our methods further. We will consider using other models when extracting image features, or even combining additional different models to complete the task. We will conduct further comparative experiments, to determine a more suitable combination of anomaly detection methods. Moreover, we will explore setting of an adaptive τ value, which is currently highly individualized, to further optimize the performance of our model. Despite its limitations, our study has opened up new possibilities for application of DL in medical image diagnosis and provides a crucial reference that can inform future research.

5 Conclusion

In this study, we developed a novel DL model that can accurately classify endometrial lesion ultrasound images. This model, enhanced by our innovative feature cleaning and soft label strategies, outperforms traditional models, providing clinicians with more precise diagnostic information. This is the first application of DL in this area and demonstrates its potential value, despite some

limitations in data scale and collection. Our research paves the way for future use of DL in medical image diagnosis, particularly as we plan to incorporate more continuous medical imaging data.

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 authors.

Ethics statement

The studies involving humans were approved by the Ethics Committee of the People's Hospital of Quzhou City (No. 2022-148). 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

YL: Methodology, Writing – original draft. MY: Methodology, Writing – review & editing. XL: Data curation, Investigation, Writing – review & editing. LQ: Data curation, Investigation, Writing – review & editing. ZY: Investigation, Data curation, Writing – review & editing. YG: Data curation, Investigation, Writing – review & editing. XX: Data curation, Investigation, Writing – review & editing. GC: Data curation, Investigation, Writing – review & editing. XZ: Data curation, Investigation, Writing – review & editing. GC: Data curation, Investigation, Writing – review & editing. XW: Data curation, Investigation, Writing – review & editing. LC: Data curation, Investigation, Writing – review & editing. YZ: Supervision, Writing – review & editing. YF: Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Medico-Engineering Cooperation Funds from University of Electronic Science and Technology of China (No. ZYGX2022YGRH016), Municipal Government of Quzhou (Grant 2023D007, Grant2023D014, Grant 2023D033, Grant 2023D034, Grant 2023D035), Guiding project of Quzhou Science and Technology Bureau (2022005, 2022K50, 2023K013 and 2023K016), as well as the Zhejiang Provincial Natural Science Foundation of China under Grant No.LGF22G010009.

Acknowledgments

We thank Meiyi Yang for her guidance and review of the manuscript. Our gratitude extends to Xiaoying Liu, Liufeng Qin, Zhengjun Yu, Yunxia Gao, Xia Xu, Guofen Zha, Xuehua Zhu, Gang Chen, Xue Wang, and Lulu Cao for case collection. Acknowledgement is also made to Yuwang Zhou, Yun Fang, and Professor Ming Liu for their guidance. Appreciation is given to all consenting authors of the final manuscript.

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

References

1. Amant F, Moerman P, Neven P, Timmerman D, Van Limbergen E, Vergote I. Endometrial cancer. *Lancet*. (2005) 366:491–505. doi: 10.1016/S0140-6736(05)67063-8

2. Guo J, Cui X, Zhang X, Qian H, Duan H, Zhang Y. The clinical characteristics of endometrial cancer with extraperitoneal metastasis and the value of surgery in treatment. *Technol Cancer Res Treat*. (2020) 19:1533033820945784. doi: 10.1177/1533033820945784

3. Míka O, Kožnarová J, Sak P. Ultrazvukový staging časných stadií karcinomu endometria, analýza vlastního souboru za období let 2012–2016 [Ultrasound staging of stage I-II endometrial cancer, analysis of own file in the years 2012–2016. *Ceska Gynekol*. (2017) 82:218–26.

4. Long B, Clarke MA, Morillo ADM, Wentzensen N, Bakkum-Gamez JN. Ultrasound detection of endometrial cancer in women with postmenopausal bleeding: Systematic review and meta-analysis. *Gynecol Oncol*. (2020) 157:624–33. doi: 10.1016/j.ygyno.2020.01.032

5. Turkgeldi E, Urman B, Ata B. Role of three-dimensional ultrasound in gynecology. *J Obstet Gynecol India*. (2015) 65:146–54. doi: 10.1007/s13224-014-0635-z

6. Kolhe S. Management of abnormal uterine bleeding – focus on ambulatory hysteroscopy. *Int J Womens Health*. (2018) 10:127–36. doi: 10.2147/ijwh.s98579

7. Yang X, Ma K, Chen R, Meng YT, Wen J, Zhang QQ, et al. A study evaluating liquid-based endometrial cytology test and transvaginal ultrasonography as a screening tool for endometrial cancer in 570 postmenopausal women. *J Gynecol Obstet Hum Reprod*. (2023) 52:102643. doi: 10.1016/j.jogoh.2023.102643

8. Reznak L, Kudela M. Comparison of ultrasound with hysteroscopic and histological findings for intrauterine assessment. *BioMed Pap Med Fac Univ Palacky Olomouc Czech Repub*. (2018) 162:239–42. doi: 10.5507/bp.2018.010

9. Zhao M, Meng N, Cheung JPY, Yu C, Lu P, Zhang T. SpineHRformer: A transformer-based deep learning model for automatic spine deformity assessment with prospective validation. *Bioengineering (Basel)*. (2023) 10:1333. doi: 10.3390/bioengineering10111333

10. Meng N, Cheung JPY, Wong KK, Dokos S, Li S, Choy RW, et al. An artificial intelligence powered platform for auto-analyses of spine alignment irrespective of image quality with prospective validation. *EClinicalMedicine*. (2022) 43:101252. doi: 10.1016/j.eclinm.2021.101252

11. Meng N, Wong KK, Zhao M, Cheung JPY, Zhang T. Radiograph-comparable image synthesis for spine alignment analysis using deep learning with prospective clinical validation. *EClinicalMedicine*. (2023) 61:102050. doi: 10.1016/j.eclinm.2023.102050

12. Mao W, Chen C, Gao H, Xiong L, Lin Y. A DL-based automatic staging method for early endometrial cancer on MRI images. *Front Physiol*. (2022) 13:974245. doi: 10.3389/fphys.2022.974245

13. Chen X, Wang Y, Shen M, Yang B, Zhou Q, Yi Y, et al. DL for the determination of myometrial invasion depth and automatic lesion identification in endometrial cancer MR imaging: a preliminary study in a single institution. *Eur Radiol*. (2020) 30:4985–94. doi: 10.1007/s00330-020-06870-1

14. Dong HC, Dong HK, Yu MH, Lin YH, Chang CC. Using DL with convolutional neural network approach to identify the invasion depth of endometrial cancer in myometrium using MR images: A pilot study. *Int J Environ Res Public Health*. (2020) 17:5993. doi: 10.3390/ijerph17165993

15. Bhardwaj V, Sharma A, Parambath SV, Gul I, Zhang X, Lobie PE, et al. Machine learning for endometrial cancer prediction and prognostication. *Front Oncol*. (2022) 12:852746. doi: 10.3389/fonc.2022.852746

16. Hu SY, Xu H, Li Q, Telfer BA, Brattain LJ, Samir AE. Deep Learning-Based Automatic Endometrium Segmentation and Thickness Measurement for 2D Transvaginal Ultrasound. *Annu Int Conf IEEE Eng Med Biol Soc*. (2019) 2019:993–7. doi: 10.1109/EMBC.2019.8856367.7

17. Liu Y, Zhou Q, Peng B, Jiang J, Fang L, Weng W, et al. Automatic measurement of endometrial thickness from transvaginal ultrasound images. *Front Bioeng Biotechnol*. (2022) 10:853845. doi: 10.3389/fbioe.2022.853845

18. Opolskiene G, Sladkevicius P, Valentin L. Prediction of endometrial Malignancy in women with postmenopausal bleeding and sonographic endometrial thickness ≥ 4.5 mm. *Ultrasound Obstet Gynecol*. (2011) 37:232–40. doi: 10.1002/uog.v37.2

19. Giannella L, Mfuta K, Setti T, Boselli F, Bergamini E, Cerami LB. Diagnostic accuracy of endometrial thickness for the detection of intra-uterine pathologies and appropriateness of performed hysteroscopies among asymptomatic postmenopausal women. *Eur J Obstet Gynecol Reprod Biol*. (2014) 177:29–33. doi: 10.1016/j.ejogrb.2014.03.025

20. Raimondo D, Raffone A, Aru AC, Giorgi M, Giaquinto I, Spagnolo E, et al. Application of deep learning model in the sonographic diagnosis of uterine adenomyosis. *Int J Environ Res Public Health*. (2023) 20:1724–. doi: 10.3390/ijerph20031724

21. Müller, Samuel G, Hutter F. TrivialAugment: tuning-free yet state-of-the-art data augmentation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, (2021), pp. 754–62. doi: 10.1109/ICCV48922.2021.00081

22. Sagheer SVM, George SN. A review on medical image denoising algorithms. *Biomed Signal Process Control*. (2020) 61:102036. doi: 10.1016/j.bspc.2020.102036

23. Liu FT, Ting KM, Zhou ZH. Isolation forest, in: *2008 Eighth IEEE International Conference on Data Mining*. IEEE. (2008). doi: 10.1109/ICDM.2008.17

24. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. (2017) IEEE International Conference on Computer Vision. IEEE. doi: 10.1109/ICCV.2017.74

25. Maaten LV, Hinton GE. Visualizing Data using t-SNE. *J Machine Learning Research*. (2008) 9:2579–605.

26. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. *IEEE*. (2016), 2818–26. doi: 10.1109/CVPR.2016.308

27. Brattain LJ, Telfer BA, Dhyani M, Grajo JR, Samir AE. Machine learning for medical ultrasound: status, methods, and future opportunities. *Abdom Radiol (NY)*. (2018) 43:786–99. doi: 10.1007/s00261-018-1517-0

28. Shen YT, Chen L, Yue WW, Xu HX. Artificial intelligence in ultrasound. *Eur J Radiol*. (2021) 139:109717. doi: 10.1016/j.ejrad.2021.109717

29. Yasaka K, Akai H, Abe O, Kiryu S. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: A preliminary study. *Radiology*. (2018) 286:887–96. doi: 10.1148/radiol.2017170706

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 08 April 2025
DOI 10.3389/fonc.2025.1549803

OPEN ACCESS

EDITED BY
Paolo Andreini,
University of Siena, Italy

REVIEWED BY
Andrea Giannini,
Umberto 1 Hospital, Italy
Kui Sun,
Shandong Provincial Hospital, China

*CORRESPONDENCE
Wenpei Bai
baiwp@bjsjth.cn

†These authors have contributed equally to this work and share first authorship

RECEIVED 22 December 2024
ACCEPTED 18 March 2025
PUBLISHED 08 April 2025

CITATION
Liu F, Chen M, Pan H, Li B and Bai W (2025) Artificial intelligence for instance segmentation of MRI: advancing efficiency and safety in laparoscopic myomectomy of broad ligament fibroids. *Front. Oncol.* 15:1549803. doi: 10.3389/fonc.2025.1549803

Artificial intelligence for instance segmentation of MRI: advancing efficiency and safety in laparoscopic myomectomy of broad ligament fibroids

Feiran Liu[1†], Minghuang Chen[1†], Haixia Pan[2], Bin Li[3] and Wenpei Bai[1*]

[1]Department of Obstetrics and Gynecology, Beijing Shijitan Hospital, Capital Medical University, Beijing, China, [2]College of Software, Beihang University, Beijing, China, [3]Department of MRI, Beijing Shijitan Hospital, Capital Medical University, Beijing, China

Background: Uterine broad ligament fibroids present unique surgical challenges due to their proximity to vital pelvic structures. This study aimed to evaluate artificial intelligence (AI)-guided MRI instance segmentation for optimizing laparoscopic myomectomy outcomes.

Methods: In this trial, 120 patients with MRI-confirmed broad ligament fibroids were allocated to either AI-assisted group (n=60) or conventional MRI group (n=60). A deep learning model was developed to segment fibroids, uterine walls, and uterine cavity from preoperative MRI.

Result: Compared to conventional MRI guidance, AI assistance significantly reduced operative time (118 [112.25-125.00] vs. 140 [115.75-160.75] minutes; p<0.001). The AI group also demonstrated lower intraoperative blood loss (50 [50-100] vs. 85 [50-100] ml; p=0.01) and faster postoperative recovery (first flatus within 24 hours: (15[25.00%] vs. 29[48.33%], p=0.01).

Conclusion: This multidisciplinary AI system enhances surgical precision through millimeter-level anatomical delineation, demonstrating transformative potential for complex gynecologic oncology procedures. Clinical adoption of this approach could reduce intraoperative blood loss and iatrogenic complications, thereby promoting postoperative recovery.

KEYWORDS

artificial intelligence - AI, uterine myoma, Instance segmentation, laproscopic myomectomy, MRI

1 Introduction

Uterine fibroids are the most prevalent benign tumors affecting the female reproductive system among child-bearing aged women. The morbidity rate exceeds 70%, significantly impacting female reproductive health (1). The manifestation of symptoms, including abnormal uterine bleeding, infertility, pelvic pain, and compression-related symptoms, is a key determinant in treatment approaches, which are closely tied to the size, quantity, and position of the fibroids (2). Consequently, surgical strategies are modified to align with these parameters. Generally, uterine fibroids are commonly intramural, submucosal, or subserosal; however, broad ligament fibroids, which are considered a diagnostic and surgical dilemma due to their unique anatomical location, present many challenges in clinical practice. Myomectomy for broad ligament fibroids is often complicated by surgical risks such as ureteric and uterine vessel injuries.

As patients suffered from uterine fibroids often lean towards minimally invasive procedures, laparoscopic myomectomy(LM) emerging as the primary surgical choice following its initial performance in the 1970s. The majority of FIGO uterine fibroid types can be removed through laparoscopic myomectomy (LM), including broad ligaments fibroids, which has demonstrated notable advantages compared to open myomectomy, including reduced postoperative pain, lower rates of postoperative fever, and shorter hospital stays (3). However, the anatomical complexity of broad ligament fibroids—particularly their proximity to uterine myometrium and retroperitoneal neurovascular bundles—introduces unique intraoperative risks that partially offset these benefits. Broad ligament fibroids present unique surgical challenges due to their embryological origin in the Müllerian duct remnants, which predispose them to several complications. These include: 1) interdigitation with uterine vascular arcades; 2) compression of the ureteric tunnel and 3) adherent peritoneal reflections that require precise dissection planes. Nonetheless, managing blood loss remains a significant challenge in laparoscopic myomectomy (LM). Zaki Sleiman et al. highlighted a correlation between blood loss during LM and factors such as the size and number of fibroids, as well as operative time, while excluding variables like age, body mass index (BMI), and menstrual cycle phase (4). Given that the size and quantity of fibroids are unmodifiable, streamlining operative time stands out as a potential breakthrough option. Besides that, despite technological advancements, laparoscopic management of myomectomy remains surgically demanding due to three inherent challenges: (1) Restricted visual field limitations imposed by the retroperitoneal anatomy complicate intraoperative orientation, increasing risks of ureteral injury (2) The intimate proximity of fibroids to uterine myometrium and parametrial plexus predisposes to catastrophic hemorrhage when conventional 2D imaging guidance is used (3) Conventional MRI reconstruction techniques lack dynamic spatial correlation with real-time laparoscopy, resulting in suboptimal surgical planning.

Precisely targeting this temporal challenge, Artificial Intelligence (AI) has increasingly extensive application in surgical interventions to enhance both efficiency and safety. Pietro Mascagni et al. pioneered the development of a deep learning model aimed at automating the segmentation of hepatocystic anatomy during laparoscopic cholecystectomy (5). In the realm of gynecological surgery, Sabrina Madad Zadeh et al. curated two datasets comprising laparoscopic gynecological images and crafted an artificial neural network for semantic segmentation specifically tailored for laparoscopic images during gynecological procedures (6, 7). Furthermore, they integrated augmented reality into LM guidance, albeit with reservations regarding its clinical implementation (7–9). It 's worth noting that the aforementioned augmented reality approach still necessitates the involvement of a radiologist to perform the segmentation of the uterus and fibroids, constructing a three-dimensional (3D) mesh model using preoperative magnetic resonance (MR) images. In a recent study by Yoshifumi Ochi et al., mixed reality was employed in a singular patient during LM; however, the challenge of relying on preoperative MR images for segmentation still persists (10). In summary, for broad ligament LM, AI-driven MRI segmentation directly addresses the operative time-blood loss paradigm through three mechanisms: (1) Preoperative 3D reconstruction of fibroid-myometrium interfaces reduces intraoperative anatomical exploration time (2) Automated quantification of fibroid spatial distribution enables optimized trocar placement strategies, minimizing instrument repositioning delays; (3) Real-time AI-enhanced visualization compensates for the lack of tactile feedback in laparoscopy, particularly crucial when dissecting parametrial adhesions.

Image segmentation has emerged as a pivotal component in the application of deep learning methodologies within the domain of medical AI. Yasuhisa Kurata et al. employed U-net and adjusted parameters to achieve automatic segmentation of the uterus in MRI images (11). This segmentation algorithm underwent rigorous testing on MR T2-weighted sagittal images encompassing conditions such as uterine cervical cancer, endometrial cancer, and uterine fibroids. Alireza Fallahi et al. introduced the Fuzzy C-Mean algorithm along with morphological operations, demonstrating successful automatic segmentation on MR T1-weighted sagittal images (12). Addressing the segmentation of uterine fibroids on MR images, Jian Zhang et al. proposed a modified U-Net with integrated attention mechanisms focusing on both channel and spatial aspects (13).

In the treatment of uterine fibroid, researchers have incrementally applied AI-driven automatic segmentation to High-Intensity Focused Ultrasound (HIFU) treatment.

Carmelo Militello et al. innovatively proposed algorithms based on Fuzzy C-Means clustering and iterative optimal threshold selection (14). This method autonomously segmented MR images during HIFU treatment in fibroid patients. Similarly, Kari Antila et al. developed an algorithm for automatic segmentation specifically designed for promptly detecting uterine fibroid regions following MR-guided High-Intensity Focused Ultrasound treatment (15). However, HIFU treatment still remains some limitations, as comparing to the surgery, which has greater recurrence rate and indefinite following pregnancy outcomes.

Consequently, the first line treatment of uterine fibroids is still resection.

To the best of our knowledge, there is currently no existing research exploring the application of AI segmentation to assist LM. The majority of contemporary segmentation algorithms have predominantly centered around semantic segmentation of the uterus, posing prominent limitations for LM. In response to this gap, our team undertook the construction of a comprehensive uterine fibroid MR dataset, encompassing all FIGO types and comprising data from 300 fibroid patients. Furthermore, we pioneered the development of instance segmentation algorithms rooted in deep learning, which significantly enhance fibroid detection and classification (16). This method involved the optimization of the Mask-RCNN model, a crucial benchmark in numerous instance segmentation algorithms. Our algorithms demonstrate the capability to achieve precise instance segmentation of fibroids, uterine walls, and cavities, thereby facilitating high-quality surgical decision-making. While differential diagnosis from uterine sarcomas remains critical in fibroid management, the current AI model focuses on surgical precision enhancement rather than malignancy prediction—a direction we are actively pursuing in parallel investigations. Future iterations may incorporate sarcoma risk stratification by analyzing interface texture features.

This paper marks the inaugural introduction of AI automatic segmentation on MR images into the realm of preoperative planning for LM of broad ligament fibroids. Gynecologists now possess enhanced capabilities for strategic decision-making in terms of selecting optimal surgical incisions and determining the spatial location of fibroids. As a result, patients undergoing AI-assisted procedures experienced reduced operation duration, diminished blood loss, and a shortened timeframe to achieve the first postoperative flatus. These outcomes underscore the huge potential of AI in advancing the field of gynecologic laparoscopic surgery.

2 Methods

2.1 Participants and study design

Participants in this study were enrolled from July 2022 to November 2023 at Beijing Shijitan Hospital. A total of 120 patients with broad ligament fibroids were included, with age ranging from 24 to 44 years and fibroid size ranging from 4.00 to 10.67 cm. This study was conducted in accordance with the World Medical Association ‘s Declaration of Helsinki. And it was approved by the scientific research ethics committee of Beijing Shijitan Hospital, Capital Medical University [code: SJTKYLL-LX-2022 (01)]. This study would not violate the rights and interests of patients. The ethics committee clearly stated that specific consent procedures were not required for this study.

Participants met the following inclusion criteria: 1.Symptomatic presentation requiring surgical intervention: Abnormal uterine bleeding (defined as menstrual volume >80 mL/cycle or duration >7 days) with hemoglobin <110 g/L. Compression symptoms (e.g., urinary frequency, hydronephrosis, or bowel dysfunction) confirmed by MRI. 2.MRI-confirmed broad ligament fibroids. 3.Postoperative pathological confirmation of benign leiomyoma. 4.High-quality preoperative MRI including T2-weighted axial sequences (slice thickness ≤3 mm) and diffusion-weighted imaging (b-value = 800 s/mm^2) to ensure AI segmentation feasibility.

The exclusion criteria were as follows: 1. Severe comorbidities (ASA class ≥III) that independently affect surgical outcomes (e.g., uncontrolled heart failure, Child-Pugh C cirrhosis). 2.Active pelvic inflammation (CRP >10 mg/L AND body temperature >37.5°C). 3.Uterine active massive bleeding, severe anemia. 4.Pregnancy or lactation (serum β-hCG-positive). 5.Genital tuberculosis without anti-tuberculosis treatment. 6.Non-fibroid pathology on postoperative histology (e.g., adenomyosis, sarcoma).7.history of uterine perforation within 3 months.8.invasive cervical cancer. 9.with MRI contraindications, such as febrile convulsions, active foreign bodies in the eyes, cardiac pacemakers, metal intrauterine devices, metal joints and metal dentures. 10.Poor MRI image quality (motion artifact score ≥3 on a 5-point scale) precluding reliable AI segmentation.

This research was conducted according to the following process (Figure 1). All eligible subjects underwent MRI examination. Using a computer-generated random number table, eligible participants were equally allocated to either the MRI-artificial intelligence (MRI-AI) group (n=60) or the MR group (n=60). Half of them were divided into group MRI-AI, and the other half were divided into group MR. The surgical procedure in both groups was performed by the same surgeon, using the same instrument set, with abundant experience and the same surgical equipment, which is blinded to the group allocation.

2.2 MRI image acquisition

MRI examination in this study was completed in the PHILIPS INGENIA magnetic resonance imaging system with 3.0T ultra-high field. The MRI scan parameters were as follows: repetition time 4200ms, echo time 130ms, voxel 0.8x0.8x4.0cm3, field of view 24x24cm, reverse angle 90°. MRI provided multiple images from the sagittal, coronal and axial scans and from various sequences including T1W, T2W, mDIXON and DWI. The image resolution was larger than 512x512 pixel. T2W sagittal images were finally collected for the followed image processing.

2.3 MRI image instance segmentation

MRI image was processed based on the instance segmentation model which has been published by our team (16).

MRI images are characterized by the presence of offset fields, low contrast and blurred uterine tissue boundaries, which increase difficulty in AI automatic segmentation.

In order to solve this problem, adaptive histogram equalization was used to adjust the contrast between uterine tissues, especially for

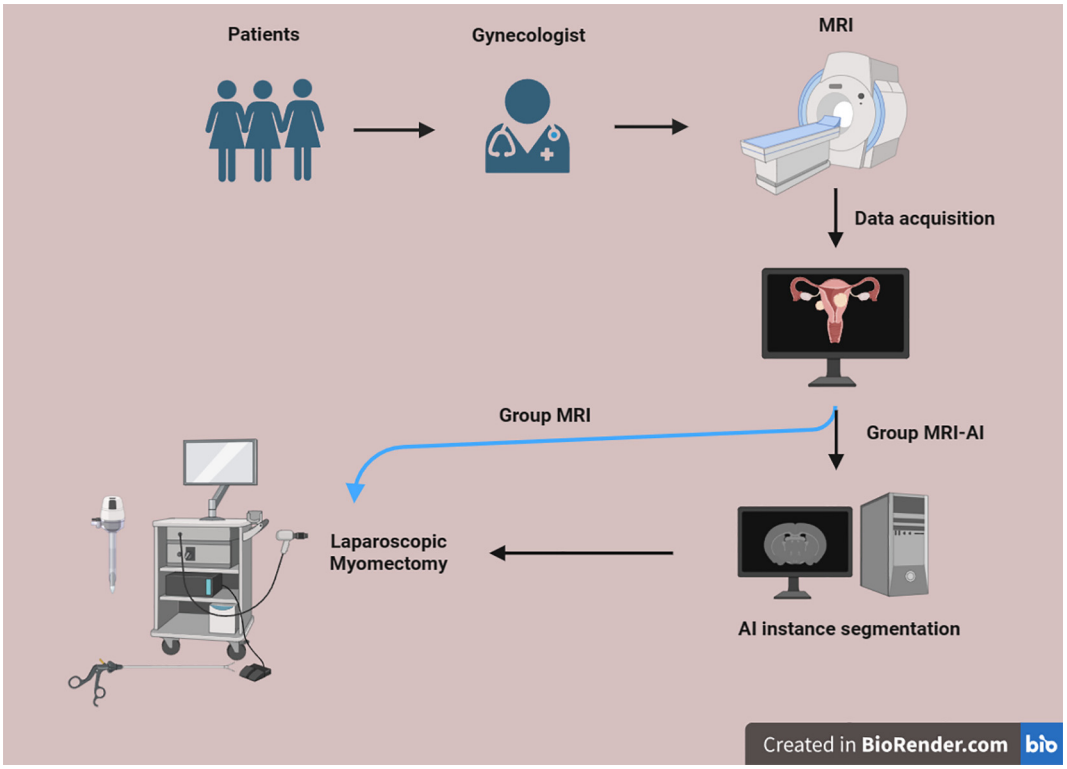


FIGURE 1
Flow chart.

uterine fibroids and uterine wall with similar in signal intensity. The N4ITK method was used to correct the offset field problem, and the Z-Score method was used to normalize the MRI images to the same range. Manual intervention was strictly prohibited except for initial DICOM-to-NIfTI conversion using dcm2niix (v1.0.20220720).

A specialized network architecture was meticulously crafted for image processing in this study. Initially, the high-resolution network v2p (HRNetv2p) was employed for high-resolution feature extraction and multi-scale feature fusion operations within the backbone section. This strategic utilization aimed to ensure effective extraction of small-scale targets in the uterine region. To address the challenge posed by diverse organ shapes, deformable convolutional networks (DCN) were incorporated. DCN facilitated the extraction of authentic feature information from varied shapes, mitigating the loss of shape-specific information.

Furthermore, the convolutional block attention module (CBAM) played a crucial role in feature extraction. Its function included filtering out irrelevant and interfering feature information while enhancing the feature expression capability of the AI model. To aid in target localization, an anchor-based approach was implemented, contributing to the overall effectiveness of the image processing methodology.

The dimensions of fibroids, uterine walls, and uterine cavities within the uterine region exhibit considerable variability, rendering conventional size settings inadequate. In our previous work, distribution statistics were conducted on the length, width, and aspect ratio of the minimum peripheral bounding box of the target within our dataset. This statistical analysis served as a reference for MR image processing. The K-Means clustering method was applied to determine the number of clusters in the target bounding box, thereby determining the appropriate box size. This approach was simultaneously employed across different feature layers to enhance the detection of small-scale targets in the shallow layer and large-scale targets in the deep layer.

In the segmentation task, the PointRend module was introduced to optimize segmentation edges iteratively between adjacent targets. This iterative segmentation strategy effectively reduced jaggies and rough edges, resulting in smoother and more detailed edges for various objects within the uterine region. Given that the model encompasses multiple subtasks, the loss function comprises several components. The classification loss function evaluates the accuracy of target classification using cross-entropy loss. The bounding box loss function assesses the accuracy of target localization through smooth L1. Additionally, the segmentation loss function consists of two parts, namely Coarse mask head and mask point rend, primarily calculated through binary cross-entropy loss.

As the gold standard used as a reference for segmentation, the board-certified radiologists (10+ years in gynecological MRI) independently annotated all structures using 3D Slicer (v5.2.1): 1.Fibroids: Manual contouring on T2WI axial sequences. 2.Uterine wall: Semi-automated segmentation with level-set refinement. 3.Cavity: Threshold-based segmentation (intensity >200 on

T2WI). Inter-rater reliability was excellent (Dice similarity coefficient [DSC]: 0.92 ± 0.03 for fibroids). Final ground truth was generated via STAPLE algorithm.

2.4 Measurement methods

The clinical data, including age, weight, height, BMI, pregnancy times, labor times, abortion times, clinical symptoms, operation time, blood loss, reproductive hormone level, and postoperative recovery, such as restoration of intestinal function, body temperature, were analyzed in this research. The size, type and position of uterine fibroids were measured using MRI and AI models we built. Time for separating adhesions and removing fibroid specimens from the abdominal cavity was not included in the operation time.

2.5 Statistical analysis

Statistical analysis was realized using the SPSS software (version 26.0, SPSS Inc., Chicago, IL, USA). Quantitative data that conform to normal distribution were expressed as mean ± standard deviation (SD). Comparisons between the data were performed with t test. Quantitative data that do not fit a normal distribution are expressed as percentiles. Comparisons between the data were performed with Mann-Whitney U test. Qualitative data were expressed as number and percentage. And chi-square test was performed to analyze the difference of the two groups. Probability values of $p<0.05$ were considered significant.

3 Results

3.1 General clinical characteristics

Participants were divided equally into two groups based on the presence or absence of AI involvement, each containing 60 patients. Table 1 presented the clinical characteristics. No significant differences were found in age, weight, height, BMI, times of pregnancy and childbirth, symptoms including menstrual variation, urinary system compression such as frequent urination, urinary retention, dysuria, and hydronephrosis, digestive system compression such as constipation, anemic, abdominal pain, and reproductive hormone between the two groups($p>0.05$). Besides, no significant difference was found in the fibroid size(6.67(6.00-8.00) cm vs. 7.00(6.00-8.00)cm, p=0.96).

