frontiers | Research Topics

# The combination of data-driven machine learning approaches and prior knowledge for robust medical image processing and analysis

**Edited by**
Jinming Duan, Chen Qin, Gongning Luo and Diwei Zhou

**Published in**
Frontiers in Medicine
Frontiers in Cardiovascular Medicine

ISSN 1664-8714
ISBN 978-2-8325-5019-9
DOI 10.3389/978-2-8325-5019-9

## 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

# The combination of data-driven machine learning approaches and prior knowledge for robust medical image processing and analysis

**Topic editors**

Jinming Duan — University of Birmingham, United Kingdom
Chen Qin — Imperial College London, United Kingdom
Gongning Luo — Harbin Institute of Technology, China
Diwei Zhou — Loughborough University, United Kingdom

**Citation**

Duan, J., Qin, C., Luo, G., Zhou, D., eds. (2024). *The combination of data-driven machine learning approaches and prior knowledge for robust medical image processing and analysis*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-5019-9

# Table of contents

frontiers | Frontiers in Medicine

TYPE Editorial
PUBLISHED 31 May 2024
DOI 10.3389/fmed.2024.1434686



OPEN ACCESS

EDITED AND REVIEWED BY
Giorgio Treglia,
Ente Ospedaliero Cantonale (EOC), Switzerland

*CORRESPONDENCE
Diwei Zhou
✉ D.Zhou2@lboro.ac.uk
Jinming Duan
✉ j.duan@bham.ac.uk

RECEIVED 18 May 2024
ACCEPTED 23 May 2024
PUBLISHED 31 May 2024

CITATION
Zhou D, Duan J, Qin C and Luo G (2024) Editorial: The combination of data-driven machine learning approaches and prior knowledge for robust medical image processing and analysis.
*Front. Med.* 11:1434686.
doi: 10.3389/fmed.2024.1434686



# Editorial: The combination of data-driven machine learning approaches and prior knowledge for robust medical image processing and analysis

Diwei Zhou[1]*, Jinming Duan[2]*, Chen Qin[3] and Gongning Luo[4]

[1]Department of Mathematical Sciences, School of Science, Loughborough University, Loughborough, United Kingdom, [2]School of Computer Science, University of Birmingham, Birmingham, United Kingdom, [3]Department of Electrical and Electronic Engineering & I-X, Imperial College London, London, United Kingdom, [4]Faculty of Computing, Harbin Institute of Technology, Harbin, Heilongjiang, China

KEYWORDS

**deep learning, medical imaging, diagnostic accuracy, data-driven approaches, prior knowledge, robustness**

Editorial on the Research Topic
The combination of data-driven machine learning approaches and prior knowledge for robust medical image processing and analysis

Combining data-driven machine learning with prior knowledge has significantly advanced medical image processing and analysis. Deep learning, driven by large datasets and powerful GPUs, excels in tasks like image reconstruction, segmentation, and disease classification. However, these models face challenges such as high resource demands, limited generalization, and lack of interpretability. In contrast, model-driven approaches offer better generalization, interpretability, and robustness but may lack accuracy and efficiency. Combining these paradigms leverages their strengths, promising superior performance and enhanced diagnostic accuracy. This Research Topic showcases how this integration enhances medical imaging, including accurate stroke onset estimation, improved COVID-19 diagnosis and recovery assessment, and enhanced cardiac imaging techniques. These advancements highlight the potential for improved diagnostic accuracy, treatment planning, and clinical decision-making in medical imaging.

A convolutional neural network (CNN) was developed by Gao et al. to identify acute ischemic stroke patients within a 6-h window for endovascular thrombectomy using computed tomography perfusion and perfusion-weighted imaging. This CNN outperformed support vector machines and random forests, demonstrating its potential for accurate stroke onset time estimation using both CT and MR imaging.

Building on the success of deep learning in stroke diagnosis, another study by Huang et al. utilized deep learning and CT scans to assess lung recovery in COVID-19 Delta variant survivors over 6 months. The findings were promising, with ground-glass opacities disappearing and mild fibrosis in most cases, alongside improved lung prognosis compared to the original COVID-19 strain. In a similar vein, a mixed-effects deep learning model was created by Bridge et al. to diagnose COVID-19 from CT scans, achieving high accuracy and robustness. With an AUROC of 0.930 in external validation, this model

outperformed other methods, showcasing potential for clinical application in automated COVID-19 diagnosis.

Transitioning to cardiac imaging, a novel Transformer-ConvNet architecture, MAE-TransRNet, was proposed by Xiao et al. for cardiac MRI registration. This method significantly improved deformable image registration accuracy by combining the strengths of convolutional neural networks (CNN) and Transformers, outperforming state-of-the-art methods on the ACDC dataset.

Extending the application of deep learning to ENT diagnostics, a multi-scale deep learning network, MIB-ANet, was developed by Bi et al. for grading adenoid hypertrophy from nasal endoscopy images. This network outperformed junior E.N.T. clinicians in accuracy and speed, demonstrating its potential for clinical application in automated adenoid hypertrophy grading.

Further advancing medical imaging, an anatomical prior-informed masking strategy for pre-training masked autoencoders was introduced by Wang et al. to enhance brain tumor segmentation. Leveraging brain structure knowledge to guide masking, this method improved efficiency and accuracy on the BraTS21 dataset, outperforming state-of-the-art self-supervised learning techniques. Similarly, a Joint 2D−3D Cross-Pseudo Supervision (JCPS) method was introduced by Zhou et al. for segmenting the carotid vessel wall in black-blood MRI images. This approach, which combines coarse and fine segmentation leveraging both labeled and unlabeled data, significantly enhanced segmentation accuracy, outperforming existing methods.

A systematic review of deep learning techniques for segmenting isointense infant brain tissues in MRI was conducted by Mhlanga and Viriri, analyzing 19 studies from 2012–2022. This review highlighted challenges due to low tissue contrast and overlapping intensity in white and gray matter, with convolutional neural networks (CNNs) being prominently used.

AI-based echocardiographic quantification of global longitudinal strain (GLS) and left ventricular ejection fraction (LVEF) in trastuzumab-treated patients was evaluated by Jiang et al.. They found moderate to strong correlations with conventional methods, suggesting AI's potential as a supplementary tool in clinical settings despite lower feasibility rates. In another study employing echocardiograms, Zhang Y. et al. introduced an automated pipeline that utilizes deep neural networks and ensemble learning to accurately quantify left ventricular ejection fraction (LVEF) and predict heart failure. Their method demonstrated high accuracy and clinical applicability, achieving a Pearson's correlation coefficient of 0.83 with expert analysis and an AUROC of 0.98 for heart failure classification. Furthermore, a semi-supervised contrastive learning network was proposed by Guo et al. for multi-structure echocardiographic segmentation. Evaluated on the CAMUS dataset, it achieved high performance, outperforming existing methods and using fewer parameters. This approach enhances cardiac disease diagnosis and reduces clinician workload.

Finally, for oncology, MRI radiomics-based machine learning models were compared for predicting glioblastoma multiforme prognosis by Zhang D. et al. The DeepSurv model outperformed traditional Cox proportional-hazards and other models, highlighting the potential of deep learning in improving GBM survival predictions.

In conclusion, the integration of data-driven machine learning approaches with prior knowledge marks a significant advancement in medical imaging. The studies reviewed herein underscore the transformative impact of these combined methodologies, offering substantial improvements in diagnostic accuracy, efficiency, and robustness across various medical imaging tasks. This Research Topic significantly contributes to the field by addressing key challenges and paving the way for more reliable and precise medical image analysis, ultimately enhancing patient outcomes and clinical decision-making.

## Author contributions

DZ: Conceptualization, Writing – original draft, Writing – review & editing. JD: Writing – review & editing. CQ: Writing – review & editing. GL: 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

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 02 February 2023
DOI 10.3389/fmed.2023.1103559



OPEN ACCESS

EDITED BY
Gongning Luo,
Harbin Institute of Technology, China

REVIEWED BY
Haiwei Pan,
Harbin Engineering University, China
Guanglu Sun,
Harbin University of Science and Technology, China

*CORRESPONDENCE
Zhaowen Qiu
✉ qiuzw@nefu.edu.cn
Mingsheng Lei
✉ mingshenglei@163.com

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Pulmonary Medicine, a section of the journal Frontiers in Medicine

RECEIVED 23 November 2022
ACCEPTED 11 January 2023
PUBLISHED 02 February 2023

CITATION
Huang J, Lin R, Bai N, Su Z, Zhu M, Li H, Chai C, Xia M, Shu Z, Qiu Z and Lei M (2023) Six-month follow-up after recovery of COVID-19 Delta variant survivors via CT-based deep learning. *Front. Med.* 10:1103559. doi: 10.3389/fmed.2023.1103559



# Six-month follow-up after recovery of COVID-19 Delta variant survivors *via* CT-based deep learning

Jianliang Huang[1†], Ruikai Lin[2†], Na Bai[2†], Zhongrui Su[1], Mingxin Zhu[1], Han Li[2], Conghai Chai[1], Mingkai Xia[1], Ziwei Shu[3], Zhaowen Qiu[2,4*] and Mingsheng Lei[1,5*]

[1]Zhangjiajie Hospital Affiliated to Hunan Normal University, Zhangjiajie, China, [2]College of Information and Computer Engineering, Northeast Forestry University, Harbin, China, [3]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore, [4]Heilongjiang Tuomeng Technology Co., Ltd., Harbin, China, [5]Zhangjiajie College, Zhangjiajie, China

**Purpose:** Using computer-aided diagnosis (CAD) methods to analyze the discharge and 6-month follow-up data of COVID-19 Delta variant survivors, evaluate and summarize the recovery and prognosis, and improve people's awareness of this disease.

**Methods:** This study collected clinical data, SGRQ questionnaire results, and lung CT scans (at both discharge and 6-month follow-up) from 41 COVID-19 Delta variant survivors. Two senior radiologists evaluated the CT scans before in-depth analysis. Deep lung parenchyma enhancing (DLPE) method was used to accurately segment conventional lesions and sub-visual lesions in CT images, and then quantitatively analyze lung injury and recovery. Patient recovery was also measured using the SGRQ questionnaire. The follow-up examination results from this study were combined with those of the original COVID-19 for further comparison.

**Results:** The participants include 13 males (31.7%) and 28 females (68.3%), with an average age of 42.2 ± 17.7 years and an average BMI of 25.2 ± 4.4 kg/m$^2$. Compared discharged CT and follow-up CT, 48.8% of survivors had pulmonary fibrosis, mainly including irregular lines (34.1%), punctuate calcification (12.2%) and nodules (12.2%). Compared with discharged CT, the ground-glass opacity basically dissipates at follow-up. The mean SGRQ score was 0.041 (0–0.104). The sequelae of survivors mainly included impaired sleep quality (17.1%), memory decline (26.8%), and anxiety (21.9%). After DLPE process, the lesion volume ratio decreased from 0.0018 (0.0003, 0.0353) at discharge to 0.0004 (0, 0.0032) at follow-up, $p < 0.05$, and the absorption ratio of lesion was 0.7147 (−1.0303, 0.9945).

**Conclusion:** The ground-glass opacity of survivors had dissipated when they were discharged from hospital, and a little fibrosis was seen in CT after 6-month, mainly manifested as irregular lines, punctuate calcification and nodules. After DLPE and quantitative calculations, we found that the degree of fibrosis in the lungs of most survivors was mild, which basically did not affect lung function. However, there are a small number of patients with unabsorbed or increased fibrosis. Survivors mainly had non-pulmonary sequelae such as impaired sleep quality and memory decline. Pulmonary prognosis of Delta variant patients was better than original COVID-19, with fewer and milder sequelae.

KEYWORDS

**follow-up, Delta variant survivors, deep lung parenchyma enhancing, sub-visual lesion, pulmonary fibrosis, COVID-19 sequelae**

# 1. Introduction

Since first detected in Wuhan, China, Coronavirus disease 2019 (COVID-19) has swept the world, threatening the world with public health concerns and social instability. As of 3rd November 2022, the cumulative number of confirmed COVID-19 cases worldwide reached 631,324,387, with more than 6,594,803 cumulative deaths (1). The major pathogen of COVID-19 has been identified as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), yet new variants kept appearing, leading to ongoing worldwide outbreaks of COVID-19 at different magnitudes. SARS-CoV-2 Delta variant (also known as lineage B.1.617.2), a variant of concern identified by the World Health Organization (WHO), became the primary strain of the COVID-19 pandemic in 2021 (2), affecting more than 75% of countries worldwide. In March 2022, a new variant named Deltacron with Delta variant as the main stem was confirmed to exist by WHO, which will exacerbate the plague of COVID-19 to humans. Therefore, it is very important and urgent to fully understand COVID-19, especially the mechanism of action and physiological effects of SARS-CoV-2 and its variants on humans.

In the mid-1960s, Lodwick first introduced the concept of using computer technology for medical image analysis and computer-aided diagnosis (CAD). However, limitations such as technology and clinical philosophy have constrained the development of CAD technology. It was not until after the 1980s, with the development of mathematics, statistics, data mining techniques, computer algorithms and other sciences, that CAD emerged in large numbers in the treatment and prognosis studies of many diseases (3). Notably, the rapid development of artificial intelligence (AI) has surged the recent CAD craze, enabling the application of technologies such as machine learning and deep learning in clinical diagnosis, treatment, and prognosis. To date, AI has gradually emerged in various medical fields and clinical challenges, such as tumor diagnosis, cardiovascular diseases, and central nervous system pathologies (4, 5). During the COVID-19 pandemic, AI approaches have been extended to understanding COVID-19 pneumonia from multiple perspectives, including prevention, diagnosis, treatment, monitoring, and follow-up examination, as such to provide an abundance of valuable clinical evidence and decision support for fighting against the disease (6).

We collected academic research on COVID-19 (SARS-CoV-2 virus) and its variants from three literature databases, the Web of Science, PubMed, and China National Knowledge Infrastructure, bringing the total number of relevant publications to 640,333 from the earliest searchable date to May 2022. While 19,873 cases were related to follow-up examination, only 101 were associated with the Delta variant. In this study, we followed up with 41 Delta variant survivors from Zhangjiajie City, China, for 6 months after discharge. We collected these patients' last CT scan and clinical data before they were discharged from the hospital and continued to collect CT scans and important clinical indicators during the 6-month follow-up. Further, we used AI approaches such as deep lung parenchyma enhancing (DLPE) to quantify follow-up CT and discharged CT (7) and to provide a comprehensive assessment of patient recovery and prognosis. Our findings provide intrinsic insights into the mechanisms underlying the prognosis of COVID-19, especially the Delta variant.

# 2. Materials and methods

## 2.1. Study design and participants

This is a retrospective study. We collected 6-month follow-up data from COVID-19 Delta variant patients admitted to Zhangjiajie City People's Hospital from July to September 2021. All diagnoses and discharges of patients conformed to the Diagnosis and Treatment of Novel Coronavirus Infection Guidelines produced by the Chinese National Health Commission (Trial Version eight or earlier versions) (8). We excluded the following patients: 1) patients who died before follow-up; 2) patients who refused to participate in follow-up; 3) patients who could not be contacted or otherwise could not participate in follow-up; 4) patients diagnosed with asymptomatic infection at discharge. A total of 41 individuals, including 13 males and 28 females, participated in this follow-up study. We classified the patients into three age groups: youngth (under 45 years), middle-aged (45–59 years), and elderly (60–89 years) in light of the WHO age classification criteria (9). Each individual' s CT scans (at discharge and follow-up) and Body Mass Index (BMI) were collected accordingly. The Chinese BMI standard defines four categories: BMI<18.4 indicates a thin body shape, 18.5<BMI<23.9 indicates a normal body shape, 24.0<BMI<27.9 indicates an overweight body shape, and BMI>28.0 indicates an obese body shape. Other clinical data including vaccination status at discharge and SGRQ scores at follow-up were recorded for analysis. Patients with no <1 dose of vaccination history were included in the vaccination cohort concerning the low availability of the COVID-19 vaccine during the Delta variant outbreak in Zhangjiajie. We conducted the hospital discharge and the follow-up CTs with the TOSHIBA Aquilion Lightning CT scanner. The tube voltage and current were set at 120 kV and 100–200 mA, respectively, with a matrix of 512 × 512. Further, we collected the lung window level. The lung window was reconstructed with a 1 mm thin layer, and the scanned lung window level and width were 600 and 1,600 HU, respectively. This retrospective study was approved by the Ethics Committee of the Zhangjiajie City People's Hospital with waived informed consent requirement.

## 2.2. Follow-up assessment

The St. George's Respiratory Questionnaire (SGRQ) (10) is a clinical measurement designed to conduct health status self-assessments for patients with chronic airflow limitation, i.e., various respiratory diseases correlated with pulmonary function (11, 12). The questionnaire contains three main sections: symptoms (respiratory discomfort), activities (impact of dyspnea on daily tasks), and psychosocial impact (psychosocial impact of the disease). Typical SGRQ scores are <1, while higher scores indicate poorer health status and more impaired pulmonary function. Only SGRQ questionnaires filled out by patients without prompting from physicians were used in this study.

Two senior radiologists in the team performed diagnosis on the follow-up and hospital discharge CT scans collected from the 41 COVID-19 Delta variant survivors. We investigated imaging features until consensuses were reached on all diagnostic findings. Further, we categorized all CT scans into two groups: Normal and Abnormal CT (including fibrotic and Non-fibrotic changes). Fibrotic changes

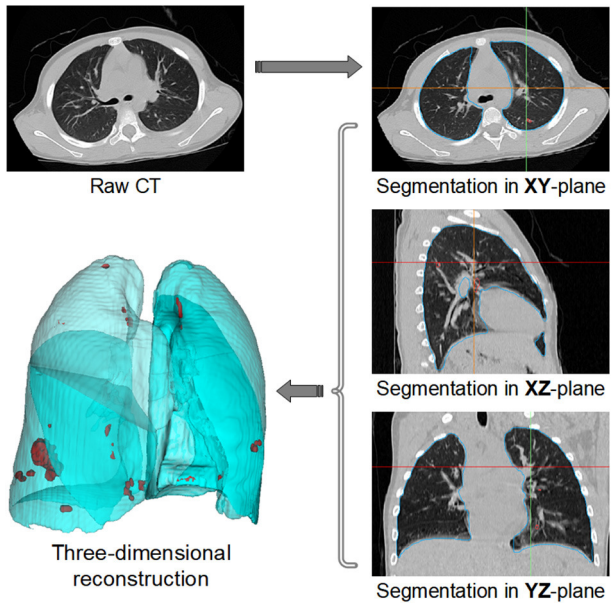


FIGURE 1
Workflow of traditional lesions segmentation.

include bronchiectasis, reticulations, nodules, punctuate calcification, irregular lines, and pulmonary bullae. Non-fibrotic changes include ground-glass opacity (GGO) and consolidation.

## 2.3. Computer-aided diagnosis

### 2.3.1. Lesions segmentation of COVID-19

Since the outbreak of COVID-19, a large number of researchers working in artificial intelligence have used deep learning models to assist in the diagnosis, treatment, and prognosis of COVID-19. Their initial goal was generally to save radiologists' time in reviewing medical images and to improve the accuracy of lesions identification. With computer-aided diagnosis methods, physicians can accurately obtain inflammation annotation in CT slices and accurately calculate the percentage of inflammation (POI) and its inflammatory density for each lung lobe or lung segment in a short time (2).

A typical workflow is shown in Figure 1. Firstly, raw CT scan is used as input for spatial normalization and signal normalization, and put into the standard space. In the standard space, the inflammation annotation is obtained using our 2.5D segmentation algorithm, i.e., the 3D data is split from three orthogonal directions (XY plane, XZ plane, YZ plane), and the segmentation is performed in each of these three directions using the U-Net network, and then the segmented results are integrated to obtain the final inflammation annotated mask. To better observe the distribution of inflammation in the lung, the results of the inflammation labeling are reconstructed in three dimensions, showing the distribution of inflammation in the lung in a clear and three-dimensional manner. This model can be deployed on an ordinary home computer to mark the inflammation and calculate the POI value of a CT scan within 1 min with the accuracy of more than 97%, which greatly improves the working efficiency of radiologists.

There are certain limitations to such an approach. Namely, the workflow of this model only allows marking visual lesions on regular CT scans, but not sub-visual lesions (i.e., it is almost impossible for a radiologist to see fibrosis lesions directly from ordinary CT scans). Among our team's latest published techniques (7), the deep lung parenchyma enhancing (DLPE) method was used to automatically mark visible and sub-visual COVID-19 lesions. In the follow-up study, we used the DLPE method to avoid the lesion-omissions issue that might occur in similar studies with traditional AI applications.studies.

### 2.3.2. Deep lung parenchyma enhancing

Deep lung parenchyma enhancing (DLPE) is a computer-aided detection (CADe) method for quantifying lung parenchymal lesions on chest CT. It can identify new lesions under the original lung window of hospitalized COVID-19 patients and survivors, whereas ordinary CT scans might neglect the sub-visual lesions. DLPE has a solid ability to predict sequelae such as pulmonary fibrosis. Its workflow includes three steps (shown in Figure 2):

(I) Segment the lung parenchyma, trachea and biood vessels. First, we used the proposed 2.5D segmentation algorithm to segment the lung. We further investigated the characteristics of the trachea and blood vessels and developed a two-stage segmentation model accordingly. The first stage determines the approximate extent of the trachea and blood vessels, reducing the search space by thousands of times, upon which the second stage achieves segmentation with higher stability and accuracy. Both stages were carried out by 2.5D segmentation models with feature-enhanced loss function. Finally, we developed a refined trachea and vascular mask.

(II) Deep lung parenchyma enhancing. We excluded the trachea and blood vessels from the lung to obtain a healthy lung parenchyma area. We further determined the position and width of the optimal window by calculating the median and standard deviation of the healthy lung parenchyma CT signal, which is generally used for observing lung parenchymal lesions. Finally, enhanced CT images, namely DLP-enhanced CTs, are obtained as parenchyma abnormalities are significantly enhanced compared to pulmonary windows.

(III) Segment the abnormalities in the lung parenchyma. We compared the enhanced parenchyma with the lung window. As the lesions were enhanced dozens of times, more previously neglected lesions were identified. Based on the DLP-enhanced CT, we built a 2.5D segmentation and quantization model which produced visible and sub-visual COVID-19 lesions from DLP-enhanced CT images. For simplicity reasons, the complete algorithm workflow was called the DLPE method.

### 2.3.3. Quantitative analysis

We used lung parenchyma lesion volume ratio and median lesion severity to measure the lesion severity of the CT images after DLPE process. The lesion volume ratio is defined as lesion volume divided by lung parenchymal volume:

$$Lesion\ volume\ ratio = V_{subvisual} \div V_{lung} \quad (1)$$

Where $V_{subvisual}$ is the volume of the sub-visual lesions, $V_{lung}$ is the volume of the lung parenchyma. Median lesion severity is the

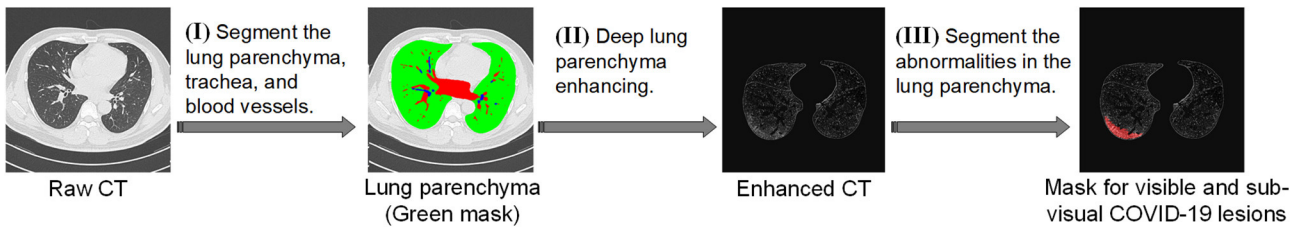


FIGURE 2
Workflow of deep lung parenchyma enhancing.

median value of the difference between the lesion and the baseline CT signal:

$$Median\ lesion\ severity = |\ Subvisual\ array - Baseline\ array\ | \quad (2)$$

Where *Baseline array* is the median CT signal value of healthy lung parenchyma, and *Subvisual array* is the CT signal value of lesion. Absorption ratio was used to describe the lesion changes from discharge to 6-month follow-up:

$$Absorption\ ratio = (Discharged - Followup) \div Discharged \quad (3)$$

Where *Discharged* represents the lesion volume ratio of hospital discharge CT, and *Followup* represents the lesion volume ratio of follow-up CT. An absorption ratio >0 reveals that the lung lesions have been absorbed since hospital discharge. As such, higher absorption ratios indicate better recovery. Vice versa, an absorption ratio less than or equal to 0 implies that the lung lesions of the patient have enlarged or remained unchanged since hospital discharge. In this case, higher absorption ratios indicate worse recovery; namely, patients may be affected to varying degrees by sequelae such as pulmonary fibrosis.

## 2.4. Statistical analysis

Statistical analyses were performed using Python 3.7. Without otherwise statement, measurement data were described by mean ± standard deviation or median (interquartile range). The Mann-Whitney U test and Kruskal-Wallis test were used to test independent samples. Count data were expressed as frequencies with percentages. $P < 0.05$ was considered to be statistically significant.

# 3. Results

## 3.1. Clinical characteristics

We retrospectively analyzed the clinical data of 41 follow-up patients. Clinical characteristics are shown in Table 1. The mean age of the patients was 42.2 ± 17.7 years, of which 13 were male patients (31.7%) and 28 were female patients (68.3%). The mean BMI of the patients was 25.2 ± 4.4 Kg/m$^2$. And there are 14.6% of patients meanwhile suffering from hypertension and 9.8% from diabetes. In the 6 months after discharge, some patients developed sequelae, which including: impaired sleep quality (17.1%), memory decline (26.8%), anxiety (21.9%), depression (4.9%), throat discomfort (9.8%), decline of vision acuity (12.2%), fatigue (7.3%), limbs weakness (4.9%), muscle or joint aches (12.2%) and hair loss (9.8%). None of the patients developed pulmonary-related sequelae such as dyspnea. We investigated the SGRQ scores of follow-up patients. The median of SGRQ scores was 0.041 and the interquartile range was (0, 0.104). All patients had SGRQ scores <1.

TABLE 1 Demographic and clinical characteristics of the enrolled COVID-19 patients.

| Characteristics | All patients ($n$ = 41) |
|---|---|
| Age, years | 42.2 ± 17.7 |
| Sex | |
| Men | 13 (31.7%) |
| Women | 28 (68.3%) |
| BMI | 25.2 ± 4.4 |
| Basic diseases | |
| Hypertension | 6 (14.6%) |
| Diabetes | 4 (9.8%) |
| Sequelae | |
| Impaired sleep quality | 7 (17.1%) |
| Memory decline | 11 (26.8%) |
| Anxiety | 9 (21.9%) |
| Depression | 2 (4.9%) |
| Throat discomfort | 4 (9.8%) |
| Decline of visual acuity | 5 (12.2%) |
| Fatigue | 3 (7.3%) |
| Arm Weakness | 2 (4.9%) |
| Muscle or joint pain | 5 (12.2%) |
| Hair loss | 4 (9.8%) |
| SGRQ score | 0.041 (0, 0.104) |

## 3.2. Chest CT evaluation

### 3.2.1. Imaging evaluation

This study collected the last CT scan before discharge and the 6-month follow-up CT scan from the 41 COVID-19 Delta variant

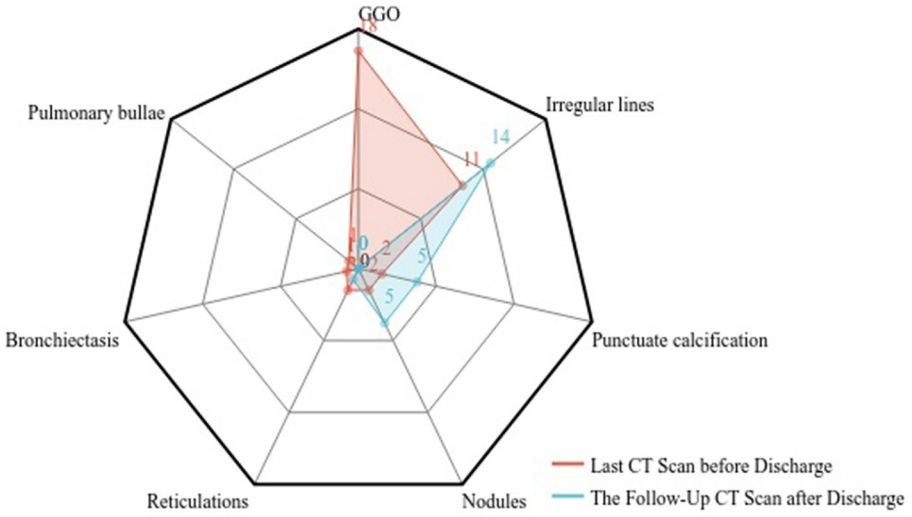


FIGURE 3
Imaging features of follow-up CT and hospital discharge CT.

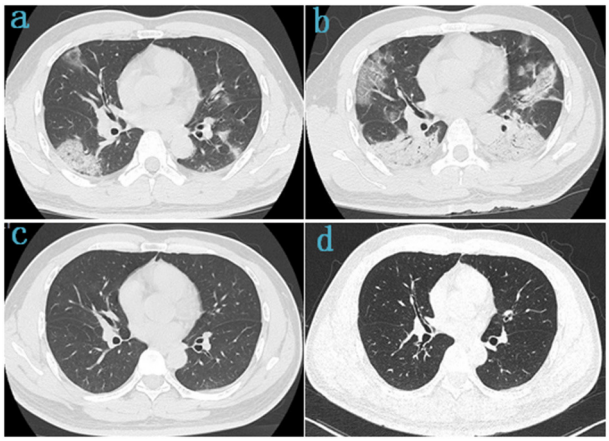


FIGURE 4
Chest CT of a 42-year-old male survivor of COVID-19 Delta variant. **(a)** On admission, baseline scan shows multiple bilateral ground-glass opacity with predominantly linear pattern and peripheral distribution, with air-bronchogram and tubular size increase of vessels in some lesions. **(b)** 4 days after admission, the lesions were significantly larger and more extensive than before, chest CT scans were subpleural ground-glass opacity that grew larger with crazy-paving pattern and consolidation. **(c)** Before discharge, the lesions in both lungs were basically absorbed. **(d)** At follow-up, the chest CT was basically normal, with a few fibrotic lesions were seen in the left lung.

survivors. Two experienced senior radiologists diagnosed all CT scans and summarized the imaging features. As shown in Figure 3, prominent abnormalities found on the CT before discharge include ground-glass opacity in 18 cases (43.9%) and irregular lines in 11 cases (26.8%). A few had punctuated calcification (4.9%), small 201 nodules (4.9%), reticulations (4.9%), and traction bronchiectasis (2.4%). In comparison, the ground-glass opacity was almost utterly unseen in the follow-up CT. Other present abnormalities include irregular lines in 14 cases (34.1%), punctuate calcification in five cases (12.2%), and small nodules in 5 cases (12.2%). Figure 4 demonstrates the lung recovery process of a typical COVID-19 Delta variant survivor.

### 3.2.2. CT slices after deep lung parenchyma enhancing

We enhanced all CT scans using the DLPE method to visualize all lesions, including sub-visual abnormalities. The comparison between the processed discharged CT and the processed follow-up CT (typical CT slices) shows that most of the lesions had been absorbed by the discharge, and the lung condition had improved considerably in 6-month (Figure 5). In addition, we measured the severity of the detected lesion using the lesion volume ratio and median lesion severity (Table 2). Of note, the lesion volume ratio and median lesion severity were significantly smaller at follow-up than at discharge ($p < 0.05$; Figure 6).

### 3.2.3. Absorption ratio at the 6-month follow-up

The mean value of the survivors' absorption ratio was 0.7147 (−1.0303, 0.9945). We grouped patients by gender, age, BMI, and COVID vaccination status, upon which we performed statistical analysis on the absorption ratio concerning median lesion severity (Table 3). While slight statistical difference was seen in the absorption ratio among patients in different BMI range groups ($p = 0.155$), the difference was stronger among different age groups ($p < 0.005$). No significant differences were seen among different gender groups or vaccination status groups.

## 3.3. Comparison with original COVID-19 follow-up

We compared the results of this follow-up study with those of the five original COVID-19 follow-up studies, as presented in Table 4. We found that Delta variant survivors had similar sequelae

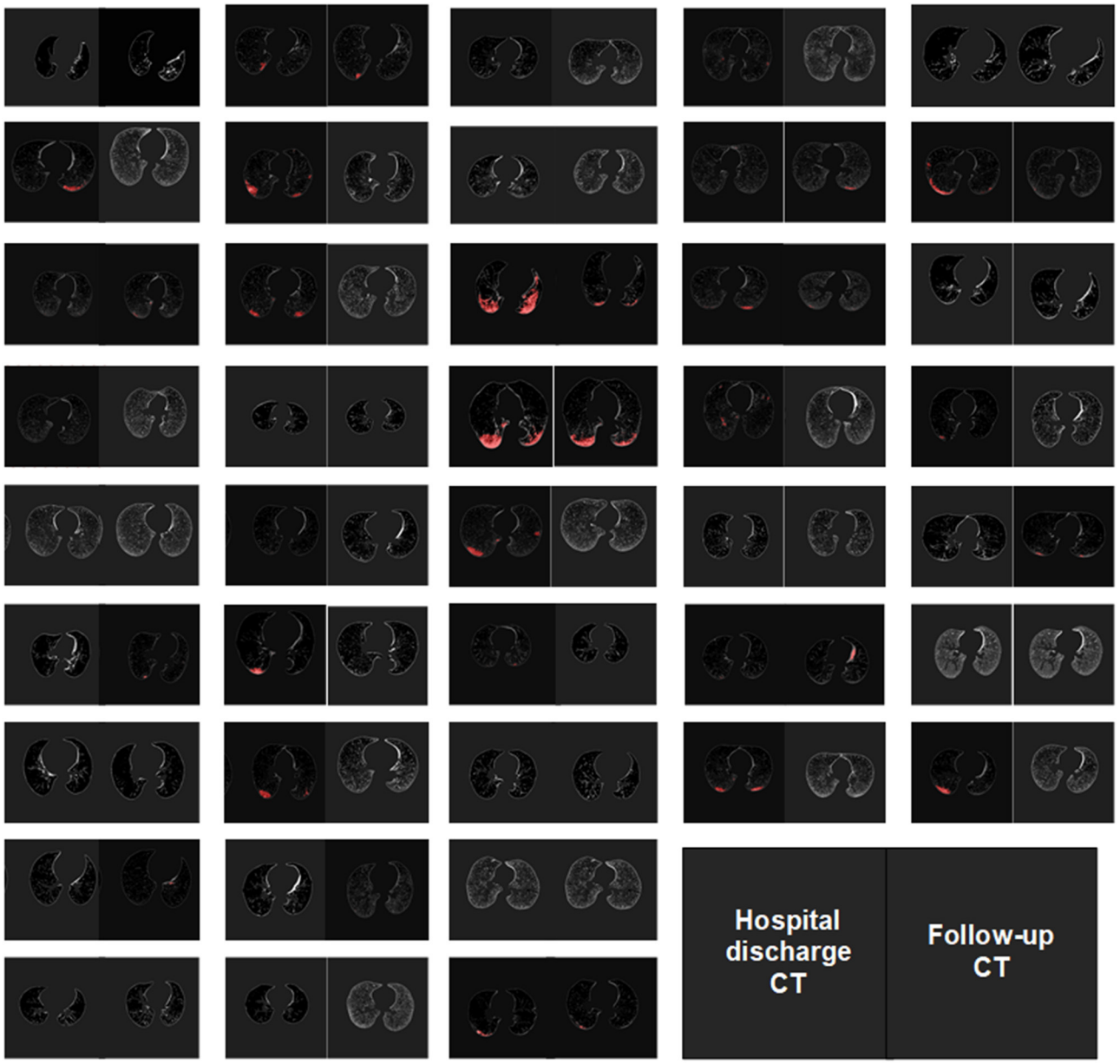


FIGURE 5
DLPE method applications in 41 patients.

TABLE 2 Changes in lesions under DLPE.

| | The hospital discharge CT ($n$ = 41) | The follow-up CT ($n$ = 41) | $P$-value |
|---|---|---|---|
| The lesion volume ratio | 0.0018 (0.0003, 0.0353) | 0.0004 (0, 0.0032) | 0.005 |
| Median lesion severity | 0.1329 (0.0632, 0.1892) | 0.0910 (0.0730, 0.1179) | 0.012 |

Data are medians, with ranges of quartiles in parentheses.

as the original COVID-19 survivors except for severe pulmonary sequelae such as chest tightness and dyspnoea. Further, we confirmed the absence of ground-glass opacity (GGO) and the mild fibrosis of lung lesions in the follow-up CT scans, suggesting that the lung prognosis of Delta variant patients is better than that of the original COVID-19 patients. Specifically, in CT imaging, 62–90% of original COVID-19 patients were discharged with GGO and 7.3–68% had consolidation. However, in our study, only 43.9% of Delta variant patients were discharged with GGO and without consolidation. And at 6-month follow-up, 27–44.8% of original COVID-19 patients still had GGO, while no GGO was found in Delta variant survivors.

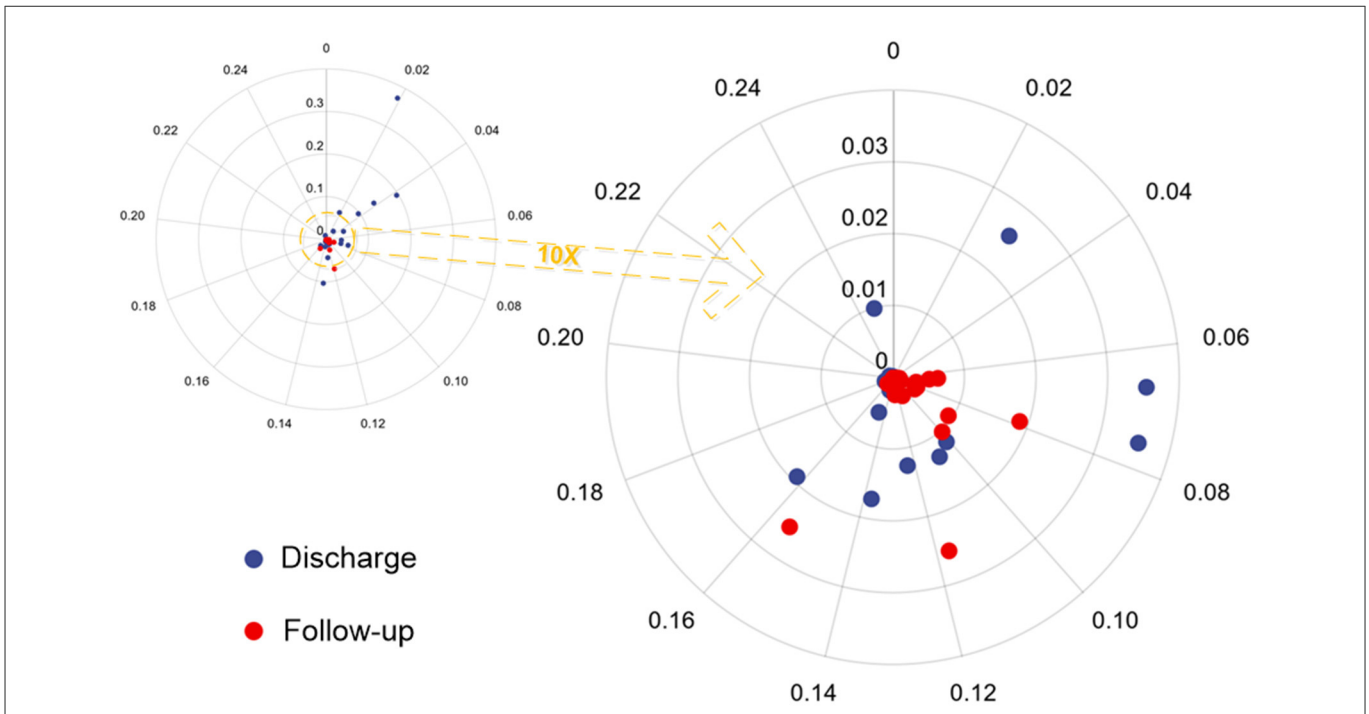


FIGURE 6
Polar scatter plot based on the lesion volume ratio and median lesion severity from hospital discharge and follow-up CT. The image on the left becomes the image on the right after enlarging 10 times.

TABLE 3 Absorption ration in lesions under DLPE.

| | Absorption rate of lesion volume ratio ($n$ = 41) | $P$-value | Absorption rate of median lesion severity ($n$ = 41) | $P$-value |
|---|---|---|---|---|
| Sex | | 0.575 | | 0.327 |
| Male ($n$ = 13) | 0.7147 (−1.3699, 0.9669) | | 0.4003 (−0.1075, 0.6171) | |
| Female ($n$ = 28) | 0.7518 (−0.7423, 0.9981) | | 0.2354 (−1.8101, 0.5783) | |
| Age range | | 0.005 | | 0.302 |
| ≤45 ($n$ = 22) | −0.7425 (−6.1833, 0.8158) | | 0.4624 (0.0100, 0.5805) | |
| 45–59 ($n$ = 12) | 0.9864 (0.3432, 0.9994) | | 0.0468 (−2.1080, 0.5253) | |
| ≥60 ($n$ = 7) | 0.9890 (0.7147, 1.0000) | | 0.3092 (−1.1145, 1.0000) | |
| BMI | | 0.155 | | 0.806 |
| ≤18.5 ($n$ = 4) | 0.9059 (−0.5578, 0.9997) | | 0.3943 (−0.5121, 0.8686) | |
| 18.5–23.9 ($n$ = 12) | 0.5655 (−1.0708, 0.9880) | | 0.1803 (−0.2579, 0.5757) | |
| ≥24 ($n$ = 18) | 0.8694 (−0.1381, 0.9997) | | 0.4271 (−0.5965, 05846) | |
| ≥28 ($n$ = 7) | −0.6000 (−10.746, 0.7147) | | 0.3059 (−2.4956, 0.5836) | |
| Vaccinated or not | | 0.626 | | 0.291 |
| Not ($n$ = 15) | 0.8238 (−1.0148, 0.9988) | | 0.3059 (−0.7877, 0.4709) | |
| Vaccinated ($n$ = 26) | 0.7077 (−1.0542, 0.9934) | | 0.4464 (−0.1467, 0.5847) | |

Data are medians, with ranges of quartiles in parentheses.

## 4. Discussion

Pulmonary fibrosis is an interstitial lung disease caused by intense fibroblast activation and extracellular matrix deposition in the lung, which often results in a range of sequelae such as reduced diffusion function of lung and labor dyspnea in patients. To date, studies on the follow-up of original COVID-19 patients have found that the most common abnormal lung changes in discharged patients are fibrosis and ground-glass opacity (GGO) (16, 18–20), consistent with SARS-related research foundings. Studies have shown that a shell

TABLE 4 Comparing the original COVID-19 follow-up to the Delta variant follow-up.

| References | Sample size | Age | Basic diseases | Sequelae | The hospital discharge CT | The follow-up CT |
|---|---|---|---|---|---|---|
| Dai et al. (13) | 50 | 48 ± 14 | Hypertension (18%), Diabetes (16%), Pulmonary disease (4%) | Decreased activity tolerance (18%), Cough (10%), Palpitation (6%), Depression (12.5%) | GGO (90%), Consolidation (54%), Reticulations (13%), Bronchiectasis (10%) | GGO (42%), Consolidation (20%), Bronchiectasis (6%), Reticulations (11%) |
| Han et al. (14) | 114 | 54 ± 12 | Hypertension (28%), Diabetes (11%), Chronic pulmonary (14%) | Dry cough (6.1%), Dyspnea (14%), Expectoration (10%) | GGO (62%), Consolidation (24%), Reticulations (14%) | Normal CT (38%), GGO (27%) Fibrotic-like changes (35%) |
| Jia et al. (15) | 205 | 56 ± 12 | Hypertension (36.9%), Diabetes (16%) | - | GGO (87%), Consolidation (7.3%), Reticulation (5.4%), Bronchiectasis (5.4%), Nodule (1%) | Normal CT (48.3%), GGO (28%), Consolidation (1.5%), Reticulation (22%), Bronchiectasis (13.7%), Nodule (7.3%) |
| Huang et al. (16) | 1,733 | 57 (47–65) | Hypertension (29%), Diabetes (12%), Pulmonary disease (2%) | Fatigue or muscle weakness (63%), Sleep difficulties (26%), Hair loss (22%), Smell disorder (11%), Palpitations (9%), Anxiety or depression (23%) | GGO (76%), Consolidation (23%), Irregular lines (30%) | Normal CT (47.3%), GGO (44.8%), Consolidation (1.1%), Irregular lines (15.9%) |
| Caruso et al. (17) | 118 | 65 ± 12 | Hypertension (34%), Diabetes (9.0%) | Cough (24%), Dyspnea (42%), Hair loss (20%), Decline of visual acuity (12%) | GGO (86%), Consolidation (68%), Fibrotic-like changes (55%) | Normal CT (28%), GGO (42%), Consolidation (1.7%), Fibrotic-like changes (72%) |
| Our study | 41 | 42.2 ± 17.7 | Hypertension (14.6%), Diabetes (9.8%) | Impaired sleep quality (17.1%), Memory decline (26.8%), Anxiety or depression (26.8%), Fatigue (7.3%), Hair loss (9.8%) | Normal CT (34.1%), GGO (43.9%), Fibrotic-like changes (34.1%) | Normal CT (51.2%), GGO (0%), Fibrotic-like changes (48.8%) |

nucleoprotein from SARS can bind to SMAD3, a cellular protein that activates a signaling pathway to promote collagen and plasminogen protein inhibitor production, further leading to the formation of fibrosis in the lungs (21). At present, no similar protein has been identified in COVID-19-related studies, so the current understanding of the prognosis of fibrotic changes in COVID-19 patients remains unclear. Caruso et al. (17) reported that residual GGO was found on lung CT in 42% of original COVID-19 patients and fibrotic changes were present in 72%; a proportion of patients were discharged with dry cough (24%), dyspnoea (42%) and many other lung-related sequelae. In a 6-month follow-up study of 114 original COVID-19 patients, Han et al. (14) found that 35% of patients had residual fibrotic changes in lungs and 14% had dyspnoea. Besides, seriously ill hospital patients developed more severe fibrosis, which was found to restain even at the 1-year follow-up. Similar results have been seen in a 15-year follow-up study of SARS patients (22). Pan et al. (23) found that 61% of patients had complete resolution of abnormal lung changes by 3 months after discharge; at the 1-year follow-up, 25% of patients still had residual fibrotic changes, but it was unclear whether this fibrosis can be further absorbed.

There are few follow-up studies on the Delta variant; hence we conducted this study to raise awareness of pulmonary fibrotic changes and other COVID-19 Delta variant sequelae. We collected hospital discharge CT scans and 6-month follow-up CT scans from 41 Delta variant survivors. We found that more than half of the patients (51.2%) had no residual fibrosis in their follow-up CT of lung after 6-month discharge and that the GGO was almost completely absorbed. The changes of fibrotic presented in follow-up CT of the remaining patients were predominantly irregular lines (34.1%) and small nodules (12.2%), and the patients had a very mild degree of fibrosis. Artificial intelligence and deep learning techniques are widely adopted in current radiology research as they enable physicians to segment infected lesions accurately and implement precision medicine. This research used the previously proposed deep lung parenchyma enhancing (DLPE) model (7) to automatically outline all lung lesions, including conventional and sub-visual lesions. Quantitative assessments was further conducted to evaluate patients' recovery, comparing the calculations of lung lesions on discharged CT and follow-up CTs. We found that most patients had largely dissipated lung lesions at discharge. And after 6 months, re-quantification of lung lesions on follow-up CT revealed a small lesion volume ratio (mean = 0.04%), leading us to assume that the lung fibrosis had been slowly absorbed over time. In the meantime, a proportion of patients developed increased fibrosis (i.e., negative absorption ratio), yet the observed fibrosis levels were less notable and the pulmonary diffusion function remained unaffected. This suggests that DLPE might be deficient in capturing some existing lung damage (early lung damage). It is also possible that the patients experienced

other lung damage after discharge, which also caused fibrosis but was unrelated to the COVID-19 infection.

There are multiple factors that affect the prognosis of COVID-19 survivors, including gender, age, body mass index (BMI), and vaccination status. Sylvester et al. (24) have shown that female patients are at a higher risk of developing long COVID syndrome due to differences in the immune system between the sexes. Obesity, measured by BMI, is considered a risk factor strongly associated with the severity of COVID-19 infection and mortality (25). Vaccination is vital in preventing and treating COVID-19 (26), as it reduces the risk of hospitalization and severe sequelae after infection. Age is also strongly associated with patient prognosis, as Huang et al. (16) found that a 10-year increase in age of COVID-19 patients was associated with a 27% increase in pulmonary diffusion dysfunction and a 4% decrease in the absorption ratio. Results of comparing the absorption ratio by gender, age, BMI, and vaccination status show that the absorption rate of lesion volume ratio was significantly different for different age groups ($p < 0.005$) and slightly different for different BMI range groups ($p = 0.155$). No differences were seen between gender and vaccination status groups, presumably due to the small sample size. In addition, sequelae such as impaired sleep quality, memory loss and anxiety were found in Delta variant survivors, similar to those noted in the follow-up study of 1,733 original COVID-19 patients by Huang et al. (16).

This study used the St. George's Respiratory Questionnaire (SGRQ) (10) to assess the Delta variant survivors and obtained a median SGRQ of 0.041 (0, 0.104), which is within the normal range. The result confirmed that the participating survivors had good pulmonary recovery with no significant pulmonary sequelae, and the infection did not significantly affect their quality of life.

The main contributions of this paper are as follow. 1) This is a 6-month follow-up study on discharged COVID-19 Delta variant patients, with data gathered from an earlier cohort of Delta variant patients in China. As limited follow-up studies have been done on the Delta variant, this research is prevailing in broadening the understanding of the Delta variant. 2) Multiple computer-aided techniques, such as deep learning and quantitative analysis, are used to compare the follow-up outcomes of Delta variant survivors and original COVID-19 survivors. Critical findings include Delta variant survivors had a better prognosis than original COVID-19 survivors. 3) In this study, we applied the previously proposed sub-visual lesion observation method (i.e., DLPE) for the first time. This novel lesion segmentation method enabled clinicians to observe and analyze lung lesion changes in Delta variant survivors in greater detail, which is of great value and guidance for the COVID-19 prognostic assessment. Indeed, this research has some shortcomings. Firstly, we lacked direct information showing the patients' pulmonary diffusion functions because the selected cohort did not undergo a complete pulmonary function test during hospitalization and follow-up examination. Secondly, this study was geared toward Delta variant survivors diagnosed in Zhangjiajie city, China, in 2021, resulting in a small sample size.

## 5. Conclusion

In this study, we analyzed the discharged CT scans, 6-month follow-up CT scans, and some clinical indicators of 41 COVID-19 Delta variant survivors to assess fibrosis absorption and sequelae comprehensively. This paper marked the first application of the deep lung parenchyma enhancing method to quantify the extent of lung lesions on hospital discharge and follow-up CTs. We found that the lung lesions had primarily dissipated by discharge and that the lesion volume ratio in follow-up CT was generally small in most cases. We also compared the absorption ratios by gender, age, BMI, and vaccination status. Results have shown that the absorption ratios were significantly different for patients in different age groups and slightly different for different BMI range groups. Statistics and analysis of Delta variant sequelae are also provided, pointing out the primarily experienced sequelae, including impaired sleep quality, memory loss, and anxiety. In conclusion, this study aims to use computer-aided AI methods to raise awareness of the COVID-19 Delta variant and promote the prognosis of the disease. While confounding progress has been made in understanding pulmonary sequelae associated with the Delta variant, it is absolutely necessary to carry on the investigation of COVID-19 and the evolution of prognosis clinical care continuously in the future.

## Data availability statement

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

## Ethics statement

The studies involving human participants were reviewed and approved by Medical Ethics Committee of Zhangjiajie City People's Hospital. The patients/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

ML and ZQ conceived this study. JH, ZSu, MZ, HL, CC, and MX collected all the chest CT scans and clinical data. RL, NB, and HL were the developers of computer-aided diagnosis methods. RL and NB completed the data analysis. JH, ZSu, and MZ wrote about the imaging findings of patients. JH, RL, NB, and ZSh drafted the manuscript. All authors were involved in the finalization of the manuscript and approved the manuscript.

## Funding

This work was financially supported by the Key R&D project in Heilongjiang Province (No. 2022ZX01A30), the General Program of Natural Science Foundation of Hunan Province of China (No. 2017JJ2261), the Science and Technology Program of Suzhou (Nos. ZXL2021431 and RC2021130), the Zhangjiajie Yongding District Science and Technology Innovation Program Project (2022), and the Fundamental Research Funds for the Central Universities (No. 2572020DR10).

## Acknowledgments

All authors pay tribute to COVID-19 patients involved in this study and all frontline healthcare workers involved in the diagnosis, treatment and follow-up in Zhangjiajie, China. And the authors would also like to thank Rui Ding for advice on language polishing.

## Conflict of interest

ZQ was employed by Heilongjiang Tuomeng Technology Co., Ltd.

The remaining 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. Johns Hopkins University. *COVID-19 Dashboard*. (2022). Available online at: https://coronavirus.jhu.edu/map.html

2. Bai N, Lin R, Wang Z, Cai S, Huang J, Su Z, et al. Exploring new characteristics: using deep learning and 3D reconstruction to compare the original COVID-19 and its delta variant based on chest CT. *Front Mol Biosci*. (2022) 9:836862. doi: 10.3389/fmolb.2022.836862

3. Briganti G, le Moine O. Artificial intelligence in medicine: today and tomorrow. *Front Med*. (2020) 7:27. doi: 10.3389/fmed.2020.00027

4. Park HJ, Park B, Lee SS. Radiomics and deep learning: hepatic applications. *Korean J Radiol*. (2020) 21:387–401. doi: 10.3348/kjr.2019.0752

5. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med*. (2019) 25:65–9. doi: 10.1038/s41591-018-0268-3

6. Mondal MRH, Bharati S, Podder P. Diagnosis of COVID-19 using machine learning and deep learning: a review. *Curr Med Imaging*. (2021) 17:1403–18. doi: 10.2174/1573405617666210713113439

7. Zhou L, Meng X, Huang Y, Kang K, Zhou J, Chu Y, et al. An interpretable deep learning workflow for discovering subvisual abnormalities in CT scans of COVID-19 inpatients and survivors. *Nat Mach Intell*. (2022) 4:494–503. doi: 10.1038/s42256-022-00483-7

8. National Health Commission and State Administration of Traditional Chinese Medicine. *Diagnosis and Treatment Protocol for Novel Coronavirus Pneumonia (Trial Version 8)*. National Health Commission of the People's Republic China (2020). Available online at: http://www.gov.cn/zhengce/zhengceku/2021-04/15/content_5599795.htm

9. World Health Organization. *Provisional Guidelines on Standard International Age Classifications*. Geneva: WHO (1982).

10. Jones PW, Quirk FH, Baveystock CM, Littlejohns P. A self-complete measure of health status for chronic airflow limitation. The St. George's respiratory questionnaire. *Am Rev Respirat Dis*. (1992) 145 6:1321–7. doi: 10.1164/ajrccm/145.6.1321

11. Santus P, Tursi F, Croce G, Simone CD, Frassanito F, Gaboardi P, et al. Changes in quality of life and dyspnoea after hospitalization in COVID-19 patients discharged at home. *Multidisc Respirat Med*. (2020) 15:713. doi: 10.4081/mrm.2020.713

12. Liu R, peng Du Y, He B. [Relationship between SGRQ score and pulmonary function and its influencing factors in patients with chronic obstructive pulmonary disease]. *Zhonghua yi xue za zhi*. (2011) 91:1533–7. doi: 10.3760/CMA.J.ISSN.0376-2491.2011.22.007

13. Dai S, Zhao B, Liu D, Zhou Y, Liu Y, Lan L, et al. Follow-up study of the cardiopulmonary and psychological outcomes of COVID-19 survivors six months after discharge in Sichuan, China. *Int J Gen Med*. (2021) 14:7207–17. doi: 10.2147/IJGM.S337604

14. Han X, Fan Y, Alwalid O, Li N, Jia X, Yuan M, et al. Six-month follow-up chest CT findings after severe COVID-19 pneumonia. *Radiology*. (2021) 299:E177-E186. doi: 10.1148/radiol.2021203153

15. Jia X, Han X, Cao Y, Fan Y, Yuan M, Li Y, et al. Quantitative inspiratory-expiratory chest CT findings in COVID-19 survivors at the 6-month follow-up. *Sci Rep*. (2022) 12:7402. doi: 10.1038/s41598-022-11237-1

16. Huang C, Huang L, ming Wang Y, Li X, Ren L, Gu X, et al. 6-month consequences of COVID-19 in patients discharged from hospital: a cohort study. *Lancet*. (2021) 397:220–32. doi: 10.1016/S0140-6736(20)32656-8

17. Caruso D, Guido G, Zerunian M, Polidori T, Lucertini E, Pucciarelli F, et al. Postacute sequelae of COVID-19 pneumonia: 6-month chest CT follow-up. *Radiology*. (2021) 301:E396-E405. doi: 10.1148/radiol.2021210834

18. Zhou F, Tao M, Shang L, Liu Y, Pan G, Jin Y, et al. Assessment of sequelae of COVID-19 nearly 1 year after diagnosis. *Front Med*. (2021) 8:717194. doi: 10.3389/fmed.2021.717194

19. Huang L, Yao Q, Gu X, Wang Q, Ren L, ming Wang Y, et al. 1-year outcomes in hospital survivors with COVID-19: a longitudinal cohort study. *Lancet*. (2021). 398:747–58. doi: 10.1016/S0140-6736(21)01755-4

20. Wu X, Liu X, Zhou Y, ying Yu H, Li R, Zhan Q, et al. 3-month, 6-month, 9-month, and 12-month respiratory outcomes in patients following COVID-19-related hospitalisation: a prospective study. *Lancet Respirat Med*. (2021) 9:747–54. doi: 10.1016/S2213-2600(21)00174-0

21. Zhao X, Nicholls JM, Chen YG. Severe acute respiratory syndrome-associated coronavirus nucleocapsid protein interacts with Smad3 and modulates transforming growth factor-$\beta$ signaling. *J Biol Chem*. (2008) 283:3272–80. doi: 10.1074/jbc.M708033200

22. Zhang P, Li J, Liu H, Han N, Ju J, Kou YH, et al. Long-term bone and lung consequences associated with hospital-acquired severe acute respiratory syndrome: a 15-year follow-up from a prospective cohort study. *Bone Res*. (2020) 8:8. doi: 10.1038/s41413-020-00113-1

23. Pan F, Yang L, Liang B, Ye T, Li L, Li L, et al. Chest CT patterns from diagnosis to 1 year of follow-up in patients with COVID-19. *Radiology*. (2021) 302:709–19. doi: 10.1148/radiol.2021211199

24. Sylvester SV, Rusu R, Chan B, Bellows M, O'Keefe C, Nicholson SC. Sex differences in sequelae from COVID-19 infection and in long COVID syndrome: a review. *Curr Med Res Opin*. (2022) 38:1391–9. doi: 10.1080/03007995.2022.2081454

25. Piernas C, Patone M, Astbury NM, Gao M, Sheikh A, Khunti KK, et al. Associations of BMI with COVID-19 vaccine uptake, vaccine effectiveness, and risk of severe COVID-19 outcomes after vaccination in England: a population-based cohort study. *Lancet Diabetes Endocrinol*. (2022) 10:571–80. doi: 10.1016/S2213-8587(22)00158-9

26. Al-Aly Z, Bowe B, Xie Y. Long COVID after breakthrough SARS-CoV-2 infection. *Nat Med*. (2022) 28:1461–7. doi: 10.1038/s41591-022-01840-0

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 22 February 2023
DOI 10.3389/fmed.2023.1085437



OPEN ACCESS

EDITED BY
Jinming Duan,
University of Birmingham, United Kingdom

REVIEWED BY
Zhaowen Qiu,
Northeast Forestry University, China
J. Jianbo,
University of Birmingham, United Kingdom

*CORRESPONDENCE
Qi Yang
✉ yangyangqiqi@gmail.com
Lihua Wang
✉ wanglh211@163.com

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Nuclear Medicine, a section of the journal Frontiers in Medicine

RECEIVED 31 October 2022
ACCEPTED 03 February 2023
PUBLISHED 22 February 2023

CITATION
Gao H, Bian Y, Cheng G, Yu H, Cao Y, Zhang H, Wang J, Li Q, Yang Q and Wang L (2023) Identifying patients with acute ischemic stroke within a 6-h window for the treatment of endovascular thrombectomy using deep learning and perfusion imaging. *Front. Med.* 10:1085437. doi: 10.3389/fmed.2023.1085437



# Identifying patients with acute ischemic stroke within a 6-h window for the treatment of endovascular thrombectomy using deep learning and perfusion imaging

Hongyu Gao[1†], Yueyan Bian[2†], Gen Cheng[3], Huan Yu[4], Yuze Cao[5], Huixue Zhang[1], Jianjian Wang[1], Qian Li[1], Qi Yang[2*] and Lihua Wang[1*]

[1]Department of Neurology, The Second Affiliated Hospital, Harbin Medical University, Harbin, Heilongjiang, China, [2]Department of Radiology, Beijing Chaoyang Hospital, Capital Medical University, Beijing, China, [3]Neusoft Medical System Co., Beijing, China, [4]Department of Radiology, Liangxiang Teaching Hospital, Capital Medical University, Beijing, China, [5]Department of Neurology, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, China

**Introduction:** It is critical to identify the stroke onset time of patients with acute ischemic stroke (AIS) for the treatment of endovascular thrombectomy (EVT). However, it is challenging to accurately ascertain this time for patients with wake-up stroke (WUS). The current study aimed to construct a deep learning approach based on computed tomography perfusion (CTP) or perfusion weighted imaging (PWI) to identify a 6-h window for patients with AIS for the treatment of EVT.

**Methods:** We collected data from 377 patients with AIS, who were examined by CTP or PWI before making a treatment decision. Cerebral blood flow (CBF), time to maximum peak (Tmax), and a region of interest (ROI) mask were preprocessed from the CTP and PWI. We constructed the classifier based on a convolutional neural network (CNN), which was trained by CBF, Tmax, and ROI masks to identify patients with AIS within a 6-h window for the treatment of EVT. We compared the classification performance among a CNN, support vector machine (SVM), and random forest (RF) when trained by five different types of ROI masks. To assess the adaptability of the classifier of CNN for CTP and PWI, which were processed respectively from CTP and PWI groups.

**Results:** Our results showed that the CNN classifier had a higher performance with an area under the curve (AUC) of 0.935, which was significantly higher than that of support vector machine (SVM) and random forest (RF) ($p = 0.001$ and $p = 0.001$, respectively). For the CNN classifier trained by different ROI masks, the best performance was trained by CBF, Tmax, and ROI masks of Tmax $>$ 6 s. No significant difference was detected in the classification performance of the CNN between CTP and PWI (0.902 vs. 0.928; $p = 0.557$).

**Discussion:** The CNN classifier trained by CBF, Tmax, and ROI masks of Tmax $>$ 6 s had good performance in identifying patients with AIS within a 6-h window for the treatment of EVT. The current study indicates that the CNN model has potential to be used to accurately estimate the stroke onset time of patients with WUS.

KEYWORDS

**acute ischemic stroke, endovascular thrombectomy, stroke onset time, deep learning, perfusion imaging**

# Introduction

In the guidelines for the early management of patients with acute ischemic stroke (AIS) published by the American Heart Association/American Stroke Association (AHA/ASA) in 2019, recombinant tissue-type plasminogen activator (rt-PA) thrombolysis and endovascular thrombectomy (EVT) are recommended to treat patients with AIS (1). Both of these are performed mainly for patients within a specific window of time from stroke onset, which are 4.5 h for rt-PA thrombolysis and 6-h for EVT. However, because 14–29.6% of patients with AIS are attacked during their sleep, which is called wake-up stroke (WUS) (2), their accurate stroke onset time cannot be ascertained to calculate this window. This means that other examinations are needed to estimate the stroke onset time of patients with WUS before treatment of rt-PA thrombolysis or EVT.

In previous studies, multi-modality imaging has been shown to have strong potential for accurately estimating the stroke onset time (3–8). In rt-PA thrombolysis treatment, the imaging biomarker of intensity mismatch between diffuse weighted imaging (DWI) and fluid-attenuated inversion recovery (FLAIR) is used to detect patients within a 4.5 h window (4), which means that the stroke onset time of patients with an unknown time and with a DWI–FLAIR mismatch biomarker is within a 4.5 h window for rt-PA thrombolysis treatment. In order to further explore the relationship between imaging biomarker and stroke onset time, Kong et al. constructed a decoder–encoder network to extract features using DWI, FLAIR, and time to maximum peak (Tmax) images, which can classify patients within a 4.5 h window for rt-PA thrombolysis treatment (8). This means that a machine learning classifier based on an imaging biomarker can accurately estimate the stroke onset time.

However, there is not a typical imaging biomarker to identify a 6-h treatment window for EVT. Some potential imaging biomarkers were found in previous works (9–12), such as a reduction in cerebral blood flow (CBF) and a delayed Tmax. The progression of AIS can be directly expressed by changes of an infarct core and ischemic region (12–14). An infarct core and penumbra region can be estimated using perfusion map images, which include CBF, cerebral blood volume (CBV), mean transit time (MTT), and Tmax. The infarct core is defined as the region of CBF reductions to <30% compared contralateral hemispheres (CBF < 30%) for computed tomography perfusion (CTP), or apparent diffusion coefficient (ADC) values <620. The ischemic region includes the infarct core and penumbra region, which is the region of Tmax > 6 s (9). Furthermore, Olivot et al. (15) estimated the benign hypoperfusion, ischemic, and infarct core regions only by different Tmax thresholds, which are, respectively, >4, >6, and >10 s. Thus, CBF and Tmax are significantly related to the stroke onset time.

The present study sought to combine the deep learning technique with perfusion map images (CBF and Tmax), which was processed from CTP or perfusion weighted imaging (PWI), to identify patients with AIS within a 6-h window for the treatment of EVT. We constructed a classifier based on a convolutional neural network (CNN), which was trained by CBF, Tmax, and a region of interest (ROI) mask. Compared to previous studies, to classify patients within a 4.5 h window for rt-PA thrombolysis treatment, our method is able to identify them within a 6-h window for the treatment of EVT. Meanwhile, our method has stable performance for both CTP and PWI. It means that our method enables compatible with both magnetic resonance (MR) and computed tomography (CT) devices, rather than only MR devices. Thus, our method has more potential to be used widely in hospitals, especially primary hospitals.

# Methods

## Patients

The local institutional review board approved this retrospective analysis, and the patient had signed the informed consent form. Also, patient records and images (including the source or raw imaging data) were anonymized before image analysis. Anonymized data are available on reasonable request to the corresponding author, and the data collected in the repository will be made accessible to qualified researchers worldwide, based on the recommendations of a scientific committee that will evaluate proposed research projects. The confidentiality of patients' information will be rigorously protected.

We recruited patients with AIS between April 2020 and April 2021 from the eStroke China national thrombolytic and thrombectomy imaging platform. Thirteen subcenters are registered on the platform and upload CTP or PWI images examined from patients with AIS before treatment to the eStroke platform. In addition, clinical information, including age, sex, national institute of health stroke scale (NIHSS), and exact stroke onset time are recorded. In order to align the examination performance among subcenters, we adjusted imaging protocols based on different device types, which are summarized in Table 1. To avoid the bias of the stroke onset time of patients with AIS, the data were collected by neurologists with more than 5 years of clinical experience, and they were recorded fully on the eStroke platform. Patients were recruited into this study based on the following criteria: (1) AIS due to anterior circulation artery (ACA) occlusion; (2) the recorded exact stroke onset time; (3) the recorded time of initial pretreatment imaging; (4) examined CTP or PWI before treatment; and (5) complete clinical information. All patients were anonymously recruited, and they were informed of and agreed to the study. The dataset will be released on the website https://github.com/bianyueyan/CNN-EVT.

## Experimental design

According to previous works, the stroke onset time is correlated with CBF/ADC, Tmax, and changes in the benign hypoperfusion, ischemic, and infarct core regions. These regions can be estimated by different thresholds in CBF and Tmax (9, 15). Therefore, three factors including CBF/ADC, Tmax and the region of diseased hemispheres, are correlated with the identification of the stroke onset time. In order to enable to be compatible with both CT and MR examinations, we chose CBF, Tmax and the region of diseased hemispheres as input images. In this study, we constructed three types of classifiers, namely, support vector machine (SVM), random

TABLE 1 List of imaging protocols.

| CTP protocols | | | | | |
|---|---|---|---|---|---|
| Subcenter | Slice thickness (mm) | No. of slices | Total coverage (mm, cc) | kVp | mAs |
| Center1 | 5 | 480 | 80 | 80 | 200 |
| Center2 | 5 | 1,080 | 80 | 80 | 223 |
| Center3 | 5 | 460 | 80 | 80 | 176 |
| Center4 | 5 | 360 | 80 | 80 | 211 |
| Center5 | 5 | 336 | 80 | 80 | 200 |
| Center6 | 5 | 1,566 | 80 | 80 | 124 |
| Center7 | 10 | 506 | 80 | 80 | 350 |
| Center8 | 5 | 864 | 80 | 80 | 158 |
| Center9 | 5 | 704 | 80 | 80 | 176 |
| Center10 | 5 | 360 | 80 | 80 | 264 |
| PWI protocols | | | | | |
| Subcenter | Slice thickness (mm) | FOV ($mm^2$) | Bandwidth (kHz) | TR/TE (ms) | Acquisition matrix |
| Center11 | 5 | 230 × 230 | 28.3 | 1,590/32 | 128 × 128 |
| Center12 | 5 | 230 × 230 | 31.2 | 1,500/19.2 | 96 × 128 |
| Center13 | 5 | 230 × 230 | 29.4 | 1,740/32 | 128 × 128 |

cc, craniocaudal; mAs, milliampere-seconds; kVp, kilovoltage peak; FOV, field of view; TR, repetition time; TE, echo time.

forest (RF), and CNN, to identify patients with AIS within a 6-h window for the treatment of EVT. These classifiers were trained by three channels of images. The first channel was CBF images, the second was Tmax images, and the third was ROI mask, which was one of the regions of CBF < 30%, Tmax > 4 s, Tmax > 6 s, Tmax > 8 s, and Tmax > 10 s.

In order to compare the performance among the different classifiers (SVM, RF, and CNN), each classifier was trained by three channels of images, consisting of CBF, Tmax, and ROI masks of Tmax > 6 s. Meanwhile, for observing the differences from the ROI masks (CBF < 30%, Tmax > 4 s, Tmax > 6 s, Tmax > 8 s, and Tmax > 10 s), the CNN classifier was trained by CBF, Tmax, and each ROI mask. Through the above process, the classifier with the best performance was selected. Finally, we trained the best classifier using CBF, Tmax, and ROI mask, respectively, from CTP and PWI to compare their agreement.

## Image preprocessing

The CTP and PWI of patients with AIS were examined before the treatment, and then intra-phase rigid registration was performed to correct motion artifacts. After this, the images were smoothed using a Gaussian filter with a kernel with a width of 2.5 mm. In order to reduce disturbance of skull and cerebrospinal fluid (CSF), the images were segmented using BET2 (16) and the thresholding method, respectively, and then the ROI was selected while the rest of the image was excluded. Perfusion parameter maps, including CBF, CBV, MTT, and Tmax, were constructed by block-circulant singular value decomposition (bSVD) provided by the eStroke platform. Perfusion parameter maps were resampled to the spacing of 1 mm in the *x*, *y*, and *z* directions to reduce the impact of image resolutions. The resampled images were chosen as the analytical basis of feature extraction, training, and testing datasets.

According to previous studies, ROI masks segmented by different thresholds based on CBF and Tmax express the progression of AIS, which are strongly related to the stroke onset time (3–8). In order to compare their performances in estimating the stroke onset time, we segmented the ROI masks by CBF < 30%, Tmax > 4 s, Tmax > 6 s, Tmax > 8 s, and Tmax > 10 s.

## Feature extraction

Features for training machine learning methods, including SVM and RF, were generated based on CBF, Tmax images, and ROI masks, which mainly included first-order descriptive statistics, features of shape, gray level co-occurrence matrix (GLCM) features, gray level dependance matrix (GLDM) features, and gray level size zone matrix (GLSZM) features. All of the features are shown in Table 2. They were extracted with the Radiomics module in the 3D Slicer software, version 4.11 (NA-MIC, NAC, BIRN, NCIGT, and the slicer community, USA). After extracting the initial features, the principal component analysis (PCA) approach was applied to reduce dimensionality and decrease the dependance on the number of training data.

## Classifier construction

We compared the performance of three types of classifiers, namely, SVM, RF, and CNN in identifying a 6-h window for the

TABLE 2 List of features.

| Feature class | No. of features | Feature name |
|---|---|---|
| Shape | 9 | Maximum 2D diameter, maximum 3D, diameter, mesh volume, minor axis length, sphericity, surface area, surface volume ratio, voxel volume |
| First order descriptive statistics | 15 | Energy, entropy, interquartile range, kurtosis, maximum, mean absolute deviation, mean, median, minimum, robust mean absolute deviation, root mean squared, skewness, total energy, uniformity, variance |
| GLCM | 12 | Autocorrelation, cluster prominence, cluster shade, cluster tendency, contrast, correlation, difference average, difference entropy, difference variance, joint average, sum entropy, sum squares |
| GLDM | 5 | Dependence entropy, dependence variance, gray level non-uniformity, gray level variance, high gray level emphasis |
| GLSZM | 10 | Gray level non-uniformity, gray level non-uniformity normalized, gray level variance, high gray level zone emphasis, large area emphasis, large area high gray level emphasis, large area low gray level emphasis, low gray level zone emphasis |

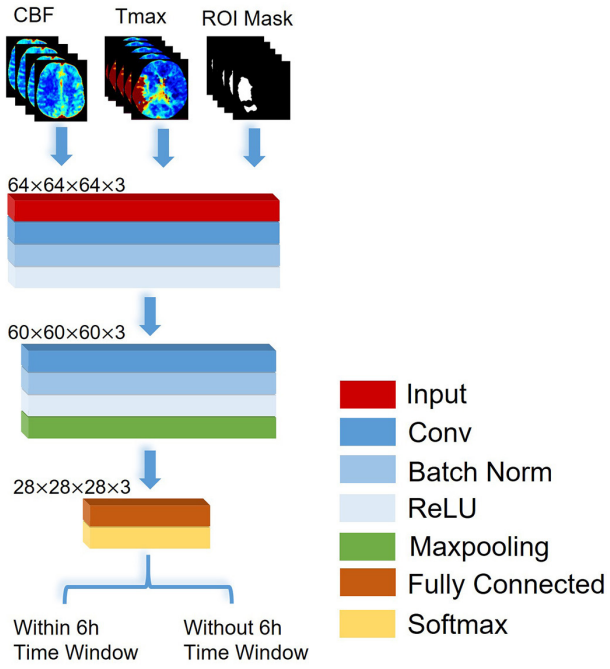


FIGURE 1
The architecture of the CNN proposed to identify patients with AIS within a 6-h window for the treatment of EVT.

treatment of EVT. Briefly, SVM is a supervised machine learning algorithm, mainly used to process classification and regression tasks. The objective of SVM is to find a hyperplane in a N-dimensional space that is defined by the number of features in order to classify the dataset (17). RF is an ensemble learning method that can operate a variety of tasks, including regression and classification. It commonly constructs a multitude of decision trees during the training time. In a classification task, RF creates many decision trees on data samples, each of which votes based upon the results of the prediction. Finally, the output of RF means the class selected by the most trees (18). A CNN is a feed-forward neural network, which is used to handle computer vision tasks such as image classification, object detection, and image recognition (19).

In this study, a CNN was constructed based on VGGNet with 2 convolutional blocks (20), which consisted of a structure of eleven layers: an input layer, three convolutional layers, two batch normalization layers, two rectified linear unit (ReLU) layers, a max pooling layer, a fully connected layer, and a soft-max layer, which are shown in Figure 1. According to the previous works, the stroke onset time of patients with AIS was correlated with the severity and range of CBF reduction and Tmax delay. Thus, the input layer in our network was designed as a three-channel layer, which included CBF, Tmax and ROI mask respectively. The CBF and Tmax channels of the input layer can provide the detail features about the severity of CBF reduction and Tmax delay, and the ROI mask channel can present a weight map to express the range of CBF reduction and Tmax delay. The input layer was separated into blocks with the size of 64 × 64 × 64. The convolutional layer contained 16 filters with a receptive field of 5 × 5 × 5 voxels in a one-voxel stride sliding. The batch normalization layer and ReLU layer which followed the convolutional layer, batch-normalized and rectified the feature map. The max pooling layer reduced the number of rectified features, and they were flattened into a single linear vector by the fully connected layer. Finally, the classification was processed in the soft-max layer. Binary cross-entropy loss was used as loss function. Comparing VGGNet with 2 convolutional blocks, the input layer in our network included three channels, and each channel was 3D images. Apart from that, we removed a max-pooling layer in the first convolutional block in order to decrease the loss of the detail features. All classifiers were trained by fivefold cross-validation to avoid overfitting bias.

## Statistical analysis

We computed the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC), which can compare the ability of all classifiers to identify patients with AIS within a 6-h window. To determine the significance of differences among classifiers in the task of identification, we used the DeLong test to compare the AUCs of the classifiers (21). We also computed patient-wise accuracy, sensitivity, specificity, and precision for each classifier. SPSS version 22.0 (IBM, USA) and GraphPad Prism version 6.0 (GraphPad, USA) powered all of the statistical computations, with significance set at $p < 0.05$.

TABLE 3 Patient characteristics.

| Characteristics | Values |
|---|---|
| No. of patients | 377 |
| Age (year) | 66.0 ± 11.9 |
| Male sex* | 263 (69.8) |
| Stroke onset time (h) | 6.7 ± 5.7 |
| NIHSS on admission | 11.5 ± 7.2 |

*Data are the number (percentage) of patients. Except where indicated, data are mean ± SD.

TABLE 4 The patient characteristics in the training and testing datasets.

| Patient characteristics | Training dataset | Testing dataset | *P*-value |
|---|---|---|---|
| Age (year) | 66.81 ± 11.63 | 68.06 ± 11.64 | 0.3737 |
| Sex (female/male) | 98/223 | 16/40 | 0.8454 |
| NIHSS on admission | 12.08 ± 7.13 | 11.00 ± 7.51 | 0.4152 |
| Stroke onset time (h) | 5.87 ± 5.46 | 5.98 ± 4.60 | 0.0552 |

# Results

## Patient characteristics

We recruited 2,500 patients from the eStroke platform; 426 were excluded due to loss of original data, and 922 were excluded because of poor image quality, such as motion artifacts during scanning. Additionally, 775 with an onset time exceeding 24 h were excluded. Finally, a total of 377 patients (263 men and 114 women; mean age = 66.0 ± 11.9 years) were included in this study. The stroke onset time was 6.7 ± 5.7 h (range = 0–24 h). All patients had ACA occlusion. The patients' baseline and NIHSS are listed in Table 3.

## Training and testing dataset analysis

Training and testing datasets were selected randomly, which were grouped by the onset time of stroke. Table 4 shows the patient characteristics in the training and testing datasets. All *p*-values for each patient characteristic between the training and testing datasets were estimated. We observed that all *p*-values were higher than 0.05, which means that there were no significant differences in each patient characteristic between the training and testing datasets.