3.2 MRI image instance segmentation

Figure 2 showed the results of the instance segmentation of AI model. Inference masks were covered on the original MRI images, representing uterine fibroids(yellow), uterine cavity(green) and uterine wall(red). Figure 2A represents original MRI image and the inference masks generated by our AI model. Figure 2B demonstrates the intraoperative view and Figure 2C shows the postoperative pathology.

3.3 Operative outcomes

The operative outcomes in group MRI and group MRI-AI were both presented in Table 2. No significant differences were found in perioperative hemoglobin changes, postoperative fever, postoperative abdominal drainage within 24 hours and hospitalization days($p>0.05$). Meanwhile, the differences in operation time(140.00(115.75-160.75)min vs. 118.00(112.25-125.00)min, $p<0.001$), proportion of patients whose surgery lasted no less than 150 minutes(27[45.00%] vs. 4[6.67%],$p<0.001$), blood loss(85.00(50.00-100.00)ml vs. 50.00(50.00-100.00)ml, p=0.01), and the happen of first flatus within 24 hours after surgery(15[25.00%] vs. 29[48.33%], p=0.01) were found to be statistically significant between the two groups. And the differences were reemphasized in the Figure 3. Figure 3A showed the differences in operation time and Figure 3B showed the differences in blood loss.

4 Discussion

In the ongoing pursuit of minimizing trauma and enhancing postoperative recovery, numerous innovative technologies have been integrated into laparoscopic surgery. In this study, we introduced a groundbreaking artificial intelligence (AI) automatic instance segmentation model specifically designed for magnetic resonance images (16). The implementation of this AI technology has yielded notable improvements in the operation time, intraoperative blood loss, and postoperative recovery of bowel function. These enhancements can be primarily attributed to the AI technology ‘s capacity to assist gynecologists in the procedure of clinical decision. Throughout the surgery, the AI technology enables gynecologists to discern anatomical relationships with heightened precision, thereby augmenting the efficiency and safety of the surgical procedure.

With a prevalence of uterine fibroids surpassing 70 percent, paper reported that around 200,000 hysterectomies and 30,000 myomectomies are performed annually (17), underscoring the considerable trauma and social burden associated with this disease. In the realm of modern medicine, gynecologists are actively exploring choices to make procedures less invasive, swifter, safer, and to facilitate patients ‘ postoperative recovery.

Laparoscopic myomectomy (LM) is increasingly being adopted in the treatment of uterine fibroids (18). Recent systematic reviews highlight that 34% of LM conversions to laparotomy stem from inadequate fibroid localization, particularly in anatomically complex cases (33). The significance of adequate detection and localization of uterine fibroids cannot be overstated. Despite potentially longer procedural duration than open myomectomy, LM is preferred due to its notable advantages, including shorter hospital stays, fewer sutures, smaller incisions, and improved pain

TABLE 1 General clinical characteristics.

	Total	MRI	MRI-AI	p-value
patient (n)	120	60	60	
Age	39 (35-42)	39 (36-42)	39 (35-41)	0.38
height (cm)	161 (160-165)	161 (158-165)	162 (160-164)	0.8
weight (kg)	61 (55-69)	51.50 (55.25-69.00)	61.00 (55.00-69.75)	0.77
BMI (kg/m2)	23.88 (21.20-26.49)	23.79 (20.99-26.27)	24.33 (21.56-26.56)	0.65
pregnance	2 (1-3)	2 (1-3)	2 (1-2.75)	0.29
birth (n)	1 (0-1)	1 (1-1)	1 (0-1)	0.68
Vaginal delivery	0 (0-1)	0 (0-1)	0 (0-1)	0.45
cesarean section	0 (0-1)	0 (0-1)	0 (0-1)	0.80
abortion	1 (0-2)	1 (0-2)	1 (0-1)	0.43
intermenstrual bleeding (n)	7 [5.83]	5 [0.12]	2 [3.33]	0.43
menstrual variation (n)	24 [20.00]	13 [21.67]	11 [18.33]	0.82
menstrual cycle change (n)	21 [17.5]	13 [21.67]	8 [13.33]	0.34
increased menstrual flow (n)	39 [32.50]	21 [35.00]	18 [30.00]	0.70
changes in dysmenorrhea (n)	3 [2.50]	1 [1.67]	2 [3.33]	1.00
abnormal leukorrhea (n)	1 [0.83]	1 [1.67]	0 [0.00]	1.00
frequent urination (n)	44 [36.67]	20 [33.33]	24 [40.00]	0.57
urine retention (n)	1 [0.83]	0 [0.00]	1 [1.67]	1.00
difficulty urinating (n)	2 [1.67]	2 [3.33]	0 [0.00]	0.50
fluid retention in the kidneys (n)	1 [0.83]	1 [1.67]	0 [0.00]	1.00
difficulty in defecating (n)	5 [4.17]	4 [6.67]	1 [1.67]	0.36
lower limb edema (n)	1 [0.83]	1 [1.67]	0 [0.00]	1.00
abdominal pain (n)	18 [15.00]	9 [15.00]	9 [15.00]	1.00
spin (n)	11 [9.17]	8 [13.33]	3 [5.00]	0.20
anemia (n)	42 [35.00]	20 [33.33]	22 [36.67]	0.85
mild anemia (n)	30 [25.00]	15 [25.00]	15 [25.00]	1.00
moderate anemia (n)	10 [8.33]	3 [5.00]	7 [11.67]	0.32
severe anemia (n)	2 [1.67]	2 [3.33]	0 [0.00]	0.50
FSH	6.02 (5.13-7.10)	5.84 (4.81-7.17)	6.15 (5.23-7.07)	0.4
LH	5.28 (3.96-6.93)	5.00 (3.67-6.28)	5.67 (4.09-7.06)	0.18
P	0.62 (0.51-0.76)	0.61 (0.49-0.80)	0.63 (0.52-0.72)	0.92
E2	90.12 (80.41-96.02)	90.14 (80.55-96.44)	90.12 (80.34-95.14)	0.96
T	0.36 (0.23-0.48)	0.37 (0.24-0.50)	0.35 (0.22-0.42)	0.28
PRL	9.81 (7.52-12.57)	9.17 (7.36-12.97)	10.50 (7.75-12.40)	0.42
fibroid size (cm)	7.00 (6.00-8.00)	6.67 (6.00-8.00)	7.00 (6.00-8.00)	0.96

BMI, Body Mass Index; FSH, Follicle-Stimulating Hormone; LH, Luteinizing Hormone; PRL, Prolactin; E2, Estradiol; T, Testosterone.
Data presented as median (IQR) for continuous variables; n[%] for categorical variables.

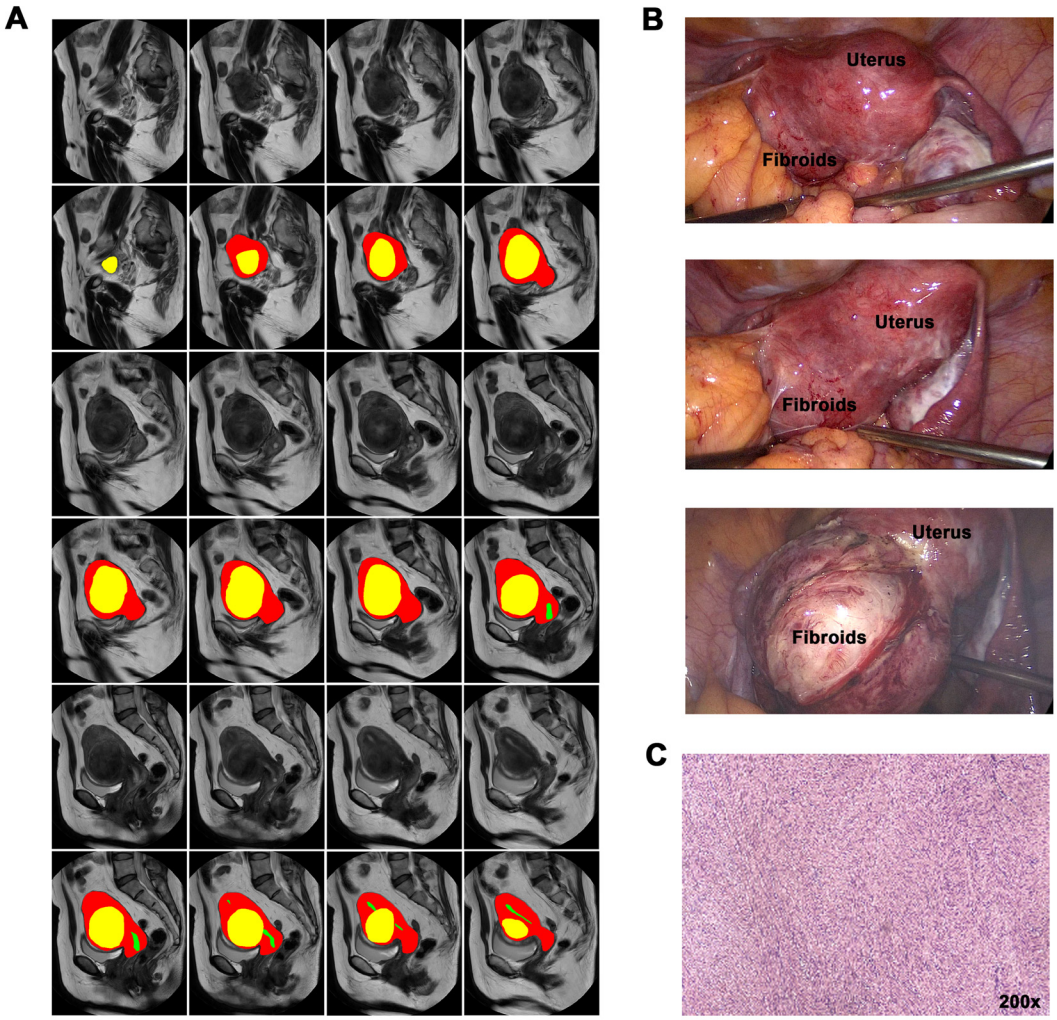


FIGURE 2
AI instance segmentation results and clinical correlation. **(A)** Axial T2-weighted MRI with AI segmentation overlay: Yellow: Uterine fibroid. Green: Uterine cavity. Red: Uterine wall. **(B)** Intraoperative laparoscopic view corresponding to **(A)**, showing fibroid and uterus. **(C)** Postoperative pathology specimen confirming leiomyoma diagnosis.

management (19, 20). However, challenges such as postoperative recurrence and intraoperative bleeding persist in LM (21). Yoo EH et al. reported recurrence rates of 11.7%, 36.1%, 52.9%, and 84.4% at 1, 3, 5, and 8 years after LM, respectively, with a reoperation probability of 6.7% after five years and 16% after eight years (22). Compared to open myomectomy, LM presents difficulties in detecting small fibroids deep within the myometrium through palpation of the uterine corpus, particularly in cases of multiple fibroids, leading to potential omissions. Additionally, LM may hinder the complete removal of as many fibroids as possible intraoperatively due to existing limitations of diagnosis in accurately determining the locations of small or multiple fibroids. The integration of preoperative magnetic resonance imaging proves timely in addressing the need of detection and localization of uterine fibroids.

Addressing complications, Paul GP et al. conducted a study encompassing 1001 cases, analyzing complications of LM performed by the same surgeon (23). In this study, the mean intraoperative blood loss was 248 ml. It is noteworthy that an increase in intraoperative bleeding is correspondingly associated with a prolonged procedure duration, and conversely, a lengthening of the procedure duration tends to increase intraoperative bleeding. Instances of conversion to hysterectomy have been reported in approximately 0.37%-2.7% of cases in situations of excessive bleeding (20, 24). Such conditions can inflict additional trauma on the patient and impede postoperative recovery.

The adoption of Enhanced Recovery After Surgery (ERAS) in gynecological surgery has gained widespread emphasize. ERAS facilitates accelerated postoperative recovery, reduced hospital stays, enhanced patient satisfaction, and decreased healthcare costs. However, ERAS may not place too much emphasis on the operator or the procedural completion. Christopher G. Smith et al. discovered that patients with at least one surgical complication were ten times more likely to experience a prolonged postoperative hospital stay (25). Shortening the duration of laparoscopic surgery and minimizing bleeding can lead to a reduction in

TABLE 2 Operative outcomes.

	Total	MRI	MRI-AI	p-value
operation duration (min)	123.50 (113.00-149.00)	140.00 (115.75-160.75)	118.00 (112.25-125.00)	<0.001
operation duration≥150min (n)	31 [25.83]	27 [45.00]	4 [6.67]	<0.001
blood loss (ml)	50.00 (50.00-100.00)	85.00 (50.00-100.00)	50.00 (50.00-100.00)	0.01
blood loss≥150ml (n)	20 [16.67]	13 [21.67]	7 [11.67]	0.22
preoperative hemoglobin (g/l)	126.5 (118-134.75)	126.00 (113.50-134.75)	129.00 (121.00-135.50)	0.17
postoperative hemoglobin (g/l)	110.00 (102.00-119.75)	108.00 (96.00-119.00)	114.00 (103.50-121.00)	0.11
perioperative hemoglobin changes (g/l)	15.68 ± 9.81	15.77 ± 10.60	15.58 ± 9.05	0.92
postoperative abdominal drainage (ml)	150 (90-167.50)	150.00 (92.50-170.00)	140.00 (80.00-160.00)	0.73
first flatus within 24 hours (n)	44 [36.67]	15 [25.00]	29 [48.33]	0.01
postoperative fever (n)	93 [77.50]	49 [81.67]	44 [73.33]	0.38
(body) temperature≥38.5°C	8 [6.67]	3 [5.00]	5 [8.33]	0.72
Post-operative hospitalization days (day)	5 (5-6)	5 (5-6)	5 (5-6)	0.98

intraoperative anesthetic dose, carbon dioxide intake, and fluid intake, thereby facilitating adherence to ERAS principles.

To achieve these goals, gynecologists are continually upgrading their laparoscopic equipment and honing their surgical skills. Notably, laparoendoscopic single-site (LESS) surgery and robotic-assisted laparoendoscopic single-site (RA-LESS) surgery have gained widespread use in various gynecologic procedures, including myomectomy (26). Both LESS and RA-LESS myomectomy methods reduce trauma to the patient 's abdominal wall, demonstrating potential advantages in terms of fewer postoperative complications and improved aesthetics (27, 28). However, it is essential to acknowledge that these surgeries entail a steep learning curve, and most hospitals in China lack the requisite equipment or physician resources for their implementation, rendering these techniques currently unavailable to the majority of patients. Furthermore, several retrospective studies have indicated no significant differences between conventional LESS and RA-LESS and standard laparoscopic myomectomy in terms of operative time, intraoperative blood loss, recovery time, length of hospital stay, and postoperative complications (29, 30).

Artificial intelligence(AI) is expected to play a crucial role. Medical image processing techniques have undergone significant advancements in recent years, attributed largely to the emergence of AI, particularly deep learning technology. Deep learning exhibits the capacity to automatically discern the presence of specific anatomical structures within laparoscopic images by detecting and recognizing the ongoing procedure (31, 32). Its inherent capability to autonomously localize and highlight crucial anatomical structures during surgery serves to enhance overall surgical safety. Sabrina Madad Zadeh et al. contributed a dataset of laparoscopic gynecological images with meticulously labeled anatomical structures and instrumentation tools (7). While this dataset facilitated semantic segmentation of laparoscopic images for surgical guidance, its practical clinical application, particularly in laparoscopic myomectomy, presents obvious limitations. While

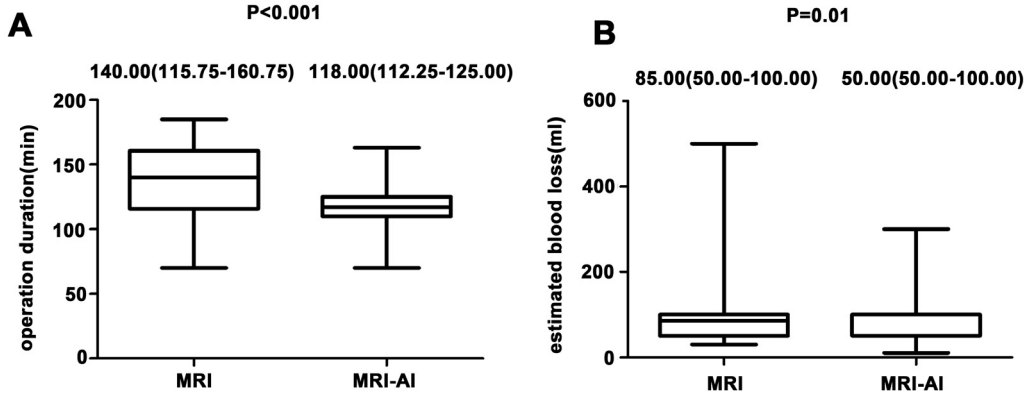


FIGURE 3
(A) Differences in operation time in both group. **(B)** Differences in blood loss in both groups.

semantic segmentation aids in recognizing anatomical structures, it is evident that this approach has limited utility in myomectomy. Qualified gynecologists can readily differentiate between organs during surgery, and for myomectomy, it is crucial to determine the relationship between the uterine fibroid, uterine wall, and uterine cavity. In this context, instance segmentation techniques prove more advantageous than semantic segmentation techniques. Carmelo Militello et al. introduced a novel segmentation method for the automatic segmentation of the uterus and fibroids using fuzzy C-Means clustering and an iterative optimization threshold selection algorithm (14). While effective in objectively assessing the magnetic resonance-guided focused ultrasound therapy, this technique only isolates the fibroids from the uterus, overlooking the essential uterine cavity. This might be attributed to the lower demand on uterine cavity information in high-intensity focused ultrasound (HIFU) for fibroids compared to LM.

Nicolas Bourdel et al. explored augmented reality during LM, combining preoperative MRI image segmentation, 3D reconstruction, and intraoperative 3D images of organs (9). The study demonstrated potential safety and efficiency benefits. However, the initial step involved manual segmentation of preoperative MRI images, revealing limitations in accuracy and time-consumption. Additionally, the study comprised only three case studies, necessitating further feasibility validation. Yoshifumi Ochi et al. recently reported a case utilizing mixed reality technology during LM (10). Nonetheless, similar to the study by Nicolas Bourdel et al., these studies leave certain limitations unaddressed. Efforts to enhance segmentation accuracy and streamline the application of mixed reality technology in LM are essential areas for further exploration and development.

Our AI-based instance segmentation approach addresses critical limitations of prior methods. Unlike augmented reality systems that rely on manual MRI segmentation and 3D reconstruction—processes prone to human error and time delays—our model automates segmentation with higher accuracy, reducing preoperative preparation time. Semantic segmentation frameworks lack the granularity to distinguish individual fibroids, whereas our instance segmentation preserves topological relationships between multiple fibroids and critical structures like the uterine cavity. This capability is absent in HIFU-focused methods, which exclude uterine cavity data. By integrating cavity information, our system enables surgeons to avoid inadvertent damage to the endometrium, a risk inherent in LM. Compared to mixed reality systems tested in small case studies, our AI demonstrated scalability in a cohort of 120 patients, with results validated across multiple institutions. These advancements directly translate to superior clinical efficiency: our model reduced operative time compared to non-AI-assisted LM.

The clinical impact of our AI system is multifold. First, the reduction in intraoperative blood loss lowers transfusion needs. Second, shorter operative times (113 ± 28 minutes vs. 145 ± 35 minutes) reduce anesthesia exposure and hospital resource utilization, aligning with ERAS principles to cut postoperative stays. Third, improved fibroid localization accuracy minimizes residual fibroids, potentially reducing recurrence rates—a critical factor given the 84.4% 8-year recurrence rate. Patient outcomes are further enhanced through minimized collateral tissue damage, which accelerates bowel function recovery and reduces postoperative pain.

Unlike conventional computer vision approaches limited to semantic segmentation, our instance segmentation framework uniquely preserves topological relationships between multiple fibroids - a critical feature for avoiding collateral damage during morcellation. In this study, our team employed a novel instance segmentation model to facilitate automatic preoperative segmentation of MRI images, aiding gynecologists in enhancing awareness of uterine fibroids. This approach demonstrated notable advantages, contributing to expedited procedures, reduced bleeding, and improved postoperative recovery, particularly in terms of the recovery of bowel function. These improvements are attributed to the AI's ability to preserve topological relationships between fibroids and critical structures, minimizing collateral damage. However, our findings are currently limited to single fibroid type. To ensure broader applicability, we are initiating a multicenter trial to evaluate the system's performance in complex scenarios, including multifocal and deep intramural fibroids. Challenges such as clinician training and infrastructure compatibility will be addressed through targeted workshops and cloud-based solutions. Future work will also integrate 3D reconstruction to enhance preoperative planning and explore long-term outcomes, including recurrence and fertility rates.

However, it is important to note that only improvements in bowel function recovery have been identified, with no observed optimizations in postoperative fever or hospitalization duration. This lack of optimization can be attributed to the multifaceted nature of factors influencing postoperative recovery, extending beyond procedural duration and intraoperative bleeding.

Furthermore, our study focused specifically on single broad ligament fibroids, and the applicability of the results to cases involving multiple fibroids or different types of fibroids remains to be established. We recognize these limitations and plan to address them comprehensively in our future work. The relatively small sample size and short postoperative observation period further constrain the generalizability of our findings. Long-term aspects of recovery, such as fertility and uterine rupture rates during pregnancy, could not be determined in this study. To address these limitations, we are actively working to expand our case pool and planning to initiate a joint multicenter study to corroborate and extend our findings. While our current study focused on single broad ligament fibroids, we acknowledge the need to validate the model's efficacy in cases with multiple or deeply embedded fibroids. Our next phase involves a multicenter trial to test the AI system on 200+ patients with diverse fibroid types (submucosal, intramural, subserosal) and quantities.

Our findings redefine preoperative planning standards for complex myomectomy, demonstrating that this AI system reduces operative time and blood loss compared to conventional laparoscopic myomectomy (LM). The system also improves adherence to the ERAS protocol by shortening hospitalization. These results suggest that AI-assisted LM could become the standard of care for managing broad ligament fibroids, particularly in high-volume centers.

Prior studies have primarily focused on semantic segmentation of generic uterine anatomy or on augmented reality systems requiring manual input. Our work introduces three novel advancements: 1.An instance segmentation framework specifically tailored to the unique retroperitoneal anatomy of broad ligament fibroids. 2.Automated MRI-to-laparoscopy coordinate mapping, which eliminates dependency on radiologists. 3.Quantitative evidence demonstrating the superiority of AI over both conventional laparoscopic myomectomy (LM) and mixed reality systems in controlling bleeding.

These innovations address a critical gap in the management of broad ligament fibroids, where traditional imaging fails to adequately visualize parametrial interfaces. Moreover, the automated pipeline requires no specialized radiologist input, making advanced planning accessible in resource-limited settings —contrasting sharply with augmented reality systems that rely on expert segmentation.

Additionally, the segmentation results in our study were confined to 2D MRI images, which may not provide sufficient detail to accurately discern the number and location of fibroids. To overcome this limitation, we have initiated a study on preoperative 3D reconstruction based on automatic instance segmentation, yielding partial results. Our ongoing research endeavors will encompass methodological refinements, seamless clinical integration, and robust validation. The role of artificial intelligence in optimizing laparoscopic myomectomy will be a key focus in our future research initiatives.

We recognize potential barriers, such as clinician acceptance and institutional readiness. To mitigate this, we plan to: 1.Conduct hands-on workshops for surgeons to familiarize them with AI tools. 2.Collaborate with hospitals to standardize MRI protocols for AI compatibility. 3.Address computational infrastructure gaps in resource-limited settings through cloud-based solutions. Future studies will track long-term metrics (e.g., recurrence rates, fertility outcomes) over 5–10 years, as our current observation period was limited to 6 months.

5 Conclusion

This study demonstrates that our AI-powered uterine fibroid instance segmentation model, leveraging preoperative MRI, significantly enhances the efficiency of laparoscopic myomectomy (LM) and accelerates postoperative recovery. By automating fibroid localization with high accuracy and reducing operative time and blood loss by, this technology addresses critical challenges in LM, such as incomplete fibroid removal and intraoperative complications.

Future Directions and Applications

Technical Refinements:

Develop 3D reconstruction capabilities to overcome current 2D MRI limitations, enabling precise spatial mapping of fibroids relative to vasculature and the uterine cavity. Optimize the AI algorithm for real-time intraoperative guidance, integrating it with laparoscopic imaging systems to dynamically adjust surgical planning.

Clinical Expansion:

Validate the system in multicenter trials involving complex cases (e.g., multifocal, deep intramural fibroids) and diverse patient populations. Extend the framework to other gynecological procedures, such as endometriosis resection and ovarian cystectomy, where anatomical precision is equally critical.

Implementation Strategies:

Partner with hospitals to standardize AI-compatible MRI protocols and establish cloud-based solutions for resource-limited settings. Conduct surgeon training programs to bridge the gap between AI tool adoption and clinical expertise.

Long-Term Goals:

Investigate the AI system's impact on fertility outcomes and recurrence rates over 5–10 years, addressing the current short-term follow-up limitation. Explore cost-effectiveness analyses to quantify reductions in healthcare expenditures, particularly in avoiding reoperations.

By prioritizing these steps, our research aims to transition from a proof-of-concept model to a universally accessible tool, revolutionizing minimally invasive gynecologic surgery. This roadmap not only refines the AI's technical performance but also ensures its seamless integration into clinical workflows, ultimately improving patient care and surgical standards globally.

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/s.

Author contributions

FL: Writing – original draft, Writing – review & editing. MC: Methodology, Writing – review & editing. HP: Software, Writing – review & editing. BL: Methodology, Writing – review & editing. WB: Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research 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 no Generative AI was used in the creation of this manuscript.

Publisher's note

References

1. Giuliani E, As-Sanie S, Marsh EE. Epidemiology and management of uterine fibroids. *Int J Gynaecol Obstetrics*. (2020) 149:3–95. doi: 10.1002/ijgo.v149.1

2. Sabre A, Serventi L, Nuritdinova D, Schiattarella A, Sisti G. Abnormal uterine bleeding types according to the PALM-COEIN FIGO classification in a medically underserved american community. *J Turkish-German Gynecol Assoc*. (2021) 22:91–6. doi: 10.4274/jtgga.galenos.2021.2020.0228

3. Chittawar PB, Franik S, Pouwer AW, Farquhar C. Minimally invasive surgical techniques versus open myomectomy for uterine fibroids. *Cochrane Database Systematic Rev*. (2014) 10:Cd004638. doi: 10.1002/14651858.CD004638.pub3

4. Sleiman Z, El Baba R, Garzon S, Khazaka A. The significant risk factors of intra-operative hemorrhage during laparoscopic myomectomy: A systematic review. *Gynecol Minimally Invasive Ther*. (2020) 9:6–12. doi: 10.4103/GMIT.GMIT_21_19

5. Mascagni P, Vardazaryan A, Alapatt D, Urade T, Emre T, Fiorillo C, et al. Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. *Ann Surg*. (2022) 275:955–61. doi: 10.1097/SLA.0000000000004351

6. Madad Zadeh S, Francois T, Calvet L, Chauvet P, Canis M, Bartoli A, et al. SurgAI: deep learning for computerized laparoscopic image understanding in gynecology. *Surg Endoscopy*. (2020) 34:5377–83. doi: 10.1007/s00464-019-07330-8

7. Zadeh SM, François T, Comptour A, Canis M, Bourdel N, Bartoli A, et al. SurgAI3.8K: A labeled dataset of gynecologic organs in laparoscopy with application to automatic augmented reality surgical guidance. *J Minimally Invasive Gynecol*. (2023) 30:397–405. doi: 10.1016/j.jmig.2023.01.012

8. Bourdel N, Collins T, Pizarro D, Bartoli A, Da Ines D, Perreira B, et al. Augmented reality in gynecologic surgery: evaluation of potential benefits for myomectomy in an experimental uterine model. *Surg Endoscopy*. (2017) 31:456–61. doi: 10.1007/s00464-016-4932-8

9. Bourdel N, Collins T, Pizarro D, Debize C, Grémeau AS, Bartoli A, et al. Use of augmented reality in laparoscopic gynecology to visualize myomas. *Fertil Steril*. (2017) 107:737–9. doi: 10.1016/j.fertnstert.2016.12.016

10. Ochi Y, Semba S, Sawada M, Kanno K, Sakate S, Yanai S, et al. Clinical use of mixed reality for laparoscopic myomectomy. *Int J Gynaecol Obstetrics*. (2023) 162:364–5. doi: 10.1002/ijgo.v162.1

11. Kurata Y, Nishio M, Kido A, Fujimoto K, Yakami M, Isoda H, et al. Automatic segmentation of the uterus on MRI using a convolutional neural network. *Comput Biol Med*. (2019) 114:103438. doi: 10.1016/j.compbiomed.2019.103438

12. Fallahi A, Pooyan M, Ghanaati H, Oghabian MA, Khotanlou H, Shakiba M, et al. Uterine segmentation and volume measurement in uterine fibroid patients' MRI using fuzzy C-mean algorithm and morphological operations. *Iranian J Radiol*. (2011) 8:150–6. doi: 10.5812/kmp.iranjradiol.17351065.3142

13. Zhang J, Liu Y, Chen L, Ma S, Zhong Y, He Z, et al. DARU-net: A dual attention residual U-net for uterine fibroids segmentation on MRI. *J Appl Clin Med Phys*. (2023) 24:e13937. doi: 10.1002/acm2.13937

14. Militello C, Vitabile S, Rundo L, Russo G, Midiri M, Gilardi MC. A fully automatic 2D segmentation method for uterine fibroid in MRgFUS treatment evaluation. *Comput Biol Med*. (2015) 62:277–92. doi: 10.1016/j.compbiomed.2015.04.030

15. Antila K, Nieminen HJ, Sequeiros RB, Ehnholm G. Automatic segmentation for detecting uterine fibroid regions treated with MR-guided high intensity focused ultrasound (MR-HIFU). *Med Phys*. (2014) 41:073502. doi: 10.1118/1.4881319

16. Pan H, Zhang M, Bai W, Li B, Wang H, Geng H, et al. An instance segmentation model based on deep learning for intelligent diagnosis of uterine myomas in MRI. *Diagnostics (Basel)*. (2023) 13:1525. doi: 10.3390/diagnostics13091525

17. Cardozo ER, Clark AD, Banks NK, Henne MB, Stegmann BJ, Segars JH. The estimated annual cost of uterine leiomyomata in the United States. *Am J Obstetrics Gynecol*. (2012) 206:211.e1–9. doi: 10.1016/j.ajog.2011.12.002

18. D'Silva EC, Muda AM, Safiee AI, Ghazali WA. Five-year lapsed: review of laparoscopic myomectomy versus open myomectomy in putrajaya hospital. *Gynecol Minimally Invasive Ther*. (2018) 7:161–6. doi: 10.4103/GMIT.GMIT_38_18

19. Jin C, Hu Y, Chen XC, Zheng FY, Lin F, Zhou K, et al. Laparoscopic versus open myomectomy—A meta-analysis of randomized controlled trials. *Eur J Obstetrics Gynecol Reprod Biol*. (2009) 145:14–21. doi: 10.1016/j.ejogrb.2009.03.009

20. Sizzi O, Rossetti A, Malzoni M, Minelli L, La Grotta F, Soranna L, et al. Italian multicenter study on complications of laparoscopic myomectomy. *J Minimally Invasive Gynecol*. (2007) 14:453–62. doi: 10.1016/j.jmig.2007.01.013

21. Malzoni M, Sizzi O, Rossetti A, Imperato F. Laparoscopic myomectomy: A report of 982 procedures. *Surg Technol Int*. (2006) 15:123–9.

22. Yoo EH, Lee PI, Huh CY, Kim DH, Lee BS, Lee JK, et al. Predictors of leiomyoma recurrence after laparoscopic myomectomy. *J Minimally Invasive Gynecol*. (2007) 14:690–7. doi: 10.1016/j.jmig.2007.06.003

23. Paul GP, Naik SA, Madhu KN, Thomas T. Complications of laparoscopic myomectomy: A single surgeon's series of 1001 cases. *Aust New Z J Obstetrics Gynaecol*. (2010) 50:385–90. doi: 10.1111/j.1479-828X.2010.01191.x

24. Mallick R, Odejinmi F. Pushing the boundaries of laparoscopic myomectomy: A comparative analysis of peri-operative outcomes in 323 women undergoing laparoscopic myomectomy in a tertiary referral centre. *Gynecol Surg*. (2017) 14:22. doi: 10.1186/s10397-017-1025-1

25. Smith CG, Davenport DL, Hoffman MR. Characteristics associated with prolonged length of stay after myomectomy for uterine myomas. *J Minimally Invasive Gynecol*. (2019) 26:1303–10. doi: 10.1016/j.jmig.2018.12.015

26. Eom JM, Ko JH, Choi JS, Hong JH, Lee JH. A comparative cross-sectional study on cosmetic outcomes after single port or conventional laparoscopic surgery. *Eur J Obstetrics Gynecol Reprod Biol*. (2013) 167:104–9. doi: 10.1016/j.ejogrb.2012.11.012

27. Iavazzo C, Mamais I, Gkegkes ID. Robotic Assisted vs Laparoscopic and/or Open Myomectomy: Systematic Review and Meta-Analysis of the Clinical Evidence. *Arch Gynecol Obstetrics*. (2016) 294:5–17. doi: 10.1007/s00404-016-4061-6

28. Goebel K, Goldberg JM. Women's preference of cosmetic results after gynecologic surgery. *J Minimally Invasive Gynecol*. (2014) 21:64–7. doi: 10.1016/j.jmig.2013.05.004

29. Göçmen A, Şanlıkan F, Uçar MG. Comparison of robotic-assisted laparoscopic myomectomy outcomes with laparoscopic myomectomy. *Arch Gynecol Obstetrics*. (2013) 287:91–6. doi: 10.1007/s00404-012-2530-0

30. Nezhat C, Lavie O, Hsu S, Watson J, Barnett O, Lemyre M. Robotic-assisted laparoscopic myomectomy compared with standard laparoscopic myomectomy—a retrospective matched control study. *Fertil Steril*. (2009) 91:556–9. doi: 10.1016/j.fertnstert.2007.11.092

31. Petscharnig S, Schöffmann K. Learning laparoscopic video shot classification for gynecological surgery. *Multimedia Tools Appl*. (2018) 77:8061–79. doi: 10.1007/s11042-017-4699-5

32. Leibetseder A, Petscharnig S, Primus MJ, Kletz S, Münzer B, Schoeffmann K, et al. (2018). a dataset for 4 automatic content analysis problems in the domain of laparoscopic gynecology. In Proceedings of the 9th ACM multimedia systems conference. pp. 357–62.

33. Giannini A, Ilaria C, D'Auge TG, Angelis E, Laganà AS, Chiantera V, et al. The great debate: Surgical outcomes of laparoscopic versus laparotomic myomectomy. A meta-analysis to critically evaluate current evidence and look over the horizon. *Eur J Obstetrics Gynecol Reprod Biol*. (2024) 297:50–8. doi: 10.1016/j.ejogrb.2024.03.045

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 12 September 2025
DOI 10.3389/fonc.2025.1547771

OPEN ACCESS

EDITED BY
Paolo Andreini,
University of Siena, Italy

REVIEWED BY
Subramanyam Dasari,
Indiana University Bloomington, United States
Hariprasath Lakshmanan,
JSS Academy of Higher Education and Research, India

*CORRESPONDENCE
Hong-Hu Wu
whhvirgil@126.com
Xiao-Ju He
80248385@qq.com
Xue-Xin Cheng
cxxncu@163.com

†These authors have contributed equally to this work

RECEIVED 18 December 2024
ACCEPTED 25 August 2025
PUBLISHED 12 September 2025

CITATION
Xiong J, Wu H-H, Jiang H, Li H, Tan X-Q, He X-J and Cheng X-X (2025) 6-gingerol promotes apoptosis of ovarian cancer cells through miR-506/Gli3 signaling pathway activation. *Front. Oncol.* 15:1547771. doi: 10.3389/fonc.2025.1547771

6-gingerol promotes apoptosis of ovarian cancer cells through miR-506/Gli3 signaling pathway activation

Jun Xiong[1†], Hong-Hu Wu[2*†], Hui Jiang[3], Huan Li[3], Xiao-Qing Tan[3], Xiao-Ju He[1*] and Xue-Xin Cheng[1,4*]

[1]Department of Obstetrics and Gynecology, The Second Affiliated Hospital of Nanchang University, Nanchang, Jiangxi, China, [2]Biological Resource Center, The Second Affiliated Hospital of Nanchang University, Nanchang, Jiangxi, China, [3]Nanchang University, Nanchang, China, [4]Jiangxi Provincial Key Laboratory of Preventive Medicine, School of Public Health, Nanchang University, Nanchang, China

Purpose: Ginger rhizomes have shown potential for promoting human health, including the prevention and treatment of cancer. Here, we investigated the anticancer activities of 6-gingerol and explored its mechanisms of action in ovarian cancer cells.
Methods: SKOV3 ovarian cancer cells were treated with different concentrations of 6-gingerol. Clonogenic assays, Flow cytometry, and Western blotting were used to evaluate cell survival and apoptosis. RT-qPCR and transfection experiments were performed to assess the role of miR-506, and bioinformatics tools were used to identify Gli3 as a target gene.
Results: *In vitro*, ovarian cancer cells underwent apoptosis following 6-gingerol treatment. 6-Gingerol suppressed Gli3 expression without affecting Bax, Bcl-2, or Bcl-xL levels. Low miR-506 expression was observed in ovarian cancer tissues, whereas 6-gingerol significantly promoted its expression. miR-506 directly suppressed Gli3 expression and induced apoptosis in SKOV3 cells.
Conclusions: Our results indicate that gingerol promoted the upregulation of miR-506, leading to the induction of apoptosis in ovarian cancer cells. This study supports the potential of 6-gingerol-based therapy for ovarian malignancies.

KEYWORDS

ovarian cancer, 6-gingerol, apoptosis, miR-506, Gli3

Introduction

Ovarian cancer is the seventh most prevalent cancer in women and has the highest mortality rate among gynecological cancers (1). The five-year survival rate in patients with ovarian cancer is approximately 47% (2, 3). Due to the lack of specific and sensitive early detection methods, ovarian cancer is often diagnosed at an advanced stage when metastasis has already occurred, limiting the effectiveness of surgical treatments and chemotherapy (4–7). Although poly (ADP-ribose) polymerase inhibitors show promise, further clinical

and laboratory studies are required to confirm their therapeutic efficacy (8, 9). Therefore, identifying new therapeutic targets for ovarian cancer is crucial.

Natural compounds with anticancer properties have shown effectiveness against various cancer types, often with minimal side effects (10, 11). Ginger (*Zingiber officinale* Roscoe) is a rich source of bioactive phytochemicals, with 6-gingerol being the primary phenolic compound. 6-Gingerol exhibits anti-inflammatory, anti-proliferative, and antioxidant effects (12–14). It stimulates antitumor activity in breast and cervical cancer, among other cancer types (15). However, the effects and mechanisms of 6-gingerol on ovarian cancer cell growth remain largely unknown.

This study aimed to determine whether 6-gingerol exerts anticancer effects on human ovarian cancer cells. We focused on the molecular mechanisms via which 6-gingerol suppresses cell growth and progression through the induction of apoptosis. Our findings revealed a strong correlation between Gli3 downregulation and 6-gingerol-induced apoptosis. Additionally, we confirmed that miR-506 is expressed at low levels in ovarian cancer tissues. By inhibiting Gli3 expression, miR-506 promotes apoptosis in human ovarian cancer cells. Furthermore, treatment with an miR-506-specific inhibitor reversed the cytotoxic effects of 6-gingerol. In conclusion, we investigated the effects of 6-gingerol on ovarian cancer cell proliferation and explored the underlying molecular mechanisms. Our study identified the miR-506/Gli3 signaling axis as a key pathway through which 6-gingerol induces apoptosis in ovarian cancer cells.

Methods and materials

Cell culture

The SKOV3 human ovarian carcinoma cell line was obtained and authenticated by the American Type Culture Collection (Manassas, VA, USA). The cells were cultured in Dulbecco's modified Eagle medium (Invitrogen, USA) supplemented with 10% fetal bovine serum (Invitrogen), 1% streptomycin, and 1% ampicillin. Cells were maintained at 37°C in a humidified incubator with 5% CO_2. 6-Gingerol was purchased from Sigma-Aldrich (G1046).

Cell transfection

Transfection was performed using Lipofectamine 3000 (Invitrogen) following the manufacturer's protocol. Specifically, 2 μg of plasmids were transfected into cells that had been seeded on a six-well plate in the log phase 24 h prior. The transfection was performed using Lipofectamine 2000, and GFP transfection was used in parallel to estimate transfection efficiency. The pcDNA3.1-miR-506 plasmid and its scrambled negative control were obtained from GenePharma (Shanghai, China).

Clonogenic survival assay

Cells (1000 per dish) were seeded in triplicate in 100 mm Petri dishes and cultured in RPMI-1640 medium for 9 consecutive days. The medium was completely replaced on the day of seeding. Cells were fixed in 100% cold methanol for 15 min and stained with 0.25% crystal violet for another 15 min at room temperature. Colonies were washed with PBS and counted in three random fields.

PCR analysis

Total RNA was extracted using a HiPure Universal miRNA kit (Magen, Guangzhou, China) according to the manufacturer's instructions. RNA quality and quantity were verified using a BioAnalyzer 2100 (Agilent, Santa Clara, CA, USA). cDNA was synthesized using a miScript Reverse Transcription Kit (Qiagen, Valencia, CA, USA). Real-time PCR was performed using a CFX Connect™ Real-Time System (Bio-Rad, Inc., Hercules, CA, USA) and a miScript PCR Kit (Qiagen) according to the manufacturers' instructions. Relative miR-506 expression was normalized to that of U6 rRNA and calculated using the $2^{-\Delta\Delta Ct}$ method. Moreover, 5s rRNA was used for normalization to determine relative expression. Primers were synthesized by GenePharma (Shanghai, China). The following qPCR primers were used: miR-506 forward: 5′-GATCCTCTACTCAGAAGGGTGCCTTATTTTTG-3′; miR-506 reverse: 5′-AATTCAAAAATAAGGCACCCTTCTGAGTAGAG-3′; U6 forward: 5′-CTCGCTTCGGCAGCACA-3′; and U6 reverse: 5′-CGAATTTGCGTGTCATCCT-3′.

Western blotting

Total protein was extracted using a radioimmunoprecipitation assay, and concentrations were determined using a Pierce BCA Protein Assay kit (Thermo Fisher Scientific, Inc.), according to the manufacturer's instructions. Proteins (30 μg/lane) were separated using 10% sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and transferred to polyvinylidene difluoride membranes (EMD Millipore, Billerica, MA, USA). Membranes were blocked with 5% non-fat milk in PBS with 0.05% Tween-20 (PBST) and incubated overnight with primary antibodies at 4°C. Detection was performed using enhanced chemiluminescence (ECL, Millipore) after incubation with the secondary antibodies and a wash with Tris-buffered saline. The antibodies used were anti-rabbit (ab6721, 1:2500) and anti-mouse (ab6789, 1:2500) (both from Abcam).

Cell apoptosis analysis

Apoptosis was analyzed using annexin V/propidium iodide (PI) staining and flow cytometry (BD Biosciences, Franklin Lakes, NJ, USA).

Cells in a single-cell suspension were incubated in the dark for 15 min in HEPES buffer and analyzed using ModFit software (BD Biosciences).

Caspase inhibition assay

To determine whether apoptosis induced by 6-gingerol is caspase-dependent, SKOV3 cells were pre-treated with 20 μM Z-VAD-FMK (Selleck Chemicals) for 2 hours, followed by treatment with 20 μM 6-gingerol. Apoptosis was then assessed using Annexin V-FITC/PI staining.

Statistical analysis

Unless otherwise stated, all experiments were performed at least three times independently. Data are presented as mean ± standard deviation (SD). Statistical analyses were performed using SPSS 11.5 (SPSS Inc., Chicago, IL, USA). One-way ANOVA and multiple t-tests were used to assess significance, with $P < 0.05$ considered statistically significant.

Results

6-gingerol induced apoptosis in SKOV3 cells

We conducted an *in vitro* evaluation to determine the potential cytotoxic effects of 6-gingerol on human ovarian carcinoma SKOV3 cells. SKOV3 cells were treated with 5 μM,10 μM,15 μM and 20 μM concentrations of 6-gingerol for 6 days, and their survival rates were assessed using a clonogenic assay. Figure 1a shows a significant decrease in clonogenic survivors at both concentrations. In the 5 μM group, the survival rates were91%, 3.2%, 0.9% and0.07% on the 2^{nd}, 4^{th},6^{th} and8^{th} days of culture, respectively. In the 10 μM group, the survival rates were 61%, 9.1%, and 0.07% on the 2^{nd}, 4^{th}, and 6^{th} days of culture, respectively. In the 15 μM group, the survival rates were 52%, 0.39%, and 0.023% on the 2^{nd}, 4^{th}, and 6^{th} days of culture, respectively. In the 20 μM group, all cells died by the 6^{th} day of culture. To further confirm apoptosis, we analyzed the levels of cleaved caspase-3 and cleaved poly (ADP-ribose) polymerase (PARP) in response to 6-gingerol treatment, using endogenous tubulin as a loading control. As shown in Figure 1b, caspase-3 and cleaved PARP levels increased with higher 6-gingerol concentrations. To assess the dose-dependent effects of 6-gingerol on ovarian cancer cell apoptosis, we treated SKOV3 cells with 0, 10, and 20 μM of 6-gingerol for 2 days and analyzed the results using flow cytometry. The data (Figures 1c, d) show that the extent of apoptosis in SKOV3 cells increased proportionally with the 6-gingerol concentration. To further confirm the caspase dependence of 6-gingerol-induced apoptosis,SKOV3 cells were pre-treated with 20 μM Z-VAD-FMK (Selleck Chemicals) for 2 hours, followed by treatment with 20 μM 6-gingerol and treated with 20 μM 6-gingerol directly for 2 days. The data (Figures 1e, f) show that the extent of apoptosis in SKOV3 cells decreased proportionally with the Z-VAD-FMK treatment. These findings provide valuable insight into the caspase dependence of 6-gingerol to induce significant apoptotic responses in ovarian cancer cells, suggesting its effectiveness as a therapeutic agent.

6-gingerol reduces Gli3 expression

Given that GL13 knockdown inhibits the growth and migration of ovarian cancer cells (16), we investigated Gli3 expression in 6-gingerol-induced apoptosis. As shown in Figures 2a, b, treatment with 6-gingerol significantly reduced Gli3 expression in SKOV3 cells. However, no notable changes in the levels of other apoptosis-related proteins, such as Bcl-2, Bcl-w, and Bik, were observed. These results suggest that Gli3 downregulation plays a critical role in 6-gingerol-induced apoptosis in ovarian cancer cells.

6-gingerol upregulates miR-506

Evidence suggests that miRNAs are key regulators involved in cancer cell proliferation, differentiation, metastasis, and apoptosis. Therefore, we hypothesized that miRNAs might mediate the regulation of Gli3 expression by 6-gingerol. Using bioinformatics algorithms, including TargetScan, miRWalk, and miRDB, we identified seven candidate miRNAs that could potentially regulate Gli3 expression in response to 6-gingerol treatment. The relative expression of these miRNAs was determined using PCR and normalized to that of endogenous 5s rRNA. As shown in Figure 3, 6-gingerol treatment significantly upregulated miR-506 expression compared to other candidate miRNAs [(3.5 ± 0.6)-fold].

miR-506 directly inhibits Gli3 and induces apoptosis in SKOV3 cells

To verify the effect of miR-506 on Gli3 expression and apoptosis, we transfected SKOV3 cells with miR-506. As shown in Figure 4a, upregulation of miR-506 significantly increased apoptosis in SKOV3 cells (45.2% ± 5.1%) compared to that in the scramble control (3.7% ± 0.3%, Figure 4b). Western blot analysis further showed that excessive miR-506 levels suppressed Gli3 protein expression (Figure 4c).

6-gingerol induces apoptosis in SKOV3 cells via miR-506

We found that both 6-gingerol and miR-506 induced apoptosis in ovarian cancer cells. To investigate whether miR-506 mediates the apoptosis effects of 6-gingerol, we used an miR-506-specific antagonist (antago-miR-506). As shown in Figure 5a, treatment with 20 μM 6-gingerol significantly reduced the survival rate of

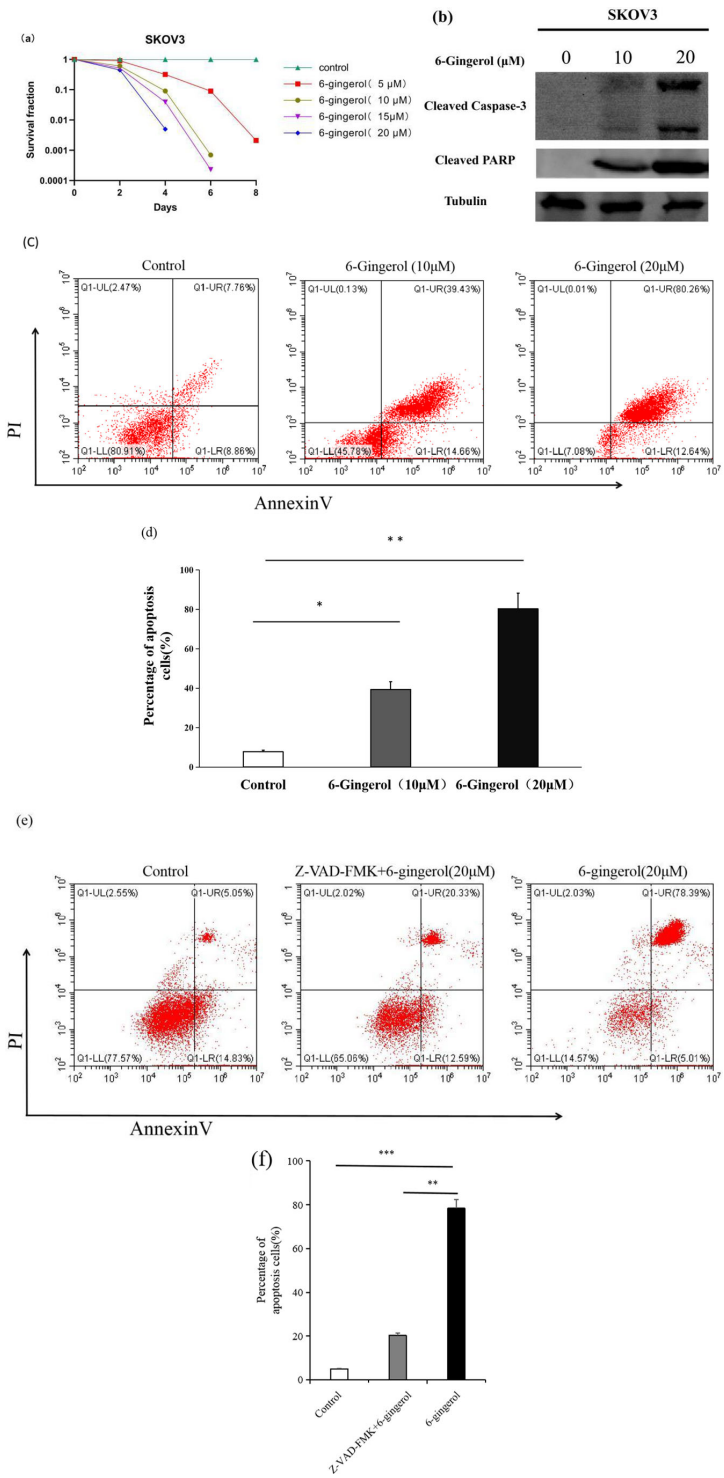


FIGURE 1
6-gingerol induces apoptosis in SKOV3 cells. **(a)** Clonogenic survival assay showing the survival rates of SKOV3 cells treated with 5 μM,10 μM,15 μM and 20 μM 6-gingerol for different durations (1st, 2nd, 4th, and 6th days). Results are based on independent experiments (n = 3). **(b)** Western blot analysis of cleaved caspase-3 or and cleaved PARP levels in SKOV3 cells treated with 6-gingerol. Tubulin was used as the loading control. **(c)** Flow cytometry analysis of apoptosis in SKOV3 cells treated with different 6-gingerol concentrations, using an Annexin V-FITC & propidium iodide (PI) apoptosis kit. Results are from three independent experiments (n = 3). **(d)** Quantification of apoptotic cells (double-positive for PI and Annexin V) from panel **(c)**. Results are presented as mean ± SD (n = 3). *$P < 0.05$, **$P < 0.001$. **(e)** Flow cytometry analysis of apoptosis in SKOV3 cells treated with 20 μM Z-VAD-FMK (Selleck Chemicals) for 2 hours, followed by treatment with 20 μM 6-gingerol and treated with 20 μM 6-gingerol directly for 2 days, using an Annexin V-FITC & propidium iodide (PI) apoptosis kit. Results are from three independent experiments (n = 3). **(f)** Quantification of apoptotic cells (double-positive for PI and Annexin V) from panel **P < 0.01, ***P < 0.001 **(e)**. Results are presented as mean ± SD (n = 3). *$P < 0.05$, **$P < 0.001$.

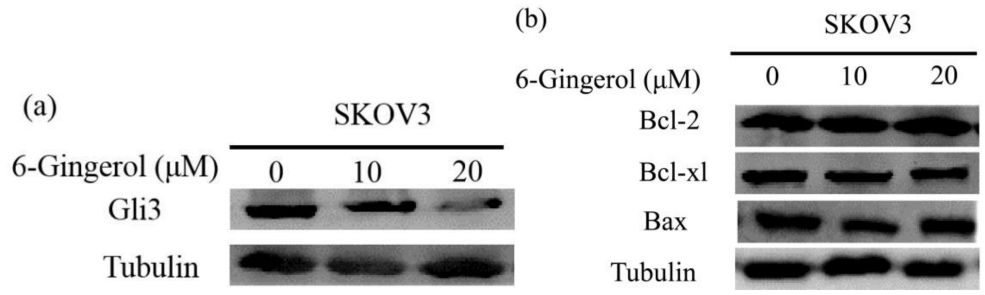


FIGURE 2
6-gingerol inhibits SKOV3 cells by reducing Gli3 expression. **(a)** Western blot analysis showing Gli3 protein levels in SKOV3 cells treated with 6-gingerol. Tubulin was used as a loading control. **(b)** Western blot analysis of apoptosis-related proteins (Bcl-xL, anti-Bcl-2, and Bax) in SKOV3 cells treated with 6-gingerol. Tubulin was used as a loading control.

SKOV3 cells. This effect was reversed by co-treatment with antago-miR-506. Similarly, flow cytometry analysis showed that the apoptosis induced by 6-gingerol in SKOV3 cells (68.2% ± 3.1%) was significantly reduced (9.4% ± 0.9%) when antago-miR-506 was introduced (P<0.05, Figures 5b, c). To elucidate the molecular mechanism, we performed western blot analysis to assess Gli3 expression in three groups: control, 6-gingerol, and 6-gingerol + antago-miR-506. As shown in Figure 5d, 6-gingerol treatment suppressed Gli3 expression; however, this suppression was reversed by antago-miR-506. These findings suggest that 6-gingerol induces apoptosis in SKOV3 cells by upregulating miR-506, which downregulates Gli3.

Discussion

Conventional anticancer therapies often lack specificity, targeting not only cancer cells but also healthy cells, leading to severe side effects. For example, platinum-based chemotherapy for ovarian cancer frequently causes gastrointestinal distress, bone marrow suppression, and liver and kidney damage (17, 18). Targeted therapies, while more specific, can still produce adverse effects, such as hypertension, proteinuria, and reduced blood cell counts. Natural compounds have emerged as promising alternatives to traditional treatments, offering increased efficiency with fewer side effects. These compounds can specifically target oncogenes and may also synergize with other chemotherapeutic agents (19, 20).

Throughout history, plant-based remedies have been widely used to treat various diseases, a practice that remains relevant today. Currently, herbal drugs account for over 50% of therapies in clinical trials (21). 6-Gingerol, the most abundant and biologically active phenolic compound present in the roots of ginger (*Zingiber officinale*), which has been more studied and more bioavailable than other phenolic compounds in ginger, exemplifies the medicinal potential of such natural products. Ginger has been used for centuries in China as a culinary spice and medicinal remedy.

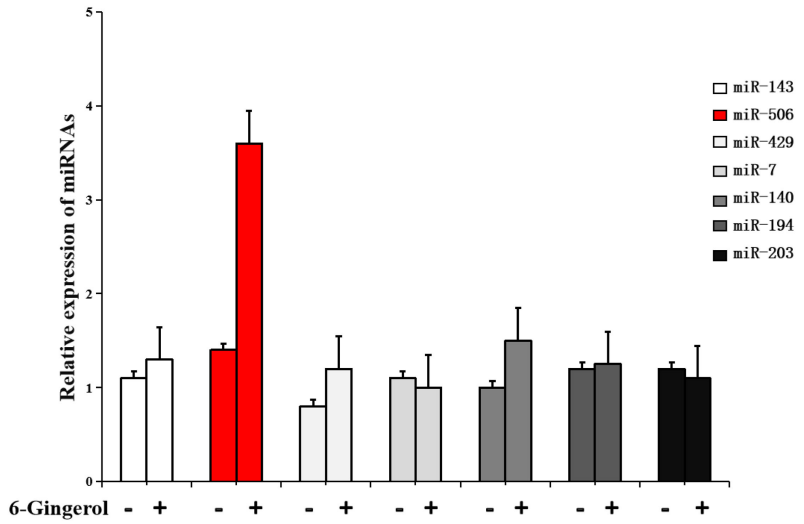


FIGURE 3
6-gingerol increases microRNA (miR)-506 expression in SKOV3 cells. RT-PCR analysis showing the expression levels of candidate microRNAs predicted to target Gli3 in SKOV3 cells treated with 6-gingerol. Data are normalized to the levels of 5s rRNA.