## Performance analysis of the classifiers

Figure 2 shows the ROC curves of the classifiers (SVM, RF, and CNN) for identifying patients with AIS within a 6-h treatment window for EVT. All of the AUCs of the classifiers were higher than 0.76, which was the highest AUC for identifying patients with AIS within a 4.5 h window for rt-PA thrombolysis treatment in a previous study (8). The AUC of RF was the lowest at 0.775 (0.732–0.818), while the AUC of the CNN was the highest at 0.935 (0.893–0.975). The AUC of the CNN was significantly higher than that of the SVM ($p = 0.001$) and RF ($p = 0.001$). The AUCs of the classifiers compared with the previous work are depicted in Table 5.

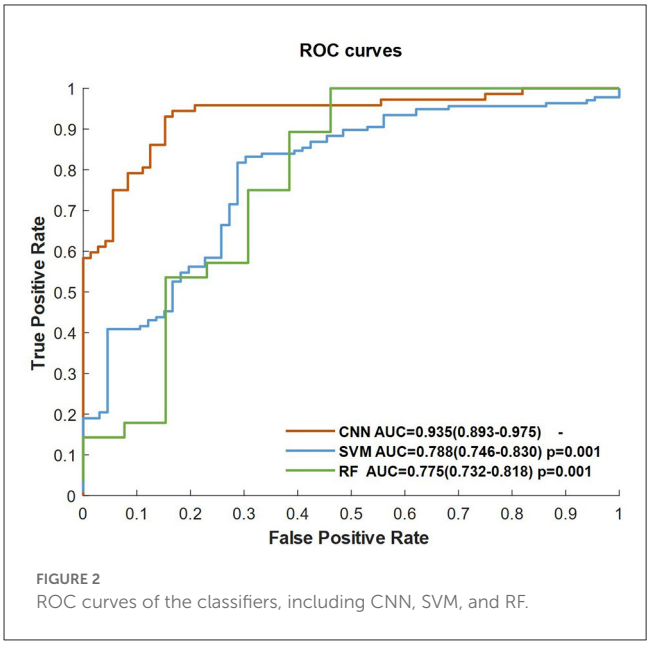


FIGURE 2
ROC curves of the classifiers, including CNN, SVM, and RF.

TABLE 5 The AUCs of classifiers of the identification of patients with AIS within a 6- and 4.5-h window.

| Classifier | Identifying patients within 4.5-h window | | Identifying patients within 6-h window |
|---|---|---|---|
| | **Ho et al. (7)** | **Kong et al. (8)** | **CBF + Tmax + ROI** |
| RF | 0.624 | 0.690 | 0.775 (0.732–0.818) |
| SVM | 0.669 | 0.746 | 0.788 (0.746–0.830) |
| CNN | – | – | **0.935 (0.893–0.975)** |

Bold indicated the highest AUC for a given classifier.

## Performance analysis of the ROI masks

The CNN classifier was trained by CBF, Tmax, and each ROI mask (respectively, CBF < 30%, Tmax > 4 s, Tmax > 6 s, Tmax > 8 s, and Tmax > 10 s), each ROC curve of which is shown in Figure 3. The AUC of Tmax > 6 s was the maximum value (AUC = 0.935), which was significantly higher than that of Tmax > 8 s and Tmax > 10 s ($p = 0.017$ and $p = 0.002$, respectively). Although the AUC of Tmax > 6 s was higher than that of Tmax > 4 s, there was no significant difference between them ($p = 0.285$). Comparing the ROI masks segmented by Tmax, the AUC of CBF < 30% was only 0.796 (0.723–0.867).

## Performance analysis of scanning devices

We separated the training dataset into two groups (CTP and PWI), and the CNN classifier was trained by CBF, Tmax, and ROI mask of Tmax > 6 s in each group. Figure 4 shows the ROC curves

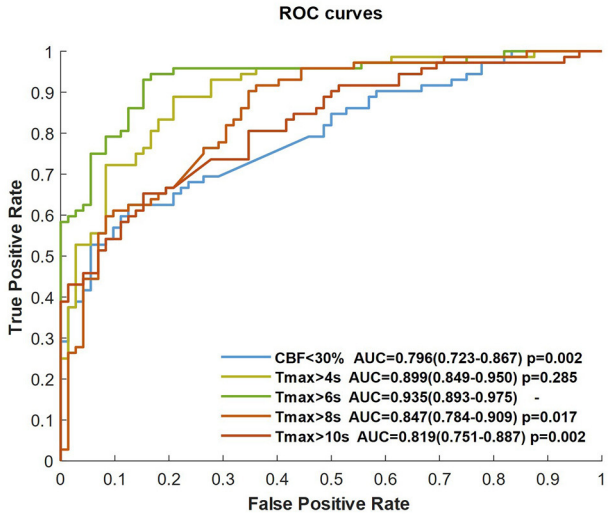


FIGURE 3
ROC curves of the CNN trained by CBF, Tmax, and each ROI mask.

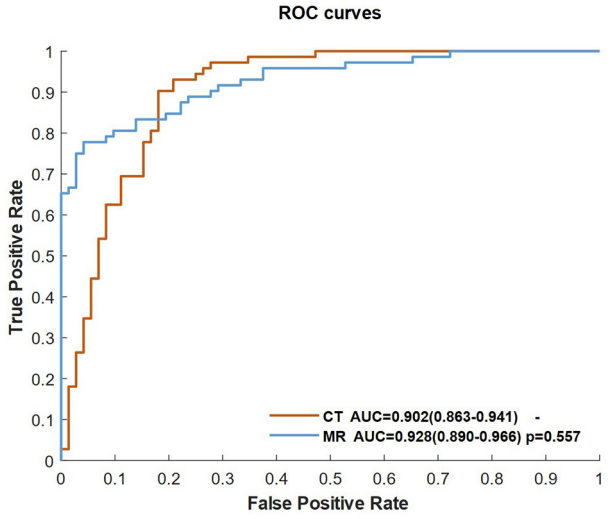


FIGURE 4
ROC curves of the classifiers of the CNN trained by CBF, Tmax, and ROI masks of Tmax > 6 s in the CT and MR groups.

of two groups. The AUCs of the two groups were higher than 0.9, and there was no significant difference between them ($p = 0.557$).

## Examples of identification

Figure 5 shows four examples for identifying patients with AIS within a 6-h window for the treatment of EVT using our method. The classifier was CNN-trained by CBF, Tmax, and ROI masks of Tmax > 6 s. The results of the classifier identification were matched with the ground truth, which was the accurate stroke onset time of patients. DWI and FLAIR are listed in Figure 5 for comparison with a previous study (8), which detected patients with AIS within a 4.5 h window for rt-PA thrombolysis treatment using the machine learning method and the imaging biomarker of DWI–FLAIR mismatch.

# Discussion

In this study, we proposed to use a CNN framework based on a perfusion map (CBF and Tmax) to identify patients with AIS within a 6-h window for the treatment of EVT. We compared the performance of each classifier (SVM, RF, and CNN) and differences from each ROI mask (CBF < 30%, Tmax > 4 s, Tmax > 6 s, Tmax > 8 s, and Tmax > 10 s). Our results showed that the CNN classifier trained by CBF, Tmax, and ROI masks of Tmax > 6 s had a higher performance in terms of identification within a 6-h window. Apart from this, our method had stable performance for both CTP and PWI, which means that the proposed method has higher potential to be used widely in stroke centers.

In a previous study, the progression of AIS could be directly expressed by changes in the infarct core and ischemic region (12–14). Thomalla et al. proposed that DWI–FLAIR mismatch can be deemed an imaging biomarker for identifying patients with AIS within a 4.5 h treatment window for rt-PA thrombolysis (4). Meanwhile, in the study of DIFFUSE 3, the infarct core and penumbra region could be estimated using CBF and Tmax (8). Because DWI, FLAIR, and Tmax are related to the progression of AIS, Kong at el. constructed a decoder–encoder network trained by DWI, FLAIR, and Tmax to identify patients with AIS within a 4.5 h window for rt-PA thrombolysis treatment (8). In fact, Kong's decoder–encoder network has the potential to detect this within a 6-h treatment window. However, because this network was trained only by MR examination, it was hard to be widely used in hospitals, especially primary hospitals. Thus, in order to be used for both CT and MR examination, we chose CBF and Tmax as two of the three channels of input images of classifiers instead of DWI and FLAIR, and we pulled ROI masks into the third channel of input images because their changes were correlated with the progression of AIS. This means that our method has more potential to be performed in primary hospitals.

In identifying patients with AIS within a 4.5 h window for rt-PA thrombolysis treatment, the AUC of the best classifier was 0.780 (8). The best classifier in this study was the CNN trained by CBF, Tmax, and ROI masks of Tmax > 6 s. The AUC of our method was 0.935, which is much higher than that of previous works. The reason is that the progression of AIS over time mainly influences cerebrovascular hemodynamic changes (9–11). For instance, in Figure 5, changes in CBF and Tmax had a significant relationship with the stroke onset time among patients A, B, C, and D. Although patient D was attacked by a stroke for 10.1 h, the intensity between DWI and FLAIR was not mismatched, which would have been misestimated in previous works. Apart from this, our results showed that the CNN has a stronger ability to capture hidden features and signal changes from CBF and Tmax, compared to machine learning methods such as SVM and RF. Moreover, by comparing the performance of classifiers trained by different ROI masks, our results showed that the AUC of Tmax > 6 s was the highest in all ROI masks, although it was not significantly higher than that of Tmax > 4 s ($p = 0.285$). According to a previous

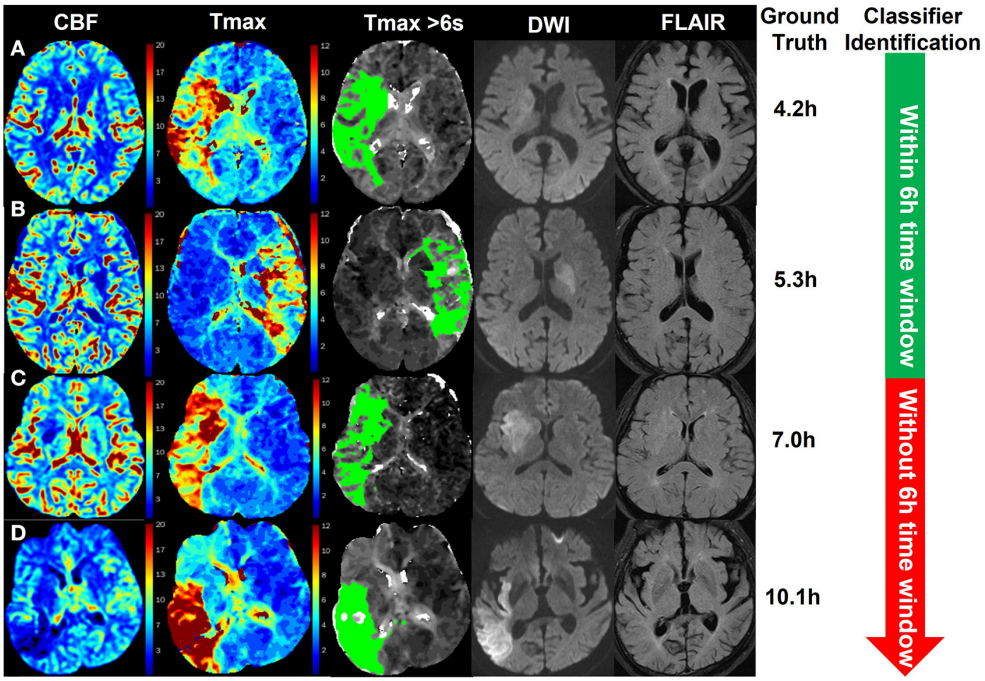


FIGURE 5
Examples of the identification of patients with AIS within a 6-h window for the treatment of EVT using the CNN classifier trained by CBF, Tmax, and ROI masks of Tmax > 6 s. The green region in the third column is the mask of Tmax > 6 s. The stroke onset times of patients **(A, B)** were, respectively, 4.2 and 5.3 h, which was identified within a 6-h time window for the treatment of EVT by the classifier. For patients **(C, D)**, their stroke onset time was 7.0 and 10.1 h, respectively, which was identified without a 6-h window by the classifier.

work (9), the region of Tmax > 6 s includes an infarct core and penumbra, while the regions of CBF < 30% and Tmax > 10 s only include an infarct core, and the region of Tmax > 8 s includes an infarct core and a part of penumbra. For the region of Tmax > 4 s, it includes benign hypoperfusion, an infarct core and penumbra, which should include more features than that of Tmax > 6 s, but the benign hypoperfusion in the region of Tmax > 4 s is always misestimated because of personalizing. For example, Tmax values in the deep area of white matter without lesions are commonly more than 4 s for patients with AIS. Therefore, we recommend the CNN classifier trained by CBF, Tmax, and ROI masks of Tmax > 6 s rather than Tmax > 4 s.

This study has some methodological limitations that need to be addressed. First, the sample size was relatively lower than that of other studies based on deep learning algorithms. However, data were collected from 13 centers, with eight types of CT and MR scanners, uniformly distributed between 0 and 24 h from the stroke onset time. Thus, the sample size was enough to support the training of the CNN model in this study. Second, data were collected retrospectively, and some inaccurate information was involved. In fact, a prospective study to evaluate the performance of our method in clinical use is a future avenue for investigation, but it does not enable to assume the clinical potential of this study. In the future, a larger, randomized, and prospective study will be designed to evaluate the performance of this method.

## Conclusion

In this study, a CNN classifier trained by CBF, Tmax, and ROI masks of Tmax > 6 s, has good performance to identify patients with AIS within a 6-h window for the treatment of EVT. Comparing with existing works to classify patients within a 4.5-h window for the treatment of rt-PA thrombolysis, to the best of our knowledge, this is the first work to assist the treatment of EVT. Meanwhile, our method performs the identifying task using CBF and Tmax, which can be acquired by CTP or PWI. It means that our method is compatible with both CT and MR devices, while previous works only support MR devices because their inputs rely on DWI and FLAIR images which are examined only by MR devices. Commonly, CT examination is faster than MR, which benefits to bring the patients out of danger. Therefore, it has the potential to be widely used to accurately estimate the stroke onset time of patients with WUS.

## 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

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the [patients/ participants OR patients/participants legal guardian/next of kin] was not required to participate in this study in accordance with the national legislation and the institutional requirements.

## Author contributions

HG and YB designed the study. HG, YC, and JW collected the data. HG, YB, and HY were involved in the interpretation of data. YB, HZ, and GC analyzed and visualized the data. HG and YB drafted the manuscript. QY and LW revised the manuscript. All authors read and approved the final manuscript.

## Funding

This work was partially supported by grants from the Capital's Funds for Health Improvement and Research (CHF), and Beijing Hospitals Authority's Ascent Plan, and Clinical Research Incubation Project, Beijing Chao-Yang Hospital, Capital Medical University (CYF202213), and National Natural Science Foundation of China (82171396 and 81820108014), and National Key Research and Development Project (2018YFE0114400).

## Conflict of interest

GC was employed by Neusoft Medical System Co., Haidian, Beijing, China.

The remaining 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. Powers WJ, Rabinstein AA, Ackerson T, Adeoye OM, Bambakidis NC, Becker K, et al. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *J Stroke.* (2019) 50:e344–418. doi: 10.1161/STR.0000000000000211

2. Elfil M, Eldokkmak M, Baratloo A, Ahmed N, Koo BB. Pathophysiologic mechanisms, neuroimaging and treatment in wake-up stroke. *CNS Spect.* (2019) 25:460–7. doi: 10.1017/S1092852919001354

3. Lee H, Lee E-J, Ham S, Lee H-B, Lee JS, Kwon SU, et al. Machine learning approach to identify stroke within 45 h. *J Stroke.* (2020) 51:860–6. doi: 10.1161/STROKEAHA.119.027611

4. Stoessl AJ, Martin WW, Mckeown MJ, Sossi V. Dwi-flair mismatch for the identification of patients with acute ischaemic stroke within 4–5 h of symptom onset (pre-flair): a multicentre observational study. *J Lancet Neurol.* (2011) 10:951–2. doi: 10.1016/S1474-4422(11)70192-2

5. Zhu H, Jiang L, Zhang H, Luo L, Chen Y, Chen Y. An automatic machine learning approach for ischemic stroke onset time identification based on dwi and flair imaging. *NeuroImage Clin.* (2021) 31:102744. doi: 10.1016/j.nicl.2021.102744

6. Zhang Y-Q, Liu A-F, Man F-Y, Zhang Y-Y, Li C, Liu Y-E, et al. Mri radiomic features-based machine learning approach to classify ischemic stroke onset time. *J Neurol.* (2021) 2021:1–11. doi: 10.1007/s00415-021-10638-y

7. Ho KC, Speier W, El-Saden S, Arnold CW, editors. Classifying acute ischemic stroke onset time using deep imaging features. In: *AMIA Annual Symposium Proceedings*. Bethesda: American Medical Informatics Association (2017).

8. Ho KC, Speier W, Zhang H, Scalzo F, El-Saden S, Arnold CW, et al. machine learning approach for classifying ischemic stroke onset time from imaging. *IEEE Trans Med Imag.* (2019) 38:1666–76. doi: 10.1109/TMI.2019.2901445

9. Albers GW, Lansberg MG, Kemp S, Tsai JP, Lavori P, Christensen S, et al. *A Multicenter Randomized Controlled Trial of Endovascular Therapy Following Imaging Evaluation for Ischemic Stroke (Defuse 3)*. London: SAGE Publications Sage (2017). doi: 10.1177/1747493017701147

10. Jovin TG, Saver JL, Ribo M, Perreira V, Furlan A, Bonafe A, et al. Diffusion-weighted imaging or computerized tomography perfusion assessment with clinical mismatch in the triage of wake up and late presenting strokes undergoing neurointervention with trevo (dawn) trial methods. *Int J Stroke Off J Int Stroke Soc.* (2017) 2017:1747493017710341. doi: 10.1177/1747493017710341

11. Campbell BC, Mitchell PJ, Kleinig TJ, Dewey HM, Churilov L, Yassi N, et al. Endovascular therapy for ischemic stroke with perfusion-imaging selection. *N Engl J Med.* (2015) 372:1009–18. doi: 10.1056/NEJMoa1414792

12. Nogueira RG, Jadhav AP, Haussen DC, Bonafe A, Budzik RF, Bhuva P, et al. Thrombectomy 6–24 h after stroke with a mismatch between deficit and infarct. *J N Engl J Med.* (2018) 378:11–21. doi: 10.1056/NEJMoa1706442

13. Wouters A, Dupont P, Norrving B, Laage R, Thomalla G, Albers GW, et al. Prediction of stroke onset is improved by relative fluid-attenuated inversion recovery and perfusion imaging compared to the visual diffusion-weighted imaging/fluid-attenuated inversion recovery mismatch. *Stroke.* (2016) 47:2559–64. doi: 10.1161/STROKEAHA.116.013903

14. Campbell B, Christensen S, Levi CR, Desmond PM, Donnan GA, Davis SM, et al. Comparison of computed tomography perfusion and magnetic resonance imaging perfusion-diffusion mismatch in ischemic stroke. *Stroke.* (2012) 43:2648–53. doi: 10.1161/STROKEAHA.112.660548

15. Olivot J-M, Mlynash M, Thijs VN, Kemp S, Lansberg MG, Wechsler L, et al. Optimal tmax threshold for predicting penumbral tissue in acute stroke. *J Stroke.* (2009) 40:469–75. doi: 10.1161/STROKEAHA.108.526954

16. Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansen-Berg H, et al. Advances in functional and structural Mr image analysis and implementation as Fsl. *Neuroimage.* (2004) 23:S208–S19. doi: 10.1016/j.neuroimage.2004.07.051

17. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* (1995) 20:721–8.

18. Breiman L. Random forests. *Mach Learn.* (2001) 45:5–32. doi: 10.1023/A:1010933404324

19. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol.* (1962) 160:106.

20. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *J arXiv preprint arXiv.* (2014).

21. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *J Biomet.* (1988) 1988:837–45.

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 09 March 2023
DOI 10.3389/fmed.2023.1114571



OPEN ACCESS

EDITED BY
Diwei Zhou,
Loughborough University, United Kingdom

REVIEWED BY
Wufeng Xue,
Shenzhen University, China
Xiaowei Han,
Nanjing Drum Tower Hospital, China

*CORRESPONDENCE
Suyu Dong
✉ dongsuyu@nefu.edu.cn
Yan Li
✉ wemn@sina.com
Guangyuan Yang
✉ ygy665@126.com
Zhaowen Qiu
✉ qiuzw@nefu.edu.cn

SPECIALTY SECTION
This article was submitted to Nuclear Medicine, a section of the journal Frontiers in Medicine

RECEIVED 02 December 2022
ACCEPTED 14 February 2023
PUBLISHED 09 March 2023

CITATION
Xiao X, Dong S, Yu Y, Li Y, Yang G and Qiu Z (2023) MAE-TransRNet: An improved transformer-ConvNet architecture with masked autoencoder for cardiac MRI registration. *Front. Med.* 10:1114571. doi: 10.3389/fmed.2023.1114571



# MAE-TransRNet: An improved transformer-ConvNet architecture with masked autoencoder for cardiac MRI registration

Xin Xiao[1], Suyu Dong[1]*, Yang Yu[2], Yan Li[1]*, Guangyuan Yang[3]* and Zhaowen Qiu[1]*

[1]College of Information and Computer Engineering, Northeast Forestry University, Harbin, China, [2]Department of Cardiovascular Surgery, Beijing Anzhen Hospital, Capital Medical University, Beijing, China, [3]First Affiliated Hospital, Jiamusi University, Jiamusi, China

The heart is a relatively complex non-rigid motion organ in the human body. Quantitative motion analysis of the heart takes on a critical significance to help doctors with accurate diagnosis and treatment. Moreover, cardiovascular magnetic resonance imaging (CMRI) can be used to perform a more detailed quantitative analysis evaluation for cardiac diagnosis. Deformable image registration (DIR) has become a vital task in biomedical image analysis since tissue structures have variability in medical images. Recently, the model based on masked autoencoder (MAE) has recently been shown to be effective in computer vision tasks. Vision Transformer has the context aggregation ability to restore the semantic information in the original image regions by using a low proportion of visible image patches to predict the masked image patches. A novel Transformer-ConvNet architecture is proposed in this study based on MAE for medical image registration. The core of the Transformer is designed as a masked autoencoder (MAE) and a lightweight decoder structure, and feature extraction before the downstream registration task is transformed into the self-supervised learning task. This study also rethinks the calculation method of the multi-head self-attention mechanism in the Transformer encoder. We improve the query-key-value-based dot product attention by introducing both depthwise separable convolution (DWSC) and squeeze and excitation (SE) modules into the self-attention module to reduce the amount of parameter computation to highlight image details and maintain high spatial resolution image features. In addition, concurrent spatial and channel squeeze and excitation (scSE) module is embedded into the CNN structure, which also proves to be effective for extracting robust feature representations. The proposed method, called MAE-TransRNet, has better generalization. The proposed model is evaluated on the cardiac short-axis public dataset (with images and labels) at the 2017 Automated Cardiac Diagnosis Challenge (ACDC). The relevant qualitative and quantitative results (e.g., dice performance and Hausdorff distance) suggest that the proposed model can achieve superior results over those achieved by the state-of-the-art methods, thus proving that MAE and improved self-attention are more effective and promising for medical image registration tasks. Codes and models are available at https://github.com/XinXiao101/MAE-TransRNet.

KEYWORDS

**deformable image registration, vision transformer, masked autoencoder, self-supervised learning, multi-head self-attention**

## 1. Introduction

Medical image registration has been considered a vital analytical task in medical image processing, especially for the registration of deformable non-rigid organs. It is capable of providing doctors with a wide variety of complementary information regarding lesions (1). The systole and diastole of the heart chambers play a vital role in maintaining the ejection function of the heart. Certain heart diseases can lead to changes in the shape of the ventricles, thus resulting in abnormal motion. For instance, hypertrophic cardiomyopathy may cause the localized thinning of the ventricular wall. Inadequate aortic valve closure can cause lesions (e.g., enlarged ventricular chambers). Thus, the study on cardiac registration takes on a critical significance in quantifying cardiac motion, which helps doctors predict the progression of patients diseases in future and conduct precise medical treatment. Moreover, cardiovascular magnetic resonance imaging (CMRI) presents accurate morphological information and a better soft tissue contrast ratio of the human heart (2), which contributes to the diagnosis of a wide variety of cardiac abnormalities. CMRI has become the gold standard in the analysis of cardiac motor function, viability, and abnormalities.

The registration of cardiac images is considered a complex task, which is primarily indicated by two aspects:

(1) Non-rigid and complex motion. The heart undergoes very complex motion and deformation in the cardiac cycle. In addition to the well-known overall deformation (e.g., expansion or contraction), the heart also undergoes overall rigid motion and local deformation, thus making it have a more complex non-rigid periodic motion than other soft tissues (3). Furthermore, due to this motion, the morphology of the slices of the heart varies significantly within continuous time frames of a cardiac cycle, thus making accurate tracking of cardiac motion a difficult task.

(2) Scarcity of anatomical landmarks. There are fewer precise anatomical landmarks required to characterize cardiac motion than to resize other soft tissue structures. Moreover, the labels are more difficult to obtain. Notably, the lack of reliable identifiable landmarks in the myocardial wall makes it difficult for registration (4).

The registration of cardiac images is significantly more complicated than that of other tissues and organs' images due to the aforementioned two major problems.

However, with the rise of deep learning technology over the past few years, traditional registration methods with low accuracy, complex and tedious iterative processes, and high time costs have been unable to reduce the difficulties of today's medical image registration. Thus, deep learning methods based on deep neural networks have become the key to solving the bottleneck of medical image registration performance (5–7). Different training methods are largely divided into three types, namely, supervised learning, unsupervised learning, and weakly supervised learning. In existing research, Rohe et al. (8) proposed SVF-Net, a fully convolutional network based on the U-Net structure. This network replaces all layers in the conventional U-Net (9) network with convolutional layers. In addition, the model combines global semantic information from the deep network and local positional information from the shallow network, and it predicts the SVF 3D velocity field using ROI from the segmentation to supervise 3D cardiac image registration. Unsupervised learning methods have been a research hotspot in the field of registration since there have been rare labels related to cardiac tissue motion analysis. Krebs et al. (10) proposed a low-dimensional multiscale probabilistic deformation network based on conditional variational autoencoder (CVAE). This network is capable of learning from unlabeled cardiac data, which can be used for the registration of deformable soft tissue structures (e.g., heart and brain). Balakrishnan et al. (11) optimized a simple U-Net network, named VoxelMorph, which can be trained in an unsupervised or supervised manner to achieve MRI registration results by defining a loss function consisting of a mean square error (MSE). The loss function comprises a similarity measure and a smoothing constraint on the deformation field. Some researchers, inspired by the above-unsupervised methods, also proposed a weakly supervised strategy to solve the problem of sparse anatomical signatures of tissues and organs. Hu et al. (12) proposed a method to infer the registration field parameters from the high-level information contained in a small number of existing anatomical labels. These researchers introduced existing annotations in the region around the target at the training stage to introduce additional information for optimizing the network parameters and increasing the registration accuracy. Deep learning based on medical image registration methods, especially using convolutional neural networks, have shown more significant improvements in registration performance over the past few years. Increasing methods have been proposed to solve the problems of slow computation and less information captured using existing 2D/3D registration methods (13). However, the current mainstream frameworks primarily use convolutional neural networks as the backbone, and the conventional convolutional operation is to extract features by sliding a window with a convolutional kernel size. Moreover, the perceptual field is limited to a fixed-size region, which is only effective in extracting local features and has some limitations in acquiring global information (14). The Transformer, originally applied in the field of NLP, has gradually become a novel alternative architecture for extracting global features in recent years since it is effective in capturing long-range global location information (15). Nevertheless, since the Transformer is insufficient to extract local detailed features, relevant research has emerged to fuse the advantages of Transformer in extracting global information and CNN in extracting local information to complement each other. Vision transformer (16) is capable of dividing the image data into patches and then interpreting these patches as sequences to take them as input. The above tokens are handed it over to the Transformer encoder for processing. Thus, Chen et al. (17) first proposed a hybrid model of Transformer and CNN (TransUnet), thus preserving the U-shaped structure of U-Net and introducing the Transformer encoder structure. The input image is first passed through a series of convolution operations to generate feature maps of different resolutions. In addition, the network serializes the feature maps output from the last layer as the patches. These patches are input to the Transformer layer for encoding. Subsequently, a feature sequence with self-attentive weights is obtained through Transformers encoding, and it is reshaped to the image size and then upsampled, which is

combined with different high-resolution CNN features derived from the encoding path in the upsampling process to achieve a more accurate medical image segmentation tasks. Chen et al. (18) proposed ViT-V-Net fusing the basic registration framework-VoxelMorph based on the V-Net (19) structure with the vision transformer-based encoder to fully use the spatial correspondence obtained from the 3D volume for more accurate registration. As a result, the network can be better in extracting registration field features and extracting global features. The network is capable of extracting global features, while preserving as many local features as possible between image contexts.

Although introducing the Transformer has been very effective in solving problems (e.g., the loss of deep local feature information), numerous Transformer baselines and hybrid models have been proposed to solve the above problems. In fact, the Transformer is transferred from the NLP to the CV field, and a relatively large gap exists between the above two fields in understanding images and texts. Compared with the high information density of linguistic text information, the image information is highly redundant, thus making it relatively difficult for the model to predict the information density. In addition, considerable information irrelevant to the task objective may be included in the scope of the model learning, so the model should spend a lot of parameter capacity in learning. Moreover, a significant gap exists in the design of Transformer-based structures for NLP tasks and CV tasks. Decoding linguistic information may be easier than images, and reconstructing pixels is more complex than reconstructing words, so the design of the Transformer's internal structure is significantly correlated with the learning effect of implicit semantic representation during image decoding. Due to the above analysis and the emergence of the problem, He et al. (20) transferred the method with masked operation from NLP to the CV field. They developed a relatively simple strategy, i.e., randomly masking a certain percentage of the image patches, so the model can learn more useful features and can predict the information of the missing pixels. This architecture is capable of effectively achieving good results in classification, segmentation, and detection tasks.

In the meantime, the self-attention mechanism plays a crucial role in Transformer encoders, and its variants have been used to varying degrees in text, image, speech, and video tasks (21–23). The self-attention mechanism can filter out the features which are useful for the target task, and improve the model computation efficiency to a certain extent. It can pay more attention to the feature correlation between the data, to solve the problems of network information redundancy, gradient dispersion, and the difficulty of handling variable-length sequences. The current multi-head self-attention mechanism used in the traditional vision transformer maps each sequence into three different feature spaces (Q, K, V), and then calculates the attention weights by scaled dot product, which selects parallel multiple features from the input features for fusion. The attention mechanism based on the scaled dot product can capture the global contextual information of the feature sequence. However, in terms of computational complexity, assuming that the sequence length is set to N of dimension D, the dot product computation is essentially a multiplication between a matrix of dimension $N \times D$ and a matrix of $D \times N$ with a time complexity of $O(n^2d)$. In natural language processing tasks and some speech recognition tasks, many related studies have simplified the computation of the self-attention mechanism. It is necessary to consider some strategies to make it better for vision tasks and to reduce the computational complexity of self-attention.

Inspired by their research, we propose a novel Transformer-ConvNet model (MAE-TransRNet) using the MAE's strategy for cardiac MRI registration.

This study aims to enhance the performance of cardiac MRI registration by combining the advantages of CNN and Transformer. In this study, the transformer structure, which is currently popular, is primarily adopted to fuse the basic structure of the existing unsupervised registration baseline-VoxelMorph. We also explore the effect of the improved self-attention mechanism on the effect of feature aggregation. In addition, the attention mechanism and the superiority of the currently proposed Transformer structure with a MAE in increasing the registration accuracy of 3D medical images are investigated. The main contributions of this study are summarized into the following aspects:

(1) We propose a new hybrid multi-head self-attention module (HyMHSA) for vision tasks. The original query-key-value-based dot product computation unit is replaced with a dense synthesis unit that directly computes the attention weights. Meanwhile, the attention module restricts the interactions between sequences by exploiting the correlation between adjacent contexts of sequences, which makes the attention weights interact only between a portion of adjacent tokens and fuses them with the dot product form of the computation unit to reduce the computational burden.

(2) We introduce the concurrent spatial and channel squeeze and excitation (scSE) module (24) in the CNN's downsampling structure. In the Transformer encoder, squeeze and excitation module (25) is introduced after the attention to the Transformer structure, so as to reduce the feature redundancy in the self-attention mechanism in the ViT model, while increasing the richness of the cardiac image features.

(3) The structure of the conventional ViT model is improved based on Masked Autoencoder (MAE) (20). The application of the Transformer combined with VoxelMorph is deeply considered in medical image registration based on the existing research, and the Transformer is employed as a baseline to make corresponding model improvements. The proposed model is named MAE-TransRNet.

## 2. Related work

### 2.1. Deformable image registration baseline–VoxelMorph

Convolutional neural networks have progressively replaced the conventional registration methods based on mutual information with the development of deep learning in recent years. VoxelMorph (11) was proposed in 2019 and has been extensively used as a baseline in medical image registration. The VoxelMorph framework can learn registration field parameters from 3D volumetric data, and the encoder-decoder structure based on U-Net (9) structure is adopted to combine shallow features and deep features and reduce the information loss of features. Moreover, VoxelMorph provides two training strategies. One training strategy

is based on the grayscale value of the image to make the similarity to maximize the similarity loss and smoothing loss. This part is primarily pure unsupervised learning method for iterative optimization. The other training strategy introduces additional segmentation labels of the image as the auxiliary information based on the unsupervised method by obtaining the dice performance between the image pairs of segmentation labels at the training stage, thus increasing the registration effect. In this study, the superiority of the VoxelMorph baseline framework in the medical image registration is considered, and the skip connection structure of the VoxelMorph model architecture is redesigned and transferred into a long-range skip connection structure containing CNN encoder and decoder. This design is capable of combining the local information of feature maps at different scales more effectively and increasing the feature extraction capability.

## 2.2. Multi-head self-attention in transformer encoder

The multi-head self-attention module selects multiple pieces of information from the inputs and learns feature representations from different representation subspaces at different locations. The operation of multi-head attention can be described as mapping a query and a set of key-value pairs to the output, where the query, key, and value are denoted by Q, K, and V. Then, the three-part linear mapping is input to the attention mechanism based on scaled dot product to perform $h$ attentions in parallel computation ($h$ refers to multiple heads). The formula for each dot product attention computation is as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{D_k}})V \quad (1)$$

$\frac{1}{\sqrt{D_k}}$ is the attention deflator that mitigates the gradient disappearance. Then the results of h-heads scaled dot product attention are concatenated to obtain the final multi-head attention output feature vector:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \quad (2)$$

$$where \quad head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

$W^Q$, $W^K$, and $W^V$ denote the weight parameter matrixes corresponding to Q, K, and V. Figure 1 illustrates the structure of the traditional multi-head self-attention mechanism.

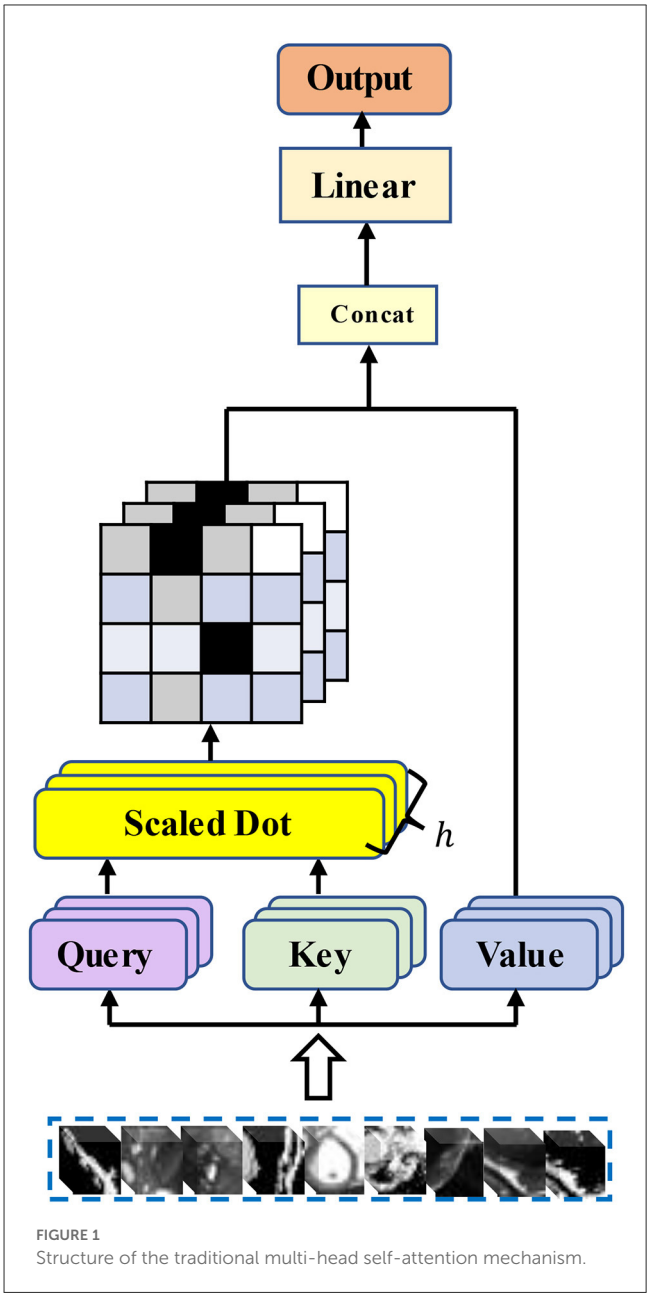


FIGURE 1
Structure of the traditional multi-head self-attention mechanism.

Self-attention models have been widely used in various fields. For query, key, and value, sequences of three vectors generated by tokens through linear layers, a considerable number of researchers have considered how to reduce the computation of attention. Quadratic, for a matrix with $N \times N$, we may not need the value of each position on the matrix to participate in the attention computation. Furthermore, for the global context information of the Transformer, we do not need to consider all the information from the beginning to the end. A sequence obtained from the cut image patches has a very long length. That is, when calculating the value of the current position, only its neighboring positions are considered. The range of neighboring positions to consider, the choice of position, and the choice of the sequence length to be calculated are all important factors that currently affect the complexity of attention computation in the vision domain. Chiu and Raffel (26) introduce a scalable and variable sliding window for attention computation, and Tay et al. (27) abandon the query-key-based interactive attention weight learning approach and propose a dense synthesizer that uses two feed-forward linear layers to predict the attention weight parameters. Xu et al. (28) further proposed a local dense synthesizer. They restrict the attention computation to a local area around the current central frame. However, improved works based on the self-attention mechanism are rarely found in medical image registration tasks. Our work is a new attempt. We introduce an improved self-attention mechanism into our Transformer encoder and explore whether the attention module

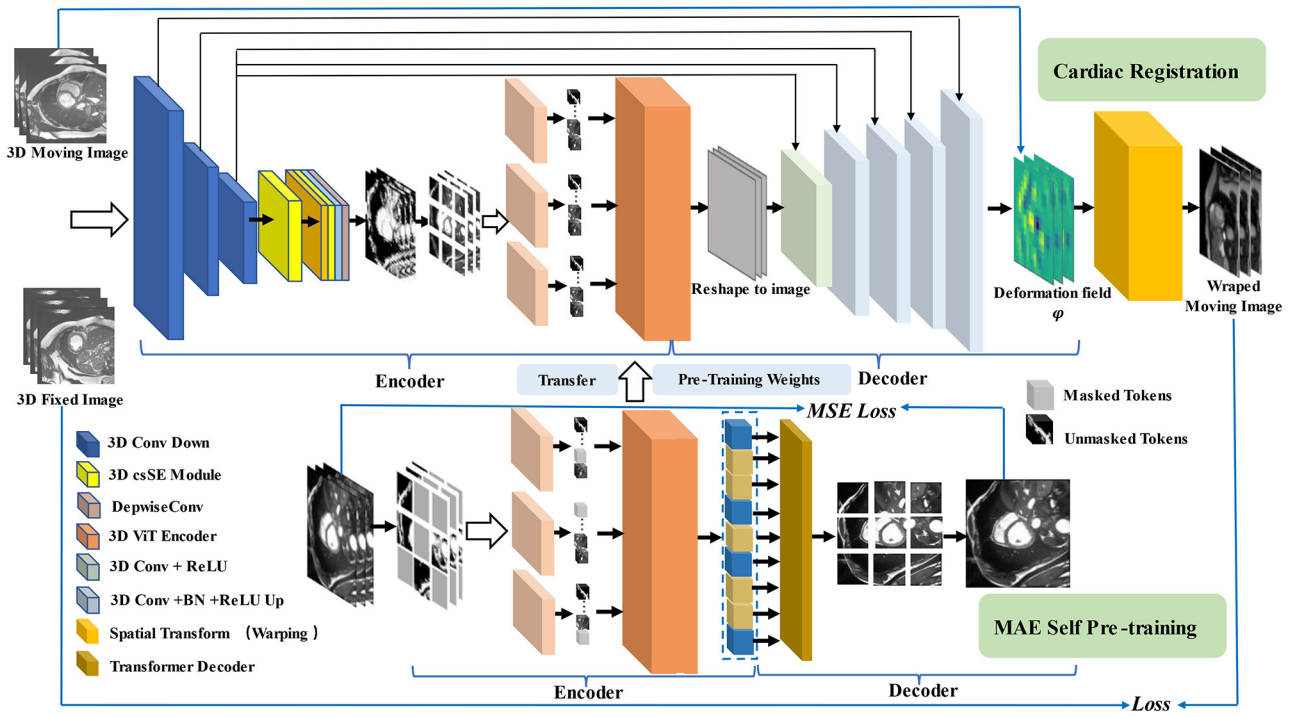


FIGURE 2
Pipeline for our registration method with MAE Pre-training. In the first stage (shown at the **bottom**), we use multi-branch MAE for pre-training, The input of the Transformer encoder is a subset of unobscured tokens from multiple branches, with a lightweight decoder to reconstruct the complete image. In the second stage (shown at the **top**), the weights obtained in the pre-training stage are used for the initialization of the registration model, the network takes a two-channel 3D volume tensor as our input, which is down-sampled by three convolutions, and then recalibrated with attention weights by scSE module, CNN decoder recovers potential features to image size and up-sample back to the original resolution size, using the spatial transformation function to warp moving image.

applicable to text, as well as speech tasks, can be well applied to our registration tasks.

## 2.3. Squeeze and excitation block in feature extraction

The convolutional operation is the core of conventional convolutional neural networks, which are based on local perceptual fields to fuse features in spatial and channel dimensions. The squeeze and excitation block proposed by Hu et al. (25) in 2018 places a focus on the research relating to the channel dimension and explores the feature relationships between channels, which can adaptively adjust the features on the channel dimension. The squeeze and excitation block can be stacked in many classical classification network structures (e.g., AlexNet and ResNet), and it has high performance on datasets (e.g., ImageNet). Inspired by the squeeze and excitation module, Guha Roy et al. (24) explored a fusion module combining channel dimension features and spatial dimension features to "reconstruct" features both in space and channel. Thus, the network can focus more on learning features that are more significance in downstream tasks, and it exhibits high applicability in 2D and 3D scenes. For the common tasks in current medical image analysis, (e.g., segmentation and registration), more insights should be gained into the spatial information at the pixel level of the image. Now, the embedding structure of such modules has been extensively used in the field of medical images (e.g., brain MRI and enhanced CT's segmentation tasks). Based on the above research, squeeze and excitation module and concurrent spatial and channel squeeze and excitation (scSE) module are embedded into the proposed model, and the role of the above two modules in improving the performance of the registration model is explored. The importance of different levels of features is adjusted, so the model can learn more valuable high-level features, and features that are less important for the target task are given less attention. Thus, richer spatial and channel information can be obtained at the pixel level. The relative importance of attention in both dimensions is calibrated simultaneously, which leads to further accuracy improvements in downstream registration tasks.

## 2.4. Transformers in vision and self-supervised learning

With the prevalence of Transformer architectures migrated from the NLP domain, increasing variants of Transformer-ConvNet have high performance in computer vision tasks. Transformer structures are now extensively employed in vital tasks (e.g., medical image segmentation, medical image registration, and reconstruction) because of their superiority in capturing global

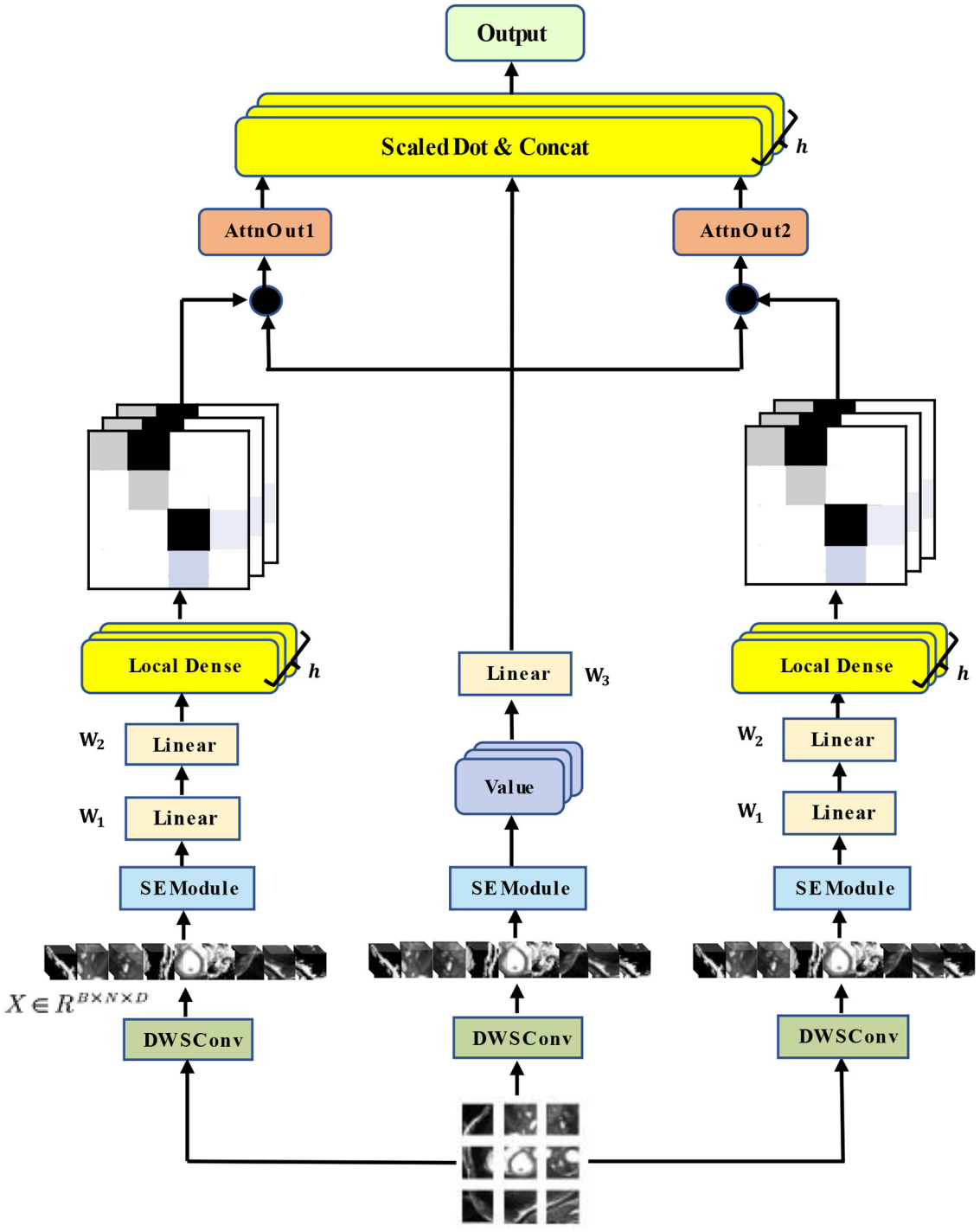


FIGURE 3
Architecture of our HyMHSA module. It is a hybrid version of the traditional multi-head self-attention mechanism and local dense self-attention mechanism.

contextual information and the localization of CNN convolutional operations for fine feature extraction. The TransUnet proposed by Chen et al. (17) is the first attempt at the Transformer-ConvNet structure, and it has achieved effective results in the segmentation tasks of cardiac and abdominal multiple organs. Several important works have also emerged in registration tasks, suggesting that the splicing and Transformer-ConvNet structures can effectively consider the advantages of both in their respective fields. However, with the emergence of some relevant in-depth studies, several problems are caused as follows:

(1) Numerous studies have suggested that the critical factor for learning efficiency is the scale of dataset, besides some problems of

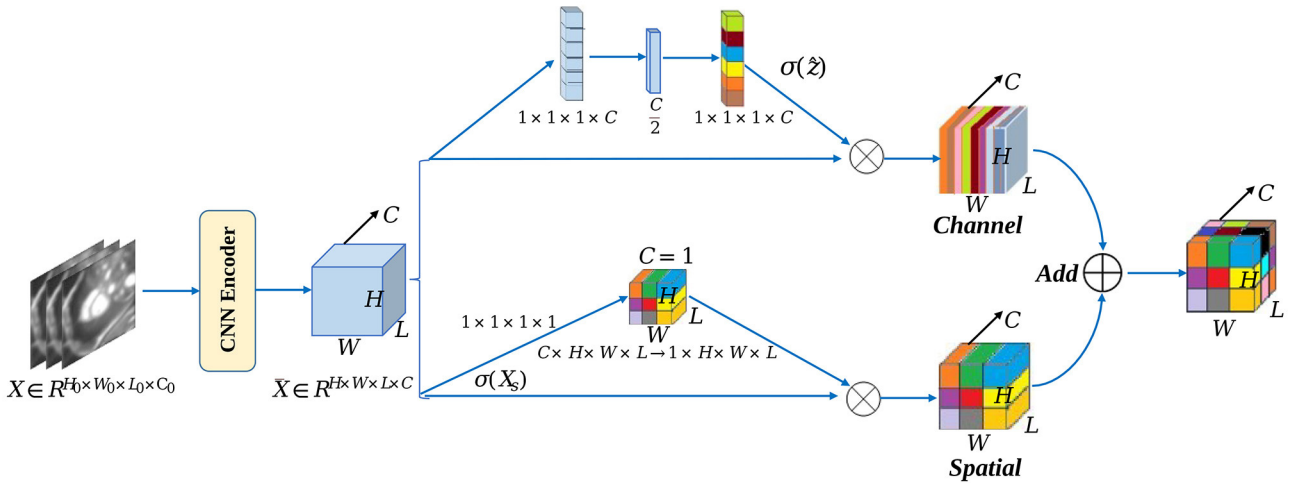


FIGURE 4
scSE module in CNN encoder.

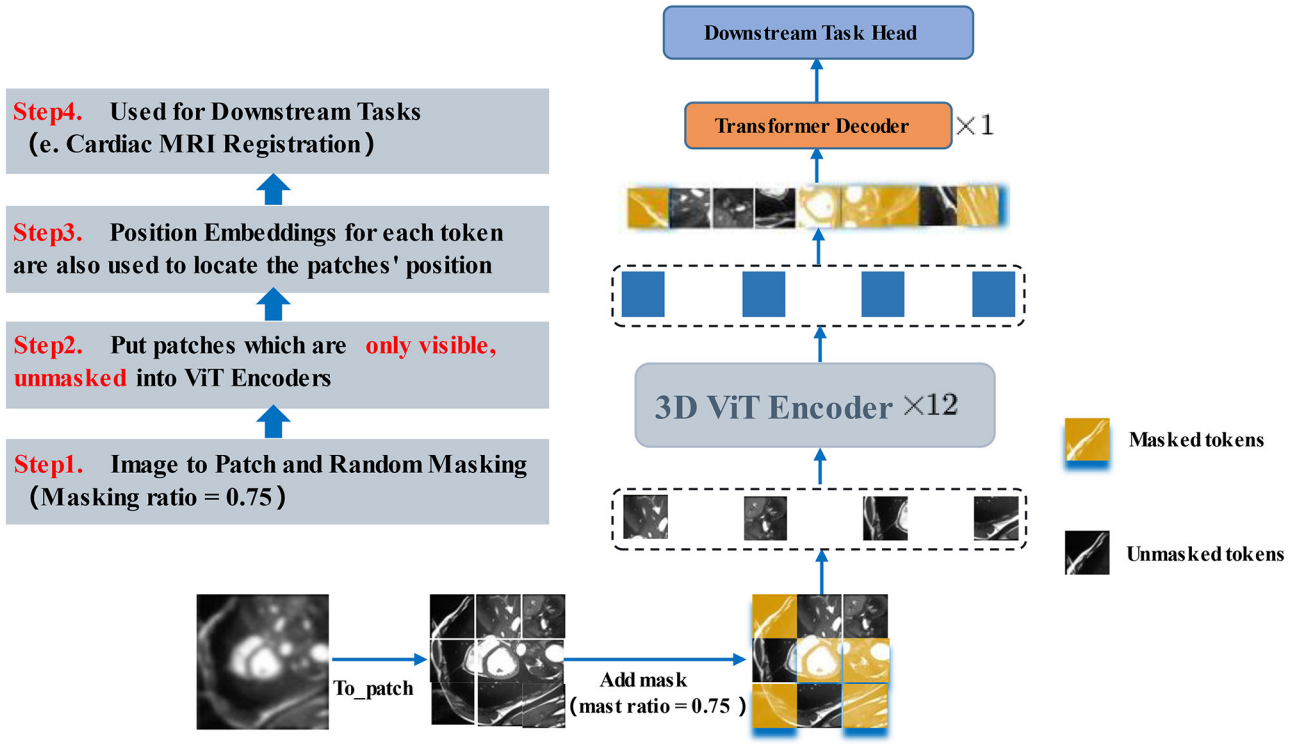


FIGURE 5
3D MAE Transformer. After the masked operation, only the unmasked patches are fed into ViT. After the linear mapping into one-dimensional tokens, the blank positions (yellow parts) are all filled by the same vector of the same dimension and then decoded by a layer of Transformer decoder. The next layer will be fine-tuned in accordance with the downstream task.

the model. The ability to learn valid information from considerable unlabeled data has been a crucial research topic in medical image analysis tasks. The number of data required to train the vision Transformer is significantly higher than that of a conventional convolutional neural network, especially the standard dataset with annotations. However, for medical images with a small sample size, it is undoubtedly challenging to obtain many labeled datasets, and the problem of data starvation always exists in the research on the vision Transformer architecture.

(2) The Transformer structures adopted to fuse CNNs are primarily migrated versions of structures based on NLP tasks, and the information density contained in the text is significantly different from the images. The features extracted by the Transformer encoder may be too complete and

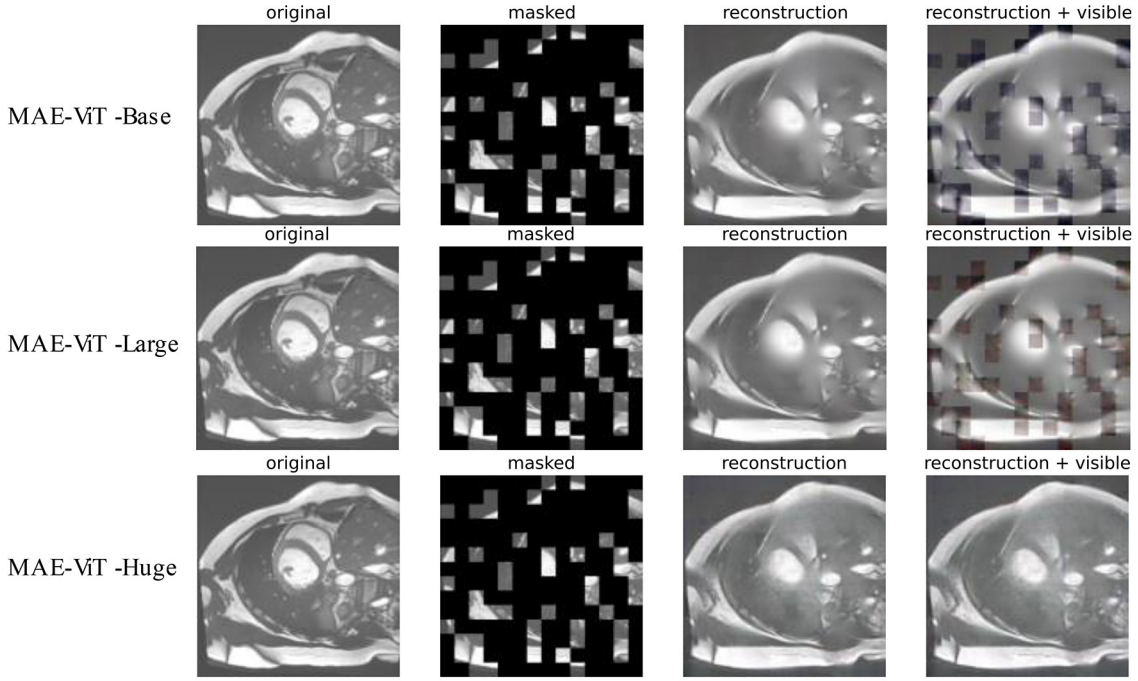


FIGURE 6
Reconstruction effect of cardiac images on MAE's three variants. The pre-processed original images, masked images, reconstructed images, and MAE reconstruction pasted with visible patches are presented from left to right.

**TABLE 1** Comparison of image registration performance (including dice performance and Hausdorff distance) of five different methods on the ACDC dataset.

| Methods | LV | | Myo | | RV | | Avg | |
|---|---|---|---|---|---|---|---|---|
| | DSC | HD | DSC | HD | DSC | HD | DSC | HD |
| VoxelMorph | 0.847 | 5.75 | 0.743 | 6.23 | 0.754 | 9.32 | 0.781 | 7.10 |
| CoTr-Based | 0.847 | 5.59 | 0.776 | 6.12 | 0.768 | 9.25 | 0.797 | 6.99 |
| PVT-Based | 0.848 | 5.37 | 0.745 | 6.08 | 0.778 | 9.12 | 0.79 | 6.86 |
| ViT-V-Net | 0.856 | 5.51 | 0.789 | 5.96 | 0.783 | 8.78 | 0.809 | 6.75 |
| **The proposed method** | **0.858** | **5.49** | **0.792** | **5.93** | **0.785** | **8.65** | **0.812** | **6.69** |

The best results are achieved and highlighted by the bold values.

contain some redundant information, so it is imperative to remove redundancy.

In medical images, the anatomical structure of the respective organ has a certain correlation between different contextual slice information, and it is also correlated with the features of the neighboring regions around the target region. The learning of the neighboring information and contextual information between pixels can facilitate the representation of advanced features. With the continuous development of self-supervised learning, the Transformer structure combined with the self-attentive mechanism (29) can break through the state of the art continuously. Self-supervised learning essentially provides a reliable learning path that allows the network to learn from large amounts of unlabeled data to be more capable of feature extraction. In fact, self-supervised learning is divided into several processes. (1) First, the basic structure or characteristics of the large amount of unlabeled data (which can be interpreted as built-in prior knowledge) are employed. Together with the relevant requirements of the task definition, some certain properties of the data are adopted instead of manual labeling, which can be interpreted as generating pseudo-labels for the images and initially training the network. Thus, it can extract features, i.e., the initial learning ability. (2) Second, the network is fine-tuned with a small amount of labeled data, so the network can further satisfy other tasks such as classification, segmentation, and registration.

The Transformer refers to an encoder -decoder integration based purely on an attention mechanism. In the current vision tasks, more novel strategies are urgently required to help models learn image features with powerful representations due to the different natures of visual information and textual information. Moreover, the MAE recently proposed by Kaiming He et al., has been well-adapted to the vision transformer and has achieved better results in tasks (e.g., classification). We consider that masked autoencoder can be effective in computer vision tasks by destroying

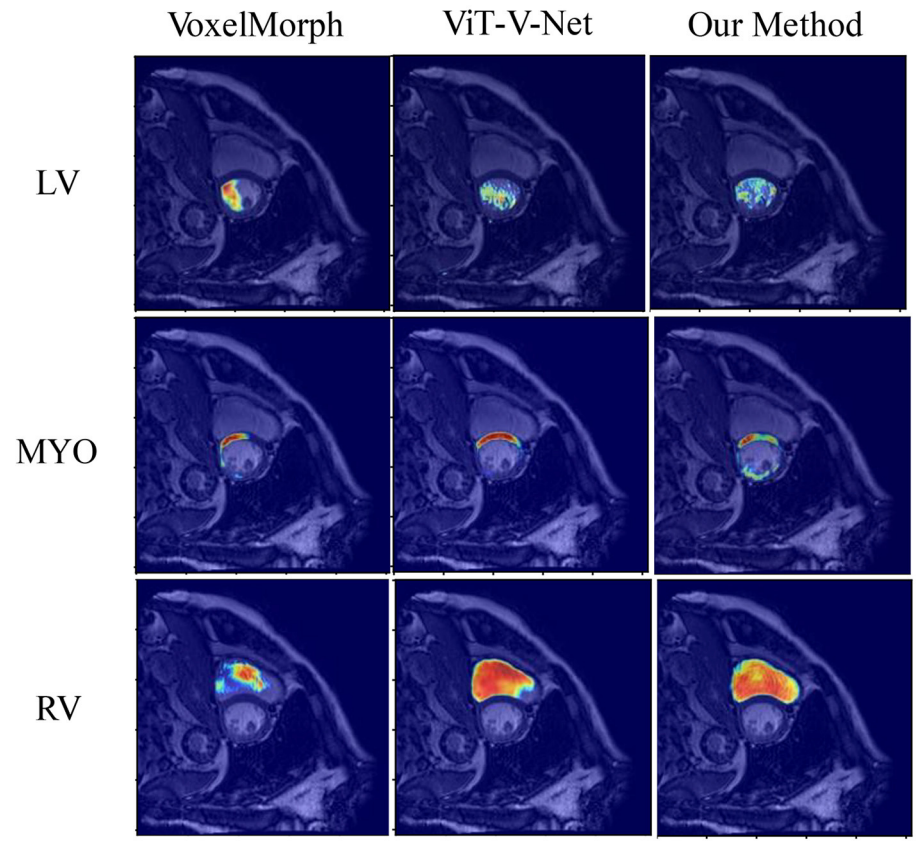


FIGURE 7
Visualization results of the attention heat map of the ACDC dataset in several models.

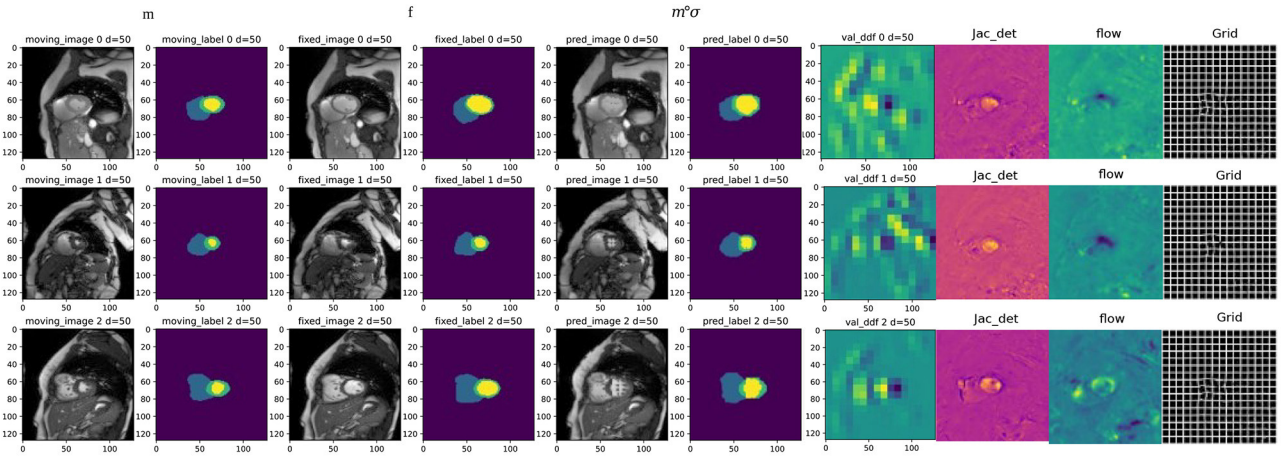


FIGURE 8
Examples of registration results from the proposed method, columns 1 and 3 are the moving image and fixed image from three different periods; columns 2 and 4 are the triple classification labels for the left ventricle, left ventricular myocardium, and right ventricle; columns 5 and 6 represent the warped original image and the warped image with labels, respectively; column 7 is the dense deformable field generated from fixed image and moving image; column 8 is the visualization result of the Jacobian determinant, as the dense displacement vector field (DVF); columns 9 and 10 are the registration flow filed and displacement field generated from the deformed images.

most of the patches of the image data and forcing the model to adapt to this defective feature structure when learning the image representation, which significantly reduces the redundancy of the image and creates a more challenging assignment. Finally, the model is enabled to learn the essential features of the image, so a powerful representation of the whole image data is obtained. The

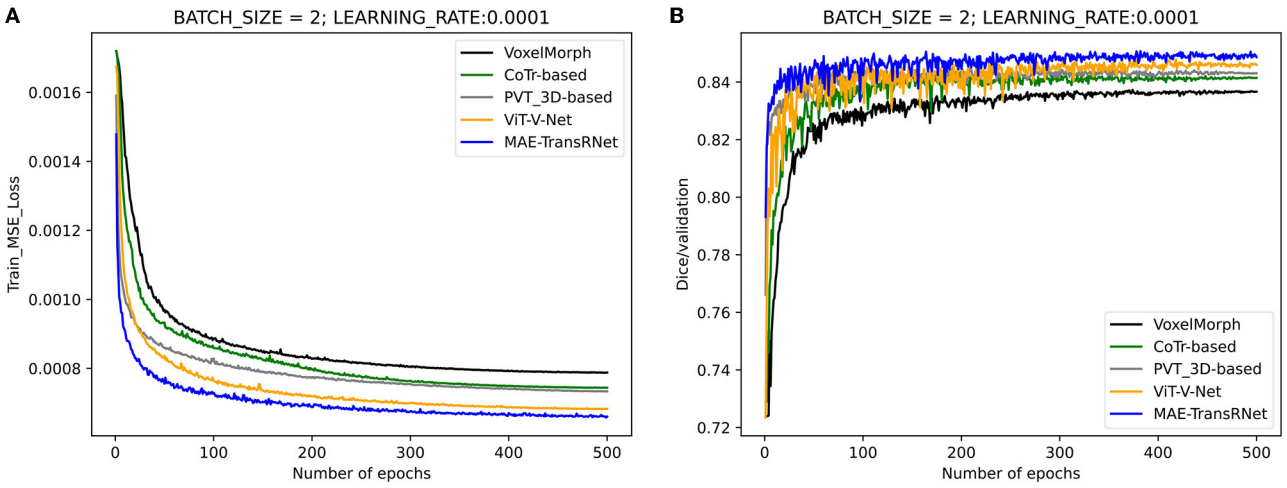


FIGURE 9
Performance comparison of the training and validation stage of different methods. Compared with other methods, the loss values of the proposed method are kept at a lower level during the training process, and the dice performance values obtained from the validation set are higher. **(A)** Training loss values under the comparison of different registration methods. **(B)** Validation dice performances under the comparison of different registration methods.

design of an asymmetric encoder -decoder structure saves model overhead, in which the encoder accounts for learning high-level feature representations by learning only the visible, unobscured patches, and the obscured patches are represented by a set of shared, learnable latent vectors. Self-supervised learning is further introduced into the visual transformer based on existing research, and the self-encoder with mask operations is applied to the heart image registration task, which can effectively solve the problem of sparsely labeled data and large information density differences between images and texts. Furthermore, applying the expandable MAE to our task and increasing the feature learning difficulty can instead lead to a stronger learning capability of the model.

# 3. Proposed method

## 3.1. Overview

Our MAE-TransRNet is a two-stage registration pipeline. In the first stage (bottom half of the figure), we use masked autoencoder as the encoder for pre-training. The encoder input is a subset of random masking of the image after patch chunking, and a modified self-attention mechanism is used in the Transformer encoder for simplifying the attention weight calculation. It calculates the attention parameter by selecting local contextual location information in the sequence. We reconstruct the complete image with a lightweight Transformer decoder, and the pre-trained model weights contain the powerful global latent features learned by the MAE pre-trained model on the cardiac image. In the second stage (top half of the figure), the pre-trained weights generated in the first stage are passed to the encoder of the registration model, and the registration network is initialized. Our input is 3D cardiac MRI ($\Omega \in R^3$), which consists of a single-channel grayscale image of the initial time frame $F \in R^{H\times W\times L}$ (fixed image) and a single-channel grayscale image of the end time frame $M \in R^{H\times W\times L}$ (moving image). The proposed model aims to learn the mapping transformations between the image pairs of the initial frame and the final frame. The resolution of the original image is first reduced to a suitable size through the downsampling operation of three convolutional neural networks to obtain a high-level feature representation, and then the spatial features are combined with the channel features by a concurrent spatial and channel squeeze and excitation (scSE) module, and the obtained high-level attention features are fed into the Transformer coding layer with the same structure as the pre-training stage. Going through the CNN decoder, the high-level features are reshaped to the image format. The deformation field is CNN's output, which is applied to the moving image through the warping layer. Here, the model uses the weights learned in the pre-training stage to train the registration network by calculating loss for backpropagation to generate a registration network model with optimal weight parameters. Subsequently, the objective function is minimized, as expressed in Equation (4)

$$\widehat{\varphi} = \arg\min_{\varphi} \mathcal{L}(F, M, \varphi) \tag{4}$$

$\varphi$ is obtained as the vector field offset from F to M as a feature of the registration image pair, i.e., $\varphi = Id + u$. $Id$ represents the constant transformation operation, and $u$ represents the displacement vector field. Figure 2 illustrates the overall pipeline of the proposed method.

## 3.2. Novel multi-head self-attention with SE module

In the Transformer encoder, the core of multi-head self-attention is to map query, key, and value in their respective representation subspaces and merge them back after processing

TABLE 2 Comparison of proposed methods with different attention mechanisms including time complexity and registration performance.

| Method | Complexity | DSC Avg | HD Avg |
|---|---|---|---|
| MHSA | $\mathcal{O}\left(N^2D\right)$ | 0.807 | 6.81 |
| LDSA | $\mathcal{O}\left(Ncn\right)$ | 0.801 | 6.71 |
| HyMHSA | $\mathcal{O}\left(N(N+cn)\right)$ | 0.812 | 6.69 |

$N$ is the length of the input feature and $cn$ is the context neighbor's value.

in their respective spaces, which is essentially the decomposition process and re-aggregation of attention features. For the visual domain, each location of each feature mapping contains information about the features at other locations in the same image, which makes the model more adept at capturing the dependencies between features with long spatial intervals. However, in practice, the input of image data is generally high resolution, especially 3D images, which makes the model need to learn longer feature sequences without losing too many fine-grained features of the image, and neither direct processing of the whole image nor downsampling can solve such problems significantly. The presence of inductive bias in CNN structures allows such models to be good at extracting local information. The Transformer structure remains desperate for extensive sample-size medical training data. In the face of such scarce data, we can only achieve this by exploring more powerful feature extractors, introducing some of the properties of inductive bias inherent in CNNs into Transformer, and in particular, embedding efficient convolution modules in the structure of self-attention computation to enhance the attention to small-scale local information in the dataset, which is one of the aims of our study.

We introduce the SE module into the computation of attention. Meanwhile, we embed the depthwise separable convolution (30) into our attention and the feed-forward layer. Given a 3D image as the input $X \in R^{B\times C\times H\times W\times L}$, the input is mapped into three subspaces representing Q, K, and V by a module consisting of deep convolution and point convolution, respectively. The depthwise convolution aggregates the spatial information, and the pointwise convolution aggregates the feature information along the channel dimension. Then we flatten the image features into a long sequence for Transformer encoding by patch embedding and position encoding. SE module is introduced in the respective transformer block to solve the problem that many channels in many current ViT models contain excessive redundant information, as well as to increase the efficiency of the model. After SE modules, the long token ($X \in \mathbb{R}^{B\times N\times D}$) is compressed into a $1\times1\times1\times D$ token, which is equivalent to compressing all global attention features into a high-level feature representation. Moreover, a series of nonlinear mappings are performed for the respective channel of the high-level features. Finally, the weight parameter corresponding to each channel is obtained, and a weight value representing the degree of attention is generated for the respective feature channel. This part aims to learn the nonlinear interaction between each token channel, and the weights are weighted with the original token to obtain the reconstructed attention to the feature representation with shape $B \times N \times D$. After the above operation, our input changes from $X \in \mathbb{R}^{B\times C\times H\times W\times L}$ to $X_Q/X_K/X_V \in \mathbb{R}^{B\times N\times D}$, formulating as:

$$\begin{aligned} X_1 &= PoiW(DepW(X, K_1), K_0) \\ X_2 &= PoiW(DepW(X, K_2), K_0) \\ X_3 &= PoiW(DepW(X, K_3), K_0) \end{aligned} \tag{5}$$

$$\begin{aligned} X_Q &= SE(Patch\_PosEmd(X_1, H, W, L, C, P), r) \\ X_K &= SE(Patch\_PosEmd(X_2, H, W, L, C, P), r) \\ X_V &= SE(Patch\_PosEmd(X_3, H, W, L, C, P), r) \end{aligned} \tag{6}$$

where $DepW$ and $PoiW$ denote depthwise convolution and pointwise convolution, $K_0$, $K_1$, $K_2$, and $K_3$ are different kernel sizes, $r$ is the reduction ratio of SE module, $X_Q$, $X_K$, and $X_V$ denote the vector representations mapped from the original input to three different subspaces, and $P$ denotes the patch size.

The design of the attention module affects the computational efficiency of the vision transformer. Currently, self-attention in vision transformer establishes global long-distance dependencies by interacting information between all regions in the image, which requires neighborhood and global context to achieve. Our approach does not entirely discard the decomposition and aggregation model of multi-headed self-attention while further setting the model's scope to consider neighborhoods. Our hybrid attention reduces the computational effort by restricting the current frame from interacting with its finite neighboring frames. We take one attention head as an example to explain our approach. First, we generate three weight matrixes $W_1$, $W_2$, and $W_3$ for computing attention using the linear layer of SELU. $W_1$ and $W_2$ are used to directly generate the attention weights corresponding to $X_Q$ and $X_K$, and $W_3$ is used to generate the attention weights for "values." In $W_2$, we introduce a hyperparameter $cn$, which represents the contextual neighbors. This parameter restricts the contexts around the current location considered in the attention calculation. Thus, the dimension of the original weight calculation is reduced from $N$ to $cn$. The model shares attention weights among only a limited number of locally adjacent contexts, significantly reducing time complexity. The input token ($X_Q/X_K/X_V \in \mathbb{R}^{B\times N\times D}$) is computed by attention weighting to obtain the query token ($S_{XQ} \in \mathbb{R}^{B\times N\times cn}$) with local contextual information and key token ($S_{XK} \in \mathbb{R}^{B\times N\times cn}$), value token ($S_{XV} \in \mathbb{R}^{B\times N\times D}$) are obtained directly by $W_3$ weighting. Furthermore, we introduce the variable $j$ to compute the weights of each $cn$ position above and below the $j$-centered position in the token with local attention, weight it with the value token, and sum it to obtain two vector outputs $AttnQ\_V$, $AttnQ\_K$ containing local adjacency context information. Since query and key are obtained by locally computing the full dense attention synthesized directly, we call this part the local dense attention computation module. The output of local dense attention is calculated by:

$$\begin{aligned} S_{XQ} &= \text{Soft max}\left(SELU(X_Q W_1) W_2\right) \\ S_{XK} &= \text{Soft max}\left(SELU(X_K W_1) W_2\right) \\ S_{XV} &= X_V W_3 \end{aligned} \tag{7}$$

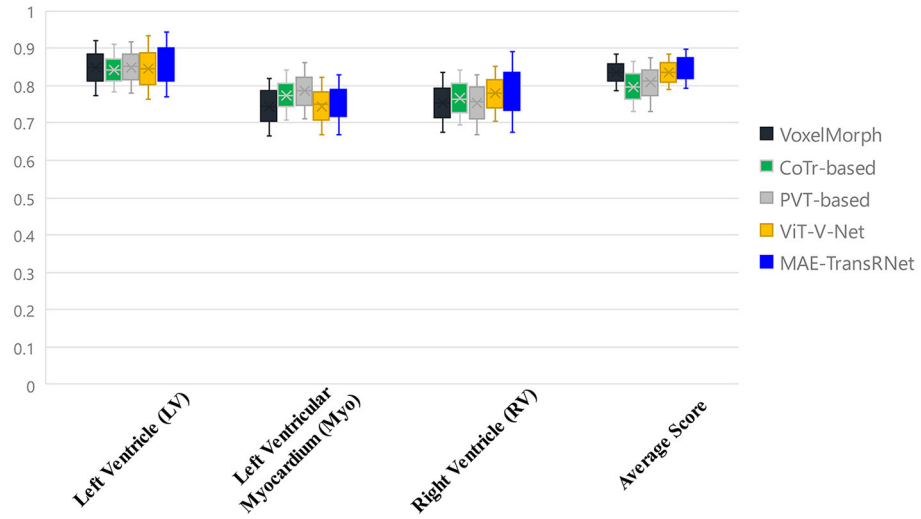


FIGURE 10
Boxplots to describe the variability of dice performance obtained by different registration methods for the same anatomical structure.

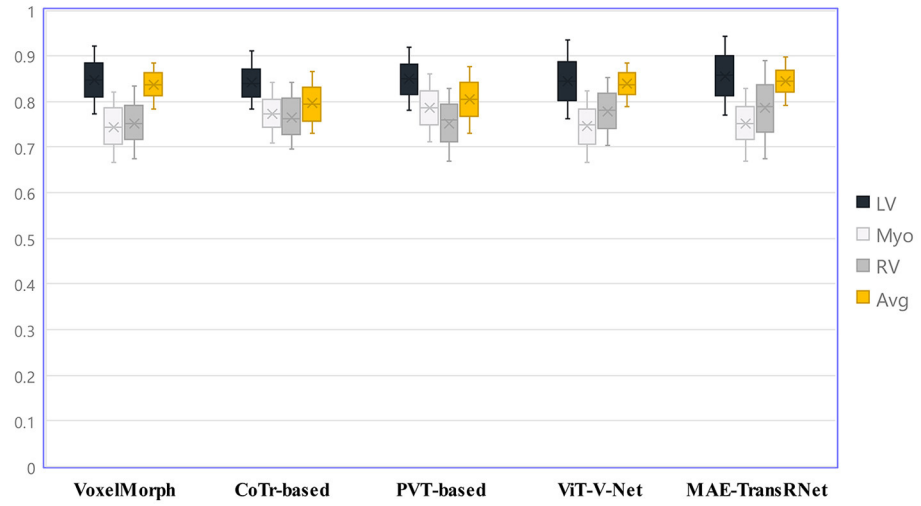


FIGURE 11
Boxplots to describe the variability of dice performance for various anatomical structures obtained under the same registration method.

TABLE 3 Ablation study about different masking ratios in the ACDC dataset for registration.

| Masking ratio | LV | | Myo | | RV | | Avg | |
|---|---|---|---|---|---|---|---|---|
| | DSC | HD | DSC | HD | DSC | HD | DSC | HD |
| 0.85 | 0.854 | 5.52 | 0.773 | 6.12 | 0.783 | 8.75 | 0.803 | 6.8 |
| **0.75** | **0.858** | **5.49** | **0.792** | **5.93** | **0.785** | **8.65** | **0.812** | **6.69** |
| 0.65 | 0.857 | 5.42 | 0.776 | 5.89 | 0.783 | 8.62 | 0.812 | 6.64 |
| 0.375 | 0.858 | 5.42 | 0.794 | 5.85 | 0.786 | 8.62 | 0.813 | 6.63 |
| **0.125** | **0.859** | **5.39** | **0.795** | **5.78** | **0.788** | **8.58** | **0.814** | **6.58** |

The best results are achieved and highlighted by the bold values.

TABLE 4 Ablation experiments on the location of SE module embedding.

| SE module position | LV | | Myo | | RV | | Avg | |
|---|---|---|---|---|---|---|---|---|
| | DSC | HD | DSC | HD | DSC | HD | DSC | HD |
| Transformer | 0.856 | 5.51 | 0.789 | 5.98 | 0.783 | 8.68 | 0.811 | 6.72 |
| CNN Block | 0.854 | 5.65 | 0.793 | 6.07 | 0.782 | 8.73 | 0.809 | 6.82 |
| **Trans+CNN** | **0.858** | **5.49** | **0.792** | **5.93** | **0.785** | **8.65** | **0.812** | **6.69** |

The best results are achieved and highlighted by the bold values.

$$Out_Q^n = \sum_{j=0}^{cn-1} S_{XQ}^{n,j} S_{XV}^{n+j-\lfloor \frac{cn}{2} \rfloor}$$
$$Out_K^n = \sum_{j=0}^{cn-1} S_{XK}^{n,j} S_{XV}^{n+j-\lfloor \frac{cn}{2} \rfloor} \quad (8)$$

$$AttnQ\_V = Out_Q^n W_3$$
$$AttnK\_V = Out_K^n W_3 \quad (9)$$

where $W_1 \in \mathbb{R}^{D\times D}$, $W_2 \in \mathbb{R}^{D\times cn}$, and $W_1 \in \mathbb{R}^{D\times D}$ are three learnable weights and $n$ denotes the number of tokens.

Finally, we aggregate the attention of the three components Q, K, and V by the traditional multi-head self-attention computation module to obtain the feature representation of hybrid attention in one attention head, and then we concatenate the outputs of all the $h$ heads and calculate the output of the HyMHSA block, formulating as:

$$AttnOut = MHSA(AttnQ\_V \cdot W^Q, AttnQ\_K \cdot W^K, S_{XV} \cdot W^V) \quad (10)$$

$$HyMHSA(X) = Concat(AttnOut_1, ..., AttnOut_h)W^m \quad (11)$$

Our architecture of the hybrid multi-head self-attention mechanism is shown in Figure 3.

## 3.3. Squeeze and excitation module in 3D CNN encoder

The channel and spatial dimensional parallel attention mechanism modules are introduced in the CNN encoder before the Transformer structure to operate on convolutional features using a dual-dimensional parallel extraction of attention features, with input feature maps of $\overline{X} \in R^{H\times W\times L\times C}$. Moreover, the attention mechanism modules [e.g., spatial squeeze and channel excitation block (cSE) and channel squeeze and spatial excitation block (sSE)] are applied to 3D CNN (Figure 4) The spatial attention module consists of a global average pooling layer and a fully connected, ReLU activation layer ($\hat{z} = W_1(\delta(W_2 z))$) behind it. This module generates the intermediate feature vector $z \in R^{1\times 1\times 1\times C}$ *via* the pooling layer while generating $n$-th element, which is expressed as follows:

$$z_n = \frac{1}{H \times W \times L} \sum_i^H \sum_j^W \sum_k^L \mathbf{u}_n(i,j,k) \quad (12)$$

In this step, the global spatial information of the image features is also embedded into the feature vector $z$. With the variation of the squeeze and excitation module, the entire attention recalibration process is expressed as follows:

$$\overline{X}_{cSE} = F_{cSE}(\overline{X}) = [\sigma(\widehat{z_1})x_1, \sigma(\widehat{z_2})x_2, \cdots, \sigma(\widehat{z_c})x_c] \quad (13)$$

where $c$ denotes the attention weight of each channel, emphasizing high-importance features and suppressing low-importance features, assigning different levels of importance to the respective channel. The other part targets the fine-grained pixel information in cardiac MRI. This part is enabled us to deeply mine the channel information of the feature map and then spatially excite it. The feature vector is expressed as $\overline{X} = \left[x^{1,1,1}, x^{1,1,2}, \cdots, x^{i,j,k}, \cdots, x^{H,W,L}\right]$, and the linear representation of the feature projection ($X_s = W_{sq} \cdot \overline{X}$) is obtained through convolution operation. The attention recalibration process is illustrated as follows:

$$\overline{X}_{sSE} = F_{sSE}(\overline{X}) = \left[\sigma\left(X_{s(1,1,1)}\right)x^{1,1,1}, \sigma\left(X_{s(i,j,k)}\right)x^{i,j,k}, \cdots, \sigma\left(X_{s(H,W,L)}\right)x^{H,W,L}\right] \quad (14)$$

The value of each $\sigma$ represents the relative importance of the spatial information $(i,j,k,c)$ for a given feature map. Accordingly, the combination of the two modules allows features on channel and spatial aspects to be considered more often in the learning process of the network. The formula is:

$$\tilde{\mathbf{X}}_{scSE} = \tilde{\mathbf{X}}_{cSE} + \tilde{\mathbf{X}}_{sSE} \quad (15)$$

## 3.4. 3D vision transformer with MAE as deformable registration core architecture

### 3.4.1. 3D vision transformer architecture

The conventional 3D ViT architecture is borrowed as the backbone for pre-training and downstream registration tasks, and the feature maps containing some high-level feature information obtained from three downsampling operations are employed as the input of ViT: $\overline{X} \in \mathbb{R}^{H\times W\times L\times C}$. The size of $P \times P \times P$ non-overlapping patches is adopted to slice the high-dimensional images to get $N = \frac{HWL}{P^3}$ patches. These patches are flattened into $P^3C$-dimension vectors, and the serialized representation with high-level features is obtained. To preserve the position information, position embedding is introduced after patch embedding, and the vector of flattened patches and the vector of position information are added for a serialized representation of the global information of the image.

TABLE 5 Comparison of image registration performance in three variants of MAE pre-training model (including dice performance and Hausdorff distance) on the ACDC dataset.

| Pretrain-Model | LV | | Myo | | RV | | Avg | |
|---|---|---|---|---|---|---|---|---|
| | DSC | HD | DSC | HD | DSC | HD | DSC | HD |
| Base | 0.858 | 5.49 | 0.792 | 5.93 | 0.785 | 8.65 | 0.812 | 6.69 |
| Large | 0.859 | 5.47 | 0.795 | 5.96 | 0.786 | 8.6 | 0.813 | 6.68 |
| Huge | 0.861 | 5.42 | 0.794 | 5.87 | 0.786 | 8.37 | 0.814 | 6.56 |

TABLE 6 Three variants of MAE's detail about configuration and parameters.

| Model | Patch size | Encoder dim | Mlp dim | ViT layers | ViT head | Params |
|---|---|---|---|---|---|---|
| Base | 16 | 768 | 3,072 | 12 | 12 | 63.837M |
| Large | 16 | 1,024 | 4,096 | 24 | 16 | 244.455M |
| Huge | 14 | 1,280 | 5,120 | 32 | 16 | 387.248M |

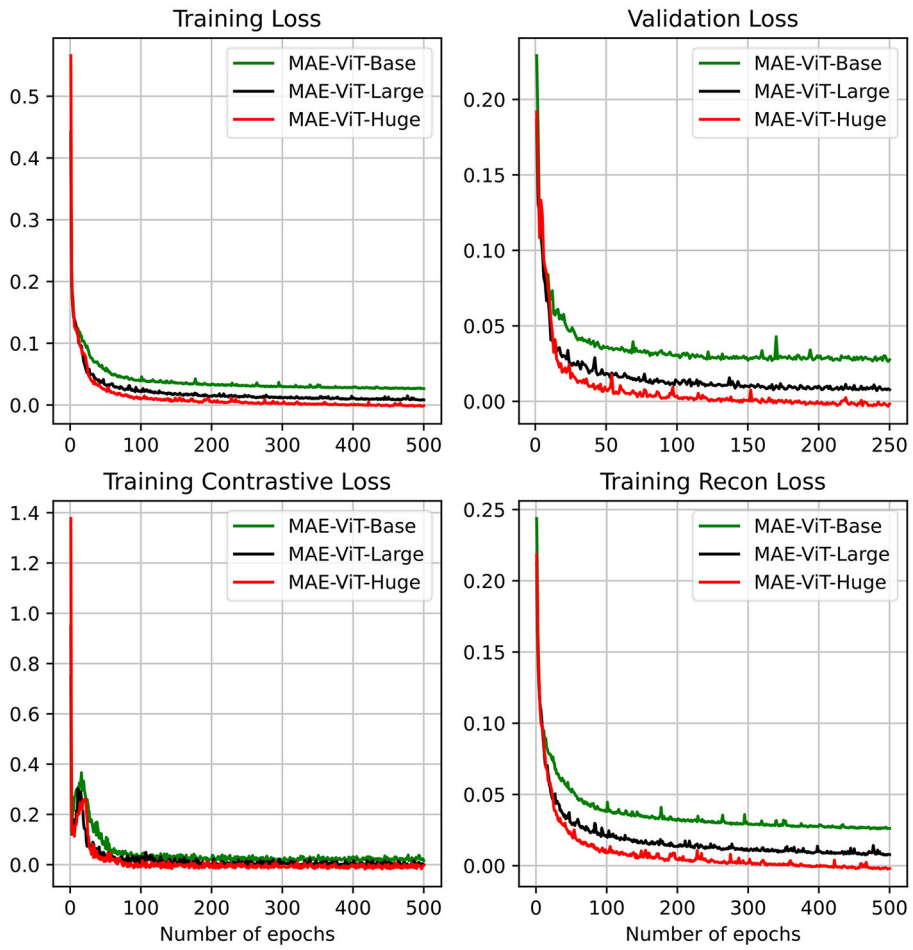


FIGURE 12
MAE's pre-training result on the ACDC dataset, including training loss, validation loss, training contrastive loss, and training reconstruction loss. We consider three different MAE variants.

### 3.4.2. Pre-training with MAE

The core part of the proposed method introduces a self-supervised learning strategy by designing 3D ViT as an autoencoder structure with random masked operations to allow the encoder to learn more high-level abstract features and by employing an asymmetric encoder -decoder structure as the core structure of the registration network. Figure 5 illustrates the 3D MAE framework adopted. The feature map is sliced into overlap patches (patch size

= 8) in the conventional ViT approach, accessing the position embedding. For the above patches, the patches above half of the ratio (masking ratio = 0.75) are masked, only a small portion of patches that are visible to the encoder are kept, and then the patches required to be masked are calculated. Next, random indices are obtained and divided into the masked and unmasked parts. The unmasked part is the shallow representation of the high-level features, while the masked part is represented by a shared and learnable vector. Each masked patche can be represented as the same vector. As depicted in Figure 5, only the unmasked patches are fed into ViT after the masked operation. After the linear projection, the patches are converted to one-dimensional tokens, and the blank positions (yellow parts) are all filled by the same vector of the same dimension and then decoded by a layer of Transformer decoder. In the MAE, the MSE is used as the reconstruction loss function, and the reconstruction effect is measured by obtaining the MSE between the reconstructed image and the original image in the pixel space.

#### 3.4.3. Designed architecture applied to downstream tasks

A simple layer is designed as the registration head according to the downstream registration task. Before this layer, five CNN decoder layers are also designed to reconstruct the feature representation obtained by the Transformer Block. Subsequently, the feature representation is recovered to the image data format and gradually upsampled back to the original resolution, as presented in Figure 5.

### 3.5. Loss functions in the registration model

The loss function in the registration model consists of a mean square error (MSE) similarity loss and a regularized smoothing loss based on a folding penalty and the sum of the two is used as the loss between the moving image $M$, the fixed image $F$, and the deformation field $\varphi$. The loss function is given by:

$$\mathcal{L}(F, M, \varphi) = \mathcal{L}_{MSE}(F, M, \varphi) + \alpha P \tag{16}$$

where $\mathcal{L}_{MSE}(F, M, \varphi)$ is the mean square error similarity loss, $\alpha$ is the regularization parameter, $P$ is the regularization loss based on the folding penalty, and the two parts of the loss function are formulated as:

$$\mathcal{L}_{MSE}(F, M, \varphi) = \mathcal{L}(\Theta) = \frac{1}{\Omega} \sum_{\Theta \in \omega} [F(\Theta) - M \circ \varphi(\Theta)]^2 \tag{17}$$

$$P = \frac{1}{V} \int_0^X \int_0^Y \int_0^Z \left[ \left( \frac{\partial^2 \mathbf{T}}{\partial x^2} \right)^2 + \left( \frac{\partial^2 \mathbf{T}}{\partial y^2} \right)^2 + \left( \frac{\partial^2 \mathbf{T}}{\partial z^2} \right)^2 + 2 \left( \frac{\partial^2 \mathbf{T}}{\partial xy} \right)^2 + 2 \left( \frac{\partial^2 \mathbf{T}}{\partial xz} \right)^2 + 2 \left( \frac{\partial^2 \mathbf{T}}{\partial yz} \right)^2 \right] dxdydz \tag{18}$$

In the mean square error similarity loss function, $\Theta$ is the network parameter to be learned, $\Omega$ represents the image domain, and $M \circ \varphi(\Theta)$ denotes the moving image after spatial transform (warped layer); in the regularized loss function based on the folding penalty, the essence of the function is to penalize the folding region of the deformation field, $V$ is the volume of the 3D image domain, $T$ is the local spatial transformation, and adding this term minimizes the second-order derivative of the local transformation of the deformation field, which leads to an affine transformation of the local deformation field and thus enhances the smoothness of the global deformation field.

## 4. Experiments

### 4.1. Preparation of datasets and related setting details

The dataset used for the experiment and the related settings are described. In this study, the dataset applied is the publicly available benchmark dataset from the automated cardiac diagnosis challenge (31) (ACDC) in 2017. This dataset contains short-axis cardiac 3D MR images from a total of 150 cases for two-time frames of initial frame-end frame, and the public dataset applied provides standard segmentation labels for three parts (including the left ventricle (LV), the left ventricular myocardium (Myo), and the right ventricle (RV)) for the registration task, which involves five categories of cases (including normal, heart failure with infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, and right ventricular abnormalities). Hundred cases of the above 150 cases contain the triple segmentation labels, while 50 cases do not contain labels. The same dataset is employed for the pre-training task and the downstream task. For the pre-training task, 250 cases are employed for training, and 50 cases are applied for validation (only include images). For the downstream task, the image parts are extracted in 40 cases (1–40), and the complete cardiac cycle images of 50 cases (101–150) for a total of 90 cases are extracted as the training set, 20 cases (41–60) containing images and labels are extracted as the validation set, and 40 cases of data from cases (61–100) are extracted as the test set. At the data preprocessing stage, all the images are cropped to 64 × 128 × 128, the random flip is adopted as the data augment method for the training set to increase the sample size of the dataset. Furthermore, the label pixel normalization method is applied for the validation and test sets to preprocess the data.

### 4.2. MAE architecture for pre-training

In the pre-training task, the three variants of MAE architecture (MAE-ViT-Base, MAE-ViT-Large, and MAE-ViT-Huge) are adopted to pre-train the heart dataset to compare how well the model learns cardiac image features. Unlike the original method of pre-training with the ImageNet-1K (32) dataset, the ACDC dataset is employed, which is divided into 250 cases and 50 cases for training and validation, respectively. The learning rate is set to 1e-4, and MAE pre-training is run for 500 epochs. Moreover, the batch size is set to 2, and the masking ratio is set to 0.75 (default setting) to save the pre-trained MAE model obtained in the pre-training stage for testing some subsequent results.

### 4.3. Downstream task–Cardiac MRI registration

The method proposed in this study is a hybrid network of CNN and Transformer structures, while some structures with Transformer structures are employed as the main backbone network to access in the registration task for comparison. Thus, VoxelMorph is adopted as the baseline network. The PyTorch framework is employed to implement all methods for the comparison experiments. The MONAI framework is used to visualize the registration results, and the methods of the experiments are completed on an NVIDIA RTX 3090 GPU. The Adam optimizer and the step decay (power = 0.9) learning rate reduction strategy are employed in all neural networks. nii format is converted into 3D volume npz format for two time frames from the dataset, and the two-time frames in the respective 3D format in the training sets are converted into fixed image and moving image, respectively. Subsequently, the validation sets and test sets let the image and label of each of the two-time frames form a 3D image pair. On that basis, the respective image is matched with the image of another time frame in a random combination, thus forming four pairs of fixed image and moving image (360 pairs, 80 pairs, and 160 pairs). The proposed framework is compared primarily with several typical methods based on deep learning, which include the baseline framework for registration, VoxelMorph, and three networks [CoTr-based (33) registration network, PVT-based (34) registration network, and ViT-V-Net)] for several applications of the Transformer backbone. The single-channel fixed image and the moving image are combined into a 3D grayscale image with a channel number of 2 as the input of the network. All inputs are subjected to the same preprocessing. The batch size is set to 2, the initial learning rate is set to 0.0001, and the training rounds of 500 epochs are set. The whole process is performed by downsampling the input five-dimensional tensor. Subsequently, the obtained high-level feature representation is divided into equal-sized patches through patch embedding operation. For patches, the remaining visible unmasked patches are fed into the encoder of the 3D vision transformer, so the deformation is achieved from the input image to the predicted densely aligned deformation field using the spatial transformer network. The proposed model is trained by optimizing the loss function for the similarity between the fixed and moving images. For the metrics to evaluate the registration effect, dice coefficient (DSC) and hausdorff distance (HD) are selected to evaluate the 3D registration results.

## 5. Results

### 5.1. Cardiac image in MAE reconstruction

The reconstruction effect for cardiac MRI is tested by pre-training the model on the ACDC dataset. Figure 6 presents the results of three variants of MAE architecture's reconstruction at a mask ratio of 0.75. As revealed by the results, although the resolution of visible patches in the reconstructed image is reduced, and the three model variants differ in their reconstruction of cardiac images. The MAE can still recover the lost information from the pixels around the missing patches effectively. The recovered features can be better applied to downstream tasks.

### 5.2. Cardiac MRI registration

The method applied takes the dice coefficient and hausdorff distance as measurement metrics. The proposed method is compared with several advanced registration methods currently available, and the experiments are performed on 150-cases ACDC dataset. The comparison results achieved for dice performance and Hausdorff distance are listed in Table 1. Some representative registration methods are based on deep learning, including the unsupervised registration baseline -VoxelMorph, as well as the registration network with 3D PVT-based, CoTr-based, and ViT-V-Net.

The visualization results of the attention heat map of the ACDC dataset in several models are shown in Figure 7. We compared the proposed method with VoxelMorph and ViT-V-Net to compare the models in terms of feature aggregation. In the visualization results, brighter regions indicate a higher degree of feature aggregation. These visualization results of the attention heat map show that all three methods can aggregate different features in three regions of the left and right ventricles and ventricular walls. In contrast, our method can wrap the target contour region more comprehensively.

Figure 8 presents the registration results of the whole cardiac organ and the left and right ventricles obtained by the proposed method and the generated deformation fields, including three different periods of registration. The proposed method is capable of enhancing the dice performance by nearly 0.01 and decreasing the Hausdorff distance by about 0.1, respectively, compared with other methods, and the loss values of the proposed method are kept at a lower level during the training process, and the dice performance values obtained from the validation set are higher (Figures 9A, B). In the meantime, we set the value of contextual neighbors in self-attention to 100 by default and compare the time complexity of traditional self-attention, naive local intensive attention, and hybrid local dense attention in our model. The results show that the model can improve model performance while maintaining a low time complexity. These results are shown in Table 2. In brief, the MAE-TransRNet achieves better registration results and verifies the effectiveness of MAE, SE, and HyMHSA modules introduced into the registration task. In the meantime, we used boxplots to describe the variability of dice performance obtained by different registration methods for the same anatomical structure and also the variability of dice performance for various anatomical structures obtained using the same registration method (Figures 10, 11).

## 6. Ablation study

To evaluate the effect of our proposed MAE-TransRNet more accurately, a series of ablation experiments are set to verify the performance of the model under different settings, including the masking ratio size, and whether to add different dimensions to the Transformer encoder and CNN modules, and the effect of the MAE model size on the effect of the registration task. All the training epoch is set to 500.

### 6.1. Masking ratio

The default value of the masking ratio is set to 0.75 as the baseline framework for the experiment, and the experimental settings are used when the masking ratio is set to other values to explore the effect of the masking ratio on the final registration effect. The result is listed in Table 3.

### 6.2. SE module's position

The SE module is introduced in the MAE-TransRNet architecture. Because of the Transformer-ConvNet architecture, the SE module is embedded into the Transformer block and the scSE module into the CNN block, and the effect of SE embedding on the model is compared at the above two positions. It is found that the dice coefficients and HD of the registration are slightly improved by introducing SE module either in the Transformer block or in the CNN block, and the results are better when SE module is introduced in both parts of the architecture, thus suggesting that the attention mechanism based on the channel and spatial dimensions in the Transformer block and CNN block is beneficial. The results of our experiments are listed in Table 4.

### 6.3. Model scaling

Finally, we provide an ablation study on different model sizes of MAE pre-training model. In particular, three different configurations, including the "Base," "Large," and "Huge" models, are investigated. For the "base" model, the patch size, encoder dim, MLP dim, number of ViT layers, and number of ViT heads are set to be 16, 768, 3,072, 12, and 12. It is concluded that larger model results in a better performance. For the huge computation cost, the MAE-ViT-Base model is applied to all the experiments. The result and the related configuration are shown in Tables 5, 6. Moreover, the related train loss value is presented in Figure 12.

## 7. Discussion

An unsupervised learning deformable image registration model is proposed based on Transformer-ConvNet. It has been implemented to predict the spatial transformation parameters between input image pairs by introducing ViT. There are two differences between most deep learning-based methods, especially some methods that introduce the Transformer as follows:

(1) The proposed model is trained by continuously optimizing the image similarity metric without any label as ground truth, while the label is used to support validation and testing. Thus, the registration effectiveness is measured.

(2) We designed the core of the Transformer as a self-encoder and lightweight decoder structure with a MAE, turning the feature extraction prior to the registration downstream task into a self-supervised learning task.

The cardiac MRI dataset for the ACDC is evaluated. The experimental results suggest that the model can outperform the baseline model of deep learning-based deformable registration and slightly outperform some other Transformer-based registration methods. A MAE is applied to the heart registration task first from the difference between text and image information. The method of masking more than half of the patches significantly reduces the redundancy of images, making the feature extraction task more challenging and forcing the model to learn more deeply hidden and better representations. Our purpose in introducing two SE modules is to enhance the feature representation capability of the Transformer structure and the CNN structure. The purpose of introducing the scSE module in the CNN structure is to help us dig deeper into the fine-grained information of the feature map by considering the importance of features in the channel and spatial dimensions to the fine-grained pixel information in the heart image; we introduce the SE module in the self-focus mechanism, hoping to analogize the application scenario of SE in convolution to do query, key, and value in the self-focus computation, respectively. We successfully introduce some convolutional induction bias in the Transformer module to enhance the extraction of local information. Also, we are the first to use this kind of local dense attention in the vision domain, especially in the alignment task. We believe that this self-attention mechanism based on local neighbor context is useful for medical image analysis tasks. The results of several comparison experiments and ablation studies suggest that using the MAE for medical image registration tasks is of great significance in the effect improvement, and the MAE with different scales has a slight difference in the reconstruction effect of cardiac images. It is more appropriate to select "Base" as the baseline model to avoid a high cost of computation. It is worth discussing that, unlike the results when the MAE with a high masking ratio is applied to natural images (e.g., ImageNet-1K dataset), a high masking ratio does not make the MAE achieve the optimal result in medical image tasks. Since the masking ratio is continuously adjusted downward, the effectiveness of our registration tasks is increased slightly, which also suggests that the masking ratio of MAE has different effects on different image analysis tasks. Moreover, the embedding of the SE module in Transformer-ConvNet structure plays a positive role in feature extraction to a certain extent.

However, the effect of the proposed method compared with other methods on the cardiac registration task does not differ significantly between models, probably because the dataset size is relatively small and the model parameters are great. In addition, for the part of MAE, before feeding into the decoder, a part of the token in the blank position is filled in by sharing the learnable vector, which essentially generates non-existent content and is easy to mislead the original features of the image. Accordingly, if the potential impact is further considered, our future research is devoted to the design of the model to be more lightweight, considering the realism of the underlying information representation, while trying to scale up a certain amount of dataset size to further enhance the registration performance.

## 8. Conclusion

An unsupervised learning deformable image registration method is proposed based on Transformer-ConvNet structure,

which changes the original ViT structure, introduces mask operations, and does not require segmentation labels as registration information. Furthermore, we introduce a new multi-head self-attention mechanism that sets the range of the model considering neighbors so that the attention module only computes contextual information within a limited distance from the current location. The result of this study verifies that the MAE-TransRNet can achieve results comparable to several popular methods at present and still has much room for improvement. Future research may be extended to multimodal cardiac image registration tasks.

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://www.creatis.insa-lyon.fr/Challenge/acdc/index.html. Codes and models are available at: https://github.com/XinXiao101/MAE-TransRNet.

## Author contributions

XX, SD, and ZQ conceived this study. YY and YL were the developers of computer-aided diagnosis methods. XX and GY completed the data analysis. XX and SD drafted the manuscript. All authors were involved in the finalization of the manuscript and approved the manuscript.

## Funding

This work was financially supported by the National Natural Science Foundation of China under Grant (62202092), the Key R&D Project of Heilongjiang Province (No. 2022ZX01A30), the Science and Technology Program of Suzhou (Nos. ZXL2021431 and RC2021130), the Fundamental Research Funds for the Central Universities (No. 2572016BB12), People's Republic of China, the Fundamental Research Funds for the Central Universities (No. 2572020DR10), Beijing Hospitals Authority's Ascent Plan (Code: DFL20220605), and the Beijing Nova Program (No. 20220484174).

## 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. Sotiras A, Davatzikos C, Paragios N. Deformable medical image registration: a survey. *IEEE Trans Med Imaging*. (2013) 32:1153–90. doi: 10.1109/TMI.2013.2265603

2. Shewaye T. Cardiac MR image segmentation techniques: an overview. *arXiv:1502.04252 [cs.CV]* (2015). doi: 10.48550/arXiv.1502.04252

3. Potel M, Mackay S, Rubin J, Aisen A, Sayre R. Three-dimensional left ventricular wall motion in man. *Invest Radiol*. (1984) 19:499–509. doi: 10.1097/00004424-198411000-00006

4. Ye M, Kanski M, Yang D, Chang Q, Yan Z, Huang Q, et al. DeepTag: an unsupervised deep learning method for motion tracking on cardiac tagging magnetic resonance images. In: *Proceedings - 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2021*. IEEE (2021). p. 7257–67.

5. Mäkelä T, Clarysse P, Sipilä O, Pauna N, Pham QC, Katila T, et al. A review of cardiac image registration methods. *IEEE Trans Med Imaging*. (2002) 21:1011–21. doi: 10.1109/TMI.2002.804441

6. Geert, Litjens, Thijs, Kooi, Babak, Ehteshami, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. (2017) 42:60–88. doi: 10.1016/j.media.2017.07.005

7. Haskins G, Kruger U, Yan P. Deep learning in medical image registration: a survey. *Mach Vis Appl*. (2019) 31:8. doi: 10.1007/s00138-020-01060-x

8. Rohe MM, Datar M, Heimann T, Sermesant M, Pennec X. SVF-Net: learning deformable image registration using shape matching. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins D, Duchesne S, editors. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017. MICCAI 2017. Lecture Notes in Computer Science*(), *Vol. 10433*. Cham: Springer (2017). p. 266–74.

9. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, editors. *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, Vol. 9351*. Cham: Springer (2015).

10. Krebs J, Delingette H, Mailhe B, Ayache N, Mansi T. Learning a probabilistic model for diffeomorphic registration. *IEEE Trans Med Imaging*. (2019) 2:7112. doi: 10.1109/TMI.2019.2897112

11. Balakrishnan G, Zhao A, Sabuncu M, Guttag J, Dalca A. VoxelMorph: a learning framework for deformable medical image registration. *IEEE Trans Med Imaging*. (2019) 2019:964. doi: 10.1109/CVPR.2018.00964

12. Hu Y, Modat M, Gibson E, Li W, Ghavami N, Bonmati E, et al. Weakly-supervised convolutional neural networks for multimodal image registration. *Med Image Anal*. (2018) 49:2. doi: 10.1016/j.media.2018.07.002

13. Miao S, Wang Z, Liao R. A CNN regression approach for real-time 2D/3D registration. *IEEE Trans Med Imaging*. (2016) 35:1800. doi: 10.1109/TMI.2016.2521800

14. Naseer M, Hayat M, Zamir SW, Khan F, Shah M. Transformers in vision: a survey. *ACM Comput Surveys*. (2022) 54:1–41. doi: 10.1145/3505244

15. Han K, Wang Y, Chen H, Chen X, Tao D. A survey on visual transformer. *arXiv:2012.12556 [cs.CV]*. (2020). doi: 10.48550/arXiv.2012.12556

16. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Houlsby N. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv:2010.11929 [cs.CV]* (2020). doi: 10.48550/arXiv.2010.11929

17. Chen J, Lu Y, Yu Q, Luo X, Zhou Y. TransUNet: transformers make strong encoders for medical image segmentation. *arXiv:2102.04306 [cs.CV]*. (2021). doi: 10.48550/arXiv.2102.04306

18. Chen J, He Y, Frey EC, Li Y, Du Y. ViT-V-Net: vision transformer for unsupervised volumetric medical image registration. *Med Image Anal*. (2021) 82:102615. doi: 10.1016/j.media.2022.102615

19. Milletari F, Navab N, Ahmadi SA. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: *2016 Fourth International Conference on 3D Vision (3DV)*. Stanford, CA: IEEE (2016). p. 565–71.

20. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE (2022). p. 16000–9.

21. Wu C, Wu F, Ge S, Qi T, Huang Y, Xie X. Neural news recommendation with multi-head self-attention. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. (2019). p. 6390–5.

22. Alamri F, Dutta A. Multi-head self-attention via vision transformer for zero-shot learning. *arXiv:2108.00045 [cs.CV*]. (2021). doi: 10.48550/arXiv.2108.00045

23. Hong Y, Zhang Y, Schindler K, Martin R. Context-aware multi-head self-attentional neural network model for next location prediction. *arXiv:2212.01953 [physics.soc-ph*] (2022). doi: 10.48550/arXiv.2212.01953

24. Guha Roy A, Navab N, Wachinger C. Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. In: Frangi A, Schnabel J, Davatzikos C, Alberola-López C, Fichtinger G, editors. *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018. MICCAI 2018. Lecture Notes in Computer Science, Vol. 11070*. Cham: Springer (2018). p. 421–9.

25. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE (2018). p. 7132–41.

26. Chiu CC, Raffel C. Monotonic chunkwise attention. *arXiv [Preprint]*. (2017). arXiv: 1712.05382.

27. Tay Y, Bahri D, Metzler D, Juan DC, Zhao Z, Zheng C. Synthesizer: Rethinking self-attention for transformer models. In: *International Conference on Machine Learning*. PMLR (2021). p. 10183–92.

28. Xu M, Li S, Zhang XL. Transformer-based end-to-end speech recognition with local dense synthesizer attention. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON: IEEE (2021). p. 5899–903.

29. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems*, Vol. 30. (2017).

30. Chollet F. Xception: deep learning with depthwise separable convolutions. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE (2017). p. 1800–7.

31. Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng PA, et al. Deep Learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Trans Med Imaging*. (2018) 37:2514–25. doi: 10.1109/TMI.2018.2837502

32. Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, FL: IEEE (2009). p. 248–55.

33. Xie Y, Zhang J, Shen C, Xia Y. CoTr: efficiently bridging CNN and transformer for 3D Medical image segmentation. In: *Medical Image Computing and Computer Assisted Intervention-MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, Vol. 12903*. Cham: Springer (2021). p. 171–80.

34. Wang W, Xie E, Li X, Fan DP, Song K, Liang D, et al. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC: IEEE (2021). p. 548–58.

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 14 April 2023
DOI 10.3389/fmed.2023.1142261



OPEN ACCESS

EDITED BY
Jinming Duan,
University of Birmingham,
United Kingdom

REVIEWED BY
Yanda Meng,
University of Liverpool,
United Kingdom
Qingjie Meng,
Imperial College London,
United Kingdom

*CORRESPONDENCE
Yunping Fan
✉ zhfanyp@163.com
Linlin Shen
✉ llshen@szu.edu.cn

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Nuclear Medicine, a section of the journal Frontiers in Medicine

RECEIVED 11 January 2023
ACCEPTED 23 March 2023
PUBLISHED 14 April 2023

CITATION
Bi M, Zheng S, Li X, Liu H, Feng X, Fan Y and Shen L (2023) MIB-ANet: A novel multi-scale deep network for nasal endoscopy-based adenoid hypertrophy grading. *Front. Med.* 10:1142261. doi: 10.3389/fmed.2023.1142261



# MIB-ANet: A novel multi-scale deep network for nasal endoscopy-based adenoid hypertrophy grading

Mingmin Bi[1†], Siting Zheng[2,3†], Xuechen Li[2,3], Haiyan Liu[1], Xiaoshan Feng[1], Yunping Fan[1*] and Linlin Shen[2,3*]

[1]Department of Otolaryngology, The Seventh Affiliated Hospital of Sun Yat-sen University, Shenzhen, China, [2]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, [3]AI Research Center for Medical Image Analysis and Diagnosis, Shenzhen University, Shenzhen, China

**Introduction:** To develop a novel deep learning model to automatically grade adenoid hypertrophy, based on nasal endoscopy, and asses its performance with that of E.N.T. clinicians.

**Methods:** A total of 3,179 nasoendoscopic images, including 4-grade adenoid hypertrophy (Parikh grading standard, 2006), were collected to develop and test deep neural networks. MIB-ANet, a novel multi-scale grading network, was created for adenoid hypertrophy grading. A comparison between MIB-ANet and E.N.T. clinicians was conducted.

**Results:** In the SYSU-SZU-EA Dataset, the MIB-ANet achieved 0.76251F1 score and 0.76807 accuracy, and showed the best classification performance among all of the networks. The visualized heatmaps show that MIB-ANet can detect whether adenoid contact with adjacent tissues, which was interpretable for clinical decision. MIB-ANet achieved at least 6.38% higher F1 score and 4.31% higher accuracy than the junior E.N.T. clinician, with much higher (80× faster) diagnosing speed.

**Discussion:** The novel multi-scale grading network MIB-ANet, designed for adenoid hypertrophy, achieved better classification performance than four classical CNNs and the junior E.N.T. clinician. Nonetheless, further studies are required to improve the accuracy of MIB-ANet.

KEYWORDS

**adenoid hypertrophy, nasal endoscopy, deep learning, medical image classification, convolutional neural networks**

## 1. Introduction

Adenoid hypertrophy is a common disease in children with otolaryngology diseases. A meta-analysis showed that the prevalence of adenoid hypertrophy in children and adolescents was 34.46% (1). Adenectomy or adenotomy is the first-recommended therapy for sleep disordered breathing in children, with "adenoid faces" (2) and other growth and development problems. Clinically, surgical indication is on the basis of the grading of adenoid hypertrophy. There are four main grading standard of adenoid hypertrophy based on nasal endoscopy, i.e.,

Clemens grading standard (3), Cassano grading standard (4), Parikh grading standard (5), and ACE grading system (6). Among which, Parikh grading standard, which was reported on Otolaryngol Head Neck Surg. in 2006, grades adenoid hypertrophy by evaluating the adjacent structure of adenoid tissue contact, which can reflect the degree of blockage in the Eustachian tube and can be related to the meaning of the surgery. However, long-time reading of different images is a tedious work and may cause misdiagnosis, especially for interns without experiences. Creating an artificial intelligence deep network for nasal endoscopy-based adenoid hypertrophy grading is meaningful.

In recent years, many deep learning methods, especially convolutional neural networks (CNNs), have been applied in the medical image domain (7–12). For adenoid hypertrophy, Shen et al. (13) collected 688 lateral cranial X-ray images of patients with adenoid hypertrophy, and divided these images into training set (488), validation set (64) and test set (116). This deep learning model calculated the AN ratio (AN ratio, where A is the absolute size of the adenoid and N is the size of the nasopharyngeal space) to grade adenoid hypertrophy. Liu et al. (14) collected 1,023 lateral cranial X-ray images, and proposed a deep learning model based on VGG16 to grade adenoid hypertrophy. In the clinic, nasoendoscope is a simple, economical, readily available, and reproducible way to diagnose adenoid hypertrophy. Compared to lateral cranial X-ray, nasoendoscope requires no radiation and provides good view to investigate the distance relationship between adenoid and adjacent structures. However, to the best of our knowledge, there is no deep learning research available to help grade endoscopic images of adenoid hypertrophy.

Inspired by the success of previous works in detection and classification of medical endoscopic images, in this study, we assumed that the adenoid hypertrophy grading could also benefit from deep learning techniques. Toward this end, we acquired a large collection of nasal endoscopic images to build a novel MIB-ANet model and assessed its performance.

# 2. Materials and methods

## 2.1. SYSU-SZU-EA dataset

We reviewed the nasoendoscopic images of patients who underwent routine clinical screening for nasal diseases at the Seventh Affiliated Hospital of Sun Yat-sen University (Shenzhen, China), between December 2019 and May 2021. All of the images in SYSU-SZU-EA Dataset were original nasoendoscopic images, without artificial light, zoom, and optical amplification restrictions. We only choose images capturing adenoid residue or adenoid hypertrophy. There was no limitation for age, gender, or whether to combine chronic rhinosinusitis or other diseases. This dataset consists of 3,179 images. All images were captured using a rigid 0-degree 2.7 mm nasoendoscope and endoscopic capture recorder (Wolf, Tuttlingen, Germany), equipped with high-performance medical imaging workstation. All of the images were saved with JPG format consisting of red, green, and blue color channels and had widths and heights ranging from 700 and 1,000 pixels. All the patients had signed informed consent before nasoendoscopy.

## 2.2. Grading method of adenoid hypertrophy

There are four main grading standard of adenoid hypertrophy, i.e., Clemens grading standard (3), Cassano grading standard (4), Parikh grading standard (5), and ACE grading system (6). Among which, Parikh grading standard grades adenoid hypertrophy by evaluating the adjacent structure of adenoid tissue contact, which can reflect 3D structure and requires few parameters, and is convenient for clinical evaluation and deep learning. Therefore, in this work Parikh grading standard were chosen as the grading method. Table 1 shows the grading method of adenoid hypertrophy and the detailed numbers of images of four grades. Adenoid hypertrophy is divided into 1–4 grades according to whether the adenoid tissue contacted or pressed the Eustachian tube pillow, vomer bone, and soft palate in a relaxed state. Figure 1 shows four example adenoid images with grades 1 to 4 in the SYSU-SZU-EA Dataset. Three E.N.T. clinicians, including one senior E.N.T. clinician, one intermediate E.N.T. clinician and one junior E.N.T. clinician were employed for data annotation.

## 2.3. Preprocessing

*Computer implementation environment*: The neural network models were coded in Python (version 3.7.6, 64 bit) using the open-source Pytorch (version 1.8.1) library and tested on Intel (R) Xeon (R) Gold 6,132 CPU @2.60GHz and a Tesla V100. Due to limited GPU resources, all images were resized to 256 × 256 pixels. In the training phase, we used a learning rate of 0.0001 and a batch size of 32 in the Adam optimizer, and used the "StepLR" with step size of 10, gamma of 0.9 to decay the learning rate. In addition, we employed random vertical flip, random horizontal flip, and random rotation on the input images to augment the dataset in training.

*Data distribution of training set, validation set and test set*: We randomly divided the 2,183 adenoid images into training set and validation set. The ratio of the image number of training set to the validation set is 4:1. In order to ensure that the number of adenoid images at each grade in the training set is sufficient, the dividing ratio for grade 1 and 2 was set as approximately 4:1, and the dividing ratio for grade 3 and 4 was set as approximately 5:1. For testing set, 996 images were graded by 3 E.N.T. clinicians with different experiences and the final result was determined based on majority voting. The detailed distribution of adenoid hypertrophy images with different grades in training set, validation set, and test set is shown in Table 1.

## 2.4. The novel multi-scale grading network: MIB-ANet

In this paper, we designed a framework, MIB-ANet, for adenoid hypertrophy classification. As shown in Figure 2A, the proposed MIB-ANet consisted of two modules, the backbone network—ANet and Modified Inception Block (MIB). MIBs and ANet were integrated as MIB-ANet by replacing the first two layers of ANet (red dotted box) with MIBs (blue box), whose details are shown in Figures 2B,C and Supplementary B.

TABLE 1 Details of Parikh grading standard and data distribution of training set, validation set, and test set in SYSU-SZU-EA dataset.

| Grade | Adjacent structure of adenoid tissue contact | Training set | Validation set | Test set | Number |
|---|---|---|---|---|---|
| 1 | None | 428 | 122 | 228 | 778 |
| 2 | Torus tubarius | 576 | 158 | 276 | 1,010 |
| 3 | Torus tubarius, vomer | 492 | 104 | 355 | 951 |
| 4 | Torus tubarius, vomer, palate (at rest) | 250 | 53 | 137 | 440 |
| Total | | 1746 | 437 | 996 | 3,179 |

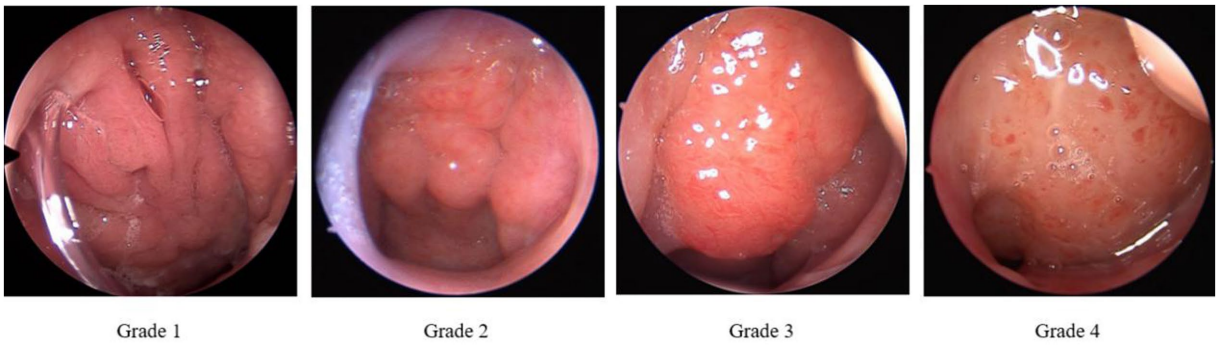


FIGURE 1
Examples of 4 grades adenoid nasoendoscopic images according to Parikh grading standard in the SYSU-SZU-EA dataset.

## 2.5. Performance evaluation

In this study, four classic CNNs, i.e., AlexNet (15), VGG16 (16), ResNet50 (17), and GoogleNet (18), were employed for performance comparison. Details of the structure of four classic CNNs and ANet are described in Supplementary A. Accuracy, F1 score and confusion matrix were adopted as the evaluation metrics of classification performance. Definition of Evaluation Metrics are described in Supplementary B. Details of ablation study for Classification Performance evaluation are described in Supplementary C. We also used the Class Activation Map (CAM) (19) to visualize the attention map of different CNNs, which can highlight the regions of interest of different models. The comparison of the performance of MIB-ANet, ANet and four classic CNNs are showed in Supplementary D.

## 2.6. Comparison between MIB-ANet and E.N.T. clinicians

We compared the diagnostic performance of MIB-ANet with three E.N.T. clinicians. While the senior E.N.T. clinician has more than 20 years of experience in nasal endoscopy, the intermediate and junior E.N.T. clinician has approximately 8 years and 5 years of experience in nasal endoscopy, respectively. They conducted blind assessments of 996 images in testing set and the final result was determined based on majority voting. We compared MIB-ANet with human experts using F1 score and accuracy.

## 2.7. Ethics

The study was approved by the ethical review board of the Seventh affiliated Hospital of Sun Yat-sen University (no. KY-2022-008-01).

## 2.8. Statistical analysis

ROC curves were adopted as the evaluation metrics of classification performance, which were coded in Python (version 3.7.6, 64 bit). Wilcoxon signed-rank test was used to analyze the difference between two paired samples of ordinal categorical variables, which was performed by SPSS 17.0. All tests were two-sided, and $p < 0.05$ was considered as statistically significant.

# 3. Results

## 3.1. Comparison based on F1 score and accuracy

We compared the performance of MIB-ANet to E.N.T. clinicians. Since the test set was annotated by 3 E.N.T. clinicians independently, the ground truth was determined based on majority voting and a face-to-face discussion of these 3 E.N.T. clinicians. Therefore, we evaluated the performance of each doctor by calculating the F1 score and accuracy of their diagnostic results with the voted ground truth. Table 2 shows the performance of MIB-ANet and 3 E.N.T. clinicians. From

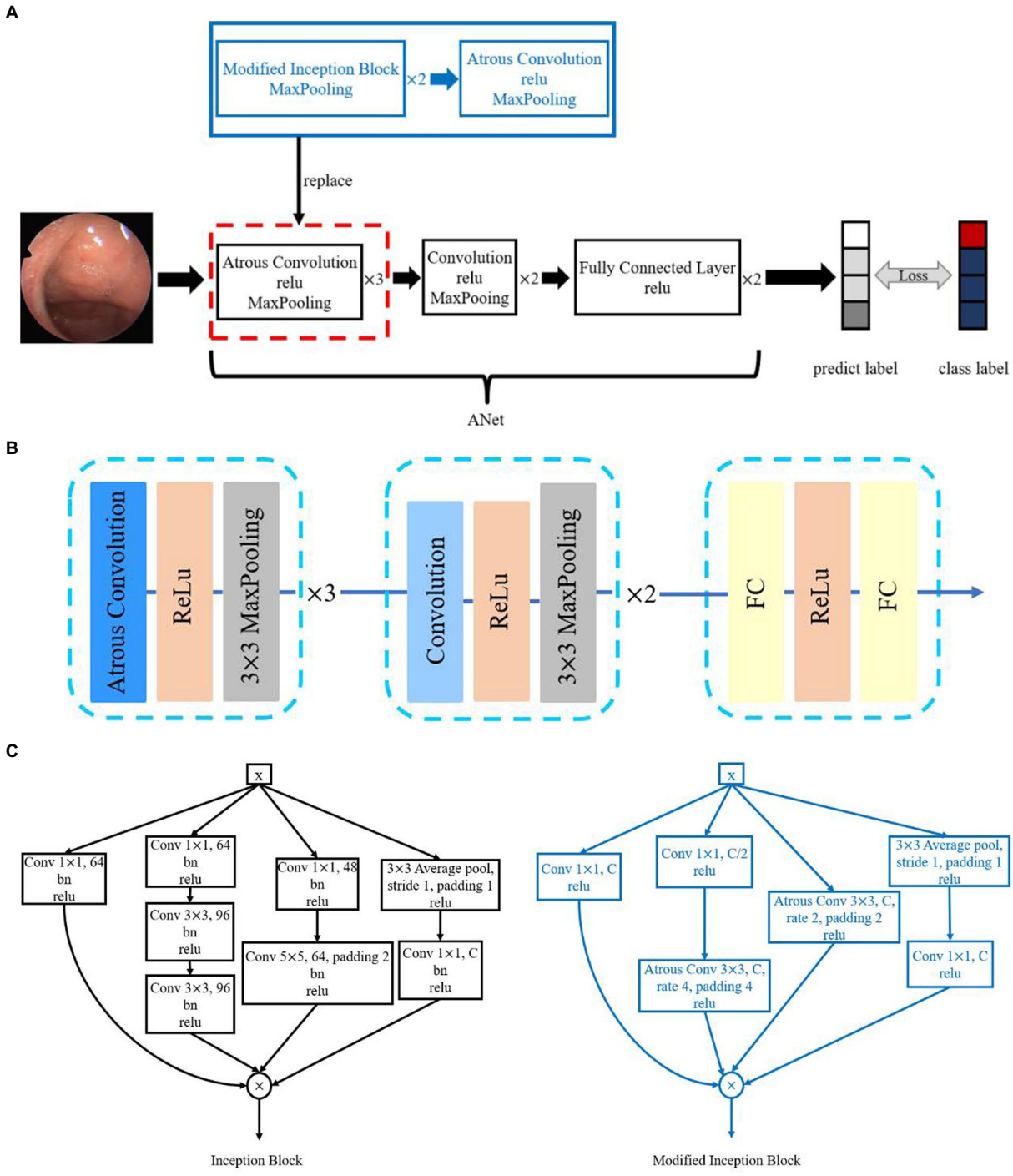


FIGURE 2
**(A)** The overview of the proposed MIB-ANet architecture. **(B)** The architecture of ANet. **(C)** The architecture of Inception Block and Modified Inception Block.

Table 2, we can see that MIB-ANet achieved at least 6% higher F1 score and 4% higher accuracy than the junior clinician, and achieved much higher diagnosing speed than human experts, e.g., at least 80 times faster than the senior clinician. Table 2 also shows the detailed *Z* and *p* value between the voted ground truth and MIB-ANet or 3 E.N.T. clinicians. Since the classification results were ordinal categorical variables, two-sided Wilcoxon signed-rank test was employed to analyze the difference between two paired samples. As we know, *p* value indicates the statistical significance and *Z* value indicates the tendentiousness. The *p* value of MIB-ANet was 0.188, which showed that there was no significant statistical difference between the voted ground truth and MIB-ANet. However, the *p* values of 3 E.N.T. clinicians were smaller than 0.05, which meant that there were significant statistical differences between the voted ground truth and 3 E.N.T. clinicians. The *Z* values of both MIB-ANet and 3 E.N.T. clinicians were smaller than zero, which meant that both clinicians and deep

TABLE 2 Performance of MIB-ANet to E.N.T. clinicians.

| | Evaluation metrics | | Time (s) | vs. Ground truth | |
|---|---|---|---|---|---|
| | F1 score | Accuracy | | *Z* | *p* value |
| Senior clinician | **0.89013** | **0.89558** | 4 ~ 8 | −6.962 | 0.000* |
| Intermediate clinician | 0.80555 | 0.80422 | 5 ~ 11 | −8.307 | 0.000* |
| Junior clinician | 0.69867 | 0.72490 | 7 ~ 13 | −5.618 | 0.000* |
| MIB-ANet | 0.76251 | 0.76807 | **0.05** | −1.316 | 0.188 |

We used "bold" to highlight the best performance of the variable.

model tended to make prediction of higher grade. Compared to 3 clinicians, MIB-ANet achieved the smallest absolute *Z* value, which meant that the prediction of MIB-ANet was more objective.

## 3.2. Comparison based on ROC curve and confusion matrices

Figure 3 shows the micro-average ROC curve of MIB-ANet and different grade. True Positive Rate (TPR) as well as False Positive Rate (FPR) of 3 E.N.T. clinicians. For points in ROC curve, the closer to the upper left corner, the better grading performance. From Figure 3A,

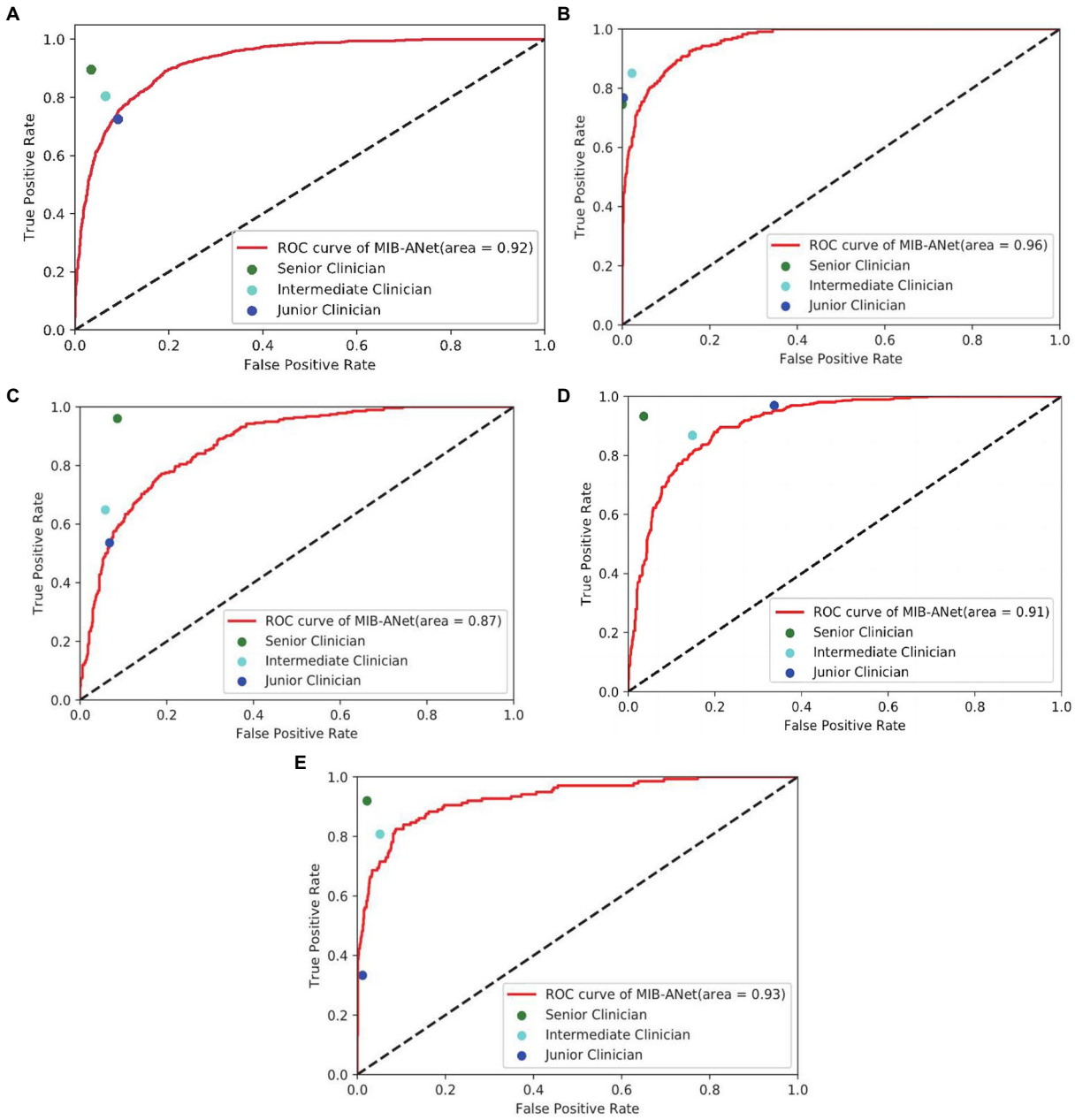


FIGURE 3
**(A–E)** show the overall micro-average ROC curve of MIB-ANet and that for grade 1 to 4 adenoid hypertrophy respectively, compared with TPR and FPR of 3 E.N.T. clinicians.

we can see that the senior clinician (green point) showed the best grading performance. MIB-ANet (red curve) showed performance between intermediate clinician (aqua point) and junior clinician (blue point). From Figures 3B,D, we can see that for grade 1 and grade 3 adenoid images, 3 E.N.T. clinicians showed better performance than MIB-ANet (All of points are located above the curve of MIB-ANet). From Figures 3C,E, we can see that for grade 2 and grade 4 adenoid images, MIB-ANet showed better performance than junior clinician (blue point is located below the red curve), while showed worse performance than senior clinician and intermediate clinician (green point and aqua point are located above the red curve).

Figure 4 shows the confusion matrices of MIB-ANet and human experts. From these matrices, we can calculate that for grade 1 adenoid images, the accuracy of 3 E.N.T. clinicians were roughly the same and higher than that of MIB-ANet. For grades 2, 3, and 4 adenoid images, senior clinician achieved the best accuracy, which were 0.92671, 0.95281, and 0.96988, respectively. MIB-ANet achieved better accuracy (0.86747) than intermediate clinician (0.85743) for grade 3 adenoid images. And for grades 2, 3, and 4 adenoid images, MIB-ANet achieved better accuracy (0.83534/0.86747/0.92771) than junior clinician (0.82229/0.77209/0.91064).

## 3.3. Comparison based on heatmap visualization

Figure 5 shows the heatmaps overlaid on adenoid nasoendoscopic images, which denotes attention map of different neural networks according to weighting of all pixels dictated by CAM. From Figure 5, we can see that, for grade 1 and grade 2, AlexNet, VGG16, ANet, and MIB-ANet tended to focus on whether the adenoid tissue is in contact with torus tubarius; ResNet50 and GoogleNet tended to focus on the adenoid area and whether adenoids were in contact with vomer. For grade 3 and grade 4, VGG16 and ResNet50 tended to focus on whether adenoids were in contact with soft palate. For grade 3, AlexNet, GoogleNet, and ANet tended to focus on the size of the airway (to some extent, the larger the adenoid, the smaller the airway space). For grade 4, AlexNet, GoogleNet, and ANet tended to focus on the adenoid area. In contrast, MIB-ANet can always focus on whether adenoids were in contact with adjacent tissues, which meant that the prediction made by MIB-ANet was based on the contact between adenoids and adjacent tissues, which was the same as how E.N.T. clinician make a decision14. The heatmaps intuitively explain why MIB-ANet has the best performance among all networks.

## 3.4. Performance of different grades

Figure 6 shows the F1 score of different grades for MIB-ANet, senior clinician, intermediate clinician, and junior clinician. From Figure 6, we can see that senior clinician showed the best classification performance among 3 E.N.T. clinicians. For grades 2, 3, and 4 adenoid images, senior clinician achieved 10–30% higher F1 score than intermediate clinician and junior clinician, and for grade 1 adenoid images, senior clinician showed comparable F1 score to intermediate clinician and junior clinician. Compared to 3 E.N.T. clinicians,

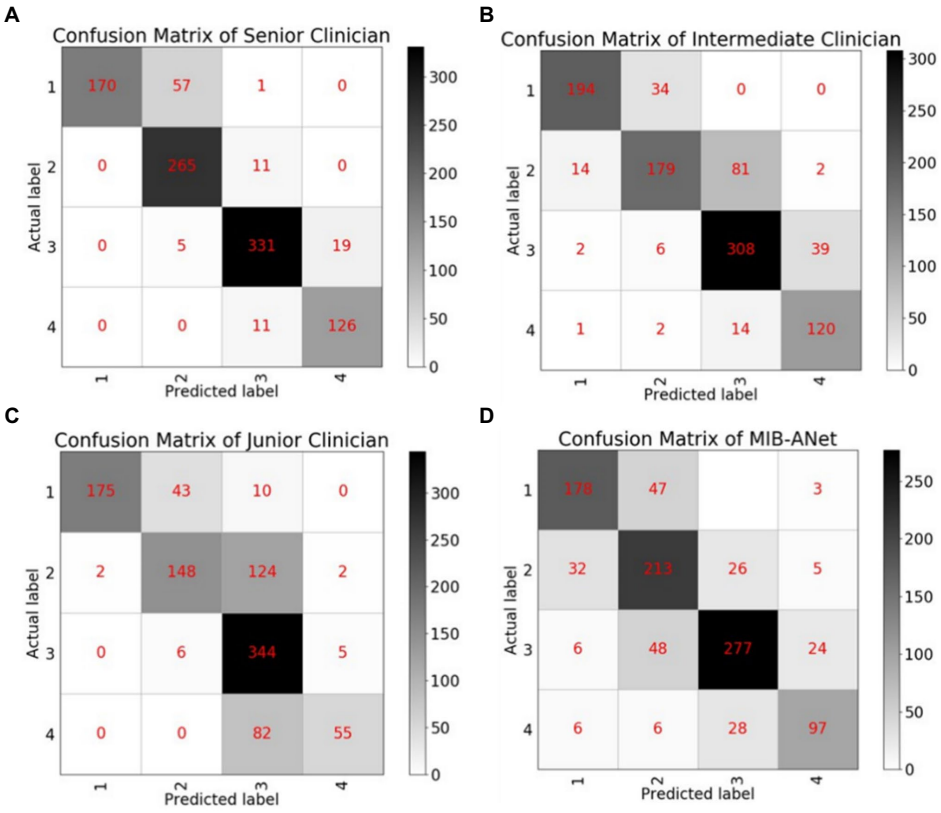


FIGURE 4
The confusion matrices of MIB-ANet and 3 clinicians. **(A–D)** Show the confusion matrix of senior clinician, intermediate clinician, junior clinician, and MIB-ANet, respectively.

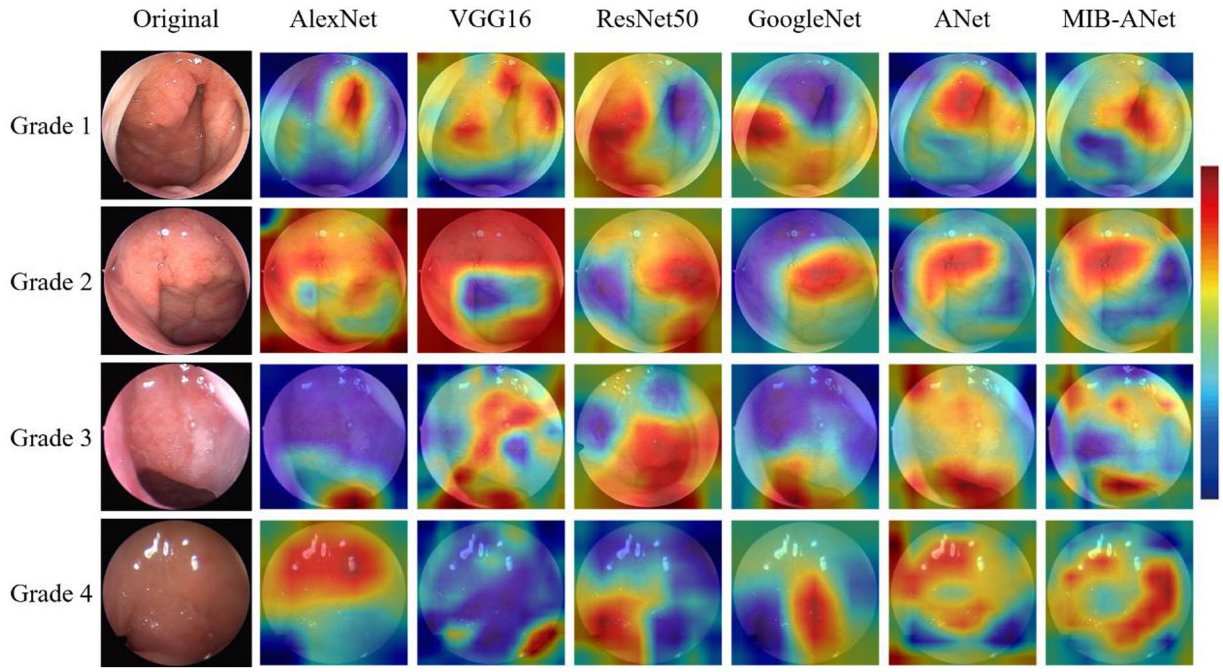


FIGURE 5
The heatmaps of different deep networks for adenoid hypertrophy prediction. The first column shows the original adenoid images. The second, third, fourth, fifth, sixth, and seventh columns show the heatmaps of AlexNet, VGG16, ResNet50, GoogleNet, ANet and MIB-ANet, respectively.

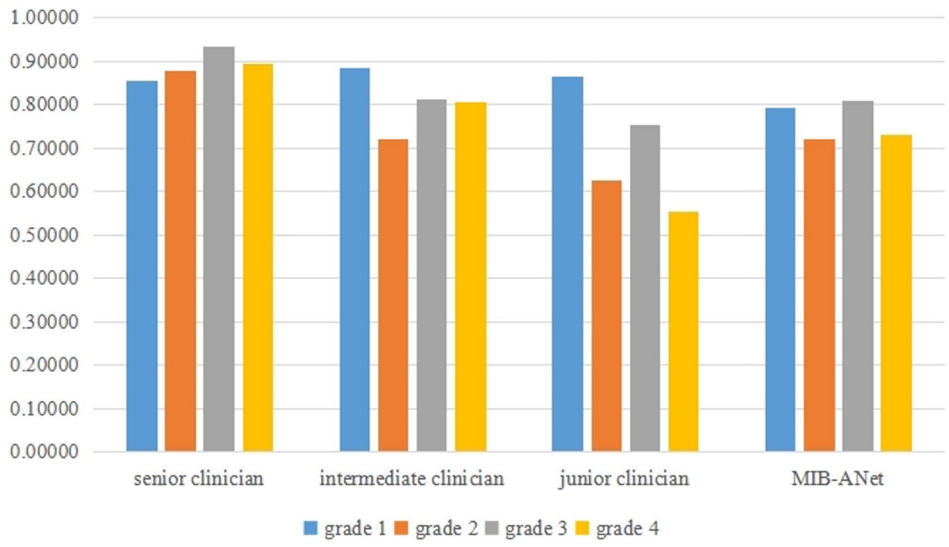


FIGURE 6
The F1 score of different grades for MIB-ANet, senior clinician, intermediate clinician, and junior clinician.

MIB-ANet achieved comparable F1 score to intermediate clinician for grade 2 and 3 adenoid images, which was 9 and 5% higher than junior clinician, respectively. For grade 4 adenoid images, MIB-ANet achieved 7% lower F1 score than intermediate clinician, but 17% higher than junior clinician. For grade 1 adenoid images, MIB-ANet achieved lower F1 score than 3 E.N.T. clinicians, but only 5% lower than senior clinician. Overall, the performance of MIB-ANet was better than junior clinician and close to intermediate clinician.

## 4. Discussion

Clinically, the grading of adenoid hypertrophy is important for surgical indication. There are several medical examinations to evaluate adenoid hypertrophy, such as lateral cranial X-ray, nasoendoscopy, cone-beam computed tomography (CBCT) (20), MRI, and 3D printed model (21). Nasal endoscopy is a radiation-free, safe, and convenient operation, which is routinely used for adenoid hypertrophy grading examination. In

this work, we built SYSU-SZU-EA nasoendoscopic image dataset and proposed a novel efficient deep neural network, MIB-ANet, for adenoid hypertrophy classification. To the best of our knowledge, this is the first deep learning research to address the grading of endoscopic images of adenoid hypertrophy. The experimental results showed that our network achieved better classification performance than four classical CNNs, i.e., AlexNet, VGG16, ResNet50, and GoogleNet. When compared to three E.N.T. clinicians, MIB-ANet achieved much higher (80× faster) diagnosing speed, with a grading performance better than the junior E.N.T. clinician.

In recent years, many deep learning methods, especially convolutional neural networks (CNNs), have been applied in the medical image domain. Girdler et al. (22) categorized 297 nasoendoscopic images by using the CNN model of ResNet-152 for automated detection and classification of nasal polyps and inverted papillomas. Overall accuracy of $0.742 \pm 0.058$ was achieved. Yang et al. (23) developed a cascaded under-sampling ensemble learning method (CUEL) to prevent and diagnose clinical rhinitis, which achieved 90.71% average accuracy on 2,231 clinical rhinitis instances. The current deep learning network is mostly used for the diagnosis of diseases. Even in the field of capsule endoscopic images with a large number of deep learning researches, little work is conducted to classify the degree of disease. In this study, we focused on the clinical requirement of adenoid hypertrophy grading, rather than disease diagnosis. At the same time, more detailed assessment, such as the nasal mucosal inflammation state, the size degree of polyps, and grading of adenoid hypertrophy, can lead to the creation of an automatic nasal endoscopy reporting system, which can reduce the burden of E.N.T clinicians and improve efficiency and accuracy of reading caused by visual fatigue.

Usually, different network models are suitable for different data sets, and the design of network structure should be based on the characteristics of data sets. Medical data sets are different from data sets collected in daily life, such as ImageNet, and contain much smaller number of images. However, the classical deep learning model has a large number of parameters, which is easy to over fit when these models are trained using small data sets in the medical field. Therefore, in order to avoid the over fitting problem in the classification of adenoid hypertrophy, we tried to reduce the amount of model parameters when designing the network structure. In addition, compared with natural images, nasoendoscopic images are characterized by more concentrated color distribution (overall red color), more abundant texture features (tissue blood vessels, dense tissue distribution), and large differences in size and shape among different types of adenoid. The classical deep learning model cannot well extract both low-level and high-level adenoid hypertrophy features. In order to solve this problem, we proposed ANet to extract high-level adenoid hypertrophy features using dilated convolutions. Based on ANet, we proposed MIB-ANet with convolution kernels of different sizes to extract both low-level and high-level adenoid hypertrophy features. The performance of ANet and MIB-ANet was better than four classic CNNs. In addition, the experimental results showed that MIB-ANet can achieve a grading performance better than the junior E.N.T. clinician with much higher diagnosing speed.

However, some limitations in our study should be mentioned. Firstly, we annotate the ground truth label of testing set according to the evaluation results of 3 E.N.T. experts with the principle of majority voting, which might still generate some incorrect labels. Further manual data cleaning and more reasonable annotation process, e.g., intraoperative evaluation of adenoid size, are required (24). Secondly, when MIB-ANet was used to grade adenoid hypertrophy, the model tended to fit the size of adenoid. When the image of adenoid collected by endoscopic technician is not standard (for example, the endoscope is close to the adenoid when collecting the adenoid image), MIB-ANet is easy to predict a higher grade. Therefore, the E.N.T clinicians are suggested to draw the boundary of the designated anatomical structure or attention area by using some software like imageScope, which can further improve the performance. At the same time, enlarging the database and building up a multicenter data platform are also helpful to improve the model. Finally, in this study, we only focused on adenoid hypertrophy grading on nasoendoscopic images. In the future, we can further add labels of other nasopharyngeal diseases, such as nasopharyngeal carcinoma and nasopharyngitis, and develop a comprehensive classification model for nasal disease diagnosis.

## Data availability statement

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

## Ethics statement

The studies involving human participants were reviewed and approved by Ethical Review Board of the Seventh affiliated Hospital of Sun Yat-sen University. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the minor(s)' legal guardian/next of kin for the publication of any potentially identifiable images or data included in this article.

## Author contributions

MB was involved in conception, design, data collection, data analysis, and interpretation, and drafted manuscripts. SZ was involved in conception, network design, data analysis, and interpretation, and drafted manuscripts. YF, XL, and LS was involved in conception and design, and made critical revisions to the manuscript. HL and XF participated in data collection and adenoid hypertrophy grading. All authors have given final recognition and agreed to be responsible for all aspects of the work.

## Funding

This work was supported by Sanming Project of Medicine in Shenzhen under Grant No. SZSM202111005; Shenzhen Fundamental Research Program (Grant No. JCYJ20190809143601759); the National Natural Science Foundation of China under Grant 82261138629; Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515010688 and Shenzhen Municipal Science and Technology Innovation Council under Grant JCYJ20220531101412030.

## Acknowledgments

The authors would like to thank Yueqi Sun and Kanghua Wang for revisions to the 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

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/fmed.2023.1142261/full#supplementary-material

## References


1. Pereira L, Monyror J, Almeida FT, Almeida FR, Guerra E, Flores-Mir C, et al. Prevalence of adenoid hypertrophy: a systematic review and meta-analysis. *Sleep Med Rev*. (2018) 38:101–12. doi: 10.1016/j.smrv.2017.06.001

2. Peltomäki T. The effect of mode of breathing on craniofacial growth--revisited. *Eur J Orthod*. (2007) 29:426–9. doi: 10.1093/ejo/cjm055

3. Clemens J, McMurray JS, Willging JP. Electrocautery versus curette adenoidectomy: comparison of postoperative results. *Int J Pediatr Otorhinolaryngol*. (1998) 43:115–22. doi: 10.1016/s0165-5876(97)00159-6

4. Cassano P, Gelardi M, Cassano M, Fiorella ML, Fiorella R. Adenoid tissue rhinopharyngeal obstruction grading based on fiberendoscopic findings: a novel approach to therapeutic management. *Int J Pediatr Otorhinolaryngol*. (2003) 67:1303–9. doi: 10.1016/j.ijporl.2003.07.018

5. Parikh SR, Coronel M, Lee JJ, Brown SM. Validation of a new grading system for endoscopic examination of adenoid hypertrophy. *Otolaryngol Head Neck Surg*. (2006) 135:684–7. doi: 10.1016/j.otohns.2006.05.003

6. Varghese AM, Naina P, Cheng AT, Asif SK, Kurien M. ACE grading-a proposed endoscopic grading system for adenoids and its clinical correlation. *Int J Pediatr Otorhinolaryngol*. (2016) 83:155–9. doi: 10.1016/j.ijporl.2016.02.002

7. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. (2017) 42:60–88. doi: 10.1016/j.media.2017.07.005

8. He Y, Carass A, Liu Y, Jedynak BM, Solomon SD, Saidha S, et al. *Fully Convolutional Boundary Regression for Retina OCT Segmentation*, Springer Internation. (2019): 120–128.

9. Li X, Shen L, Shen M, Tan F, Qiu CS. Deep learning based early stage diabetic retinopathy detection using optical coherence tomography. *Neurocomputing*. (2019) 369:134–44. doi: 10.1016/j.neucom.2019.08.079

10. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*. (2016). doi: 10.48550/arXiv.1606.05718

11. Li X, Shen L, Xie X, Huang S, Xie Z, Hong X, et al. Multi-resolution convolutional networks for chest X-ray radiograph based lung nodule detection. *Artif Intell Med*. (2020) 103:101744. doi: 10.1016/j.artmed.2019.101744

12. Li Y, Shen L. Skin lesion analysis towards melanoma detection using deep learning network. *Sensors (Basel)*. (2018) 18:556. doi: 10.3390/s18020556

13. Shen Y, Li X, Liang X, Xu H, Li C, Yu Y, et al. A deep-learning-based approach for adenoid hypertrophy diagnosis. *Med Phys*. (2020) 47:2171–81. doi: 10.1002/mp.14063

14. Liu JL, Li SH, Cai YM, Lan DP, Lu YF, Liao W, et al. Automated radiographic evaluation of adenoid hypertrophy based on VGG-lite. *J Dent Res*. (2021) 100:1337–43. doi: 10.1177/00220345211009474

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

16. Simonyan K, Zisserman A. Very deep convolutional networks forLarge-Scale image recognition. Computer Science (2014). doi: 10.48550/arXiv.1409.1556

17. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2016): 770–8. doi: 10.48550/arXiv.1512.03385

18. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going Deeper with Convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2015): 1–9. doi: 10.1109/CVPR.2015.7298594

19. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2016): 2921–9. doi: 10.1109/CVPR.2016.319

20. Thereza-Bussolaro C, Lagravère M, Pacheco-Pereira C, Flores-Mir C. Development, validation and application of a 3D printed model depicting adenoid hypertrophy in comparison to a Nasoendoscopy. *Head Face Med*. (2020) 16:5. doi: 10.1186/s13005-020-00216-4

21. Girdler B, Moon H, Bae MR, Ryu SS, Bae J, Yu MS. Feasibility of a deep learning-based algorithm for automated detection and classification of nasal polyps and inverted papillomas on nasal endoscopic images. *Int Forum Allergy Rhinol*. (2021) 11:1637–46. doi: 10.1002/alr.22854

22. Yang J, Zhang M, Yu S. A novel rhinitis prediction method for class imbalance. *Biomed Signal Proces*. (2021) 69:102821. doi: 10.1016/j.bspc.2021.102821

23. Soldatova L, Otero HJ, Saul DA, Barrera CA, Elden L. Lateral neck radiography in preoperative evaluation of adenoid hypertrophy. *Ann Otol Rhinol Laryngol*. (2020) 129:482–8. doi: 10.1177/0003489419895035

24. Pachêco-Pereira C, Alsufyani NA, Major MP, Flores-Mir C. Accuracy and reliability of oral maxillofacial radiologists when evaluating cone-beam computed tomography imaging for adenoid hypertrophy screening: A comparison with nasopharyngoscopy. *Oral Surg Oral Med Oral Pathol Oral Radiol*. (2016) 121:e168–74. doi: 10.1016/j.oooo.2016.03.010

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 23 August 2023
DOI 10.3389/fmed.2023.1113030



OPEN ACCESS

EDITED BY
Jinming Duan,
University of Birmingham, United Kingdom

REVIEWED BY
Yuexiang Li,
Tencent Jarvis Lab, China
Fei He,
Coventry University, United Kingdom

*CORRESPONDENCE
Yalin Zheng
✉ Yalin.Zheng@liverpool.ac.uk

RECEIVED 30 November 2022
ACCEPTED 08 August 2023
PUBLISHED 23 August 2023

CITATION
Bridge J, Meng Y, Zhu W, Fitzmaurice T, McCann C, Addison C, Wang M, Merritt C, Franks S, Mackey M, Messenger S, Sun R, Zhao Y and Zheng Y (2023) Development and external validation of a mixed-effects deep learning model to diagnose COVID-19 from CT imaging.
*Front. Med.* 10:1113030.
doi: 10.3389/fmed.2023.1113030



# Development and external validation of a mixed-effects deep learning model to diagnose COVID-19 from CT imaging

Joshua Bridge[1], Yanda Meng[1], Wenyue Zhu[1], Thomas Fitzmaurice[1,2], Caroline McCann[3], Cliff Addison[4], Manhui Wang[4], Cristin Merritt[5], Stu Franks[5], Maria Mackey[6], Steve Messenger[6], Renrong Sun[7], Yitian Zhao[8] and Yalin Zheng[1]*

[1]Institute of Life Course and Medical Sciences, University of Liverpool, Liverpool, United Kingdom, [2]Department of Respiratory Medicine, Liverpool Heart and Chest Hospital NHS Foundation Trust, Liverpool, United Kingdom, [3]Department of Radiology, Liverpool Heart and Chest Hospital NHS Foundation Trust, Liverpool, United Kingdom, [4]Advanced Research Computing, University of Liverpool, Liverpool, United Kingdom, [5]Alces Flight Limited, Bicester, United Kingdom, [6]Amazon Web Services, London, United Kingdom, [7]Department of Radiology, Hubei Provincial Hospital of Integrated Chinese and Western Medicine, Hubei University of Chinese Medicine, Wuhan, China, [8]Cixi Institute of Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China

**Background:** The automatic analysis of medical images has the potential improve diagnostic accuracy while reducing the strain on clinicians. Current methods analyzing 3D-like imaging data, such as computerized tomography imaging, often treat each image slice as individual slices. This may not be able to appropriately model the relationship between slices.

**Methods:** Our proposed method utilizes a mixed-effects model within the deep learning framework to model the relationship between slices. We externally validated this method on a data set taken from a different country and compared our results against other proposed methods. We evaluated the discrimination, calibration, and clinical usefulness of our model using a range of measures. Finally, we carried out a sensitivity analysis to demonstrate our methods robustness to noise and missing data.

**Results:** In the external geographic validation set our model showed excellent performance with an AUROC of 0.930 (95%CI: 0.914, 0.947), with a sensitivity and specificity, PPV, and NPV of 0.778 (0.720, 0.828), 0.882 (0.853, 0.908), 0.744 (0.686, 0.797), and 0.900 (0.872, 0.924) at the 0.5 probability cut-off point. Our model also maintained good calibration in the external validation dataset, while other methods showed poor calibration.

**Conclusion:** Deep learning can reduce stress on healthcare systems by automatically screening CT imaging for COVID-19. Our method showed improved generalizability in external validation compared to previous published methods. However, deep learning models must be robustly assessed using various performance measures and externally validated in each setting. In addition, best practice guidelines for developing and reporting predictive models are vital for the safe adoption of such models.