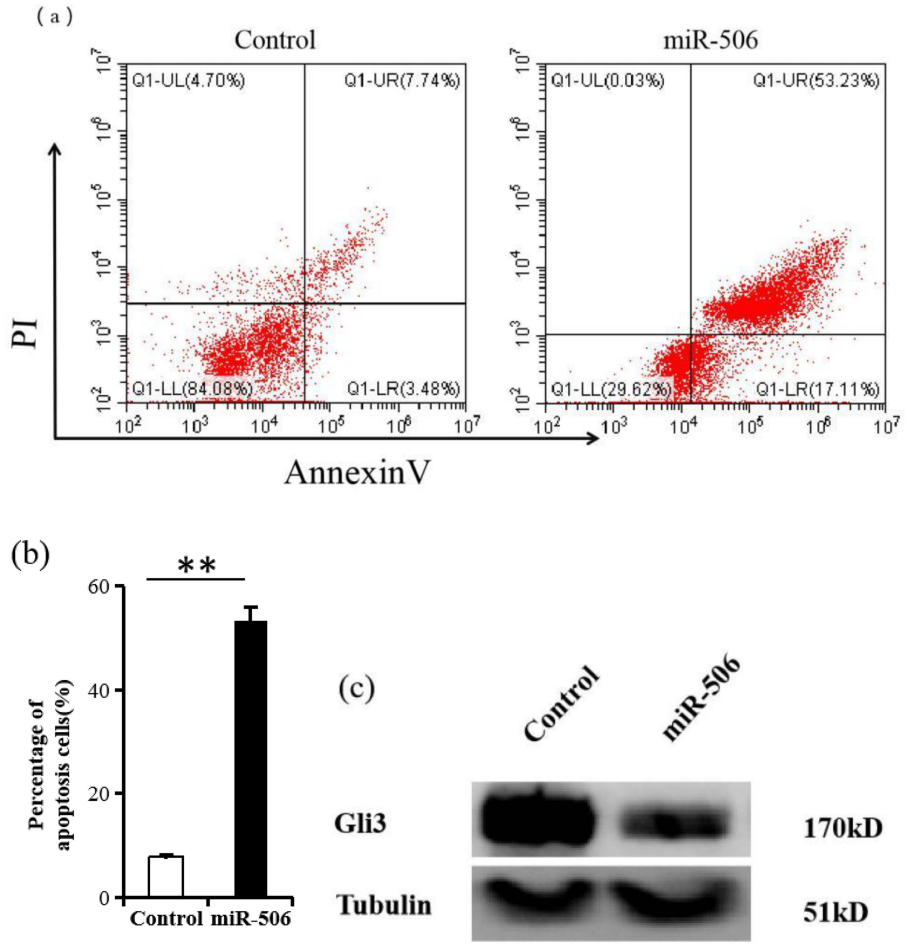


FIGURE 4
miR-506 suppresses Gli3 and induces apoptosis in SKOV3 cells. **(a)** Flow cytometry analysis of apoptosis in SKOV3 cells after transfection with miR-506, using Annexin V-FITC and propidium iodide (PI) staining. Results are based on three independent experiments (n = 3). **(b)** Quantification of apoptotic cells from panel **(a)**. The data show the percentage of double-positive Annexin V and PI cells. Results are presented as mean ± SD (n = 3). $^{**}P < 0.01$. **(c)** Western blot analysis showing Gli3 protein levels. Tubulin was used as a loading control.

Ginger has been a staple in traditional Chinese medicine for centuries, valued for its anti-inflammatory, antibacterial, and anticancer properties. Notably, 6-gingerol induces apoptosis in breast cancer cells by activating Bax transcription and caspase-7 (22).

The ability of 6-gingerol to arrest the cell cycle and induce apoptosis has been shown in human cervical and oral cancer cells (23, 24). Furthermore, 6-gingerol exhibits cytoprotective effects by reducing apoptosis and oxidative stress, potentially via the activation of Nrf2 pathways and inhibition of p38/NF-κB signaling (25). However, the mechanisms underlying the cytotoxic effects of 6-gingerol in ovarian cancer cells were previously unclear. Our study demonstrates that a concentration of 10 μM 6-gingerol effectively suppresses the clonogenic capacity of SKOV3 cells, leading to apoptosis.

We identified Gli3, a zinc-finger transcription factor, as a key player in this process. Gli3 has been implicated in the growth and metastasis of several cancer types. Knockdown of Gli3 suppresses the proliferation and migration of androgen receptor-positive breast and ovarian cancer cells, which does not occur for androgen receptor-negative cells (16). Additionally, loss of Gli3 in fibroblasts reduces suppressor cells derived from myeloid lineages and enhances natural killer cell activity, thereby inhibiting tumor growth (26). In colorectal cancer, Gli3 knockdown reduces cell migration and invasion by affecting epithelial-mesenchymal transition through the ERK signaling pathway. Elevated Gli3 expression correlated with poor prognosis in patients with colorectal cancer (27, 28). These results complicate the role of Gli3 expression in tumor tissues. In our study, 6-gingerol treatment significantly reduced Gli3 protein levels in SKOV3 cells. Interestingly, the expression of other apoptosis-related proteins, such as Bcl-2, Bax, and Bcl-xL, remained unchanged. To further explore the regulation of Gli3, we examined the role of miR-506, a microRNA known to regulate cell growth, differentiation, and metastasis, in SKOV3 cells treated with 6-gingerol. Bioinformatics analysis predicted miR-506 as a potential regulator of Gli3 expression, and our results confirmed that 6-gingerol upregulates

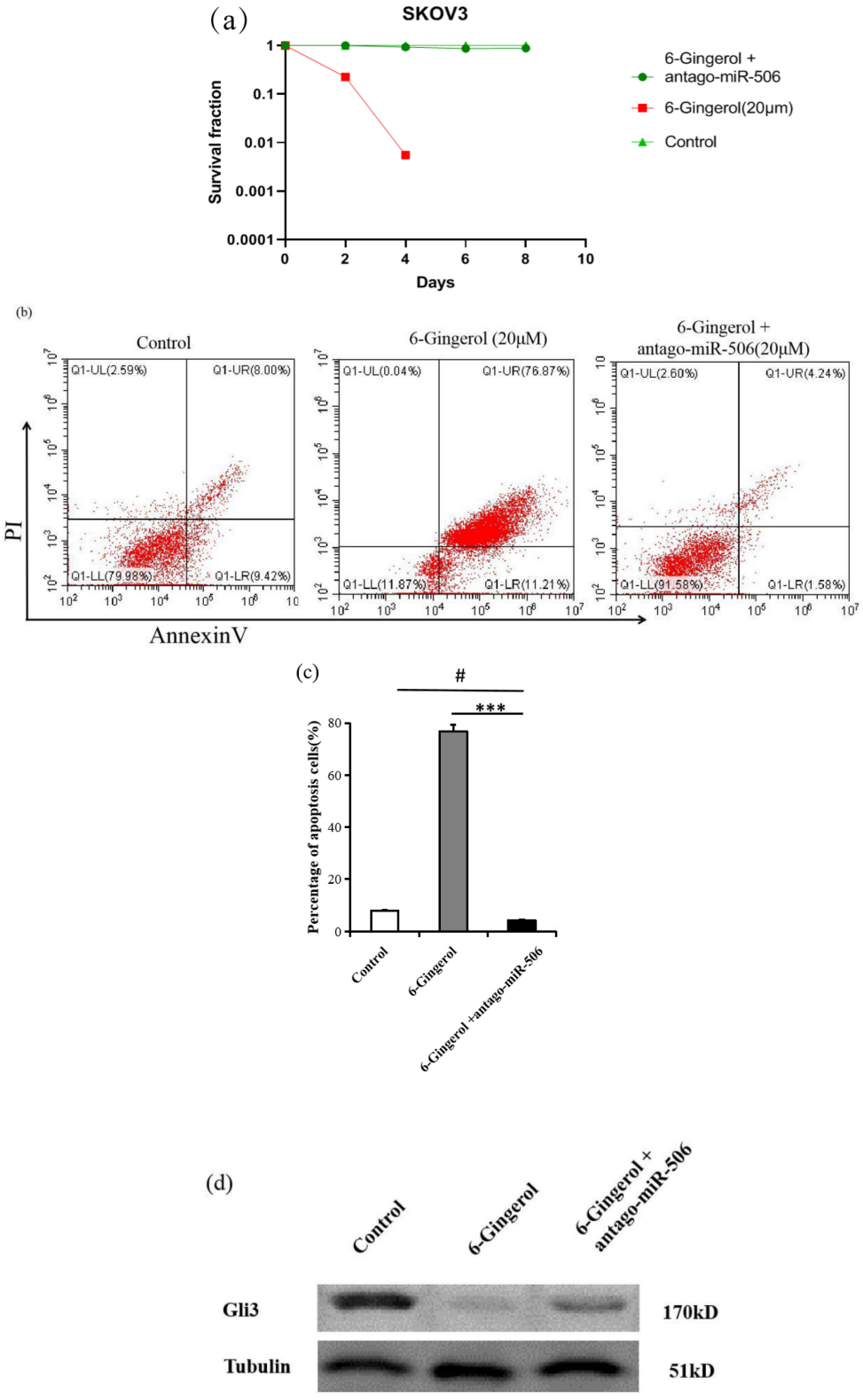


FIGURE 5
6-gingerol induces apoptosis in SKOV3 cells via miR-506. **(a)** Clonogenic survival assay showing the percentage of SKOV3 cells surviving after treatment with 20 μM 6-gingerol or 20 μM 6-gingerol + antago-miR-506 over different time points (days 1, 2, 4, 6, and 8). Results are based on three independent experiments (n = 3). **(b)** Flow cytometry analysis of apoptosis in SKOV3 cells treated with 6-gingerol or 6-gingerol + antago-miR-506 using Annexin V-FITC and propidium iodide (PI) staining (n = 3) # $P > 0.05$, *** $P < 0.001$. **(c)** Quantification of apoptotic cells (double-positive for PI and Annexin V) from panel **(b)**. Results are presented as mean ± SD (n = 3). ***$P < 0.001$, #$P > 0.05$. **(d)** Western blot was performed with anti-Gli3 antibody. Tubulin was used as a loading control.

miR-506, which in turn suppresses Gli3 expression and induces ovarian cancer cell apoptosis.

The role of miR-506 in cancer is context-dependent. In some cancer types, miR-506 acts as a tumor suppressor, whereas in others, it may function as an oncogene (29). For instance, Tong et al. (30) reported a high miR-506 expression in HCPT-resistant SW1116/HCPT colon cancer cells, suggesting its role in tumor suppression. Similarly, Streicher et al. (31) showed that the miR-506–514 cluster is consistently overexpressed in most melanomas, independent of the presence of B-raf or N-ras mutations. This cluster, or one of its sub-clusters (Sub-cluster A) comprising six mature miRNAs, can inhibit cell growth, promote apoptosis, and reduce invasiveness and colony formation in melanoma cell lines by reducing the expression of its target genes. Conversely, Luo et al. (32) found that miR-506 expression is reduced in glioblastoma. Overexpression of miR-506 in these cells suppressed cell growth, blocked the G1/S cell cycle transition, and inhibited cell invasion into glioblastoma cells. Zhang et al. (33) reported that cancer tissues and cultured cells exhibited lower miR-506 levels. They found that miR-506 expression was negatively correlated with EZH2 expression, lymph node invasion, tumor growth, metastasis, and TNM stage. Higher miR-506 levels were associated with a more favorable prognosis in patients. Consistent with these findings, we observed that miR-506 expression was significantly downregulated in ovarian cancer tissues. Our results showed that upregulation of miR-506 reduces ovarian cancer cell proliferation by targeting the transcription factor Gli3.

This study has several limitations. First, although SKOV3 cells are representative of high-grade serous ovarian cancer, validation in additional cell lines (e.g., CAOV3, OVCAR3) would strengthen the findings. Second, the functional role of Gli3 in migration/invasion was not examined, which should be addressed in future studies given its known metastatic functions. These limitations do not affect the core mechanistic conclusions but highlight directions for further research.

In summary, Our findings demonstrate that 6-gingerol induces ovarian cancer cell apoptosis through miR-506-mediated Gli3 suppression, providing an alternative to conventional Bax/Bcl-2-targeting approaches. Interestingly, while 6-gingerol has shown promise in combination with cisplatin (34), our work reveals its equally potent single-agent activity through this newly identified pathway. The clinical relevance of miR-506 downregulation in patient tumors further supports the therapeutic potential of 6-gingerol, particularly for tumors with impaired miR-506/Gli3 regulation.

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

JX: Conceptualization, Writing – original draft. H-HW: Conceptualization, Writing – original draft. HJ: Visualization, Writing – review & editing. HL: Visualization, Writing – review & editing. X-QT: Visualization, Writing – review & editing. X-JH: Funding acquisition, Supervision, Writing – review & editing. X-XC: Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research and/or publication of this article. This work was supported by Research Project of Traditional Chinese Medicine in Jiangxi Province (2019B028), Jiangxi Province Science and Technology Infrastructure Platform Construction Project (20203CCD46007), the Science and Technology Plan Fund of Jiangxi Provincial Health and Family Planning Commission (No. 202210036), Science and technology plan project of Jiangxi Administration of Traditional Chinese Medicine (2021B672), Science and Technology Research Project of Education Department of Jiangxi Province (190141), National Natural Science Foundation of China (#81760504), Jiangxi Provincial Natural Science Foundation Senior Project (No. 20242BAB25485), and General Project of Science and Technology Plan of Jiangxi Provincial Administration of Traditional Chinese Medicine (No. 2024B0268). We hereby declare that: All funders (including the Jiangxi Provincial Natural Science Foundation and Jiangxi Provincial Administration of Traditional Chinese Medicine) had no involvement in the study design, data collection/analysis, manuscript preparation, or publication decision.

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

References

1. Boyd J. Specific keynote: hereditary ovarian cancer: what we know. *Gynecologic Oncol*. (2003) 88:S8–S10. doi: 10.1006/gyno.2002.6674

2. Moufarrij S, Dandapani M, Arthofer E, Gomez S, Srivastava A, Lopez-Acevedo M, et al. Epigenetic therapy for ovarian cancer: promise and progress. *Clin Epigenetics*. (2019) 11:7. doi: 10.1186/s13148-018-0602-0

3. Lheureux S, Braunstein M, Oza AM. Epithelial ovarian cancer: Evolution of management in the era of precision medicine. *CA Cancer J Clin*. (2019) 69:280–304. doi: 10.3322/caac.21559

4. Stewart C, Ralyea C, Lockwood S. Ovarian cancer: an integrated review. *Semin Oncol Nurs*. (2019) 35:151–6. doi: 10.1016/j.soncn.2019.02.001

5. Bonifácio VDB. Ovarian cancer biomarkers: moving forward in early detection. *Adv Exp Med Biol*. (2020) 1219:355–63. doi: 10.1007/978-3-030-34025-4_18

6. Sun Y, Meng C, Liu G. MicroRNA-506-3p inhibits ovarian cancer metastasis by down-regulating the expression of EZH2. *J Cancer*. (2022) 13:943–50. doi: 10.7150/jca.66959

7. Kuroki L, Guntupalli SR. Treatment of epithelial ovarian cancer. *BMJ*. (2020) 371: m3773. doi: 10.1136/bmj.m3773

8. Yang C, Xia BR, Zhang ZC, Zhang YJ, Lou G, Jin WL. Immunotherapy for ovarian cancer: adjuvant, combination, and neoadjuvant. *Front Immunol*. (2020) 11:577869. doi: 10.3389/fimmu.2020.577869

9. O'Malley DM. New therapies for ovarian cancer. *J Natl Compr Canc Netw*. (2019) 17:619–21. doi: 10.6004/jnccn.2019.5018

10. Kang DY, Sp N, Kim DH, Joung YH, Lee HG, Park YM, et al. Salidroside inhibits migration, invasion and angiogenesis of MDA-MB 231 TNBC cells by regulating EGFR/Jak2/STAT3 signaling via MMP2. *Int J Oncol*. (2018) 53:877–85. doi: 10.3892/ijo.2018.4430

11. Sp N, Kang DY, Jo ES, Rugamba A, Kim WS, Park YM, et al. Tannic acid promotes TRAIL-induced extrinsic apoptosis by regulating mitochondrial ROS in human embryonic carcinoma cells. *Cells*. (2020) 9:282. doi: 10.3390/cells9020282

12. Wen J, Wang J, Li P, Wang R, Wang J, Zhou X, et al. Protective effects of higenamine combined with [6]-gingerol against doxorubicin-induced mitochondrial dysfunction and toxicity in H9c2 cells and potential mechanisms. *BioMed Pharmacother*. (2019) 115:108881. doi: 10.1016/j.biopha.2019.108881

13. Kubra IR, Rao LJ. An impression on current developments in the technology, chemistry, and biological activities of ginger (Zingiber officinale Roscoe). *Crit Rev Food Sci Nutr*. (2012) 52:651–88. doi: 10.1080/10408398.2010.505689

14. Hong MK, Hu LL, Zhang YX, Xu YL, Liu XY, He PK, et al. 6-Gingerol ameliorates sepsis-induced liver injury through the Nrf2 pathway. *Int Immunopharmacol*. (2020) 80:106196. doi: 10.1016/j.intimp.2020.106196

15. de Lima RMT, Dos Reis AC, de Menezes APM, Santos JVO, Filho JWGO, Ferreira JRO, et al. Protective and therapeutic potential of ginger (Zingiber officinale) extract and [6]-gingerol in cancer: A comprehensive review. *Phytother Res*. (2018) 32:1885–907. doi: 10.1002/ptr.6134

16. Lin M, Zhu H, Shen Q, Sun LZ, Zhu X. Gli3 and androgen receptor are mutually dependent for their Malignancy-promoting activity in ovarian and breast cancer cells. *Cell Signal*. (2022) 92:110278. doi: 10.1016/j.cellsig.2022.110278

17. Yang L, Xie H-J, Li Y-Y, Wang X, Liu X-X, Mai J. Molecular mechanisms of platinum-based chemotherapy resistance in ovarian cancer. *Oncol Rep*. (2022) 47:82. doi: 10.3892/or.2022.8293

18. Garrido MP, Fredes AN, Lobos-González L, Valenzuela-Valderrama M, Vera DB, Romero C. Current treatments and new possible complementary therapies for epithelial ovarian cancer. *Biomedicines*. (2022) 10:77. doi: 10.3390/biomedicines10010077

19. Naus PJ, Henson R, Bleeker G, Wehbe H, Meng F, Patel T. Tannic acid synergizes the cytotoxicity of chemotherapeutic drugs in human cholangiocarcinoma by modulating drug efflux pathways. *J Hepatol*. (2007) 46:222–9. doi: 10.1016/j.jhep.2006.08.012

20. Joung YH, Na YM, Yoo YB, Darvin P, Sp N, Kang DY, et al. Combination of AG490, a Jak2 inhibitor, and methylsulfonylmethane synergistically suppresses bladder tumor growth via the Jak2/STAT3 pathway. *Int J Oncol*. (2014) 44:883–95. doi: 10.3892/ijo.2014.2250

21. Sekiwa Y, Kubota K, Kobayashi A. Isolation of novel glucosides related to gingerdiol from ginger and their antioxidative activities. *J Agric Food Chem*. (2000) 48:373–7. doi: 10.1021/jf990674x

22. Wala K, Szlasa W, Sauer N, Kasperkiewicz-Wasilewska P, Szewczyk A, Saczko J, et al. Anticancer efficacy of 6-gingerol with paclitaxel against wild type of human breast adenocarcinoma. *Molecules*. (2022) 27:2693. doi: 10.3390/molecules27092693

23. Kapoor V, Aggarwal S, Das SN. 6-Gingerol Mediates its Anti Tumor Activities in Human Oral and Cervical Cancer Cell Lines through Apoptosis and Cell Cycle Arrest. *Phytother. Res*. (2016) 30:588–95. doi: 10.1002/ptr.5561

24. Rastogi N, Duggal S, Singh SK, Porwal K, Srivastava VK, Maurya R, et al. Proteasome inhibition mediates p53 reactivation and anti-cancer activity of 6-Gingerol in cervical cancer cells. *Oncotarget*. (2015) 6:43310. doi: 10.18632/oncotarget.6383

25. Han X, Liu P, Zheng B, Zhang M, Zhang Y, Xue Y, et al. 6-Gingerol exerts a protective effect against hypoxic injury through the p38/Nrf2/HO-1 and p38/NF-κB pathway in H9c2 cells. *J Nutr Biochem*. (2022) 104:108975. doi: 10.1016/j.jnutbio.2022.108975

26. Scales MK, Velez-Delgado A, Steele NG, Schrader HE, Stabnick AM, Yan W, et al. Combinatorial Gli activity directs immune infiltration and tumor growth in pancreatic cancer. *PloS Genet*. (2022) 18:e1010315. doi: 10.1371/journal.pgen.1010315

27. Shen M, Zhang Z, Wang P. Gli3 promotes invasion and predicts poor prognosis in colorectal cancer. *BioMed Res Int*. (2021) 2021:8889986. doi: 10.1155/2021/8889986

28. Iwasaki H, Nakano K, Shinkai K, Kunisawa Y, Hirahashi M, Oda Y, et al. Hedgehog Gli3 activator signal augments tumorigenicity of colorectal cancer via upregulation of adherence-related genes. *Cancer Sci*. (2013) 104:328–36. doi: 10.1111/cas.12073

29. Li J, Ju J, Ni B, Wang H. The emerging role of miR-506 in cancer. *Oncotarget*. (2016) 7:62778–88. doi: 10.18632/oncotarget.11294

30. Tong JL, Zhang CP, Nie F, Xu XT, Zhu MM, Xiao SD, et al. MicroRNA 506 regulates expression of PPAR alpha in hydroxycamptothecin-resistant human colon cancer cells. *FEBS Lett*. (2011) 585:3560–8. doi: 10.1016/j.febslet.2011.10.021

31. Streicher KL, Zhu W, Lehmann KP, Georgantas RW, Morehouse CA, Brohawn P, et al. A novel oncogenic role for the miRNA-506–514 cluster in initiating melanocyte transformation and promoting melanoma growth. *Oncogene*. (2012) 31:1558–70. doi: 10.1038/onc.2011.345

32. Luo Y, Sun R, Zhang J, Sun T, Liu X, Yang B. miR-506 inhibits the proliferation and invasion by targeting IGF2BP1 in glioblastoma. *Am J Transl Res*. (2015) 7:2007–14.

33. Zhang Y, Lin C, Liao G, Liu S, Ding J, Tang F, et al. MicroRNA-506 suppresses tumor proliferation and metastasis in colon cancer by directly targeting the oncogene EZH2. *Oncotarget*. (2015) 6:32586–601. doi: 10.18632/oncotarget.5309

34. Salari Z, Khosravi A, Pourkhandani E, Molaakbari E, Salarkia E, Keyhani A, et al. The inhibitory effect of 6-gingerol and cisplatin on ovarian cancer and antitumor activity: In silico, *in vitro*, and *in vivo*. *Front Oncol*. (2023) 13:1098429. doi: 10.3389/fonc.2023.1098429

frontiers | Frontiers in Medicine

TYPE Systematic Review
PUBLISHED 07 August 2024
DOI 10.3389/fmed.2024.1414582

OPEN ACCESS

EDITED BY
Monica Bianchini,
University of Siena, Italy

REVIEWED BY
Massimo Salvi,
Polytechnic University of Turin, Italy
Eric Munger,
United States Department of Veterans Affairs, United States
Jingwen Deng,
Guangzhou University of Chinese Medicine, China

*CORRESPONDENCE
Huihui Li
✉ lihh@gpnu.edu.cn
Chunlin Xu
✉ xuchunlin@gpnu.edu.cn

RECEIVED 09 April 2024
ACCEPTED 02 July 2024
PUBLISHED 07 August 2024

CITATION
Li H, Chen G, Zhang L, Xu C and Wen J (2024) A review of psoriasis image analysis based on machine learning. *Front. Med.* 11:1414582. doi: 10.3389/fmed.2024.1414582

A review of psoriasis image analysis based on machine learning

Huihui Li[1]*, Guangjie Chen[1], Li Zhang[2,3], Chunlin Xu[1]* and Ju Wen[2,3]

[1]School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China, [2]The Second School of Clinical Medicine, Southern Medical University, Guangzhou, China, [3]Department of Dermatology, Guangdong Second Provincial General Hospital, Guangzhou, China

Machine Learning (ML), an Artificial Intelligence (AI) technique that includes both Traditional Machine Learning (TML) and Deep Learning (DL), aims to teach machines to automatically learn tasks by inferring patterns from data. It holds significant promise in aiding medical care and has become increasingly important in improving professional processes, particularly in the diagnosis of psoriasis. This paper presents the findings of a systematic literature review focusing on the research and application of ML in psoriasis analysis over the past decade. We summarized 53 publications by searching the Web of Science, PubMed and IEEE Xplore databases and classified them into three categories: (i) lesion localization and segmentation; (ii) lesion recognition; (iii) lesion severity and area scoring. We have presented the most common models and datasets for psoriasis analysis, discussed the key challenges, and explored future trends in ML within this field. Our aim is to suggest directions for subsequent research.

KEYWORDS

machine learning, deep learning, dermatology, psoriasis, review

1 Introduction

Psoriasis is a chronic, inflammatory and hyperproliferative skin disease with a genetic basis (1). It can appear in any form on the arms, legs, scalp, buttocks, the folds of the skin and the trunk of the body (2). Awareness is increasing that psoriasis as a disease is more than skin deep and that it is associated with systemic disorders, including Crohn's disease, diabetes mellitus (notably type 2), metabolic syndrome, depression, and cancer (3). The disease follows a lengthy course and is prone to relapse, sometimes persisting for a lifetime. Psoriasis is characterized by scaling, silver shavings, protrusion and erythema. Its severity is evaluated based on the degree of infiltration, erythema, area, epidermal desquamation/scaling and other indicators, each of which is scored according to different clinical manifestations (4). Worldwide, approximately 125 million people have psoriasis, and psoriasis prevalence is highly variable across regions, ranging from 0.5% in parts of Asia to as high as 8% in Norway. In most regions, women and men are affected equally (5).

ML has been widely developed to analyse health data, particularly medical images, to assist professionals in making decisions and reducing medical errors. In particular, DL applications have shown promising results in dermatology and other specialties, including radiology, cardiology, and ophthalmology (6). ML technologies can be broadly classified into TML and DL. In TML, data features are obtained through a feature engineering process and then fed into a classifier for result prediction. Common TML classifiers include Random Forest (RF) (7), K-means (8), Decision Tree (9) K-Nearest Neighbor (KNN) (10)

and Support Vector Machine (SVM) (11). For instance, a random forest is a decision-making process, whereas KNN classifies vectors with similar distances in a feature space into the same class. Although these techniques are easy to explain and intuitive, they become less effective as the complexity of the data increases.

With the upgrading of algorithms and hardware, researchers began to focus on DL and explore its advantages in medical image analysis (12). DL has significant advantages in dermatological medical image processing: (1) Automatic feature extraction; (2) Handle complex data; (3) High performance. Convolutional neural networks (CNNs) are commonly used in the selection of DL models for dermatological diagnosis. Several CNNs-based models, including U-Net (13) and ResNet (14), have been used for psoriasis analysis. However, despite the strong potential of deep learning in skin medical image processing, it also faces challenges, such as data scarcity leading to model overfitting, complex models leading to long training times, and inexplicable models making it difficult for doctors to trust their results (15). Moreover, for DL, the deeper the layers of the model, the higher the hardware requirements, and the DL spend will be higher compared to TML.

Although recent studies have reviewed the application of AI in psoriasis diagnosis (16–19), these reviews did not conduct a thorough analysis of the ML models and the associated datasets. Therefore, this paper provides a detailed review of the use and advantages and disadvantages of ML models (including TML and DL models) in the application of psoriasis diagnosis. The contributions of this review can be summarized as follows:

- Provides a comprehensive overview of ML models used in psoriasis diagnosis, including TML models and DL models, and provides a detailed analysis of the advantages and disadvantages of each model.
- Evaluates existing psoriasis datasets and discusses their limitations in model development and evaluation.
- Proposes some future research directions to improve the accuracy and efficiency of psoriasis diagnosis.

The rest of this article is organized as follows: Section 2 introduces the methods adopted in this paper to conduct systematic review research; Section 3 introduces the results of paper retrieval. In Section 3.1, we introduce several publicly accessible datasets; The key content of this review, that is, the tasks of machine learning in various psoriasis analyses, are presented in Section 3.2, of which Section 3.2.1 is the segmentation task, Section 3.2.2 is the recognition task, and Section 3.2.3 is the assessment task. Section 4 is the discussion, including the challenges in Section 4.2 and future developments in Section 4.3; Finally, a systematic summary of this paper is given in Section 5.

2 Methods

We performed a literature search for relevant publications in 3 databases: Web of Science, PubMed, and IEEE Xplore. We chose these databases in order to cover general resources (Web of Science), medical (PubMed), and computing (IEEE Xplore). Relevant articles published in English between 2014 and April 2024, were considered. We use "and/or" operators to combine multiple keywords with "psoriasis", including "Machine Learning (ML)", "Deep Learning (DL)", "segmentation", "recognition", "assessment", and "review". To avoid missing keywords, we expanded the search scope of keywords to the entire text. Search expressions are shown in Table 1.

TABLE 1 Search expressions used in the systematic review.

Database	Query statement	Year of release
Web of Science	ALL=(psoriasis) AND (ALL=(ML) OR ALL=(DL))	2014–2024.04
PubMed	ALL=(psoriasis) AND (ALL=(ML) OR ALL=(DL)) AND (ALL=(segmentation) OR ALL=(recognition) OR ALL=(assessment))	
IEEE Xplore	ALL=(skin) AND ALL=(review) AND (ALL=(ML) OR ALL=(DL))	

We reviewed all retrieved papers from all platforms and removed duplicates, non-English papers, papers published before 2014, inaccessible papers, papers not related to machine learning, and papers not related to psoriasis. The remaining papers were confirmed by the authors to meet the requirements and were finally included in the review. Figure 1 reports our systematic review process using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses framework (20).

3 Results

Our search method identified 830 citations. After following the review protocal, 53 full-text articles were included for qualitative synthesis (Figure 1). Following the models used in the papers and the year of publication (Figure 2A), we found that the number of studies on psoriasis on machine learning has increased in recent years, a trend that can be attributed to the increase in datasets and advances in modeling. In all, we summarized a total of 10 papers on psoriasis lesion segmentation, 22 papers on psoriasis lesion recognition, and 21 papers on psoriasis severity scoring (Figure 2B). This review provides a comprehensive analysis of these papers and the datasets they use, describing the progress, limitations, and future directions of psoriasis in ML research.