KEYWORDS

**CT, COVID-19, deep learning, diagnosis, imaging**

# 1. Background

Coronavirus disease 2019 (COVID-19) is an infectious respiratory disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Virus clinical presentation ranges from mild cold-like symptoms to severe viral pneumonia, which can be fatal (1). While some countries have achieved relative control through lockdowns, future outbreaks and new strains are expected to continue, with many experts believing the virus is here to stay (2). Detection and isolation is the most effective way to prevent further spread of the virus. Even with effective vaccines becoming widely available, with the threat of continued waves and new potentially vaccine-resistant variants, it is vital to further develop diagnostic tools for COVID-19. These tools will likely also apply to future outbreaks of other similar diseases as well as common diseases such as pneumonia.

The diagnosis of COVID-19 is usually determined by Reverse Transcription Polymerase Chain Reaction (RT-PCR), but this is far from being a gold standard. A negative test does not necessarily indicate a negative diagnosis, with one recent review finding that RT-PCR has a real-world sensitivity of around 70% and a specificity of 95% (3). Furthermore, an individual patient data systematic review (4) found that RT-PCR often fails to detect COVID-19, and early sampling is key to reducing false negatives. Therefore, these tests are often more helpful to rule in COVID-19 rather than ruling out. If a patient presents with symptoms of COVID-19, but an RT-PCR test is negative, then further tests are often required (1). Consecutive negative tests with at least a one-day gap are recommended; however, this still does not guarantee that the patient is negative for COVID-19 (5). Computed tomography (CT) can play a significant role in diagnosing COVID-19 (6). Given the excessive number of COVID-19 cases worldwide and the strain on resources expected, automated diagnosis might reduce the burden on reporting radiologists.

CT images are made up of many slices, creating a three dimensional (3D)-like structure. Previous approaches, such as those used by Li et al. (7) and Bai et al. (8) treat the image as separate slices and use a pooling layer to concatenate the slices. An alternative approach assumes the slices form a 3D structure and use a 3D CNN, such as that proposed in CoviNet (9). A fundamental limitation of these methods is the need for the same number of slices as their inputs, but the number of slices often varies between different CT volumes. Instead, we propose using a novel mixed-effects layer to consider the relationship between slices in each scan. Mixed-effects models are commonly used in traditional statistics (10, 11), but we believe this is the first time that mixed-effects models have been utilized in such a way. It has been observed that some lobes of the lung are more often affected by COVID-19 than others (12, 13) with lower lobe distribution being a prominent feature of COVID-19 (14), the fixed-effects take this into account by considering where each slice is located within the scan.

Deep learning has shown great potential in the automatic classification of disease, often achieving expert-level performance. Such models could screen and monitor COVID-19 by automatically analyzing routinely collected CT images. As observed by Wynants et al. (15) and Roberts et al. (16) many models are already developed to diagnose COVID-19, which often obtain excellent discriminative performance; however, very few of these models, if any, are suitable for clinical use, mainly due to a lack of robust analysis and reporting. These models often suffer from common pitfalls, making them unsuitable for broader adoption. Roberts et al. (16) identified three common areas in which models often fail these are: (1) a lack of adequately documented methods for reproducibility, (2) failure to follow established guidelines and best practices for the development of deep learning models, and (3) an absence of external validation displaying the model's applicability to a broader range of data outside of the study sample. Failure to address these pitfalls leads to profoundly flawed and biased models, making them unsuitable for deployment.

In this work, we aim to address the problems associated with previous models by following guidelines for the reporting (17, 18) and development (19) of prediction models to ensure that we have rigorous documentation allowing the methods developed here to be replicated. In addition, we will make code and the trained model publicly available at: github.com/JTBridge/ME-COVID19 to promote reproducible research and facilitate adoption. Finally, we use a second dataset from a country other than the development dataset to externally validate the model and report a range of performance measures evaluating the model's discrimination, calibration, and clinical usefulness.

Hence, our main aim is to develop a mixed-effects deep learning model to accurately classify images as healthy or COVID-19, following best practice guidelines. Our secondary aim is to show how deep learning predictive algorithms can satisfy current best practice guidelines to create reproducible and less biased models.

# 2. Methods

Our proposed method consists of a feature extractor and a two-stage generalized linear mixed-effects model (GLMM) (20), with all parameters estimated within the deep learning framework using backpropagation. First, a series of CT slices forming a CT volume is input to the model. In our work, we use 20 slices. Next, a convolutional neural network (CNN) extracts relevant features from the model and creates a feature vector for each CT slice. Then, a mixed-effects layer concatenates the feature vectors into a single vector. Finally, a fully connected layer followed by a sigmoid activation gives a probability of COVID-19 for the whole volume. The mixed effects and fully connected layer with sigmoid activation are analogous to a linear GLMM in traditional statistics. The overall framework is shown in Figure 1.

## 2.1. Feature extractor

For the feature extractor, we use a CNN. In this work, we chose InceptionV3 (21) as it is relatively efficient and commonly used. InceptionV3 outputs a feature vector of length 2048. To reduce the time needed to reach convergence, we pretrained the CNN on ImageNet (22). A CNN is used for each slice, with shared weights between CNNs; this reduces the memory footprint of the model. Following the CNN, we used a global average pooling layer to reduce each image to a feature vector for each slice. We then added a dropout of 0.6 to improve generalizability to unseen images. We form the feature vectors into a matrix of shape $20 \times 2{,}048$. Although we used InceptionV3 (21) here, other networks would also work and may provide better performance on other similar tasks. We then need to concatenate these feature vectors into a single feature vector for the

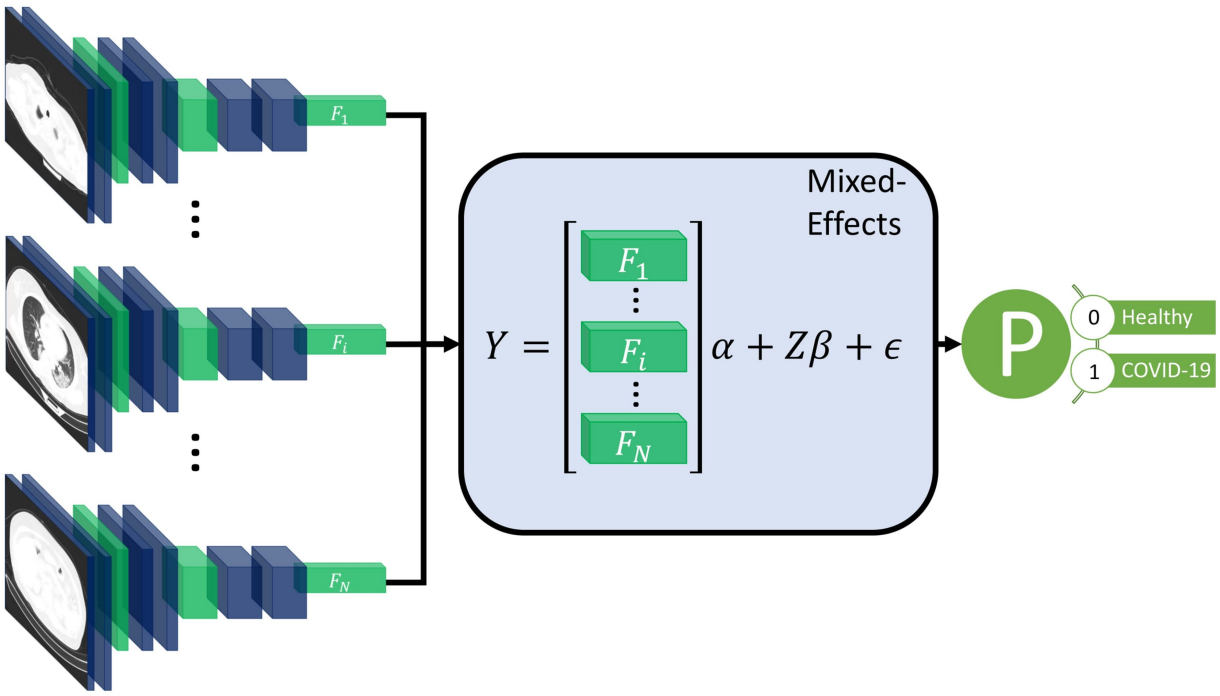


FIGURE 1
Diagram of the overall framework. Twenty slices are chosen from a CT volume. Each slice is fed into a CNN with shared weights, which outputs a feature vector of length 2048 for each image. The feature vectors form a 20-by-2048 fixed effects matrix, X, for the GMM model with a random-effects matrix, Z, consisting of an identity matrix. A mixed-effects model is used to model the relationship between slices. Finally, a fully connected layer and sigmoid activation return a probability of the diagnosis.

whole volume; normally, pooling is used, in our work we propose using a mixed-effects models.

## 2.2. Mixed-effects network

We propose utilizing a novel mixed-effects layer to model the relationship between slices. Mixed-effects models are a statistical model consisting of a fixed-effects part and a random-effects part. The fixed-effects part models the relationship within the CT slice; the random effects can model the spatial correlation between CT slices within the same image (11). For volumetric data, the number of slices may differ significantly due to various factors such as imaging protocol and the size of the patient. Some CT volumes in the dataset may have fewer images than the model is designed to use, which leads to missing data. The number of slices depends upon many factors including the scanning protocol and the size of the patient. Mixed-effects models can deal with missing data provided the data are missing at random (23). It is likely that the data here is missing at random, although not completely at random. The mixed-effects model is given by

$$Y_i = X_i\alpha + Z_i\beta + e_i$$

where $Y_i, X_i, Z_i, e_i$ are vectors of outcomes, fixed effects design matrix of shape $slices \times features$, random effects design matrix of shape $slices \times slices$, and vector of error unknown random errors of the $i$th patient of shape $slices$, respectively, and $\alpha, \beta$ are fixed and random effects parameters, both of length $features$ and $slices$, respectively. In our work, we have 20 slices and 2048 features and use the identity matric for the random effects design matrix. The values in the random effects design matric can be changed to reflect a non-uniform distance between slices. We assume that the random effects $\beta$ are normally distributed with mean 0 and variance $G$

$$\beta \sim N(0, G).$$

We also assume independence between the random effects and the error term.

The fixed effects design matrix, $X$, is made up of the feature vectors output from the feature extraction network. For the random effects design matrix, $Z$, we use an identity matrix with the same size as the number of slices; in our experiments, this is 20. The design matrix is then given by

$$Z_{20\times 20} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ \vdots & & \ddots & \vdots & \\ 0 & 0 & & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

This matrix easily generalizes to any number of slices. If the distance between slices is not uniform, the values can be altered accordingly. We assumed no particular correlation matrix. We included the fixed and random intercept in the model. All parameters for the mixed-effects layer were initialized using the Gaussian distribution with mean 0 and standard deviation 0.05.

A type of mixed-effects modeling has previously been combined with deep learning for gaze estimation (24). However, their mixed-effects method is very different from our proposed method; they used the same design matrix for fixed and random effects. In addition, they also estimated random-effects parameters with an expectation–maximization algorithm, which was separate from the fixed effects estimation, which used deep learning. In our work, we utilize a spatial design matrix to model the spatial relationship between slices and estimate parameters within the deep learning framework using backpropagation without the need for multiple stages.

## 2.3. Loss function

As the parameters in the model are all estimated using backpropagation, we must ensure that the assumption of normally distributed random effects parameters with mean zero is valid. We achieve this by introducing a loss function for the random effects parameters, which enforces a mean, skewness, and excess kurtosis of 0. We measure skewness using the adjusted Fisher–Pearson standardized moment coefficient

$$Skew(\beta) = \frac{\sqrt{n(n-1)}}{n-2} \frac{E\left[\left(\beta - \bar{\beta}\right)^3\right]}{\left(E\left[\left(\beta - \bar{\beta}\right)^2\right]\right)^{3/2}}$$

and the excess kurtosis using

$$Kurt(\beta) - 3 = \frac{1}{n^2} \sum_{i=1}^{n} \left( \frac{E\left[\left(\beta - \bar{\beta}\right)^4\right]}{\left(E\left[\left(\beta - \bar{\beta}\right)^2\right]\right)^2} - 3 \right),$$

where $n$ is the length of $\beta$, $\bar{\beta}$ is the mean of $\beta$ and $E[\cdot]$ is the expectation function. The Gaussian distribution has a kurtosis of 3; therefore, the excess kurtosis is given by the kurtosis minus 3. This formula for this random-effects parameters loss function which we aim to minimize, is then given by

$$L_{random} = |E(\beta)| + |Skew(\beta)| + |Kurt(\beta) - 3|.$$

For the classification, we use the Brier Score (25) as the loss function, which is given by

$$L_{Brier} = \frac{1}{N} \sum_{i=1}^{N} (p_i - o_i)^2$$

where $N$ is the total number of samples, $p_i$ is the predicted probability of sample $i$ and $o_i$ is the observed outcome of sample $i$. The Brier score is the same as the mean squared error of the predicted probability.

We chose to use the Brier Score over the more commonly used binary cross-entropy because it can be decomposed into two components: refinement and calibration. Calibration is often overlooked in deep learning models but is vital to assess the safety of any prediction model. The refinement component combines the model's resolution and uncertainty and measures the model's discrimination. The calibration component can be used as a measure of the model calibration. Therefore, the Brier Score can be used to optimize both the discrimination and calibration of the model. The overall loss function is given by

$$L = L_{Brier} + L_{random}.$$

A scaling factor could be introduced to weight one part of the loss function as more important than the other; however, we give both parts of the loss function equal weighting in our work.

We also transformed the labels as suggested by Platt (26) to reduce overfitting. The negative and positive labels become

$$o_- = \frac{1}{N_- + 2}$$

and

$$o_+ = \frac{N_+ + 1}{N_+ + 2}$$

respectively, where $N_-$ and $N_+$ are the total number of negative and positive cases in the training set. This is similar to label smoothing as commonly used in deep learning, but the new targets are chosen by applying Bayes' Rule to the out-of-sample data to prevent overfitting.

## 2.4. Classification layer

The output of the mixed-effects layer is a single vector, which is the same length as the number of slices used. For example, in our work, we had a vector of length 20. Furthermore, we used a fully connected layer with sigmoid activation to obtain a probability of the scan showing COVID-19; the sigmoid activation is analogous to the logistic link function in traditional statistics. Finally, we added an L1 regularization term of 0.1 and an L2 regularization term of 0.01 to the kernel to reduce overfitting.

## 2.5. Model performance

Many deep learning models focus on assessing discriminative performance only, using measures such as the area under the receiver operating characteristic curve (AUROC), sensitivity, and specificity. To better understand the model performance and impact, we report performance measures in three broad areas: discrimination, calibration, and clinical usefulness (27). Discrimination assesses how well a model can discriminate between healthy and COVID-19 positive patients. Models with excellent discriminative performance can still produce unreliable results, with vastly overestimated probabilities regardless of the true diagnosis (28). Model calibration is often overlooked and rarely reported in deep learning, if at all; however, poorly calibrated models can be misleading and lead to dangerous clinical decisions (28). Calibration can be assessed using

four levels, with each level indicating better calibration than the last (29). The fourth and most stringent level (strong calibration) requires the correct model to be known, which in turn requires predictors to be non-continuous, and an infinite amount of data to be used and is therefore considered utopic. We consider the third level (moderate calibration) using calibration curves. Moderate calibration will ensure that the model is at least not clinically harmful. Finally, measures of clinical usefulness assess the clinical consequences of the decision and acknowledge that a false positive may be more or less severe than a false negative.

Firstly, the discriminative performance is assessed using AUROC using the pROC package in R (30), with confidence intervals constructed using DeLong's (31) method. For sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), we use the epiR (32) package in R (30); with 95% confidence intervals constructed using Jeffrey's prior (33). We report performance at a range of probability thresholds to demonstrate how the thresholds can be adjusted to reduce false positives or false negatives depending on the setting (34). Secondly, we assess model calibration using calibration curves created using the CalibrationCurves package (29), which is based on the rms (35) package. Finally, we assess the clinical usefulness of the model using decision curve analysis (36). Net benefits are given at various thresholds, and models which reach zero net benefit at higher thresholds are considered more clinically useful. Two brief sensitivity analyses are performed, one assessing the model's ability to deal with missing data and the other assessing its ability to deal with noise. To improve the model's interpretability and reduce the black-box nature, we produce saliency maps (37) that show which areas of the image are helpful to the model in the prediction. We also check the assumption of normally distributed random-effects parameters.

## 2.6. Comparison models

To assess the added benefit of using our mixed-effects method, we compare against networks that use alternative methods. Both COVNet (7) and a method proposed by Bai et al. (8) propose deep learning models that consider the slices separately before concatenating the features using max pooling. COVNet uses a ResNet50 (38) CNN to extract features and pooling layers to concatenate the features before a fully connected classification layer. The model proposed by Bai et al. uses EfficientNetB4 (39) to extract features followed by a series of full-connected layers with batch normalization and dropout; average pooling is then used to concatenate the feature vectors before classification. While max pooling is simple and computationally efficient, it cannot deal with pose variance and does not model the relationship between slices.

An alternative method to pooling is treating the scans as 3D, such as in CoviNet (40). CoviNet takes the whole scan and uses a 16-layer 3D CNN followed by pooling and fully connected layers. We implemented these models as described in their respective papers.

In all comparison experiments, we kept hyperparameters, such as learning rate, learning rate decay, and data augmentation, the same to ensure the comparisons were fair. For COVNet (7) and the model proposed by Bai et al. (8) we pretrained the CNNs on ImageNet as they also did; however, no pretrained models were available for CoviNet. For the loss function, we also used the Brier score (25).

## 2.7. Computing

Models were developed using an Amazon Web Services p3.8xlarge node with four Tesla V100 16GiB GPUs and 244GiB available memory. Model inference was performed on a local Linux machine running Ubuntu 18.04, with a Titan X 12GiB GPU and 32GiB available memory. Model development and inference were performed using Tensorflow 2.4 (41, 42), and R 4.0.5 (30) was used to produce evaluation metrics (43, 44) and graphs (35, 45). We used mixed precision to reduce the computational cost, which uses 16-bit floating-point precision in all layers, except for the mixed-effects and classification layers, where 32-bit floating-point precision is used.

We used the Adam optimizer (46) with an initial learning rate of 1e-4; if the internal validation loss did not improve for three epochs, we reduced the learning rate to 20%. In addition, we assumed convergence and stopped training if the loss did not improve for 10 epochs to reduce the time spent training and the energy used.

## 2.8. Data

There is currently no established method for estimating the sample size estimate in deep learning. We propose treating the final fully connected classification layer as the model and treating previous layers as feature extraction. We can then use the number of parameters in the final layer to estimate the required sample size. Using the "pmsampsize" package (47) in R, we estimate the required minimum sample size in the development set. We use a conservative expected C-statistic of 0.8, with 21 parameters and an estimated disease prevalence of 80% based on datasets used in other studies. This gives a minimum required sample size of 923 patients in the training set. For model validation, around 200 patients with the disease and 200 patients without the disease are estimated to be needed to assess calibration (29).

All data used here is retrospectively collected and contains hospital patients with CT scans performed during the COVID-19 pandemic. The diagnosis was determined by examining radiological features of the CT scan for signs of COVID-19, such as ground-glass opacities. For model development, we use the MosMed dataset (48), which consists of a total of 1,110 CT scans displaying either healthy or COVID-19 infected lungs. The scans were performed in Moscow hospitals between March 1, 2020, and April 25, 2020. We split the dataset into two sets for training and internal validation on the patient level. The training set is used to train the model, and the internal validation set is used to select the best model based on the loss at each epoch; this helps prevent overfitting on the training set. In addition, we obtained images from a publicly available dataset published by Zhang et al. (49) consisting of CT images from a consortium of Chinese hospitals.

Overall, this allows us to perform external geographical validation in another country and to better evaluate the developed model. In addition, we will be able to assess how well a deep learning model generalizes to other populations. A summary of all the datasets used

is shown in Table 1. We have 923 patients in the training set and at least 200 patients in each class for the external validation set.

## 2.9. Patient and public involvement

Patients or the public were not involved in the design, conduct, reporting, or dissemination of our research.

## 2.10. Data pre-processing and augmentation

The MosMed dataset was converted from Dicom image format into PNG, normalized to have a mean of 120 and a variance of 95. Images were ordered from the top of the lungs to the bottom. During training, we applied random online data augmentation to the images. This alters the image slightly and gives the effect of increasing the training dataset size, although this is not as good as expanding the training dataset with more samples. First, we adjusted the brightness and contrast between 80 and 120%. We then rotated the image plus or minus 5 degrees and cropped the image up to 20% on each side. Finally, we flipped the image horizontally and vertically with a probability of 50% each. All random values were chosen using the uniform distribution except for the flips, which were chosen using a random bit. Example images are shown in Figure 2A.

The dataset taken from Zhang et al. (49) required a large amount of sorting to be made suitable for use. Some of the scans were pre-segmented and only showed the lung areas, while others showed the whole CT scan. We removed any pre-segmented images. Identifying information on some images had to be cropped to reduce bias in the algorithm. In addition, many of the scans were duplicates but were not labeled as such, and many scans were incomplete, only showing a few lung slices or not showing any lung tissue at all. We only used complete scans with one scan per patient. Finally, some scans needed to be ordered top to bottom. Using the bilinear sampling algorithm, all images were resized to 256 by 256 pixels, and image values were divided by 255 to normalize between 0 and 1. Example images are shown in Figure 2B.

The MosMed dataset has a median of 41 slices, a minimum of 31 slices and a maximum of 72 slices. The Zhang et al. dataset has much greater variability in scan size with a median of 61 slices, a minimum of 19 slices, and a maximum of 415 slices. We present histograms showing the number of slices per scan in Figure 3. We require a fixed number of slices as input, and we chose to use 20 slices. For all scans, we included the first and last images. If scans had more than 20 slices, we sampled uniformly to select 20. Only one scan in the Zhang et al. dataset had less than 20 slices; a blank slice replaced this slice; the mixed-effects model can account for missing data.

While removing slices may waste some information available to us, using the full 415 slices that some images have would be impractical due to the large memory footprint. An alternative to removing slices would be to reduce the resolution of each slice; however, this again would waste information. Choosing to use 20 slices of each CT image is a compromise between the amount of information used and the practicality of processing the CT scans.

TABLE 1 Summary of the datasets used.

| Dataset | Location | Use | Healthy/ COVID19 |
|---|---|---|---|
| MosMed training | Moscow, Russia | Training | 169/856 |
| MosMed validation | Moscow, Russia | Internal validation | 85/285 |
| Zhang et al. (48) | China | External validation | 243/553 |

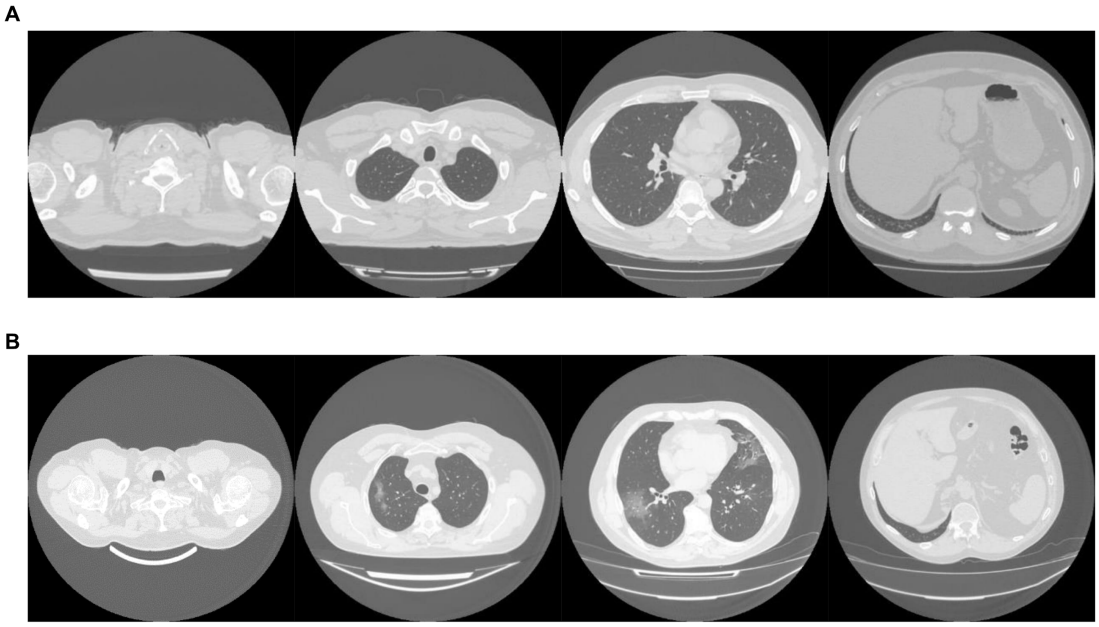


FIGURE 2
Example images showing **(A)** healthy and **(B)** COVID-19 lungs taken from the Mosmed dataset.

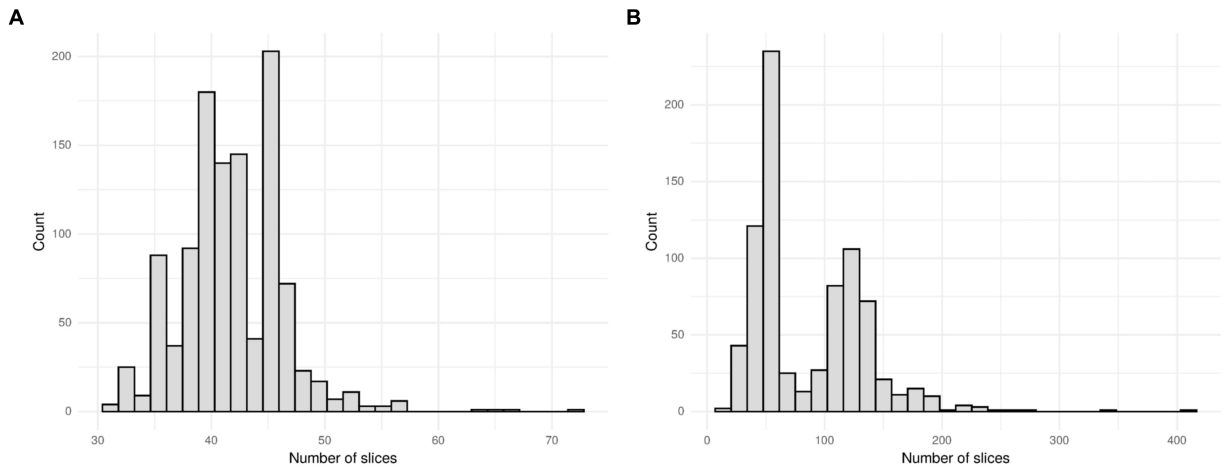


FIGURE 3
Histogram showing the number of slices per scan for **(A)** the MosMed dataset and **(B)** the Zhang et al. dataset. The MosMed dataset has much fewer slices on average with a much smaller spread.

## 3. Results

On the internal validation dataset, the proposed model attained an AUROC of 0.936 (95%CI: 0.910, 0.961). Using a probability threshold of 0.5, the sensitivity, specificity, NPV, and PPV were 0.753 (0.647, 0.840), 0.909 (0.869, 0.940), 0.711 (0.606, 0.802), and 0.925 (0.888, 0.953), respectively. The model proposed by Bai et al. (8) attained an AUROC of 0.731 (0.674, 0.80). However, despite attaining a reasonably AUC value, the model was badly calibrated, and the predicted probabilities of COVID-19 were all clustered around 0.42, meaning that the sensitivity, specificity, PPV, and NPV are meaningless. We tried to retrain the model and rechecked the code implementation; however, we could not obtain more meaningful results. Covinet (9) attained an AUROC of 0.810 (0.748, 0.853). Using a probability threshold of 0.5, the sensitivity, specificity, NPV, and PPV were 0.824 (0.726, 0.898), 0.596 (0.537, 0.654), 0.378 (0.308, 0.452), and 0.919 (0.870 0.954), respectively. COVNet (7) attained an AUROC of 0.935 (0.912, 0.959). Using a probability threshold of 0.5, the sensitivity, specificity, NPV, and PPV were 1.0 (0.958, 1.0), 0.796 (0.745, 0.842), 0.594 (0.509, 0.676), and 1.0 (0.984, 1.0), respectively. Full results for a range of probability thresholds are shown in Table 2, with ROC curves shown in Figure 4.

Calibration curves in Figure 5 show reasonable calibration for the mixed-effects model, although the model may still benefit from some recalibration. The other models do not have good calibration and likely provide harmful predictions. The decision curve in Figure 6 shows that the proposed model is of great clinical benefit compared to the treat all and treat-none approach.

It is important to remember that the model was selected using this internal testing set to avoid overfitting on the training set; therefore, these results are biased, and the external validation results are more representative of the true model performance.

On the external geographical validation dataset, the proposed model attained an AUROC of 0.930 (0.914, 0.947). With a probability threshold of 0.5, the sensitivity, specificity, NPV, and PPV were 0.778 (0.720, 0.828), 0.882 (0.853, 0.908), 0.744 (0.686, 0.797), and 0.90 (0.872, 0.924), respectively. The model proposed by Bai et al. (8) again attained a reasonable AUROC of 0.805 (0.774, 0.836); however, the sensitivity, specificity, NPV, and PPV were meaningless. Covinet (9) attained an AUROC of 0.651 (0.610, 0.691). Using a probability threshold of 0.5, the sensitivity, specificity, NPV, and PPV were0.008 (0.001, 0.029), 0.991 (0.979, 0.997), 0.286 (0.037, 0.710), and 0.695 (0.661, 0.727), respectively. COVNet (7) attained an AUROC of 0.808 (0.775, 0.841). With a cut-off point of 0.5, the sensitivity, specificity, NPV, and PPV were 0.387 (0.325, 0.451), 0.940 (0.917, 0.959), 0.740 (0.655, 0.814), and 0.777 (0.744, 0.808), respectively. Full results are shown in Table 3.

Similar to the internal validation, Figure 7 shows reasonable calibration for the mixed-effects model, although some recalibration may improve performance. Again, the comparison models could give harmful predictions as they are poorly calibrated. The decision curve in Figure 8 shows that the model is of great clinical benefit compared to the treat all and treat-none approach.

Although our proposed method and the Covnet model showed comparable performance on the internal validation set, the Covnet model could not generalize to the external geographical validation set, and calibration showed that the Covnet model would provide harmful risk estimates. This highlights the need for robust external validation in each intended setting. Nevertheless, the results show that the proposed method better generalizes to external geographical datasets and provides less harmful predictions when compared to the four previously proposed methods based on the calibration curves.

### 3.1. Saliency maps

It is vital to understand how the algorithm makes decisions and to check that it identifies the correct features within the image. Saliency maps can be used as a visual check to see what features the algorithm is learning. For example, the saliency maps in Figure 9 show that the model correctly identifies the diseased areas of the scans. We used 100 samples with a smoothing noise of 0.05 to create these saliency maps.

TABLE 2 Area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) on the internal validation dataset.

| Model | AUROC | Threshold | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|---|---|
| Bai et al. | 0.731 (0.674, 0.80) | 0.3 | 0 0.0 (0.0, 0.042) | 1.0 (0.987, 1.0) | NA | 0.77 (0.724, 0.812) |
| | | 0.4 | 0.012 (0, 0.064) | 0.996 (0.981, 1.0) | 0.50 (0.013, 0.987) | 0.772 (0.725, 0.814) |
| | | 0.5 | 1.0 (0.958, 1.0) | 0.0 (0.0, 0.013) | 0.230 (0.188, 0.276) | NA |
| | | 0.6 | 1.0 (0.958, 1.0) | 0.0 (0.0, 0.013) | 0.230 (0.188, 0.276) | NA |
| | | 0.7 | 1.0 (0.958, 1.0) | 0.0 (0.0, 0.013) | 0.230 (0.188, 0.276) | NA |
| CoviNet | 0.801 (0.748, 0.853) | 0.3 | 0.459 (0.350, 0.570) | 0.898 (0.857, 0.931) | 0.574 (0.448, 0.693) | 0.848 (0.802, 0.886) |
| | | 0.4 | 0.706 (0.597, 0.80) | 0.761 (0.708, 0.810) | 0.469 (0.380, 0.559) | 0.897 (0.851, 0.932) |
| | | 0.5 | 0.824 (0.726, 0.898) | 0.596 (0.537, 0.654) | 0.378 (0.308, 0.452) | 0.919 (0.870 0.954) |
| | | 0.6 | 0.918 (0.838, 0.966) | 0.446 (0.387, 0.505) | 0.331 (0.271, 0.394) | 0.948 (0.895, 0.979) |
| | | 0.7 | 0.965 (0.90, 0.993) | 0.246 (0.197, 0.30) | 0.276 (0.226, 0.331) | 0.959 (0.885, 0.991) |
| CovNet | 0.935 (0.912, 0.959) | 0.3 | 0.941 (0.868, 0.981) | 0.839 (0.791, 0.879) | 0.635 (0.544, 0.719) | 0.98 (0.953, 0.993) |
| | | 0.4 | 0.965 (0.90, 0.993) | 0.825 (0.775, 0.867) | 0.621 (0.533, 0.704) | 0.987 (0.964, 0.997) |
| | | 0.5 | 1.0 (0.958, 1.0) | 0.796 (0.745, 0.842) | 0.594 (0.509, 0.676) | 1.0 (0.984, 1.0) |
| | | 0.6 | 1.0 (0.958, 1.0) | 0.779 (0.726, 0.826) | 0.574 (0.490, 0.655) | 1.0 (0.984, 1.0) |
| | | 0.7 | 1.0 (0.958, 1.0) | 0.761 (0.708, 0.810) | 0.556 (0.473, 0.636) | 1.0 (0.984, 1.0) |
| Mixed-effects (ours) | 0.936 (0.910, 0.961) | 0.3 | 0.588 (0.476 0.694) | 0.961 (0.932, 0.981) | 0.820 (0.70, 0.906) | 0.887 (0.846, 0.920) |
| | | 0.4 | 0.659 (0.548, 0.758) | 0.933 (0.898, 0.959) | 0.747 (0.633, 0.840) | 0.902 (0.862, 0.933) |
| | | 0.5 | 0.753 (0.647, 0.840) | 0.909 (0.869, 0.940) | 0.711 (0.606, 0.802) | 0.925 (0.888, 0.953) |
| | | 0.6 | 0.812 (0.712, 0.888) | 0.884 (0.841, 0.919) | 0.676 (0.577, 0.766) | 0.940 (0.905 0.960) |
| | | 0.7 | 0.906 (0.823 0.958) | 0.832 (0.783, 0.873) | 0.616 (0.525, 0.702) | 0.967 (0.937, 0.986) |

Point estimates and 95% confidence intervals were calculated using De Long's method for AUROC and Jeffrey's interval for sensitivity, specificity, PPV, and NPV. Results are shown at a range of probability thresholds.

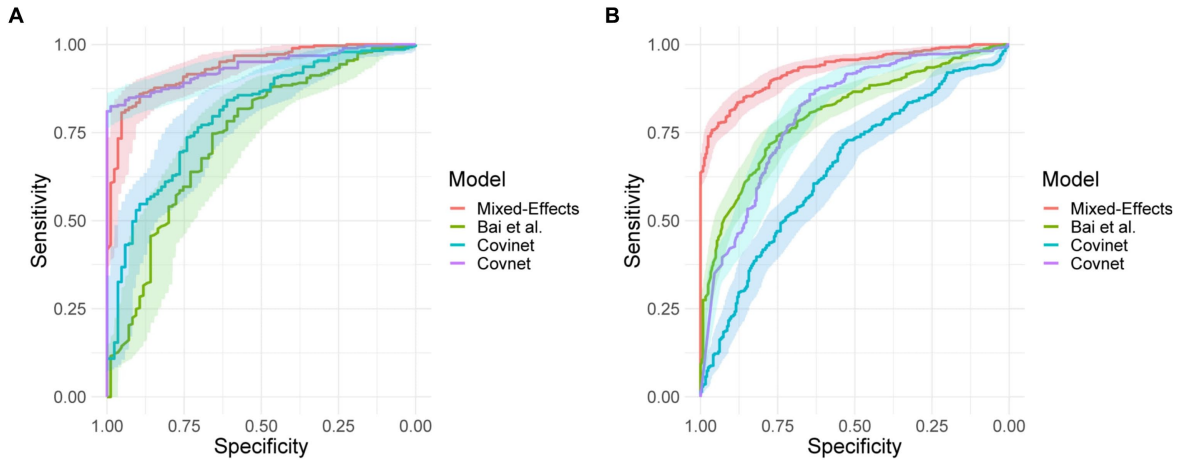


FIGURE 4
Receiver operating characteristic curves for **(A)** the MosMed internal validation set and **(B)** the Zhang et al. external validation set.

## 3.2. Sensitivity analysis

Mixed-effects models are capable of accounting for missing data. However, only one image had less than 20 slices; hence, we could not adequately assess if our model can indeed maintain good performance with missing data. Here, we rerun the analysis using the same dataset, using the same model and weights; however, we reduce the number of slices available as testing data inputs to simulate missing data. Blank images replace these slices. We uniformly sampled the slices choosing between 10 and 19 slices; this equates to between 5 and 50% missing data for the model. We ran inference at each level of missingness and briefly show the AUROC to determine at which point the predictive performance is significantly reduced.

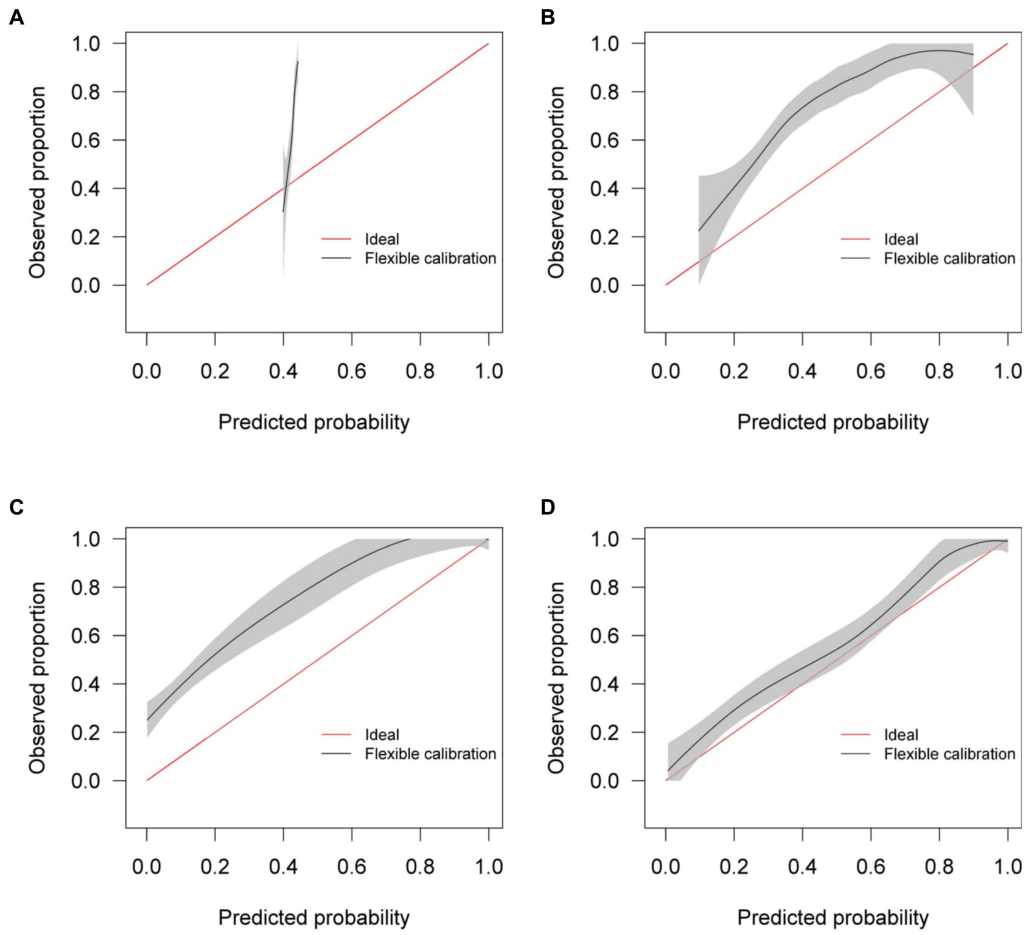


FIGURE 5
Calibration curves for **(A)** the Bai et al. model **(B)** the Covinet model, **(C)** the Covnet model, **(D)** the proposed mixed-effects model on the Mosmed internal validation dataset.

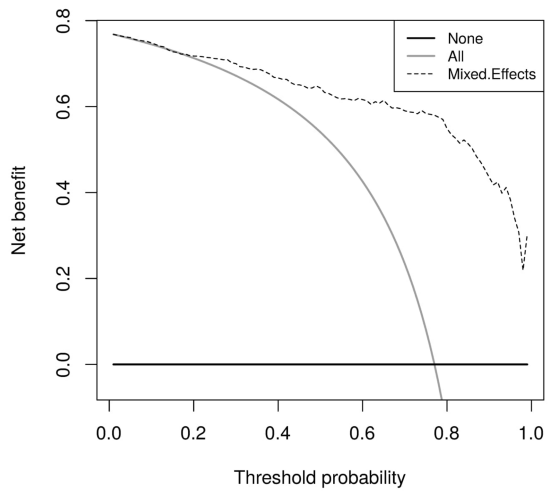


FIGURE 6
Decision curves for the proposed mixed-effects model on the Mosmed internal validation dataset.

The plot of AUROCs at different levels of missingness is shown in Figure 10, along with 95% confidence intervals. We can see that at 20% missingness, there is a statistically significant decrease in predictive performance. Although, even at 50% missingness, the model still performs relatively well, with an AUROC of 0.890 (95% CI: 0.868, 0.912). It should be noted that this does not mean that there is no reduction in performance at 5–15% missingness, only that the reduction was not statistically significant at the 95% confidence level.

Deep learning models can be susceptible to adversarial attacks (50), where minor artifacts or noise on an image can cause the image to be misclassified, even when the image does not look significantly different to a human observer. Here, we perform a brief sensitivity analysis by adding a small Gaussian noise to the image. We tested the model performance on the external dataset, with each image having a random Gaussian noise added. Experiments were conducted with standard deviations of 0 up to 0.005 in increments of 0.001 added to the normalized image. We did not add Gaussian noise in the data augmentation so that the model is not explicitly trained to deal with this kind of attack.

When using a variance of 0, the images are unchanged, and the results are the same as the standard results above. We present results on the Zhang et al. (49) dataset. Example images for each level of variance are shown in Figure 11, and a graph showing the reduction in AUROC is shown in Figure 12.

TABLE 3 Area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) on the external validation dataset.

| Model | AUROC | Threshold | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|---|---|
| Bai et al. | 0.805 (0.774, 0.836) | 0.3 | 0.0 (0.0, 0.015) | 1.0 (0.993, 1.0) | NA | 0.695 (0.661, 0.727) |
| | | 0.4 | 0.0 (0.0, 0.015) | 1.0 (0.993, 1.0) | NA | 0.695 (0.661, 0.727) |
| | | 0.5 | 1.0 (0.985, 1.0) | 1.0 (0.0, 0.007) | 0.305 (0.273, 0.339) | NA |
| | | 0.6 | 1.0 (0.985, 1.0) | 1.0 (0.0, 0.007) | 0.305 (0.273, 0.339) | NA |
| | | 0.7 | 1.0 (0.985, 1.0) | 1.0 (0.0, 0.007) | 0.305 (0.273, 0.339) | NA |
| CoviNet | 0.651 (0.610, 0.691) | 0.3 | 0.0 (0.0, 0.015) | 1.0 (0.993, 1.0) | NA | 0.695 (0.661, 0.727) |
| | | 0.4 | 0.0 (0.0, 0.015) | 1.0 (0.993, 1.0) | NA | 0.695 (0.661, 0.727) |
| | | 0.5 | 0.008 (0.001, 0.029) | 0.991 (0.979, 0.997) | 0.286 (0.037, 0.710) | 0.695 (0.661, 0.727) |
| | | 0.6 | 0.160 (0.117, 0.213) | 0.929 (0.905, 0.949) | 0.50 (0.385, 0.615) | 0.716 (0.681, 0.749) |
| | | 0.7 | 0.551 (0.487, 0.615) | 0.694 (0.654, 0.733) | 0.442 (0.385, 0.50) | 0.779 (0.740, 0.815) |
| CovNet | 0.808 (0.775, 0.841) | 0.3 | 0.305 (0.247, 0.367) | 0.969 (0.951, 0.982) | 0.813 (0.718, 0.887) | 0.760 (0.727, 0.791) |
| | | 0.4 | 0.354 (0.294, 0.418) | 0.955 (0.934, 0.971) | 0.775 (0.686, 0.849) | 0.771 (0.737, 0.802) |
| | | 0.5 | 0.387 (0.325, 0.451) | 0.940 (0.917, 0.959) | 0.740 (0.655, 0.814) | 0.777 (0.744, 0.808) |
| | | 0.6 | 0.432 (0.369, 0.497) | 0.937 (0.913, 0.956) | 0.750 (0.670, 0.819) | 0.790 (0.756, 0.820) |
| | | 0.7 | 0.473 (0.409, 0.538) | 0.931 (0.907, 0.951) | 0.752 (0.675, 0.818) | 0.801 (0.768, 0.831) |
| Mixed-effects (ours) | 0.930 (0.914, 0.947) | 0.3 | 0.675 (0.612, 0.733) | 0.935 (0.911, 0.954) | 0.820 (0.760, 0.871) | 0.867 (0.838, 0.894) |
| | | 0.4 | 0.741 (0.681, 0.795) | 0.904 (0.877, 0.927) | 0.773 (0.713, 0.825) | 0.888 (0.859, 0.913) |
| | | 0.5 | 0.778 (0.720, 0.828) | 0.882 (0.853, 0.908) | 0.744 (0.686, 0.797) | 0.90 (0.872, 0.924) |
| | | 0.6 | 0.827 (0.774, 0.873) | 0.859 (0.827, 0.887) | 0.720 (0.664, 0.772) | 0.919 (0.892, 0.941) |
| | | 0.7 | 0.885 (0.838, 0.922) | 0.828 (0.794, 0.859) | 0.694 (0.639, 0.744) | 0.942 (0.918, 0.961) |

Point estimates and 95% confidence intervals were calculated using De Long's method for AUROC and Jeffrey's interval for sensitivity, specificity, PPV, and NPV. Results are shown at a range of probability thresholds.

### 3.3. Fixed-effects only

To show that the mixed-effects method improves prediction over the fixed-effects method alone, we removed the random-effects part of the model to leave the fixed effects only. This was the only change to the model and allowed us to see the added benefit of the mixed-effects part. The full results are shown in Tables 4, 5. This experiment shows much worse performance when the random effects are removed from the model.

## 4. Discussion

Artificial intelligence is set to revolutionize healthcare, allowing large amounts of data to be processed and analyzed automatically, reducing pressure on stretched healthcare services. These tools can aid clinicians in monitoring and managing both common conditions and outbreaks of novel diseases. However, these tools must be assessed adequately, and best practice guidelines for reporting and development must be followed closely to increase reproducibility and reduce bias. We have developed a deep learning model to classify CT scans as healthy or COVID-19 using a novel mixed-effects model. Following best practice guidelines, we have externally validated the model. In addition, we robustly externally geographically validated the developed model in several performance areas, which are not routinely reported. For example, discriminative performance measures show that the model can discriminate between healthy and COVID-19 CT scans well, calibration shows that the model is not clinically harmful. Finally, the clinical usefulness measures show that the model may be useful in a clinical setting. From the results presented here, it would seem that our deep learning model outperforms the RT-PCR tests as shown in the review by Watson et al. (3); however, those results are conservative estimates and were conducted under real-world clinical settings. A prospective study is required to determine if this is the case.

Compared to previously proposed models, our model showed similar discriminative performance to one existing method; however, our method generalized better to an external geographical validation set and showed improved calibration performance. Interestingly, in both internal and external validation, the sensitivity and NPV are similar in all models. However, the specificity and PPV are statistically significantly improved for the mixed-effects model in the external validation dataset. The performance of the proposed model in the external validation set is similar to that reported by PCR testing (3). However, a direct comparison should not be made as PCR testing on this exact dataset is unavailable.

There are several limitations of the study that should be highlighted and improved in future work. Firstly, we have only performed external geographical validation in a single dataset. Further external validation, both geographical and temporal, is needed on many datasets to determine if the model is correct in each intended setting. Although we performed a brief sensitivity analysis

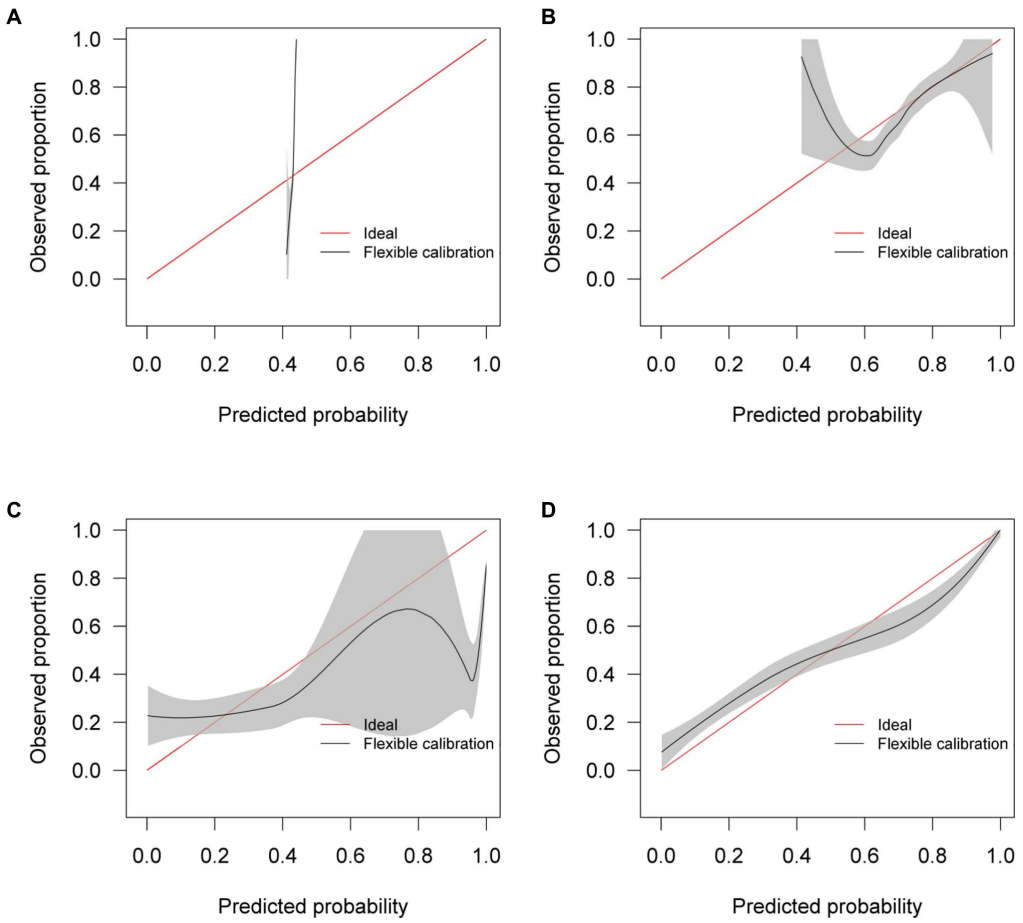


FIGURE 7
Calibration curves for **(A)** the Bai et al. model **(B)** the Covinet model, **(C)** the Covnet model, **(D)** the proposed mixed-effects model on the Zhang et al. external validation dataset.

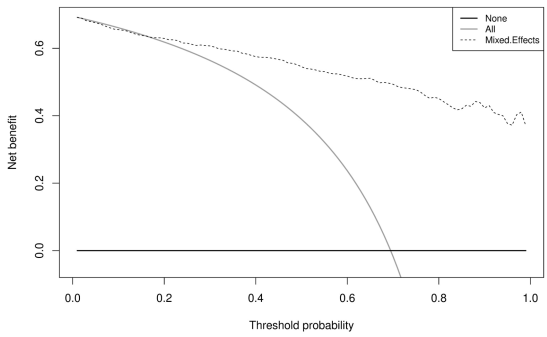


FIGURE 8
Decision curves for the proposed mixed-effects model on the Zhang et al. external validation dataset.

here, more extensive work on adversarial attacks is needed. Future studies could consider following the method proposed by Goodfellow et al. (50) to improve robustness against adversarial examples. Patient demographic data were not available for this study, but future studies could incorporate this data into the model to improve results. Finally, rules of thumb for assessing sample size calculations in the validation set can lead to imprecise results (51). Simulating data is a better alternative; however, it is difficult to anticipate the distribution of the model's linear predictor. Therefore, we were required to revert to the rule of thumb using a minimum of 200 samples in each group (29).

Initial experiments used the Zhang et al. (49) dataset for training; this showed promising results on the internal validation set; however, external validation showed random results. In addition, saliency maps showed that the model was not using the features of COVID-19 to make the diagnosis and was instead using the area around the image. We concluded that the images for each class were slightly different, perhaps due to different imaging protocols, and the algorithm was learning the image format rather than the disease. We then used the MosMed dataset for training and the Zhang et al. (49) dataset for external validation. This highlights the need for good quality training data and external validation and visualization.

Future studies should validate models and follow reporting guidelines such as TRIPOD (17) or the upcoming QUADAD-AI (52) and TRIPOD-AI (53) to bring about clinically useful and deployable models. Further research could look deeper into the areas of images identified by the algorithm as shown on the

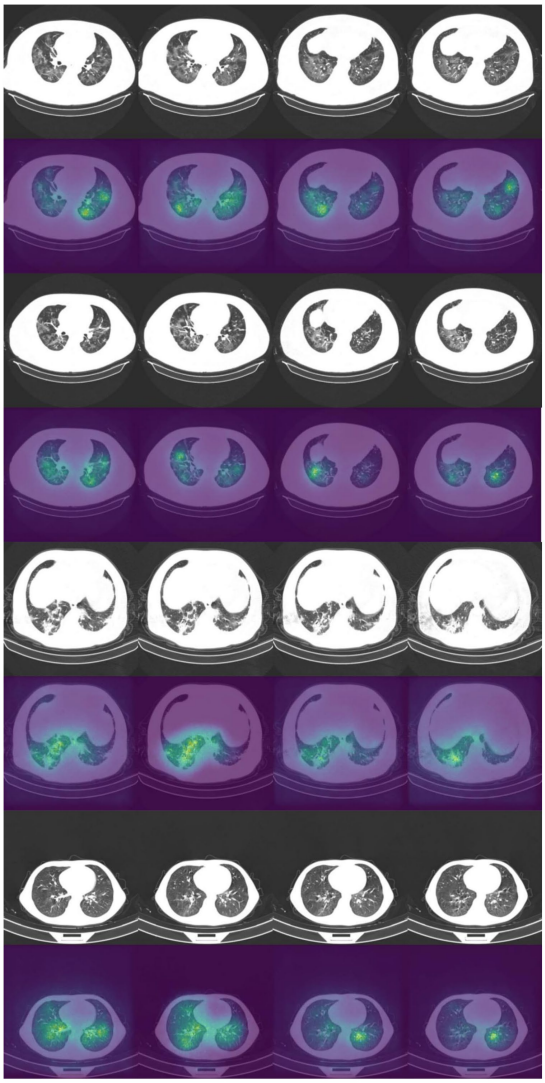

FIGURE 9
Example of original images and saliency maps showing highlighted regions on four patients in the Zhang et al. dataset. Four consecutive images display how the diseased areas differ between slices. All images are taken from the external validation set.

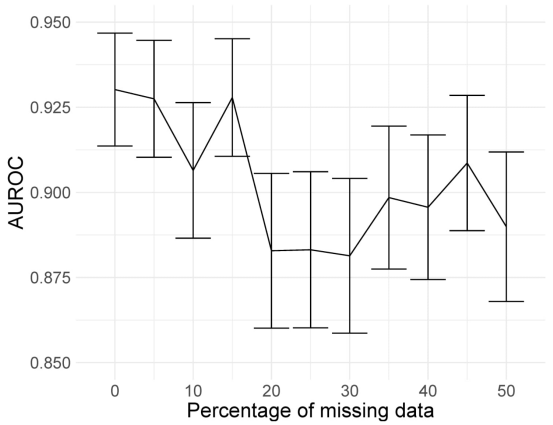


FIGURE 10
AUROC values at different levels of missingness. At 20% missingness, the loss in performance becomes statistically significant; however, even with 50% missing images, the model still has a reasonably high AUROC.

saliency maps; this could potentially identify new features of COVID-19 which have gone unnoticed. Before any model can be fully deployed, clinical trials are needed to study the full impact of using such algorithms to diagnose COVID-19 and the exact situations in which such a model may be used. In-clinic prospective studies comparing the performance deep learning models with RT-PCR and lateral flow tests should be carried out to determine how deep learning compares; this will show whether deep learning could be used as an automated alternative to RT-PCR testing.

This study indicates that deep learning could be suitable for screening and monitoring of COVID-19 in a clinical setting; however, validation in the intended setting is vital, and models should not be adopted without this. It has been observed that the quality of reporting of deep learning prediction models is usually very poor; however, with a bit of extra work and by following best practice guidelines, this problem can be overcome. This study highlights the importance of robust analysis and reporting of models with external validation.

## Data availability statement

The source code and datasets presented in this study can be found in an online repository. The accession link to the source code is: https://github.com/JTBridge/ME-COVID19. Publicly available datasets were analyzed in this study. These data can be found at: the CNCB and MosMed repositories, available from doi: 10.17816/DD46826 and doi: 10.1016/j.cell.2020.04.045.

## Ethics statement

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

## Author contributions

JB, YM, and YaZ: conception. JB, YM, YiZ, and YaZ: methodology. CA, MW, CMe, SF, MM, SM, and YaZ: administration. JB, YM, WZ, TF, CMc, RS, YiZ, and YaZ: investigation. JB, YM, CA, MW, CMe, SF, MM, SM, RS, and YiZ: data curation. JB: analysis and validation. JB, YM, WZ, and YaZ: writing of first draft. All authors made substantial contributions to the reviewing and editing of the manuscript.

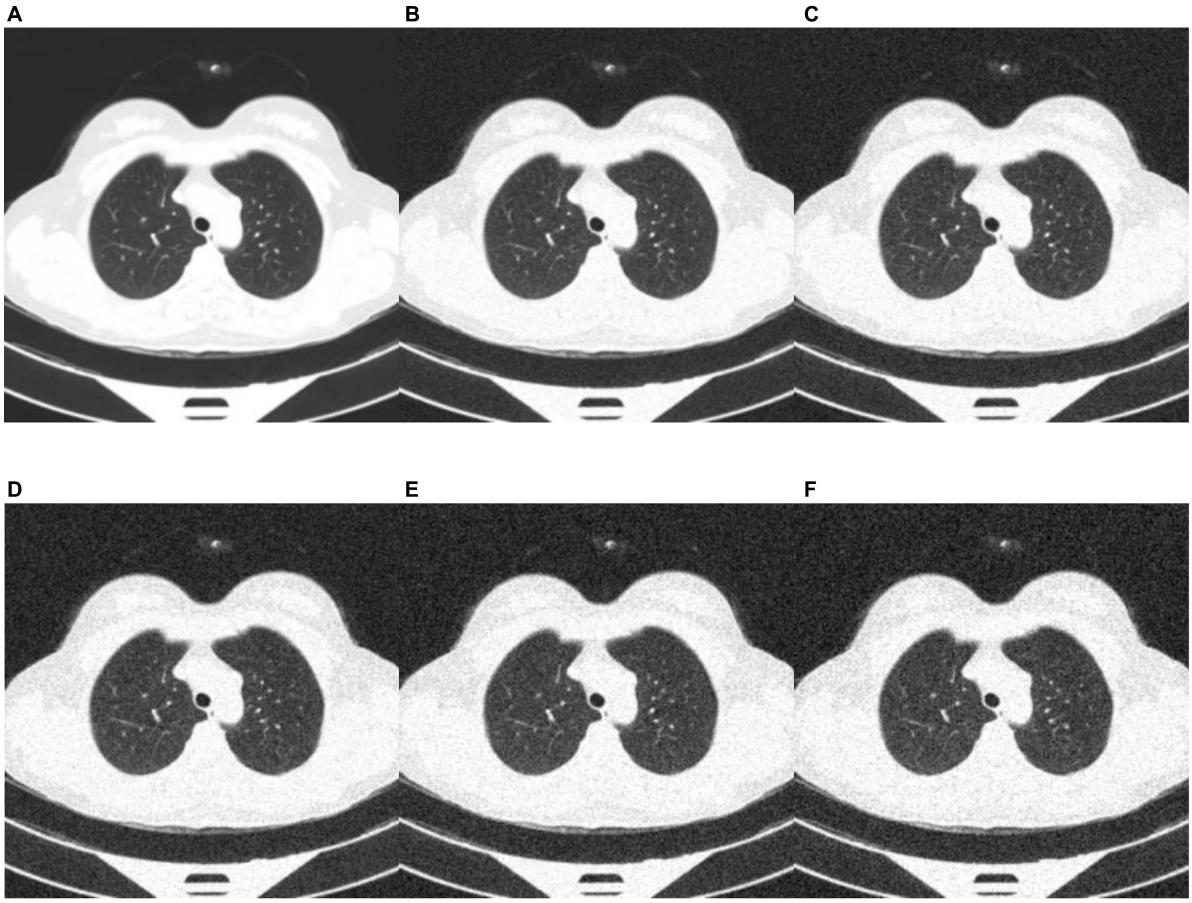


FIGURE 11
Example images showing the effect of increasing the amount of noise in the image input. **(A)** no noise; **(B)** deviation = 0.001; **(C)** deviation = 0.002; **(D)** deviation = 0.003; **(E)** deviation = 0.004; **(F)** deviation = 0.005.

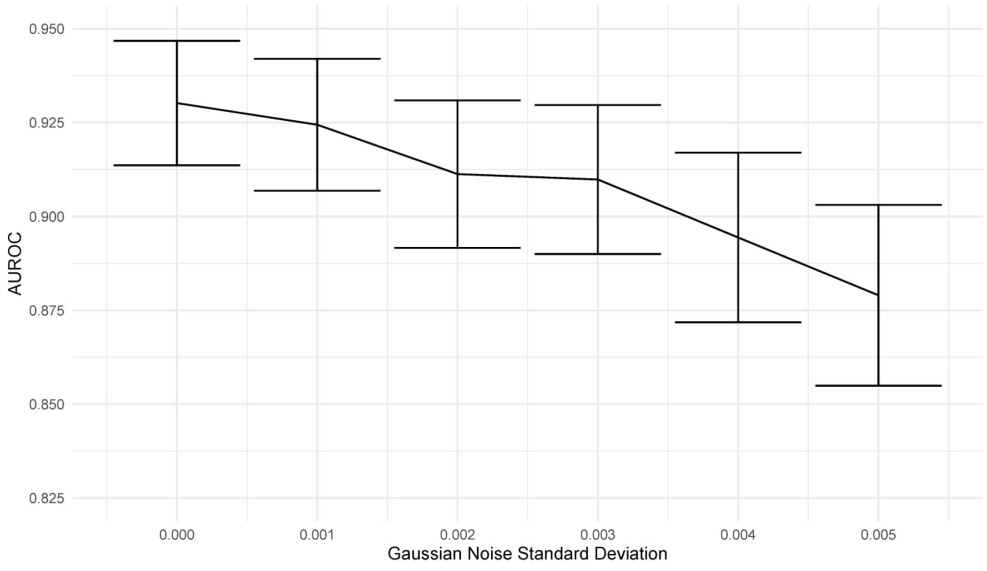


FIGURE 12
Graph showing the drop in AUROC as the amount of noise in the image input increases. The AUROC falls steadily with increased noise in the image.

TABLE 4 Area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) on the internal validation dataset for our proposed model and the fixed effects model.

| Model | AUROC | Threshold | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|---|---|
| Fixed effects | 0.494 (0.427, 0.561) | 0.3 | 0.859 (0.766, 0.925) | 0.165 (0.124, 0.213) | 0.235 (0.189, 0.286) | 0.797 (0.672, 0.890) |
| | | 0.4 | 0.953 (0.884, 0.987) | 0.046 (0.025, 0.077) | 0.229 (0.187, 0.277) | 0.765 (0.501, 0.932) |
| | | 0.5 | 0.988 (0.936, 1.0) | 0.014 (0.004, 0.036) | 0.231 (0.188, 0.277) | 0.80 (0.284, 0.995) |
| | | 0.6 | 1.0 (0.958, 1.0) | 0.0 (0.0, 1.0) | 0.230 (0.188, 0.276) | NA (NA, NA) |
| | | 0.7 | 1.0 (0.958, 1.0) | 0.0 (0.0, 1.0) | 0.230 (0.188, 0.276) | NA (NA, NA) |
| Mixed-effects (fixed + random) | 0.936 (0.910, 0.961) | 0.3 | 0.588 (0.476 0.694) | 0.961 (0.932, 0.981) | 0.820 (0.70, 0.906) | 0.887 (0.846, 0.920) |
| | | 0.4 | 0.659 (0.548, 0.758) | 0.933 (0.898, 0.959) | 0.747 (0.633, 0.840) | 0.902 (0.862, 0.933) |
| | | 0.5 | 0.753 (0.647, 0.840) | 0.909 (0.869, 0.940) | 0.711 (0.606, 0.802) | 0.925 (0.888, 0.953) |
| | | 0.6 | 0.812 (0.712, 0.888) | 0.884 (0.841, 0.919) | 0.676 (0.577, 0.766) | 0.940 (0.905 0.960) |
| | | 0.7 | 0.906 (0.823 0.958) | 0.832 (0.783, 0.873) | 0.616 (0.525, 0.702) | 0.967 (0.937, 0.986) |

Point estimates and 95% confidence intervals were calculated using De Long's method for AUROC and Jeffrey's interval for sensitivity, specificity, PPV, and NPV. Results are shown at a range of probability thresholds.

TABLE 5 Area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) on the external validation dataset for our proposed model and the fixed effects model.

| Model | AUROC | Threshold | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|---|---|
| Fixed effects | 0.630 (0.590, 0.670) | 0.3 | 0.794 (0.738, 0.843) | 0.374 (0.334, 0.416) | 0.358 (0.317, 0.40) | 0.805 (0.752, 0.852) |
| | | 0.4 | 0.971 (0.942, 0.988) | 0.159 (0.130, 0.192) | 0.337 (0.302, 0.373) | 0.926 (0.854, 0.970) |
| | | 0.5 | 1.0 (0.984, 1.0) | 0.063 (0.044, 0.087) | 0.319 (0.286, 0.354) | 1.0 (0.90, 1.0) |
| | | 0.6 | 1.0 (0.985, 1.0) | 0.018 (0.277, 0.343) | 0.309 (0.277, 0.343) | 1.0 (0.692, 1.0) |
| | | 0.7 | 1.0 (0.985, 1.0) | 0.004 (0.0, 0.013) | 0.306 (0.274, 0.339) | 1.0 (0.158, 1.0) |
| Mixed-effects (fixed + random) | 0.930 (0.914, 0.947) | 0.3 | 0.675 (0.612, 0.733) | 0.935 (0.911, 0.954) | 0.820 (0.760, 0.871) | 0.867 (0.838, 0.894) |
| | | 0.4 | 0.741 (0.681, 0.795) | 0.904 (0.877, 0.927) | 0.773 (0.713, 0.825) | 0.888 (0.859, 0.913) |
| | | 0.5 | 0.778 (0.720, 0.828) | 0.882 (0.853, 0.908) | 0.744 (0.686, 0.797) | 0.90 (0.872, 0.924) |
| | | 0.6 | 0.827 (0.774, 0.873) | 0.859 (0.827, 0.887) | 0.720 (0.664, 0.772) | 0.919 (0.892, 0.941) |
| | | 0.7 | 0.885 (0.838, 0.922) | 0.828 (0.794, 0.859) | 0.694 (0.639, 0.744) | 0.942 (0.918, 0.961) |

Point estimates and 95% confidence intervals were calculated using De Long's method for AUROC and Jeffrey's interval for sensitivity, specificity, PPV, and NPV. Results are shown at a range of probability thresholds.

## Funding

This study received funding from EPSRC studentship (No. 2110275), EPSRC Impact Acceleration Account (IAA) Awards, and Amazon Web Services. The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

## Conflict of interest

CMe and SF were employed by Alces Flight Ltd. MM and SM were employed by Amazon Web Services.

The remaining 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/fmed.2023.1113030/full#supplementary-material

# References


1. *Coronavirus disease 2019 (COVID-19) - symptoms, diagnosis and treatment | BMJ Best Practice* BMJ Publishing Group (2020) Available at: https://bestpractice.bmj.com/topics/en-gb/3000201.

2. Torjesen I. Covid-19 will become endemic but with decreased potency over time, scientists believe. *BMJ*. (2021) 372:n494. doi: 10.1136/bmj.n494

3. Watson J, Whiting PF, Brush JE. Interpreting a covid-19 test result. *BMJ*. (2020) 369:m1808. doi: 10.1136/bmj.m1808

4. Mallett S, Allen AJ, Graziadio S, Taylor SA, Sakai NS, Green K, et al. At what times during infection is SARS-CoV-2 detectable and no longer detectable using RT-PCR-based tests? A systematic review of individual participant data. *BMC Med*. (2020) 18:346. doi: 10.1186/s12916-020-01810-8

5. Ruan Z-R, Gong P, Han W, Huang M-Q, Han M. A case of coronavirus disease 2019 with twice negative nucleic acid testing within 8 days. *Chin Med J*. (2020) 133:1487–8. doi: 10.1097/CM9.0000000000000788

6. Pontone G, Scafuri S, Mancini ME, Agalbato C, Guglielmo M, Baggiano A, et al. Role of computed tomography in COVID-19. *J Cardiovasc Comput Tomogr*. (2020) 15:27–36. doi: 10.1016/j.jcct.2020.08.013

7. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, et al. Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy. *Radiology*. (2020) 296:E65–71. doi: 10.1148/radiol.2020200905

8. Bai HX, Wang R, Xiong Z, Hsieh B, Chang K, Halsey K, et al. Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. *Radiology*. (2020) 296:E156–65. doi: 10.1148/radiol.2020201491

9. Mittal B, Oh J. CoviNet: Covid-19 diagnosis using machine learning analyses for computerized tomography images. *Thirteenth International Conference on Digital Image Processing (ICDIP 2021)*. SPIE (2021).

10. MacCormick IJC, Zheng Y, Czanner S, Zhao Y, Diggle PJ, Harding SP, et al. Spatial statistical modelling of capillary non-perfusion in the retina. *Sci Rep*. (2017) 7:16792. doi: 10.1038/s41598-017-16620-x

11. Zhu W, Ku JY, Zheng Y, Knox PC, Kolamunnage-Dona R, Czanner G. Spatial linear mixed effects modelling for OCT images: SLME model. *J Imaging*. (2020) 6:44. doi: 10.3390/jimaging6060044

12. Albtoush OM, Al-Shdefat RB, Al-Akaileh A. Chest CT scan features from 302 patients with COVID-19 in Jordan. *Eur J Radiol Open*. (2020) 7:100295. doi: 10.1016/j.ejro.2020.100295

13. Haseli S, Khalili N, Bakhshayeshkaram M, Sanei Taheri M, Moharramzad Y. Lobar distribution of COVID-19 pneumonia based on chest computed tomography findings; a retrospective study. *Arch Acad Emerg Med*. (2020) 8:e55-e. doi: 10.22037/aaem.v8i1.665

14. Xiang C, Lu J, Zhou J, Guan L, Yang C, Chai C. CT findings in a novel coronavirus disease (COVID-19) pneumonia at initial presentation. *Biomed Res Int*. (2020) 2020:1–10. doi: 10.1155/2020/5436025

15. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*. (2020) 369:m1328. doi: 10.1136/bmj.m1328

16. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell*. (2021) 3:199–217. doi: 10.1038/s42256-021-00307-0

17. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. (2015) 350:g7594. doi: 10.1136/bmj.g7594

18. Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell*. (2020) 2:e200029. doi: 10.1148/ryai.2020200029

19. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. (2019) 170:51–8. doi: 10.7326/M18-1376

20. Jiang J, Nguyen T. *Linear and generalized linear mixed models and their applications*. New York, NY: Springer (2007).

21. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE (2016);2818–2826.

22. Deng J, Dong W, Socher R, Li L, Kai L, Li F-F. ImageNet: a large-scale hierarchical image database. *2019 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE (2009); 248–255.

23. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. (2009) 338:b2393. doi: 10.1136/bmj.b2393

24. Xiong Y, Kim HJ, Singh V. Mixed effects neural networks (MeNets) with applications to gaze estimation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE (2019);7735–7744.

25. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev*. (1950) 78:1–3. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

26. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv Large Marg Classif*. (1999) 10:61–74.

27. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. (2010) 21:128–38. doi: 10.1097/EDE.0b013e3181c30fb2

28. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med*. (2019) 17:230. doi: 10.1186/s12916-019-1466-7

29. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. (2016) 74:167–76. doi: 10.1016/j.jclinepi.2015.12.005

30. R Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. (2021).

31. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. (1988) 44:837–45. doi: 10.2307/2531595

32. Stevenson M, Sergeant E, Nunes T, Heuer C, Marshall J, Sanchez J, et al. *epiR: tools for the analysis of epidemiological data*. (2022). Availabe at: https://CRAN.R-project.org/package=epiR

33. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Stat Sci*. (2001) 16:101–17. doi: 10.1214/ss/1009213286

34. Wynants L, van Smeden M, McLernon DJ, Timmerman D, Steyerberg EW, Van Calster B, et al. Three myths about risk thresholds for prediction models. *BMC Med*. (2019) 17:192. doi: 10.1186/s12916-019-1425-3

35. Harrell FE Jr. *rms: regression modeling strategies*. (2021). Available at: https://CRAN.R-project.org/package=rms

36. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak*. (2006) 26:565–74. doi: 10.1177/0272989X06295361

37. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. Smoothgrad: removing noise by adding noise. *arXiv*. (2017) arXiv:170603825. doi: 10.48550/arXiv.1706.03825

38. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE (2016);770–778.

39. Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning* (2019) PMLR.

40. Mittal B, Oh J. CoviNet: Covid-19 diagnosis using machine learning analyses for computerized tomography images. SPIE Proceeding (2021).

41. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensorflow: large-scale machine learning on heterogeneous distributed systems. *SPIE* (2016) *arXiv:160304467*. doi: 10.48550/arXiv.1603.04467

42. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al., editors. Tensorflow: a system for large-scale machine learning. *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. USENIX Association. (2016).

43. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. (2011) 12:77. doi: 10.1186/1471-2105-12-77

44. Du Z, Hao Y. *reportROC: an easy way to report ROC analysis*. R package version 3.5. (2020). Available at: https://CRAN.R-project.org/package=reportROC

45. Wickham H. *ggplot2: elegant graphics for data analysis*. New York, NY: Springer-Verlag (2016).

46. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *arXiv*. (2014) 14126980. doi: 10.48550/arXiv.1412.6980

47. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE Jr, Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med*. (2019) 38:1276–96. doi: 10.1002/sim.7992

48. Morozov SP, Andreychenko AE, Blokhin IA, Gelezhe PB, Gonchar AP, Nikolaev AE, et al. MosMedData: data set of 1110 chest CT scans performed during the COVID-19 epidemic. *Digit Diagn*. (2020) 1:49–59. doi: 10.17816/DD46826

49. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cells*. (2020) 181:1423–33.e11. doi: 10.1016/j.cell.2020.04.045

50. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. *arXiv*. (2014) 14126572. doi: 10.48550/arXiv.1412.6572

51. Snell KIE, Archer L, Ensor J, Bonnett LJ, Debray TPA, Phillips B, et al. External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. *J Clin Epidemiol*. (2021) 135:79–89. doi: 10.1016/j.jclinepi.2021.02.011

52. Sounderajah V, Ashrafian H, Rose S, Shah NH, Ghassemi M, Golub R, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nat Med*. (2021) 27:1663–5. doi: 10.1038/s41591-021-01517-0

53. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. (2021) 11:e048008. doi: 10.1136/bmjopen-2020-048008

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 13 September 2023
DOI 10.3389/fmed.2023.1211800



OPEN ACCESS

EDITED BY
Diwei Zhou,
Loughborough University, United Kingdom

REVIEWED BY
Zeju Li,
University of Oxford, United Kingdom
Safa Elsheikh,
Loughborough University, United Kingdom

*CORRESPONDENCE
Shuo Wang
✉ shuowang@fudan.edu.cn
Zhijian Song
✉ zjsong@fudan.edu.cn

†These authors have contributed equally to this work

RECEIVED 25 April 2023
ACCEPTED 21 August 2023
PUBLISHED 13 September 2023

CITATION
Wang K, Li Z, Wang H, Liu S, Pan M, Wang M, Wang S and Song Z (2023) Improving brain tumor segmentation with anatomical prior-informed pre-training.
*Front. Med.* 10:1211800.
doi: 10.3389/fmed.2023.1211800



# Improving brain tumor segmentation with anatomical prior-informed pre-training

Kang Wang[1,2†], Zeyang Li[3†], Haoran Wang[1,2], Siyu Liu[1,2], Mingyuan Pan[4], Manning Wang[1,2], Shuo Wang[1,2*] and Zhijian Song[1,2*]

[1]Digital Medical Research Center, School of Basic Medical Sciences, Fudan University, Shanghai, China, [2]Shanghai Key Lab of Medical Image Computing and Computer Assisted Intervention, Fudan University, Shanghai, China, [3]Department of Neurosurgery, Zhongshan Hospital, Fudan University, Shanghai, China, [4]Radiation Oncology Center, Huashan Hospital, Fudan University, Shanghai, China

**Introduction:** Precise delineation of glioblastoma in multi-parameter magnetic resonance images is pivotal for neurosurgery and subsequent treatment monitoring. Transformer models have shown promise in brain tumor segmentation, but their efficacy heavily depends on a substantial amount of annotated data. To address the scarcity of annotated data and improve model robustness, self-supervised learning methods using masked autoencoders have been devised. Nevertheless, these methods have not incorporated the anatomical priors of brain structures.

**Methods:** This study proposed an anatomical prior-informed masking strategy to enhance the pre-training of masked autoencoders, which combines data-driven reconstruction with anatomical knowledge. We investigate the likelihood of tumor presence in various brain structures, and this information is then utilized to guide the masking procedure.

**Results:** Compared with random masking, our method enables the pre-training to concentrate on regions that are more pertinent to downstream segmentation. Experiments conducted on the BraTS21 dataset demonstrate that our proposed method surpasses the performance of state-of-the-art self-supervised learning techniques. It enhances brain tumor segmentation in terms of both accuracy and data efficiency.

**Discussion:** Tailored mechanisms designed to extract valuable information from extensive data could enhance computational efficiency and performance, resulting in increased precision. It's still promising to integrate anatomical priors and vision approaches.

KEYWORDS

**masked autoencoder, anatomical priors, transformer, brain tumor segmentation, magnetic resonance image, self-supervised learning**

## 1. Introduction

Glioblastoma (GBM) is one of the most aggressive brain cancers among adults (1). Multi-parameter magnetic resonance imaging (MRI) provides valuable information for characterizing the size, invasiveness, and intrinsic heterogeneity of brain tumors (2, 3). Accurate delineation of GBM on multi-parameter MRI is crucial for clinical diagnosis and treatment, such as assisting surgical planning for maximum glioblastoma resection while preserving neurological function. However, the current clinical routine still relies on manual delineation, which is time-consuming and requires expert knowledge. There is a high demand for automatic brain tumor segmentation to enhance the efficiency of diagnostic procedures, facilitate surgical planning, and contribute to prognostic analyses (4).