3.1 Datasets

To conduct psoriasis analysis using ML, psoriasis data and various labels are necessary. After reviewing a significant amount of psoriasis-related literature, we discovered that most of it is produced in collaboration with hospitals and the datasets are private. As can be seen from the Table 2, from paper to paper they vary in the number of images, the source of the images and even the

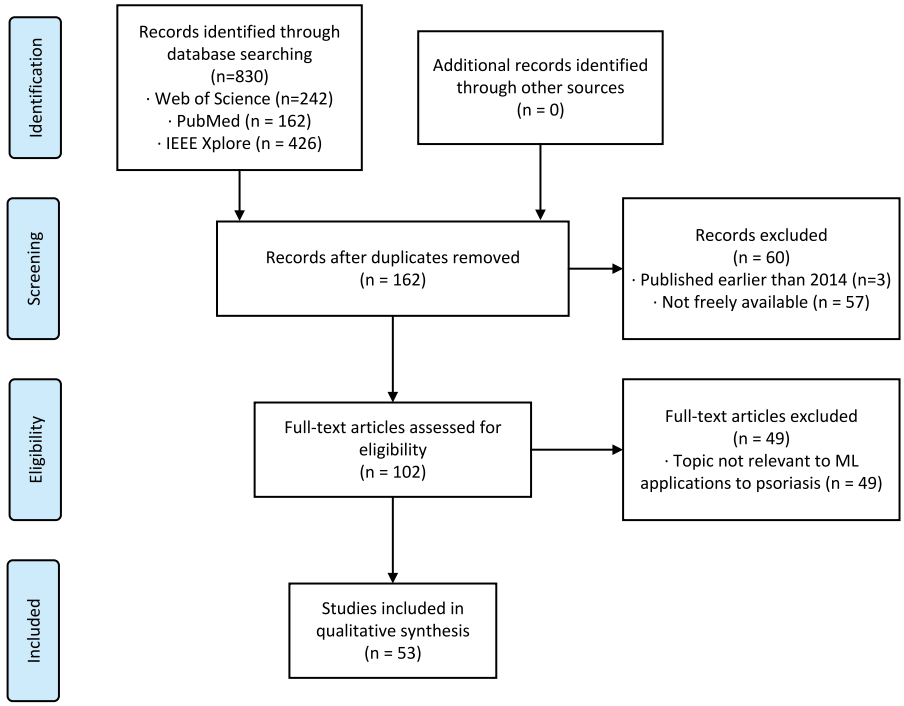


FIGURE 1
Systematic review flowchart according to the PRISMA framework. PRISMA indicates Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

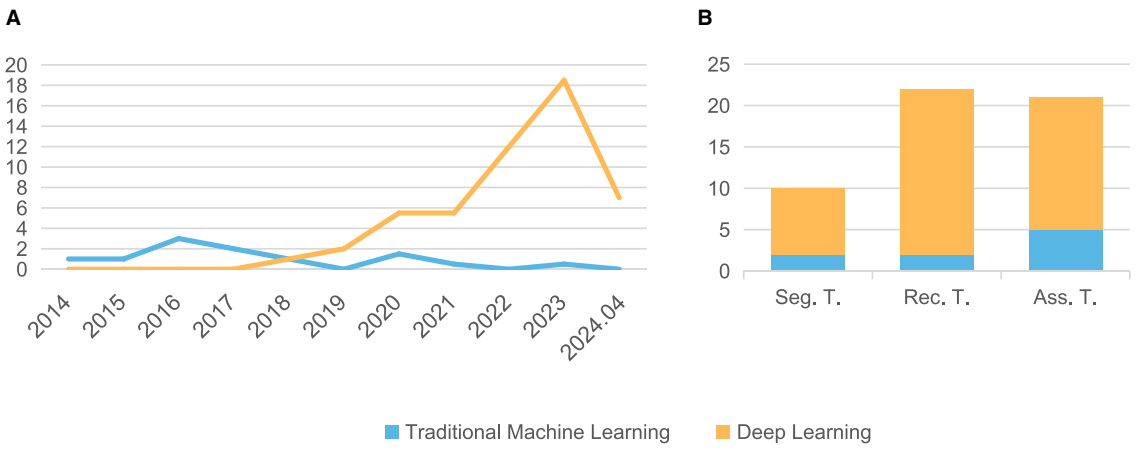


FIGURE 2
The distribution of the papers summarized in this article. **(A)** Number of papers published each year from 2014 to 2024.04; **(B)** Number of papers related to three different tasks. Seg, Segmentation; Rec, Recognition; Ass, Assessment; T, Task.

way the images are captured. This makes it impossible to compare these studies peer-to-peer, but only independently.

In addition to private datasets, there are also publicly accessible psoriasis datasets summarized in Table 3. One thing to note is that these publicly available datasets for psoriasis can only be applied to recognition tasks as they do not have segmentation masks and evaluation score labels. We have showcased some images from these publicly available datasets in Figure 3. Among them, the XiangyaDerm (29) and Kaggle[1] datasets not only include psoriasis but also cover other types of skin diseases such as Melanoma, Atopic Dermatitis, Basal Cell Carcinoma (BCC), and Benign

1 https://www.kaggle.com/datasets/ismailpromus/skin-diseases-image-dataset

TABLE 2 Statistics of private datasets adopted by the reviewed articles.

References	Number of images for various tasks and classes							
	Seg. task	Rec. task		Ass. task				
	Images	Pso	No-Pso	H.	Mi.	Mo.	Se.	V.Se.
George et al. (21)	676	-	-	-	-	-	-	-
Dash et al. (22)	5,179	-	-	-	-	-	-	-
Shrivastava et al. (23)	-	270	270	-	-	-	-	-
Zhao et al. (24)	-	900	7,121	-	-	-	-	-
Hammad et al. (25)	-	2,055	1,677	-	-	-	-	-
Shrivastava et al. (26)	-	-	-	383	47	245	145	28
Shrivastava et al. (27)	-	-	-	218	29	138	165	121
Dash et al. (28)	5,000	5,000	5,000	5,000	845	1,404	1,465	1,286

Pso, Psoriasis; H., Health; Mi., Mild; Mo., Moderate; Se., Severe; V.Se., Very Severe.

TABLE 3 Public dataset related to psoriasis and their description.

Dataset	Description
XiangyaDerm (29)	It contains **107,565** clinical images, covering **541** types of skin diseases. The largest amount of data in the dataset is psoriasis, **67,066** images, accounting for **62%** of the total dataset.
Skin diseases image dataset in Kaggle (see text footnote 1)	There are **10** types of skin diseases. Among them, **2,055** cases of psoriasis were included.
DermNetNZ (30)	It contains **11** different types of psoriasis, including but not limited to facial psoriasis, nail psoriasis, scalp psoriasis, etc.
Dermatology Atlas (31)	It contains **6** different types of psoriasis, including but not limited to arthropathic psoriasis, nail psoriasis, etc.
Hellenic Dermatology Atlas (32)	It contains **15** different types of psoriasis, including but not limited to generalized psoriasis, guttate psoriasis, inverse psoriasis, etc.

Keratosis-like Lesions (BKL). These two datasets are primarily used for multi-class skin disease recognition rather than being limited to the study of psoriasis alone. In the DermNetNZ (30), Dermatology Atlas (31), and Hellenic Dermatology Atlas (32) databases, we can observe various types of psoriasis with examples of their categories shown in the figure. The dataset available to the public contains information on different types of psoriasis, such as chronic plaque psoriasis, facial psoriasis, flexural psoriasis, and guttate psoriasis. These datasets can be used to train models to identify various types of psoriasis. Additionally, they offer a plethora of data on other skin conditions.

It can be clearly found in the Figure 3 that the most obvious problem of the psoriasis image is the lack of standardization of the data. The lesions appear in different positions, such as skin folds, hands, and joints. Some are even found in cluttered backgrounds. Therefore, it is difficult for doctors and even researchers to be confident whether the model, when recognizing these images of lesions, is extracting features from the lesion areas, or from other, distracting elements. As discussed in Yan et al. (33), there may be the same confusion concept in images of the same category, and the model is likely to refer to this confusion concept to classify this type of lesion, which we know is incorrect. We will discuss this in detail in the Challenges section.

3.2 ML application in psoriasis

In this section, we thoroughly describe the collected papers and summarize them in a table according to the research methodology. We also discuss the aims and results of these papers in detail. We classify the papers based on the real-world problems they address, including segmentation, recognition, and severity assessment of psoriasis.

3.2.1 Lesion segmentation

The accurate segmentation of lesion areas from skin images is essential for the development of effective computer-aided diagnosis (CAD) systems for skin diseases (34). In dermatology, common skin lesions include, but are not limited to, skin cancer, acne, eczema, and psoriasis. These lesions usually have different shapes, sizes, and colors, thus requiring specific algorithms to accurately segment them (35). Commonly used lesion segmentation methods include edge-based segmentation methods, region-based segmentation methods, and DL-based segmentation methods. Among them, DL-based methods have achieved good results in many fields due to their powerful feature extraction capabilities and adaptability. We summarize and present papers that apply ML to the task of psoriasis segmentation (Table 4).

For the evaluation indicators for segmentation task, the main indicators are the Dice Similarity Index (DSC) and Jaccard Index (JI). The DSC (44) metric represents the efficiency of the segmentation model by measuring the similarity between ground truth lesion (L_{gt}) and predicted segmented lesion (L_p) (45). Whereas, the JI (46) metric provides the overlapping measure between L_{gt} and L_p (38). Other performance metrics such as pixel accuracy (ACC), sensitivity (SE) and specificity (SP) are also available, where ACC indicates the proportion of image pixels classified correctly. In this paper, only their ACC metrics are

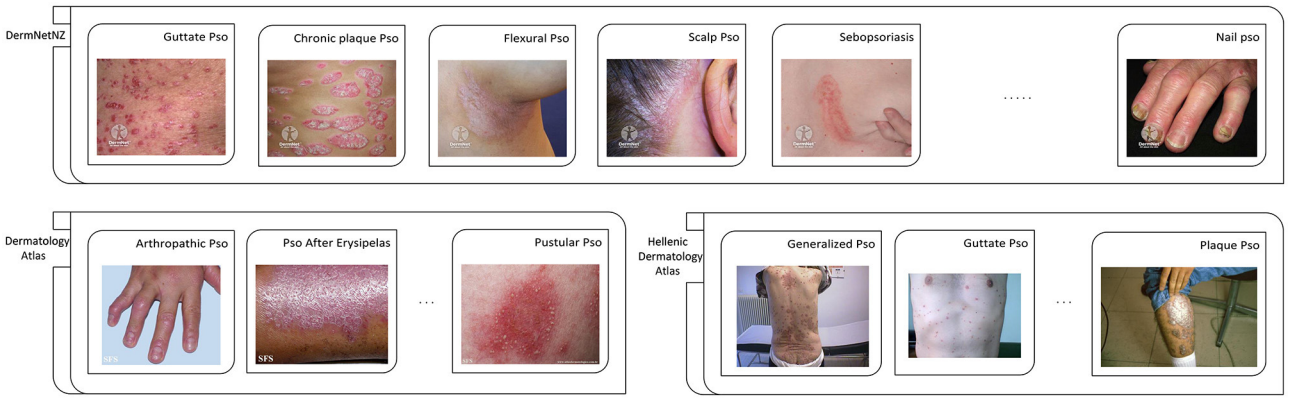


FIGURE 3
Partial examples of images from each exposed data set. Pso, Psoriasis.
Reprinted with permission of six watermarked images from the DermNetNZ dataset, which is labeled as Guttate Pso, Chronic plaque Pso, Flexural Pso, Scalp Pso, Sebopsoriasis, and Nail Pso, are from https://dermnetnz.org, © DermNet®, licensed under CC BY-NC-ND 3.0 NZ. For the DermNetNZ dataset, the links to the individual images are as follows: Guttate Pso, https://dermnetnz.org/topics/guttate-psoriasis; Chronic plaque Pso, https://dermnetnz.org/topics/chronic-plaque-psoriasis; Flexural Pso, https://dermnetnz.org/topics/flexural-psoriasis; Scalp Pso, https://dermnetnz.org/topics/scalp-psoriasis; Sebopsoriasis, https://dermnetnz.org/topics/sebopsoriasis; Nail Pso, https://dermnetnz.org/topics/nail-psoriasis.

Reprinted with permission of three watermarked images from the Dermatology Atlas dataset, which is labeled as Artropathic Pso, Pso After Erysipelas, and Pustular Pso, are from https://www.atlasdermatologico.com.br. For the Dermatology Atlas dataset, the links to the individual images are as follows: Artropathic Pso, https://www.atlasdermatologico.com.br/disease.jsf?diseaseId=43; Pso After Erysipelas, https://www.atlasdermatologico.com.br/disease.jsf?diseaseId=397; Pustular Pso, https://www.atlasdermatologico.com.br/disease.jsf?diseaseId=398.

Reprinted with permission of three images from the Hellenic Dermatology Atlas dataset, which is labeled as Generalized Pso, Guttate Pso, and Palque Pso, are from http://www.hellenicdermatlas.com/en/. For the Hellenic Dermatology Atlas dataset, the links to the individual images are as follows: Generalized Pso, http://www.hellenicdermatlas.com/en/search/advancedSearch/28/528/0/; Guttate Pso, http://www.hellenicdermatlas.com/en/search/advancedSearch/28/529/0/; Palque Pso, http://www.hellenicdermatlas.com/en/search/advancedSearch/28/535/0/.

TABLE 4 **Lesion segmentation.**

Methods	Remarks	References	Quantity of data	Evaluation metrics*		
				DSC↑	JI↑	ACC↑
Clustering	Image segmentation of lesion images using clustering algorithms from TML models	(21)	676	0.783	0.698	0.870
		(36)	780	-	0.830	0.909
CNN	The vast majority of CNN studies on psoriasis use U-Net as a segmentation model. Some papers also modify it to improve metrics	(22)	5179	0.930	0.864	0.948
		(37)	350	0.910	0.837	0.986
		(38)	500	0.948	0.901	0.992
		(39)	255	0.655	0.536	0.976
		(40)	580	0.919	-	-
Object detection backbone	Utilize object detection models as feature extraction modules in their proposed models before performing psoriasis segmentation	(41)	400	-	-	0.972
Optimization algorithm	These studies leverage CNNs where the weights and biases are optimized using optimization algorithms, for psoriasis segmentation	(42)	4200	0.960	0.905	0.970
		(43)	-	0.970	0.920	0.980

*DSC, Dice Similarity Index; JI, Jaccard Index; ACC: Pixel Accuracy.

counted. The formulas for the performance indicators are shown in Table 5.

Upon investigation, we found that the majority of papers utilizing traditional machine learning for psoriasis segmentation tasks employ clustering model algorithms (21, 36), such as K-means (8). Clustering algorithms group similar vectors in high-dimensional space and label them as the same class, excelling in both efficiency and interpretability. However, these algorithms

are primarily designed for numerical datasets, necessitating modifications to the images for their application. For instance, George et al. (21) adopted a strategy of segmenting images into superpixels of varying sizes, subsequently clustering these superpixels into lesion and non-lesion regions. Ultimately, they achieved a pixel accuracy of 86.99% on 100 test images. However, with the growth of the scale and complexity of datasets, traditional methods have become inadequate. This has led to the emergence of technologies such as DL.

U-Net (13) is a very popular DL model for medical image segmentation (47). It has demonstrated superior performance in medical segmentation tasks, capable of producing accurate segmentation results even with limited training data. Therefore, researchers favor the U-Net architecture and its variants as the backbone (22, 37, 38). Raj et al. (37) proposed a model for psoriasis lesion segmentation from the raw RGB color images having complex backgrounds and challenging surroundings. Taking advantage of residual networks and migration learning, Raj et al. (38) proposed a model with a residual encoder for segmenting psoriasis lesions from digital images with uneven backgrounds, based on U-Net. Czajkowska et al. (40) used DeepLab (48) for epidermal segmentation, which is a crucial first step for detecting changes in epidermal thickness, shape, and intensity. In psoriasis diagnosis, it is also necessary to score the elevation level of lesions. However, conventional computer vision models can only process 2D images and are not well-suited for training on 3D elevation data. Therefore, this method is worth studying.

Using object detection models as a backbone for segmentation tasks is also an alternative approach compared to using conventional segmentation models (41). Their main approach is to use object detection models [e.g., Lin et al. (41) using Mask R-CNN (49)] as a backbone such as a feature extractor for the segmentation model, followed immediately by a segmentation output branch to perform the segmentation task.

Unlike proposing new CNNs, in order to guide the training of CNNs that can move toward more excellence, Mohan et al.

TABLE 5 Formulas for different performance indicators for segmentation task.

Performance metric	Formula*
DSC	$DSC = \frac{2\times\lvert L_{gt}\cap L_p\rvert}{\lvert L_{gt}\rvert+\lvert L_p\rvert} = \frac{2\times TP}{FP+FN+(2\times TP)}$
JI	$JI = \frac{\lvert L_{gt}\cap L_p\rvert}{\lvert L_{gt}\cup L_p\rvert} = \frac{TP}{TP+FN+FP}$
ACC	$ACC = \frac{TP+TN}{TP+FP+TN+FN}$

*TP, ture positive; FP, false positive; TN, true negative; FN: false negative.

TABLE 6 Lesion recognition.

Methods	Remarks	References	Quantity of data	Evaluation metrics*		
				ACC↑	F1↑	AUC↑
PCA; SVM	Traditional machine learning methods.	(23)	540	1.0	-	1.0
		(51)	90	0.90	-	-
CNNs	Classify psoriasis vs. other skin disease (including healthy skin)	(52)	1,358	-	-	0.922
		(53)	3,570	0.801	-	-
		(54)	312	0.942	0.942	0.990
		(55)	1,876	0.910	-	-
		(56)	2,101	0.919	0.894	0.959
	A publicly available dataset was used for the study.	(57)	938	0.653	0.655	0.904
		(24)	8,021	0.960	-	0.981
	Identify psoriasis from skin lesion such as eczema and pityriasis rosea that are extremely similar to it.	(58)	4,740	0.959	-	0.987
		(59)	11,031	0.920	-	-
		(60)	292	0.896	-	-
		(25)	3,732	0.962	0.958	0.971
		(61)	869	0.857	-	-
	Identify nail psoriasis from healthy nails.	(62)	1,155	0.957	-	-
	Light-weighted CNN	(63)	33,904	0.70	-	-
	CNN + ViT	(64)	8,000	0.977	0.965	-
	Classify different types of psoriasis.	(65)	30,000	-	0.890	0.920
		(66)	1,836	0.987	0.958	-
		(56)	814	0.933	0.919	-
	CNN vs. LSTM	(67)	1,838	0.842	-	-
	Light-weighted CNN	(68)	12,015	0.998	-	0.99

*ACC, Accuracy; F1, F1-Score; AUC, Area Under Curve.

TABLE 7 Formulas for different performance indicators for recognition and assessment task.

Performance metric	Formula*
ACC	$ACC = \frac{TP+TN}{TP+FP+TN+FN}$
Recall	$Recall = \frac{TP}{TP+FN}$
Precision	$Precision = \frac{TP}{TP+FP}$
F1-Score	$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall}$

*TP, ture positive; FP, false positive; TN, true negative; FN, false negative.

(42) proposed a convolutional neural network (CNN) based on the Adaptive Chimpanzee Optimization Algorithm (AChOA) for automated segmentation of psoriasis skin images, which utilizes the AChOA to optimize the weights and bias values of the CNN. Similarly, Panneerselvam et al. (43) proposed Adaptive Golden Eagle Optimization (IGEO) to tune the weights and bias parameters of the CNN.

The segmentation task plays a crucial role in the application of computer technology to the medical field. It not only helps eliminate interference from non-lesion regions, but also provides a solid foundation for subsequent recognition or assessment tasks.

3.2.2 Lesion recognition

The process of diagnosing skin cancer is intricate and involves visual examination and judgment by a physician, followed by microscopic examination of a biopsy. Therefore, developing more accurate algorithms for skin lesion recognition could greatly facilitate timely diagnosis of skin cancer. Automated classification of lesions is used in clinical examination to help physicians and allow rapid and affordable access to lifesaving diagnoses (50). Lesion recognition aims to differentiate psoriasis from other common skin diseases (or healthy skin) or to distinguish between different types of psoriasis, primarily through techniques such as feature extraction and segmentation. We summarize and present papers that apply ML to the task of psoriasis recognition (Table 6).

Four performance metrics are used to evaluate the performance of the recognition models: Accuracy(ACC), recall, precision and F1-score(F1). We summarize the ACC and F1 in the paper (since F1 then already makes use of recall and precision). The formulas for the performance indicators are shown in Table 7. In addition, we also summarized the Area Under Curve(AUC) metrics from the papers. In the task, "psoriasis" was represented as a positive category and "non-psoriasis" as a negative category, and a threshold was set to distinguish positive or negative cases. By constantly adjusting this threshold, we were able to obtain multiple sets of different sensitivities and specificities. These sets were then labeled in coordinates and Receiver Operating Characteristic (ROC) curves were plotted (24). AUC is the area of the ROC curve, which is used to measure the performance of machine learning algorithms for "classification problems" (generalization ability).

When using TML models for psoriasis classification, researchers extract color and texture features from the images, corresponding to the erythema and silver desquamation attributes of psoriasis, respectively, since these models cannot actively analyze images (23, 51). Among them, Texture features are the most traditional way to explore specific pattern information in images, and they can quantify the texture present in lesions. Common texture analysis techniques include: Gray Level Co-occurrence Matrix(GLCM), Gray Level Run Length Matrix (GLRLM) (69), etc. For the obtained features, they can be fed into Principal Component Analysis (PCA) (70) for dimensionality reduction, which is a feature dimensionality reduction technique. From the experimental results of Shrivastava et al. (23), the best classification result was obtained by using the features of Higher Order Spectra (HOS) (71), texture and color together for classification, and the binary classification accuracy can reach 100%.

However, to achieve classification between different skin diseases, or even between different types of psoriasis, it is not enough to use TML. From the CNNs section of the table we can see that there are two main tasks in psoriasis recognition. For the former, the focus of the psoriasis identification task is on distinguishing psoriasis from skin diseases that are very similar to psoriasis compared to common classification tasks such as the ISIC dermatology dataset (72), e.g., to distinguish scalp psoriasis from scalp seborrheic, which have the same region of onset and a small difference in the lesion appearance but have completely different treatment approaches, CAD comes in handy in order to avoid incorrect diagnoses by doctors (52). Lichen planus, parapsoriasis, lupus erythematosus and eczema are also particularly similar but differently treated skin conditions which, in addition to all being characterized by a reddish color, also have papules or plaques (25, 58–61). Because of Inflammatory skin diseases, such as psoriasis (Pso), eczema (Ecz), and atopic dermatitis (AD), are very easily to be mis-diagnosed in practice, Wu et al. (58) developed an end-to-end deep learning model. Yang et al. (59) aimed to train an efficient deep-learning network to recognize dermoscopic images of psoriasis (and other papulosquamous diseases), improving the accuracy of the diagnosis of psoriasis. While they have similar symptoms, Psoriasis and Eczema have vastly different underlying causes and behaviors, Chatterjee et al. (60) explores state of the art Deep Learning techniques for distinguishing Psoriasis and Eczema. Hammad et al. (25) presents an enhanced deep learning approach for the accurate detection of eczema and psoriasis skin conditions. Zhu et al. (61) propose a novel abscissa-ordinate focused network (AOFNet) with active label smoothing for the identification of psoriasis and eczema from images.

Using models from the natural language processing (NLP) domain to extract image features is a very popular approach. This is because these models, when applied to sentences, are able to capture the distant relationships between sentences and thus calculate the relationships between words. The researchers want to try to use this idea to capture long distance relationships between images to make up for the fact that the computation of convolution can only capture local information. Aijaz et al. (67) innovatively used Long Short-Term Memory (LSTM) (73) for classification in addition to CNNs. However, LSTM only obtained an accuracy of 0.723 on the results (CNN obtained 0.842), proving that CNN is still superior to models from NLP for image processing. Vishwakarma et al. (64) proposed a model that combines the features of a CNN and a Vision Transformer (ViT) (74) with the aim of building a high-performance, lightweight hybrid model for the intended task. In this, ViT processes the convolutional feature maps to capture long-term dependencies that represent global features.

The use of deeper neural networks is a straightforward and effective way to deal with the increase in the amount of data, but this can lead to a very fatal problem - an increase in the number of parameters, resulting in the need for better hardware. However, instead of opting for a larger model, Arunkumar et al. (63) proposed their own lightweight CNN when solving tens of thousands of datasets, and obtained relatively good results. The model proposed by Rashid et al. (68) is very easy to be used and deployed as a smartphone application in a real-time decision-making environment due to its lightweight nature. The model can handle recognition and classification of psoriasis types for low or high resolution images.

Zhao, Aggarwal, and Rashid et al. (24, 57, 68) used the psoriasis dataset (Table 3) from a public dataset for identification of common skin diseases and psoriasis. The study using the public dataset can enhance the confidence of the diagnosis as all images were verified by pathological examination and history and labeling was done by experienced dermatologists. We believe that psoriasis research will become more comprehensive as more and more papers conduct research on public datasets.

3.2.3 Lesion severity assessment

Psoriasis severity assessment refers to the objective and accurate evaluation of the severity of a patient's psoriasis, so that the doctor can develop a reasonable treatment plan and monitor its effectiveness. Commonly assessment methods include the PASI scoring system, DLQI scoring system (75), etc. Among them, the PASI score system is used to score psoriasis patients based on factors such as lesion area, erythema, scaling, and infiltration, with a total score of 0 to 72. The higher the score, the more severe the condition. In the process of using ML to evaluate the severity of psoriasis, feature selection is a very important step, including the extraction of features such as lesion area, erythema, scaling, and infiltration. Before this, it is necessary to segment and identify the image, especially to prevent the background interference from affecting the extraction of color features. We summarize and present papers that apply ML to the task of psoriasis severity assessment (Table 8).

Similar to the psoriasis classification task, the task of psoriasis severity assessment using TML models also requires the extraction of various features such as color and texture in the image, which are then fed into various classifiers for severity assessment. In this regard, Shrivastava et al. (26, 27). conducted two different experiments on two different datasets, one on the 848 psoriasis dataset, which achieved 99.92% accuracy, and one on the 670 psoriasis dataset, which was first segmented by Bayesian modeling and then classified, which achieved 99.84% accuracy. It can be noticed that although the dataset has become smaller, the accuracy can still be kept high by segmentation followed by classification.

In the experiments of Moon et al. (79), they used and compared automatic [Simple linear iterative clustering (SLIC) superpixel-based segmentation (21) and U-Net model] and semi-automatic [level set method (LSM) (94) and interactive graph cuts (IGC) (95)] segmentation algorithms. It was found that the semi-automatic segmentation models are particularly subjective and time consuming, while the automatic models are less effective in segmenting the curved, illuminated or shadowed parts of the image. From the results, the LSM from semi-automated segmentation was able to achieve a DICE of 0.945 and the SLIC from automated segmentation a DICE of 0.915 (Other segmentation metrics are noted in the paper). Taking into consideration time efciency and reproducibility, the paper finally chose SLIC as the segmentation task model before the evaluation task.

The work of Dash et al. (28) is the most consistent with the physician's diagnostic process within all the papers. Specifically, they distinguished 5,000 healthy skin from 5,000 psoriasis with 99.08% accuracy, then, segmented the lesion areas in the psoriasis images with 94.76% accuracy, and, ultimately, assessed the segmented images at four levels of severity with 99.21% accuracy. Raj et al. (84) extended the work of Dash et al. (22) by broadening the scope of lesion detection to segment healthy skin, psoriatic lesions, and background regions simultaneously from full-body areas.

Training out a segmentation model requires relevant data with labels, and how well it is trained affects the subsequent tasks, with errors at each stage accumulating to be very catastrophic in the end (77). Thus, Huang et al. (88) avoided the use of segmentation models and instead added various attention modules after the backbone output, allowing the model to localize the lesion area without going through the segmentation model. Schaap et al. (87) utilized a special CNN (96) for the assessment task. This CNN is assessed for psoriasis with a decreasing probability from 0 to 5, with a final threshold set to arrive at a score for that psoriasis. Moon et al. (92) used CutMix to generate multiple-severity disease images and proposed a hierarchical Multiscale Deformable Attention Module (MS-DAM) that adaptively detects representative regions of irregular and complex patterns in multi-severe disease analyses.

You Only Look Once (YOLO) (97) is a deep neural network-based target recognition and localization algorithm with fast processing speed and suitable for real-time systems. YOLO-v4, which builds on the original YOLO target detection architecture, employs state-of-the-art optimization strategies in the field of CNNs. Thus, Yin et al. (93) used the YOLO-v4 algorithm as a feature extractor for images to detect the severity and lesion area of each disease in a specific portion of an image and perform a comprehensive assessment.

ViT's input adaptive weighting and global information learning can show good performance in vision related tasks. Raj et al. (85) put ViT into a classification module for computation, where the feature vectors output from the backbone are computed globally, and then the output is collapsed back into the dimensions of the feature representations produced by the convolution operation.

4 Discussion

4.1 Methods statistical analysis

We have summarized the methods used in the collected papers (Figure 4). We found that when researchers select TML models, for segmentation tasks, clustering models such as K-Means are usually used to achieve segmentation of diseased regions by clustering diseased pixels together. Whereas for lesion recognition and

TABLE 8 Lesion severity assessment.

Methods	Remarks	References	Quantity of data	Evaluation metrics*		
				ACC↑	F1↑	AUC↑
PCA; SVM; NB; DT	Traditional machine learning methods	(76)	17	0.920	-	-
		(26)	848	0.999	-	0.999
		(27)	670	0.998	-	0.998
Dic. L; BoVWs	A novel image representation and unsupervised feature extractor method	(77)	676	-	0.710	-
		(78)	676	0.808	-	-
CNNs	Segmentation was performed before severity assessment	(28)	5,000	0.926	0.926	0.992
	Semi-automatic vs. automatic segmentation algorithms	(79)	80	-	0.989	-
	Segmenting and scoring nail psoriasis	(80)	705	0.765	-	-
		(81)	300+	0.915	-	-
		(82)	1,154	0.55	0.55	0.63
	Segmenting and scoring pustular psoriasis (PP)	(83)	611	0.667	-	-
	Segmenting and scoring large areas of psoriasis	(84)	500	0.942	-	-
	CNN + ViT	(85)	1,018	0.795	0.792	0.950
	Direct assessment of psoriasis severity using CNNs	(86)	705	-	-	-
		(87)	1,731	-	-	-
		(88)	14,096	-	-	-
		(89)	5,951	-	0.940	-
		(90)	792	0.910	-	-
		(91)	2,700	-	-	-
	Attention	(92)	792	0.908	0.930	-
	YOLO	(93)	2,657	-	-	-

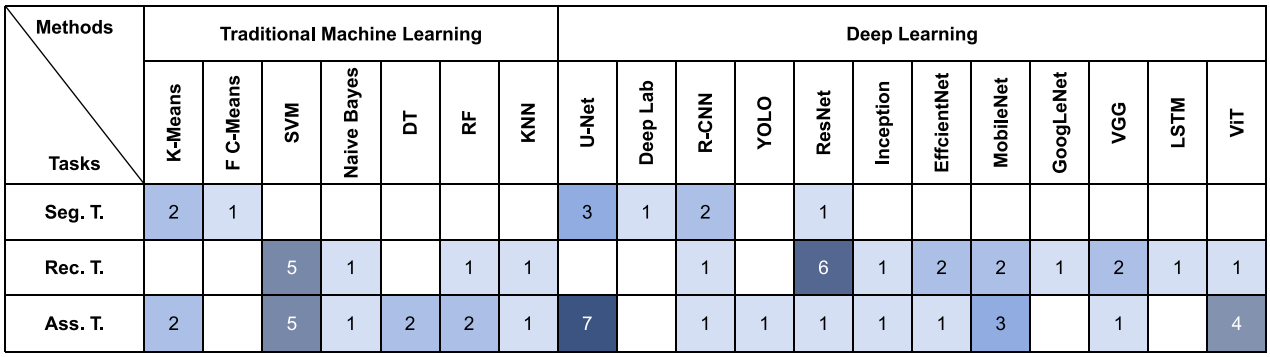


FIGURE 4
Quantitative distribution of different ML methods on the three tasks.

assessment tasks, given the limited datasets available for psoriasis, researchers tend to favor support vector machines as it performs well with small datasets.

In DL model selection, U-Net is widely used for its high accuracy in medical segmentation (98). Segmentation models are also utilized in psoriasis recognition or assessment tasks, where only by locating and segmenting the diseased regions, the model is able to avoid interference from non-diseased regions (99).

Some methods originally used for NLP (e.g., LSTM and Transformer) have been widely used in the field of computer vision

in recent years (100), and have also been applied to medical image analysis. However, there are fewer papers using these methods to analyse psoriasis, and their scalability in medical images needs to be further investigated. In addition, many other methods are not shown in the diagram, and we have only summarized the most commonly used ones.

4.2 Challenges

Through a comprehensive analysis of collected papers, including data collection, preprocessing, modeling approaches and experiments, we analyse the current challenges of machine learning in psoriasis.

4.2.1 Lack of data sources

ML (especially DL) algorithms require large amounts of data to effectively train models (101). However, since very few people study psoriasis in the field of ML, the amount of data available for analysis then becomes very limited, making it difficult to build accurate and reliable models. In addition, most psoriasis datasets are not publicly available, and most of the datasets used in the papers listed in the table above were obtained through collaboration with hospitals. Moreover, different tasks require different annotations, which adds to the complexity of ML for research in the field of psoriasis. To use ML for psoriasis research, access to sufficient data is critical. However, this may not always be feasible due to the high cost of physician annotation time or the difficulty of obtaining consistent images (102). In addition, the acquired images may have unevenly distributed categories or incorrect labels, which can lead to training the model in the wrong direction or overfitting.

4.2.2 Data inconsistency

Even if there is enough data, its inconsistency and irregularity can lead to poor model performance. That is, if the data come from different databases or are taken by different doctors with different angles, lighting or resolutions, then the integration and analysis of these data will be a big challenge. Although the International Skin Imaging Collaboration (ISIC) has attempted to address the issue of data standardization by developing a set of technical standards for skin lesion imaging (103), psoriasis differs from common dermatological datasets in that the site of onset can be systemic (e.g., body depressions), which leads to the analysis not being able to train the model exactly according to the characteristics of the dermatological condition (rounded, localized, more regular, flattened). At the same time, some features are difficult to obtain through machine such as the sclerotic height of psoriasis, and most of the commonly used DL is applied to flat images, which can only obtain features that are accessible to flat vision, such as color and texture. Although skin thickness segmentation was proposed in Czajkowska et al. (40), it is particularly demanding on the dataset.

4.2.3 The inexplicability of methods

Selection of appropriate methods and improvement of existing methods to improve the accuracy of psoriasis analyses are common threads in existing papers, but doctors and patients are most concerned about the accuracy of psoriasis analyses and whether the researchers can explain how the proposed models arrive at their conclusions. However, from the collected papers, most of them only propose a model with good diagnostic results for psoriasis, while little research has been done on the interpretability of the model.

4.3 Future development

In response to these challenges to the application of ML in psoriasis, we propose solutions and summarize the future development of ML.

4.3.1 Few-shot learning

Model training using a small amount of data is also a current research hotspot in ML, especially DL. For example, Folle et al. (82) used a small number of samples to study the diagnosis of psoriasis, and the BEiT model, which they used, was designed to train models with fewer samples. Few-shot learning is a ML paradigm designed to enable efficient training of models with a small number of samples. In Xiao, Liu and Chen et al. (104–107), they classified and segmented lesion data with fewer lesion images. Data collection for psoriasis is also difficult, especially labeling, and requires overcoming a variety of subjective factors. In today's era of predominantly data-driven model training, smaller, more granular datasets may produce better results than larger, more extensive datasets.

4.3.2 Feature consistency

Differences between images can also worsen the model, especially in feature extraction. Therefore, we would like to unify the images before training the model, or, in other words, extract common features. For example, Diaz et al. (108) aim to pixelate images using a segmentation model that labels pixels belonging to the same lesion features (e.g., pigment networks, blue-white stripes, dots, bubbles, blood vessels) as belonging to the same category in skin lesions. This reduces the differences in image-level features by extracting pixel-level features, while directing the model to use these features for further training and avoiding image differences that cause the model to recognize the same features as different features. However, segmentation requires labeling, which leads to a relatively poor feasibility of this approach. To solve this problem, Pathak et al. (109) used the idea of weak segmentation, which does not require prior labeling, but automatically obtains the segmentation labels through learning. Using this idea, when faced with psoriasis images that are extremely different at the image level, the model can recognize the same attributes or features between them, thus enabling the model to better assess psoriasis. In addition, preprocessing features of skin lesions (e.g. color) is also an aspect that could be considered.Barata et al. (110–112) have shown that image preprocessing techniques (e.g. color constancy) can improve the performance of AI systems for segmentation and classification of skin lesions. Using such techniques, when assessing the severity

of a feature of psoriasis (e.g. erythema), it may be possible to avoid situations where the assessment of erythema is different due to the difference in the psoriasis, if we can first normalize the psoriasis.

4.3.3 Model explainability

Currently, there is an increasing amount of interpretable research in the field of AI in medicine (113). These papers essentially use techniques that are intuitively capable of interpreting the model to enable interpretable research. For example, a class activation map (CAM) (114) is used to visualize the regions of interest of the model, just as Ding et al. (115) used a CAM to direct the model's attention to the lesion region while explaining the model's focus in the middle layer. Concept activation vectors (CAV) (116), a technique that converts high-level concepts that can be understood by humans (e.g., whether or not there are hairs in the area of the lesion, etc.) into vectors that can be understood by a computer. It is therefore feasible to use CAM or CAV to interpret the model. Using CAM, it is possible to understand which areas on the image the model focuses on, and using CAV, it is possible to direct the model's attention to which important high-level concepts. Of course, there are many more interpretable techniques waiting to be discovered, all aimed at increasing physician or patient trust in the model and its outputs.

5 Conclusion

This review provides an overview of the application of ML (especially DL) to psoriasis diagnosis over the last decade, including segmentation, recognition and assessment tasks. However, we have identified a number of challenges in this area, the most important of which are data inconsistency and the issue of data privacy. It is also worth noting that not all DL models are best suited for every task. TML algorithms have also shown good results in feature extraction, and different models should be selected depending on the specific task at hand.

In conclusion, we hope that this review will encourage research in this area and stimulate more advanced techniques to help physicians in their work.

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/s.

Author contributions

HL: Funding acquisition, Supervision, Writing – review & editing. GC: Methodology, Visualization, Writing – original draft. LZ: Conceptualization, Data curation, Methodology, Writing – original draft. CX: Funding acquisition, Supervision, Writing – review & editing. JW: Resources, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was funded by National Natural Science Foundation of China (No. 62006049), Basic and Applied Basic Research Foundation of Guangdong Province (No. 2023A1515010939), and Project of Education Department of Guangdong Province (Nos. 2022KTSCX068 and 2021ZDZX1079).

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

1. Gudjonsson JE, Elder JT. Psoriasis: epidemiology. *Clin Dermatol*. (2007) 25:535–46. doi: 10.1016/j.clindermatol.2007.08.007

2. Habif T. *Psoriasis and Other Papulosquamous Diseases. Clinical? Dermatology 4th ed*. New York: Mosby. (2004). p. 209–45.

3. Griffiths C E M BJNWN. Pathogenesis and clinical features of psoriasis. *Lancet*. (2007) 370:263–71. doi: 10.1016/S0140-6736(07)61128-3

4. Gottlieb AB, Chaudhari U, Baker DG, Perate M, Dooley LT. The National Psoriasis Foundation Psoriasis Score (NPF-PS) system versus the Psoriasis Area Severity Index (PASI) and Physician's Global Assessment (PGA): a comparison. *J Drugs Dermatol*. (2003) 2:260–6.

5. Armstrong AW, Read C. Pathophysiology, clinical presentation, and treatment of psoriasis: a review. *JAMA*. (2020) 323:1945–60. doi: 10.1001/jama.2020.4006

6. Puri P, Comfere N, Drage LA, Shamim H, Bezalel SA, Pittelkow MR, et al. Deep learning for dermatologists: part II current applications. *J Am Acad Dermatol*. (2022) 87:1352–60. doi: 10.1016/j.jaad.2020.05.053

7. Dhivyaa CR, Sangeetha K, Balamurugan M, Amaran S, Vetriselvi T, Johnpaul P. Skin lesion classification using decision trees and random forest algorithms. *J Amb Intellig Human Comp*. (2020) 12:18134. doi: 10.1007/s12652-020-02675-8

8. Muthukannan K, Moses MM. Color image segmentation using k-means clustering and optimal fuzzy C-means clustering. In: *2010 International Conference on Communication and Computational Intelligence (INCOCCI)*. Erode: IEEE. (2010). p. 229–234.

9. Charbuty B, Abdulazeez A. Classification based on decision tree algorithm for machine learning. *J Appl Sci Technol Trends*. (2021) 2:20–8. doi: 10.38094/jastt20165

10. Peterson LE. K-nearest neighbor. *Scholarpedia*. (2009) 4:1883. doi: 10.4249/scholarpedia.1883

11. Pitchiah MS, Rajamanickam T. Efficient feature based melanoma skin image classification using machine learning approaches. *Traitement du Signal*. (2022) 39:24. doi: 10.18280/ts.390524

12. Shen D, Wu G. Deep learning in medical image analysis. *Ann Rev Biomed Eng*. (2017) 19:221–48. doi: 10.1146/annurev-bioeng-071516-044442

13. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Cham: Springer. (2015). p. 234–241. doi: 10.1007/978-3-319-24574-4_28

14. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2016). p. 770–778. doi: 10.1109/CVPR.2016.90

15. Saraf V JA Chavan P. Deep learning challenges in medical imaging. In: *Advanced Computing Technologies and Applications: Proceedings of 2nd International Conference on Advanced Computing Technologies and Applications–ICACTA 2020*. Singapore: Springer (2020). doi: 10.1007/978-981-15-3242-9_28

16. Yu K, Syed MN, Bernardis E, Gelfand JM. Machine learning applications in the evaluation and management of psoriasis: a systematic review. *J Psori Psoriat Arthr*. (2020) 5:147–59. doi: 10.1177/2475530320950267

17. Havelin A, Hampton P. *Telemedicine and e-Health in the Management of Psoriasis: Improving Patient Outcomes-A Narrative Review*. Psoriasis: Targets and Therapy. (2022). p. 15–24.

18. Liu Z, Wang X, Ma Y, Lin Y, Wang G. Artificial intelligence in psoriasis: where we are and where we are going. *Exp Dermatol*. (2023). doi: 10.1111/exd.14938

19. Lunge SB, Shetty NS, Sardesai VR, Karagaiah P, Yamauchi PS, Weinberg JM, et al. Therapeutic application of machine learning in psoriasis: a Prisma systematic review. *J Cosmet Dermatol*. (2023) 22:378–82. doi: 10.1111/jocd.15122

20. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev*. (2015) 4:1–9. doi: 10.1186/2046-4053-4-1

21. George Y, Aldeen M, Garnavi R. Automatic psoriasis lesion segmentation in two-dimensional skin images using multiscale superpixel clustering. *J Med Imag*. (2017) 4:044004–044004. doi: 10.1117/1.JMI.4.4.044004

22. Dash M, Londhe ND, Ghosh S, Semwal A, Sonawane RS. PsLSNet: automated psoriasis skin lesion segmentation using modified U-Net-based fully convolutional network. *Biomed Signal Process Control*. (2019) 52:226–37. doi: 10.1016/j.bspc.2019.04.002

23. Shrivastava VK, Londhe ND, Sonawane RS, Suri JS. Computer-aided diagnosis of psoriasis skin images with HOS, texture and color features: a first comparative study of its kind. *Comput Methods Programs Biomed*. (2016) 126:98–109. doi: 10.1016/j.cmpb.2015.11.013

24. Zhao S, Xie B, Li Y. Zhao Xy, Kuang Y, Su J, et al. Smart identification of psoriasis by images using convolutional neural networks: a case study in China. *J Eur Acad Dermatol Venereol*. (2020) 34:518–24. doi: 10.1111/jdv.15965

25. Hammad M, Pławiak P, ElAffendi M, El-Latif AAA, Latif AAA. Enhanced deep learning approach for accurate eczema and psoriasis skin detection. *Sensors*. (2023) 23:7295. doi: 10.3390/s23167295

26. Shrivastava VK, Londhe ND, Sonawane RS, Suri JS. A novel approach to multiclass psoriasis disease risk stratification: machine learning paradigm. *Biomed Signal Process Control*. (2016) 28:27–40. doi: 10.1016/j.bspc.2016.04.001

27. Shrivastava VK, Londhe ND, Sonawane RS, Suri JS. A novel and robust Bayesian approach for segmentation of psoriasis lesions and its risk stratification. *Comput Methods Programs Biomed*. (2017) 150:9–22. doi: 10.1016/j.cmpb.2017.07.011

28. Dash M, Londhe ND, Ghosh S, Raj R, Sonawane RS. A cascaded deep convolution neural network based CADx system for psoriasis lesion segmentation and severity assessment. *Appl Soft Comput*. (2020) 91:106240. doi: 10.1016/j.asoc.2020.106240

29. Xie B, He X, Zhao S, Li Y, Su J, Zhao X, et al. XiangyaDerm: a clinical image dataset of Asian race for skin disease aided diagnosis.In: *LABELS 2019, HAL-MICCAI 2019, CuRIOUS 2019*. (2019).

30. Library DI. *DermNet NZ*. (2019). Available online at: https://www.dermnetnz.org/image-library/ (accessed April 9, 2024).

31. da Silva SF. *Dermatology Atlas*. (2019). Available online at: http://www.atlasdermatologico.com.br/ (accessed April 9, 2024).

32. Verros CD. *Hellenic Dermatology Atlas*. (2019). Available online at: http://www.hellenicdermatlas.com/en/ (accessed April 9, 2024).

33. Yan S, Yu Z, Zhang X, Mahapatra D, Chandra SS, Janda M, et al. Towards trustable skin cancer diagnosis via rewriting model's decision. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2023). p. 11568–11577.

34. Mirikharaji Z, Abhishek K, Bissoto A, Barata C, Avila S, Valle E, et al. A survey on deep learning for skin lesion segmentation. *Med Image Anal*. (2023) 88:102863. doi: 10.1016/j.media.2023.102863

35. Li H, He X, Zhou F, Yu Z, Ni D, Chen S, et al. Dense deconvolutional network for skin lesion segmentation. *IEEE J Biomed Health Informat*. (2019) 23:527–37. doi: 10.1109/JBHI.2018.2859898

36. Dash M, Londhe ND, Ghosh S, Shrivastava VK, Sonawane RS. Swarm intelligence based clustering technique for automated lesion detection and diagnosis of psoriasis. *Comput Biol Chem*. (2020) 86:107247. doi: 10.1016/j.compbiolchem.2020.107247

37. Raj R, Londhe ND, Sonawane RS. Automatic psoriasis lesion segmentation from raw color images using deep learning. In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Seoul: IEEE. (2020). p. 723–728.

38. Raj R, Londhe ND, Sonawane R. Automated psoriasis lesion segmentation from unconstrained environment using residual U-Net with transfer learning. *Comput Methods Programs Biomed*. (2021) 206:106123. doi: 10.1016/j.cmpb.2021.106123

39. Lin YL, Huang A, Yang CY, Chang WY. Measurement of body surface area for psoriasis using U-net models. *Comp Mathem Methods Med*. (2022) 2022:7960151. doi: 10.1155/2022/7960151

40. Czajkowska J, Badura P, Korzekwa S, Płatkowska-Szczerek A. Automated segmentation of epidermis in high-frequency ultrasound of pathological skin using a cascade of DeepLab v3+ networks and fuzzy connectedness. *Comp Mathem Methods Med*. (2022) 95:102023. doi: 10.1016/j.compmedimag.2021.102023

41. Lin GS, Lai KT Syu JM, Lin JY, Chai SK. Instance segmentation based on deep convolutional neural networks and transfer learning for unconstrained psoriasis skin images. *Appl Sci*. (2021) 11:3155. doi: 10.3390/app11073155

42. Mohan S, Kasthuri N. Automatic segmentation of psoriasis skin images using adaptive chimp optimization algorithm-based CNN. *J Digit Imaging*. (2023) 36:1123–36. doi: 10.1007/s10278-022-00765-x

43. Panneerselvam K, Nayudu PP. Improved golden eagle optimization based CNN for automatic segmentation of psoriasis skin images. *Wireless Pers Commun*. (2023) 131:1817–31. doi: 10.1007/s11277-023-10522-0

44. Dice LR. Measures of the amount of ecologic association between species. *Ecology*. (1945) 26:297–302. doi: 10.2307/1932409

45. Al-Masni MA, Al-Antari MA, Choi MT, Han SM, Kim TS. Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. *Comput Methods Programs Biomed*. (2018) 162:221–31. doi: 10.1016/j.cmpb.2018.05.027

46. Niwattanakul S, Singthongchai J, Naenudorn E, Wanapu S. Using of Jaccard coefficient for keywords similarity. In: *The 2013 IAENG International Conference on Internet Computing and Web Services (ICICWS'13)*. Hong Kong (2013). p. 380–4.

47. Liu L, Mou L, Zhu XX, Mandal M. Skin Lesion Segmentation Based on Improved U-net. In: *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*. Edmonton, AB: IEEE (2019). p. 1–4.

48. Wang Y, Sun S, Yu J, Yu DL. Skin lesion segmentation using atrous convolution via DeepLab V3. *arXiv [Preprint]*. arXiv:180708891. (2018). doi: 10.48550/arXiv.1807.08891

49. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: *Proceedings of the IEEE International Conference on Computer Vision*. Venice: IEEE. (2017). p. 2961–2969.

50. Foraker RE, Kite B, Kelley MM, Lai AM, Roth C, Lopetegui MA, et al. EHR-based visualization tool: adoption rates, satisfaction, and patient outcomes. *eGEMs*. (2015) 3:1159. doi: 10.13063/2327-9214.1159

51. Wei Ls, Gan Q, Ji T, et al. Skin disease recognition method based on image color and texture features. *Comp Mathem Methods Med*. (2018) 2018:8145713. doi: 10.1155/2018/8145713

52. Yu Z, Kaizhi S, Jianwen H, Guanyu Y, Yonggang W, A. deep learning-based approach toward differentiating scalp psoriasis and seborrheic dermatitis from dermoscopic images. *Front Med*. (2022) 9:965423. doi: 10.3389/fmed.2022.965423

53. Nieniewski M, Chmielewski LJ, Patrzyk S, Woźniacka A. Studies in differentiating psoriasis from other dermatoses using small data set and transfer learning EURASIP. *J Image Video Proc*. (2023) 2023:7. doi: 10.1186/s13640-023-00607-y

54. Singh A, KC NK, Kumar MA, Negi HS. An improved deep learning framework approach for detecting the psoriasis disease. In: *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*. Chennai: IEEE. (2023). p. 1–5.

55. Ji H, Li J, Zhu X, Fan L, Jiang W, Chen Y. Enhancing assisted diagnostic accuracy in scalp psoriasis: a Multi-Network Fusion Object Detection Framework for dermoscopic pattern diagnosis. *Skin Res Technol*. (2024) 30:e13698. doi: 10.1111/srt.13698

56. Yaseliani M, Maghsoodi AI, Hassannayebi E, Aickelin U. Diagnostic clinical decision support based on deep learning and knowledge-based systems for psoriasis: from diagnosis to treatment options. *Comp Indust Eng*. (2024) 187:109754. doi: 10.1016/j.cie.2023.109754

57. Aggarwal SLP. Data augmentation in dermatology image recognition using machine learning. *Skin Res Technol*. (2019) 25:815–20. doi: 10.1111/srt.12726

58. Wu H, Yin H, Chen H, Sun M, Liu X, Yu Y, et al. A deep learning, image based approach for automated diagnosis for inflammatory skin diseases. *Ann Transl Med*. (2020) 8:39. doi: 10.21037/atm.2020.04.39

59. Yang Y, Wang J, Xie F, Liu J, Shu C, Wang Y, et al. A convolutional neural network trained with dermoscopic images of psoriasis performed on par with 230 dermatologists. *Comput Biol Med*. (2021) 139:104924. doi: 10.1016/j.compbiomed.2021.104924

60. Chatterjee A. Optimization of image recognition for the identification of psoriasis and eczema. In: *Applications of Machine Learning*. San Diego, CA: SPIE Digital Library (2022). p. 213–7.

61. Zhu W, Lai H, Zhang H, Zhang G, Luo Y, Wang J, et al. Abscissa-ordinate focused network for psoriasis and eczema healthcare cyber-physical system with active label smoothing. *IEEE Access*. (2024). doi: 10.1109/ACCESS.2024.3384310

62. Zhu X, Zheng B, Cai W, Zhang J, Lu S, Li X, et al. Deep learning-based diagnosis models for onychomycosis in dermoscopy. *Mycoses*. (2022) 65:466–72. doi: 10.1111/myc.13427

63. Arunkumar T, Jayanna H. A novel light weight approach for identification of psoriasis affected skin lesion using deep learning. *J Phys*. 2062:012017. doi: 10.1088/1742-6596/2062/1/012017

64. Vishwakarma G, Nandanwar AK, Thakur GS. Optimized vision transformer encoder with cnn for automatic psoriasis disease detection. *Multimedia Tools Appl*. (2023) 2023:1–20. doi: 10.1007/s11042-023-16871-z

65. Peng L, Na Y, Changsong D, Sheng L, Hui M. Research on classification diagnosis model of psoriasis based on deep residual network. *Digi Chin Med*. (2021) 4:92–101. doi: 10.1016/j.dcmed.2021.06.003

66. Goswami A, Singh S, Tarekar P, Sharma N. Intra-class classification of psoriasis using deep learning. *J Biomed Eng Soc India*. (2023) 2023:14.

67. Aijaz SF, Khan SJ, Azim F, Shakeel CS, Hassan U. Deep learning application for effective classification of different types of psoriasis. *J Healthc Eng*. (2022) 2022:7541583. doi: 10.1155/2022/7541583

68. Rashid MS, Gilanie G, Naveed S, Cheema S, Sajid M. Automated detection and classification of psoriasis types using deep neural networks from dermatology images. *Signal Image Video Proc*. (2024) 18:163–72. doi: 10.1007/s11760-023-02722-9

69. Kalyan K, Jakhia B, Lele RD, Joshi M, Chowdhary A. Artificial neural network application in the diagnosis of disease conditions with liver ultrasound images. *Adv Bioinformatics*. (2014) 2014:708279. doi: 10.1155/2014/708279

70. Maćkiewicz A, Ratajczak W. Principal components analysis (PCA). *Comp Geosci*. (1993) 19:303–42. doi: 10.1016/0098-3004(93)90090-R

71. Chua KC, Chandran V, Acharya UR, Lim CM. Application of higher order statistics/spectra in biomedical signals–A review. *Med Eng Phys*. (2010) 32:679–89. doi: 10.1016/j.medengphy.2010.04.009

72. Codella N, Rotemberg V, Tschandl P, Celebi ME, Dusza S, Gutman D, et al. Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (isic). *arXiv [Preprint]*. arXiv:190203368. (2019). doi: 10.48550/arXiv.1902.03368

73. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. (1997) 9:1735–80. doi: 10.1162/neco.1997.9.8.1735

74. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. *arXiv [Preprint]*. arXiv:170603762. (2017).

75. Mattei P L KAB. Corey K C. Psoriasis Area Severity Index (PASI) and the Dermatology Life Quality Index (DLQI): the correlation between disease severity and psychological burden in patients treated with biological therapies. *J Eur Acad Dermatol Venereol*. (2014) 28:333–7. doi: 10.1111/jdv.12106

76. Banu S, Toacse G, Danciu G. Objective erythema assessment of Psoriasis lesions for Psoriasis Area and Severity Index (PASI) evaluation. In: *2014 International Conference and Exposition on Electrical and Power Engineering (EPE)*. Iasi: IEEE. (2014). p. 052–056.

77. George Y, Aldeen M, Garnavi R. Psoriasis image representation using patch-based dictionary learning for erythema severity scoring. *Comp Med Imag Graph*. (2018) 66:44–55. doi: 10.1016/j.compmedimag.2018.02.004

78. George Y, Aldeen M, Garnavi R. Automatic scale severity assessment method in psoriasis skin images using local descriptors. *IEEE J Biomed Health Informat*. (2019) 24:577–85. doi: 10.1109/JBHI.2019.2910883

79. Moon CI, Lee J, Yoo H, Baek Y, Lee O. Optimization of psoriasis assessment system based on patch images. *Sci Rep*. (2021) 11:18130. doi: 10.1038/s41598-021-97211-9

80. Ji B, Wang Y, Zuo D. Automatic detection and evaluation of nail psoriasis based on deep learning: A preliminary application and exploration. In: *International Conference on Computer Application and Information Security (ICCAIS 2021)*. Bellingham: SPIE (2022). p. 311–317.

81. Hsieh KY, Chen HY, Kim SC, Tsai YJ, Chiu HY, Chen GY, et al. mask R-CNN based automatic assessment system for nail psoriasis severity. *Comput Biol Med*. (2022) 143:105300. doi: 10.1016/j.compbiomed.2022.105300

82. Folle L, Fenzl P, Fagni F, Thies M, Christlein V, Meder C, et al. DeepNAPSI multi-reader nail psoriasis prediction using deep learning. *Sci Rep*. (2023) 13:5329. doi: 10.1038/s41598-023-32440-8

83. Paik K, Kim BR, Youn SW. Evaluation of the area subscore of the Palmoplantar Pustulosis Area and Severity Index using an attention U-net deep learning algorithm. *J Dermatol*. (2023). doi: 10.1111/1346-8138.16752

84. Raj R, Londhe ND, Sonawane R. PsLSNetV2: End to end deep learning system for measurement of area score of psoriasis regions in color images. *Biomed Signal Process Control*. (2023) 79:104138. doi: 10.1016/j.bspc.2022.104138

85. Raj R, Londhe ND, Sonawane RS. Objective scoring of psoriasis area and severity index in 2D RGB images using deep learning. *Multimedia Tools Appl*. (2024) 2024:1–27. doi: 10.1007/s11042-024-18138-7

86. Okamoto T, Kawai M, Ogawa Y, Shimada S, Kawamura T. Artificial intelligence for the automated single-shot assessment of psoriasis severity. *J Eur Acad Dermatol Venereol*. (2022) 36:2512–5. doi: 10.1111/jdv.18354

87. Schaap MJ, Cardozo NJ, Patel A, de Jong EMGJ, van Ginneken B, Seyger MMB. Image-based automated Psoriasis Area Severity Index scoring by Convolutional Neural Networks. *J Eur Acad Dermatol Venereol*. (2022) 36:68–75. doi: 10.1111/jdv.17711

88. Huang K, Wu X, Li Y, Lv C, Yan Y, Wu Z, et al. Artificial intelligence-based psoriasis severity assessment: real-world study and application. *J Med Internet Res*. (2023) 25:e44932. doi: 10.2196/44932

89. Arunkumar T, Jayanna H, A. Machine learning approach for the estimation of severity of psoriasis disorder using depth-wise convolution neural network. *Indian J Sci Technol*. (2023) 16:318–30. doi: 10.17485/IJST/v16i5.1723

90. Moon CI, Kim EB, Baek YS, Lee O. Transformer based on the prediction of psoriasis severity treatment response. *Biomed Signal Process Control*. (2024) 89:105743. doi: 10.1016/j.bspc.2023.105743

91. Xing Y, Zhong S, Aronson SL, Rausa FM, Webster DE, Crouthamel MH, et al. Deep learning-based psoriasis assessment: harnessing clinical trial imaging for accurate psoriasis area severity index prediction. *Digit Biomark*. (2024) 8:13–21. doi: 10.1159/000536499

92. Moon CI, Lee J, Baek YS, Lee O. Psoriasis severity classification based on adaptive multi-scale features for multi-severity disease. *Sci Rep*. (2023) 13:17331. doi: 10.1038/s41598-023-44478-9

93. Yin H, Chen H, Zhang W, Zhang J, Cui T, Li Y, et al. Image-based remote evaluation of PASI scores with psoriasis by the YOLO-v4 algorithm. *Exp Dermatol*. (2024) 33:e15082. doi: 10.1111/exd.15082

94. Rodtook A, Kirimasthong K, Lohitvisate W, Makhanov SS. Automatic initialization of active contours and level set method in ultrasound images of breast abnormalities. *Pattern Recognit*. (2018) 79:172–82. doi: 10.1016/j.patcog.2018.01.032

95. Badrinarayanan V, Kendall A, Cipolla R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell*. (2017) 39:2481–95. doi: 10.1109/TPAMI.2016.2644615

96. Cao W, Mirjalili V, Raschka S. Consistent rank logits for ordinal regression with convolutional neural networks. *arXiv [Preprint]*. arXiv:190107884. (2019). doi: 10.1016/j.patrec.2020.11.008

97. Redmon J, Divvala S, Girshick R, Farhadi A. 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. (2016). p. 779–788.