In the last decade, there have been extensive studies on automatic brain tumor segmentation (5), and most of them are based on convolutional neural networks (CNNs) (6–8). However, due to limited receptive field, CNNs often struggle to capture long-range dependencies and global context (9, 10), potentially leading to inaccurate segmentation predictions. The recent success of transformer architecture in vision tasks (11, 12) has shown benefits in learning global contextual information. New network designs with vision transformers have emerged for medical image segmentation (13, 14) and achieved state-of-the-art (SOTA) performance in brain tumor segmentation (15–17). However, the supervised training of vision transformers typically requires a large amount of densely annotated images, otherwise there is a high risk of overfitting.

To combat the challenge of data scarcity in medical image segmentation, self-supervised learning (SSL) has proven to be a promising solution (18). In general, a pretext SSL task is designed to pre-train the network using unannotated data, and the learned encoder weights are further optimized in the downstream segmentation task. Since no manual annotation is needed for SSL, it can be applied to utilize large unannotated datasets. Recently, one of the most successful SSL frameworks is the masked language modeling (MLM), which has achieved great success in numerous natural language processing tasks with transformer-based architecture (19–21). Motivated by MLM, masked image modeling (MIM) was also proposed for pre-training vision transformers. In MIM, the model predicts masked image patches from unmasked patches. The prediction target can be either token features or raw pixel values of the masked patches. BEiT (22) utilizes a discrete variational autoencoder (dVAE) to transform all image patches into discrete tokens, which are then used to pre-train a vision transformer at the token level. However, tokenizing the image patches requires additional training of a dVAE. In contrast, He et al. (23) introduced the masked autoencoder (MAE), which randomly masks a subset of image patches and reconstructs the masked pixels from unmasked patches. The high masking ratio of MAE enables efficient pre-training of vision transformers with large annotated datasets. The success of MAE has motivated a series of variants in vision tasks (24–27) and applications in medical image analysis using MIM techniques. For instance, Tang et al. (28) utilized masked inpainting for the pre-training of a Swin UNETR (Shifted-window UNet transformer) in abdominal segmentation tasks. Chen et al. (29) compared multiple MIM approaches in abdominal segmentation. Zhou et al. (30) applied MAE pre-training with UNETR (UNet Transformer) and obtained performance gains in both abdominal and brain tumor segmentation.

Building a masked image is a crucial step in MIM pre-training. As shown in Figure 1, the smallest masking unit of MLM, such as BERT (19), is typically the vocabulary, which preserves contextual information. However, MIM employs random masking, which can disrupt the spatial context and regions with the same semantic meaning, given the absence of the concept of words commonly observed in MLM. This, in turn, makes it challenging for the representation learning process to obtain high-quality pretrained network, especially when the masking ratio reaches a high percentage. Recently, several studies demonstrated that the masking strategy has a substantial effect on model performance in downstream tasks (31, 32). Although random masking is widely used, recent advances have shown that appropriate masking strategies can achieve better performance, such as region-based masking (33), attention-based masking (34), and adaptive masking (AdaMAE) (31). These masking strategies take the patch context into account, leading to more effective and efficient pre-training.

In the context of medical images, anatomical knowledge could help improve the pre-training. Huang et al. (35) incorporated the symmetry characteristics of brain structures into the pre-training by constructing symmetric positional encodings. However, few studies have integrated the more precise brain atlas (36) into the masking strategy. Inspired by the performance gains achieved by weighted masking strategies, we propose an anatomical prior-informed masking strategy for the MAE pre-training. We hypothesize that the tumor distribution among brain structures can guide the MAE pre-training, therefore improving the downstream brain tumor segmentation. To achieve this, we analyze the tumor occurrence in the SRI-24 space and establish an anatomical prior-informed probability map for image masking. This strategy allows us to select more informative patches for MAE pre-training. By combining the data-driven MAE with anatomical knowledge, we aim to improve the accuracy and data-efficiency of brain tumor segmentation.

In this study, our contributions are as follows:

(1) An anatomical prior-informed masking strategy is proposed to enhance the pre-training of masked autoencoder. This strategy is designed to preserve contextual information in 3D medical images and allows the pre-training process to concentrate on regions that are more relevant to the downstream segmentation task.
(2) By incorporating prior-informed weighted sampling, we construct an anatomical prior-informed masked autoencoder, referred to as API-MAE. This self-supervised pre-training approach utilizes 6,415 skull-stripped brain T1 MR images and combines data-driven reconstruction with anatomical priors.
(3) Inheriting the pretrained encoder weights, our method demonstrates superior performance in the downstream segmentation task on the BraTS21 dataset, outperforming several transformer models and surpassing state-of-the-art self-supervised learning methods. Subsequent experiments demonstrate that our method exhibits greater efficiency compared with a regular masked autoencoder and maintains a satisfactory trade-off between segmentation accuracy and computational consumption.

# 2. Methodology

## 2.1. Overview of proposed method

We propose a novel masking strategy for improved MAE pre-training and downstream brain tumor segmentation in

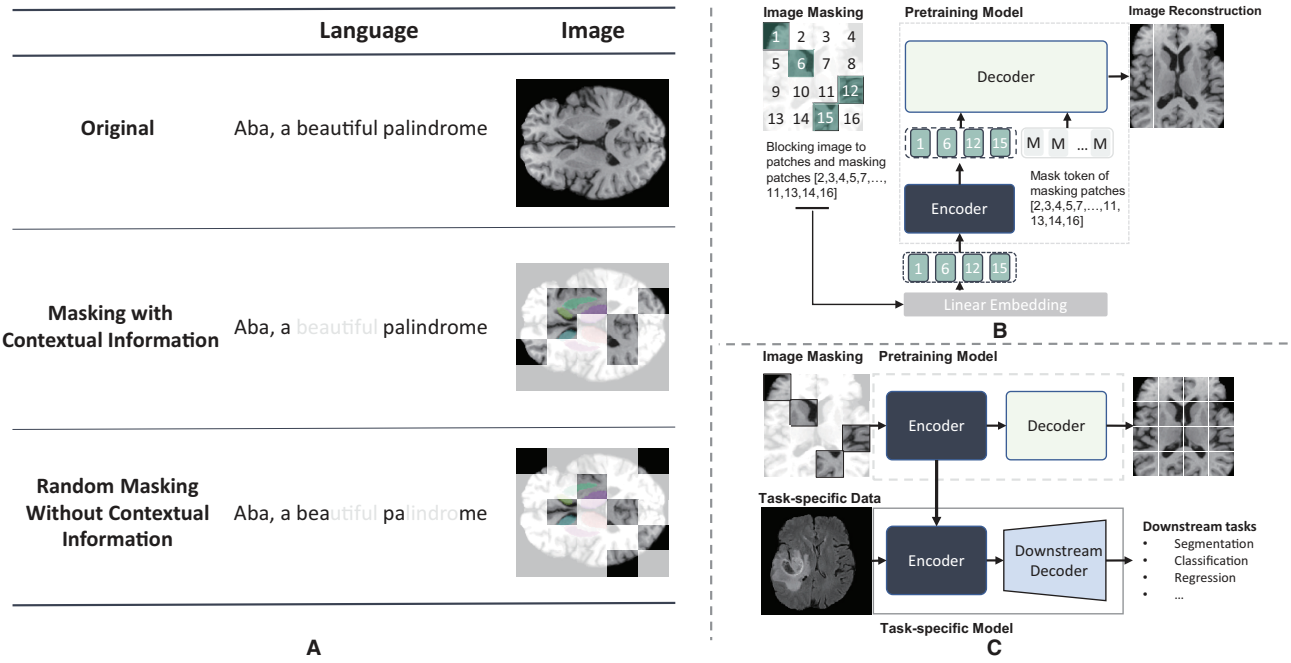


FIGURE 1
Illustration of maksed modeling pre-training and masked image modeling(MIM). **(A)** Different masking result of masked modeling pre-training. The light white regions represents the masked regions and the colored reigons of axial brain MR image represents different independent brain areas. **(B)** The pre-training illustraion of masked autoencoder. **(C)** the whole procedure of mim.

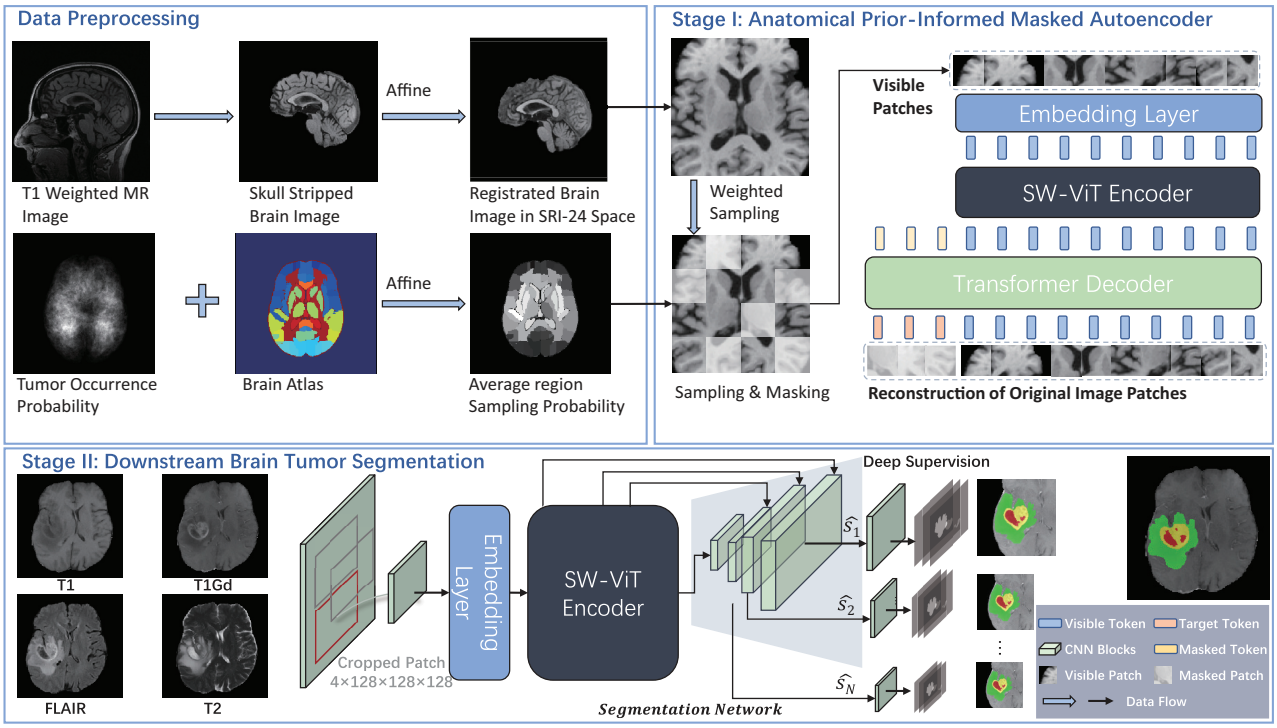


FIGURE 2
Overview of our proposed method.

MRI. As shown in Figure 2, our proposed method consists of two stages: (1) pre-training a masked autoencoder with anatomical prior-informed masking strategy on the unannotated dataset and (2) transferring the pre-trained weights of the encoder and fine-tuning the segmentation network on the annotated dataset.

## 2.2. Statistical analysis of tumor occurrence

### 2.2.1. Registration to standard brain template

To represent the anatomical priors, we first align all images with the standard brain template. The DICOM image data are transformed into Nifti format, and the brain is extracted using FSL tools (37). After that, we transform each image into the SRI-24 standard space (36) *via* affine registration. Using the optimized affine transformation matrix $M^*$, all images are aligned in the SRI-24 space.

$$\begin{aligned} M^* &= \arg\min_{M} C(I_f, \text{Affine}(I_m; M)) \\ I &= \text{Affine}(I_m; M^*) \end{aligned} \tag{1}$$

where $I_m$ represents the moving image, which corresponds to the MRI image of each sample. The fixed image, denoted as $I_f$, refers to the T1 template of the SRI-24 standard space. In this study, the operation $C(I_m, I_f)$ represents the cost function used to quantify disparities between the fixed image $I_m$ and the moving image during the registration optimization process, where a correction ratio is applied (38). The notation Affine($I$; $M$) signifies the affine operation that maps the floating image $I$ to the fixed image using the affine matrix $M$. Moreover, $I$ represents the output registrated image.

### 2.2.2. Sampling weight map derived from brain tumor occurrence

We conduct a statistical analysis of enhanced tumor (ET) across BraTS21 dataset (39–41) and obatin a distribution map of ET occurrence in the SRI-24 standard space. To implement this analysis, we utilize a brain parcellation atlas building upon the parc116plus atlas (36). Some excessively small regions are merged into larger ones, resulting in 128 parcellation regions of the entire skull-stripped brain. To obtain the sampling probability of each voxel, the average sampling probability for each parcellation is defined as follows:

$$P_{R_i} = \frac{\sum_j f_{i,j}}{V_{R_i} \cdot \sum_i \sum_j f_{i,j}} \quad (i = 1, 2, \ldots, 128; j = 1, 2, \ldots, N_{R_i}) \tag{2}$$

where $R_i$ represents the $i$-th brain parcellation, $P_{R_i}$ denotes the average sampling probability per volume of region $R_i$, $f_{i,j}$ is the occurrence frequency of the ET region in the $j$-th voxel within the $i$-th parcellation, $V_{R_i}$ represents the volume of $R_i$, and $N_{R_i}$ represents the number of voxel in $R_i$. Consequently, the sampling weight map $W$, depicted in Figure 3, can be generated by assigning voxels within the parcellation region $R_i$ the identical probability value $P_{R_i}$.

## 2.3. Anatomical prior-informed masked auto-encoder

As shown in Figure 3, our proposed Anatomical Prior-Informed Masked AutoEncoder (API-MAE) consists of five components as follows: (1) Anatomical Prior-informed Masking, (2) Patch embedding, (3) Transformer Encoder, (4) Transformer Decoder, and (5) Discriminator.

### 2.3.1. Anatomical prior-informed masking strategy

Instead of the random masking strategy used in standard MAE pre-training, we propose a dedicated masking strategy to select informative patches based on the derived sampling weight map. The input image $I$ and sampling weights map $W$ are center-cropped with a size of 128, i.e., $I \in \mathbb{R}^{128\times128\times128}$, $W \in \mathbb{R}^{128\times128\times128}$. Subsequently, $I$ and $W$ are transformed into patches represented as $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$ and $\mathcal{W} = \{\mathbf{w}_i\}_{i=1}^n$, respectively. Here, $n$ signifies the quantity of patches, and the patch size is configured at 8, a choice consistent with previous studies (35). This configuration leads to $n = 16 \times 16 \times 16$, aligning with the concept of vision transformers (12) splitting the 2D image into $16 \times 16$ tokens. The sample probability of each patch is determined by the probability vector $\boldsymbol{p} = [p_1, p_2, \ldots, p_n]^\intercal$, where $p_i = \sum_j \mathbf{w}_{i,j} / \sum_{i,j} \mathbf{w}_{i,j}$, and $\mathbf{w}_{i,j}$ denotes the sampling weight of the $j$-th voxel within the $i$-th patch corresponding to the voxels $\mathbf{x}_{i,j}$ of the image patch. Consequently, the visible patches that are fed into the encoder can be sampled as follows:

$$\mathcal{X}_{\text{vis}} = \text{Sampling}(\mathcal{X}, \boldsymbol{p}) \tag{3}$$

where $\mathcal{X}_{\text{vis}} = \{\mathbf{x}_i\}_{i=1}^k$ represents visible patches sampled from the original image patches $\mathcal{X}$, and $k = \eta \cdot n$ represent the number of visible patches, $\eta = 0.25$ is the sampling ratio which aligned with the 75% masking ratio of MAE. The Sampling($\mathcal{X}, \boldsymbol{p}$) operation involves utilizing a multinomial probability distribution with the probability vector $\boldsymbol{p}$ to select tokens from $\mathcal{X}$ for sampling, which then constitute the visible tokens. The sampling procedure is implemented using the multinomial API from PyTorch. As depicted in Figure 4, the prior-informed sampling maintains superior structural consistency compared to random masking, which is advantageous for the calculation of region-based sampling weights.

### 2.3.2. Patch embedding

The input visible patches in $\mathcal{X}_{\text{vis}}$ are first flattened into one-dimensional vectors, then mapped to the feature dimension $D$ *via* learnable patch tokenizer g($\cdot$). The input of the transformer encoder $\mathbf{x}_{\text{enc}}$ is calculated as follows:

$$\mathbf{x}_{\text{enc}} = \text{g}(\mathbf{x}_i) + \text{PE} \in \mathbb{R}^D \tag{4}$$

where $\mathbf{x}_i \in \mathcal{X}_{vis}$, and PE is the sinusoidal positional encoding.

$$\begin{aligned} \text{PE}(pos, 2i) &= \sin\left(\frac{pos}{10000^{2i/D}}\right) \\ \text{PE}(pos, 2i+1) &= \cos\left(\frac{pos}{10000^{2i/D}}\right) \end{aligned} \tag{5}$$

where $pos = 1, 2, \ldots, T$ represents the token position, $i$ represents the $i$-th dimension.

### 2.3.3. Transformer encoder

We adopt a shifted window vision transformer, known as SW-ViT (35), as the transformer encoder in API-MAE. As shown in Figures 3C, D, the multi-head self-attention (MSA) in the original

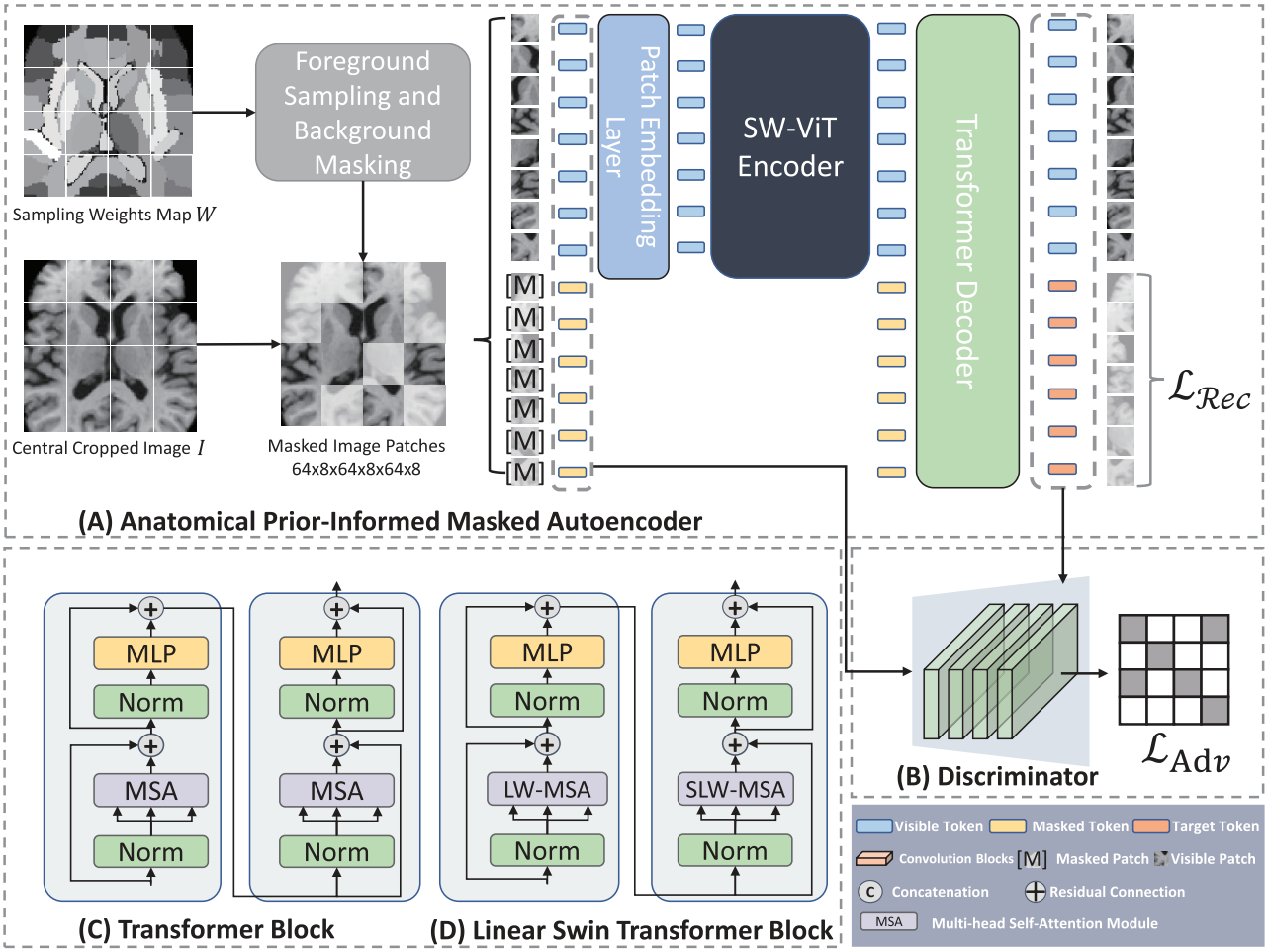


FIGURE 3
Architecture of Anatomical Prior-Informed Masked Autoencoder (API-MAE). **(A)** Is the architecture of Anatomical Prior-Informed Masked Autoencoder. **(B)** Is the Discriminator used for reconstruction. **(C)** If the Transformer block. **(D)** Is the linear Swin Transformer block.

transformer block is replaced with linear window-based multi-head self-attention (LW-MSA) and shifted linear window-based multi-head self-attention (SLW-MSA) in the Swin transformer block. Both LW-MSA and SLW-MSA reduce parameters and computations among each head, which improves the network efficiency without significant accuracy loss. The transformer encoder serves as the feature extractor in API-MAE and the segmentation network. The output of the transformer encoder will undergo a linear projection to fit the higher feature dimension of the transformer decoder.

### 2.3.4. Transformer decoder

We use a shallow transformer decoder to reconstruct the original image in API-MAE. The inputs to the decoder consist of both visible tokens and masked tokens with positional encodings. The output of the decoder is the reconstructed image tokens $\hat{\mathbf{y}}_i$ for each input patch. The reconstruction loss function is the standard L2 loss:

$$\mathcal{L}_{\text{Rec}} = \frac{1}{2}\sum_{i} ||\hat{\mathbf{y}}_i - \mathbf{x}_i||_2, i = 1, 2, \ldots, m \tag{6}$$

where $\mathbf{x}_i$ denotes the $i$-th image patch and $m$ represents the number of masked tokens. It should be noted that only masked tokens are calculated for reconstructed loss.

### 2.3.5. Reconstruction Discriminator

Recent advancements in self-supervised learning, such as DiRA (42), have demonstrated that the collaborative learning of self-supervised and adversarial tasks can lead to a more generalizable representation, encompassing fine-grained semantic representation. Moreover, discriminators have been proven beneficial for the masked autoencoder (32, 43). In API-MAE, we introduced a reconstruction discriminator, envisioning its potential synergistic effect when integrated into MAE decoder. This combination aims to enhance the learning representation and improve visual quality of the reconstructed output. The discriminator is constructed as a shallower convolutional neural

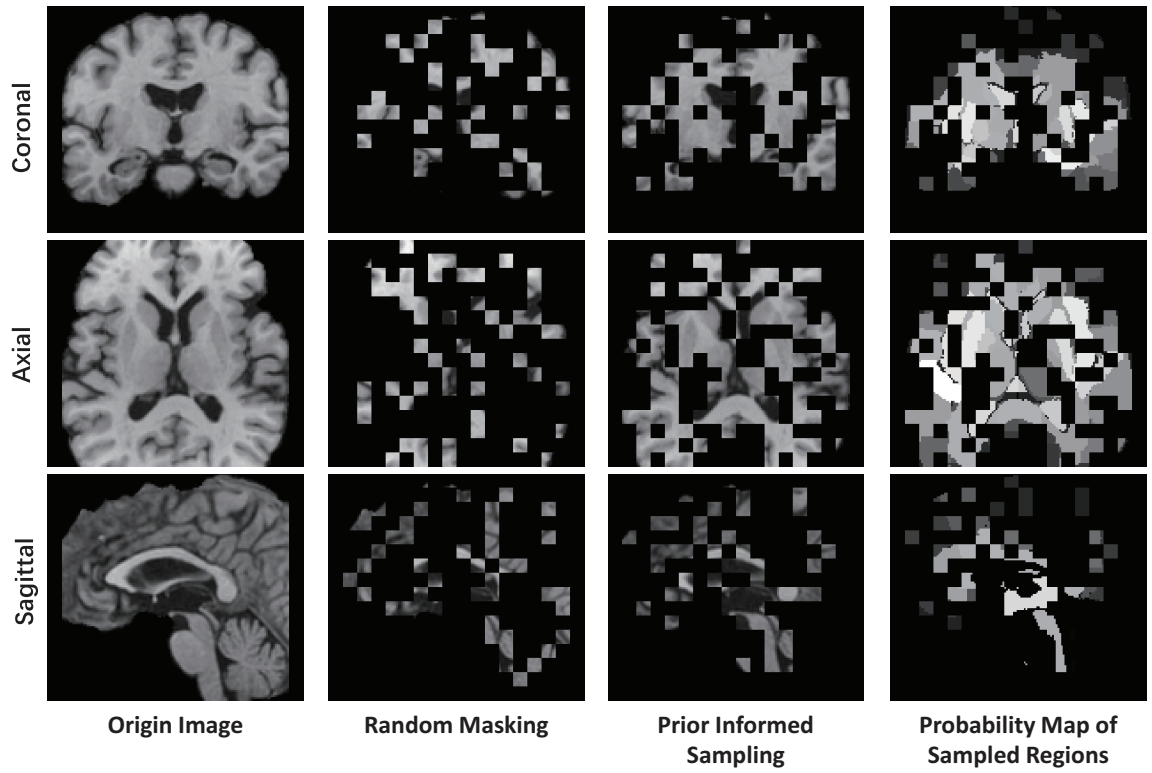


FIGURE 4
Example of visualizing a T1 MR image using different masking strategies with a masking ratio of 0.75. This image is center-cropped with a shape of 128 × 128 × 128, and each token has a patch size of 8 × 8 × 8.

network, comprising five convolutional layers tasked with distinguishing between the reconstructed and real images. The adversarial loss employed for the discriminator is represented as an L2 loss as follows:

$$\mathcal{L}_{\text{Adv}} = \frac{1}{2}\sum_{i}(||\mathcal{D}(\mathbf{x}_i) - 1||_2 + ||\mathcal{D}(\hat{\mathbf{y}}_i)||_2), \quad i = 1, 2, \ldots, n \tag{7}$$

where $\mathbf{x}_i$ is the $i$-th image patch, $\hat{\mathbf{y}}_i$ is the corresponding reconstructed patch, and $n$ is the token number of the original image. Thus, the total loss of API-MAE is a combination of reconstruction loss and adversarial loss as follows:

$$\mathcal{L}_{\text{API-MAE}} = \mathcal{L}_{\text{Rec}} + \mathcal{L}_{\text{Adv}} \tag{8}$$

## 2.4. Segmentation network

After the pre-training of API-MAE, we discard the transformer decoder and keep the transformer encoder for the brain tumor segmentation task. The architecture of the segmentation network is shown in Figure 5. The segmentation network contains three parts as follows: (1) encoder, which contains patch embedding and transformer blocks, (2) encoder propagation, and (3) decoder. The patch embedding layer maps the input multi-parameter MRI (i.e., T1, T1Gd, T2-FLAIR, and T2 image) patches to the embedding features. The transformer blocks share the same architecture and are initialized with the pre-training weight of the transformer encoder in API-MAE. The encoder propagation and decoder parts utilize features from the original image (i.e., $\mathbf{z}_0$) and specific transformer layers (2nd, 4th, 6th, 8th, and last layer, i.e., $\mathbf{z}_2, \mathbf{z}_4, \mathbf{z}_6, \mathbf{z}_8, \mathbf{z}_{12}$) to propagate features and segment the image into three target classes as follows: whole tumor (WT), tumor core (TC), and enhanced tumor (ET). To obtain better segmentation, the segmentation network adopts cross-entropy and Dice loss with deep supervision as the segmentation loss as follows:

$$\mathcal{L}_{\text{Seg}} = \sum_{i=1}^{4} \frac{1}{2^{i-1}} \cdot (\text{CrossEntropy}(S_i, \widehat{S}_i) + \text{Dice}(S_i, \widehat{S}_i)) \tag{9}$$

where $i$ represents the stage of deep supervision, $\widehat{S}_i$ denotes the prediction of stage $i$, and $S_i$ represents the ground truth resized to match the corresponding prediction.

# 3. Experiments

We pre-train the MAE model on an unannotated brain MRI dataset and evaluate the segmentation performance on an annotated brain tumor MRI dataset.

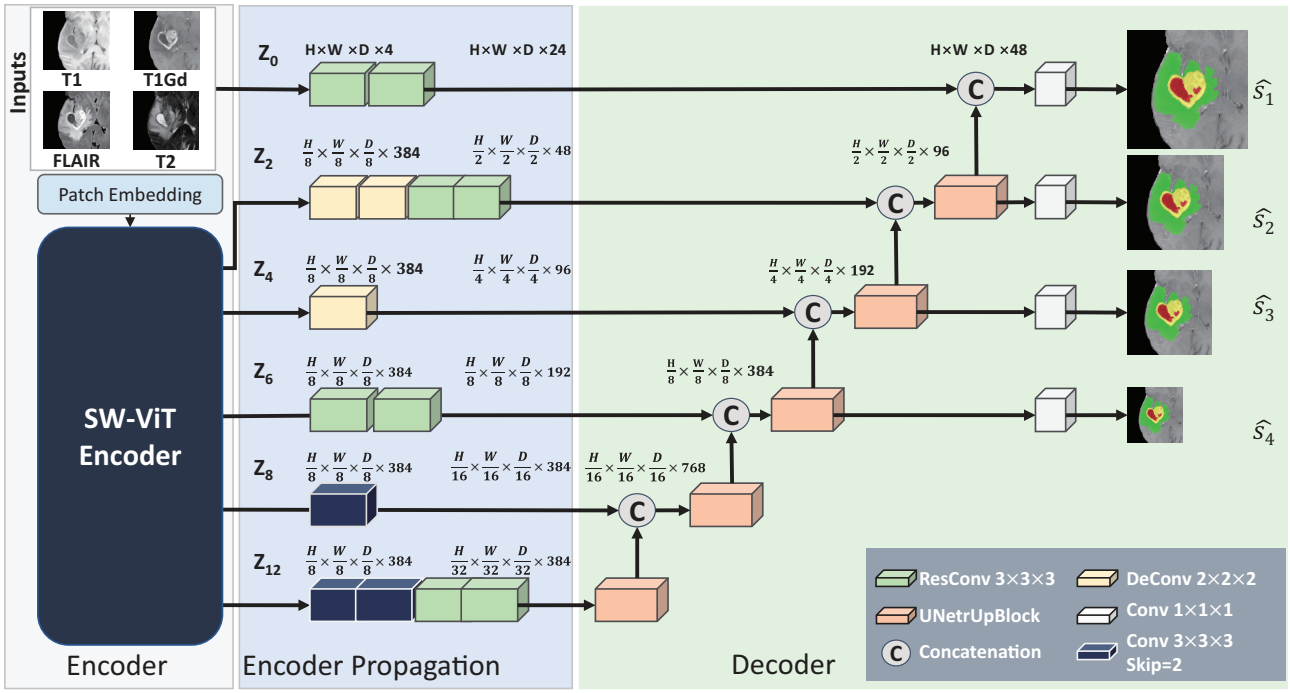


FIGURE 5
Architecture of the baseline segmentation network. This network is made up of three parts, i.e., the Encoder part for feature extraction, the Encoder Propagation part used for channel and spatial normalization and skip connection, and the Decoder parts used for upsampling and predicting the segmentation results. The convolution blocks with skip=2 in the Encoder Propagation part are used for downsampling, and the UNetrUpBlock used in the decoder part is used for upsampling and each block contains a deconvolution block and two residual convolution blocks.

## 3.1. Datasets

### 3.1.1. ADNI dataset

Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (44) is derived from a longitudinal multicenter study aimed at early detection and tracking of Alzheimer's disease (AD). In this study, we collected 7,945 skull-stripped T1 MR images and subsequently handpicked 6,415 images of superior visual quality for utilization in the pre-training dataset. This selection was made following a visual inspection of the registration results.

### 3.1.2. BraTS21 dataset

The BraTS21 dataset (39–41) consists of 1,251 multi-parameter MRI scans. Each case includes four different modalities as follows: a) native (T1), b) post-contrast T1-weighted (T1Gd), c) T2-weighted (T2), and d) T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) images, acquired from various protocols and scanners across multiple institutions. Each scan has been annotated by experienced radiologists with three different subregions as follows: enhancing tumor (ET), peritumoral edematous/invaded tissue (ED), and necrotic tumor core (NCR). In this study, we divide the 1,251 samples into training, validation, and testing sets at a ratio of 7:1:2, following previous studies (35).

## 3.2. Evaluation metrics

Both the volumetric metric dice similarity coefficient (DSC) and surface metric Hausdorff distance (HD) are used for performance evaluation. DSC quantifies the overlap between segmentation results and annotations in voxel space, while the $95_{th}$ percentile of Hausdorff distance (HD95) measures the distances between the segmentation surface and ground-truth surface. The calculation of HD95 is performed by the MedPy package using the analysis framework from nnFormer (45).

## 3.3. Implementation details

**Experimental settings**: All the experiments are implemented using the PyTorch 1.2 framework. We use 4 NVIDIA A100 GPUs (40 GB VRAM) for MAE pre-training and NVIDIA RTX3090 GPU (24 GB VRAM) for segmentation training and inference.

**Data preprocessing**: In the preprocessing section, we employ affine registration to align individual images with the standard space. Here, the cost function during the image registration optimization is correlation ratio (38). To prevent the registration results from being flipped upside down, we defined the rotation search space for affine registration as follows: $[-30°, 30°]$ for X-axis rotation, $[-30°, 30°]$ for Y-axis rotation, and $[-180°, 180°]$ for Z-axis rotation. This configuration is aimed to emphasize rotation in the X-Y plane and prevent upside-down flipping along the

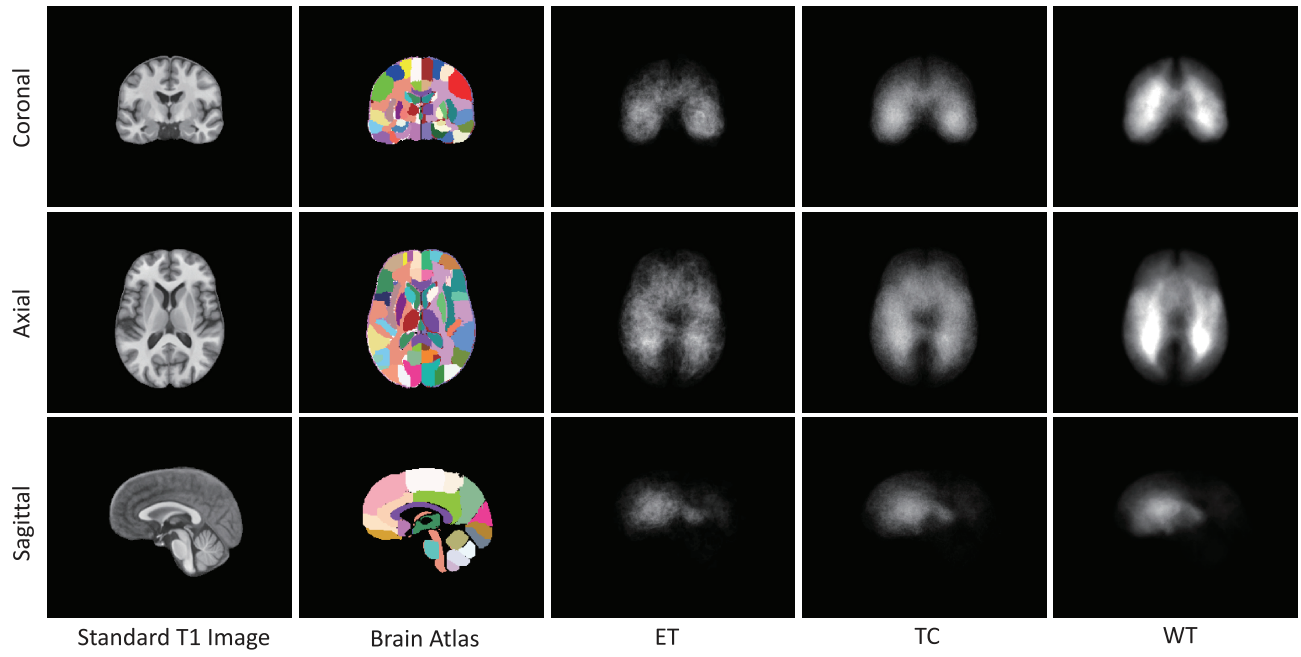


FIGURE 6
The occurrence frequency in SRI-24 standard space among 1,251 cases from BraTS21 dataset. The five columns represent the standard brain T1 MR image, brain atlas (enhanced parc116 plus) in SRI-24 Space, Enhanced Tumor (ET) occurrence, Tumor Core (TC) Tumor occurrence, and Whole Tumor (WT) occurrence, respectively.

Z-axis. It performed effectively with our dataset of 6,415 pre-training samples. The registration optimization and transformation processing were executed using the FLIRT (46) toolbox from FSL. Trilinear interpolation was utilized to compute the intensity of new voxels during affine mapping. For the pre-training data, we employ the MONAI (47) library for data normalization and cropping. Additionally, we utilize the segmentation data preprocessing pipeline provided by nnUNet (7), to handle the multi-modality segmentation data.

**Model architecture**: In API-MAE, the transformer encoder contains 12 layers of linear swin transformer blocks with a feature dimension $D = 384$. The transformer decoder comprises 8 layers of vanilla transformer blocks with a feature dimension of 384. The discriminator consists of four convolution blocks with a kernel size of $k = 3$ and a convolution block with a kernel size of $k = 1$. In the segmentation network, the weights of encoder propagation and decoder parts are initialized with the He initialization (48).

**Model training**: For MAE training, the AdamW optimizer with a batch size of 12 is trained for 300 epochs. The initial learning rate is 1e-3. Weight decay of 5e-2 is also adopted for model regularization. For the segmentation procedure, we apply the (45) training framework and default parameter for 1,000 epochs.

# 4. Results

## 4.1. Pre-training results of anatomical prior-informed MAE

As presented in Figure 6, we note distinct differences in the spatial distribution of tumor occurrence within the SRI24 space. Specifically, gliomas are more frequently observed in the white matter regions of the middle and posterior sections of the brain, with comparatively lower frequencies in the brainstem and cerebellar regions. Table 1 shows the normalized probability of tumor occurrence among all 128 brain parcellations. Considering that ET is the most challenging region to segment, we employ the probability of the ET region for probabilistic masking.

The masking and reconstruction results are shown in Figure 7. It can be observed that random masking tends to distribute masked patches uniformly across the entire image, whereas our proposed weighted sampling strategy enables concentration on more valuable, concentrated, and relatively contiguous regions. The disruption of contextual information in random masking makes the reconstruction task challenging and results in a blurry reconstructed image. In contrast, the proposed weighted sampling method can maintain the integrity of semantic regions, allowing for better reconstruction results.

## 4.2. Segmentation results on BraTS21 dataset

### 4.2.1. Segmentation performance on BraTS21 dataset

To validate the effectiveness of the proposed SSL pre-training approach in downstream segmentation task, we conducted validation experiments using the BraTS21 dataset. The downstream brain tumor segmentation network is initialized with the pre-trained API-MAE encoder weights and subsequently fine-tuned using the BraTS21 dataset. We conducted a comparison of

TABLE 1 The normalized occurrence of tumor regions within different brain parcellations in enhanced SRI-24 atlas analyzed from 1,251 training cases of the BraTS21 dataset.

| Atlas No. | Occurrence (‰) | | | Atlas No. | Occurrence (‰) | | | Atlas No. | Occurrence (‰) | | | Atlas No. | Occurrence (‰) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ET | TC | WT | | ET | TC | WT | | ET | TC | WT | | ET | TC | WT |
| 1 | 4.752 | 6.232 | 7.804 | 33 | 9.670 | 9.221 | 8.695 | 65 | 6.924 | 6.237 | 6.478 | 97 | 0.234 | 0.371 | 0.451 |
| 2 | 6.076 | 7.487 | 9.324 | 34 | 10.018 | 8.920 | 9.074 | 66 | 5.039 | 5.836 | 6.568 | 98 | 0.566 | 0.627 | 0.660 |
| 3 | 4.488 | 5.667 | 6.900 | 35 | 8.562 | 6.641 | 5.224 | 67 | 5.838 | 5.624 | 5.486 | 99 | 0.432 | 0.359 | 0.365 |
| 4 | 7.042 | 7.193 | 7.949 | 36 | 7.492 | 5.694 | 4.343 | 68 | 5.379 | 4.854 | 4.737 | 100 | 0.346 | 0.406 | 0.432 |
| 5 | 4.258 | 4.047 | 4.198 | 37 | 22.961 | 19.311 | 16.840 | 69 | 3.144 | 3.872 | 4.267 | 101 | 0.511 | 0.250 | 0.221 |
| 6 | 2.253 | 3.679 | 3.666 | 38 | 20.812 | 18.402 | 15.837 | 70 | 3.148 | 3.553 | 4.698 | 102 | 0.006 | 0.020 | 0.108 |
| 7 | 5.058 | 5.902 | 7.232 | 39 | 10.437 | 8.829 | 8.304 | 71 | 17.277 | 18.222 | 16.062 | 103 | 0.443 | 0.244 | 0.238 |
| 8 | 7.983 | 8.575 | 8.887 | 40 | 8.959 | 8.447 | 7.602 | 72 | 17.632 | 19.147 | 18.173 | 104 | 0.077 | 0.102 | 0.154 |
| 9 | 4.697 | 4.000 | 4.069 | 41 | 20.511 | 17.402 | 15.713 | 73 | 21.153 | 20.498 | 20.578 | 105 | 0.234 | 0.234 | 0.324 |
| 10 | 4.302 | 4.591 | 4.235 | 42 | 19.540 | 18.639 | 16.355 | 74 | 22.365 | 23.167 | 22.596 | 106 | 0.297 | 0.236 | 0.338 |
| 11 | 10.660 | 11.250 | 12.304 | 43 | 3.381 | 3.224 | 3.248 | 75 | 16.388 | 16.856 | 17.372 | 107 | 0.000 | 0.048 | 0.104 |
| 12 | 12.720 | 13.980 | 14.909 | 44 | 2.699 | 2.668 | 2.584 | 76 | 19.246 | 19.615 | 18.951 | 108 | 0.109 | 0.091 | 0.130 |
| 13 | 6.800 | 6.684 | 7.671 | 45 | 5.139 | 4.690 | 4.523 | 77 | 10.394 | 10.765 | 11.793 | 109 | 0.571 | 0.502 | 0.674 |
| 14 | 8.131 | 8.618 | 8.160 | 46 | 4.073 | 3.794 | 3.657 | 78 | 16.606 | 14.655 | 15.036 | 110 | 0.332 | 0.496 | 0.603 |
| 15 | 8.506 | 7.635 | 7.178 | 47 | 1.932 | 1.842 | 1.673 | 79 | 20.442 | 19.811 | 21.291 | 111 | 0.649 | 0.531 | 0.665 |
| 16 | 6.499 | 6.783 | 6.170 | 48 | 2.241 | 1.949 | 1.706 | 80 | 23.945 | 25.019 | 22.553 | 112 | 0.385 | 0.244 | 0.307 |
| 17 | 13.663 | 14.048 | 16.044 | 49 | 6.592 | 5.891 | 5.982 | 81 | 13.914 | 14.137 | 15.073 | 113 | 0.513 | 0.320 | 0.336 |
| 18 | 17.307 | 17.339 | 17.131 | 50 | 4.727 | 4.686 | 5.243 | 82 | 15.253 | 15.150 | 14.987 | 114 | 0.212 | 0.342 | 0.395 |
| 19 | 3.260 | 4.923 | 5.407 | 51 | 4.827 | 4.206 | 4.233 | 83 | 18.236 | 16.137 | 15.620 | 115 | 17.195 | 15.945 | 14.050 |
| 20 | 3.897 | 4.862 | 5.722 | 52 | 4.906 | 4.636 | 4.816 | 84 | 13.306 | 13.089 | 12.785 | 116 | 8.231 | 8.323 | 9.206 |
| 21 | 7.835 | 8.926 | 9.186 | 53 | 1.949 | 1.836 | 1.902 | 85 | 11.423 | 10.749 | 11.322 | 117 | 6.310 | 7.011 | 9.164 |
| 22 | 7.424 | 9.238 | 8.825 | 54 | 2.332 | 2.017 | 1.918 | 86 | 9.425 | 10.035 | 10.850 | 118 | 6.761 | 7.458 | 9.569 |
| 23 | 5.723 | 7.193 | 6.791 | 55 | 7.666 | 6.745 | 6.341 | 87 | 11.912 | 10.302 | 11.184 | 119 | 19.636 | 18.910 | 16.956 |
| 24 | 8.352 | 8.358 | 8.053 | 56 | 6.645 | 6.265 | 5.671 | 88 | 9.728 | 8.737 | 8.556 | 120 | 21.037 | 20.192 | 18.556 |
| 25 | 5.593 | 5.984 | 6.282 | 57 | 5.545 | 6.162 | 7.175 | 89 | 9.253 | 8.937 | 9.267 | 121 | 2.881 | 7.415 | 6.751 |
| 26 | 5.275 | 5.768 | 6.391 | 58 | 7.017 | 7.076 | 8.576 | 90 | 7.208 | 7.937 | 7.986 | 122 | 3.018 | 7.112 | 6.942 |
| 27 | 3.380 | 3.943 | 3.972 | 59 | 6.488 | 6.475 | 6.568 | 91 | 0.382 | 0.253 | 0.251 | 123 | 15.209 | 15.058 | 16.662 |
| 28 | 2.602 | 3.767 | 3.784 | 60 | 5.566 | 5.746 | 7.283 | 92 | 0.046 | 0.066 | 0.137 | 124 | 16.987 | 16.541 | 18.001 |
| 29 | 22.308 | 22.074 | 20.767 | 61 | 7.063 | 6.733 | 6.843 | 93 | 0.425 | 0.228 | 0.177 | 125 | 7.655 | 7.821 | 7.360 |
| 30 | 22.077 | 23.383 | 20.966 | 62 | 7.861 | 7.588 | 9.150 | 94 | 0.000 | 0.033 | 0.110 | 126 | 7.681 | 8.149 | 7.542 |
| 31 | 15.676 | 14.889 | 13.405 | 63 | 6.903 | 6.783 | 7.793 | 95 | 0.233 | 0.461 | 0.587 | 127 | 3.154 | 3.684 | 3.294 |
| 32 | 17.266 | 15.530 | 13.671 | 64 | 8.235 | 7.420 | 7.827 | 96 | 0.462 | 0.396 | 0.585 | 128 | 3.571 | 4.070 | 3.562 |

The occurrence is expressed in permillage format. ET, enhanced tumor; TC, means tumor core; WT, the whole tumor.



10.3389/fmed.2023.1211800

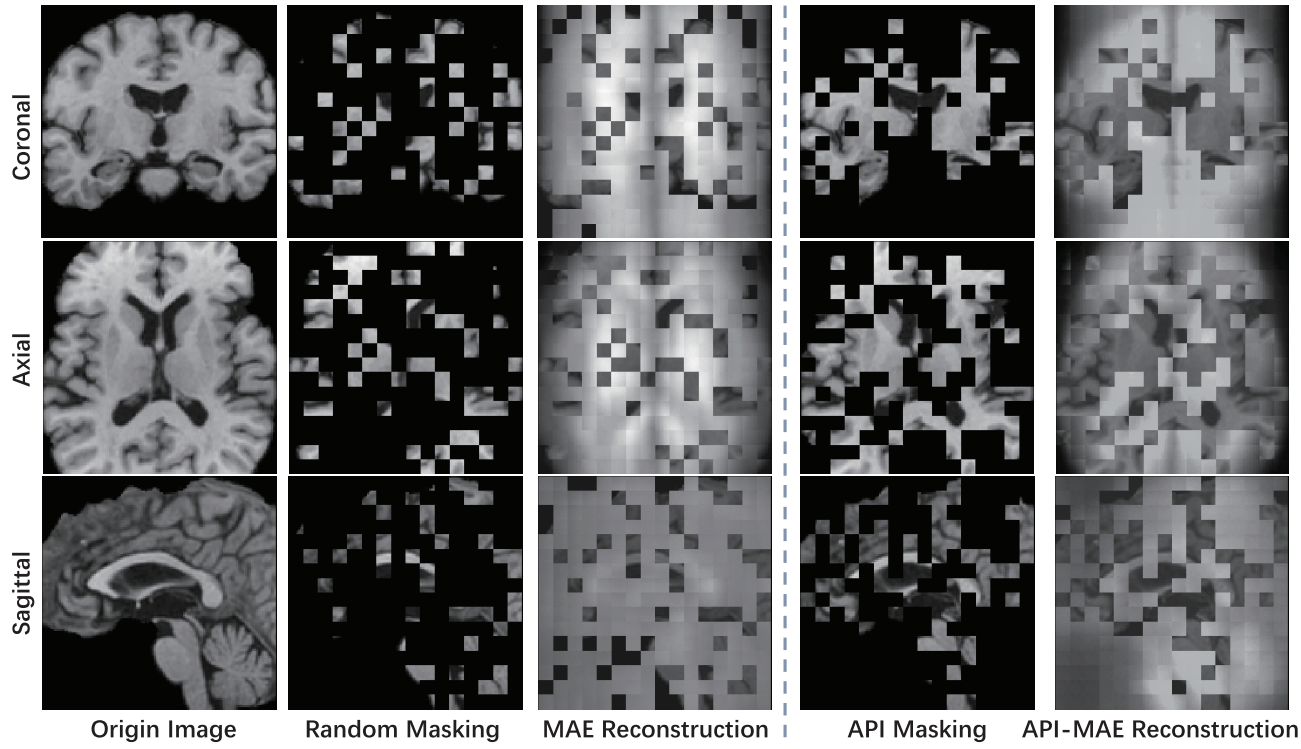


FIGURE 7
The visible example of masking images and the reconstruction results of MAE and API-MAE. The five columns represent the origin brain T1 MR image, the random masking strategy used in MAE, the masking image generated from API token sampling, and the reconstruction results of API-MAE, respectively.

TABLE 2 Efficiency analysis.

| Metric | nnFormer | TransBTS | UNETR | SW-ViT |
|---|---|---|---|---|
| FLOPs (G) | 271.64 | 527.46 | 2141.32 | 860.03 |
| Params (M) | 37.48 | 30.62 | 91.04 | 85.29 |
| CPU inference time (s) | 1.425 | 16.011 | 21.953 | 5.030 |
| GPU inference time (s) | 0.010 | 0.006 | 0.008 | 0.106 |

FLOPs stands for Floating Point Operations, which are recorded in units of gigaflops. Params refers to the learnable parameters of different network architectures, recorded in units of millions. Inference time is computed using an input tensor with dimensions of $2 \times 128 \times 128 \times 128$.

our method against several transformer-based models, including nnFormer (45), TransBTS (16), and UNETR (13) without pre-training. Additionally, we compared against several SSL pre-training methods used in medical imaging, namely, 3D-RPL and 3D-Jig (49), as well as the current state-of-the-art ASA in brain tumor segmentation (35).

As shown in Table 2, we observed that the pre-trained models demonstrate better performance, and our proposed API-MAE achieved the best performance in terms of the Dice similarity coefficient (DSC) metrics for whole tumor (WT) and tumor core (TC) and the best average performance of all three regions.

### 4.2.2. Ablation study on masking strategies

To evaluate the effectiveness of our proposed masking strategy, we conduct an ablation study on different MAE masking strategies. The comparison methods include the baseline without pre-training, MAE pre-trained with random masking, and our proposed API-MAE pre-trained with anatomical prior-informed masking strategy. Table 3 shows that our proposed API-MAE showed improved performance for all regions compared with vanilla MAE and baseline. This demonstrates the effectiveness of our anatomical prior-informed masking compared with the random masking strategy. However, the marginal improvement indicates that in the presence of enough annotated data (more than 1,000 cases in BraTS21), transformer-based models already achieve satisfactory performance, and the benefit of pre-training is not substantial.

### 4.2.3. Data-efficiency analysis

To validate the data efficiency of our pre-trained model, we further train the segmentation model on a small subset of the whole training dataset. We randomly sampled 100 cases from the

TABLE 3 Ablation study on the segmentation performance trained on the BraTS21 dataset.

| Methods | DSC (%)↑ | | | | HD95 (mm)↓ | | | |
|---|---|---|---|---|---|---|---|---|
| | WT | TC | ET | Mean | WT | TC | ET | Mean |
| Baseline | 93.96 | 91.06 | 85.83 | 90.28 | **3.700** | 3.600 | **2.566** | 3.288 |
| MAE | 93.84 | 90.78 | 86.20 | 90.27 | 3.977 | 3.619 | 2.758 | 3.451 |
| Ours | **94.07** | **91.47** | **86.53** | **90.69** | 3.825 | **3.172** | 2.680 | **3.225** |

DSC means the Dice similarity coefficient, and HD95 means the 95th percentile Hausdoff distance. ↑ indicates higher is better and ↓ indicates lower is better. Bold indicates the best performance.

TABLE 4 Comparison of model performance trained with 100 cases sampling from BraTS21 dataset.

| Metric | Methods | WT | TC | ET | Mean |
|---|---|---|---|---|---|
| DSC (%)↑ | Baseline | 91.83 ± 0.16 | 87.43 ± 1.88 | 83.11 ± 1.44 | 87.46 |
| | MAE | **91.96 ± 0.03** | 87.89 ± 0.53 | 83.75 ± 0.59 | 87.87 |
| | API-MAE | 91.95 ± 0.19 | **88.02 ± 0.68** | **84.25 ± 0.67** | **88.07** |
| HD95 (mm)↓ | Baseline | 6.489 ± 0.426 | 5.563 ± 0.377 | 3.985 ± 0.226 | 5.346 |
| | MAE | **6.262 ± 0.358** | 5.513 ± 0.722 | 4.093 ± 0.692 | 5.289 |
| | API-MAE | 6.285 ± 0.694 | **4.979 ± 0.424** | **3.856 ± 0.384** | **5.040** |

Results from four independent sampling processes are reported with mean±std.↑ indicates higher is better and ↓ indicates lower is better. Bold indicates the best performance.

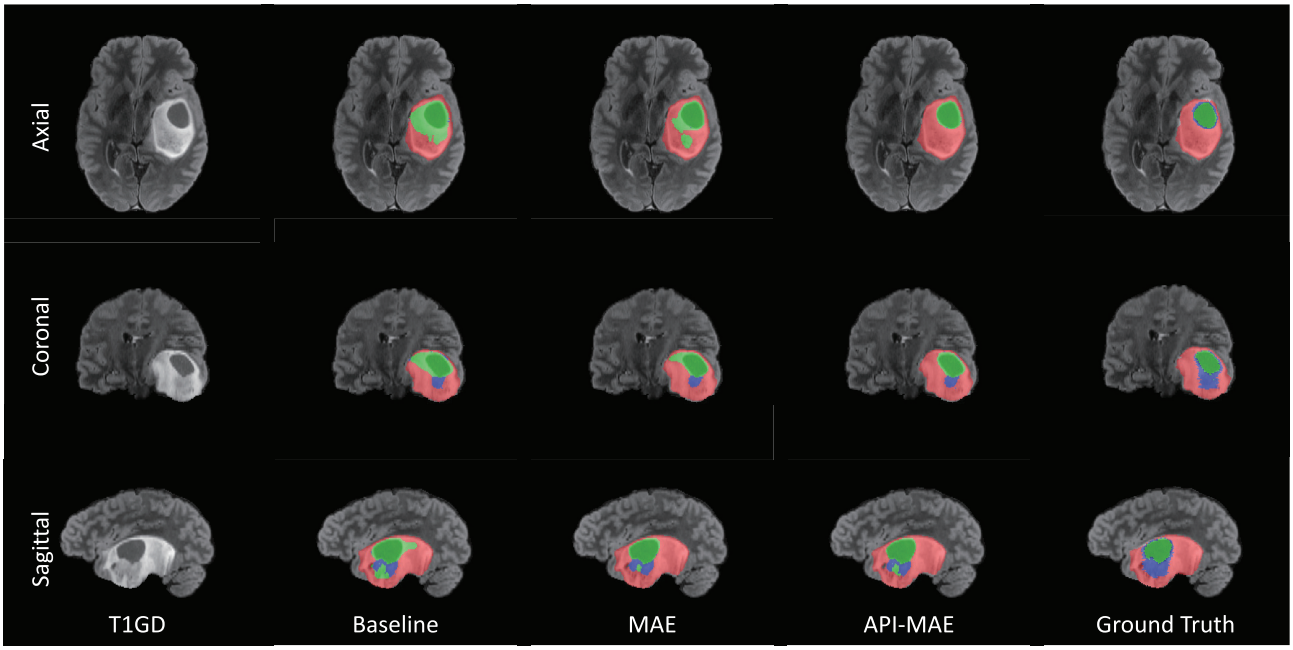


FIGURE 8
Example of tumor segmentation results from a testing image with 100 training cases. The three rows are from the axial, coronal, and sagittal views. The green region represents the necrotic tumor core (NCR), the blue region represents the Gd-enhancing tumor (ET), and the red region represents the peritumoral edematous/invaded tissue (ED).

original training cases, while the validation and testing sets were kept the same as the whole dataset. The compared methods include the baseline without pre-training, MAE pre-trained with random masking, and our proposed API-MAE pre-trained with anatomical prior-informed masking strategy. The sampling process is repeated four times to mitigate the selective bias.

The segmentation results on the small training set are shown in Table 4. It is observed that MAE pre-training benefits the segmentation performance and improves the model robustness in most scenarios. The improvement by pre-training is more prominent in this small-dataset setting compared with the whole dataset. The best segmentation performance for ET and TC regions is obtained by API-MAE, in terms of DSC metrics, which matches the purpose of using ET occurrence map for weighted sampling. As shown in Figure 8, training with the MAE paradigm tends to reduce the erroneous falsely predicted regions and reduce the prediction error of ET regions, particularly in difficult-to-segment regions.

TABLE 5 Comparision results on BraTS21 dataset.

| Methods | DSC (%)↑ | | | | HD95 (mm)↓ | | | |
|---|---|---|---|---|---|---|---|---|
| | WT | TC | ET | Mean | WT | TC | ET | Mean |
| nnFormer (45) | 91.46 | 87.42 | 82.22 | 87.03 | 10.15 | 9.59 | 16.78 | 12.17 |
| TransBTS (16) | 92.06 | 88.20 | 79.46 | 86.57 | 4.98 | 4.86 | 16.32 | 8.72 |
| UNETR (13) | 92.12 | 88.32 | 79.61 | 86.68 | 4.91 | 4.67 | 16.32 | 8.63 |
| 3D-RPL (49) | 93.92 | 90.13 | 85.92 | 89.99 | 3.74 | 3.98 | 13.71 | 7.14 |
| 3D-Jig (49) | 93.87 | 90.14 | 86.01 | 90.01 | 3.85 | 3.94 | 11.79 | 6.53 |
| ASA (35) | 94.03 | 90.29 | **86.76** | 90.36 | **3.61** | 3.78 | 10.25 | 5.88 |
| Ours | **94.07** | **91.47** | 86.53 | **90.69** | 3.82 | **3.17** | **2.68** | **3.23** |

DSC means the Dice similarity coefficient, and HD95 means the 95th percentile Hausdoff distance. ↑ indicates higher is better and ↓ indicates lower is better. Bold indicates the best performance and the results of previous studies are adopted from (35).

To further investigate the efficiency of proposed method, we conducted an efficiency analysis of the segmentation phase for the methods, as shown in Table 5. Since different SSL methods share the same segmentation network, specifically SW-ViT, the variations in performance arise from the encoder weights inherited from diverse SSL pre-training tasks. This comparison involves distinct network architectures, namely, nnFormer, TransBTS, UNETR, and SW-ViT. All the methods were reproduced using the original code on a local server equipped with an AMD Ryzen 9 5900X CPU (3.7 GHz), 128 GB RAM (DDR4 2400MT/s), and an NVIDIA RTX3090 GPU. For fair comparison, we modified UNETR by adjusting its input channels to 4 and configuring the patch size as $8 \times 8 \times 8$, in alignment with SW-ViT. The computation consumption was calculated utilizing the thop package. This process entails inputting a tensor with dimensions of $2 \times 128 \times 128 \times 128$ into the network for computation and the standard segmentation procedure.

Combining the data from Tables 2, 4, we observe that nnFormer exhibits the best inference efficiency. This superiority can be attributed to the dimension of the embedding feature in the Transformer module of the network, which is [96, 192, 384, 768]. In contrast, other Transformer models often have embedding feature dimensions of 384 or 768. This relatively shallower transformer architecture contributes to its enhanced computational efficiency. However, it may result in slightly lower segmentation performance. Higher segmentation accuracy can be achieved in both WT and TC components in models with increased transformer layers. However, when using a high-layer transformer encoder such as UNETR, the number of floating point operations (FLOPs) and learnable parameters will increase rapidly. While the SW-ViT could reduce the FLOPs and parameters with the help of shifted window-based linear transformer modules. Enhanced with SSL pre-training tasks, particularly our proposed API-MAE, the methods using SW-ViT obtain the best segmentation performance while maintaining a favorable balance in terms of segmentation time consumption. Due to the presence of certain operations within the network architecture that do not parallelize efficiently during GPU computation, the proposed method does not achieve optimal computational efficiency on the GPU. However, the proposed method could attain decent CPU time consumption, which maintains a reasonable balance between accuracy and efficiency.

## 5. Discussion

Recently, transformer-based models have emerged as state-of-the-art methods for 3D medical image segmentation, owing to their superiority in modeling long-range dependencies and leveraging global contextual information over fully convolutional neural networks. However, such methods often rely on a vast of training data for network optimization. A major challenge in training such models is the limited availability of annotated data. In this study, we address this challenge by utilizing 6,415 unannotated T1-weighted MR images from the ADNI dataset for pre-training. Our approach consistently improved the segmentation accuracy in scenarios with both large and small training sets. Although only T1-weighted images are used for pre-training, the learned weights benefit the downstream brain tumor segmentation on multi-parameter MRI. This highlights the potential of pre-training for improved medical image segmentation.

The MAE used in computer vision typically employs random masking with a high masking ratio of 0.75 and utilizes 25% unmasked patches for encoder training. The high masking ratio can lead to the loss of contextual information in high-dimensional medical images, making image reconstruction challenging and potentially affecting the learning of generalizable features. Therefore, it is important to consider tailored sampling strategies that take into account the specific characteristics and requirements of the task at hand. In this study, we introduce an anatomical prior-informed masking strategy, where brain regions with higher tumor occurrence are more frequently sampled for pre-training. The experiments demonstrate that our proposed pre-training method enhances the performance of brain tumor segmentation, which outperforms other self-learning approaches. This indicates that incorporating anatomical priors into the pre-training stage leads to performance improvements in downstream tasks.

Additionally, our anatomical prior-informed sampling strategy can be considered as an attention mechanism in selecting valuable and task-related patches for MAE pre-training. In general, attention mechanisms usually help models filter out high-value information from large amount of data, thereby improving computational efficiency and performance and making computing more precise and efficient. Given a large number of image patches in the unannotated dataset, it is important to let the pre-training process

attend the informative patches. By incorporating the tumor occurrence rate and brain template into the construction of an attentive sampling strategy, our approach integrates anatomical priors with masked image modeling pre-training. This enables efficient sampling and the most use of unannotated data.

There are some limitations of this study. Our proposed method requires the pre-registration of the sampling weighting map for each individual, a process typically executed on the CPU and incurring a time cost. In future study, this procedure can be expedited through the utilization of deep learning-based networks, enabling accurate and rapid registration. We showcase the advantage of integrating anatomical priors during the pre-training stage, leveraging only tumor occurrence information. In future, the exploration of more advanced anatomical priors, such as symmetric brain structure or active learning strategies (50), holds potential for further investigation.

## 6. Conclusion

In this study, we introduce a novel pre-training technique for brain tumor segmentation utilizing transformer networks. This technique involves the integration of an anatomical prior-informed masking strategy into the masked image modeling process. Informative image patches from brain parcellations with higher tumor occurrence are sampled more frequently, facilitating the mask autoencoder to focus on the regions of interest. The proposed approach demonstrates promising performance in the brain tumor segmentation task, surpassing compared self-learning methods.

## Data availability statement

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

## Author contributions

KW, HW, MW, and SW: main idea and study design. ZL and MP: organized the database, data inspection, and analysis. KW, ZL, and SW wrote the first draft of the manuscript. SL: built the initial code of the network training framework. MW, SW, and ZS: supervised, supported, and revised the manuscripts. All authors contributed to the manuscript revision, read, and approved the submitted version.

## Funding

This work was supported by National Natural Science Foundation of China under Grant 82072021 and SW was supported by Shanghai Sailing Programs of Shanghai Municipal Science and Technology Committee (22YF1409300).

## 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. Janjua TI, Rewatkar P, Ahmed-Cox A, Saeed I, Mansfeld FM, Kulshreshtha R, et al. Frontiers in the treatment of glioblastoma: past, present and emerging. *Adv Drug Deliv Rev*. (2021) 171:108–38. doi: 10.1016/j.addr.2021.01.012

2. Li C, Wang S, Yan JL, Piper RJ, Liu H, Torheim T, et al. Intratumoral heterogeneity of glioblastoma infiltration revealed by joint histogram analysis of diffusion tensor imaging. *Neurosurgery*. (2019) 85:524–34. doi: 10.1093/neuros/nyy388

3. Li C, Wang S, Liu P, Torheim T, Boonzaier NR, van Dijken BR, et al. Decoding the interdependence of multiparametric magnetic resonance imaging to reveal patient subgroups correlated with survivals. *Neoplasia*. (2019) 21:442–9. doi: 10.1016/j.neo.2019.03.005

4. Wang S, Dai C, Mo Y, Angelini E, Guo Y, Bai W. Automatic brain tumour segmentation and biophysics-guided survival prediction. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part II 5*. Cham: Springer. (2020) p. 61–72.

5. Zhao X, Wu Y, Song G, Li Z, Zhang Y, Fan Y, et al. deep learning model integrating FCNNs and CRFs for brain tumor segmentation. *Med Image Anal*. (2018) 43:98–111. doi: 10.1016/j.media.2017.10.002

6. 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.

7. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. (2021) 18:203–11. doi: 10.1038/s41592-020-01008-z

8. Futrega M, Milesi A, Marcinkiewicz M, Ribalta P. Optimized U-Net for brain tumor segmentation. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part II*. Cham: Springer. (2022) p. 15–29.

9. Azad R, Heidari M, Wu Y, Merhof D. Contextual attention network: transformer meets U-net. In: Lian C, Cao X, Rekik I, Xu X, Cui Z, editors. *Machine Learning in Medical Imaging*. Cham: Springer Nature Switzerland (2022). p. 377–86. doi: 10.1007/978-3-031-21014-3_39

10. Wu H, Pan J, Li Z, Wen Z, Qin J. Automated skin lesion segmentation via an adaptive dual attention module. *IEEE Trans Med Imaging*. (2021) 40:357–70. doi: 10.1109/TMI.2020.3027341

11. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems*. (2017) p. 6000–10.

12. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: *International Conference on Learning Representations*. Ithaca, NY (2021). Available online at: https://openreview.net/forum?id=YicbFdNTTy

13. Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, et al. Unetr: Transformers for 3d medical image segmentation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. Waikoloa, HI: IEEE (2022) p. 574–584.

14. Lin A, Chen B, Xu J, Zhang Z, Lu G, Zhang D. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Trans Instrum Meas*. (2022) 71:1–15. doi: 10.1109/TIM.2022.3178991

15. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I*. Cham: Springer. (2022) p. 272–284.

16. Wang W, Chen C, Ding M, Yu H, Zha S, Li J. Transbts: Multimodal brain tumor segmentation using transformer. In: *Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part I 24*. Cham: Springer. (2021) p. 109–119.

17. Dobko M, Kolinko DI, Viniavskyi O, Yelisieiev Y. Combining CNNs with transformer for multimodal 3D MRI brain tumor segmentation. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part II*. Cham: Springer. (2022) p. 232–241. doi: 10.1007/978-3-031-09002-8_21

18. Liu X, Zhang F, Hou Z, Mian L, Wang Z, Zhang J, et al. Self-supervised learning: generative or contrastive. *IEEE Trans Knowl Data Eng*. (2021) 35:857–76. doi: 10.1109/TKDE.2021.3090866

19. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv*. (2018). doi: 10.48550/arXiv.1810.04805

20. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. In: *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeuraIPS)*. MIT Press (2020). p. 1877–901.

21. Chung YA, Zhang Y, Han W, Chiu CC, Qin J, Pang R, et al. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In: *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Cartagena: IEEE. (2021) p. 244-250.

22. Bao H, Dong L, Piao S, Wei F. Beit: Bert pre-training of image transformers. *arXiv*. (2021). doi: 10.48550/arXiv.2106.08254

23. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, LA: IEEE (2022) p. 16000–16009.

24. Feichtenhofer C, Fan H, Li Y, He K. Masked autoencoders as spatiotemporal learners. In: *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeuraIPS)*. New Orleans, LA: MIT Press (2022).

25. Yu X, Tang L, Rao Y, Huang T, Zhou J, Lu J. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, LA: IEEE (2022) p. 19313–22.

26. Chen X, Ding M, Wang X, Xin Y, Mo S, Wang Y, et al. Context autoencoder for self-supervised representation learning. *arXiv*. (2022). doi: 10.48550/arXiv.2202.03026

27. Tong Z, Song Y, Wang J, Wang L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv*. (2022). doi: 10.48550/arXiv.2203.12602

28. Tang Y, Yang D, Li W, Roth HR, Landman B, Xu D, et al. Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA: IEEE (2022) p. 20730–40.

29. Chen Z, Agarwal D, Aggarwal K, Safta W, Balan MM, Brown K. Masked image modeling advances 3D medical image analysis. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, HI: IEEE (2023) p. 1970–80.

30. Zhou L, Liu H, Bae J, He J, Samaras D, Prasanna P. Self pre-training with masked autoencoders for medical image analysis. *arXiv*. (2022). doi: 10.48550/arXiv.2203.05573

31. Bandara WGC, Patel N, Gholami A, Nikkhah M, Agrawal M, Patel VM. AdaMAE: adaptive masking for efficient spatiotemporal learning with masked autoencoders. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC: IEEE (2023). p. 14507–17.

32. Chen H, Zhang W, Wang Y, Yang X. Improving masked autoencoders by learning where to mask. *arXiv*. (2023). doi: 10.48550/arXiv.2303.06583

33. Qing Z, Zhang S, Huang Z, Wang X, Wang Y, Lv Y, et al. Mar: masked autoencoders for efficient action recognition. *IEEE Trans Multimed*. (2023) 1–16. doi: 10.1109/TMM.2023.3263288

34. Xu H, Ding S, Zhang X, Xiong H, Tian Q. Masked autoencoders are robust data augmentors. *arXiv*. (2022). doi: 10.48550/arXiv.2206.04846

35. Huang J, Li H, Li G, Wan X. Attentive symmetric autoencoder for brain MRI segmentation. In: *Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, September 18-22, 2022, Proceedings, Part V*. Cham: Springer. (2022) p. 203–213.

36. Rohlfing T, Zahr NM, Sullivan EV, Pfefferbaum A. The SRI24 multichannel atlas of normal adult human brain structure. *Hum Brain Mapp*. (2010) 31:798–819. doi: 10.1002/hbm.20906

37. Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. Fsl. *Neuroimage*. (2012) 62:782–90. doi: 10.1016/j.neuroimage.2011.09.015

38. Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. *Med Image Anal*. (2001) 5:143–56. doi: 10.1016/S1361-8415(01)00036-6

39. Baid U, Ghodasara S, Mohan S, Bilello M, Calabrese E, Colak E, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv*. (2021). doi: 10.48550/arXiv.2107.02314

40. Bjoern HM, Andras J, Stefan B, Jayashree KC, Keyvan F, Justin K, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging*. (2015) 34:1993–2024. doi: 10.1109/TMI.2014.2377694

41. Lloyd CT, Sorichetta A, Tatem AJ. High resolution global gridded data for use in population studies. *Scient Data*. (2017) 4:1–17. doi: 10.1038/sdata.2017.1

42. Haghighi F, Taher MRH, Gotway MB, Liang J. DiRA: discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE. (2022) p. 20792–802.

43. Fei Z, Fan M, Zhu L, Huang J, Wei X, Wei X. Masked auto-encoders meet generative adversarial networks and beyond. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC: IEEE (2023) p. 24449–59.

44. Jack Jr CR, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J Magnet Reson*. (2008) 27:685–91. doi: 10.1002/jmri.21049

45. Zhou HY, Guo J, Zhang Y, Han X, Yu L, Wang L, et al. nnFormer: volumetric medical image segmentation via a 3D transformer. *IEEE Trans Image Proc*. (2023) 32:4036–45. doi: 10.1109/TIP.2023.3293771

46. Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*. (2002) 17:825–41. doi: 10.1006/nimg.2002.1132

47. MONAI Consortium. *MONAI: Medical Open Network for AI*. (2020). Available online at: https://github.com/Project-MONAI/MONAI (accessed April 24, 2023).

48. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *Proceedings of the IEEE International Conference on Computer Vision*. Santiago: IEEE (2015) p. 1026–1034.

49. Taleb A, Loetzsch W, Danz N, Severin J, Gaertner T, Bergner B, et al. 3d self-supervised methods for medical imaging. In: *Proceedings of the Conference on Advances in neural information processing systems (NeuraIPS)*. MIT Press (2020). p. 1815872.

50. Dai C, Wang S, Mo Y, Angelini E, Guo Y, Bai W. Suggestive annotation of brain MR images with gradient-guided sampling. *Med Image Anal*. (2022) 77:102373. doi: 10.1016/j.media.2022.102373

frontiers | Frontiers in Cardiovascular Medicine

TYPE Methods
PUBLISHED 22 September 2023
DOI 10.3389/fcvm.2023.1266260



OPEN ACCESS

EDITED BY
Gongning Luo,
Harbin Institute of Technology, China

REVIEWED BY
Guanglu Sun,
Harbin University of Science and Technology, China
Haiwei Pan,
Harbin Engineering University, China
Amir Faisal,
Sumatra Institute of Technology, Indonesia

*CORRESPONDENCE
Zhaowen Qiu
qiuzw@nefu.edu.cn
Yan Li
wemn@sina.com
Caijuan Li
caijuanli123@163.com

†These authors have contributed equally to this work and share first authorship

RECEIVED 25 July 2023
ACCEPTED 12 September 2023
PUBLISHED 22 September 2023

CITATION
Guo Z, Zhang Y, Qiu Z, Dong S, He S, Gao H, Zhang J, Chen Y, He B, Kong Z, Qiu Z, Li Y and Li C (2023) An improved contrastive learning network for semi-supervised multi-structure segmentation in echocardiography. Front. Cardiovasc. Med. 10:1266260. doi: 10.3389/fcvm.2023.1266260



# An improved contrastive learning network for semi-supervised multi-structure segmentation in echocardiography

Ziyu Guo[1†], Yuting Zhang[2†], Zishan Qiu[3], Suyu Dong[1], Shan He[2], Huan Gao[1], Jinao Zhang[1], Yingtao Chen[1], Bingtao He[1], Zhe Kong[1], Zhaowen Qiu[1*], Yan Li[1*] and Caijuan Li[4*]

[1]College of Computer and Control Engineering, Northeast Forestry University, Harbin, China, [2]School of Computer Science, University of Birmingham, Birmingham, United Kingdom, [3]College of Art and Science, New York University Shanghai, Shanghai, China, [4]Department of Medical Ultrasonics, Hongqi Hospital of Mudanjiang Medical University, Mudanjiang, China

Cardiac diseases have high mortality rates and are a significant threat to human health. Echocardiography is a commonly used imaging technique to diagnose cardiac diseases because of its portability, non-invasiveness and low cost. Precise segmentation of basic cardiac structures is crucial for cardiologists to efficiently diagnose cardiac diseases, but this task is challenging due to several reasons, such as: (1) low image contrast, (2) incomplete structures of cardiac, and (3) unclear border between the ventricle and the atrium in some echocardiographic images. In this paper, we applied contrastive learning strategy and proposed a semi-supervised method for echocardiographic images segmentation. This proposed method solved the above challenges effectively and made use of unlabeled data to achieve a great performance, which could help doctors improve the accuracy of CVD diagnosis and screening. We evaluated this method on a public dataset (CAMUS), achieving mean Dice Similarity Coefficient (DSC) of 0.898, 0.911, 0.916 with 1/4, 1/2 and full labeled data on two-chamber (2CH) echocardiography images, and of 0.903, 0.921, 0.928 with 1/4, 1/2 and full labeled data on four-chamber (4CH) echocardiography images. Compared with other existing methods, the proposed method had fewer parameters and better performance. The code and models are available at https://github.com/gpgzy/CL-Cardiac-segmentation.

KEYWORDS
echocardiography, deep learning, semi-supervised learning, images semantic segmentation, contrastive learning

## 1. Introduction

Cardiovascular diseases (CVDs) are increasing threats to global health and have become the leading cause of death in industrialized countries (1). The American Society of Echocardiography (ASE) and the European Association of Cardiovascular Imaging (EACVI) have emphasized the importance of cardiac chamber quantification by echocardiography in the diagnosis and treatment of cardiovascular diseases (2). Echocardiography is one of the most widely utilized diagnostic tests in cardiology, offering clear visualizations of left ventricular size during end systole and end diastole, along with the thickness of the myocardium (3). In echocardiography, a heart afflicted by disease might display enlarged atrial and ventricular volumes or an augmented thickness

of the myocardium (4). However, these chamber quantifications, such as chamber sizes, volumes, and etc., are usually based on precise segmentation of certain critical structures (such as ventricle, atrium and myocardium), and then measuring key metrics that indicate the heart's functionality (2, 5). Practically, this process requires cardiologists to manually describe the anatomy and takes measurements of relevant biological parameters, which can be tedious, time-consuming, and subjective (5). Therefore, there exists a genuine requirement in clinical settings for an efficient and precise automated echocardiographic segmentation technique that can enhance the efficacy and reduce the burden of the physician in clinical imaging screening, track disease progression and make informed decisions about treatment and intervention.

There are two main categories of automated segmentation techniques for cardiac structures: traditional techniques and neural network-based methods. Traditional methods include contour models (6), level sets (7), and atlas-based methods (8). Barbosa et al. (6) put forward a B-spline active contour formulation that employs explicit functions for real-time segmentation of 3D echocardiography and liver computer tomography. This method overcomes the limitations of the initial Active Geometric Functions (AGF) framework introduced by Real-time segmentation by Active Geometric Functions while preserving computational speed. Yang et al. (7) proposed a two-layer level set method along with a circular shape constraint to segment the left ventricle (LV) from short-axis cardiac magnetic resonance images (CMRI) without relying on any pre-trained models. This technique can be applied to other level set methods and effectively addresses common issues in LV segmentation, such as intensity overlap between Trabeculations and Papillary Muscles (TPM) and the myocardium, and the existence of outflow track in basal slices. Zhuang et al. (8) developed a fully automated framework for whole-heart segmentation that relies on the locally affine registration method (LARM) and free-form deformations with adaptive control point status (ACPS FFDs) for automatic segmentation of CMRI. However, these methods are unable to surmount the challenges of low contrast and noise that are inherent in echocardiography. As a result, they are unable to produce accurate segmentation results based on echocardiography.

Neural network-based methods have demonstrated enhanced segmentation accuracy in echocardiography as well, and they can be classified further as supervised methods and semi-supervised methods. Cui et al. (9) proposed a multitask model with Task Relation Spatial Co-Attention for joint segmentation and quantification on 2D echocardiography. This method integrated the Boundary-aware Structure Consistency (BSC) and Joint Indices Constraint (JIC) into the multitask learning optimization objective to guide the learning of segmentation and quantification paths. It was validated on the CAMUS dataset and demonstrated outstanding performance, achieving an overall mean Dice score of 0.912 and 0.923, as well as average precision scores of 0.931 and 0.941 for the two-chamber views (A2C) and apical four-chamber views (A4C). Cui et al. (10) utilized a training strategy named multi-constrained aggregate learning (MCAL) for the segmentation of myocardium in 2D echocardiography. This method leveraged anatomical knowledge learned through ground-truth labels to infer segmented parts and discriminate boundary pixels. It was validated on CAMUS dataset and had performance with segmentation Dice of $0.853 \pm 0.057$ and $0.859 \pm 0.560$ for the apical A4C and A2C views, respectively. Hamila et al. (11) proposed a novel convolution neural network (CNN) that combines denoising and feature extraction techniques for automatic LV segmentation of echocardiography. 2D echocardiographic images from 70 patients were used to train this network, and it was then tested on 12 patients, achieving a segmentation Dice of 0.937. While these supervised methods can achieve excellent performance, they all require a sufficient number of pixel-wise annotations to train the model, which can be a time-consuming and tedious process.

Semi-supervised methods have shown effectiveness in reducing the need for a large number of annotated samples in echocardiography segmentation. Wu et al. (5) integrated a novel adaptive spatiotemporal semantic calibration module into the mean teacher semi-supervised architecture to determine spatiotemporal correspondences based on feature maps for echocardiography segmentation. The proposed method was evaluated using the EchoNet-Dynamic and CAMUS datasets, resulting in average Dice coefficients of 0.929 and 0.938, respectively, for the segmentation of the left ventricular endocardium. Additionally, based on these two datasets, El Rai et al. (12) presented a new semi-supervised approach called GraphECV for the segmentation of the LV in echocardiography by using graph signal processing, respectively resulting in Dice coefficients of 0.936 and 0.940 with 1/2 labeled data for the left ventricular segmentation. Wei et al. (13) used a co-learning mechanism to explore the mutual benefits of cardiac segmentation, therefore alleviating the noisy appearance. It was validated on the training set of CAMUS dataset using 10-fold cross-validation, achieving a Dice of 0.923, 0.948 and 0.895 for the segmentation of LV, myocardium and left atrium (LA). Chen et al. (14) proposed a framework for cross-domain echocardiography segmentation that incorporated multi-space adaptation-segmentation-joint based on a generative adversarial architecture with a generator and multi-space discriminators. The CAMUS dataset was used to evaluate this method, and the experiments show that this method attained the mean Dice coefficients of 0.890 for the segmentation of LV endocardium and LV epicardium. However, only a few semi-supervised methods have been used for the segmentation of the LV, LA, and myocardium, and although some methods attempt to segment all three regions simultaneously, the accuracy of multi-structure segmentation remains to be improved, especially for the LV and LA segmentation.

To help improve the accuracy of diagnosing and screening for cardiovascular diseases while also easing the workload associated with evaluating LV images. In this paper, we proposed a novel semi-supervised method for multi-structure segmentation of echocardiography, and the main contributions could be summarized as follows:

(1) We first applied contrastive learning in the multi-structure segmentation of echocardiography, and could accurately

segment LV, LA and myocardium without requiring a large number of annotated samples, which explored the feasibility of contrastive learning in echocardiography multi-structure segmentation.

(2) We made two improvements to the existing model, building upon the work by Lai et al. (15): replacing DeeplabV3+ (16) with u-net (17) and modifying the structure of the projector. These changes aimed to tackle challenges in echocardiography, such as low contrast, unclear boundaries, and incomplete cardiac structures.

(3) Our method was evaluated on the CAMUS dataset, and it demonstrated excellent performance in both two-chamber (2CH) and four-chamber (4CH) echocardiographic images.

# 2. Method

## 2.1. Overview

The cardiac structure segmentation network in this study was built upon the contrastive learning, taking inspiration from the work of Lai et al. (15) (**Figure 1**). The proposed network consists of two branches: a supervised branch and an unsupervised branch. The supervised branch was first trained with labeled data to acquire basic features within the echocardiographic images. Next, these parameters were shared with the unsupervised branch, which was then continually optimized with unlabeled data. More details about the supervised branch, the unsupervised

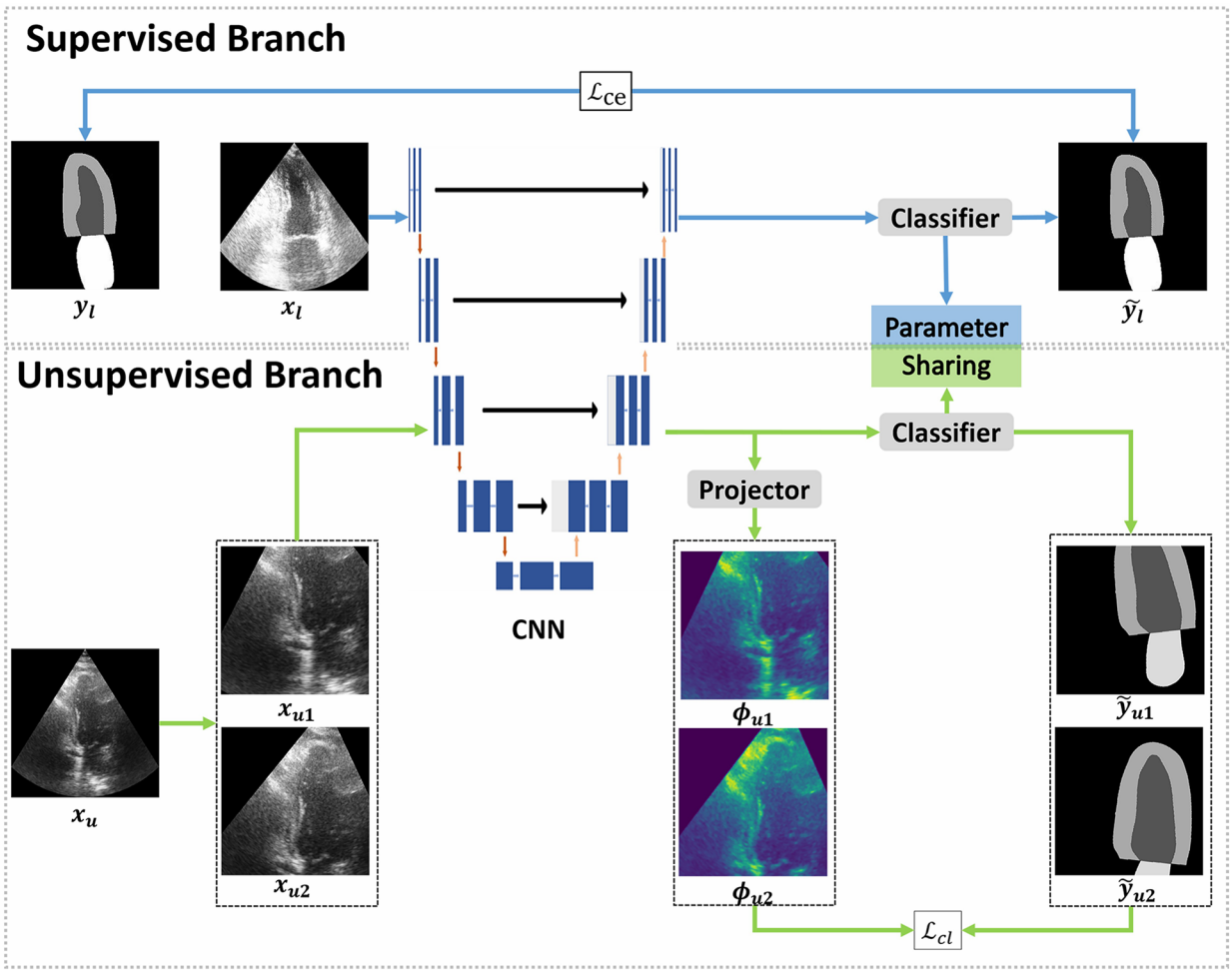


FIGURE 1
Overview of the proposed framework. The proposed framework can be divided into two branches: supervised and unsupervised. The supervised branch was trained with a few epochs firstly, followed by training the unsupervised branch with a lot of epochs. The supervised branch consists of a convolutional neural network (CNN) to extract the feature map of labeled images and a classifier to make predictions. In the unsupervised branch, the same CNN and classifier from the supervised branch were used, the classifiers in two branches share parameters with each other. In CNN, each blue box represents a multi-channel feature map and each white box represents a copied feature map. The downward red arrows indicate the downsampling stages and the upward red arrows indicate the upsampling stages. The blue arrow to the right represents a convolution of $1 \times 1$ and the black arrow to the right indicates a skip connection. The dilated convolutions were implemented in the downsampling stages. Furthermore, a projector was introduced to modify the feature dimensions. In the supervised branch, the loss function employed a standard cross-entropy loss ($\mathcal{L}_{ce}$), which measures the discrepancy between the predictions and the ground truth. On the other hand, the unsupervised branch employed a contrastive learning loss ($\mathcal{L}_{cl}$) to quantify the difference between pseudo labels and feature maps.

branch and loss functions were introduced specifically in Sections 2.2, 2.3 and 2.4.

## 2.2. Supervised branch

The supervised branch of our network, like other supervised networks, was composed of a convolutional neural network (CNN) and a classifier. The CNN was employed to extract features from images, whereas the classifier was responsible for mapping these features to predictions. More specifically, we employed the u-net as the backbone in supervised branch to convert training images into feature vectors, which was the CNN referred to in Figure 1. U-net (17) is a type of neural network that follows an encoder-decoder structure, where the encoder is able to capture context information and the decoder can perform precise localization. The encoder and decoder are connected through a skip connection. While the skip connection in u-net can help prevent shallow features from being lost, the use of multiple pooling layers in the contracting path can result in information loss in the images. Dilated convolutions can be used to solve this issue by increasing the field of perception without adding more parameters, minimizing information loss during downsampling. These convolutions inflate the kernel with holes between the kernel elements, and the dilation rate parameter indicates the amount the kernel is widened (18). To enhance the performance of the proposed network, we incorporated dilated convolutions into the downsampling process of u-net, and used this modified u-net as the CNN. The dilated convolutions were implemented in the second, third and fourth downsampling stages of u-net, with dilation rates of 1, 2 and 4 respectively. To evaluate the effectiveness of the dilated convolutions, we conducted an ablation experiment which was described in Section 3.6.

## 2.3. Unsupervised branch

The proposed unsupervised branch was based on the contrastive learning strategy, which is a type of framework for learning discriminative representations. The main focus of contrastive learning was to compare pairs of sample examples that are considered to be either similar (positive samples) or dissimilar (negative samples) in terms of their semantic content.

In our work, each image was randomly cropped twice and then done some different augmentations to create two different transformation views, which were considered as positive samples. In this process, we made sure the two positive samples have an overlap region. The negative samples were images from the training set, but without including the given image. As shown in Figure 1, we transformed the $x_u$ image to create $x_{u1}$ and $x_{u2}$, which serve as positive samples, and randomly selected images from our training set (excluding $x_u$) as negative samples. The training process aimed to bring the positive samples ($x_{u1}$ and $x_{u2}$) closer together and separated the negative samples that belong to other classes. To maintain consistency in the representation of the overlap region, we employed the loss function $\mathcal{L}_{cl}$. Finally, the unsupervised branch was able to learn the deep features of ventricle, myocardium and atrium and discriminate them well without labeled images.

As shown in Figure 2, the unsupervised branch consisted of a CNN, a classifier and a projector. Among them, the CNN is the same one shared with the supervised branch. The classifier in unsupervised branch had the same architecture as that in the supervised branch, and they shared the same parameters. The projector was comprised of two linear layers, with the first layer followed by batch normalization and rectified linear units (ReLU). The purpose of the projector was to change the feature dimension and prevent the loss of useful information for segmentation. An ablation experiment was performed to evaluate the significance of this projector, and more details have been shown in Section 3.6. The loss function was a pixel-wise

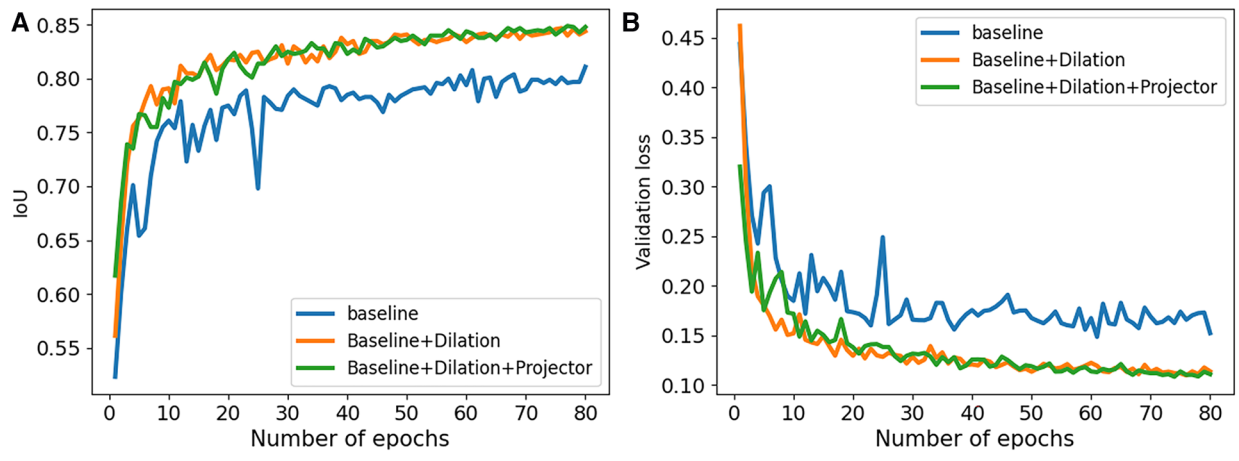


FIGURE 2
Performance comparison of our ablation studies on 2CH images from CAMUS dataset. (A) The IoU values were examined to assess the performance of our ablation studies. (B) The validation loss values were analyzed to evaluate the performance of our ablation studies.

contrastive loss, and it will be elaborated upon, as shown in Section 2.4.

## 2.4. Loss functions

### 2.4.1. Supervised loss

Cross entropy loss was used in supervised networks to measure the dissimilarity between the predicted probability distribution of class labels and the actual distribution of class labels (19). Compared with other loss functions, this loss function is differentiable and easy to optimize using gradient based methods. Therefore, in the supervised branch, we used cross entropy loss as the loss function, which is frequently employed for image semantic segmentation tasks.

The cross-entropy loss $\mathcal{L}_{ce}$ can be written as follows:

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{i}\sum_{c=1}^{M} y_{ic} log(p_{ic}) \tag{1}$$

where $M$ and $N$ are the number of classes and samples; $y_{ic}$ depends on the truth value of $i$, if it is equals to $c$, $y_{ic}$ is 1, else $y_{ic}$ is 0. $p_{ic}$ is the probability of sample $i$ belongs to class $c$.

### 2.4.2. Unsupervised loss

In the unsupervised branch, followed the previous work Lai et al. (15), the proposed loss function $\mathcal{L}_{cl}$ is:

$$\mathcal{L}_{cl} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{dc}^{ns,pf} \tag{2}$$

where $\lambda$ is used to control the contribution of $\mathcal{L}_{dc}^{ns,pf}$, and the range of $\lambda$ is 0-1. In our experiment, we set $\lambda$ with 0.1. $\mathcal{L}_{ce}$ is a standard cross entropy loss, which is the same as the loss function in supervised branch.

$\mathcal{L}_{dc}^{ns,pf}$ was a Directional Contrastive loss (DC Loss), which was used to minimize the distance between positive feature pairs (features with the same class) and maximize the distance between negative feature pairs (features with different classes). Specifically, as shown in Figure 1, we regarded two features of $\phi_{u1}$ and $\phi_{u2}$ as a positive pair, because both of them corresponded to the same pixels in $x_u$ but with different transformations. In addition, any two images in the training dataset were regarded as a negative pair.

The DC Loss ($\mathcal{L}_{dc}^{ns,pf}$) can be written as follows:

$$\mathcal{L}_{dc}^{ns,pf} = \frac{1}{B}\sum_{b=1}^{B}(l_{dc}^{b,ns,pf}(\phi_{u1}, \phi_{u2}) + l_{dc}^{b,ns,pf}(\phi_{u2}, \phi_{u1})) \tag{3}$$

where $B$ represents the batch size of training.

The $l_{dc}^{b,ns,pf}$ can be written as follows:

$$l_{dc}^{b,ns,pf}(\phi_{u1}, \phi_{u2}) = -\frac{1}{N}\sum_{h,w}\mathcal{M}_{d,pf}^{h,w} \cdot log\frac{r(\phi_{u1}^{h,w}, \phi_{u2}^{h,w})}{r(\phi_{u1}^{h,w}, \phi_{u2}^{h,w}) + \sum_{\phi_n \in \mathcal{F}_u}\mathcal{M}_{n,1}^{h,w} \cdot r(\phi_{u1}^{h,w}, \phi_n)} \tag{4}$$

where $\mathcal{M}_{d,pf}^{h,w}$ is the binary mask and can filter those uncertain positive samples. $r$ denotes the exponential function of the cosine similarity $s$ between two features with a temperature $\tau$. $r(\phi_1, \phi_2) = \exp(s(\phi_1, \phi_2)/\tau)$; $h$ and $w$ denote the height and width of 2-D images; $N$ denotes the number of spatial locations of $x_u$; $\phi_n \in \mathbb{R}^c$ represents the negative counterpart of the feature $\phi_{u1}^{h,w}$, and $F_u$ represents the set of negative samples.

The binary mask $\mathcal{M}_{d,pf}^{h,w}$ can be written as follows:

$$\mathcal{M}_{d,pf}^{h,w} = \mathcal{M}_{d}^{h,w} \cdot 1\{max\mathcal{C}(f_{u2}^{h,w}) > \gamma\} \tag{5}$$

where $\gamma$ is a threshold. If the confidence of a positive sample is lower than $\gamma$, this positive pair will not contribute to the final loss.

The directional mask $\mathcal{M}_{d}^{h,w}$ can be written as:

$$\mathcal{M}_{d}^{h,w} = 1\{max\mathcal{C}(f_{u1}^{h,w}) < max\mathcal{C}(f_{u2}^{h,w})\} \tag{6}$$

where $C$ is the classifier. $f_{u1}^{h,w}$ and $f_{u2}^{h,w}$ are features of $x_{u1}$ and $x_{u2}$ extracted by CNN.

# 3. Experiment

## 3.1. Dataset

In this paper, the proposed method was validated on CAMUS dataset (20), which consisted of clinical exams from 500 patients. For each patient, it included two-dimension (2D) apical 2CH and 4CH view echocardiogram sequences, along with annotations for LV, LA, and LV endocardium at end diastole (ED) and end systole (ES) frames. Thereinto, 450 patients were used as the training dataset to train the proposed model, and 50 patients were used as the testing dataset to evaluate the performance of the trained model. For the training dataset, it contained 366 patients with good or medium image quality, and 84 patients with poor image quality. The testing dataset contained 40 patients with good or medium image quality, and 10 patients with poor image quality. Since the images in the dataset had different sizes, we resized all of them to a uniform resolution of $512 \times 512$ and normalized them to the range of $[-1, 1]$ before training.

## 3.2. Implementation details

The proposed network was implemented based on pytorch1.13 and trained on a single NVIDIA Tesla A40 GPU with 48GB memory. In order to reduce the GPU memory usage and improve the efficiency of training, we used automatic mixed precision (AMP) in training.

In our experiments, we initialized the network parameters randomly and opted for the SGD optimizer with a weight decay of 0.0001 and an initial learning rate of 0.01. To update the learning rate, we employed the poly decay policy, which can be expressed as follows:

$$l(iter) = lr \cdot \left(1 - \frac{iter}{\text{total}}\right)^{\text{power}} \tag{7}$$

where $power = 0.9$; $iter$ is the number of epochs we are currently training; $total$ is the sum of the epochs used for training. We trained the supervised branch in the first 5 epochs before training the unsupervised branch. In the end, we completed training for a total of 80 epochs.

## 3.3. Data augmentation

In order to avoid the overfitting and improve the robustness of the proposed network, several data augmentations were applied before training, including Gaussian blur, color jitter, gray scale, horizontal flipping. In unsupervised branch, we applied random crop and random rotation. Specifically, we randomly cropped images to a size of $320 \times 320$ and rotated them with an arbitrarily degree within the range of $[-15°, 15°]$.

## 3.4. Training process

The training process of the supervised branch can be described as follows. Firstly, the labeled image $x_l$ was processed by the CNN ($\varepsilon$) to obtain its corresponding feature map $f_l = \varepsilon(x_l)$. Then, the classifier $C$ made predictions $\sim y_l = C(f_l)$ based on the feature map. Finally, the predictions $\sim y_l$ were compared to the ground truth $y_l$ using cross entropy loss for supervision.

The training process of the unsupervised branch can be described as follows. Firstly, an unlabeled image $x_u$ was processed with two different transformations to get two images $x_{u1}$ and $x_{u2}$. These two images were then fed through the CNN ($\varepsilon$) to generate the feature maps $f_{u1} = \varepsilon(x_{u1})$ and $f_{u2} = \varepsilon(x_{u2})$. After that, the classifier $C$ made predictions and based on the feature map, while the projector $\Phi$ change the feature dimension $\varphi_{u1} = \Phi(f_{u1})$ and $\varphi_{u2} = \Phi(f_{u2})$. Finally, we calculated the loss between the low dimension feature and the pseudo labels.

## 3.5. Evaluation metrics

In our experiments, we used Dice Similarity Coefficient (DSC) and Intersection-over-Union (IoU) to evaluate the performance of the proposed method for images segmentation.

The DSC and IoU can be described as follows:

$$DSC = \frac{2 \times |A \cap B|}{|A| + |B|} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{8}$$

$$IoU = \frac{A \cap B}{A \cup B} = \frac{TP}{TP + FP + FN} \tag{9}$$

where $A$ is the predicted set of pixels; $B$ is the ground truth; $TP$ represents the true positive; $FP$ represents the false positive and $FN$ represents the false negative. Note that IoU and DSC in the following tables and figures represent the average value of segmentation results from four classes, including background, LA, LV and myocardium.

## 3.6. Ablation study

We conducted a series of ablation experiments to verify the contribution of each component in the proposed method. The ablation study was based on the full labeled data on CAMUS dataset. In our experiments, we first embed a standard u-net as the baseline. Then, we modified the u-net by adding dilated convolutions to the downsampling process and named it "Baseline+Dilation." Finally, we added a projector to "Baseline +Dilation" to complete proposed method, which we called "Baseline+Dilation+Projector." The mean IoU and mean DSC results for each method were presented in **Table 1** and **Figure 2**. Meanwhile, boxplots were employed to illustrate the variability of the mean Intersection over Union (IoU) for the aforementioned three methods in **Figure 3**.

*Dilation:* In order to demonstrate the effectiveness of the dilated convolution applied in the u-net, we made a comparison between Baseline and Baseline with dilation (Baseline+Dilation). The experimental results shown that the model incorporated dilation (Baseline+Dilation) outperformed the one without dilation (Baseline), with an average IoU of 0.026 for 2CH images and 0.014 for 4CH images (**Table 1**).

*Projector:* In **Table 1**, we can see that incorporating a projector into the model improved the mean IoU from 0.847 to 0.849 in 2CH images. From **Figure 2**, we can find that the segmentation performance became better and more stable with the influence of projector in the last 30 epochs of the training.

## 3.7. Segmentation results and comparison with other methods

The proposed method was assessed for its ability to segment multiple cardiac structures, including LV, myocardium and LA. The segmentation performance was evaluated on the testing

TABLE 1 Statistical comparison of ablation studies on 2CH and 4CH images with full labeled training data.

| Method | 2CH | | | | 4CH | | | |
|---|---|---|---|---|---|---|---|---|
| | IoU | *D*(IoU) | DSC | *D*(DSC) | IoU | *D*(IoU) | DSC | *D*(DSC) |
| Baseline | 0.811 | 0.004 | 0.893 | 0.002 | 0.854 | 0.003 | 0.919 | 0.002 |
| Baseline + Dilation | 0.847 | 0.003 | 0.916 | 0.001 | **0.868** | 0.002 | **0.928** | 0.001 |
| Baseline + Dilation + Projector | **0.849** | 0.002 | **0.917** | 0.001 | **0.868** | 0.002 | **0.928** | 0.001 |

The *D*(IoU) and *D*(DSC) represent the variance of the IoU and DSC scores, respectively.
The best results are achieved and highlighted by the bold values.

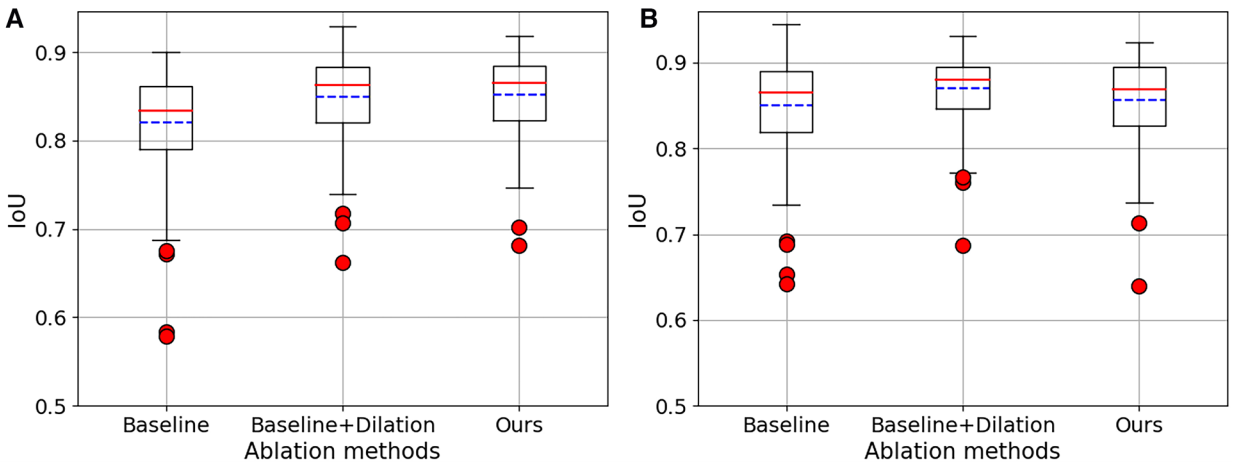


FIGURE 3
The variability of IoU about our ablation studies. (A) The results obtained from 2CH images. (B) The results obtained from 4CH images.

dataset of CAMUS dataset, comprising 40 patients with good or medium quality images and 10 patients with poor quality images.

While the CAC (Context Aware Consistency Network) method demonstrates strong performance in natural images segmentation, it falls short when it comes to accurately delineating cardiac structures in echocardiography (15). In certain cases, it may even incorrectly segment certain regions. Despite these limitations, CAC remains a widely employed semi-supervised technique for various image segmentation tasks, including those involving cardiac structures in echocardiography. Therefore, we compared the performance of our proposed method with CAC in order to demonstrate the superiority of our approach for cardiac segmentation. In addition, we compared our method with some supervised methods, including u-net and DeeplabV3+ with Resnet50 backbone, using all the labeled images available. To guarantee a fair comparison, we implemented all methods under the same conditions, including the same data augmentations and the same learning rate adjustment strategy.

The performance of each method has been presented in Tables 2 and 3, showcasing the results for the proposed

TABLE 2 Segmentation performance comparison of IoU and DSC between the proposed method and other techniques.

| Method | 2CH | | | | 4CH | | | | *N* | Params |
|---|---|---|---|---|---|---|---|---|---|---|
| | IoU | *D*(IoU) | DSC | *D*(DSC) | IoU | *D*(IoU) | DSC | *D*(DSC) | | |
| DeeplabV3+ | 0.673 | 0.190 | 0.796 | 0.170 | 0.671 | 0.012 | 0.794 | 0.009 | 1/4 | 40.347MB |
| | 0.70 | 0.014 | 0.816 | 0.012 | 0.713 | 0.011 | 0.825 | 0.009 | 1/2 | |
| | 0.786 | 0.005 | 0.876 | 0.003 | 0.818 | 0.003 | 0.896 | 0.002 | Full | |
| U-net | 0.780 | 0.007 | 0.872 | 0.005 | 0.796 | 0.006 | 0.882 | 0.003 | 1/4 | 17.267MB |
| | 0.805 | 0.006 | 0.888 | 0.004 | 0.835 | 0.006 | 0.907 | 0.003 | 1/2 | |
| | 0.832 | 0.004 | 0.906 | 0.002 | 0.854 | 0.003 | 0.919 | 0.002 | Full | |
| CAC | 0.780 | 0.006 | 0.872 | 0.004 | 0.811 | **0.002** | 0.892 | **0.001** | 1/4 | 40.348MB |
| | 0.784 | 0.005 | 0.875 | 0.003 | 0.824 | **0.002** | 0.900 | **0.001** | 1/2 | |
| | 0.787 | 0.004 | 0.877 | 0.002 | 0.836 | 0.003 | 0.908 | **0.001** | Full | |
| Ours | **0.819** | **0.004** | **0.898** | **0.002** | **0.829** | 0.003 | **0.903** | 0.002 | 1/4 | 17.268MB |
| | **0.838** | **0.003** | **0.911** | **0.002** | **0.857** | 0.003 | **0.921** | **0.001** | 1/2 | |
| | **0.849** | **0.002** | **0.917** | **0.001** | **0.868** | **0.002** | **0.928** | **0.001** | Full | |

*N* represents the ratio of labeled images that we used. The *D*(IoU) and *D*(DSC) represent the variance of the IoU and DSC scores, respectively.
The best results are achieved and highlighted by the bold values.

TABLE 3 Segmentation performance comparison of precision and recall between the proposed method and other techniques.

| Method | 2CH | | | | 4CH | | | | N |
|---|---|---|---|---|---|---|---|---|---|
| | *P* | *D*(*P*) | *R* | *D*(*R*) | *P* | *D*(*P*) | *R* | *D*(*R*) | |
| DeeplabV3+ | 0.809 | 0.010 | **0.953** | **0.002** | 0.847 | 0.005 | 0.950 | **0.001** | 1/4 |
| | 0.829 | 0.007 | **0.956** | **0.001** | 0.874 | 0.005 | **0.971** | **0.001** | 1/2 |
| | 0.896 | 0.002 | **0.965** | **0.001** | 0.908 | 0.002 | 0.957 | **0.001** | Full |
| U-Net | 0.847 | 0.004 | 0.870 | 0.004 | 0.881 | 0.004 | 0.926 | 0.002 | 1/4 |
| | 0.882 | 0.004 | 0.907 | 0.003 | 0.904 | 0.002 | 0.938 | **0.001** | 1/2 |
| | 0.902 | 0.002 | 0.924 | 0.002 | 0.911 | 0.002 | 0.939 | **0.001** | Full |
| CAC | 0.864 | 0.003 | 0.926 | **0.002** | 0.891 | **0.002** | 0.937 | **0.001** | 1/4 |
| | 0.880 | 0.002 | 0.927 | 0.002 | 0.903 | **0.001** | 0.940 | **0.001** | 1/2 |
| | 0.881 | 0.002 | 0.935 | 0.002 | 0.897 | 0.002 | 0.932 | **0.001** | Full |
| Ours | **0.894** | **0.002** | 0.927 | **0.002** | **0.905** | **0.002** | **0.961** | **0.001** | 1/4 |
| | **0.908** | **0.001** | 0.938 | **0.001** | **0.921** | **0.001** | 0.959 | **0.001** | 1/2 |
| | **0.923** | **0.001** | 0.948 | **0.001** | **0.924** | **0.001** | **0.962** | **0.001** | Full |

*N* represents the ratio of labeled images that we used. *P* represents precision and *R* represents recall. The *D*(*P*) and *D*(*R*) represent the variance of the precision and recall, respectively.
The best results are achieved and highlighted by the bold values.

TABLE 4 Segmentation performance of the proposed method on LV, LA and myocardium.

| *N* | | 2CH | | | | 4CH | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | IoU | *D*(IoU) | DSC | *D*(DSC) | IoU | *D*(IoU) | DSC | *D*(DSC) |
| 1/4 | LV | 0.824 | 0.007 | 0.904 | 0.003 | 0.832 | 0.008 | 0.908 | 0.003 |
| | LA | 0.806 | 0.014 | 0.893 | 0.007 | 0.837 | 0.012 | 0.911 | 0.011 |
| | Myocardium | 0.694 | 0.012 | 0.819 | 0.007 | 0.697 | 0.013 | 0.809 | 0.008 |
| 1/2 | LV | 0.836 | 0.007 | 0.911 | 0.003 | 0.865 | 0.005 | 0.928 | 0.002 |
| | LA | 0.830 | 0.011 | 0.907 | 0.005 | 0.843 | 0.014 | 0.915 | 0.012 |
| | Myocardium | 0.726 | 0.009 | 0.841 | 0.005 | 0.748 | 0.008 | 0.856 | 0.004 |
| Full | LV | 0.848 | 0.003 | 0.917 | 0.001 | 0.877 | 0.004 | 0.934 | 0.002 |
| | LA | 0.842 | 0.010 | 0.915 | 0.005 | 0.854 | 0.015 | 0.921 | 0.012 |
| | Myocardium | 0.744 | 0.005 | 0.851 | 0.002 | 0.766 | 0.006 | 0.867 | 0.003 |

*N* represents the ratio of labeled images that we used.

method, CAC method, and other supervised methods in both the 2CH and 4CH views. From the tables, it was evident that the proposed method outperformed the other methods in terms of segmentation performance with 1/4, 1/2, and full labeled data. We also compared the number of parameters among u-net, DeeplabV3+, the proposed method, and the

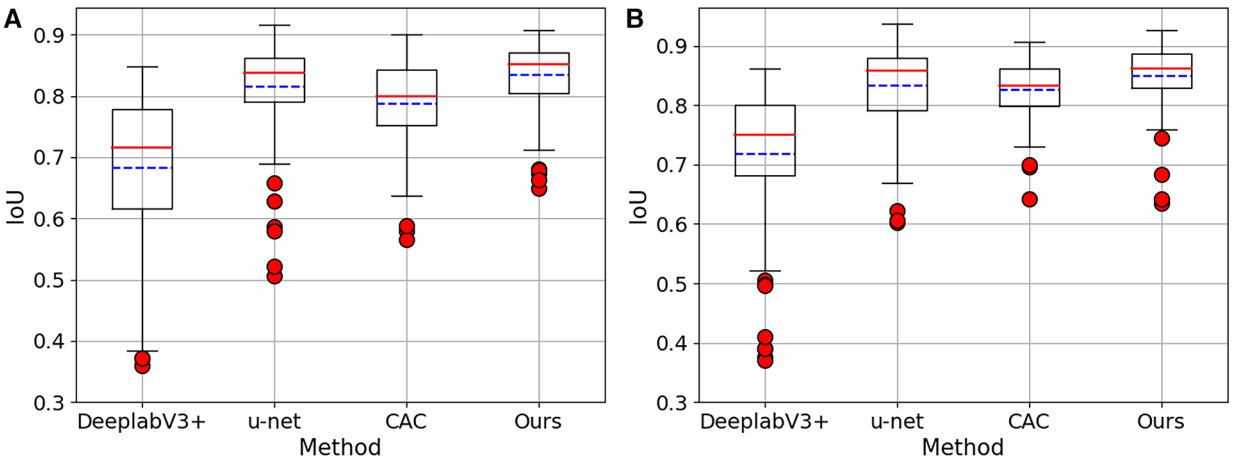


FIGURE 4
The Boxplots show the IoU values for both 2CH and 4CH images, using the four methods mentioned in Table 2. These boxplots specifically showcase the outcomes achieved when employing 1/2 labeled images. (**A**) The results obtained from 2CH images. (**B**) The results obtained from 4CH images.

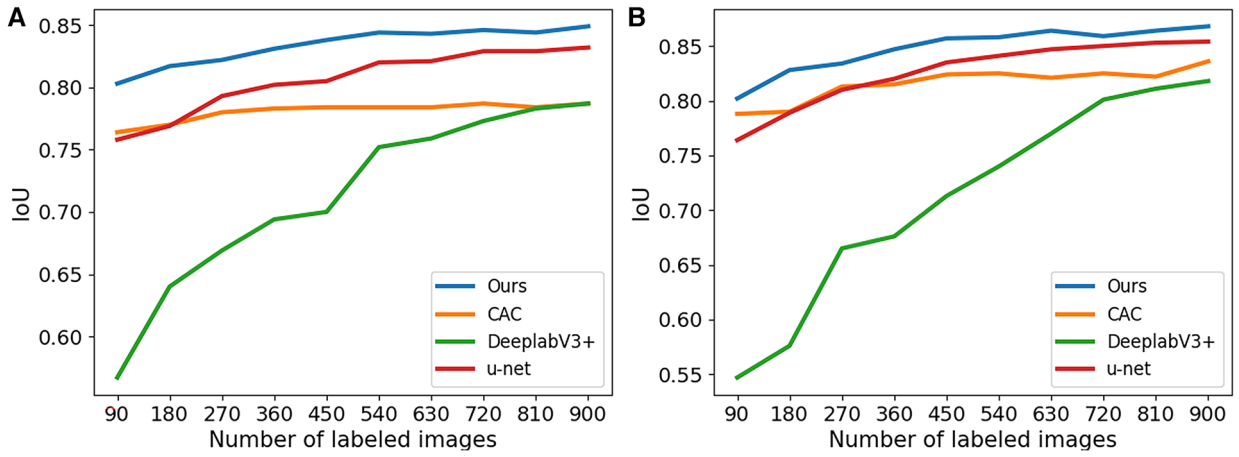


FIGURE 5
The IoU for segmentation results of the proposed method trained with different numbers of labeled images. (A) The results obtained from 2CH images. (B) The results obtained from 4CH images.

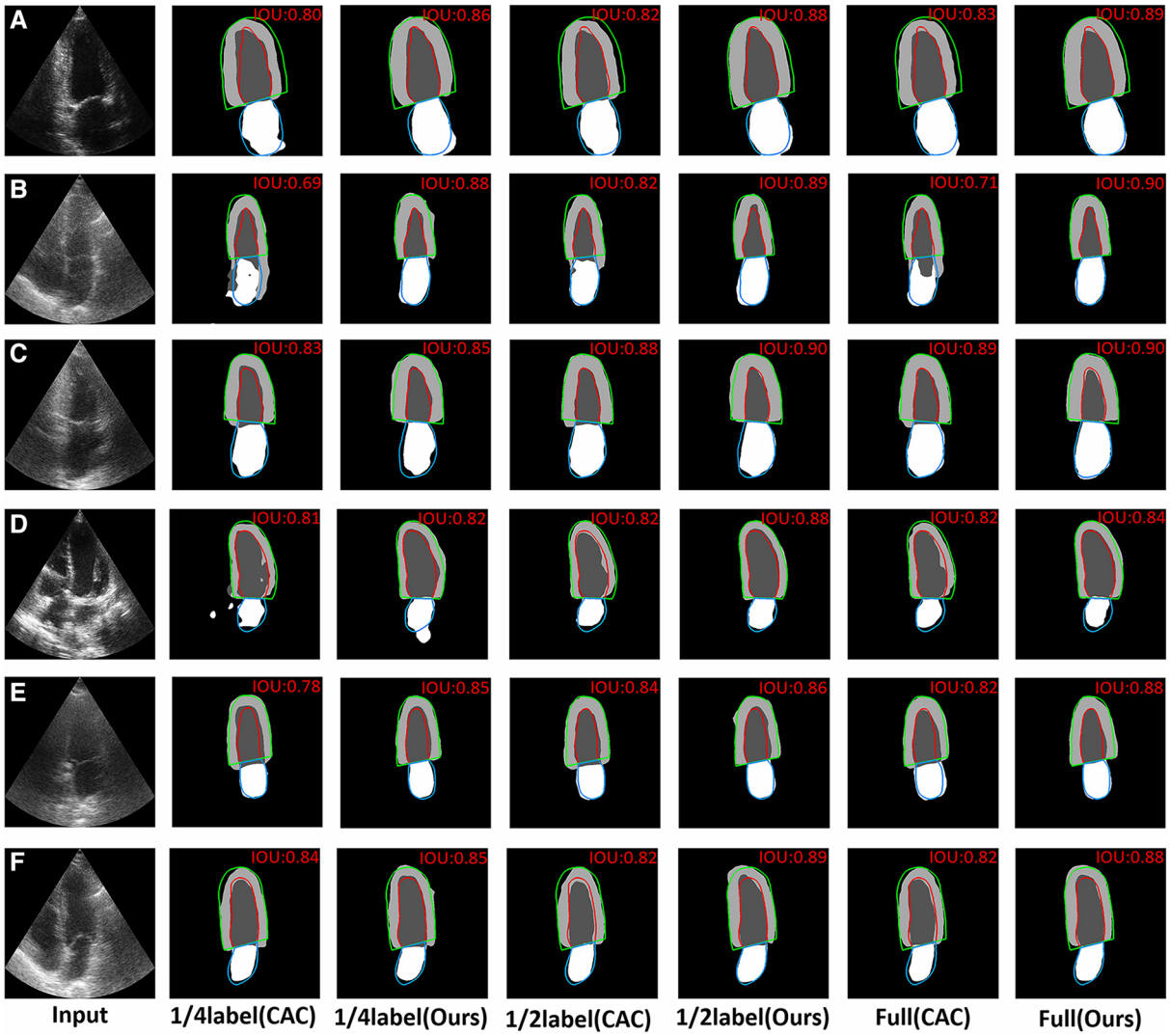


FIGURE 6
Visual comparison of CAC method and the proposed method with 1/4labeled, 1/2labeled and full labeled data on 2CH and 4CH images. The red, green and blue lines represent the ground truth. The value of IoU is marked on the right upper corner of each image. (A,B) The visual comparison of the performance about discriminating the borders of myocardium on 2CH images and 4CH images. (C,D) The visual comparison of the probability of segmentation errors in some regions on 2CH images and 4CH images. (E,F) The visual comparison of the performance about the identification of the ventricle and myocardium location.

CAC method in Table 2, showing that our method had fewer parameters. Note that the values in the tables shown the maximum of the epochs we trained. The multi-structure segmentation performance of the proposed method has been presented in Table 4. We can see with the increase in the number of labeled images that take part in the training process, the Iou and DSC are improved.

In Figure 4, boxplots have been used to visually represent the range of variation in IoU values achieved by the four methods mentioned earlier, where 1/2 labeled images were employed for training. We can see our proposed method achieved lower variation and higher mean IoU for for both 2CH and 4CH images, in comparison to the other methods. In addition, Figure 5 illustrated the trends of mean IoU as the number of labeled images increases. It was observed that as the number of labeled images utilized in the training process increased, the mean IoU also improved. Notably, the proposed method consistently outperformed the other three methods in terms of segmentation accuracy across all increments of labeled data.

The typical visual segmentation result of CAC method and the proposed method were shown in Figure 6, where the colorful line represents the ground truth. In Figures 6A,B, it shown that the proposed method could discriminate the borders of myocardium better than CAC method. Figures 6C,D shown that the proposed method could reduce the probability of segmentation errors in each region of the images. Figures 6E,F shown that the proposed method achieved a better identification of the ventricle and myocardium location than CAC method.

In Figure 7, we shown certain challenging cases with poor image quality. The IoU values are presented in the upper right corner of each segmentation result. It becomes evident that the

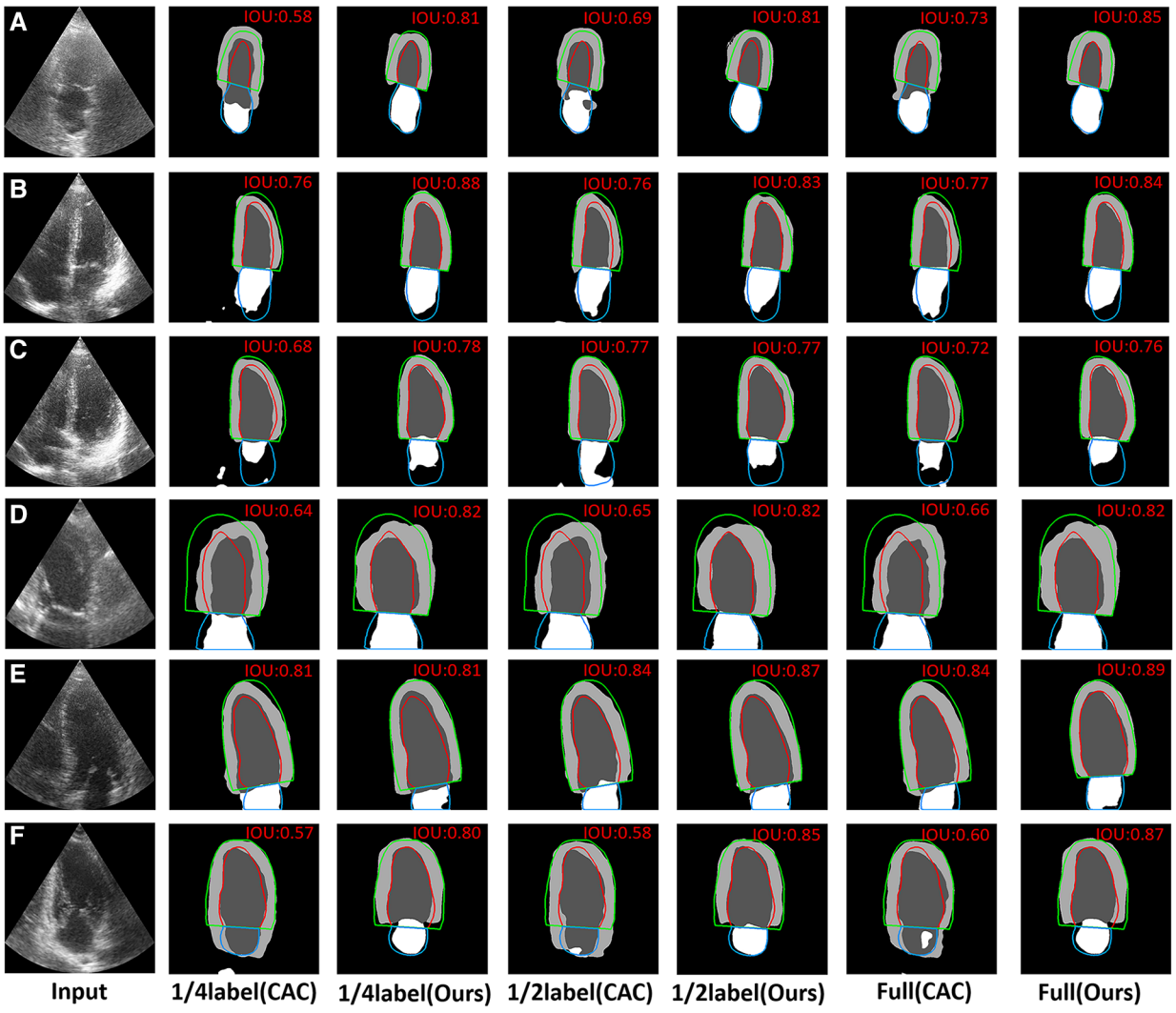


FIGURE 7
Visual comparison of CAC method and the proposed method on 6 typical challenging images. The red, green and blue lines represent the ground truth. The value of IoU is marked on the right upper corner of each image. (A–C) The visual comparison of the performance on some typical low contrast images. (D) The visual comparison of the performance on images where the complete cardiac structures are not present. (E,F) The visual comparison of the performance on images that have no a clear border between the ventricle and the atrium.

proposed method surpassed limitations of echocardiography more effectively. Specifically, **Figures 7A–C** demonstrate that the proposed method outperformed the CAC method in mitigating the disadvantage of low contrast in echocardiography. Additionally, **Figure 7D** showcases that the proposed method achieved superior heart location identification compared to the CAC method in images where complete cardiac structures are not present. Moreover, **Figures 7E,F** indicate that the proposed method reduced the likelihood of segmentation errors in regions lacking clear boundaries between the ventricle and the atrium.

## 4. Conclusion

In this paper, we proposed a semi-supervised method to segment the cardiac structures with echocardiography. The proposed method first applied contrastive learning strategy into cardiac structure segmentation, allowing for effective use of unlabeled data. The network was able to mitigate the adverse effects of low contrast, incomplete cardiac structures and unclear boundaries in certain aspects of echocardiography. A lot of experiments conducted on the CAMUS dataset shown that the proposed network can effectively employ unlabeled data for the automatic segmentation of multiple structures, resulting in outstanding performance. This advancement contributes significantly to the diagnosis and screening of cardiovascular diseases (CVD), and also reduce the burden of doctors in assessing echocardiography.

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.creatis.insa-lyon.fr/Challenge/camus/databases.html.

## Ethics statement

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

ZG: Data curation, Validation, Visualization, Writing – original draft. YZ: Formal analysis, Writing – original draft, Writing – review & editing. ZQ: Writing – original draft, Data curation, Software. SD: Writing – original draft, Investigation, Methodology, Resources, Supervision, Writing – review & editing. SH: Project administration, Investigation, Validation, Writing – original draft. HG: Conceptualization, Project administration, Visualization, Writing – review & editing. JZ: Data curation, Formal analysis, Investigation, Writing – original draft. YC: Writing – review & editing, Conceptualization, Data curation, Formal analysis. BH: Writing – review & editing, Software, Validation. ZK: Conceptualization, Writing – review & editing, Investigation. ZQ: Writing – review & editing, Conceptualization, Funding acquisition, Resources, Supervision, Writing – original draft. YL: Software, Visualization, Writing – review & editing. CL: Funding acquisition, Resources, Writing – original draft.

## Funding

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

## Conflict of interest

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

## Publisher's note

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

## References

1. Chen C, Qin C, Qiu H, Tarroni G, Duan J, Bai W, et al. Deep learning for cardiac image segmentation: a review. *Front Cardiovasc Med.* (2020) 7:25. doi: 10.3389/fcvm.2020.00025

2. Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography, the european association of cardiovascular imaging. *Eur Heart J Cardiovasc Imaging.* (2015) 16:233–71. doi: 10.1093/ehjci/jev014

3. Gottdiener JS. Overview of stress echocardiography: uses, advantages,, limitations. *Prog Cardiovasc Dis.* (2001) 43:315–34. doi: 10.1053/pcad.2001.20502

4. Capotosto L, Massoni F, De Sio S, Ricci S, Vitarelli A, et al.Early diagnosis of cardiovascular diseases in workers: role of standard and advanced echocardiography. *BioMed Res Int.* (2018) 2018:2314–6133. doi: 10.1155/2018/7354691

5. Wu H, Liu J, Xiao F, Wen Z, Cheng L, Qin J. Semi-supervised segmentation of echocardiography videos via noise-resilient spatiotemporal semantic calibration, fusion. *Med Image Anal.* (2022) 78:102397. doi: 10.1016/j.media.2022.102397

6. Barbosa D, Dietenbeck T, Schaerer J, D'hooge J, Friboulet D, Bernard O. B-spline explicit active surfaces: an efficient framework for real-time 3-D region-based segmentation. *IEEE Trans Image Process.* (2011) 21:241–51. doi: 10.1109/TIP.2011.2161484

7. Yang C, Wu W, Su Y, Zhang S. Left ventricle segmentation via two-layer level sets with circular shape constraint. *Magn Reson Imaging.* (2017) 38:202–13. doi: 10.1016/j.mri.2017.01.011

8. Zhuang X, Rhode KS, Razavi RS, Hawkes DJ, Ourselin S. A registration-based propagation framework for automatic whole heart segmentation of cardiac MRI. *IEEE Trans Med Imaging.* (2010) 29:1612–25. doi: 10.1109/TMI.2010.2047112

9. Cui X, Cao Y, Liu Z, Sui X, Mi J, Zhang Y, et al. TRSA-Net: task relation spatial co-attention for joint segmentation, quantification and uncertainty estimation on paired 2D echocardiography. *IEEE J Biomed Health Inform*. (2022) 26:4067–78. doi: 10.1109/JBHI.2022.3171985

10. Cui X, Zhang P, Li Y, Liu Z, Xiao X, Zhang Y, et al. MCAL: an anatomical knowledge learning model for myocardial segmentation in 2-D echocardiography. *IEEE Trans Ultrason Ferroelectr Freq Control*. (2022) 69:1277–87. doi: 10.1109/TUFFC.2022.3151647

11. Hamila O, Ramanna S, Henry CJ, Kiranyaz S, Hamila R, Mazhar R, et al. Fully automated 2D and 3D convolutional neural networks pipeline for video segmentation and myocardial infarction detection in echocardiography. *Multimed Tools Appl*. (2022) 81:37417–39. doi: 10.1007/s11042-021-11579-4

12. El Rai MC, Darweesh M, Al-Saad M. Semi-supervised segmentation of echocardiography videos using graph signal processing. *Electronics*. (2022) 11:3462. doi: 10.3390/electronics11213462

13. Wei H, Ma J, Zhou Y, Xue W, Ni D. Co-learning of appearance and shape for precise ejection fraction estimation from echocardiographic sequences. *Med Image Anal*. (2023) 84:102686. doi: 10.1016/j.media.2022.102686

14. Chen T, Xia M, Huang Y, Jiao J, Wang Y. Cross-domain echocardiography segmentation with multi-space joint adaptation. *Sensors*. (2023) 23:1479. doi: 10.3390/s23031479

15. Lai X, Tian Z, Jiang L, Liu S, Zhao H, Wang L, et al. Semi-supervised semantic segmentation with directional context-aware consistency. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway, New Jersey, United States: IEEE Xplore (2021). p. 1205–14.

16. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with Atrous separable convolution for semantic image segmentation. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Berlin/Heidelberg, Germany: Springer (2018). p. 801–18.

17. 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, Proceedings, Part III 18*; 2015 Oct 5–9; Munich, Germany. Springer (2015). p. 234–41.

18. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions [Preprint] (2015). Available at: https://doi.org/10.48550/arXiv.1511.07122

19. De Boer P-T, Kroese DP, Mannor S, Rubinstein RY. A tutorial on the cross-entropy method. *Ann Oper Res*. (2005) 134:19–67. doi: 10.1007/s10479-005-5724-z

20. Leclerc S, Smistad E, Pedrosa J, Østvik A, Cervenansky F, Espinosa F, et al. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. *IEEE Trans Med Imaging*. (2019) 38:2198–210. doi: 10.1109/TMI.2019.2900516

frontiers | Frontiers in Cardiovascular Medicine

TYPE Original Research
PUBLISHED 13 November 2023
DOI 10.3389/fcvm.2023.1250311



OPEN ACCESS

EDITED BY
Gongning Luo,
Harbin Institute of Technology, China

REVIEWED BY
Kamran Shamsa,
University of California, Los Angeles, United States
Wenqi Lu,
Manchester Metropolitan University, United Kingdom

*CORRESPONDENCE
S. S. Hothi
s.hothi@nhs.net

RECEIVED 05 July 2023
ACCEPTED 16 October 2023
PUBLISHED 13 November 2023

CITATION
Jiang J, Liu B, Li YW and Hothi SS (2023) Clinical service evaluation of the feasibility and reproducibility of novel artificial intelligence based-echocardiographic quantification of global longitudinal strain and left ventricular ejection fraction in trastuzumab-treated patients.
Front. Cardiovasc. Med. 10:1250311.
doi: 10.3389/fcvm.2023.1250311



# Clinical service evaluation of the feasibility and reproducibility of novel artificial intelligence based-echocardiographic quantification of global longitudinal strain and left ventricular ejection fraction in trastuzumab-treated patients

J. Jiang[1], B. Liu[2,3], Y. W. Li[4] and S. S. Hothi[1,3,5]*

[1]Heart and Lung Centre, New Cross Hospital, Royal Wolverhampton NHS Trust, Wolverhampton, United Kingdom, [2]Department of Cardiology, Manchester University NHS Foundation Trust, Manchester, United Kingdom, [3]Institute of Cardiovascular Sciences, University of Birmingham, Birmingham, United Kingdom, [4]Department of Anaesthesia, New Cross Hospital, Royal Wolverhampton NHS Trust, Wolverhampton, United Kingdom, [5]Research Centre for Health and Life Sciences, Coventry University, Coventry, United Kingdom

**Introduction:** Cardiotoxicity is a potential prognostically important complication of certain chemotherapeutic agents that may result in preclinical or overt clinical heart failure. In some cases, chemotherapy must be withheld when left ventricular (LV) systolic function becomes significantly impaired, to protect cardiac function at the expense of a change in the oncological treatment plan, leading to associated changes in oncological prognosis. Accordingly, patients receiving potentially cardiotoxic chemotherapy undergo routine surveillance before, during and following completion of therapy, usually with transthoracic echocardiography (TTE). Recent advancements in AI-based cardiac imaging reveal areas of promise but key challenges remain. There are ongoing questions as to whether the ability of AI to detect subtle changes in individual patients is at a level equivalent to manual analysis. This raises the question as to whether AI-based left ventricular strain analysis could provide a potential solution to left ventricular systolic function analysis in a manner equivocal to or superior to conventional assessment, in a real-world clinical service. AI based automated analyses may represent a potential solution for addressing the pressure of increasing echocardiographic demands within limited service-capacity healthcare systems, in addition to facilitating more accurate diagnoses.

**Methods:** This clinical service evaluation aims to establish whether AI-automated analysis compared to conventional methods (1) is a feasible method for assessing LV-GLS and LVEF, (2) yields moderate to good correlation between the two approaches, and (3) would lead to different clinical recommendations with serial surveillance in a real-world clinical population.

**Results and Discussion:** We observed a moderate correlation ($r = 0.541$) in GLS between AI automated assessment compared to conventional methods. The LVEF quantification between methods demonstrated a strong correlation ($r = 0.895$). AI-generated GLS and LVEF values compared reasonably well with conventional methods, demonstrating a similar temporal pattern throughout echocardiographic surveillance. The apical-three chamber view demonstrated the lowest correlation ($r = 0.423$) and revealed to be least successful for

acquisition of GLS and LVEF. Compared to conventional methodology, AI-automated analysis has a significantly lower feasibility rate, demonstrating a success rate of 14% (GLS) and 51% (LVEF).

KEYWORDS

cardio-oncology, trastuzumab, cardiotoxicity, artificial intelligence, strain, echocardiography

## Introduction

Cardiotoxicity is a significant, potential complication of certain chemotherapeutic agents that can lead to either preclinical or overt heart failure. In some cases, chemotherapy must be withheld when cardiac function, primarily left ventricular (LV) systolic function, becomes significantly impaired to protect cardiac function at the expense of a change in the oncological treatment plan and associated changes in prognosis (1). Accordingly, patients receiving potentially cardiotoxic chemotherapy are recommended to undergo routine surveillance before, during and following completion of therapy, usually with transthoracic echocardiography (TTE). Transthoracic echocardiography is a well-established and widely available imaging modality with an important role in determining cardiac structure and function. To date, it remains the preferred technique for assessing the development, progression and regression of cardiotoxicity among oncology patients undergoing cardiac surveillance (2).

Echocardiographic indices such as left ventricular ejection fraction (LVEF) by Simpson's Biplane method has traditionally been used to assess changes in LV systolic function. However, in the modern era of speckle tracking echocardiography (STE), strain quantification has rapidly evolved into a valuable tool for the early detection of cardiotoxicity during oncological therapy and has since been incorporated into international guidance (3, 4).

Until now, global longitudinal strain (GLS) has been the most studied strain parameter with the largest body of literature supporting its diagnostic and prognostic value (5, 6). One early study evaluated eighty-one females with newly diagnosed HER2 + breast cancer for early alterations of myocardial strain during treatment with anthracycline and/or trastuzumab. Patients received three-monthly surveillance throughout the course of a fifteen-month study period. A reduction in LVEF was observed in the overall cohort ($64 \pm 5\%$ to $59 \pm 6\%$; $p < 0.0001$); twenty-six patients [32%, (22%–43%)] developed cardiotoxicity, and of these patients, 5 [6%, (2%–14%)] developed symptoms of heart failure (HF). Significant LVEF reduction (≥8%) was detected in 15% of patients that developed subsequent cardiotoxicity, whereas upon the application of strain analysis, the incidence rate increased to 78%. Among the patients that later developed HF, all had a reported GLS of less than −19% (7).

While strain quantification with speckle tracking echocardiography represents a sensitive method for assessing LV function, this postprocessing analysis remains laborious, time-consuming and is subject to significant inter- and intra-observer variability, related to reproducibility of contouring cardiac structure by manual and even semi-automated contouring. In recent years, the emergence of artificial intelligence (AI) in echocardiography has generated much interest among the cardiac imaging community. The technology is rapidly evolving but is yet to be widely adopted into clinical practice. Recent evidence has revealed promising findings, demonstrating that the application of AI enables data analysis free from human operator bias, accelerated workflow and quantification, along with high feasibility rate in the absence of operator input. One multicentre study which assessed LVEF and longitudinal strain using visual, manual and fully AI-automated-methods (TomTec-Arena 1.2, TomTec Imaging Systems) reported a high feasibility (98%) of AI-automated assessment (8). Good correlation and levels of agreement were observed between manual and automated assessment (ICC: 0.83; bias: 0.7%; 95% CI: 0.1%–1.3%). Expectedly, bias and levels of agreement were wider when visual assessments were compared. A key advantage of automated LVEF and LV-GLS compared to manual and visual assessment was the absence of inter-measurement variability on repeated assessments with the AI method able to identify the same patterns each time. Finally, beat to-beat variability was $0.96 \pm 3.52\%$ for automated LVEF, $2.7 \pm 8.16\%$ for manual LVEF, $0.19 \pm 1.31\%$ for automated GLS, and $1.09 \pm 3.29\%$ for manual GLS (8).

In support of these findings is another recent trial by Salte et al., which reported good correlation ($R = 0.93$, $p < 0.001$) and low bias of $-1.4 \pm 0.3\%$ ($p < 0.01$) with an estimated level of agreement (LOA) of ±3.7% when comparing AI-automated vs. conventional methodology (EchoPAC v.202, GE), suggesting that the application of AI is potentially comparable to human expert performance using conventional methodology (9).

While AI-based cardiac imaging analysis appear promising, there are areas that require further assessment. AI-automated analysis must be able to perform at least as well as established methodologies to detect subtle changes in left ventricular function, whether LVEF or GLS. Hence, further research is needed to fully establish the vulnerability of automated image processing networks. Furthermore, this automated approach relies upon a large training dataset to implicitly learn features of the heart relevant to segmentation which is resource intensive, demands close clinical supervision and raises potential ethical and privacy concerns.

If AI-automated analysis of LV function can be demonstrated to be equivocal to or superior to conventional methods within a real-world clinical service, then it may represent a potential solution for the challenges of limited clinical service capacity by reducing the pressures of increasing echocardiographic demands, in addition to facilitating more accurate diagnoses. This clinical service evaluation aims to establish whether AI-automated analysis is: (1) a feasible method for assessing LV-GLS and LVEF, (2) correlates well with conventional methods, and (3) whether AI analysis would lead to different clinical recommendations during serial surveillance in a real-world clinical population.

# Materials and methods

## Patient population

This single-centre audit and service evaluation retrospectively reviewed all HER2 + breast cancer patients that underwent TTE surveillance and trastuzumab therapy between January 2019 and October 2022 at the Royal Wolverhampton NHS Trust (UK) and assessed the evaluation of cardiac function against international cardio-oncology guidance (Audit/Service evaluation number 5918, Royal Wolverhampton NHS Trust, UK). Informed consent was not required due to the retrospective nature of the clinical audit and evaluation. Patients undergoing combination therapy including anthracycline were excluded from the study. Patients with atrial fibrillation or other form of arrhythmias during the echocardiographic studies were also excluded. To reflect real-world patient population and feasibility, patients with partially suboptimal endocardial border definition were not excluded. Clinical characteristics of our cohort were collected from the image reporting system and hospital records and are summarised in **Table 1**.

## Echocardiographic imaging protocol and analysis

648 TTE studies acquired from 142 oncology patients that received trastuzumab echocardiographic surveillance between 2019 and 2022 were retrospectively evaluated. All echocardiographic studies within our British Society of Echocardiography (BSE) accredited imaging laboratory were comprehensive studies which complied with BSE cardio-oncology guidelines. Echo imaging was performed by BSE accredited echocardiographers using commercial equipment (Affiniti, EPIQ and iE33, Phillips Medical Systems, Andover, Massachusetts, USA).

## Assessment of GLS and LVEF

AI-automated and conventionally measured GLS and LVEF were assessed from standard apical four- (A4C), three- (A3C), and two-chamber (A2C) cine loops in accordance with BSE guidance.

TABLE 1 Demographic and clinical characteristics of the patient population.

| | | | |
|---|---|---|---|
| Age (years) | 59 ± 13 | | |
| Gender | 140 Female | 2 Male | |
| ECG and HR (bpm) | 142 SR 79 ± 13 | | |
| Height (cm) | 163 ± 7.5 | | |
| Weight (kg) | 76 ± 18 | | |
| BMI (kg/m$^2$) | 28.7 ± 6.4 | | |
| BSA (m$^2$) | 1.85 ± 0.2 | | |
| Blood pressure (BP) | Systolic BP 137 ± 26 mmHg | Diastolic BP 80 ± 14 mmHg | |
| Cancer type | 119 BC | 18 GC | 5 OC |

Data are expressed as mean ± standard deviation.
ECG, electrocardiogram; HR, heart rate; SR, sinus rhythm; AF, atrial fibrillation; BMI, body mass index; BSA, body surface area; BC, breast cancer; GC, gastric cancer; OC, oesophageal cancer.

AI-automated assessments (GLS and LVEF) were performed on individual echocardiographic studies using an AI-based platform (Ultromics EchoGo Core, Oxford, UK). The investigators submitted individual clinical studies required for analysis from the local hospital archiving system to the AI pipeline (Ultromics SaaS). Individual views are identified and classified with the existing convolutional neural network (CNN) model and subsequently processed by a U-Net based architecture for view-specific LV contouring, myocardial segmentation, and myocardial motion tracking to compute GLS and LVEF in the absence of manual adjustments (10).

Conventional GLS assessment was performed in a semi-automated fashion from the apical four-, three- and two-chamber LV-focused cine images in dedicated conventional software (QLab, version 15.5, Philips Medical Systems). Upon detection of the endocardial border, the software automatically established a region of interest (ROI) and calculated the strain values of the selected view. The BSE-accredited or similarly experienced operator manually adjusted the ROI to optimise tracking if deemed necessary and strain values were recalculated to reflect this adjustment. Where image quality was insufficient to permit strain assessment of all three views, then a global strain value could not be calculated. Conventional LVEF was manually performed using the Simpson's biplane method of discs (Modified Simpson's rule) for LV volumes and LVEF calculation. End-diastole was defined as the frame following mitral valve closure or the frame in which the cardiac dimension is largest, in preference to the onset of the QRS. End-systole was defined as the frame preceding mitral valve opening or the time in the cardiac cycle in which the cardiac dimension is smallest, respectively. This protocol was performed using the LV-focused A4C and A2C views.

## Statistical analysis

Continuous variables were expressed as mean ± standard deviation and categorical variables were presented as *n* (%). Linear regression analysis was performed to evaluate the relationship between GLS and LVEF when assessed by either conventional or AI-automated methods. Bland-Altman analysis was used to assess the levels of agreement and quantify systemic differences between assessments. Comparison of mean values between the automated and conventional groups were performed using the paired sample student *t*-test. Analysis of variance (ANOVA) was used to compare the means of three of more groups. For all statistical tests performed, a *p*-value less than 0.05 was regarded as statistically significant. Statistical analyses were performed using IBM SPSS Statistics version 29 (New York, USA).

# Results

## Subject characteristics

The patient cohort included 142 patients which had undergone a total of 648 echocardiographic studies as part of their oncological

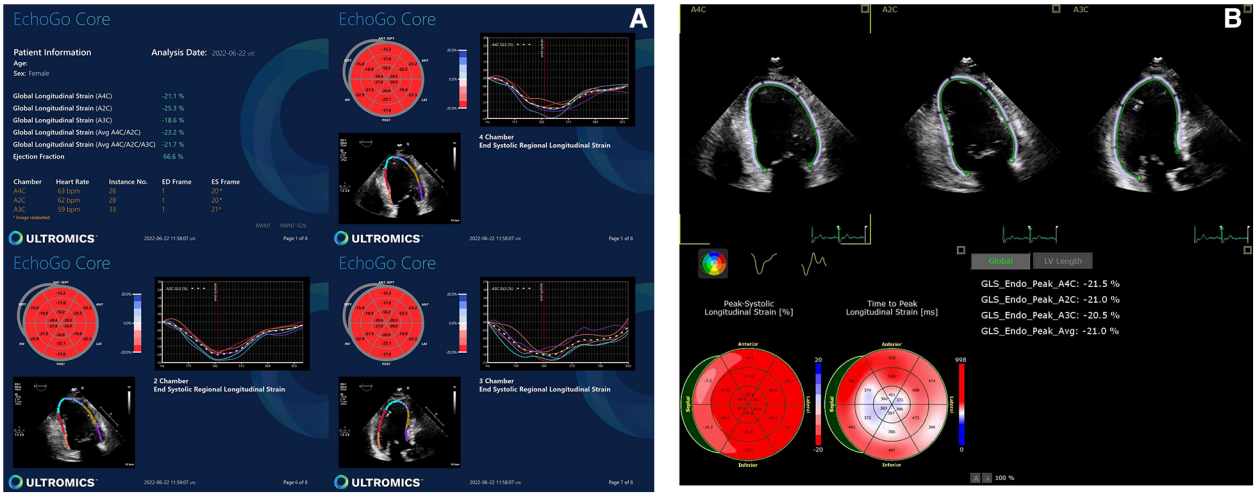


FIGURE 1
Normal GLS data yielded by (A) AI-based and (B) conventional semi-automated strain analysis.

therapy cardiac surveillance. The population comprised 140 females (99%), with mean age 59 ± 13 years (range 28–89 years). Oncological diagnoses predominantly comprised breast cancer (84%), but also included gastric (13%) and oesophageal cancer (3%). Patient demographic and clinical characteristics are summarised in Table 1.

## Technical feasibility of AI-based compared to conventional assessment in GLS and LVEF

AI-generated GLS and LVEF values were acquired in 14% and 51% of all studies, respectively. Representative examples of normal and abnormal GLS studies analysed by AI-generated and conventional assessment are shown in Figures 1, 2 respectively. The rate of success in obtaining strain results using AI vs. conventional methods for the three standard apical views were: A4C, 56% vs. 74%; A3C, 14% vs. 38%; A2C, 46% vs. 53%, respectively (Figure 3).

Technical failure to derive strain from the A3C was therefore the main reason for the low rate of success in obtaining AI-generated GLS (ANOVA $p = 0.028$). Whilst the success rate of deriving longitudinal strain from the A3C via the conventional method was also low, the failure rate was superior to that of AI. Factors contributing to suboptimal image quality, particularly affecting the A3C, included challenging body composition, tachyarrhythmias, ectopy, limited rib space and previous mastectomy.

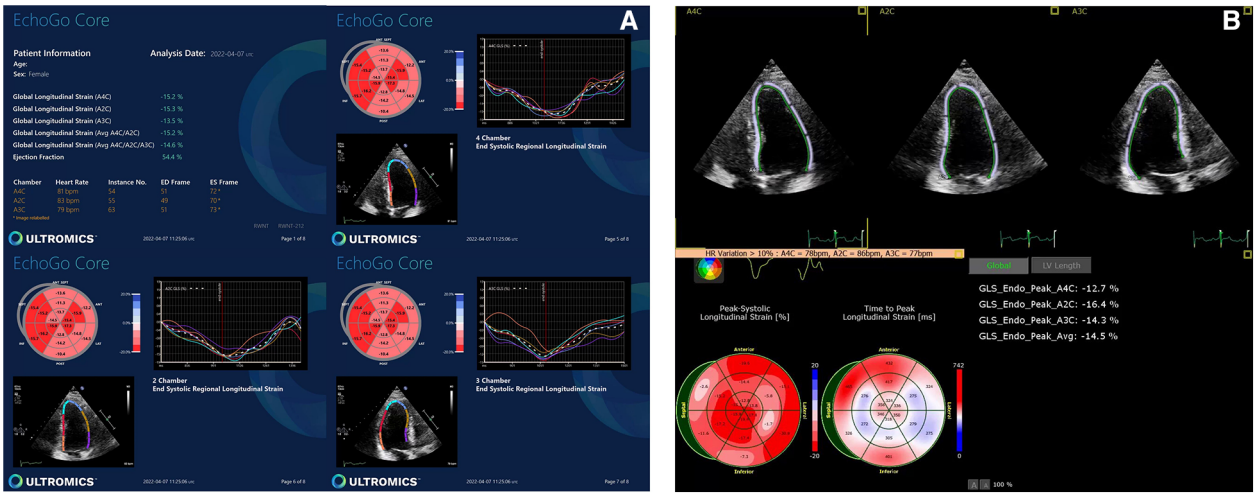


FIGURE 2
Abnormal GLS data yielded by (A) AI-based and (B) conventional semi-automated strain analysis.

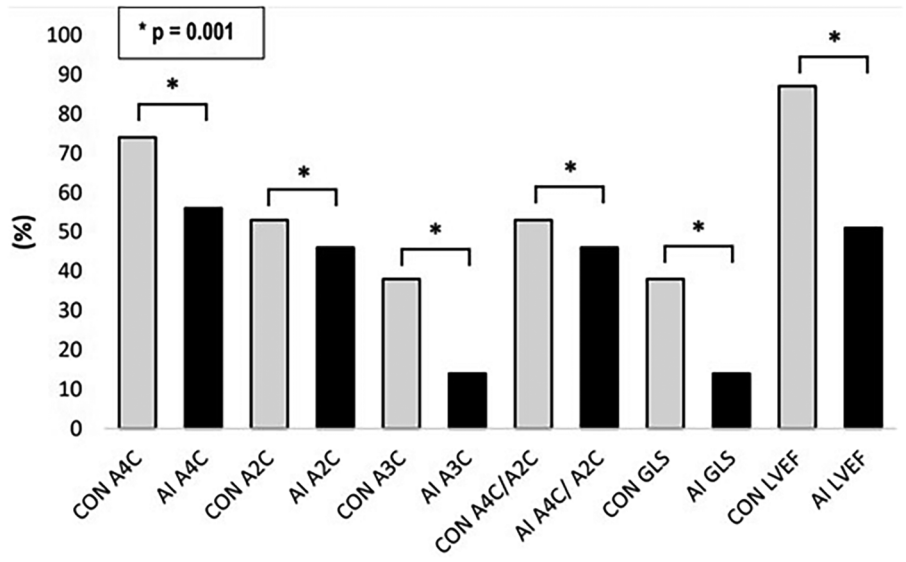


FIGURE 3
Feasibility of AI-based versus conventional semi-automated strain analysis and LVEF in the standard apical views.

## GLS and LVEF using AI vs. conventional assessment

Mean GLS in whole cohort was −17.9 ± 2.2% (AI) vs. −19.1 ± 2.0% (conventional). Mean LVEF in the whole cohort was 61.6 ± 5.7% (AI) vs. 60.7 ± 4.9% (conventional). Linear regression and Bland-Altman analysis for GLS revealed moderate correlation ($r = 0.541$, $p < 0.001$) and disagreement (mean bias −1.2%, 95% CI: −5.2% to 2.8%; **Figures 4A,B**). In contrast, LVEF showed strong correlation ($r = 0.895$, $p < .001$) with small biases (**Figures 5A,B**).

## Comparison between strain at individual apical views using AI vs. conventional assessment

Mean longitudinal strain values from specific apical views were −18.7 ± 2.9% and −19.0 ± 2.6% (A4C) (**Figures 6A,B**), −18.1 ± 2.8% and −18.6 ± 2.6% (A2C) (**Figures 7A,B**), −15.7 ± 2.6% and −16.6 ± 1.6% (A3C) (**Figures 8A,B**), and −18.2 ± 2.7% and −18.6 ± 2.6% for the AI method and the conventional method, respectively. A strong correlation and agreement was demonstrated in the A4C ($r = 0.883$, $p < .001$, 95% CI: −3.0% to 2.4%) and A4C/A2C (measurable values achieved from both A4C and A2C views within a given study) strain ($r = 0.853$, $p < .001$, 95% CI: −3.2% to 2.4%) views for strain between AI-automated and conventional methods (**Figures 9A,B**). In comparison, the A2C strain revealed a moderate correlation ($r = 0.771$, $p < .001$). The weakest correlation ($r = 0.423$, $p = 0.008$) and widest limits of agreement among each individual apical view were observed in the A3C view.

## Temporal changes in GLS and LVEF between AI vs. conventional assessments during surveillance

Serial changes in strain and LVEF during TTE surveillance are summarised in **Table 2**. Statistical differences between the conventional and AI-automated methods at each time point are illustrated in **Table 3** using the independent sample *t*-test. Conventional and AI-automated values followed a similar temporal pattern in patients receiving trastuzumab therapy for both GLS and LVEF irrespective of the cardiotoxic cohort or the total study population (**Figures 10**, **11**). At 3 months (T1), both conventional and automated method demonstrated a reduction in GLS and LVEF compared to baseline measurements (T0). By 6 months (T2), further reduction in LV function was observed to a similar degree by both methods. The GLS and LVEF were seen to be lowest at 9 months (T3) from the initiation of trastuzumab therapy. The AI-automated GLS values were consistently more negative lower at each timepoint compared to the conventional method (**Table 3** and **Figure 11**). The LVEF values at timepoint 3 to 5 were almost identical by both methods although a higher degree of variation was observed from the AI-automated method (T3: 58.9 ± 8.7%, $p = 0.422$; T4: 58.9 ± 7.4, $p = 0.638$; T5: 62 ± 6.0, $p = 0.038$). At 12- (T4) and 15-months (T5), AI-automated values demonstrated improvements in GLS and LVEF. Similar trends were observed from the conventional method although the degree of improvement is shown to be smaller in LVEF at 15-months. There were no significant differences observed between the AI-automated and conventional methods for GLS. For LVEF, there was a significantly lower LVEF from the conventional method (59.5 ± 5.7% vs. 62 ± 6.0%, $p = 0.038$). Based on the GLS and LVEF criteria (11), six patients developed cardiotoxicity; this number was considered too small to allow statistical sub-analysis. Nevertheless,

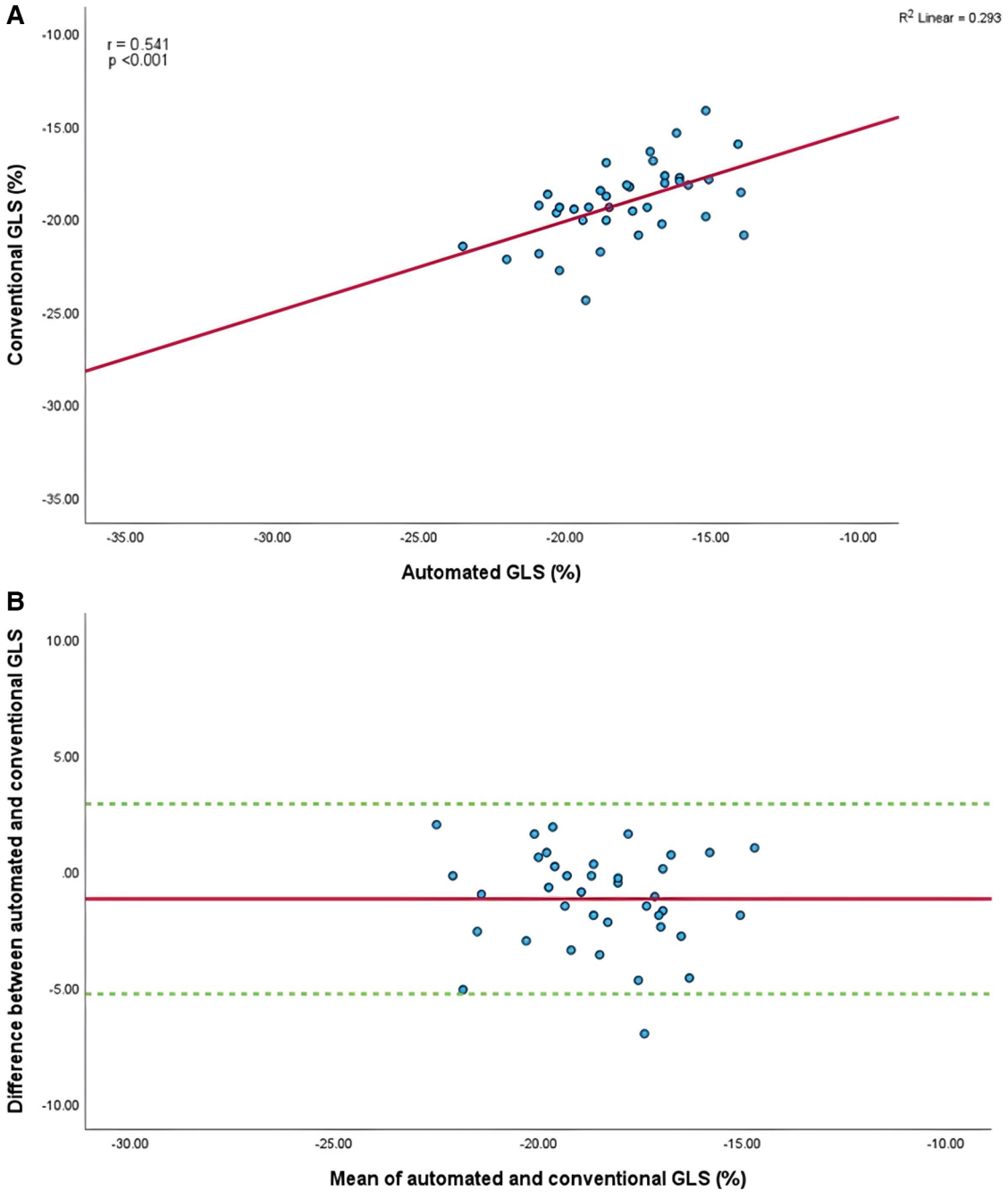


FIGURE 4
(A) Correlation between conventional and AI-automated global longitudinal strain. (B) Bland-Altman plot of conventional and AI-automated global longitudinal strain.

the limited cases have highlighted the ability for AI-automated analysis in detecting left ventricular changes among the cardiotoxic cohort.