98. Du G, Cao X, Liang J, Chen X, Zhan Y. Medical image segmentation based on U-Net: a review. *J Imag Sci Technol*. (2020) 64:020508. doi: 10.2352/J.ImagingSci.Technol.2020.64.2.020508

99. Cai L, Gao J, Zhao D. A review of the application of deep learning in medical image classification and segmentation. *Ann Transl Med*. (2020) 8:44. doi: 10.21037/atm.2020.02.44

100. Tatsunami Y, Taki M. Sequencer: Deep lstm for image classification. *Adv Neural Inf Process Syst*. (2022) 35:38204–17.

101. Zhou ZH. Learnware: on the future of machine learning. *Front Comput Sci*. (2016) 10:589–90. doi: 10.1007/s11704-016-6906-3

102. Shi X, Dou Q, Xue C, Qin J, Chen H, Heng PA. An active learning approach for reducing annotation cost in skin lesion analysis. In: *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10*. Cham: Springer. (2019). p. 628–636.

103. Katragadda C, Finnane A, Soyer HP, Marghoob AA, Halpern A, Malvehy J, et al. Technique standards for skin lesion imaging: a delphi consensus statement. *JAMA Dermatol*. (2017) 2016:3949. doi: 10.1001/jamadermatol.2016.3949

104. Xiao J, Xu H, Zhao W, Cheng C, Gao H. A prior-mask-guided few-shot learning for skin lesion segmentation. *Computing*. (2021) 2021:1–23. doi: 10.1007/s00607-021-00907-z

105. Liu XJ, Li Kl, Luan Hy, Wang Wh, Chen Zy. Few-shot learning for skin lesion image classification. *Multimedia Tools Appl*. (2022) 81:4979–90. doi: 10.1007/s11042-021-11472-0

106. Chen B, Han Y, Yan L, A. Few-shot learning approach for Monkeypox recognition from a cross-domain perspective. *J Biomed Inform*. (2023) 144:104449. doi: 10.1016/j.jbi.2023.104449

107. Xiao J, Xu H, Fang D, Cheng C, Gao H. Boosting and rectifying few-shot learning prototype network for skin lesion classification based on the internet of medical things. *Wireless Netw*. (2023) 29:1507–21. doi: 10.1007/978-3-031-32138-2

108. Díaz IG. Incorporating the knowledge of dermatologists to convolutional neural networks for the diagnosis of skin lesions. *arXiv [Preprint]*. arXiv:170301976. 2017;.

109. Pathak D, Krahenbuhl P, Darrell T. Constrained convolutional neural networks for weakly supervised segmentation. In: *Proceedings of the IEEE International Conference on Computer Vision*. (2015). p. 1796–1804. doi: 10.1109/ICCV.2015.209

110. Barata C, Celebi ME, Marques JS. Improving dermoscopy image classification using color constancy. *IEEE J Biomed Health Informat*. (2014) 19:1146–52. doi: 10.1109/ICIP.2014.7025716

111. Salvi M, Branciforti F, Veronese F, Zavattaro E, Tarantino V, Savoia P, et al. DermoCC-GAN: A new approach for standardizing dermatological images using generative adversarial networks. *Comput Methods Programs Biomed*. (2022) 225:107040. doi: 10.1016/j.cmpb.2022.107040

112. Salvi M, Branciforti F, Molinari F, Meiburger KM. Generative models for color normalization in digital pathology and dermatology: advancing the learning paradigm. *Expert Syst Appl*. (2024) 245:123105. doi: 10.1016/j.eswa.2023.123105

113. Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural networks for medical image analysis: a review of interpretability methods. *Comput Biol Med*. (2022) 140:105111. doi: 10.1016/j.compbiomed.2021.105111

114. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. *IEEE Comp Soc*. (2016). doi: 10.1109/CVPR.2016.319

115. Ding S, Wu Z, Zheng Y, Liu Z, Yang X, Yang X, et al. Deep attention branch networks for skin lesion classification. *Comput Methods Programs Biomed*. (2021) 212:106447. doi: 10.1016/j.cmpb.2021.106447

116. Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: *International Conference on Machine Learning*. New York: PMLR. (2018). p. 2668–2677.

frontiers | Frontiers in Oncology

TYPE Case Report
PUBLISHED 24 November 2025
DOI 10.3389/fonc.2025.1551792

OPEN ACCESS

EDITED BY
Monica Bianchini,
University of Siena, Italy

REVIEWED BY
Vishakha Mahajan,
The University of Auckland, New Zealand
Ahmed Abdlkadir,
King Hussein Cancer Center, Jordan

*CORRESPONDENCE
Min Kang
drkm0327@163.com
Donghui Huang
13600001163@139.com

†These authors have contributed equally to this work and share first authorship

RECEIVED 26 December 2024
REVISED 25 October 2025
ACCEPTED 10 November 2025
PUBLISHED 24 November 2025

CITATION
Chen H, Shang X, Shen Y, Huang H, Jiang Z, Wang Q, Cao Z, Yan P, Xiao S, Chen L, Huang D and Kang M (2025) High-intensity focused ultrasound as a combined approach for the treatment of recurrent low-grade endometrial stromal sarcoma: a case report and literature review.
Front. Oncol. 15:1551792.
doi: 10.3389/fonc.2025.1551792

High-intensity focused ultrasound as a combined approach for the treatment of recurrent low-grade endometrial stromal sarcoma: a case report and literature review

Huihui Chen[1,2†], Xiaonan Shang[1†], Yue Shen[1†], Huajing Huang[3], Zhebo Jiang[2], Qingyi Wang[2], Zhixing Cao[4], Peiyu Yan[1,5,6], Suying Xiao[2], Liangyu Chen[2], Donghui Huang[2*] and Min Kang[1,2*]

[1]Faculty of Chinese Medicine, Macau University of Science and Technology, Macao, Macao SAR, China, [2]Zhuhai Hospital of Integrated Traditional Chinese Western Medicine, Zhuhai, Guangdong, China, [3]Northeastern University, Boston, MA, United States, [4]Zhuhai People's Hospital, Zhuhai, Guangdong, China, [5]State Key Laboratory of Quality Research in Chinese Medicines, Macao, Macao SAR, China, [6]Macau University of Science and Technology Zhuhai MUST Science and Technology Research Institute, Macao, Macao SAR, China

Background: Surgery is the primary treatment for Endometrial Stromal Sarcoma (ESS), however, a substantial proportion of patients with ESS experience recurrence or metastasis. Currently, surgery and local ablation are the main treatments for recurrent ESS followed by chemotherapy, radiotherapy, immunotherapy, targeted therapy, and anti-estrogen therapy. Surgery and local ablation are invasive treatments and may carry risks such as intestinal damage and the risk of massive bleeding from tumor rupture. For patients who refuse or are unable to undergo surgery and local ablation, conservative treatment is not effective, and there is currently no definitive effective non-invasive or combined treatment plan.

Case presentation: This report presents a case of a patient with recurrent endometrial stromal sarcoma who refused surgical and local ablation treatments. After receiving HIFU treatment combined with chemotherapy, the progression of the tumor was effectively inhibited, the tumor volume significantly reduced, and liver function was restored during the HIFU period, providing an opportunity for chemotherapy.

Conclusions: HIFU combined with chemotherapy may provide a new treatment strategy for patients with recurrent, metastatic endometrial stromal sarcoma, or those who are unsuitable for surgery, local ablation, or those with poor baseline status unable to tolerate intensive chemotherapy.

KEYWORDS

LGESS, HIFU, tumor recurrence, combination therapy, case report

Introduction

ESS is an invasive tumor originating from endometrial stromal cells. The cells resemble proliferative phase endometrial stromal cells, manifesting as infiltrative growth, with or without lymphovascular invasion. It accounts for approximately 0.2-1% of uterine malignancies and 6-20% of uterine sarcomas (1–3). According to the WHO (2020 edition) classification of gynecological malignancies, ESS is divided into Low-Grade Endometrial Stromal Sarcoma (LGESS) and High-Grade Endometrial Stromal Sarcoma (HGESS) (4).

Due to the lack of specific clinical and radiographic manifestations, ESS is easily misdiagnosed as uterine fibroids or adenomyosis with similar symptoms (5). Therefore, a thorough evaluation must be performed on rapidly enlarging fibroid masses before surgery. High-grade stromal sarcoma carries a poor prognosis, especially when diagnosis is delayed or presented with advanced stages (6). LGESS is typically discovered during pathological examination of hysterectomy specimens (7). LGESS is a relatively indolent tumor with a good overall survival rate, but it is characterized by multiple or late recurrences (3, 8). Recurrence is more common in the pelvic and abdominal cavities, and less common in the lungs and vagina. Due to its indolent course, distant recurrence is more frequently seen in clinical practice, necessitating long-term follow-up, hence there is less research on the prognosis of recurrent LGESS (9).

Currently, hysterectomy and bilateral salpingo-oophorectomy are the first-line treatments for ESS. However, approximately 30%-50% of ESS patients experience recurrence or metastasis (10). At present, surgical treatment, anti-estrogen therapy, chemotherapy, radiotherapy, and targeted drug therapy are used to treat recurrent or metastatic ESS. However, due to the different pathological characteristics and fewer cases, there is not enough research and data, and the treatment plan for recurrent metastatic ESS is still not clearly unified.

In terms of examinations and follow-up, MRI differentiates uterine fibroids from sarcomas through its superior soft-tissue resolution, while monitoring tumor volume changes and therapeutic effects. PET-CT precisely identifies metastases or recurrent lesions, yet its phased utilization is prioritized in clinical practice due to cost and procedural constraints. MRI serving as the foundational modality, while PET-CT provides targeted assistance.

We report a case of recurrent low-grade endometrial stromal sarcoma with multiple pelvic metastases and right sacral bone metastasis. The patient had a short-term recurrence after surgery and underwent multiple rounds of combined radiochemotherapy and regular follow-up. Three years later, the patient relapsed again. After hospital evaluation, the patient was unwilling to undergo a second surgery due to concerns about surgical risks. The patient then received three cycles of chemotherapy. After chemotherapy, the patient developed abnormal liver function. After discussion by the doctors, the treatment plan was changed to HIFU and chemotherapy. This effectively inhibited tumor progression with significant results.

Case report

The patient is a 28-year-old unmarried and nulliparous female with no family history of malignancy and no prior gynecological disorders or estrogen-related medication use. She presented to the hospital in November 2019 with progressively worsening dysmenorrhea for one year and menorrhagia for six months. Gynecological ultrasound and abdominal CT indicated an enlarged uterus, suggesting uterine fibroids. On November 21, 2019, she underwent laparoscopic exploration. During the operation, a tumor approximately 9*9*8cm in size was seen on the posterior wall of the uterus, and another tumor approximately 5×4cm in size was seen on the lower segment of the posterior wall of the uterus. The intraoperative frozen pathology diagnosis was a mesenchymal malignant tumor. With the consent of the family, the operation was changed to total hysterectomy, bilateral adnexectomy, and omentectomy. Postoperative pathology and immunohistochemistry indicated low-grade endometrial stromal sarcoma with transformation to high-grade, local necrosis, enlarged and round nuclei, invasion of the uterine myometrium, involvement of the endometrium and serosal layer, tumor invasion seen in the blood vessels, no tumor invasion seen in the nerves, and no tumor seen in the bilateral adnexa and omentum (Figure 1). The postoperative pathological stage was stage IB. After the operation, she underwent three rounds of intraperitoneal hyperthermic perfusion therapy (cisplatin 110mg).

On December 17, 2019, the patient's follow-up 18F-FDG PET/CT (18F-fluorodeoxyglucose positron emission computed tomography/computed tomography) showed thickening of vaginal soft tissue with increased glucose metabolism, multiple pelvic lymph nodes with increased glucose metabolism, suggesting metastasis. The right side of the sacrum showed slightly increased bone density with increased glucose metabolism, suggesting possible metastasis. On December 23, 2019, an enhanced whole abdomen MR suggested a nodular lesion on the left margin of the vaginal stump, highly suspicious of tumor; multiple lymph nodes near bilateral iliac vessels, on both sides of the pelvis, and in the pre-sacral space, lymph node metastasis could not be excluded. The preliminary diagnosis was "vaginal recurrence of endometrial stromal sarcoma and sacral metastasis". From January 9 to January 20, 2020, the patient underwent VMAT radiotherapy (dose: GTVnd 6000cGy, CTV4500cGy). From February 28 to March 12, 2020, the patient underwent 4 sessions of brachytherapy (dose: 28Gy/4f, cisplatin as a radiosensitizer). On January 9 and January 20, 2020, she received concurrent chemotherapy with cisplatin (dose: 25mg, d1-4). On February 21, 2020, she accepted chemotherapy with cisplatin (100mg) and nivolumab (200mg) and regorafenib capsules (20mg, Bid). On March 19, 2020, she accepted a cycle of chemotherapy with paclitaxel (300mg) and lobaplatin (150mg) and nivolumab (200mg).From April 15 to June 3, 2020, she continued to receive 3 cycles of chemotherapy with lobaplatin (150mg) and paclitaxel (330mg) and bevacizumab (350mg). On April 15, 2020, she underwent a biopsy of the vaginal lesion, and the

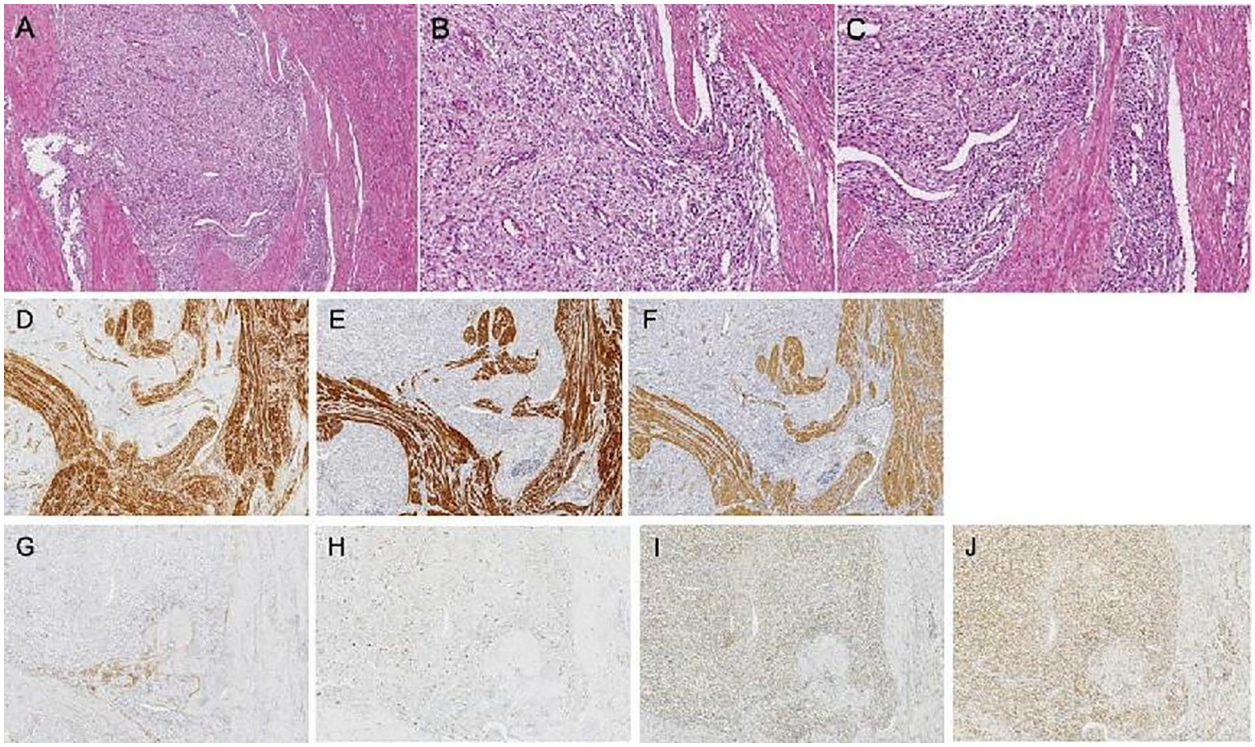


FIGURE 1
Tumor histopathology images [**(A-C)** Hematoxylin-eosin staining; magnification: **(A)** 4×; **(B, C)** 10×]. Main immunohistochemical staining results for low-grade endometrial stromal sarcoma [**(D–J)** 4×]. **(D)** Caldesmon; **(E)** Desmin; **(F)** SMA; **(G)** CD10; **(H)** Ki-67; **(I)** ER; **(J)** PR.

pathology results indicated: (vaginal orifice nodule) no endometrial stromal sarcoma seen. Serial MRI scans performed every three months between January and June 2020 revealed no abnormalities. Following the completion of chemotherapy, the patient underwent PET/CT scans every six months, with no significant abnormalities detected in the results.

On August 18, 2023, the patient experienced pain in the lower left abdomen, which gradually worsened, accompanied by left-sided back pain and fever. Outpatient ultrasound examination of the urinary system suggested: dilation of the upper segment of the left ureter with hydronephrosis of the left kidney, and a hypoechoic mass behind the bladder, measuring approximately 88×74×88mm, with clear boundaries and uneven internal echo. On August 24, 2023, a PET/CT scan showed a mass of approximately 87×83×90mm at the vaginal stump, suggesting a possible recurrence of the tumor.

The patient was admitted to the hospital for treatment on August 28, 2023, and underwent enhanced abdominal MR and urinary CTU examinations. The MR enhancement suggested an abnormal signal in the pelvic cavity, measuring approximately 99mm×88mm×116mm, suggesting local tumor recurrence, possibly involving the rectum, colon, bladder, and left ureter. After pelvic metastasis, the patient's primary symptoms included left-sided lumbar soreness, abdominal distension, and lower abdominal pain. Physical examination revealed a pelvic mass measuring approximately 9 cm×8 cm on triple examination, with a firm consistency, poor mobility, no significant tenderness, and no percussion tenderness over the sacrococcygeal region. After multidisciplinary consultation, the patient was informed of the high risk of surgery, including potential intestinal and bladder injury, and the possibility of performing intestinal and renal fistula surgery, ablation therapy may carry risks of tumor rupture and bleeding, and injury to the intestines and bladder. The patient strongly refused surgery and ablation therapy, requesting conservative treatment. After ruling out contraindications to chemotherapy, the patient underwent three cycles of systemic chemotherapy with the TC regimen (paclitaxel injection 260mg + carboplatin injection 500mg) on September 5, September 26, and October 23, 2023. The tumor size decreased from 99mm×88mm×116mm to 69.2mm×57.8mm×75.4mm (Figures 2A, B).

On November 12, 2023, the patient's liver function showed significant abnormalities (Alanine transaminase (ALT): 47.7 U/L; Aspartic amino transferase (AST): 32.8 U/L;CTCAE version 5.0: Grade 1 hepatotoxicity) and she could not receive the fourth cycle of chemotherapy as scheduled. After discussion and with the patient's consent, the treatment plan was changed to HIFU treatment and liver protection treatment, waiting for the opportunity for chemotherapy. From November 13 to November 24, 2023, the patient underwent nine intermittent HIFU treatments, after which the blood flow in the pelvic tumor significantly decreased. During this period, the patient was given liver protection treatment (silymarin capsules 140mg, bid, orally), and glutathione (1.2g, qd, intravenous infusion). On December 12, 2023, the patient's liver function recovered, and she underwent the fourth cycle of systemic chemotherapy with the TC regimen (paclitaxel injection 260mg + carboplatin injection 600mg). On January 5, 2024, the patient's liver

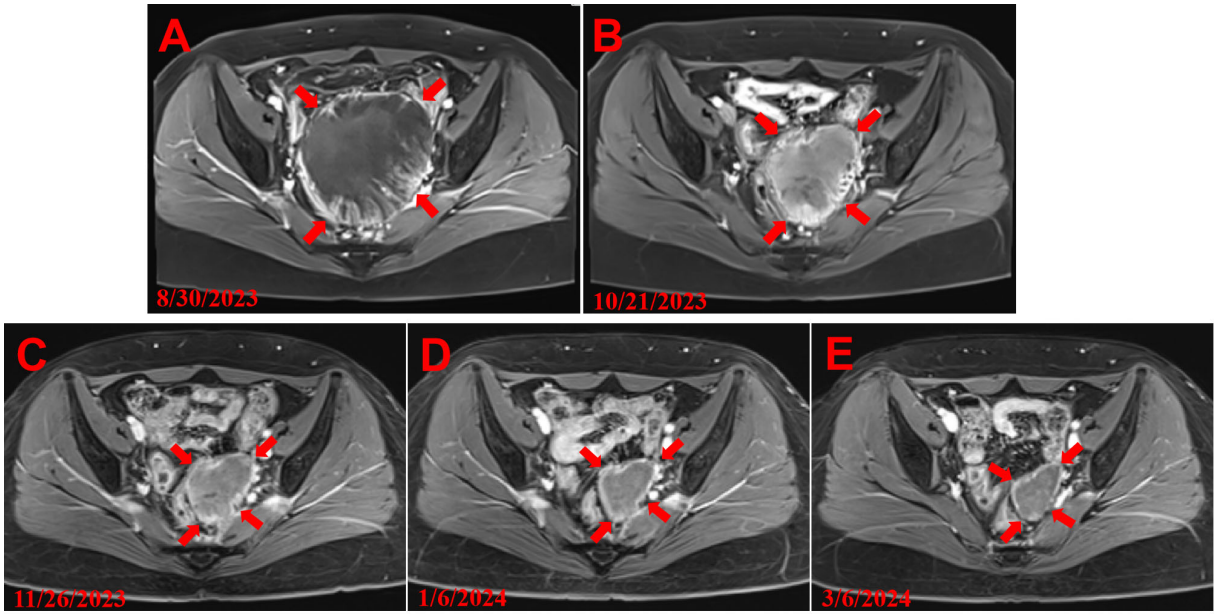


FIGURE 2
MRI images of the patient. **(A)** Tumor volume size of the patient's first MRI after recurrence. **(B)** Tumor volume size before the first HIFU treatment after three courses of chemotherapy with TC regimen. **(C)** Tumor volume size after first HIFU treatment. **(D)** Tumor volume size after the fourth course of TC regimen chemotherapy. **(E)** Tumor volume size after the second HIFU treatment.

function was abnormal again (Alanine transaminase (ALT): 118.6 U/L;Aspartic amino transferase (AST):41.6 U/L;CTCAE version 5.0: Grade 1 hepatotoxicity) and she could not receive the fifth cycle of chemotherapy as scheduled. From January 8 to January 16, 2024, the patient received eight intermittent HIFU treatments, and liver protection treatment was continued during the treatment period. On February 2, 2024, the patient's liver function recovered, and she underwent the fifth cycle of systemic chemotherapy with the TC regimen (paclitaxel injection 270mg + carboplatin injection 780mg). After 17 HIFU treatments combined with chemotherapy, the patient's lesion decreased from 69.2mm×57.8mm×75.4mm to 43mm×33mm×45mm (Figures 2C–E). The scattered small nodules in the original pelvic cavity disappeared, the dilation of the upper segment of the original left ureter improved significantly, the turbidity of the fat space in the original pelvic cavity and the pelvic effusion disappeared. The edema of the left piriformis muscle significantly improved. The level of tumor markers gradually decreased and tended to stabilize. The patient's abdominal pain and bloating symptoms disappeared, and she had no other discomfort. On March 8, 2024, she underwent the sixth cycle of systemic chemotherapy with the TC regimen (paclitaxel injection 270mg + carboplatin injection 650mg).

On April 22, 2024, a PET/CT scan suggested that the blood flow signal around the patient's pelvic mass had significantly decreased, the mass had basically shown changes after HIFU treatment (Figure 3), and the patient's tumor markers (Figure 4) had steadily decreased and trended toward stabilization. The treatment effect was satisfactory. The patient was advised to undergo surgical treatment, but she still refused. The benefit of immunotherapy for the patient was not evident at present. The patient requested regular follow-up, and there were no new lesions at present. It is recommended to continue regular HIFU maintenance treatment in the future. The patient is currently under continued follow-up observation. The disease timeline is shown in Figure 5.

Methods

In this case, the patient used the yLab Class C Ultrasound Diagnostic System (Shenzhen Baisheng Medical Equipment Co., Ltd) and the HIFUNIT9000 Focused Ultrasound Tumor Ablation Machine (Shanghai Aishen Technology Development Co., Ltd). The system consists of a main unit, motor system, control console, monitoring system, power supply, and water treatment system.

Pre-treatment preparation: The patient was instructed to abstain from high-protein food the day before the treatment. Prior to the treatment, the patient was asked to retain a small amount of urine to fill the bladder. Lactulose oral solution (Beijing Hanmei, 100ml/bottle) was administered for bowel preparation, and parecoxib sodium (Dynastat) was administered via intramuscular injection for analgesia.

During the treatment, phloroglucinol injection was administered intravenously. The patient was positioned supine, and the machine located the pelvic tumor. Throughout the procedure, the treatment intensity and duration were adjusted according to the patient's tolerance and the grayscale changes displayed on the ultrasound.

HIFU is a non-invasive therapeutic technique that does not require anesthesia, has no incisions, no radiation, and has a quick

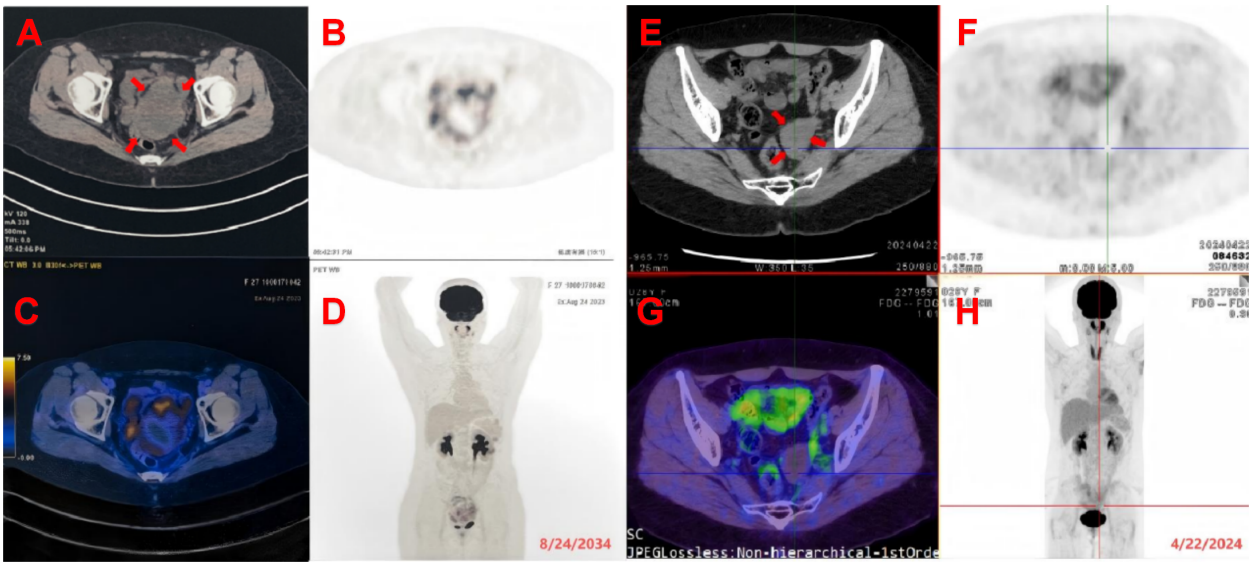


FIGURE 3
(A–D) Pre-treatment PET/CT: A pelvic soft tissue mass demonstrates heterogeneously intense radiotracer uptake with a maximum standardized uptake value (SUVmax) of approximately 7.0. The lesion measures approximately 8.7 × 8.3 × 9.0 cm, showing internal necrotic components. The mass invades the vaginal stump and exhibits ill-defined borders with the rectum and the pelvic segment of the left ureter. **(E–H)** Post-treatment PET/CT after 6 courses of chemotherapy and 2 sessions of HIFU therapy.A hypodense lesion is noted in the left pelvis, measuring approximately 4.7 cm × 3.1 cm. It demonstrates ill-defined borders with the vaginal stump and has an SUVmax of 2.8.

recovery time. It is primarily used for solid tumors that can be observed under ultrasound, such as adenomyosis, uterine fibroids, osteosarcoma, most primary and metastatic liver tumors, etc.