## Discussion

In this real-world service evaluation and audit of the assessment of left ventricular ejection fraction and strain in a cohort of patients receiving trastuzumab chemotherapy, we assessed whether an AI-automated solution to LV systolic function is a feasible and reliable methodology compared to conventional analysis. The main findings are firstly, that GLS and LVEF quantification obtained from AI-automated assessment showed moderate to strong correlation compared to conventional methods. Secondly, AI-generated GLS and LVEF values compared reasonably well with conventional methods, demonstrating a similar temporal pattern throughout the echocardiographic surveillance. Thirdly, the apical-three chamber view demonstrated the lowest correlation and revealed to be least

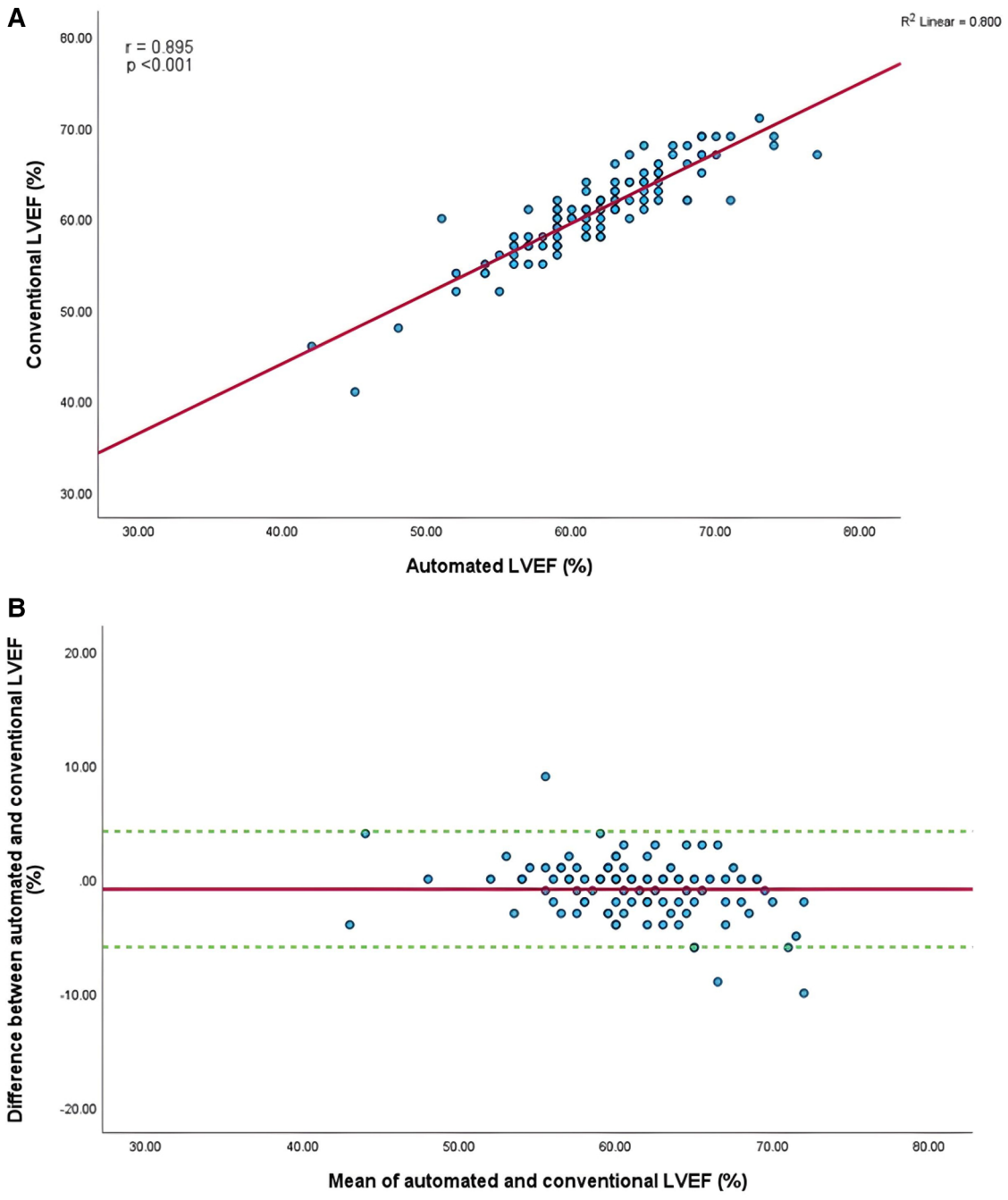


FIGURE 5
(A) Correlation between conventional and AI-automated left ventricular ejection fraction. (B) Bland-Altman plot of conventional and AI-automated left ventricular ejection fraction.

successful for acquisition of GLS and LVEF. Finally, compared to conventional methodology, AI-automated analysis has a significantly lower feasibility rate, demonstrating a success rate of 14% (GLS) and 51% (LVEF).

## Clinical demand and relevance

While the introduction of speckle tracking has provided exciting opportunities in the field of cardiac imaging, its clinical application is rendered meritless if performed by unexperienced or suboptimally trained practitioners. Like any echocardiographic technique, there is a steep learning curve with performing and interpreting echocardiograms (12). Interpretation of echocardiographic studies is demanding and this can limit workflow particularly among smaller centres with fewer trained echocardiographers. The application of AI echocardiography may potentially address these challenges by utilising an AI-based analysis of LV strain.

There is emerging data suggesting that a fully automated AI assessment could potentially reduce post-processing time with

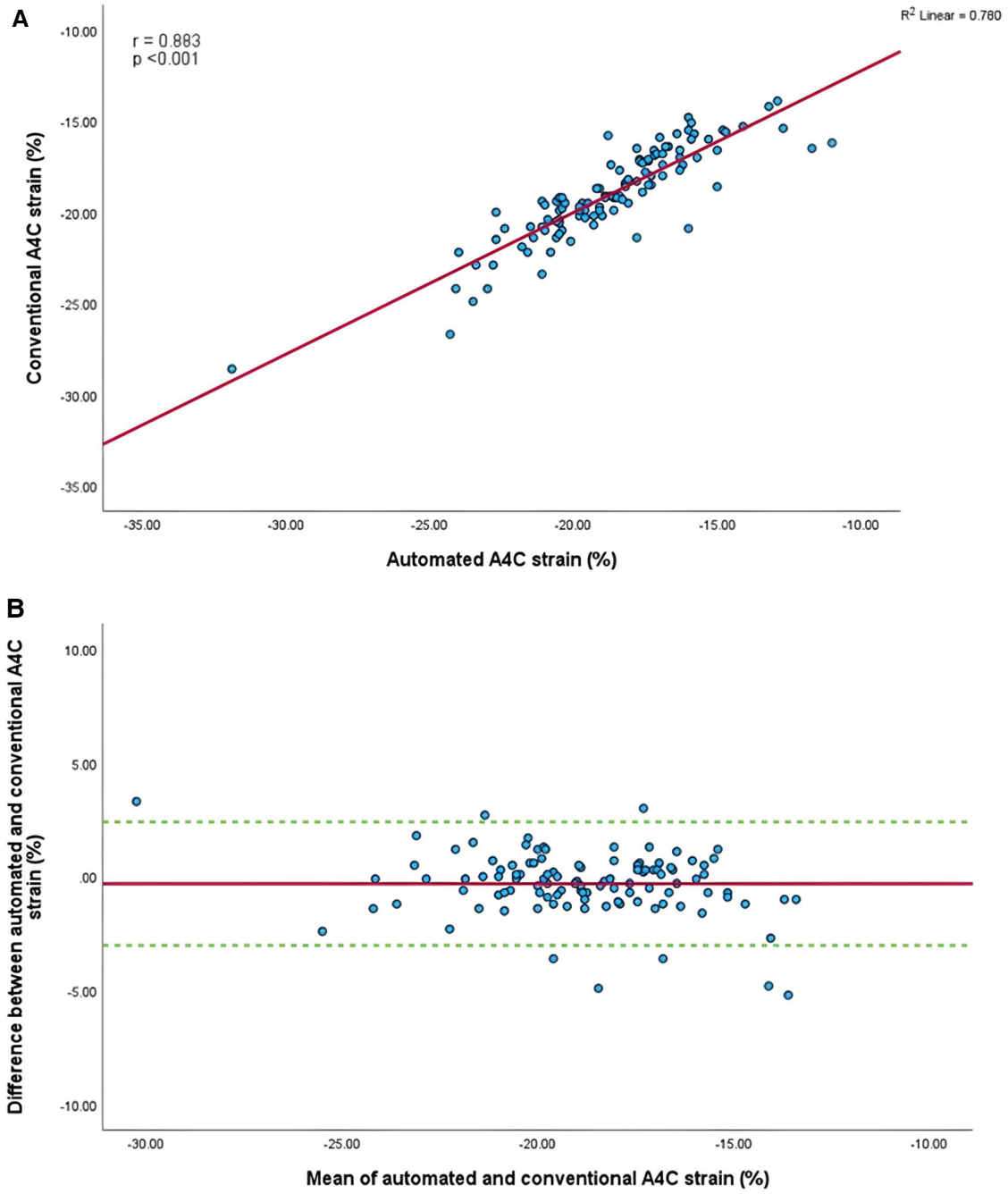


FIGURE 6
(A) Correlation between conventional and AI-automated strain in apical-four chamber view. (B) Bland-Altman plot of conventional and AI-automated strain in apical-four chamber view.

high reproducibility and reduced risk imposed by human-software interaction. However, in the presence of significant knowledge gaps the technology may fall short of this potential. Presently, semi-automated assessments are in clinical use and accepted as a standard, feasible method for LV strain assessment, supported by evidence from numerous studies have supported the use of these methods (13–16). However, the human-software interaction is such that the current semi-automated approach yields values that are highly influenced by the level of experience and training of the sonographer.

Furthermore, research to date has rarely explored the application of AI-automated assessment in cancer therapy-related cardiac dysfunction but instead has largely focused on ischaemia-related cardiac abnormalities. Given that cancer therapy-induced heart failure carries a worse prognosis compared to heart failure related to other causes (17), the need for accurate and frequent echocardiographic

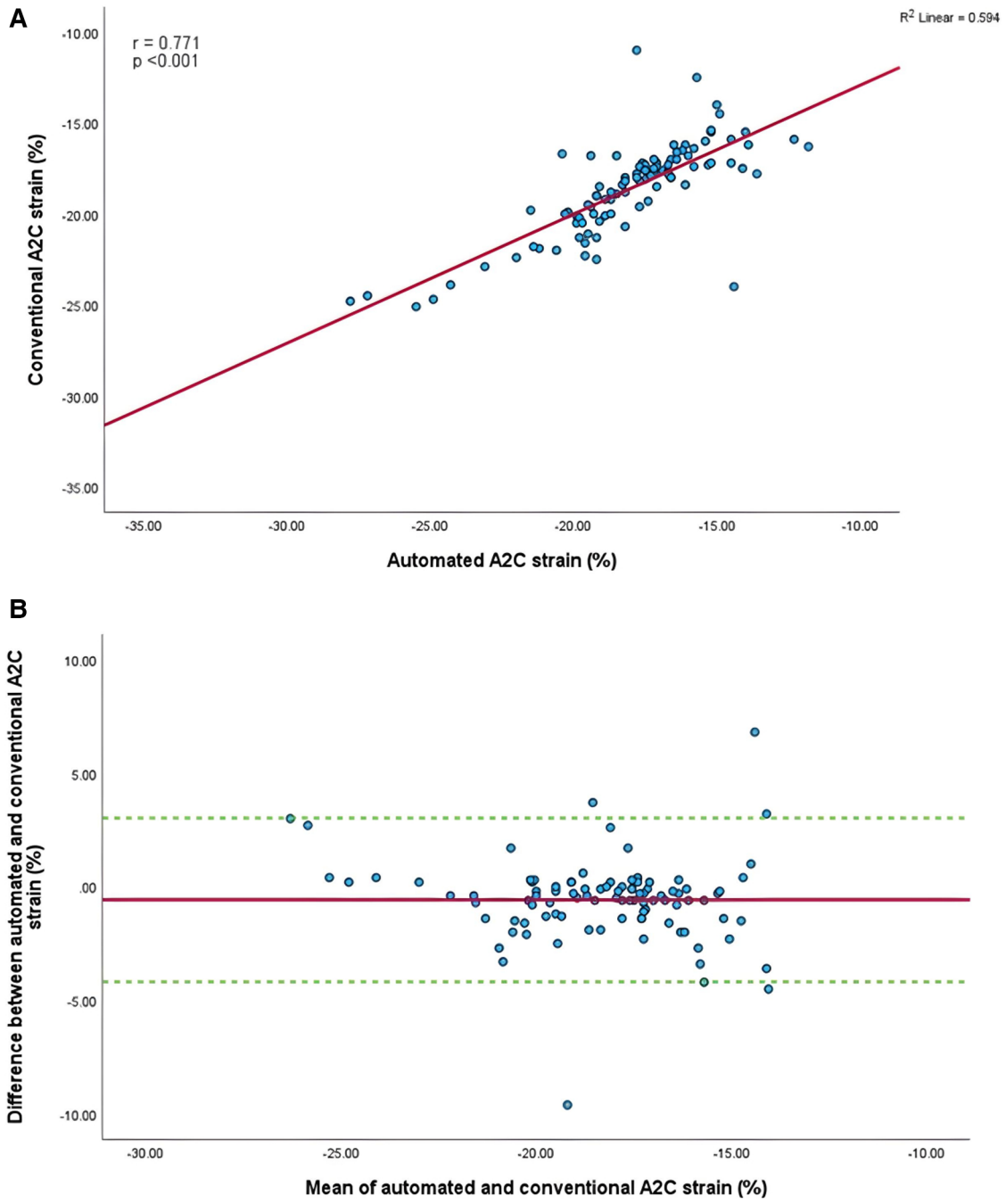


FIGURE 7
(A) Correlation between conventional and AI-automated strain in apical-two chamber view. (B) Bland-Altman plot of conventional and AI-automated strain in apical-two chamber view.

surveillance is clear and of paramount importance. It follows that there is a clinical need for research into AI-automated detection of subclinical changes in cardiac function to accurately, reliable and rapidly detect changes earlier in the disease process. To the best of our knowledge, this is the first real-world evaluation of such an approach to validate and explore the clinical feasibility of AI-automated LV assessment in this patient cohort throughout the surveillance period.

## The feasibility and accuracy of automated GLS and LVEF

The present findings reveal that the current version of AI-automated GLS possess some limitations in feasibility, achieving successful acquisition of GLS in only 14% of all studies. The higher rate of success demonstrated from conventional methods (38%) suggests either that the AI-automated approach is inferior to the semi-automated approach or that the semi-

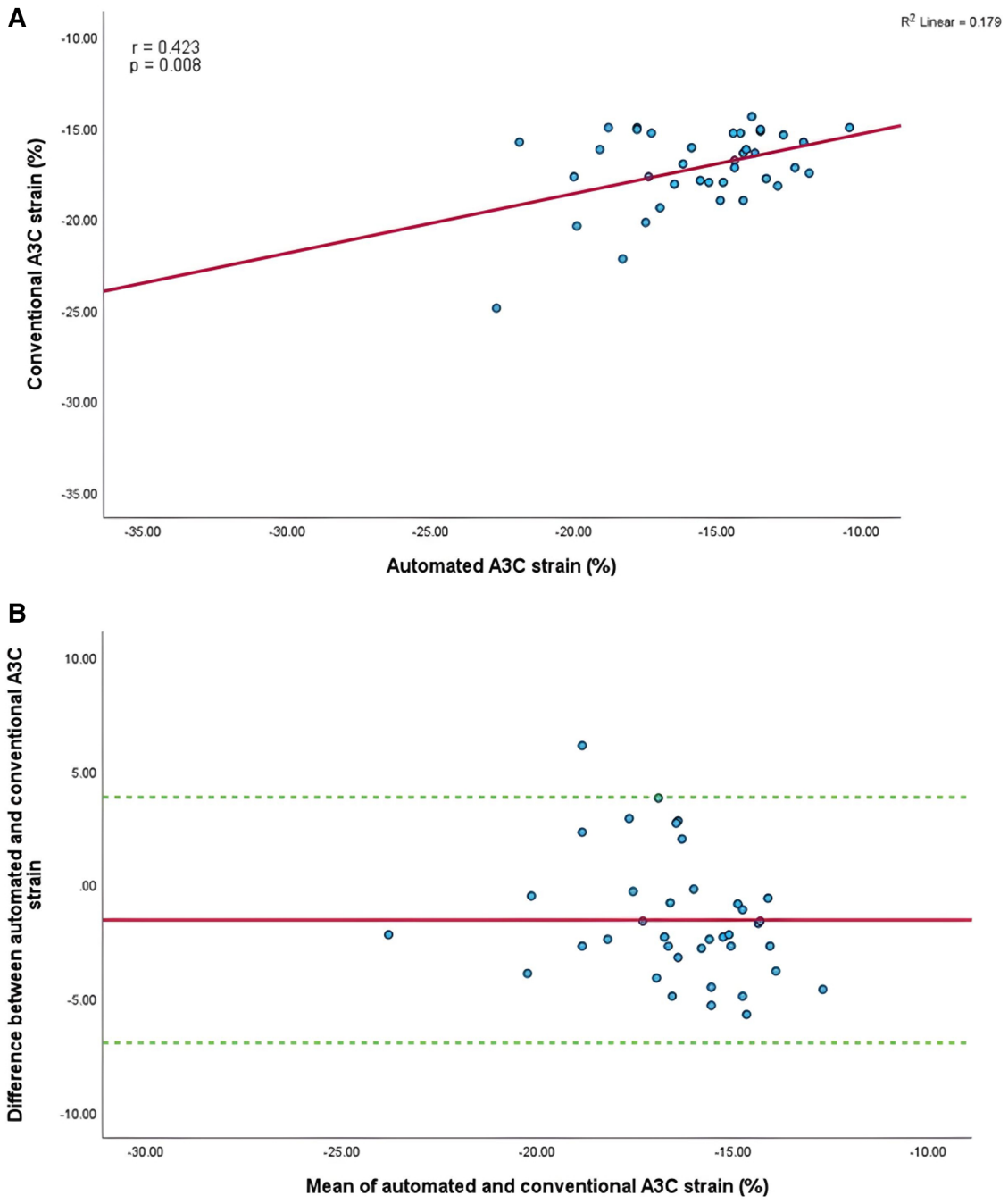


FIGURE 8
(A) Correlation between conventional and AI-automated strain in apical-three chamber view. (B) Bland-Altman plot of conventional and AI-automated strain in apical-three chamber view.

automated approach is overly generous in the studies to which it is applied. The unifying consideration here is that of a threshold for acceptability for an echo study to be amenable to either of the assessment methods. We speculate that the two approaches accept image qualities of different levels. Standardising this threshold is not necessarily a straight-forward proposition as even with a group of selected studies, the AI-automated system is using different approaches to strain assessment than in the semi-automated system.

In either analysis approach, the acquisition of GLS requires the strain values of three individual apical views. The present study found that the A3C view was the most frequently limiting view followed by the A2C (46%) in preventing a GLS assessment. These findings are in keeping with a study by Kawakami et al. which examined the automated tracking quality in each individual LV segments (14). The study found that the LV segments in these in these views are often associated with considerably poorer automated tracking compared to segments in the A4C view.

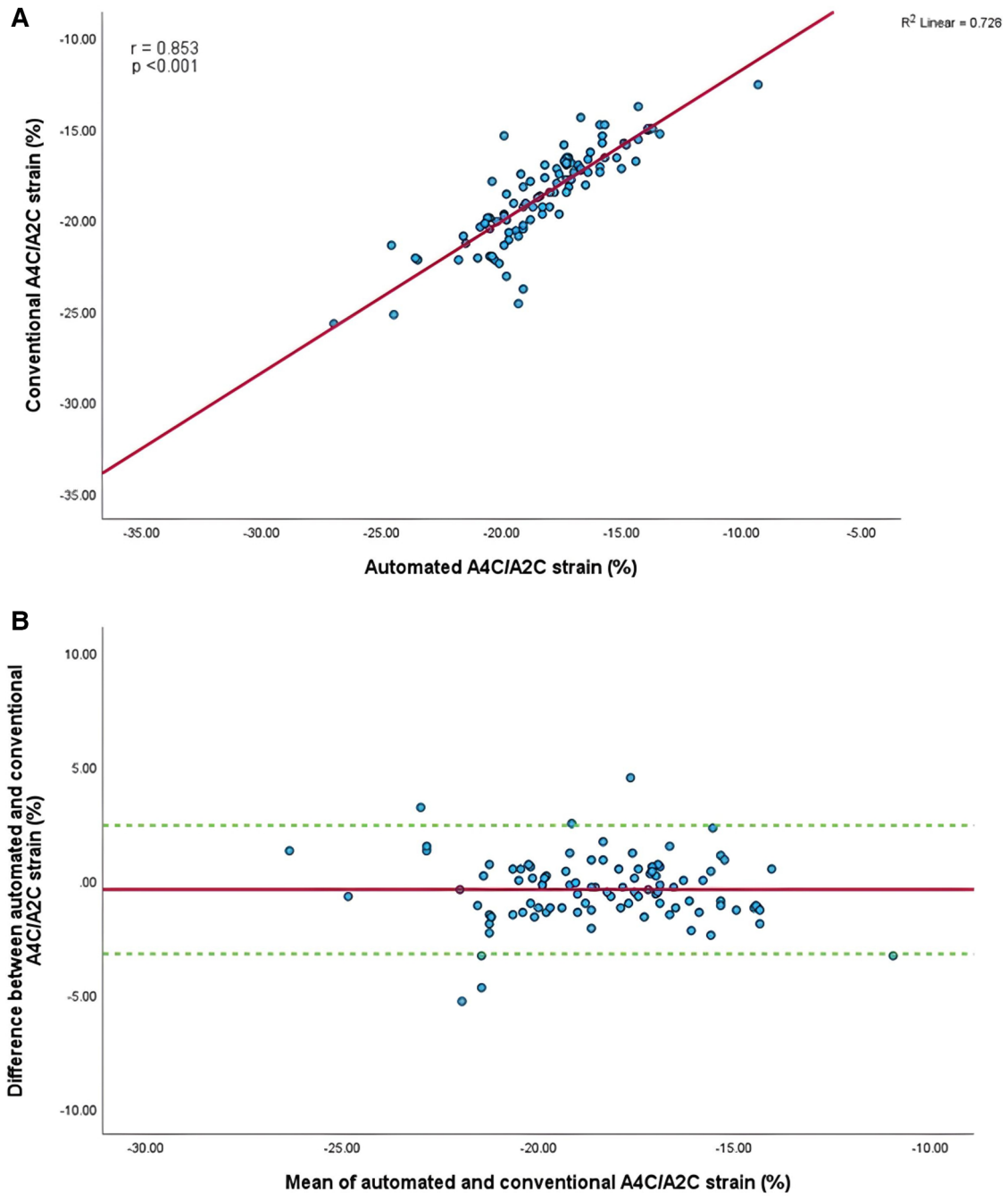


FIGURE 9
(A) Correlation between conventional and AI-automated strain in apical-four/-two chamber view. (B) Bland-Altman plot of conventional and AI-automated strain in apical-four/-two chamber view.

In contrast to previous studies that excluded echo studies where image quality were deemed substandard (14), the present analysis did not exclude these patients and is therefore relevant to real-world clinical practice. All oncology patients that were administered trastuzumab and underwent echo surveillance were included to minimise selection bias and reflect real-world patient cohorts, including known imaging challenges often specific to cardio-oncology patients such as radiotherapy, breast reconstruction surgeries, mastectomy and breast implantation (18). This might explain the lower rate of successful acquisition compared to previous trials as the availability of diagnostic quality images are reduced. Conversely, the possibility for over-analysis in potentially non-feasible images should not be excluded. The likelihood of the operator repeatedly adjusting the region of interest in the presence of limited or absence of endocardial border definition to "inaccurately" create a GLS value that is consistent with visual assessment is not uncommon and ought to be considered.

TABLE 2 Mean values and standard deviation of conventional GLS and AI-automated GLS at individual timepoints during trastuzumab therapy.

| | T0 | T1 | T2 | T3 | T4 | T5 |
|---|---|---|---|---|---|---|
| GLS (CON) | −20.1 ± 2.6 | −19.0 ± 2.8 | −18.8 ± 2.3 | −18.3 ± 2.3 | −19.1 ± 1.9 | −19.3 ± 2.5 |
| GLS (AI) | −19.0 ± 2.2 | −18.6 ± 1.8 | −18.2 ± 2.7 | −17.3 ± 3.2 | −18.1 ± 2.4 | −18.4 ± 2.0 |
| A4C (CON) | −20.0 ± 3.1 | −19.0 ± 3.2 | −18.9 ± 2.7 | −18.6 ± 3.0 | −19.1 ± 2.3 | −19.4 ± 2.6 |
| A4C(AI) | −19.8 ± 3.9 | −19.4 ± 4.0 | −18.5 ± 3.5 | −18.8 ± 3.7 | −19.3 ± 3.1 | −19.1 ± 3.1 |
| A2C CON) | −20.8 ± 2.9 | −19.2 ± 3.7 | −18.5 ± 4.1 | −18.8 ± 2.7 | −19.6 ± 2.4 | −20.0 ± 3.5 |
| A2C (AI) | −20.7 ± 4.2 | −19.2 ± 4.2 | −18.4 ± 3.9 | −18.1 ± 3.2 | −19.1 ± 4.4 | −19.7 ± 4.1 |
| A3C CON) | −19.6 ± 3.1 | −18.8 ± 3.8 | −18.8 ± 2.5 | −18.3 ± 2.9 | −19.3 ± 2.8 | −18.8 ± 3.3 |
| A3C (AI) | −15.3 ± 3.0 | −15.5 ± 2.0 | −15.7 ± 3.5 | −13.9 ± 3.7 | −15.5 ± 2.8 | −14.9 ± 3.7 |
| LVEF (CON) | 61.5 ± 4.6 | 59.8 ± 5.7 | 58.7 ± 6.6 | 58.5 ± 6.3 | 58.9 ± 5.7 | 59.5 ± 5.7 |
| LVEF (AI) | 63.4 ± 6.9 | 62.4 ± 6.7 | 58.8 ± 9.2 | 58.9 ± 8.7 | 58.9 ± 7.4 | 62 ± 6.0 |

Data are expressed as mean ± standard deviation.
AI, artificial intelligence; A4C, apical-four chamber; A2C, apical-two chamber; A3C, apical-three chamber; CON, conventional; GLS, global longitudinal strain; LVEF, left ventricular ejection fraction.

Previous validation studies (8, 9, 14, 19) comparing AI-automated and conventional methods have reported good feasibility and correlation values, often in patient groups with ischaemia-related heart diseases and other pathologies unrelated to chemotherapy. In the setting of cardio-oncology, our results are in line with previously reported evidence which demonstrated a reasonable correlation between AI-derived GLS and LVEF values to the conventional method, suggesting that there were no considerable differences between method of assessments.

Although our reported values were lower compared to the literature, this may be influenced by the preselection of subjects with segments suited for assessment in previous studies. Our findings also demonstrated that serial monitoring of trastuzumab-treated oncology patients with AI-assisted technology to detect subtle changes in LVEF and GLS may be done with similar certainty to conventional assessment with the values generated from both methods being largely similar.

A significant difference in LVEF was observed at one timepoint although this may be attributed to smaller sample size at the final follow-up. Further work will be required to assess longitudinal echocardiographic trends in addition to correlation between AI-automated and conventional analyses, and there may be systematic differences in absolute values whether related to the vendor or system used.

In the cardiotoxic cohort, while the sample size was small, both methods demonstrated a similar temporal trend highlighting the potential for AI-automated methods to reliability detect LV functional deterioration. Such findings suggest that AI-automated LV assessments represent a valuable method of serial echocardiographic monitoring in longitudinal patient care and can build a case for future prospective studies in this area.

TABLE 3 AI-Automated and conventional global longitudinal strain and left ventricular ejection fraction at each timepoint.

| | Method of assessment | | Pearson correlation coefficient | *p*-value |
|---|---|---|---|---|
| | Conventional (%) | Automated (%) | | |
| GLS (T0) | −20.1 ± 2.6 | −19.0 ± 2.2 | 0.835 | >0.001 |
| GLS (T1) | −19.0 ± 2.8 | −18.6 ± 1.8 | 0.856 | >0.001 |
| GLS (T2) | −18.8 ± 2.3 | −18.2 ± 2.7 | 0.779 | 0.004 |
| GLS (T3) | −18.3 ± 2.3 | −17.3 ± 3.2 | 0.761 | 0.020 |
| GLS (T4) | −19.1 ± 1.9 | −18.1 ± 2.4 | 0.782 | 0.017 |
| GLS (T5) | −19.3 ± 2.5 | −18.4 ± 2.0 | 0.727 | 0.023 |
| LVEF (T0) | 61.5 ± 4.6 | 63.4 ± 6.9 | 0.811 | 0.019 |
| LVEF (T1) | 59.8 ± 5.7 | 62.4 ± 6.7 | 0.715 | 0.031 |
| LVEF (T2) | 58.7 ± 6.6 | 58.8 ± 9.2 | 0.866 | >0.001 |
| LVEF (T3) | 58.5 ± 6.3 | 58.9 ± 8.7 | 0.838 | 0.011 |
| LVEF (T4) | 58.9 ± 5.7 | 58.9 ± 7.4 | 0.844 | 0.006 |
| LVEF (T5) | 59.5 ± 5.7 | 62 ± 6.0 | 0.613 | 0.032 |

Data are expressed as mean ± standard deviation.
GLS, global longitudinal strain; LVEF, left ventricular ejection fraction.

## Study limitations

There are a few potential limitations associated with the present analysis that deserves to be mentioned. First, we only studied patients in sinus rhythm, thus data could not be extrapolated from patients with irregular heart rhythms. Additionally, our study included a relatively small sample size. Despite this, our patient cohort included all patients during the study period to reflect a real-world clinical setting and is the first to study functional changes in this specific patient cohort, thereby providing valuable insight into the application of AI-automated analysis in serial echocardiographic studies in trastuzumab-treated patients. Our report and early insights thereby provide a basis for future studies to expand upon. Second, the potential vendor differences in AI-imaging software for strain and LVEF analysis due to differences in AI-algorithms should be noted. Third, is the lack of gold standard reference to compare our strain and EF measurements. However, the primary objective was to determine the level of correlation between AI-automated and conventional methods thus identifying the "true" reference value is of lesser significance. We therefore used the current clinically accepted semi-automated approach as the comparator. Finally, the analysis was conducted retrospectively which meant that it suffered from the inherent limitations of a restrospective study design. Nevertheless, this report describes a straightforward comparision of imaging as opposed to patient outcomes, thus selection bias is of lesser relevance.

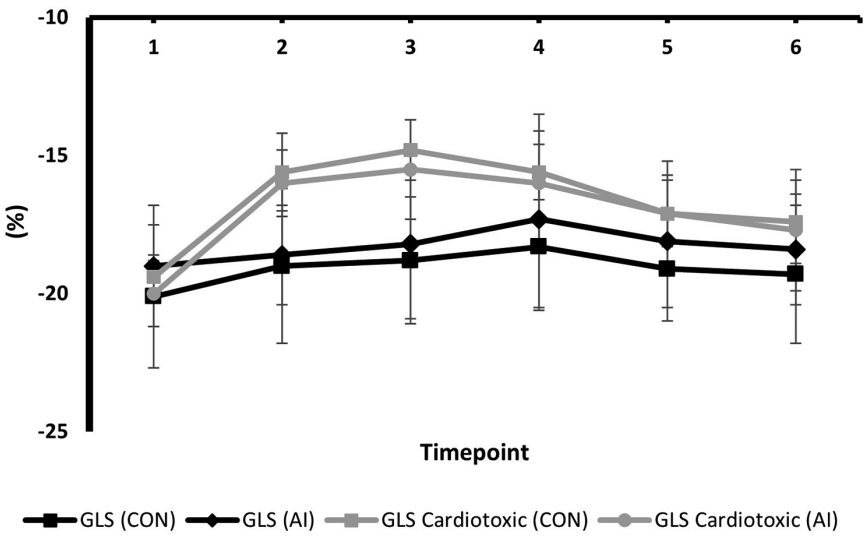


FIGURE 10
Mean values and standard deviation of conventional GLS and AI-automated GLS at individual timepoints during trastuzumab therapy in the study population.

## Future research directions

With increasing echocardiographic demands surpassing clinical capacity in the face of a shortage of echocardiographers, there is now an urgent need for the active incorporation of AI guided technology to assist, or potentially substitute the need for operator input into analysis of advanced echocardiographic techniques. Consequently, software solutions must possess the accuracy to where it could be confidently applied irrespective of the GLS experience of the operator. There are a number of challenges in the widespread clinical implementation of AI echocardiography, none of which are considered insurmountable.

The future appears positive for the application of AI in echocardiography and significant advances are anticipated to address the current knowledge gaps. Future work should explore whether: (1) AI-based assessment is superior to less experienced humans, (2) image rejection threshold appropriateness, (3) accuracy and reproducibility of automated, semi-automated and manually generated data, and (4) improvements in post-processing time and overall workflow on echocardiography services.

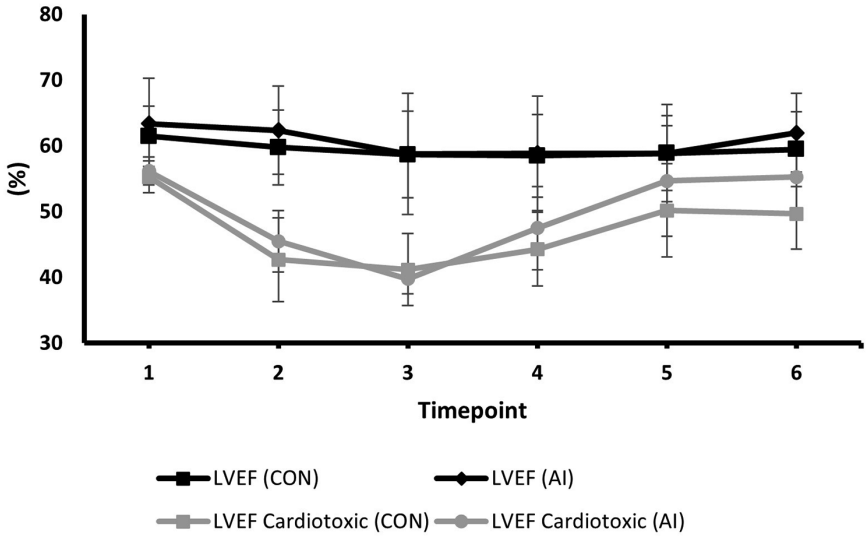


FIGURE 11
Mean values and standard deviation of conventional LVEF and AI-automated LVEF at individual timepoints during trastuzumab therapy in the study population.

## Conclusions

Despite enthusiasm for the application of AI technology in healthcare, it is yet to be widely embraced in the echocardiographic community. Due to significant limitations and knowledge gaps in automation, AI technology in echocardiography remains premature for clinical use if adopted completely independent of operator intervention. Instead, at present, it could be a useful unbiased "second opinion" for "experienced" practitioners. Our analysis is supportive of prospective studies into the utility and application of AI-based analysis of heart function by echocardiography in patients receiving potentially cardiotoxic chemotherapy.

## Data availability statement

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

## Ethics statement

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

## Author contributions

SH and YL contributed to conception and design of the study. JJ performed the statistical analysis. JJ wrote the first draft of the manuscript. SH, BL, JJ critically revised and wrote sections of the manuscript. All authors contributed to manuscript revision, read and approved the submitted version.

## Conflict of interest

The Royal Wolverhampton NHS Trust received an NHSx phase 4 award for the use of Ultromics artificial intelligence stress echo analysis software in my trust (Royal Wolverhampton NHS Trust). This comprised cost towards IT set up, clinical implementation, and cost-free provision of the novel Ultromics Echo Core Pro software for a one-year period. SSH has research agreements with Ligence Heart and Ventripoint Medical System.

The remaining 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. Sawaya H, Sebag IA, Plana JC, Januzzi JL, Ky B, Cohen V, et al. Early detection and prediction of cardiotoxicity in chemotherapy-treated patients. *Am J Cardiol.* (2011) 107(9):1375–80. doi: 10.1016/j.amjcard.2011.01.006

2. Thavendiranathan P, Negishi T, Coté MA, Penicka M, Massey R, Cho GY, et al. Single versus standard multiview assessment of global longitudinal strain for the diagnosis of cardiotoxicity during cancer therapy. *JACC Cardiovasc Imaging.* (2018) 11(8):1109–18. doi: 10.1016/j.jcmg.2018.03.003

3. Heimdal A, Støylen A, Torp H, Skjærpe T. Real-time strain rate imaging of the left ventricle by ultrasound. *J Am Soc Echocardiogr.* (1998) 11(11):1013–9. doi: 10.1016/S0894-7317(98)70151-8

4. Leitman M, Lysyansky P, Sidenko S, Shir V, Peleg E, Binenbaum M, et al. Two-dimensional strain–a novel software for real-time quantitative echocardiographic assessment of myocardial function. *J Am Soc Echocardiogr.* (2004) 17(10):1021–9. doi: 10.1016/j.echo.2004.06.019

5. Thavendiranathan P, Poulin F, Lim KD, Plana JC, Woo A, Marwick TH. Use of myocardial strain imaging by echocardiography for the early detection of cardiotoxicity in patients during and after cancer chemotherapy: a systematic review. *J Am Coll Cardiol.* (2014) 63(25):2751–68. doi: 10.1016/j.jacc.2014.01.073

6. Ye L, Yang ZG, Selvanayagam JB, Luo H, Yang TZ, Perry R, et al. Myocardial strain imaging by echocardiography for the prediction of cardiotoxicity in chemotherapy-treated patients: a meta-analysis. *JACC Cardiovasc Imaging.* (2020) 13(3):881–2. doi: 10.1016/j.jcmg.2019.09.013

7. Sawaya H, Sebag IA, Plana JC, Januzzi JL, Ky B, Tan TC, et al. Assessment of echocardiography and biomarkers for the extended prediction of cardiotoxicity in patients treated with anthracyclines, taxanes, and trastuzumab. *Circ Cardiovasc Imaging.* (2012) 5(5):596–603. doi: 10.1161/CIRCIMAGING.112.973321

8. Knackstedt C, Bekkers SC, Schummers G, Schreckenberg M, Muraru D, Badano LP, et al. Fully automated versus standard tracking of left ventricular ejection fraction and longitudinal strain: the FAST-EFs multicenter study. *J Am Coll Cardiol.* (2015) 66(13):1456–66. doi: 10.1016/j.jacc.2015.07.052

9. Salte IM, Østvik A, Smistad E, Melichova D, Nguyen TM, Karlsen S, et al. Artificial intelligence for automatic measurement of left ventricular strain in echocardiography. *JACC Cardiovasc Imaging.* (2021) 14(10):1918–28. doi: 10.1016/j.jcmg.2021.04.018

10. Upton R, Beqiri A, Parker A, Hawkes W, Gao S, Porumb M, et al. Automated echocardiographic detection of severe coronary artery disease using artificial intelligence. *JACC Cardiovasc Imaging.* (2022) 15(5):715–27. doi: 10.1016/j.jcmg.2021.10.013

11. Lyon AR, López-Fernández T, Couch LS, Asteggiano R, Aznar MC, Bergler-Klein J, et al. 2022 ESC guidelines on cardio-oncology developed in collaboration with the European Hematology Association (EHA), the European Society for Therapeutic Radiology and Oncology (ESTRO) and the International Cardio-Oncology Society (IC-OS) developed by the task force on cardio-oncology of the European Society of Cardiology (ESC). *Eur Heart J.* (2022) 43(41):4229–361. doi: 10.1093/eurheartj/ehac244

12. Khan AM, Wiegers SE. The importance of being expert: is it time to revisit the concept? *J Am Soc Echocardiogr.* (2012) 25(2):218–9. doi: 10.1016/j.echo.2011.12.001

13. Kitano T, Nabeshima Y, Abe Y, Otsuji Y, Takeuchi M. Accuracy and reliability of novel semi-automated two-dimensional layer specific speckle tracking software for quantifying left ventricular volumes and function. *PLoS One.* (2019) 14(8):e0221204. doi: 10.1371/journal.pone.0221204

14. Kawakami H, Wright L, Nolan M, Potter EL, Yang H, Marwick TH. Feasibility, reproducibility, and clinical implications of the novel fully automated assessment for global longitudinal strain. *J Am Soc Echocardiogr.* (2021) 34(2):136–45. doi: 10.1016/j.echo.2020.09.011

15. Medvedofsky D, Kebed K, Laffin L, Stone J, Addetia K, Lang RM, et al. Reproducibility and experience dependence of echocardiographic indices of left

ventricular function: side-by-side comparison of global longitudinal strain and ejection fraction. *Echocardiography*. (2017) 34(3):365–70. doi: 10.1111/echo.13446

16. Chen Y, Hua W, Yang W, Shi Z, Fang Y. Reliability and feasibility of automated function imaging for quantification in patients with left ventricular dilation: comparison with cardiac magnetic resonance. *Int J Cardiovasc Imaging*. (2022) 38 (6):1267–76. doi: 10.1007/s10554-021-02510-x

17. Nadruz W, West E, Sengeløv M, Grove GL, Santos M, Groarke JD, et al. Cardiovascular phenotype and prognosis of patients with heart failure induced by cancer therapy. *Heart*. (2019) 105(1):34–41. doi: 10.1136/heartjnl-2018-313234

18. Plana JC, Galderisi M, Barac A, Ewer MS, Ky B, Scherrer-Crosbie M, et al. Expert consensus for multimodality imaging evaluation of adult patients during and after cancer therapy: a report from the American society of echocardiography and the European association of cardiovascular imaging. *Eur Heart J Cardiovasc Imaging*. (2014) 15(10):1063–93. doi: 10.1093/ehjci/jeu192

19. Wierzbowska-Drabik K, Hamala P, Roszczyk N, Lipiec P, Plewka M, Kręcki R, et al. Feasibility and correlation of standard 2D speckle tracking echocardiography and automated function imaging derived parameters of left ventricular function during dobutamine stress test. *Int J Cardiovasc Imaging*. (2014) 30(4):729–37. doi: 10.1007/s10554-014-0386-z

frontiers | Frontiers in Cardiovascular Medicine

TYPE Original Research
PUBLISHED 17 November 2023
DOI 10.3389/fcvm.2023.1203400



OPEN ACCESS

EDITED BY
Gongning Luo,
Harbin Institute of Technology, China

REVIEWED BY
Suyu Dong,
Northeast Forestry University, China
Runnan He,
Peng Cheng Laboratory, China

*CORRESPONDENCE
Yongjiang Cai
caiyj2000@sina.cn
Yuping Duan
yuping.duan@tju.edu.cn

†These authors have contributed equally to this work and share first authorship

RECEIVED 26 April 2023
ACCEPTED 18 October 2023
PUBLISHED 17 November 2023

CITATION
Zhou Y, Yang L, Guo Y, Xu J, Li Y, Cai Y and Duan Y (2023) Joint 2D–3D cross-pseudo supervision for carotid vessel wall segmentation.
Front. Cardiovasc. Med. 10:1203400.
doi: 10.3389/fcvm.2023.1203400



# Joint 2D–3D cross-pseudo supervision for carotid vessel wall segmentation

Yahan Zhou[1,2†], Lin Yang[3†], Yuan Guo[1,4†], Jing Xu[2], Yutong Li[4], Yongjiang Cai[3*] and Yuping Duan[1*]

[1]School of Mathematical Sciences, Beijing Normal University, Beijing, China, [2]School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, China, [3]Health Management Center, Peking University Shenzhen Hospital, Peking University, Shenzhen, China, [4]Center for Applied Mathematics, Tianjin University, Tianjin, China

**Introduction:** The segmentation of the carotid vessel wall using black-blood magnetic resonance images was a crucial step in the diagnosis of atherosclerosis. The objective was to accurately isolate the region between the artery lumen and outer wall. Although supervised learning methods achieved remarkable accuracy in vessel segmentation, their effectiveness remained limited due to their reliance on extensive labeled data and human intervention. Furthermore, when confronted with three-dimensional datasets featuring insufficient and discontinuous label data, these learning-based approaches could lose their efficacy. In this paper, we proposed a novel Joint 2D–3D Cross-Pseudo Supervision (JCPS) method for accurate carotid vessel wall segmentation.

**Methods:** In this study, a vascular center-of-gravity positioning module was developed to automatically estimate the region of blood vessels. To achieve accurate segmentation, we proposed a joint 2D–3D semi-supervised network to model the three-dimensional continuity of vascular structure. In addition, a novel loss function tailored for vessel segmentation was introduced, consisting of four components: supervision loss, cross-pseudo supervision loss, pseudo label supervision loss, and continuous supervision loss, all aimed at ensuring the accuracy and continuity of the vessel structure. In what followed, we also built up a user-friendly Graphical User Interface based on our JCPS method for end-users.

**Results:** Our proposed JCPS method was evaluated using the Carotid Artery Vessel Wall Segmentation Challenge dataset to assess its performance. The experimental results clearly indicated that our approach surpassed the top 10 methods on the leaderboard, resulting in a significant enhancement in segmentation accuracy. Specifically, we achieved an average Dice similarity coefficient increase from 0.775 to 0.806 and an average quantitative score improvement from 0.837 to 0.850, demonstrating the effectiveness of our proposed JCPS method for carotid artery vessel wall segmentation.

**Conclusion:** The experimental results suggested that the JCPS method had a high level of generalization performance by producing pseudo labels that were comparable with software annotations for data-imbalanced segmentation tasks.

KEYWORDS

carotid artery wall, atherosclerosis, black-blood vessel wall MRI, semi-supervised learning, continuous prior, Graphical User Interface

# 1. Introduction

Cardio-cerebrovascular disease (CCVD) manifests as systemic vasculopathy affecting the heart and brain, making it a global public health concern and a leading cause of mortality. Vascular medical images are extensively used to visualize the three-dimensional (3D) morphology of cardiac and cerebral vessels, playing an essential role in the diagnosis and treatment of CCVD. Blood vessel segmentation is aimed at extracting well-defined vessel structures from these medical images. Therefore, computer-based automatic detection and segmentation of blood vessel walls are of great clinical significance, as they represent a crucial step in ensuring precise diagnosis, early intervention, and surgical planning for CCVD.

However, medical image segmentation has not been adequately handled due to the complexity and diversity of the medical images. Consequently, researchers have dedicated significant efforts to develop effective segmentation methods, including both traditional and deep learning–based approaches in recent years. Traditional image segmentation techniques, such as thresholding (1, 2), region growing method (3–5), active contour model (6, 7), and level set method (8–10), have been widely recognized. However, these methods have their limitations. They are often semi-automatic and rely on human input, making them prone to noise interference and intensity unevenness. Deep learning methods have shown remarkable performance in medical image segmentation tasks. For instance, the fully convolutional network (FCN) can take inputs of arbitrary sizes and produce correspondingly sized output with efficient inference and learning for image segmentation tasks. Since then, the FCN has been extensively used in the fields of medical image segmentation (11–13), e.g., the segmentation of breast tumors on MR images (13) and the segmentation of human torsos on CT images (12). However, the FCN suffered from issues such as inaccurate edges and loss of details. The U-Net architecture (14) used the jump connections to effectively realize the integration of features and performed more efficiently in training. Since then, it was widely used for medical image segmentation (15–18). To deal with small organs or tissues, a coarse-to-fine segmentation framework was established to enhance the accuracy by extracting regions of interest (ROI) during the coarse segmentation stage and using ROI as inputs for the fine segmentation network. These kinds of approaches have achieved satisfactory performance in various image segmentation tasks (19, 20) and were also successfully applied to handle vascular segmentation problems (21–24).

Indeed, the vessel segmentation had unique characteristics such as the significant imbalance of blood vessel proportions, complex structures of blood vessels, and difficulties in acquiring blood vessel labels. Samber et al. (25) applied a convolutional neural network (CNN) to segment the carotid artery after extensive manual preprocessing to improve carotid artery segmentation accuracy. Oliveira et al. (26) combined the multiscale analysis provided by the stationary wavelet transform with a multiscale FCN for the purpose of automatic vessel segmentation. Ni et al. (27) proposed a global channel attention network (GCA-Net) to segment intracranial blood vessels. Liu et al. (28) developed a novel residual depth-wise over-parameterized convolutional (ResDO-conv) network for automatic and accurate retinal vessel segmentation. Imran et al. (29) designed an intelligence-based automated shallow network with high performance and low cost named Feature Preserving Mesh Network (FPM-Net) for the accurate segmentation of retinal vessels. Tan et al. (30) proposed the U-Net using local phase congruency and orientation scores (UN-LPCOS), which showed a remarkable ability to identify and segment small retinal vessels. However, the aforementioned methods were all built up for dealing with 2D vessel segmentation tasks. Zhou et al. (31) proposed an approach that combined a voxel-based fully convolution network (Voxel-FCN) and a continuous max-flow module to automatically segment the carotid vessel wall. Tetteh et al. (32) presented the DeepVesselNet to extract vessel trees in 3D angiographic volumes. Xia et al. (33) proposed an edge-reinforced network (ER-Net) for 3D vessel-like structure segmentation, which incorporates a reverse edge attention module. Alblas et al. (34) formulated the vessel wall segmentation as a multi-task regression problem in polar coordinates to automatically segment the carotid artery wall with high accuracy. However, the performance of these methods was hindered when insufficient labeled data were available. As such, semi-supervised segmentation methods became increasingly popular to alleviate the demand for labeled data, which could be broadly classified into entropy-minimization–based methods (35) and consistency determination–based methods (36–39). Recently, a novel approach known as cross-pseudo supervision (CPS) has emerged to enhance performance in semi-supervised learning problems (40, 41). The CPS method enforces consistency among slightly different network outputs, leading to satisfactory results even with limited labeled data. More importantly, the CPS method effectively avoids confronting the strong coupling between the teacher and student networks (42).

In this paper, we presented a novel coarse-to-fine vessel wall segmentation method. In the coarse segmentation stage, we developed a modified Deeplabv3+ network to estimate both the vessel location and signed distance function. Based on the coarse segmentation, we calculated the location of the blood vessel's center of gravity using the first-order moment method. This information was then utilized to crop the original images, specifically selecting the ROI that contained the vessels. In the fine segmentation stage, we proposed a joint 2D–3D CPS network to ideally exploit the spatial information of 3D volumes and used the continuity prior of blood vessels, which helped enhance the blood vessel features and improved the segmentation accuracy. It is worth mentioning that the CPS operation involved both labeled and unlabeled data, which improved the generalizability using the lower cost of manual annotation. In comparison to existing coarse-to-fine methods, our model incorporated both the position of the center of gravity and the continuity of the target blood vessel to enhance the utilization of carotid artery features. The proposed method was evaluated on the 3D carotid black-blood MRI dataset obtained from the Carotid Artery Vessel Wall Segmentation Challenge, which was a

typical semi-supervised segmentation task with only around 20% labeled data. Through numerical experiments, we were able to demonstrate that our JCPS method surpassed the state-of-the-art results on the competition's leader board, exhibiting a significant improvement in segmentation accuracy when compared to both the baseline U-Net model and single CPS model. Furthermore, we designed an effective and user-friendly Graphical User Interface (GUI) for the automated segmentation of MRI images of black-blood carotid arteries, aimed at providing valuable assistance to clinicians in their diagnostic.

The rest of this paper is organized as follows. Section 2. introduces our joint 2D–3D cross-pseudo supervision method, including coarse and fine segmentation models, a loss function, and implementation details. Section 3. presents experimental results and ablation studies. We briefly discuss the proposed approach and conclude with a summary and possible future work in Section 4.

# 2. Materials and methods

## 2.1. Data source

The training set and test set data used in this study were both from the Carotid Artery Vessel Wall Segmentation Challenge, in which 25 cases with various carotid vessel wall conditions were used as the training set, and the other 25 cases with various carotid vessel wall conditions were used as the test set. A total of 12,920 vessel wall images of sufficient quality in the training set (2,584 images with manual contour labels) were used for training, and a total of 2,412 images with manual contour labels in the test set were used for testing. Each vessel wall image is an axial slice of a carotid black-blood MRI image, and the size of each original image is of $720 \times 720$ in order to facilitate subsequent evaluation and meet the Carotid Artery Vessel Wall Segmentation Challenge.

## 2.2. Our approach

The proposed automatic carotid artery vessel wall segmentation approach, known as the Joint 2D–3D Cross-Pseudo Supervision (JCPS), comprised two stages, as illustrated in Figure 1. The coarse segmentation model consisted of a vascular center-of-gravity positioning model, and the fine segmentation model consisted of a joint 2D–3D CPS network.

### 2.2.1. Coarse segmentation

Since the target vessel occupied only a small fraction of the whole image and varied in sizes and locations in 2D axial slices, we needed to automatically determine the approximated location of the center of gravity for the blood vessel. This was crucial for providing a region of interest specific to the local vessel area, which would be utilized for subsequent vessel wall segmentation. To achieve this, we developed a vascular center-of-gravity positioning module within the coarse segmentation

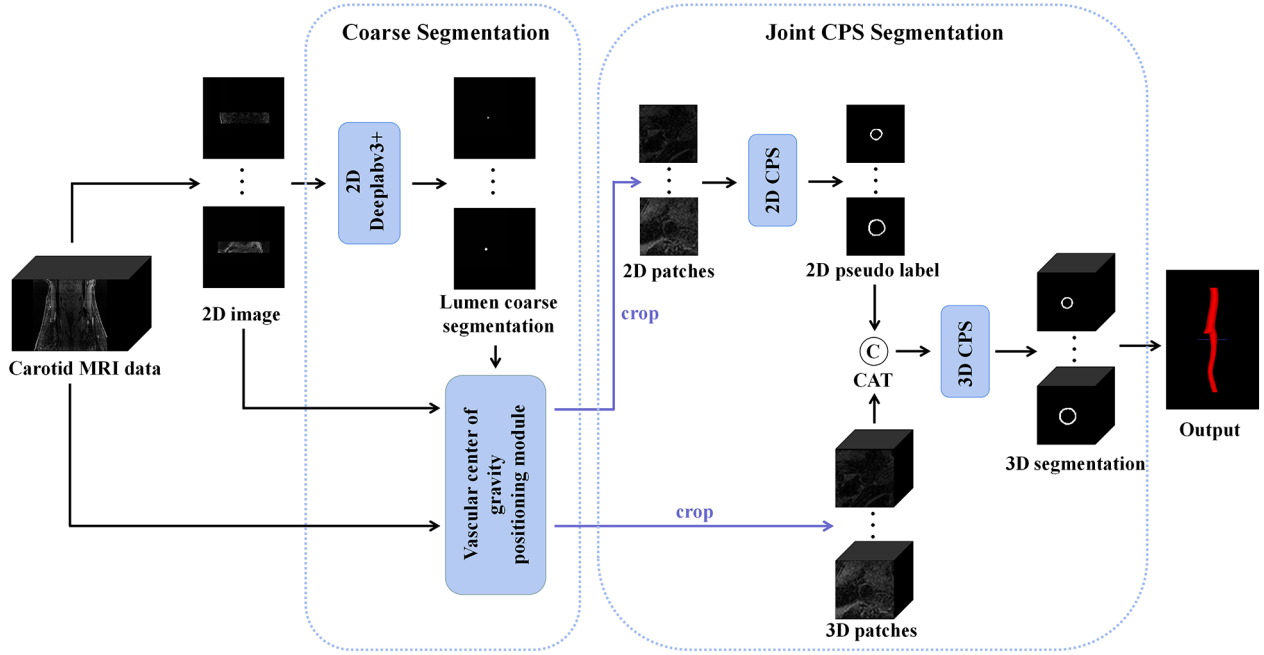


FIGURE 1
An overview of the proposed JCPS framework, where CAT is short for concatenation. In the coarse segmentation model, a 2D Deeplabv3+ network was employed to locate objects in high-resolution images. The vascular center-of-gravity positioning module, derived from the lumen coarse segmentation, was utilized to identify the vascular center of gravity in both 2D and 3D original images. For the fine segmentation model, the CPS network was adopted, enabling efficient utilization of limited labeled data and a large amount of unlabeled data to achieve precise segmentation.

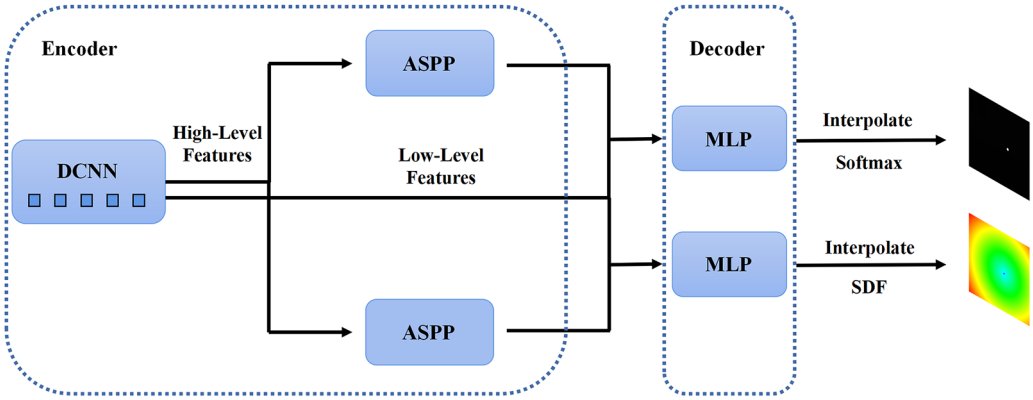


FIGURE 2
The structure of the 2D Deeplabv3+ network with the encoder module and the decoder module. The high-level feature information is first transferred through the deep convolutional network into two parallel feature pyramid modules, and each enters the Multi-Layer Perceptron (MLP) of the decoder module, which simultaneously outputs the pixel-level classification and the corresponding signed distance function.

model to estimate the center of the vessel. The backbone of the coarse segmentation model was chosen as DeepLabv3+ (43), which had been commonly used for medical segmentation (44, 45).

In the first stage, we identified 2D slices with the sufficient image quality from 3D carotid black-blood MRI images $I_{3D} \in \mathbb{R}^{D\times H\times W}$, where $D$, $H$, and $W$ represent the depth, height, and width of the 3D volume, respectively. The input and output of the coarse segmentation model were represented as $I_{2D} \in \mathbb{R}^{H\times W}$ and $Q_{2D} \in \mathbb{R}^{H\times W}$, respectively. Different from the classical Deeplabv3+ network, our approach involved learning both the pixel-level classification task and the signed distance function. These components were utilized to achieve binary classification results for the lumen area and to accurately capture the lumen boundary, respectively. For a detailed overview of the network architecture, please refer to **Figure 2**. The input image was processed utilizing a deep convolutional neural network (DCNN) to extract both low-level and high-level features. Following that, the high-level features were fed into the Atrous Spatial Pyramid Pooling (ASPP) module, which consists of parallel dilated convolutional layers and pooling layers to extend the receptive field. It is capable of extracting relevant features from original images with a relatively low proportion of vessel regions, and subsequently merging them at different scales, thereby enhancing the accuracy of the coarse segmentation stage. Within the decoder, the Multi-Layer Perceptron (MLP) module concatenated the low-level and high-level features derived from the encoder. Subsequently, the outputs were restored to the original image resolution by employing interpolation and upsampling techniques. For a detailed illustration of the network structure, please refer to **Figure 3**. Based on the coarse segmentation, we used the vessel center-of-gravity positioning model to crop the data into 2D or 3D patches, which were utilized as inputs for the fine segmentation model. Subsequently, the fine segmentation model accurately predicted binary labels for 3D carotid black-blood

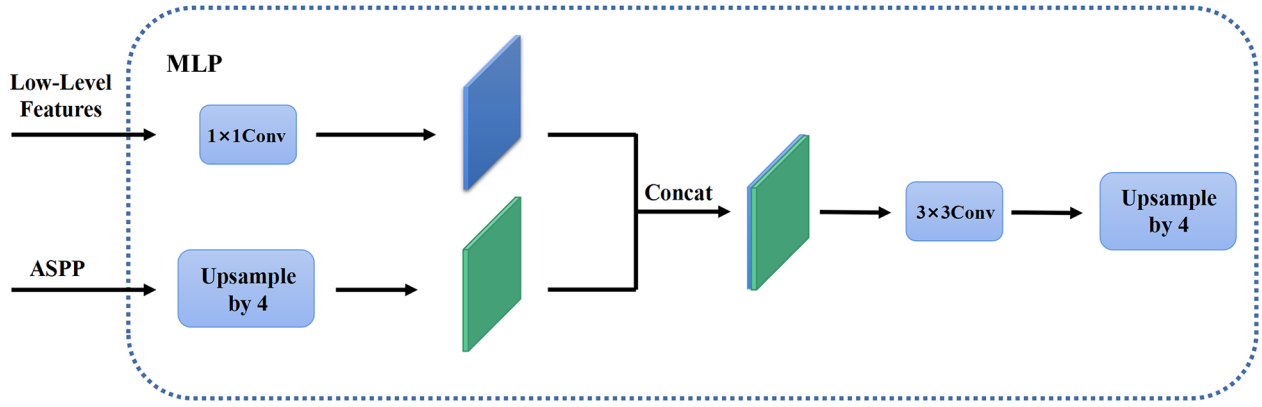


FIGURE 3
The MLP module in the decoder. The high-level feature information through the feature pyramid module is upsampled by quadruple interpolation and stacked with the low-level feature information in the channel dimension, while the output is the same as the original input image resolution after a 3 × 3 convolution and a quadruple upsampling.

MRI volumes, with "0" representing the background and "1" denoting the vessel wall.

### 2.2.2. 2D CPS network

To calculate the center of gravity of the 2D lumen area, we utilized the first-order moment as follows

$$G_{2D} = g_{2D}(Q_{2D}), \tag{1}$$

with

$$g_{2D}(Q_{2D}) = \left( \frac{\sum_I \sum_J i \cdot Q_{2D}(i,j)}{\sum_I \sum_J Q_{2D}(i,j)}, \frac{\sum_I \sum_J j \cdot Q_{2D}(i,j)}{\sum_I \sum_J Q_{2D}(i,j)} \right),$$

where $Q_{2D}(i, j)$ represents the gray value of the binary segmentation map $Q_{2D}$ at point $(i, j)$. Obviously, the center of gravity of the segmented lumen was an approximation for the centerline of the vessels. Subsequently, the estimated center of gravity was used to crop local patches $X_{2D} \in \mathbb{R}^{h \times w}$ with a fixed size $h \times w$. These patches were then employed as inputs for the fine segmentation model.

Assuming that the manual labels were randomly distributed in the 3D carotid black-blood MRI volumes, comprising approximately 20% of the total slices, we endeavored to exploit a limited amount of 2D labeled data and a substantial amount of 2D unlabeled data to generate more precise pseudo labels for the latter. To achieve this, we employed a 2D semi-supervised method CPS to integrate the pseudo labels and consistency regularization, thereby maximizing the utilization of both labeled and unlabeled data. Specifically, the U-Net architecture was adopted as the backbone of the CPS network, as depicted in Figure 4. The U-Net consisted of a contracting path and an expansive path. Notably, the number of channels in the network was halved compared to the traditional U-Net. Each convolutional layer (Conv) was followed by a batch normalization (BN) and a rectification linear unit (ReLU), denoted as a composite layer (Conv-BN-ReLU).

As depicted in Figure 4, two U-Net networks, denoted as $f(\varphi_1)$ and $f(\varphi_2)$, were initially generated. These networks shared the same structure but had different initialization parameter. The patches $X_{2D}$, obtained from the coarse segmentation stage and containing both labeled and unlabeled data, served as inputs for both U-Nets. Their objective was to estimate the segmentation confidence maps $P^n_{2D} \in \mathbb{R}^{C \times h \times w}$ $(n = 1, 2)$, which can be expressed as

$$P^n_{2D} = f(X_{2D}; \varphi_n), \tag{2}$$

where $C$ represented the number of categories, i.e., the images were divided into $C$ categories. The corresponding one-hot labels $S^n_{2D} \in \mathbb{R}^{h \times w}$ $(n = 1, 2)$ were then obtained through the argmax operation. These labels were considered the pseudo labels predicted by the two networks. During the training of unlabeled

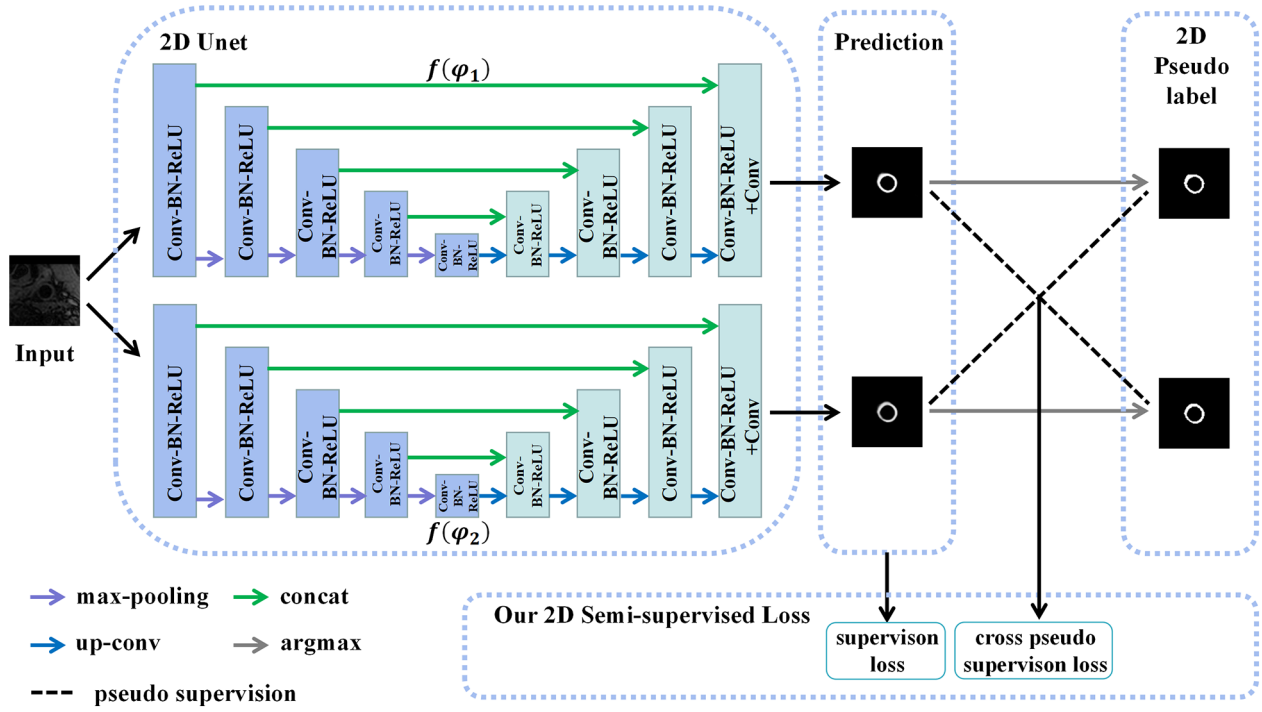


FIGURE 4
The structure of the 2D CPS network, which consists of two 2D U-Net networks with varying parameters. Our loss function comprises the supervision loss and cross-pseudo supervision loss, where supervision loss represents the loss between the output of U-Net and the ground truth, and cross-pseudo supervision loss represents the loss of mutual supervision of the pseudo labels of the two U-Net outputs.

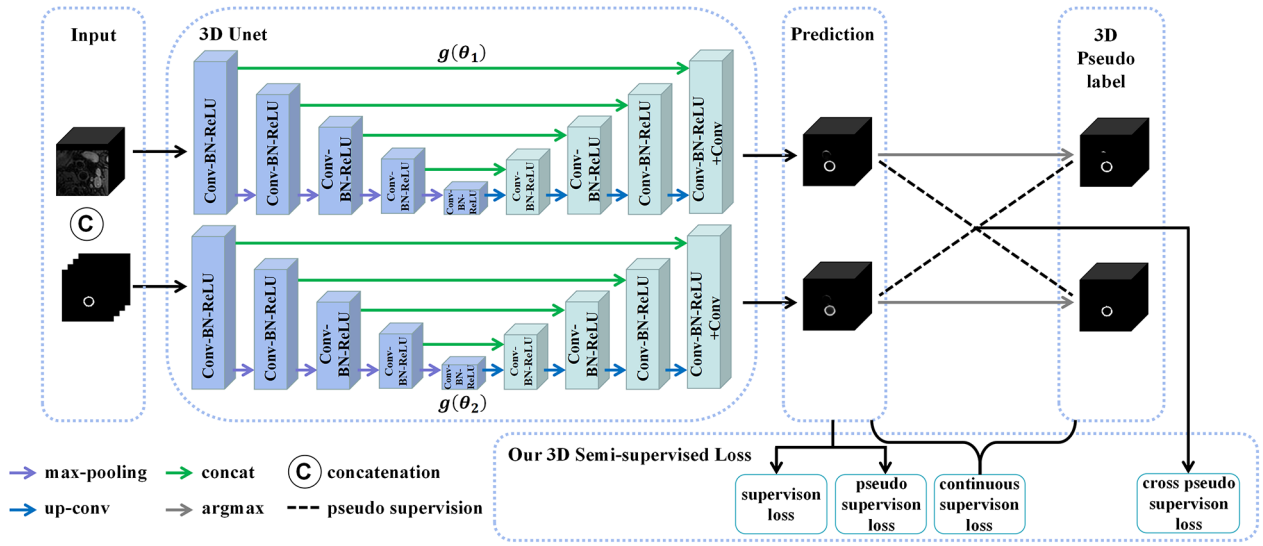


FIGURE 5
The structure of the 3D CPS network, where CAT is short for concatenation. The original image patches and the pseudo labels output by the 2D CPS network are concatenated to form the input of the 3D CPS net. It consists of two 3D U-Net networks with varying parameters. The loss function of the 3D CPS network consists of four parts: the supervision loss, pseudo supervision loss, continuous supervision loss, and cross-pseudo supervision loss.

data, we adopted the method of pseudo-label mutual supervised learning, where the pseudo labels $S_{2D}^1$ were used to supervise $P_{2D}^2$, and the pseudo labels $S_{2D}^2$ were used to supervise $P_{2D}^1$. The goal was to enforce a high degree of consistency between the predictions of the two perturbed networks. Subsequently, the continuous 2D pseudo labels $S_{2D} \in \mathbb{R}^{h \times w}$ obtained after sufficient training were concatenated as additional inputs for the 3D CPS network. These pseudo labels also provided auxiliary supervision for the outputs of the 3D CPS network.

### 2.2.3. 3D CPS network

Although the 2D CPS model estimated the 2D pseudo labels, it lacked the modeling of three-dimensional continuity. On the other hand, employing 3D methods that take 3D images as inputs often incurs high computational costs. To mitigate such issues, the use of smaller 3D patches can be considered to balance the performance and computational efficiency. Thus, we proposed a novel method for acquiring 3D patches by utilizing the vascular center-of-gravity positioning model and 2D pseudo labels to extract the relevant local vascular regions of interest. In addition, an overlapping sliding window approach was employed to preserve more contextual information within the extracted patches. Firstly, we split the 3D volumes into a series of small-size 3D patches $J_{3D} \in \mathbb{R}^{d \times H \times W}$, where $d$ represented the depth of the desired 3D patch. For each $J_{3D}$, there was a corresponding 3D lumen binary segmentation map $Q_{3D} \in \mathbb{R}^{d \times H \times W}$, which was obtained by gathering the 2D lumen coarse segmentation $Q_{2D}$. The vascular center of gravity $G_{3D}$ of the 3D image $J_{3D}$ was calculated using $Q_{3D}$ according to the following equation:

$$G_{3D} = g_{3D}(Q_{3D}), \tag{3}$$

where $g_{3D}$ denoted the 3D first-order moment function, which was a direct extension of the equation (1). Specifically, only the $x$-axis and $y$-axis coordinates of $G_{3D}$ needed to be determined since the patch depth had already been fixed to $d$. Therefore, we used the position information of the vascular center of gravity $G_{3D}$ to crop the input patch $J_{3D}$ into $X_{3D} \in \mathbb{R}^{d \times h \times w}$, where the sizes were $d \times h \times w$. Finally, the obtained 3D patches and pseudo labels were used as inputs for the newly proposed 3D CPS network to estimate the segmentation results. The specific architecture of this network is illustrated in Figure 5.

Similar to the 2D CPS model, the 3D CPS network was constructed using two 3D U-Net networks, denoted as $g(\theta_1)$ and $g(\theta_2)$, which had identical structures but different parameters. The input of the 3D CPS network consisted of both the 3D patches and the pseudo labels estimated by the 2D CPS. The output of the 3D CPS network was represented by the confidence map $P_{3D}^m \in \mathbb{R}^{C \times d \times h \times w} (m = 1, 2)$. Therefore, the relationship could be expressed as follows:

$$P_{3D}^m = g(X_{3D}, P_{2D}; \theta_m). \tag{4}$$

Consequently, we obtained the corresponding pseudo labels $S_{3D}^m \in \mathbb{R}^{d \times h \times w}$ $(m = 1, 2)$ through the argmax operation. In contrast to the 2D CPS model, the limited availability of 2D labeled data within the 3D patches, which accounted for less than 10%, posed difficulty for the semi-supervised network CPS to achieve accurate segmentation. To address this challenge, we additionally used the pseudo labels obtained by the 2D CPS network to supervise the predictions of the 3D CPS network. Simultaneously, the pseudo-label supervised learning enforced the prediction of 3D CPS to be of high consistency with the 2D

CPS model. In addition, we exploited the spatial continuity of the vessels in order to enhance the plausibility of the predictions made by the 3D CPS network.

### 2.2.4. Loss function

In the following, we will discuss the loss functions used for coarse segmentation and fine segmentation, respectively.

In the coarse segmentation stage, the network output the classification and signed distance function (SDF) simultaneously. We used the Focal Tversky (FT) loss function (46) to calculate the loss of the pixel-wise classification, given as follows:

$$L_{\mathrm{FT}} = (1 - L_T)^{\gamma},$$

with

$$L_T = \frac{|P \cap Y|}{|P \cap Y| + \alpha|P - Y| + \beta|Y - P|},$$

where $L_T$ represents the Tversky Loss, $P$ and $Y$ represent the predicted pixel-level classification results and ground truth, and $\alpha$ and $\beta$ controlled the proportion of false positives and false negatives, respectively. As can be seen, the Focal Tversky loss introduced a focal mechanism based on the Tversky index. Compared to the traditional cross-entropy loss function, it was proven to be better suited for addressing class imbalance issues in image segmentation. In addition, it can enhance penalty on boundary regions and suppress the classification of pixels being misclassified. Therefore, we adopted the Focal Tversky loss to address the challenging vessel segmentation problem in coarse segmentation.

The SDF reflected the position information and boundary information of the segmented lumen, which was defined as follows:

$$\varphi(x) = \begin{cases} -\inf\limits_{y \in \partial V} \|x - y\|_2, & \text{if } x \in V; \\ 0, & \text{if } x \in \partial V; \\ \inf\limits_{y \in \partial V} \|x - y\|_2, & \text{if } x \in \Omega \setminus V; \end{cases}$$

where $V$ represents the vascular area, $y$ was the point on the border of the vascular area, $\varphi: \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}$, the signed distance function was expressed as the infimum of the minimum value to the border of the vascular area for a given point $x$. Thus, the loss function for the coarse segmentation stage consisted of the following two terms:

$$L = L_{\mathrm{FT}} + L_{\mathrm{SDF}}.$$

In order to balance the loss contributions from both tasks, we used the homoscedastic uncertainty for weighting a dual-task loss function as follows:

$$L(\sigma_1, \sigma_2) = \frac{1}{\sigma_1^2} L_{\mathrm{FT}} + \frac{1}{\sigma_2^2} L_{\mathrm{SDF}} + \log \sigma_1 \sigma_2, \tag{5}$$

where parameters $\sigma_1$ and $\sigma_2$ corresponded to the homoscedastic uncertainties of the Focal Tversky loss and the signed distance function loss, regarding the classification task and the regression task, respectively. By minimizing the loss $L$ and the noise variables $\sigma_1$, $\sigma_2$, task-specific losses could be balanced during the training process.