The principle of HIFU treatment involves precise positioning and outlining of the tumor under ultrasound, scanning point by point and layer by layer according to the shape of the tumor. Utilizing the penetrative and focusing properties of ultrasound waves, the waves emitted from outside the body are focused on the pathological tissue inside the body. Through thermal effects, mechanical effects, and cavitation effects, the temperature of the pathological tissue rises instantly to 60-100°C, causing instantaneous irreversible cell death and coagulative necrosis of the tumor tissue. HIFU therapy achieves precise targeted ablation through real-time ultrasound imaging guidance. A 3.5–5 MHz dual-mode transducer enables simultaneous visualization of anatomical structures and blood flow distribution. During treatment, gray-scale ultrasound images are acquired at 5-minute intervals, monitoring echo intensity enhancement in the target area (indicative of coagulative necrosis).

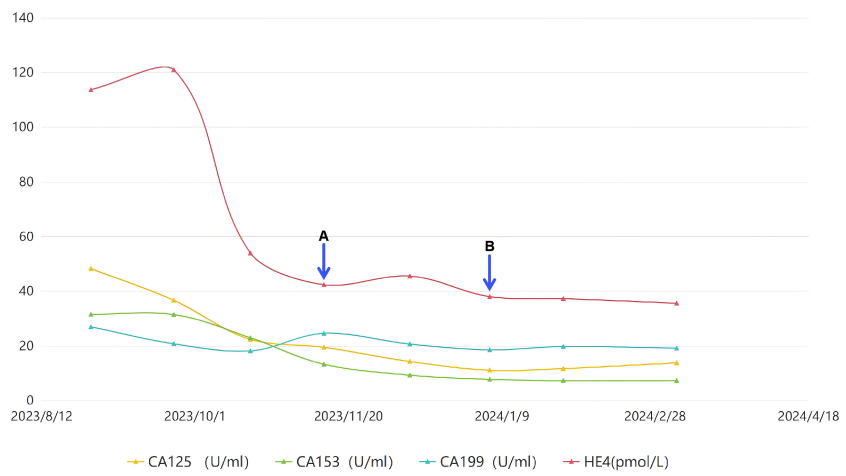


FIGURE 4
The trend chart of various tumor indicators. **(A)** First HIFU treatment; **(B)** Second HIFU treatment.

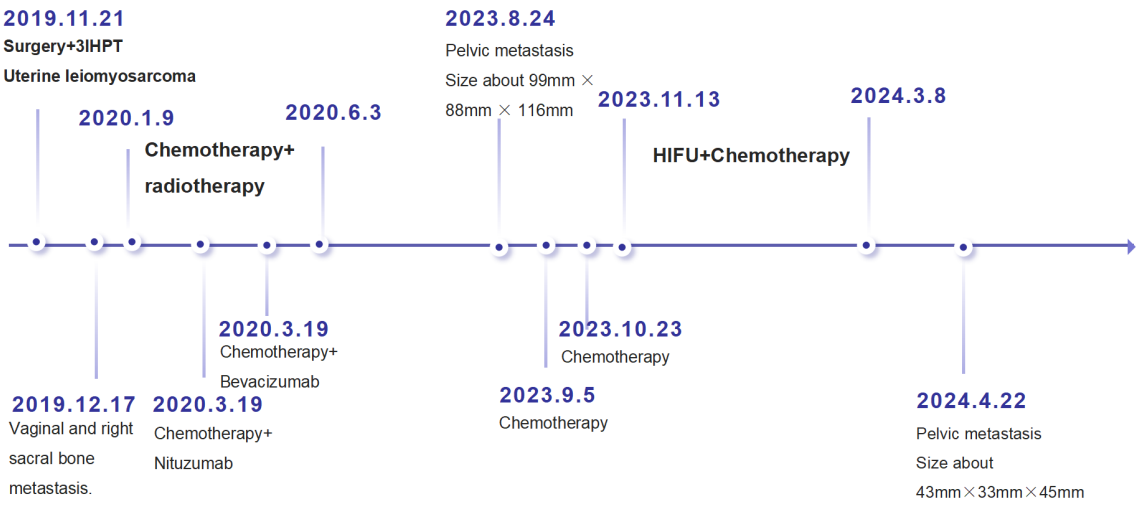


FIGURE 5
Time-event axis.

Initial acoustic intensity is set at 300–500 W/cm², with dynamic adjustments to pulse frequency (0.8–1.2 MHz) and duty cycle (30–50%) based on real-time thermal curves (target temperature 55–65°C). Should acoustic pathway deviation occur (e.g., due to bowel gas interference), immediate treatment suspension and refocalization of the acoustic energy are implemented (Figure 6).

In the assessment of patient adaptability, it is mainly based on the evaluation of subjective symptoms such as lumbosacral pain, abdominal pain, lower - limb neuralgia, and local skin burning during the patient's treatment. If the pain and skin burning are obvious, the energy intensity should be reduced or the treatment should be suspended. Regarding the efficacy assessment, there are currently no precise treatment standards and data. Our clinical experience mainly relies on ultrasound examinations. A better treatment effect is indicated when, in comparison before and after treatment, the area of enhanced echo of the mass under ultrasound exceeds 90%, the closure rate of small blood vessels exceeds 70%, and the reduction rate of blood - flow signals in local thick blood vessels exceeds 50%.

Discussion

ESS is a relatively rare gynecological malignancy. The treatment of recurrent ESS remains a challenge. The 5-year survival rate for patients with stage I and II low-grade ESS reaches 90%, while for patients with stage III and IV, it is about 50% (9). Previous studies have reported recurrence rates of LGESS ranging from 10% to 76%, which may be due to its characteristic of recurrence over 5 years, resulting in a large difference in recurrence rates (11). At present, the main treatment option for endometrial stromal sarcoma is surgery, supplemented by chemotherapy, radiotherapy, anti-estrogen therapy, etc. Due to the many adverse reactions of radiotherapy and chemotherapy and the inability to continue, for patients who cannot undergo surgery, ablation therapy can be chosen. However, the ablation process requires puncture, which may damage surrounding organs such as the intestines and bladder. The puncture process may lead to the risk of tumor rupture, bleeding, and tumor dissemination and metastasis. Therefore, the treatment of recurrent ESS remains a significant challenge.

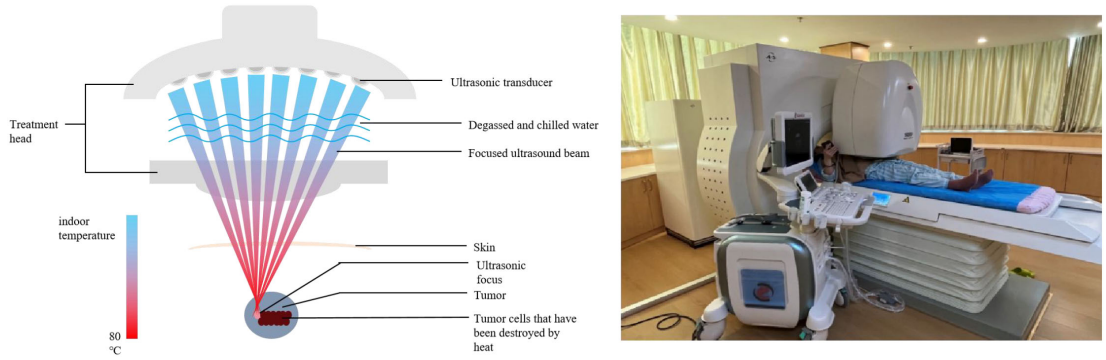


FIGURE 6
Schematic diagram of HIFU operation principle.

In this case, the patient was diagnosed with LGESS and experienced rapid recurrence shortly after surgery. After multiple rounds of combined radiotherapy and chemotherapy, the condition stabilized and regular follow-up examinations were scheduled. In August 2023, the tumor recurred again, measuring approximately 10cm, with multiple pelvic metastases and right sacral bone metastasis. The tumor was suspected of locally recurring and possibly invading the rectum, colon, bladder, and left ureter. Unable to accept the risks of surgery, the patient strongly refused surgical intervention and underwent chemotherapy. After three cycles of chemotherapy, the tumor size decreased from 99mm×88mm×116mm to 69.2mm×57.8mm×75.4mm. However, due to abnormal liver function, the fourth cycle of chemotherapy could not be administered. After evaluation and discussion, HIFU treatment was added, liver protection treatment was administered during this period, and the timing for chemotherapy was awaited. After 17 sessions of HIFU treatment combined with systemic chemotherapy, the tumor size reduced from 69.2mm×57.8mm×75.4mm to 43mm×33mm×45mm, no longer compressing the bladder and ureter, the scattered small nodules in the pelvic cavity disappeared, the level of tumor markers gradually decreased and stabilized, and the patient's abdominal pain and bloating disappeared, significantly improving her quality of life. HIFU treatment during periods when chemotherapy cannot be administered can continuously inhibit tumor progression, preventing tumor enlargement during periods without chemotherapy. Combined liver protection treatment is beneficial for the recovery of liver function, allowing the patient to receive chemotherapy on schedule.

HIFU is a novel non-invasive thermotherapy that can cause coagulative necrosis of tumor tissue. It has the advantages of high repeatability, uniform heat diffusion, virtually painless treatment process, no external injuries, rapid postoperative recovery, and no impact on patient function. It has been proven effective and safe in the treatment of solid tumors such as uterine fibroids, breast cancer, and pancreatic cancer. A prospective study suggested that the effectiveness of HIFU in treating uterine fibroids was higher than surgical treatment, and it was safer (12). MR-HIFU treatment significantly alleviates the clinical symptoms caused by uterine fibroids and effectively reduces the tumor volume (13). In addition, a retrospective review findings of HIFU treatment was more effective than secondary myoma resection, with fewer side effects, longer asymptomatic periods, and lower risk of re-intervention (14). A systematic review study showed that patients with postoperative pathological diagnosis of uterine sarcomas (including LGESS and uterine leiomyosarcoma) do not cause histological dissemination of sarcoma after receiving HIFU treatment (15). HIFU treatment has therapeutic effects on uterine fibroids and sarcomas, and also has good effects in the treatment of other pelvic tumors. Zhong Q,etc (16), retrospectively analyzed 153 patients with cervical cancer residual or recurrent after chemoradiotherapy (CRT) who received HIFU treatment from 2010 to 2021. The results showed that HIFU can significantly reduce the size of residual or recurrent lesions, improve local control rates and survival time, and even elderly or physically poor patients can tolerate it, providing a supplementary treatment method for cervical cancer patients with adverse reactions after chemotherapy. Lei T,etc (17), treated 8 patients with recurrent ovarian cancer or metastatic pelvic tumors with HIFU, and found that the pain relief rate was 60%, short-term quality of life improved, and adverse reactions after treatment were mild. Studies have shown that HIFU treatment of pelvic metastatic tumors or recurrent ovarian cancer is feasible and without serious complications. HIFU treatment is also used in breast cancer and pancreatic cancer. Zulkifli D,etc (18), included nine studies and found that HIFU can induce coagulative necrosis of local breast cancer tumors, with small side effects, good cosmetic effects, and a 5-year disease-free survival rate of more than 90%. A meta-analysis evaluated 19 studies with a total of 939 patients, and the results showed that HIFU treatment combined with drug treatment of pancreatic cancer can relieve patients' chronic pain, the incidence of adverse events is low, and it can improve the overall survival rate (19). In the treatment of prostate cancer, HIFU treatment also plays a role. Parry MG reported that after 1381 patients with prostate cancer received HIFU treatment, the tumor effectively shrank, and urinary and reproductive functions were preserved, with little impact on the quality of life (20).

HIFU is currently used for pelvic and abdominal solid tumors, and the treatment effect is good, patients with residual or recurrent tumors in the pelvis after radiotherapy and chemotherapy also benefit. These research results provide evidence for us to choose to add HIFU in this case, clinical data also prove that HIFU combined with chemotherapy for the treatment of recurrent low-grade endometrial stromal sarcoma is effective and safe.

During HIFU treatment, different tumor sizes and locations are associated with distinct side effects and limitations. To enhance treatment safety, prior to treatment, it is necessary to improve the patient's nutritional status, control underlying diseases, and establish psychological expectations. Additionally, multi - modal imaging techniques should be employed to precisely locate the lesion. During the treatment, parameters should be dynamically adjusted based on the tumor size, depth, blood supply characteristics, and the patient's adverse reactions. This ensures effective ablation of the tumor tissue while minimizing damage to the surrounding normal tissues to the greatest extent. After the treatment, measures should be taken as early as possible to address adverse reactions. Hierarchical interventions should be carried out for common problems such as fever, pain, and skin damage. Meanwhile, psychological counseling should be provided to improve the patient's treatment experience.

The combined HIFU therapy has gained increasing attention, and changes in immune-related markers and tumor biomarkers may be associated with treatment prognosis. Dong S et al. compared pancreatic cancer patients receiving HIFU-priority versus chemotherapy-priority regimens in combined therapy and found that the HIFU-priority group demonstrated significantly improved overall survival (OS) (HR = 0.38) (21). Additionally, patients with normal CRP and CA125 levels exhibited longer survival. Elevated neutrophil-lymphocyte ratio (NLR) and low lymphocyte-monocyte ratio (LMR) were associated with poor prognosis. Wang R et al. found that patients positive for CD133 and other stem cell markers may benefit from targeted nanocarrier-based therapies combined

with HIFU (22). HIFU may enhance chemotherapeutic efficacy by creating a tumor hypoxic environment that activates hypoxia-inducible factors (HIFs), thereby improving the delivery efficiency of chemotherapeutic agents such as doxorubicin. Concurrently, HIFU promotes CD4+/CD8+ lymphocyte infiltration into tumor tissues (23, 24). HIFU activates systemic immune responses by releasing tumor antigens and danger signals, with CD8+ lymphocyte infiltration correlating with regression of distant untreated lesions (24, 25). Patients with higher baseline tumor-infiltrating lymphocyte (TIL) levels are more suitable for HIFU combined with PD-1 inhibitors and chemotherapy (26, 27).

In terms of pathological characteristics, the combination of HIFU and chemotherapy significantly controls the growth of recurrent lesions in mucinous ovarian cancer (28). In advanced gastric cancer (GC) patients, HIFU-priority regimens following neoadjuvant chemotherapy significantly improve OS, particularly in stage III patients (HR = 1.61) (29). Multimodal imaging serves as the gold standard for post-HIFU chemotherapeutic response evaluation, with contrast-enhanced CT/MRI clearly delineating tumor anatomy and Extent of necrosis (30, 31). Molecular ultrasound imaging dynamically monitors tumor vascular characteristics (via the QuanTAV index), predicting treatment sensitivity (32). Translucent texture changes in ultrasound/MRI follow-up of muscularis lesions indicate therapeutic efficacy, whereas residual enhancing foci warrant caution for recurrence (33, 34). Future research should prioritize refining a multi-parameter decision model integrating tumor biomarkers, imaging features, pathological staging, and immunological status to optimize HIFU-chemotherapy combination therapy precision.

The main mechanisms by which HIFU combined with chemotherapy may exert its therapeutic effect are likely related to the following aspects. First, tumor cells are more sensitive to high temperatures than normal cells. HIFU destroys tumor tissue through its thermal effect, inducing apoptosis of tumor cells; the thermal effect can increase tumor blood flow and enhance the permeability of the tumor cell membrane, thereby accelerating the penetration and absorption of chemotherapeutic drugs (21, 35). Second, after HIFU treatment, tumor cells die and cellular components enter the bloodstream. The expression of a large number of tumor antigens in the fragments activates the immune system's anti-tumor response. Third, some studies suggest that HIFU treatment can change the tumor's resistance to chemotherapy, increasing the sensitivity of tumor cells to chemotherapeutic drugs (36, 37). The anti-tumor mechanism of HIFU treatment is still under research, especially the impact on the immune system which requires further exploration.

While offering the advantage of non-invasiveness, HIFU possesses significant limitations in clinical application. Its efficacy is constrained by tissue acoustic properties; it cannot effectively penetrate gas-containing organs (e.g., lungs) or bone, limiting its use for tumors in locations like the thorax or intracranial cavity. Furthermore, HIFU application is highly dependent on specific tumor characteristics: size, well-defined margins, and proximity to critical vasculature or nerves. Tumors that are excessively large or unfavorably located pose procedural risks, including potential damage to adjacent structures such as bowel loops or nerves.

Additionally, real-time monitoring during treatment and accurate post-procedural efficacy assessment remain challenging. The inability to obtain tissue samples for histopathological confirmation necessitates reliance on post-treatment imaging follow-up for evaluating response. Procedural success heavily depends on operator expertise, resulting in a steep learning curve. Crucially, HIFU primarily ablates localized tumor tissue; it does not target systemic tumor dissemination via hematogenous spread, lymphatic metastasis, or distant seeding. Therefore, HIFU must be integrated with systemic therapies and serves as an effective adjunct to, rather than a replacement for, conventional cancer treatments.

Conclusions

In conclusion, this case demonstrates that HIFU combined with chemotherapy is effective in treating recurrent endometrial stromal sarcoma. This combined treatment provides a new option for patients who refuse secondary surgery or cannot tolerate it. We hope that more clinical research and data will confirm its effectiveness and safety in the future, and further explore its mechanism of action in endometrial stromal sarcoma, especially its impact on immune function and the mechanism of action in increasing sensitivity and enhancing efficacy.

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 authors.

Ethics statement

The studies involving humans were approved by Zhuhai Hospital of Integrated Traditional Chinese and Western Medicine Ethics Committee (Ethic code: 20220113006). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article. Written informed consent was obtained from the participant/patient(s) for the publication of this case report.

Author contributions

HC: Supervision, Writing – original draft, Writing – review & editing, Investigation. XS: Investigation, Writing – original draft, Writing – review & editing. YS: Investigation, Writing – original draft, Writing – review & editing. HH: Investigation, Writing – review & editing. ZJ: Investigation, Writing – review & editing. QW: Investigation, Writing – review & editing. ZC: Data curation,

Writing – review & editing. PY: Writing – original draft, Investigation. SX: Data curation, Writing – original draft. LC: Data curation, Writing – review & editing. DH: Supervision, Writing – review & editing. MK: Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, 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 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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2025.1551792/full#supplementary-material

References

1. Huang W, Zhang T, Wang H, Liu Z, Zhai P, Wang X, et al. Intravenous metastasis of unexpected uterine sarcoma in the context of uterine fibroids: case report and literature review. *Front Oncol*. (2024) 14:1354032. doi: 10.3389/fonc.2024.1354032

2. Micci F, Heim S, Panagopoulos I. Molecular pathogenesis and prognostication of "low-grade" and "high-grade" endometrial stromal sarcoma. *Genes Chromosomes Cancer*. (2021) 60:160–7. doi: 10.1002/gcc.22907

3. Borella F, Bertero L, Cassoni P, Piovano E, Gallio N, Preti M, et al. Low-Grade uterine endometrial stromal sarcoma: prognostic analysis of clinico-Pathological characteristics, surgical management, and adjuvant treatments. Experience from two referral centers. *Front Oncol*. (2022) 12:883344. doi: 10.3389/fonc.2022.883344

4. Niu S, Zheng W. Endometrial stromal tumors: Diagnostic updates and challenges. *Semin Diagn Pathol*. (2022) 39:201–12. doi: 10.1053/j.semdp.2022.01.004

5. Meng LL, Jia XP, Lu LX, Zhang HZ, Shen XH, Piao ZH, et al. Unusual morphologic features of low-grade endometrial stromal sarcoma: A case report. *J Clin Lab Anal*. (2022) 36:e24502. doi: 10.1002/jcla.24502

6. Al-Ibraheem A, Abdlkadir AS. The outcome of progressive uterine sarcoma with potential bone involvement. *World J Nucl Med*. (2023) 22:48–51. doi: 10.1055/s-0042-1757285

7. Hoang L, Chiang S, Lee CH. Endometrial stromal sarcomas and related neoplasms: new developments and diagnostic considerations. *Pathology*. (2018) 50:162–77. doi: 10.1016/j.pathol.2017.11.086

8. Laufer J, Scasso S, Kim B, Shahi M, Mariani A. Fertility-sparing management of low-grade endometrial stromal sarcoma. *Int J Gynecol Cancer*. (2023) 33:1145–9. doi: 10.1136/ijgc-2023-004448

9. Dai Q, Xu B, Wu H, You Y, Wu M, Li L. The prognosis of recurrent low-grade endometrial stromal sarcoma: a retrospective cohort study. *Orphanet J Rare Dis*. (2021) 16:160. doi: 10.1186/s13023-021-01802-8

10. Liang L, Wang J, Xie J, Xu Y, Zhang L, Liu D, et al. High-dose insulin and dexamethasone combined with radiotherapy in endometrial stromal sarcoma recurring with multiple metastases: A case report. *Med (Baltimore)*. (2023) 102:e33525. doi: 10.1097/md.0000000000033525

11. Quan C, Zheng Z, Cao S, Wu Y, Zhang W, Huang Y. The value of surgery and the prognostic factors for patients with recurrent low-grade endometrial stromal sarcoma: a retrospective study of 38 patients. *J Gynecol Oncol*. (2024) 35:e98. doi: 10.3802/jgo.2024.35.e98

12. Chen J, Li Y, Wang Z, McCulloch P, Hu L, Chen W, et al. Evaluation of high-intensity focused ultrasound ablation for uterine fibroids: an IDEAL prospective exploration study. *Bjog*. (2018) 125:354–64. doi: 10.1111/1471-0528.14689

13. Lozinski T, Filipowska J, Pyka M, Baczkowska M, Ciebiera M. Magnetic resonance-guided high-intensity ultrasound (MR-HIFU) in the treatment of symptomatic uterine fibroids - five-year experience. *Ginekol Pol*. (2021) 93:185–94. doi: 10.5603/GP.a2021.0098

14. Liu X, Tang J, Luo Y, Wang Y, Song L, Wang W. Comparison of high-intensity focused ultrasound ablation and secondary myomectomy for recurrent symptomatic uterine fibroids following myomectomy: a retrospective study. *Bjog*. (2020) 127:1422–8. doi: 10.1111/1471-0528.16262

15. Wang Q, Wu X, Zhu X, Wang J, Xu F, Lin Z, et al. MRI features and clinical outcomes of unexpected uterine sarcomas in patients who underwent high-intensity focused ultrasound ablation for presumed uterine fibroids. *Int J Hyperthermia*. (2021) 38:39–45. doi: 10.1080/02656736.2021.1921288

16. Zhong Q, Tang F, Ni T, Chen Y, Liu Y, Wu J, et al. Salvage high intensity focused ultrasound for residual or recurrent cervical cancer after definitive chemoradiotherapy. *Front Immunol*. (2022) 13:995930. doi: 10.3389/fimmu.2022.995930

17. Lei T, Guo X, Gong C, Chen X, Ran F, He Y, et al. High-intensity focused ultrasound ablation in the treatment of recurrent ovary cancer and metastatic pelvic tumors: a feasibility study. *Int J Hyperthermia*. (2021) 38:282–7. doi: 10.1080/02656736.2021.1889698

18. Zulkifli D, Manan HA, Yahya N, Hamid HA. The applications of high-intensity focused ultrasound (HIFU) ablative therapy in the treatment of primary breast cancer: A systematic review. *Diagnostics (Basel)*. (2023) 13:2595. doi: 10.3390/diagnostics13152595

19. Fergadi MP, Magouliotis DE, Rountas C, Vlychou M, Athanasiou T, Symeonidis D, et al. A meta-analysis evaluating the role of high-intensity focused ultrasound (HIFU) as a fourth treatment modality for patients with locally advanced pancreatic cancer. *Abdom Radiol (NY)*. (2022) 47:254–64. doi: 10.1007/s00261-021-03334-y

20. Parry MG, Sujenthiran A, Nossiter J, Morris M, Berry B, Nathan A, et al. Prostate cancer outcomes following whole-gland and focal high-intensity focused ultrasound. *BJU Int*. (2023) 132:568–74. doi: 10.1111/bju.16122

21. Dong S, Zhong A, Zhu H, Wang K, Cheng CS, Meng Z. Sequential high-intensity focused ultrasound treatment combined with chemotherapy for inoperable pancreatic cancer: a retrospective analysis for prognostic factors and survival outcomes. *Int J Hyperthermia*. (2023) 40:2278417. doi: 10.1080/02656736.2023.2278417

22. Wang R, Yao Y, Gao Y, Liu M, Yu Q, Song X, et al. CD133-targeted hybrid nanovesicles for fluorescent/ultrasonic imaging-guided HIFU pancreatic cancer therapy. *Int J Nanomedicine*. (2023) 18:2539–52. doi: 10.2147/ijn.S391382

23. Ashar H, Singh A, Kishore D, Neel T, More S, Liu C, et al. Enabling chemo-Immunotherapy with HIFU in canine cancer patients. *Ann BioMed Eng*. (2024) 52:1859–72. doi: 10.1007/s10439-023-03194-1

24. Su S, Wang Y, Lo EM, Tamukong P, Kim HL. High-intensity focused ultrasound ablation to increase tumor-specific lymphocytes in prostate cancer. *Transl Oncol*. (2025) 53:102293. doi: 10.1016/j.tranon.2025.102293

25. Mekers V, de Visser M, Suijkerbuijk K, Bos C, Moonen C, Deckers R, et al. Mechanical HIFU and immune checkpoint inhibition: toward clinical implementation. *Int J Hyperthermia*. (2024) 41:2430333. doi: 10.1080/02656736.2024.2430333

26. Wu F, Jiang T, Chen G, Huang Y, Zhou J, Lin L, et al. Multiplexed imaging of tumor immune microenvironmental markers in locally advanced or metastatic non-small-cell lung cancer characterizes the features of response to PD-1 blockade plus chemotherapy. *Cancer Commun (Lond)*. (2022) 42:1331–46. doi: 10.1002/cac2.12383

27. Wongpattaraworakul W, Choi A, Buchakjian MR, Lanzel EA, Kd AR, Simons AL. Prognostic role of tumor-infiltrating lymphocytes in oral squamous cell carcinoma. *BMC Cancer*. (2024) 24:766. doi: 10.1186/s12885-024-12539-5

28. Guo X, Liu W, Zhou K, Zhu H, Pan L, Feng C, et al. High-intensity focused ultrasound (HIFU) assisted by a rectal Foley catheter for the treatment of recurrent mucinous ovarian cancer: a case report and literature review. *Front Oncol*. (2024) 14:1498631. doi: 10.3389/fonc.2024.1498631

29. Liao Y, Wang D, Yang X, Ni L, Lin B, Zhang Y, et al. High–intensity focused ultrasound thermal ablation boosts the efficacy of immune checkpoint inhibitors in advanced cancers with liver metastases: A single–center retrospective cohort study. *Oncol Lett*. (2025) 29:124. doi: 10.3892/ol.2025.14871

30. Ran L, Yang W, Chen X, Zhang J, Zhou K, Zhu H, et al. High-Intensity focused ultrasound ablation combined with pharmacogenomic-Guided chemotherapy for advanced pancreatic cancer: initial experience. *Ultrasound Med Biol*. (2024) 50:1566–72. doi: 10.1016/j.ultrasmedbio.2024.06.013

31. Lan JY, Yang JH, Liang YH, Lin AY, Liao JY. Comparative analysis of imaging and pathological features in diagnosis of endometrial carcinosarcoma based on multimodal MRI. *Arch Gynecol Obstet*. (2025) 311:159–61. doi: 10.1007/s00404-024-07817-3

32. Braman N, Prasanna P, Bera K, Alilou M, Khorrami M, Leo P, et al. Novel radiomic measurements of tumor-Associated vasculature morphology on clinical imaging as a biomarker of treatment response in multiple cancers. *Clin Cancer Res*. (2022) 28:4410–24. doi: 10.1158/1078-0432.Ccr-21-4148

33. Pan Y, Lin K, Hu Y, Song X, Xu L, Zhou Z, et al. Integrating high-intensity focused ultrasound with chemotherapy for the treatment of invasive hydatidiform mole in reproductive-age women. *Gynecol Minim Invasive Ther*. (2024) 13:184–8. doi: 10.4103/gmit.gmit_86_23

34. Tang R, He H, Lin X, Wu N, Wan L, Chen Q, et al. Novel combination strategy of high intensity focused ultrasound (HIFU) and checkpoint blockade boosted by bioinspired and oxygen-supplied nanoprobe for multimodal imaging-guided cancer therapy. *J Immunother Cancer*. (2023) 11:e006226. doi: 10.1136/jitc-2022-006226

35. Bachu VS, Kedda J, Suk I, Green JJ, Tyler B. High-intensity focused ultrasound: A review of mechanisms and clinical applications. *Ann Biomed Eng*. (2021) 49:1975–91. doi: 10.1007/s10439-021-02833-9

36. Li J, Chen X, Hu X. High-intensity focused ultrasound for treatment of recurrent uterine leiomyosarcoma: a case report and literature review. *J Int Med Res*. (2020) 48:300060520942107. doi: 10.1177/0300060520942107

37. She C, Li S, Wang X, Lu X, Liang H, Liu X. High-intensity focused ultrasound ablation as an adjuvant surgical salvage procedure in gestational trophoblastic neoplasia chemotherapy with chemoresistance or recurrence: two case reports. *Int J Hyperthermia*. (2021) 38:1584–9. doi: 10.1080/02656736.2021.1998659

Frontiers in Oncology

Advances knowledge of carcinogenesis and tumor progression for better treatment and management

The third most-cited oncology journal, which highlights research in carcinogenesis and tumor progression, bridging the gap between basic research and applications to imrpove diagnosis, therapeutics and management strategies.

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