The training of 2D CPS consisted of the supervision loss $L_{2\mathrm{D}}^{s}$ and cross-pseudo supervision loss $L_{2\mathrm{D}}^{\mathrm{cps}}$ such as

$$L_{2\mathrm{D}} = L_{2\mathrm{D}}^{s} + \lambda_0 L_{2\mathrm{D}}^{\mathrm{cps}}, \tag{6}$$

where $\lambda_0$ was the trade-off weight. The supervision loss for the labeled data included the cross-entropy and dice loss as given below:

$$L_{2\mathrm{D}}^{s} = \frac{1}{2} \sum_{n=1}^{2} (l_{\mathrm{ce}}(Y^n, P_{2\mathrm{D}}^n) + l_d(Y^n, P_{2\mathrm{D}}^n)),$$

where $Y^n$ represented the ground truth, $l_{\mathrm{ce}}$ was the cross-entropy loss, and $l_d$ was the dice loss. In addition, the cross-pseudo supervision loss formula for labeled data and unlabeled data was also considered

$$L_{2\mathrm{D}}^{\mathrm{cps}} = l_{\mathrm{ce}}(S_{2\mathrm{D}}^2, P_{2\mathrm{D}}^1) + l_{\mathrm{ce}}(S_{2\mathrm{D}}^1, P_{2\mathrm{D}}^2).$$

In addition, the loss function for 3D CPS included the supervision loss $L_{3\mathrm{D}}^{s}$, the cross-pseudo supervision loss $L_{3\mathrm{D}}^{\mathrm{cps}}$, the pseudo label supervision loss $L_{3\mathrm{D}}^{\mathrm{ps}}$, and the continuous supervision loss $L_{3\mathrm{D}}^{\mathrm{cs}}$, which was defined as follows:

$$L_{3\mathrm{D}} = L_{3\mathrm{D}}^{s} + \lambda_1 L_{3\mathrm{D}}^{\mathrm{cps}} + \lambda_2 L_{3\mathrm{D}}^{\mathrm{ps}} + \lambda_3 L_{3\mathrm{D}}^{\mathrm{cs}}, \tag{7}$$

where $\lambda_1$ and $\lambda_3$ were the trade-off weights, and $\lambda_2$ was the pseudo label weight. Because the labels were in 2D format, the supervision loss construction for the labels for the 3D CPS network was the same as for the 2D CPS network, i.e.,

$$L_{3\mathrm{D}}^{s} = \frac{1}{2|A|} \sum_{i \in A} \sum_{m=1}^{2} (l_{\mathrm{ce}}(Y_i^m, P_{3\mathrm{D}}^m(i)) + l_d(Y_i^m, P_{3\mathrm{D}}^m(i))),$$

where $A$ was the set of labels, $Y_i^m$ represented the ground truth, and $P_{3\mathrm{D}}^m(i)$ represents the $i$th layer of the output $P_{3\mathrm{D}}^m$ of the 3D CPS network. In addition, the cross-pseudo supervision loss was defined as follows

$$L_{3\mathrm{D}}^{\mathrm{cps}} = l_{\mathrm{ce}}(S_{3\mathrm{D}}^2, P_{3\mathrm{D}}^1) + l_{\mathrm{ce}}(S_{3\mathrm{D}}^1, P_{3\mathrm{D}}^2),$$

where $S_{3\mathrm{D}}^m$ represents the pseudo label estimated by the 3D CPS network for $m = 1, 2$. The pseudo-label supervised loss formula

for unlabeled data was described as

$$L_{3D}^{ps} = \frac{1}{2|B|}\sum_{i\in B}\sum_{m=1}^{2}(l_{ce}(S_{2D}, P_{3D}^{m}(i)) + l_{d}(S_{2D}, P_{3D}^{m}(i))),$$

where $B$ was the unlabeled dataset, and $S_{2D}$ was the pseudo labels of the segmentation from the 2D CPS network. Finally, the continuous supervision loss was defined as follows:

$$L_{3D}^{cs} = \frac{1}{2}\sum_{m=1}^{2}\left(\sum_{i=1}^{d-1} l_{ce}(S_{3D}^{m}(i+1), P_{3D}^{m}(i)) + \sum_{i=2}^{d} l_{ce}(S_{3D}^{m}(i-1), P_{3D}^{m}(i))\right).$$

## 2.3. Evaluation metrics

In the testing phase, the performance of the proposed method was evaluated using manually corrected ground truth. The segmentation effectiveness of the vessel wall, lumen, and outer wall was assessed using the following designed quantitative metrics (QS), the Dice Similarity Coefficient (DSC) of the lumen region ($DSC^L$), and the DSC of the wall region ($DSC^W$). QS was calculated based on six additional indicators: the DSC of the vessel wall region, Lumen area difference (Lad), Wall area difference (Wad), Normalized wall index difference (Nwid), Hausdorff distance on lumen normalized by radius (Hdol), and Hausdorff distance on wall normalized by radius (Hdow). The calculation of QS was as follows:

$$\begin{aligned}QS = {} & 0.5 \times DSC + 0.1 \times (f(Lad) + f(Wad)) + 0.2 \times f(Nwid) \\ & + 0.05 \times (f(Hdol) + f(Hdow)),\end{aligned}$$

where $f(x) = \max(0, 1 - x)$. As an ensemble similarity measure, DSC was computed to assess the similarity between the vessel wall segmentation result and the ground truth, which was defined as follows:

$$DSC = \frac{2(X \cap Y)}{X + Y}.$$

where $X$ and $Y$ represent the binary vessel wall segmentation result and ground truth, respectively. Therefore, DSC equaled 1 when the segmentation result was the same as the ground truth. The Lad and Wad calculated the area difference between the lumen and outer wall and the ground truth, respectively, which were defined as follows:

$$Lad = \frac{|XA^L - YA^L|}{YA^L}, \quad Wad = \frac{|XA^W - YA^W|}{YA^W}.$$

where $XA^L$, $XA^W$, $YA^L$, and $YA^W$ represent the area of the lumen segmentation, the area of the outer wall segmentation, and their corresponding ground truth areas, respectively. In addition, the Nwid represented the difference between the normalized outer wall area and the normalized outer wall ground truth area using the following formula:

$$Nwid = \frac{\left|\frac{XA^W - XA^L}{XA^L} - \frac{YA^W - YA^L}{YA^L}\right|}{\frac{YA^W - YA^L}{YA^L}}.$$

The Hdol and Hdow were calculated by the Hausdorff distance between the contours of the lumen and outer wall to the ground truth, respectively, such as

$$Hdol = \frac{\max(h(XO^L, YO^L), h(YO^L, XO^L))}{\sqrt{XA^L/\pi}},$$

and

$$Hdow = \frac{\max(h(XO^W, YO^W), h(YO^W, XO^W))}{\sqrt{XA^W/\pi}},$$

where $h(B, C) = \max_{b\in B}\{\min_{c\in C}\|b - c\|\}$, and $XO^L$, $XO^W$, $YO^L$, and $YO^W$ represent the contour point set of the lumen segmentation result, the contour point set of the outer wall segmentation result, and their corresponding ground truth contour point sets, respectively.

# 3. Experiments and results

Our method was implemented by using PyTorch, and all experiments were performed on a server with one NVIDIA Geforce RTX 3090 Founders Edition GPU. In the coarse segmentation stage, the total training time was 12 h. In the fine segmentation stage, the total training time was 7 h.

## 3.1. Data processing

Manual vessel contour labels were given by a customized vessel wall annotation software (CASCADE), so that some labels in the test set had a certain offset error, as shown in **Figure 6**. To address this issue, we manually corrected the label images to eliminate the offset errors. Specifically, a total of 526 labels with offset errors were manually adjusted to achieve the closest approximation to the ground truth, as shown in **Figure 6(B,D)**.

## 3.2. Implementation details

The training details of our proposed JCPS network are described as follows. In the coarse segmentation stage, a Deeplabv3+ coarse segmentation network was trained, and its input patch size was the original resolution $H \times W$, where $H$

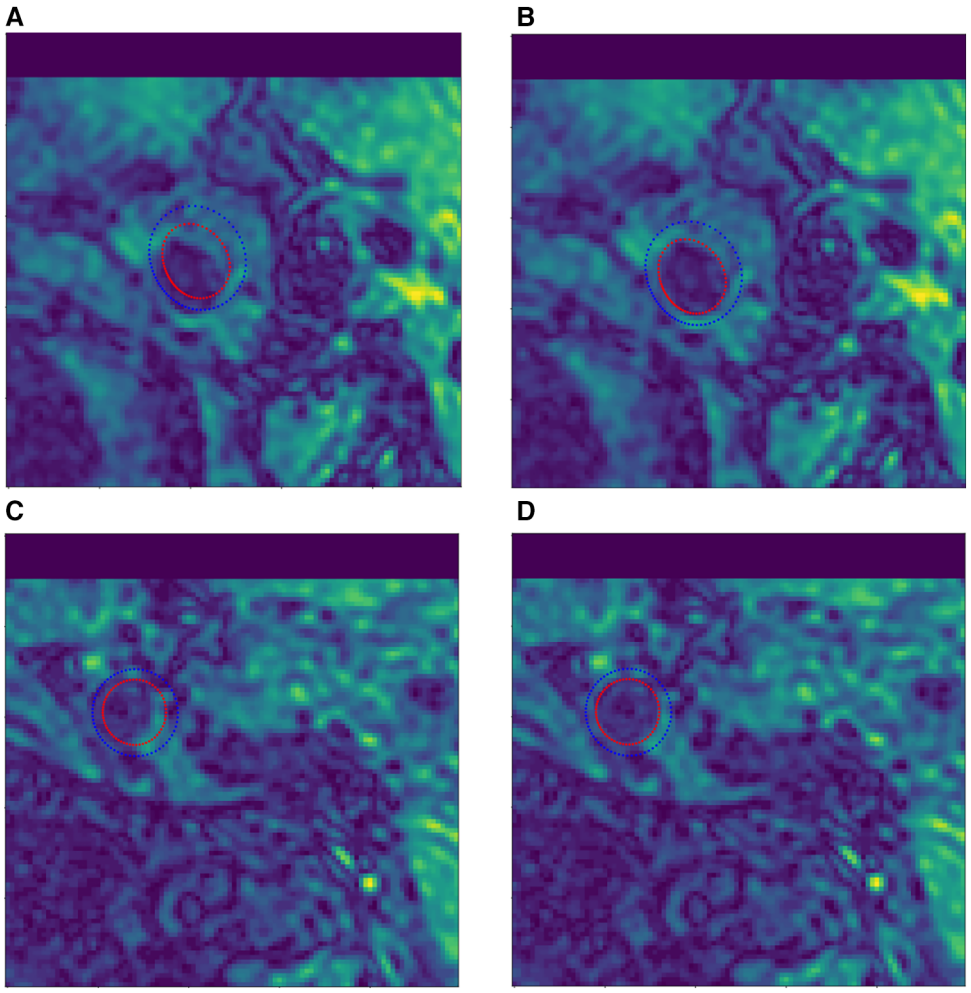


FIGURE 6
Example of manual contour label correction. (A,C) Test set label data with offset error. (B,D) Manually corrected test set label data.

and $W$ were both set to 720. The epoch number and batch size were set to 400 and 12, respectively. The Deeplabv3+ was optimized using an Adam optimizer, with a learning rate of 0.001, multiplied by 0.9 in iterations of 1,000. In the fine segmentation stage, a joint 2D–3D CPS network was trained to finely segment the vessel wall. For the 2D CPS network in JCPS, the input patch size was $h \times w$, where $h$ and $w$ were both set to 96. The iteration number, batch size, and batch size of the labeled data were set to 30,000, 4, and 2, respectively. We employed the Poly learning rate strategy, where the learning rate was set to 0.01 and was changed by the initial learning rate multiplied by $(1 - \text{iter}/\text{max\_iter})^{0.9}$ for each iteration. In addition, we employed mini-batch stochastic gradient descent (SGD) with momentum to train 2D CPS, where the momentum was fixed at 0.9 and weight decay was set to 0.0001. For the 3D CPS network in JCPS, the input patch size was $d \times h \times w$, where $d$, $h$, and $w$ were set to 32, 96, and 96, respectively. The iteration number, batch size, and batch size of the labeled data were set to 30,000, 4, and 2, respectively. The settings of the learning rate strategy and SGD were the same as in 2D CPS. In the loss function of the coarse segmentation stage, we set the weights as $\alpha = 0.7$, $\beta = 0.3$, and $\gamma = 0.7$. In the loss function of the fine segmentation stage, we empirically set the weights as $\lambda_0 = \lambda_1 = \lambda_3 = e^{-5(1-t)^2}$, $t = \text{epoch}/\text{max\_epoch} \in [0, 1]$, which were a weight ramp-up equation (37) that increased with time, and $\lambda_2 = 1$. In particular, the parameter settings of all variants of our method were the same as those described above.

Note that the erroneous segmentation in the coarse segmentation may affect the selection of central points and subsequently impact the fine segmentation stage. The failure in the first stage can be roughly divided into three cases: (1) there are scattered fragments around the vessel wall, causing the center point to deviate from its geometric center; (2) due to the inability of coarse segmentation to accurately distinguish between internal and external carotid arteries at the bifurcation of blood vessels, the central point is located in the external carotid artery region; (3) in areas of carotid artery stenosis, especially extremely narrow areas, the coarse segmentation may not even be able to identify vascular, thus unable to locate the center point. Therefore, we applied the morphological post-processing to the results of the coarse segmentation. We eliminated fragmented regions in the coarse segmentation results by selecting the largest

connected region. In the bifurcation area of the carotid artery, we used the position of the center point before and after the bifurcation to estimate the correct center point relying on the spatial continuity of vessels. Finally, we used the segmentation results of regular regions to interpolate the narrow regions.

## 3.3. Performance on the test dataset

In the first place, we used coarse segmentation to estimate the center of gravity and the local patches. As shown in Figure 7, our modified Deeplabv3+ model accurately identified the center of gravity in all slices. According to statistical analysis, we found that the diameter of carotid artery vessels is smaller than 64 pixels. Therefore, we set the patch size to $96 \times 96$ to capture sufficient information on the carotid vessels. Furthermore, we also validated that the segmentation results using $96 \times 96$ sized patches were optimal in numerical experiments.

In the fine segmentation stage, we evaluated the segmentation performance of our proposed method using the public 3D carotid black-blood MRI dataset. The segmentation accuracy of the top four methods on the leaderboard, as well as our method, is presented in Table 1. The results clearly demonstrated that our method surpassed the top-ranked team by more than 1% on quantitative scoring metrics and 3% on the Dice coefficient, while also surpassing other teams by a significant margin. In addition, the Lad, Wad, and Nwid indicators indicated a substantial reduction in errors within the segmented area using our JCPS model. Although the Hdol and Hdow indicators were slightly higher than those of the top-ranked team, the overall performance of our JCPS model was superior to all others.

TABLE 1 Performance of carotid vessel wall segmentation in comparison to the other top four teams.

| | DSC | Lad | Wad | Nwid | Hdol | Hdow | QS |
|---|---|---|---|---|---|---|---|
| Team 1 | 0.775 | 0.086 | 0.072 | 0.080 | **0.246** | **0.215** | 0.837 |
| Team 2 | 0.761 | 0.064 | 0.075 | 0.079 | 0.554 | 0.515 | 0.728 |
| Team 3 | 0.736 | 0.089 | 0.136 | 0.139 | 0.366 | 0.358 | 0.727 |
| Team 4 | 0.697 | 0.170 | 0.144 | 0.130 | 0.407 | 0.361 | 0.694 |
| Ours | **0.806** | **0.063** | **0.068** | **0.054** | 0.305 | 0.297 | **0.850** |

Teams 1–4 are the top four methods in the Carotid Artery Vessel Wall Segmentation Challenge. Evaluation indicators include DSC of the vessel wall region, Lad, Wad, Nwid, Hdol, Hdow, and QS.
The bold values represent the optimal results achieved in the respective columns for the indicators.

The effectiveness of each component in our method is demonstrated in Table 2 and Figure 8. First, we examined the effectiveness of the semi-supervised method CPS by comparing U-Net and 2D–CPS during the fine segmentation stage. The results presented in Table 2 and Figure 8 indicate a significant improvement in segmentation accuracy with 2D-CPS compared

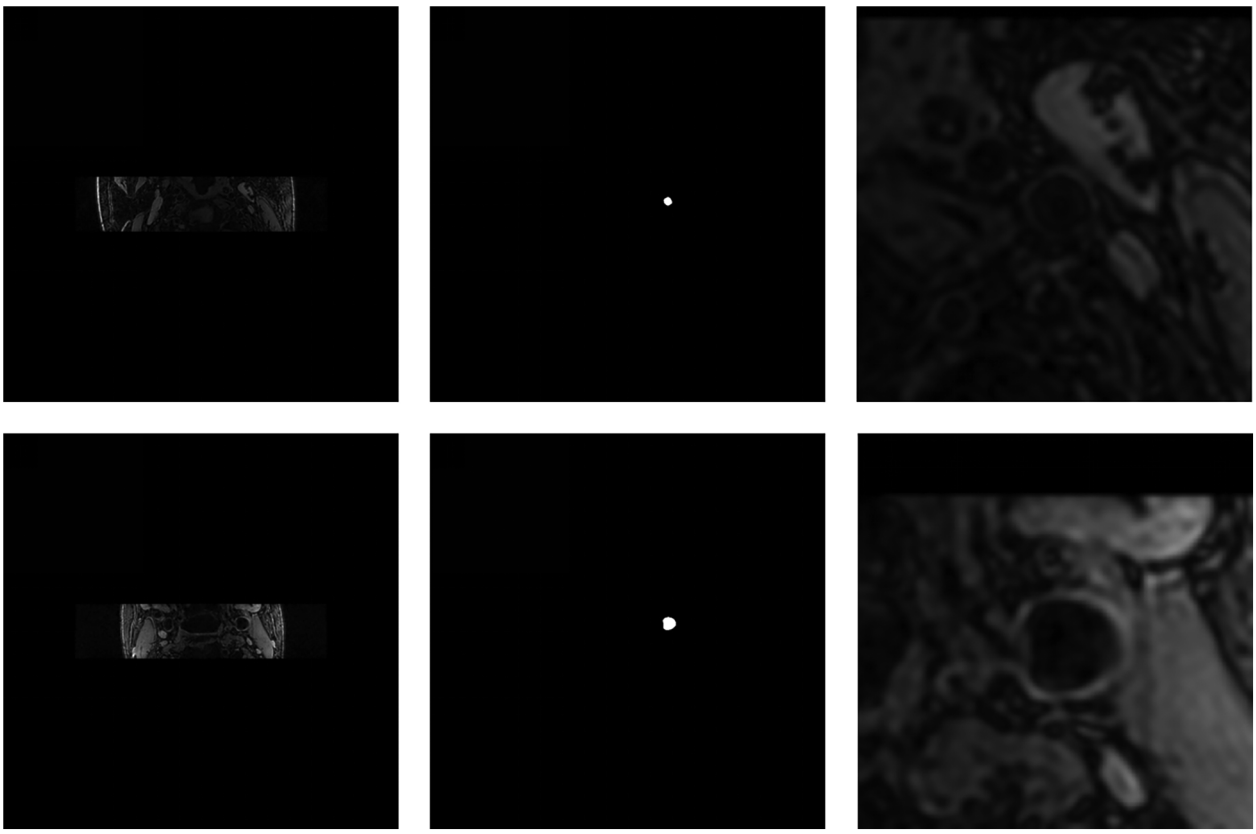

FIGURE 7
Visualization of the coarse segmentation. The first column are the original-resolution MRI images, the middle column are the centers of gravity estimated by coarse segmentation, and the last column are local patches obtained after passing through the vascular center-of-gravity positioning model.

TABLE 2 Performance comparison of the carotid vessel wall segmentation between JCPS and its variants.

| Method | DSC | $DSC^L$ | $DSC^W$ | Lad | Wad | Nwid | Hdol | Hdow | QS |
|---|---|---|---|---|---|---|---|---|---|
| U-Net | 0.766 | 0.908 | 0.911 | 0.108 | 0.112 | 0.104 | 0.491 | 0.478 | 0.791 |
| 2D-CPS | 0.778 | 0.924 | 0.927 | 0.090 | 0.108 | 0.094 | 0.416 | 0.397 | 0.810 |
| 3D-CPS | 0.784 | 0.938 | 0.933 | 0.075 | 0.097 | 0.092 | 0.364 | 0.350 | 0.821 |
| 3D-CPS-w/-CSL | 0.788 | 0.938 | 0.934 | 0.074 | 0.083 | 0.084 | 0.346 | 0.331 | 0.828 |
| JCPS-w/o-CSL | 0.799 | 0.937 | 0.935 | 0.072 | 0.080 | 0.073 | 0.315 | 0.303 | 0.839 |
| JCPS | **0.806** | **0.939** | **0.939** | **0.063** | **0.068** | **0.054** | **0.305** | **0.297** | **0.850** |

U-Net: consists of a vascular center-of-gravity positioning model and a 2D U-Net model. 2D-CPS: consists of a vascular center-of-gravity positioning model and a 2D CPS model. 3D-CPS: consists of a vascular center-of-gravity positioning model and a 3D CPS model. 3D-CPS-w/-CSL: consists of a vascular center-of-gravity positioning model and a 3D CPS model with continuous supervision loss. JCPS-w/o-CSL: consists of a vascular center-of-gravity positioning model and a joint 2D–3D CPS model without continuous supervision loss.
The bold values represent the optimal results achieved in the respective columns for the indicators.

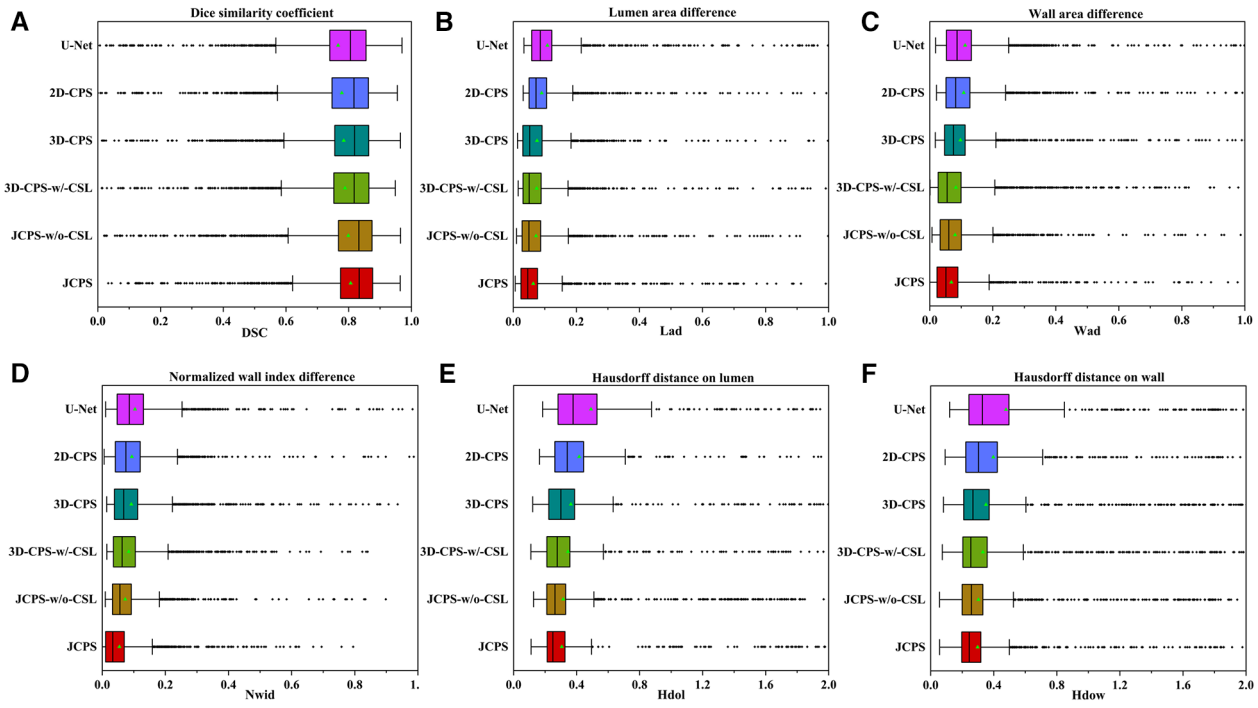


FIGURE 8
Performance metrics (DSC, Lad, Wad, Nwid, Hdol, Hdow) for JCPS and its variants, where ♦ indicates outliers. Note that the proportion of outliers in each evaluation index of our method is basically less than 3%.

to U-Net, as evidenced by higher scores across all indicators. This suggests that the CPS network is better suited for datasets with limited labeled data and exhibits superior generalization performance. In addition, the visualization results depicted in Figure 9 demonstrate a substantial enhancement in our segmentation accuracy for images containing lesions and those near the carotid bifurcation, surpassing the performance of plain U-Net models. This further confirmed the effectiveness of CPS in improving segmentation accuracy.

Through the comparison between 3D-CPS and 2D-CPS, it can be concluded that the 3D CPS network yields superior results by leveraging the information across slices. As shown in Table 2, Figures 8 and 9, the 3D-CPS outperforms the 2D-CPS in the fine segmentation stage, which produced more complete contours for the challenging images, showing the improvement brought by the 3D segmentation approaches.

We then investigated the performance of the joint 2D–3D CPS model. By using the same loss function as the 3D-CPS model, the JCPS-w/o-CSL demonstrated a significant enhancement in segmentation accuracy and yielded superior results for challenging images (refer to Table 2, Figures 8 and 9). It verified the effectiveness of using 2D CPS to generate high-quality pseudo labels that aid the 3D CPS networks in achieving accurate segmentation. Furthermore, it showcases that the joint 2D–3D semi-supervised network is well-suited for processing 3D carotid image datasets with limited 2D labels available.

Finally, we introduced the continuous supervision loss into the joint 2D–3D CPS network to ensure the continuity between adjacent slices. Through a comparison between 3D-CPS, 3D-CPS-w/-CSL, JCPS-w/o-CSL, and JCPS, it was observed that 3D-CPS-w/-CSL exhibited slightly better performance across all metrics (refer to Table 2 and Figure 8). In addition, Figure 9

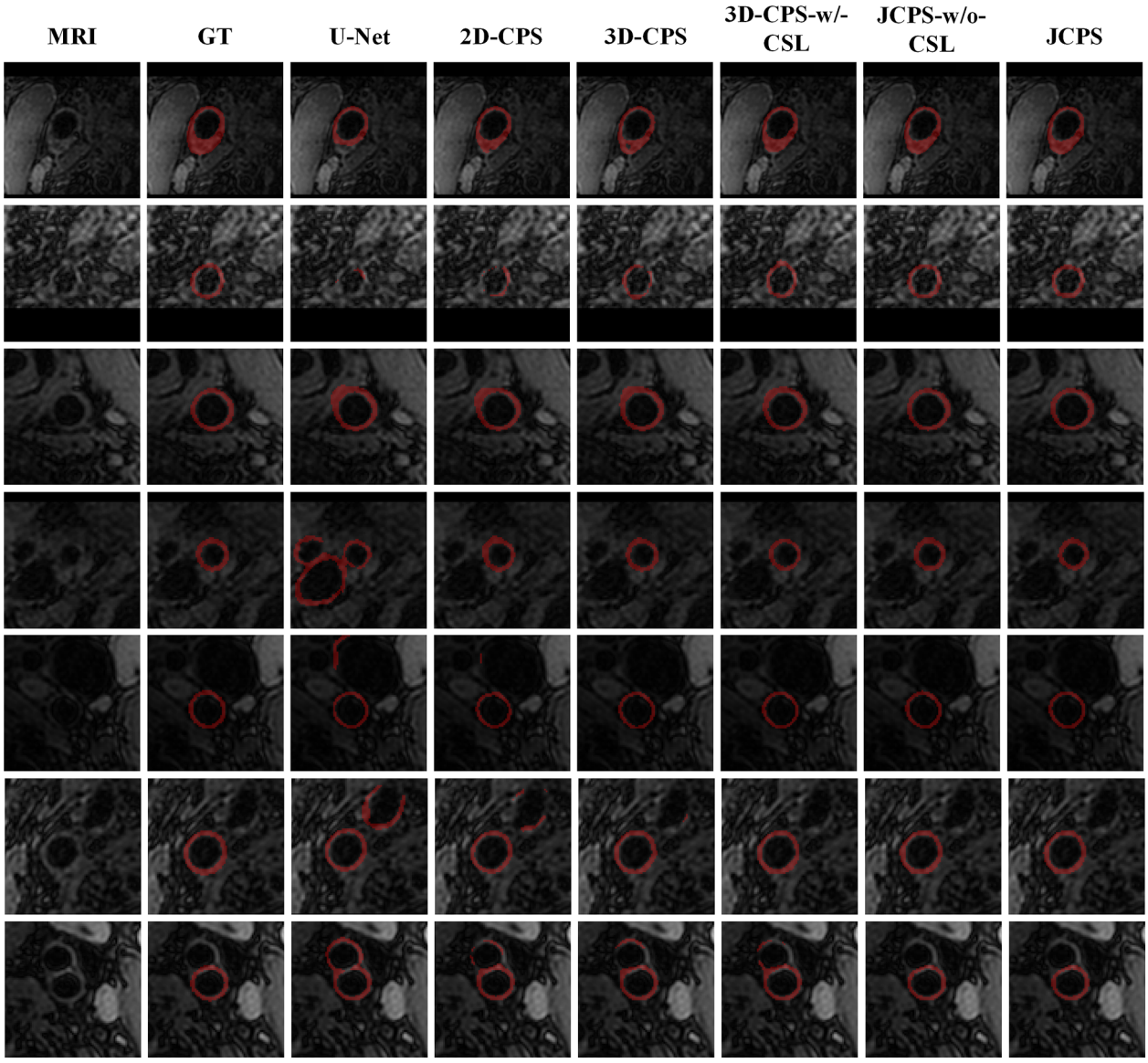


FIGURE 9
Visual comparison between JCPS and its variants, where the second column GT represents the ground truth, referring to the high-quality annotations. These red annotations represent the segmentation of the vessel wall. The case of blood vessels with the plaque are shown on the first row, blood vessels with fuzzy boundary issues are shown in the second through fifth rows, and images of the carotid artery bifurcation are shown in the sixth and seventh rows.

illustrated that 3D-CPS-w/-CSL achieved superior segmentation results compared to 3D-CPS, but both were slightly inferior to JCPS-w/o-CSL. Furthermore, the visualization of 3D carotid vessel wall segmentation in **Figure 10** demonstrated that JCPS outperformed JCPS-w/o-CSL in certain details, indicating the beneficial effect of CSL in obtaining a more continuous carotid vessel wall segmentation.

Based on the visualization results depicted in **Figure 9**, it was observed that all methods were able to accurately identify the vessel region of interest by utilizing the vascular center of gravity obtained during the initial coarse segmentation stage. This indicates the feasibility of the vascular center-of-gravity positioning model. The first two rows of **Figure 9** demonstrate that the segmentation methods encountered challenges with under-segmentation when dealing with vessel images featuring blurred boundaries and plaques. However, our method successfully mitigated these issues by leveraging the semi-supervised learning approach and ensuring continuity between adjacent layers. Consequently, our method achieved stable and precise segmentation outcomes. Furthermore, in the third and fourth rows, it was also noted that images of blood vessels with indistinct boundaries could lead to over-segmentation. Nevertheless, our approach effectively addressed such cases. In the last three examples, it was evident that accurately segmenting the target artery near the carotid bifurcation posed difficulties for other methods due to limited labeled data. Remarkably, our method overcame this problem in most instances.

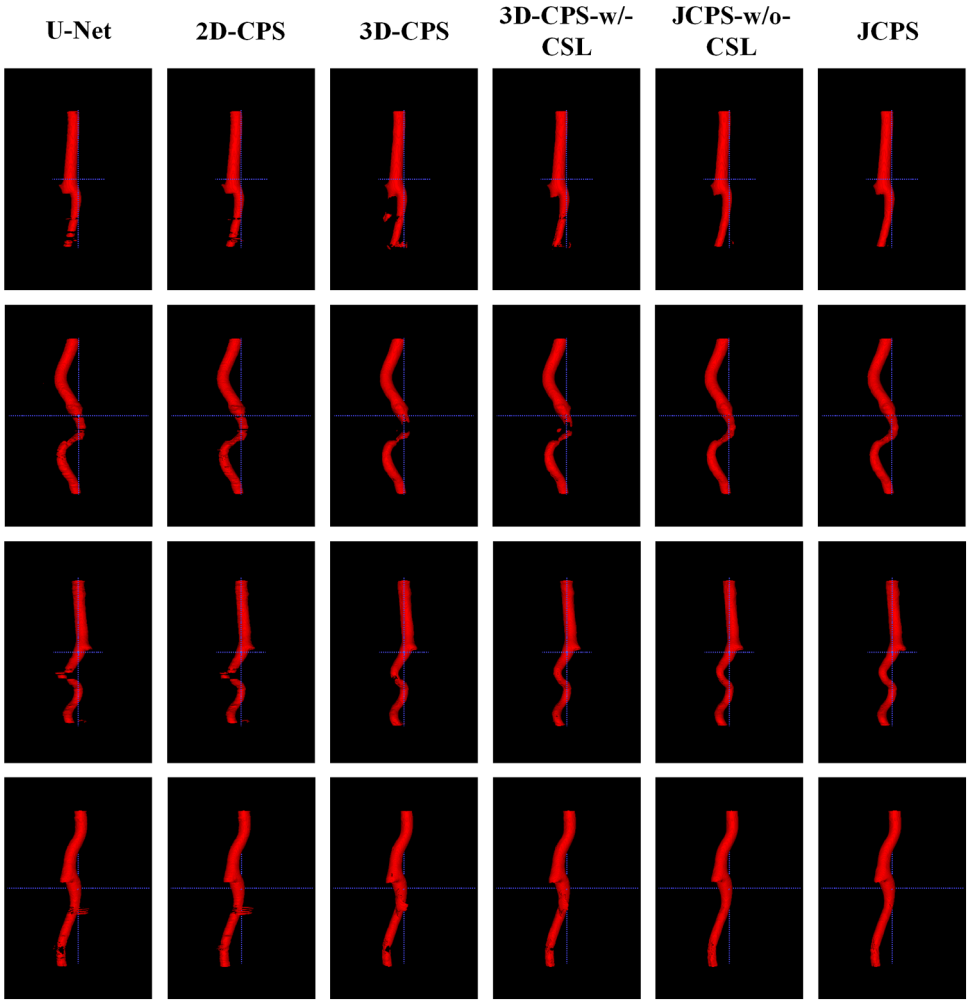


FIGURE 10
3D visualization comparison between the JCPS and its variants, where red annotations represent the segmentation of the vessel wall.

The 3D visualization results of our method and its variants are shown in Figure 10. Compared to methods using 3D networks, both U-Net and 2D-CPS produced discontinuous and incomplete blood vessels. By looking at the middle two columns, we saw that 3D-CPS provided a more complete vascular structure but might still fail in some challenging regions for such a problem with the dataset of incomplete labels. Also, it can be clearly observed that our JCPS could estimate complete and reasonable 3D segmentation results with fewer areas of poor segmentation quality.

## 3.4. Graphical user interface

In practical applications, end-users exhibit a preference toward software solutions that are user-friendly and incorporate a GUI. However, to the best of our knowledge, a comprehensive human–computer interaction (HCI) system that is exclusively dedicated to MRI black-blood carotid image segmentation and offers an effective HCI verification environment for current deep learning algorithms has yet to be established, thereby significantly limiting the clinical application of these algorithms. To address this limitation, we have developed a complete automatic vessel segmentation system founded on a deep learning model. Our system encompasses essential functions including data reading, model import, vessel segmentation, result display, and segmentation accuracy evaluation, seamlessly integrated in a pipeline fashion. The system was implemented using the Python programming language, and the vessel segmentation model based on deep learning algorithms could be executed on a single machine. Figure 11 demonstrates the GUI interface used for both coarse and fine segmentation of the lumen and outer wall. It allows for the visualization of segmentation results and evaluation indicators, thereby illustrating the accuracy of the segmentation process. We tested the CPU time to process one black-blood MRI data of size $230 \times 720 \times 720$ on an AMD Ryzen 7 5700U processor. The CPU time for coarse segmentation and fine segmentation is 80.73 and 139.42 s, respectively, which can satisfy the clinical needs.

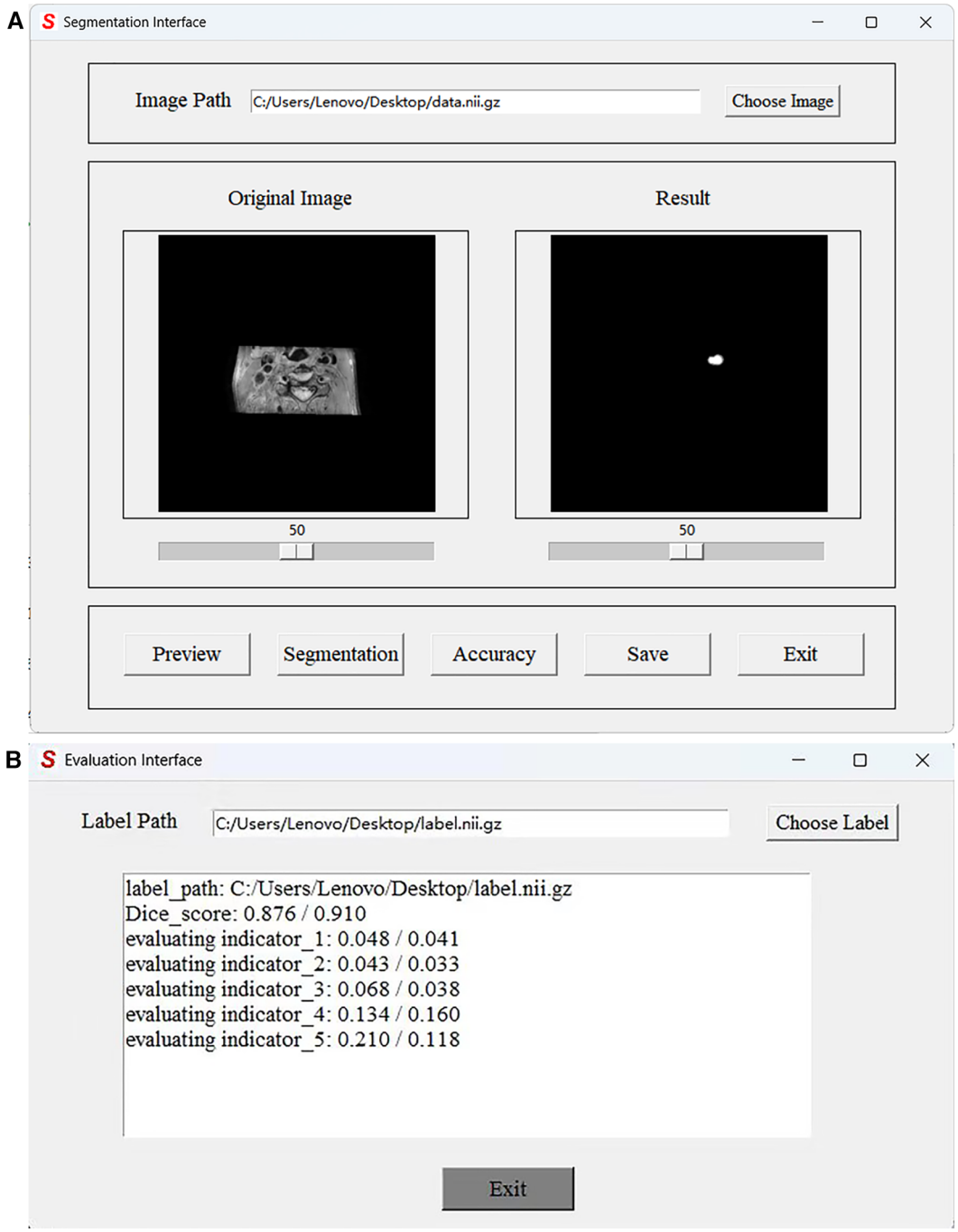


FIGURE 11
Graphical User Interface of the JCPS method. (A) The vessel segmentation interface can display the original image, run the JCPS method, and display and save the segmentation results. (B) The evaluation interface can provide six indicators to illustrate the segmentation accuracy of rough segmentation and fine segmentation.

## 4. Concluding remarks

In this study, we developed a two-stage segmentation framework for carotid vessel wall segmentation. In the coarse segmentation stage, we achieved automatic detection of the vascular center of gravity using a vascular center-of-gravity positioning model. The original images were then clipped into local patches containing vessels based on centers of gravity and used as inputs for fine segmentation modeling. Notably, our proposed approach enabled accurate localization of vascular center-of-gravity without any manual intervention. In the fine segmentation stage, we employed the joint 2D–3D CPS network to estimate the vessel wall. To ensure accurate segmentation of vascular structures, we introduced a novel hybrid loss function. In comparison to the existing approaches, our method did not require a large amount of labeled data and human interaction, and

it exhibited improved segmentation performance across a range of evaluation indicators. Therefore, with reduced costs, the proposed JCPS network could facilitate clinicians in reading vessel wall outlines and diagnosing atherosclerosis. Moreover, a user-friendly and effective graphical user interface has been created to simplify the implementation of our carotid vessel wall segmentation method.

Our JCPS can handle the task of segmenting the carotid artery vessel wall with low image qualities. Indeed, our fine segmentation network has quite good robustness to the results of coarse segmentation, which can provide reasonable segmentation results even for coarse segmentation with defects. However, its performance may deteriorate when dealing with other vessel segmentation problems. In the future, we plan to explore the domain adaptive coarse segmentation model to achieve constant performance on different vessel segmentation tasks. On the other hand, the two-stage approach we used has high complexity and the segmentation results also lack interpretability. Thus, we would like to consider incorporating more effective domain knowledge to develop reliable vessel stenosis prediction methods.

Indeed, our JCPS method is not restricted to carotid black-blood MRI images but also can be used for other blood vessel segmentation and 3D vessel reconstruction tasks. In future works, we would like to investigate automatic segmentation methods depending on even fewer manual annotations for facilitating medical diagnosis. An avenue we plan to pursue involves developing efficient methods for vessel segmentation based on few-shot learning (47) and zero-shot learning (48). In addition, we also intend to evaluate carotid stenosis on the basis of a vascular model combined with hemodynamic simulation.

## 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.

## Author contributions

JX, YC, and YD conceived this study. JX, YZ, YG, and YD developed the carotid vessel wall segmentation methods. YZ, YG, and LY completed the data analysis. YG and YL developed the GUI for the JCPS implementation. YZ, YG, and YL drafted the manuscript. All authors contributed to the article and approved the submitted version.

## Funding

This work was supported by the National Natural Science Foundation of China (NSFC 12071345), the Major Science and Technology Project of Tianjin (18ZXRHSY00160), and Zhejiang Gongshang University 'Digital Interdisciplinary Construction Management Project' (Project Number SZJ2022C007).

## 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. Cheriet M, Said J, Suen C. A recursive thresholding technique for image segmentation. *IEEE Trans Image Process.* (1998) 7:918–21. doi: 10.1109/83.679444

2. Fan H, Xie F, Li Y, Jiang Z, Liu J. Automatic segmentation of dermoscopy images using saliency combined with Otsu threshold. *Comput Biol Med.* (2017) 85:75–85. doi: 10.1016/j.compbiomed.2017.03.025

3. Adams R, Bischof L. Seeded region growing. *IEEE Trans Pattern Anal Mach Intell.* (1994) 16:641–7. doi: 10.1109/34.295913

4. Dehmeshki J, Amin H, Valdivieso M, Ye X. Segmentation of pulmonary nodules in thoracic CT scans: a region growing approach. *IEEE Trans Image Process.* (2008) 27:467–80. doi: 10.1109/TMI.2007.907555

5. Zeng Y-z, Liao S-h, Tang P, Zhao Y-q, Liao M, Chen Y, et al. Automatic liver vessel segmentation using 3D region growing, hybrid active contour model. *Comput Biol Med.* (2018) 97:63–73. doi: 10.1016/j.compbiomed.2018.04.014

6. Chan TF, Vese LA. Active contours without edges. *IEEE Trans Image Process.* (2001) 10:266–77. doi: 10.1109/83.902291

7. Al-Diri B, Hunter A, Steel D. An active contour model for segmenting, measuring retinal vessels. *IEEE Trans Med Imaging.* (2009) 28:1488–97. doi: 10.1109/TMI.2009.2017941

8. Li C, Xu C, Gui C, Fox MD. Distance regularized level set evolution and its application to image segmentation. *IEEE Trans Image Process.* (2010) 19:3243–54. doi: 10.1109/TIP.2010.2069690

9. Li C, Huang R, Ding Z, Gatenby JC, Metaxas DN, Gore JC. A level set method for image segmentation in the presence of intensity inhomogeneities with application to MRI. *IEEE Trans Image Process.* (2011) 20:2007–16. doi: 10.1109/tip.2011.2146190

10. Li BN, Chui CK, Chang S, Ong SH. Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation. *Comput Biol Med.* (2011) 41:1–10. doi: 10.1016/j.compbiomed.2010.10.007

11. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Trans Pattern Anal Mach Intell.* (2015) 39:640–51. doi: 10.1109/CVPR.2015.7298965

12. Zhou X, Ito T, Takayama R, Wang S, Hara T, Fujita H. Three-dimensional CT image segmentation by combining 2D fully convolutional network with 3D majority voting. *Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016*; 2016 Oct 21; Athens, Greece. Proceedings 1. Springer International Publishing (2016). p. 111–20. Available from: https://doi.org/10.1007/978-3-319-46976-8_12. /SEP

13. Zhang J, Saha A, Zhu Z, Mazurowski MA. Hierarchical convolutional neural networks for segmentation of breast tumors in MRI with application to radiogenomics. *IEEE Trans Med Imaging.* (2018) 38:435–47. doi: 10.1109/tmi.2018.2865671

14. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted*

*Intervention -MICCAI 2015: 18th International Conference*; 2015 Oct 5–9; Munich, Germany, Proceedings, Part III 18. Springer International Publishing (2015). p. 234–41. Available from: https://doi.org/10.1007/978-3-319-24574-4_28

15. Seo H, Huang C, Bassenne M, Xiao R, Xing L. Modified U-Net (mU-Net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in CT images. *IEEE Trans Med Imaging*. (2019) 39:1316–25. doi: 10.1109/tmi.2019.2948320

16. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. (2021) 18:203–11. doi: 10.1038/s41592-020-01008-z

17. Chen H, Zhao N, Tan T, Kang Y, Verdonschot N, Sprengers A. Knee bone and cartilage segmentation based on a 3D deep neural network using adversarial loss for prior shape constraint. *Front Med*. (2022) 9:792900. doi: 10.3389/fmed.2022.792900

18. Wang B, Yang J, Ai J, Luo N, An L, Feng H, et al. Accurate tumor segmentation via octave convolution neural network. *Front Med*. (2021) 8:653913. doi: 10.3389/fmed.2021.653913

19. Yu Q, Xie L, Wang Y, Zhou Y, Fishman EK, Yuille AL. Recurrent saliency transformation network: incorporating multi-stage visual cues for small organ segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018). p. 8280–9. Available from: https://doi.org/10.1109/cvpr.2018.00864

20. Hu P, Li X, Tian Y, Tang T, Zhou T, Bai X, et al. Automatic pancreas segmentation in CT images with distance-based saliency-aware DenseASPP network. *IEEE J Biomed Health Inform*. (2020) 25:1601–11. doi: 10.1109/jbhi.2020.3023462

21. Neto LC, Ramalho GL, Neto JFR, Veras RM, Medeiros FN. An unsupervised coarse-to-fine algorithm for blood vessel segmentation in fundus images. *Expert Syst Appl*. (2017) 78:182–92. doi: 10.1016/j.eswa.2017.02.015

22. Ma Y, Hao H, Xie J, Fu H, Zhang J, Yang J, et al., Rose: a retinal OCT-angiography vessel segmentation dataset, new model. *IEEE Trans Med Imaging*. (2020) 40:928–39. doi: 10.1109/tmi.2020.3042802

23. Thuy LNL, Trinh TD, Anh LH, Kim JY, Hieu HT, Bao PT. Coronary vessel segmentation by coarse-to-fine strategy using U-Nets. *Biomed Res Int*. (2021) 2021. doi: 10.1155/2021/5548517

24. Ye Y, Pan C, Wu Y, Wang S, Xia Y. mFI-Net: multiscale feature interaction network for retinal vessel segmentation. *IEEE J Biomed Health Inform*. (2022) 26 (9):4551–62. doi: 10.1109/jbhi.2022.3182471

25. Samber DD, Ramachandran S, Sahota A, Naidu S, Pruzan A, Fayad ZA, et al., Segmentation of carotid arterial walls using neural networks. *World J Radiol*. (2020) 12:1. doi: 10.4329/wjr.v12.i1.1

26. Oliveira A, Pereira S, Silva CA. Retinal vessel segmentation based on fully convolutional neural networks. *Expert Syst Appl*. (2018) 112:229–42. doi: 10.1016/j.eswa.2018.06.034

27. Ni J, Wu J, Wang H, Tong J, Chen Z, Wong KK, et al., Global channel attention networks for intracranial vessel segmentation. *Comput Biol Med*. (2020) 118:103639. doi: 10.1016/j.compbiomed.2020.103639

28. Liu Y, Shen J, Yang L, Bian G, Yu H. ResDO-UNet: a deep residual network for accurate retinal vessel segmentation from fundus images. *Biomed Signal Process Control*. (2023) 79:104087. doi: 10.1016/j.bspc.2022.104087

29. Imran S, Saleem M, Hameed M, Hussain A, Naqvi R, Lee S. Feature preserving mesh network for semantic segmentation of retinal vasculature to support ophthalmic disease analysis. *Front Med*. (2022) 9. doi: 10.3389/fmed.2022.1040562

30. Tan T, Kuang X, Xu X, Fang L, Kozegar E, Sun Y, et al., Improved fully convolutional neuron networks on small retinal vessel segmentation using local phase as attention. *Front Med*. (2023) 10:314. doi: 10.3389/fmed.2023.1038534

31. Zhou R, Guo F, Azarpazhooh MR, Spence JD, Ukwatta E, Ding M, et al., A voxel-based fully convolution network and continuous max-flow for carotid vessel-wall-volume segmentation from 3D ultrasound images. *IEEE Trans Med Imaging*. (2020) 39:2844–55. doi: 10.1109/tmi.2020.2975231

32. Tetteh G, Efremov V, Forkert ND, Schneider M, Kirschke J, Weber B, et al., DeepVesselNet: vessel segmentation, centerline prediction, and bifurcation detection in 3-D angiographic volumes. *Front Neurosci*. (2020) 14:1285. doi: 10.3389/fnins.2020.592352

33. Xia L, Zhang H, Wu Y, Song R, Ma Y, Mou L, et al., 3D vessel-like structure segmentation in medical images by an edge-reinforced network. *Med Image Anal*. (2022) 82:102581. doi: 10.1016/j.media.2022.102581

34. Alblas D, Brune C, Wolterink JM. Deep-learning-based carotid artery vessel wall segmentation in black-blood MRI using anatomical priors. *Medical Imaging 2022: Image Processing*. Vol. 12032. SPIE (2022). p. 237–44. Available from: https://doi.org/10.1117/12.26111127

35. Lee D-H. Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. *Workshop on Challenges in Representation Learning, ICML*. Vol. 3 (2013). p. 896. Available from: https://doi.org/10.1109/ijcnn48605.2020.9207304

36. Rasmus A, Berglund M, Honkala M, Valpola H, Raiko T. Semi-supervised learning with ladder networks. *Adv Neural Inf Process Syst*. (2015) 28. doi: 10.1186/1477-5956-9-S1-S5

37. Laine S, Aila T. Temporal ensembling for semi-supervised learning [Preprint] (2016). Available from: https://doi.org/10.48550/arXiv.1610.02242

38. Tarvainen A, Valpola H. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. *Adv Neural Inf Process Syst*. (2017) 30. doi: 10.48550/arXiv.1703.01780

39. Ouali Y, Hudelot C, Tami M. Semi-supervised semantic segmentation with cross-consistency training. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020). p. 12674–84. Available from: https://doi.org/10.1109/cvpr42600.2020.01269.

40. Chen X, Yuan Y, Zeng G, Wang J. Semi-supervised semantic segmentation with cross pseudo supervision. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021). p. 2613–22. Available from: https://doi.org/10.1109/cvpr46437.2021.00264.

41. Wu Y, Xu M, Ge Z, Cai J, Zhang L. Semi-supervised left atrium segmentation with mutual consistency training. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference*; 2021 Sep 27–Oct 1; Strasbourg, France. Proceedings, Part II 24. Springer International Publishing (2021). p. 297–306. Available from: https://doi.org/10.1007/978-3-030-87196-3_28.

42. Ke Z, Wang D, Yan Q, Ren J, Lau RW. Dual student: breaking the limits of the teacher in semi-supervised learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2019). p. 6728–36. Available from: https://doi.org/10.1109/iccv.2019.00683.

43. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*. (2018). p. 801–18. Available from: https://doi.org/10.1007/978-3-030-01234-2_49.

44. Liu Y, Duan Y, Zeng T. Learning multi-level structural information for small organ segmentation. *Signal Process*. (2022) 193:108418. doi: 10.1016/j.sigpro.2021.108418

45. Liu Y, Wang Y, Duan Y. Effective 3D boundary learning via a nonlocal deformable network. *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE (2022). p. 1–5. Available from: https://doi.org/10.1109/isbi52829.2022.9761415.

46. Abraham N, Khan NM. A novel focal Tversky loss function with improved attention U-Net for lesion segmentation. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE (2019). p. 683–7. Available from: https://doi.org/10.1109/ISBI.2019.8759329.

47. Feng R, Zheng X, Gao T, Chen J, Wang W, Chen DZ, et al. Interactive few-shot learning: limited supervision, better medical image segmentation. *IEEE Trans Med Imaging*. (2021) 40:2575–88. doi: 10.1109/tmi.2021.3060551

48. Bian C, Yuan C, Ma K, Yu S, Wei D, Zheng Y. Domain adaptation meets zero-shot learning: an annotation-efficient approach to multi-modality medical image segmentation. *IEEE Trans Med Imaging*. (2021) 41:1043–56. doi: 10.1109/tmi.2021.3131245

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 30 November 2023
DOI 10.3389/fmed.2023.1271687



OPEN ACCESS

EDITED BY
Gongning Luo,
Harbin Institute of Technology, China

REVIEWED BY
Suyu Dong,
Northeast Forestry University, China
Shaodong Cao,
The Fourth Hospital of Harbin Medical University, China

*CORRESPONDENCE
Guolin Ma
maguolin1007@qq.com
Chuanchen Zhang
zhangchuanchen666@163.com

†These authors share first authorship

RECEIVED 02 August 2023
ACCEPTED 15 November 2023
PUBLISHED 30 November 2023

CITATION
Zhang D, Luan J, Liu B, Yang A, Lv K, Hu P, Han X, Yu H, Shmuel A, Ma G and Zhang C (2023) Comparison of MRI radiomics-based machine learning survival models in predicting prognosis of glioblastoma multiforme.
*Front. Med.* 10:1271687.
doi: 10.3389/fmed.2023.1271687



# Comparison of MRI radiomics-based machine learning survival models in predicting prognosis of glioblastoma multiforme

Di Zhang[1†], Jixin Luan[2,3†], Bing Liu[2,3], Aocai Yang[2,3], Kuan Lv[4], Pianpian Hu[4], Xiaowei Han[5], Hongwei Yu[3], Amir Shmuel[6,7], Guolin Ma[3*] and Chuanchen Zhang[1*]

[1]Department of Radiology, Liaocheng People's Hospital, Shandong First Medical University & Shandong Academy of Medical Sciences, Liaocheng, Shandong, China, [2]China-Japan Friendship Hospital (Institute of Clinical Medical Sciences), Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China, [3]Department of Radiology, China-Japan Friendship Hospital, Beijing, China, [4]Peking University China-Japan Friendship School of Clinical Medicine, Beijing, China, [5]Department of Radiology, The Affiliated Drum Tower Hospital of Nanjing University Medical School, Nanjing, China, [6]McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, QC, Canada, [7]Department of Neurology and Neurosurgery, McGill University, Montreal, QC, Canada

**Objective:** To compare the performance of radiomics-based machine learning survival models in predicting the prognosis of glioblastoma multiforme (GBM) patients.

**Methods:** 131 GBM patients were included in our study. The traditional Cox proportional-hazards (CoxPH) model and four machine learning models (SurvivalTree, Random survival forest (RSF), DeepSurv, DeepHit) were constructed, and the performance of the five models was evaluated using the C-index.

**Results:** After the screening, 1792 radiomics features were obtained. Seven radiomics features with the strongest relationship with prognosis were obtained following the application of the least absolute shrinkage and selection operator (LASSO) regression. The CoxPH model demonstrated that age (HR = 1.576, $p = 0.037$), Karnofsky performance status (KPS) score (HR = 1.890, $p = 0.006$), radiomics risk score (HR = 3.497, $p = 0.001$), and radiomics risk level (HR = 1.572, $p = 0.043$) were associated with poorer prognosis. The DeepSurv model performed the best among the five models, obtaining C-index of 0.882 and 0.732 for the training and test set, respectively. The performances of the other four models were lower: CoxPH (0.663 training set / 0.635 test set), SurvivalTree (0.702/0.655), RSF (0.735/0.667), DeepHit (0.608/0.560).

**Conclusion:** This study confirmed the superior performance of deep learning algorithms based on radiomics relative to the traditional method in predicting the overall survival of GBM patients; specifically, the DeepSurv model showed the best predictive ability.

KEYWORDS

**glioblastoma multiforme, radiomics, machine learning, survival models, prognosis**

# 1 Introduction

Glioblastoma multiforme (GBM) is the most common and least prognostic primary tumour of the central nervous system, with a 5-year survival rate of 6–22% based on a combination of age at diagnosis and other risk factors (1). Prognostic models that include only the patient's age, ethnicity, whether or not they receive radiotherapy, and risk factors such as the size, location and histopathological composition of the tumour often fail to predict overall survival (OS) accurately (2, 3). Therefore, identifying risk factors for GBM prognosis and developing appropriate predictive models are essential for the individualized and precise treatment of GBM patients.

Radiomics, which transforms digital medical images into mineable high-dimensional features and builds statistical models to analyze the features, has been widely used in tumour diagnosis, prognosis prediction, and treatment selection (4). Studies have shown that GBM radiomics information is closely related to patient prognosis and recurrence (5, 6). Zhang et al. (7) developed and validated a radiomics nomogram model to determine GBM survival probabilities in a non-invasive manner, achieving superior accuracy in both the training and test set. Survival analysis (also known as time-effect analysis) methods have been widely used in medical research, such as clinical efficacy trials and disease prognosis analysis. The Cox proportional-hazards model (Cox-PH) is the most well-known method used to determine the association between clinical predictor variables and the risk of mortality events. The CoxPH model is based on the assumption of a linear combination of event risk and variables; however, it is likely to be too simplistic to fit the actual disease progression.

Machine learning is a branch of artificial intelligence that has a wide range of applications in diagnosing and prognostic assessing GBM (5, 8). Compared to CoxPH models, machine learning can identify clinically significant risks with some marginal variables that can significantly improve the model's performance (9). Deep learning (DL) is a frontier area of machine learning algorithms. Deep learning-based features are mainly extracted through convolutional neural networks (CNN), and feature learning algorithms are derived from the data itself and are more targeted to specific studies (10), and are widely used in imaging diagnosis, disease staging and prognosis, which can effectively improve outcome prediction (11–13). The Deepsurv model is a deep learning technique applied to a non-linear cox proportional risk network (14). Studies have shown that the DeepSurv model can obtain patient risk factors from multiple parameters and has achieved good predictive performance in assessing different patients, such as lung cancer and nasopharyngeal carcinoma (15, 16). Previous deep-learning algorithms that have been applied to assess the prognosis of GBM patients used traditional clinical prognostic risk factors and did not incorporate radiomics features (17). To our knowledge, no study has been conducted on the prognosis of GBM patients using radiomics combined with machine learning. Therefore, this study aimed to construct: (1) a traditional CoxPH model, (2) a tree-based SurvivalTree model, (3) an RSF model based on ensemble learning, (4) a DeepSurv, and (5) a DeepHit model based on deep learning for predicting the overall survival of GBM patients based on GBM radiomics and clinical data. Following the construction of these five models, we compared their performance.

# 2 Materials and methods

## 2.1 Clinical case data

According to the proposed inclusion criteria, (1) clinical information of The Cancer Genome Atlas (TCGA) for GBM was downloaded from the TCGA database[1] and (2) Magnetic Resonance Imaging (MRI) data were obtained from the Cancer Imaging Archive (TCIA),[2] and a total of 262 patients were enrolled. Then, 131 patients were excluded due to (1) the lack of fluid-attenuated inversion recovery (FLAIR) sequences from TCIA ($n = 114$) and (2) MRI sequences acquired with severe motion or artefacts that may have induced bias in the subsequent analysis ($n = 17$). A total of 131 patients with GBM were subsequently retrospectively enrolled in our study. In this retrospective study, the requirement for informed consent was waived, as the relevant patient data in the TCGA were publicly available. We followed the relevant policies of the TCGA and TCIA in the acquisition and use of data. The flow chart for this study is shown in Figure 1.

## 2.2 Image acquisition and segmentation

Using ITK-SNAP[3] software to segment the FLAIR images of patients in 3D, the segmentation process is shown in Figure 2. The FLAIR scan parameters were as follows: thickness = 4 ~ 5.5 mm, TR/TE = 9,000 ~ 12,500/140 ~ 157 ms, slice gap = 4 ~ 6.5 mm, flip angle = 80 ~ 90°. The area of interest covered the entire tumour and edema region, and all feature extraction methods were implemented using the Cancer Imaging Phenomics Toolkit (CaPTk www.cbica.upenn.edu/captk). To confirm the reproducibility of the features, 30 patients were randomly selected, two people performed the Region Of Interest (ROI) segmentation, and the intraclass correlation coefficient (ICC) of the two ROIs was calculated (18). A threshold of ICC > 0.8 was set for considering a good agreement between the two neuro-radiologists. Features that achieved ICC higher than this thereshold were considered as showing reproducibility. The calculated features all contain first-order statistical features and statistical-based texture features, such as grey-level co-occurrence matrices (GLCM), grey-level dependence matrix (GLDM), neighbourhood grey-tone difference matrices (NGTDM), grey-level run-length matrices (GLRLM), and grey-level size zone matrices (GLSZ), grey-level size zone matrices (GLSZM) (19, 20).

## 2.3 Establishing radiomics signature and data cleaning

The least absolute shrinkage and selection operator (LASSO) method was used to select key features from the dataset significantly associated with prognosis. The selected features were linearly combined according to their respective coefficient weights to construct

1 https://tcga-data.nci.nih.gov/
2 https://wiki.cancerimagingarchive.net/
3 https://www.itk-snap.org/

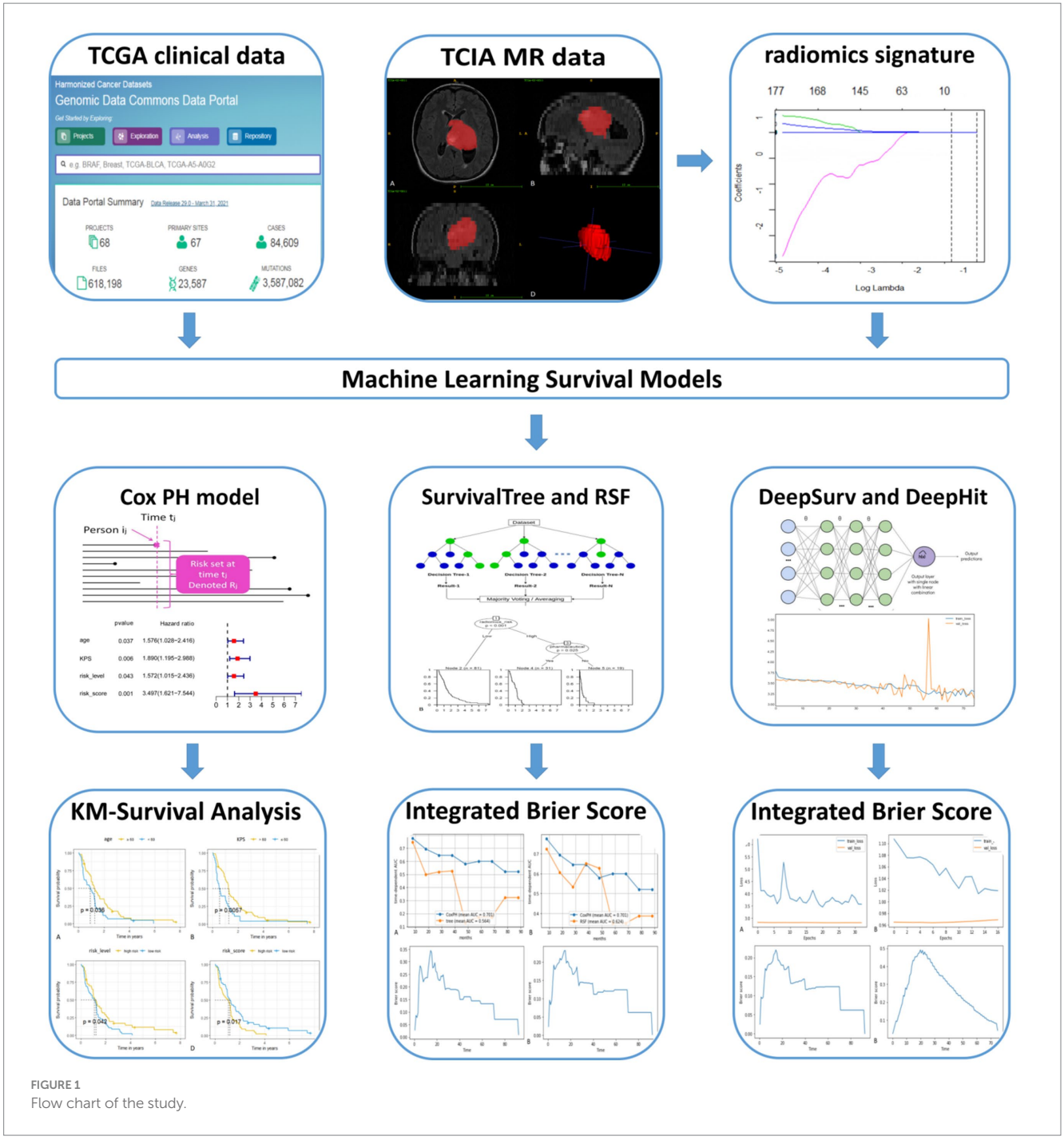


FIGURE 1
Flow chart of the study.

a radiomics signature, calculate the risk score for each patient, and determine the risk level. Subsequently, all collected data were classified as numerical or subtypes according to the input features. The missing data imputation was performed using the k-nearest neighbor (KNN) algorithm (Supplementary Table S1).

## 2.4 Feature engineering

According to Subtype, one-hot coding was performed to convert different categories of risk factors into categorical variables. This resulted in two new features called Subtype_Mesenchymal and Subtype_Proneural.

## 2.5 Construction of the model

### 2.5.1 CoxPH model

For the CoxPH model, proportional risk assumptions were made using the CoxPHFitter function. Filter-based feature selection was performed using Cox regression to select features significantly associated with prognosis in GBM patients. All comparisons were

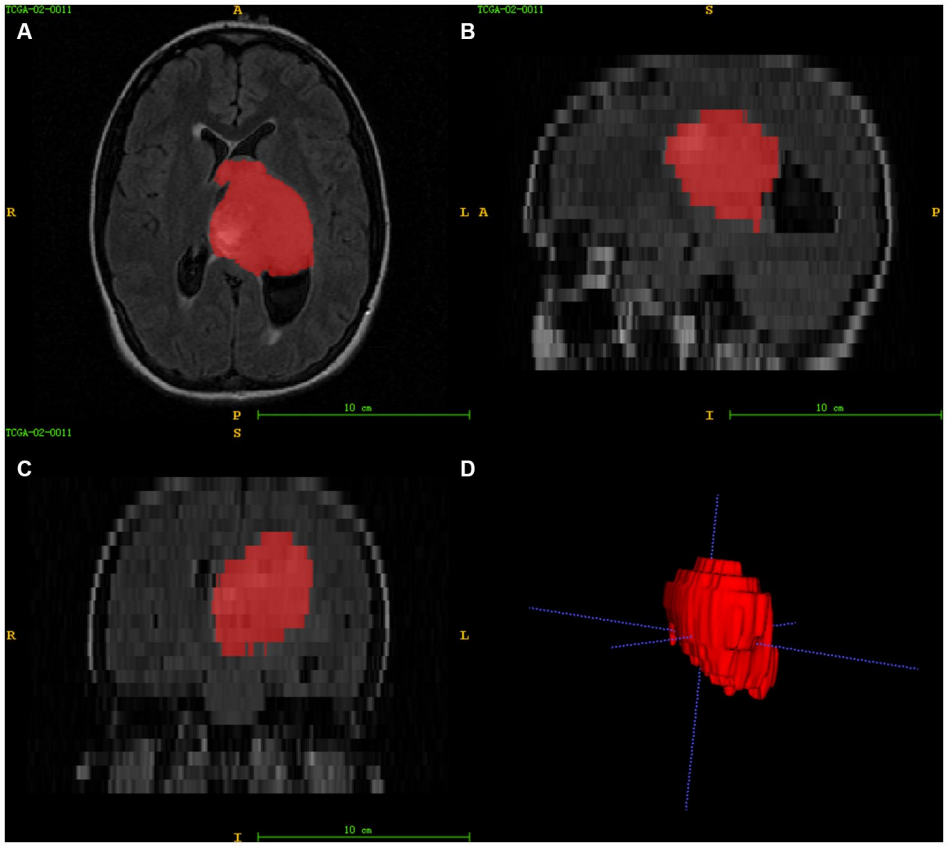


FIGURE 2
Image segmentation **(A–C)** represent the axial, sagittal, and coronal views of the images, respectively, and **(D)** shows the 3D reconstruction results of the ROI.

performed at the 95% confidence level, with $p < 0.05$ indicating statistical significance.

### 2.5.2 SurvivalTree model

SurvivalTree is based on classification and regression trees (CART) (21). The model is based on the tree structure, and the tree building mainly includes tree generation and pruning. Simple dichotomous classification problems can better perform the prognostic grouping of the method.

### 2.5.3 RSF model

Random Survival Forest is a combination of random forest (RF) and survival analysis methods. The model calculates a cumulative risk function for each tree by selecting a subset of variables at each node and splitting the node tree based on survival time and event state, and finally calculates the mean of the integrated cumulative risk function to predict the error (22).

### 2.5.4 DeepSurv model

DeepSurv is a feed-forward deep neural network for CoxPH models to model a nonlinear representation of the risk of clinical events based on input features. The model architecture includes network inputs from patient data, fully connected and hidden layers, and an output layer with linearly activated individual nodes for estimating the logarithmic risk function in the CoxPH model (14). DeepSurv can make predictions without specifying interaction terms, and in addition, the model's hyperparameters can vary depending on the model's performance.

### 2.5.5 DeepHit model

The DeepHit model was initially designed to analyze the competing risks of multiple events (23). In the present study, we considered only one event: patient survival. Therefore, we can use a simplified DeepHit model to analyze our data. We can obtain an estimated probability value with the softmax layer of the model.

## 2.6 Model training and evaluation

After data preprocessing, the data was divided into 70% training data and 30% test data. The hyperparameters of the models were selected via random search. The performance of the models was compared using Harrell's concordance index (C-index) and brier scores. C-index was used to estimate the proportion of random individuals with the same survival time ranking as their accurate survival time, with a C-index value of 1 indicating perfect discrimination and when 0.5 indicating random prediction. The brier score represents the mean squared difference between the observed patient status and the predicted probability of survival, with scores ranged from 0 (worst) to 1 (best). The overall estimate of the brier score for all available times is called the Integrated Brier Score (IBS). In practice, models with IBS below 0.25 are considered valuable. In

addition, for the SurvivalTree and RSF models, we also used the receiver operating characteristic curves (ROC) over time and calculated the area under the curve (AUC) values to evaluate the model performance.

## 2.7 Statistical analysis

Statistical analysis was performed using R 3.6.0 and the model construction was performed using Python 3.7. The R packages used are as follows: glmnet package for LASSO logistic regression, gplots and pheatmap packages for heat map analysis. The Python packages were used are as follows: CoxPH analysis using the lifelines package, SurvivalTree and RSF using the scikit-survival package, feature importance ranking using the permutation_importance function; DeepSurv and DeepHit using the Pytorch-based pycox package. The comparison of patients between training and test set was performed for continuous variables with a t-test or Mann–Whitney test. The chi-square test was performed for subtype variables. All statistics were two-tailed, and *p*-values less than 0.05 were considered statistically significant.

# 3 Results

## 3.1 Clinical characteristics of patients

The clinical characteristics of the patients in the training and test set are shown in Table 1. There were no statistically significant differences in patient age, sex, race, radiation, pharmaceutical, survival status or survival month between the training and test set ($p = 0.071–1.000$).

## 3.2 Radiomics feature extraction and construction of radiomics signature

In this study, 1792 radiomics features were obtained based on T2-FLAIR images from the TICA database, using CaPTk software. The 1792 features were brought into the LASSO cox regression model to screen the optimal radiomics features. We screened the optimal radiomics features in the full dataset using the LASSO Cox regression model with ten-fold cross-validation (24). We obtained seven radiomics features (three signal intensity features and four texture features) that were most closely related to the prognosis. A radiomics signature was constructed based on the linear combination of the screened seven radiomics features and their corresponding Cox regression coefficient products. The radiomics signature we constructed is described by a formula in the Supplementary Material.

## 3.3 Correlation between radiomics signature and clinical information

The correlation between the radiomics signature and clinical information was evaluated using heat map analysis (Supplementary Figure S1). The results show that "GLCM_Contrast_Variance" has a high correlation with survival status, mostly in red color.

## 3.4 CoxPH model

The univariate cox analysis showed that age (HR = 1.576, $p = 0.037$), KPS score (HR = 1.890, $p = 0.006$), radiomics risk score (HR = 3.497, $p = 0.001$), and radiomics risk level (HR = 1.572, $p = 0.043$) were prognostic factors for overall survival in GBM (Table 2), and the univariate analysis forest plot is shown in Figure 3; multivariate cox analysis showed that KPS score (HR = 1.864, $p = 0.008$), radiomics risk score (HR = 3.370, $p = 0.003$) were prognostic factors for overall survival of GBM (Table 2). In the training and test set, the C-index of the CoxPH model was divided into 0.663 and 0.635, with an overall C-index of 0.662, and for predicting 1-year, 3-year, and 5-year survival, the brier score was 0.225, 0.080, and 0.040, respectively, and the IBS was 0.102 (Table 3). The KM survival curves for variables that were significant for the univariate survival analysis are shown in Figure 4.

## 3.5 SurvivalTree and RSF model

GBM survival prediction models based on the SurvivalTree and RSF tree algorithms were built using the training set and validated in the test set. Figure 5 shows the AUC values of the CoxPH model, the SurvivalTree model and the RSF model over time. As can be seen from the graph, the CoxPH model has the highest AUC value of 0.701, and the SurvivalTree model has the lowest AUC of 0.564.

In the training and test set, the C-index of the SurvivalTree model was divided into 0.702 and 0.655, and the overall C-index was 0.564. For predicting 1-year, 3-year, and 5-year survival, the brier scores were 0.225, 0.080, and 0.040, respectively, and the combined brier score was 0.192. In the training and test set, the C-index of the RSF model was divided into 0.735 and 0.667, and the overall C-index was 0.642; for predicting 1-year, 3-year, and 5-year survival, the brier scores were 0.214, 0.143, and 0.124, respectively, and the IBS was 0.152 (Table 3). The IBS plots of the two models are shown in Figure 5.

The ranked importance of SurvivalTree and RSF model features are shown in Figure 6 and Supplementary Table S2; from the table, we can see that KPS, radiation and risk score are more important for the model. For both models, radiation is the most important feature, if radiation is removed from the model, the C-index of both will decrease by 0.145 and 0.101, respectively.

## 3.6 Deep learning model

DeepSurv and DeepHit survival prediction models based on deep learning algorithms were built using the training set and validated in the test set. In the training and test sets, the DeepSurv model had a C-index of 0.882 and 0.732, an overall C-index of 0.691, and a brier score of 0.203, 0.139, and 0.124 for predicting 1-year, 3-year, and 5-year survival, respectively, with a combined brier score of 0.116. In the training and test set, the DeepHit model had a C-index of 0.608 and 0.560, an overall C-index of 0.617, and a brier score of 0.347, 0.330, and 0.146 for predicting 1-year, 3-year, and 5-year survival, respectively, with an IBS of 0.261 (Table 3). The IBS plots for the two models are shown in Figure 7.

TABLE 1 Demographics of patients enrolled in the training set and test set.

| Variables | | Total ($n$ = 131) | Training set ($n$ = 91) | Test set ($n$ = 40) | $p$ |
|---|---|---|---|---|---|
| Age | | | | | 0.220 |
| | ≤60 | 73 (56%) | 47 (52%) | 26 (65%) | |
| | >60 | 58 (44%) | 44 (48%) | 14 (35%) | |
| Sex | | | | | 0.979 |
| | female | 44 (34%) | 30 (33%) | 14 (35%) | |
| | male | 87 (66%) | 61 (67%) | 26 (65%) | |
| Race | | | | | 0.462 |
| | white | 20 (15%) | 12 (13%) | 8 (20%) | |
| | others | 111 (85%) | 79 (87%) | 32 (80%) | |
| KPS | | | | | 0.645 |
| | ≤60 | 93 (71%) | 63 (69%) | 30 (75%) | |
| | >60 | 38 (29%) | 28 (31%) | 10 (25%) | |
| Subtype | | | | | 0.742 |
| | Classical | 36 (27%) | 24 (26%) | 12 (30%) | |
| | Proneural | 49 (37%) | 36 (40%) | 13 (32%) | |
| | Mesenchymal | 46 (35%) | 31 (34%) | 15 (38%) | |
| CIMP_status | | | | | 0.773 |
| | G-CIMP | 116 (89%) | 81 (89%) | 35 (88%) | |
| | Non G-CIMP | 15 (11%) | 10 (11%) | 5 (12%) | |
| Radiation | | | | | 1.000 |
| | no | 102 (78%) | 71 (78%) | 31 (78%) | |
| | yes | 29 (22%) | 20 (22%) | 9 (22%) | |
| Pharmaceutical | | | | | 0.454 |
| | no | 101 (77%) | 68 (75%) | 33 (82%) | |
| | yes | 30 (23%) | 23 (25%) | 7 (18%) | |
| Survival status | | | | | 0.071 |
| | alive | 16 (12%) | 8 (9%) | 8 (20%) | |
| | dead | 115 (88%) | 83 (91%) | 32 (80%) | |
| Survival months[#] | | 12.27 (5.5, 19.9) | 13.13 (5, 22.09) | 11.71 (6.88, 17.62) | 0.581 |

[#]Continuous variables; median (range).

TABLE 2 Univariate and multivariate cox analysis of overall survival of GBM patients.

| Variables | Univariate analysis | | Multivariate analysis | |
|---|---|---|---|---|
| | Hazard ratio (95% CI) | $p$ value | Hazard ratio (95% CI) | $p$ value |
| Age | 1.576 (1.028–2.416) | 0.037 | 1.452 (0.943–2.235) | 0.090 |
| KPS | 1.890 (1.195–2.988) | 0.006 | 1.864 (1.175–2.956) | 0.008 |
| Risk level | 1.572 (1.015–2.436) | 0.043 | 1.041 (0.580–1.850) | 0.090 |
| Risk score | 3.497 (1.621–7.544) | 0.001 | 3.370 (1.499–7.573) | 0.003 |

## 4 Discussion

Precision treatment of GBM can slow down tumour growth and help improve patient prognosis. Previous studies on GBM have used deep learning for diagnostic and prognostic assessment of tumours (17, 25). To our knowledge, this is the first study to use machine learning and radiomics approaches to assess the prognosis of GBM patients. In this study, by constructing radiomics prognostic labels, using different machine learning models and comparing the performance with the traditional CoxPH model, the results show that the DeepSurv deep learning model shows superior predictive power compared to the traditional CoxPH model.

While traditional radiography focuses on the visual presentation of images, radiomics focuses on the relationship between image

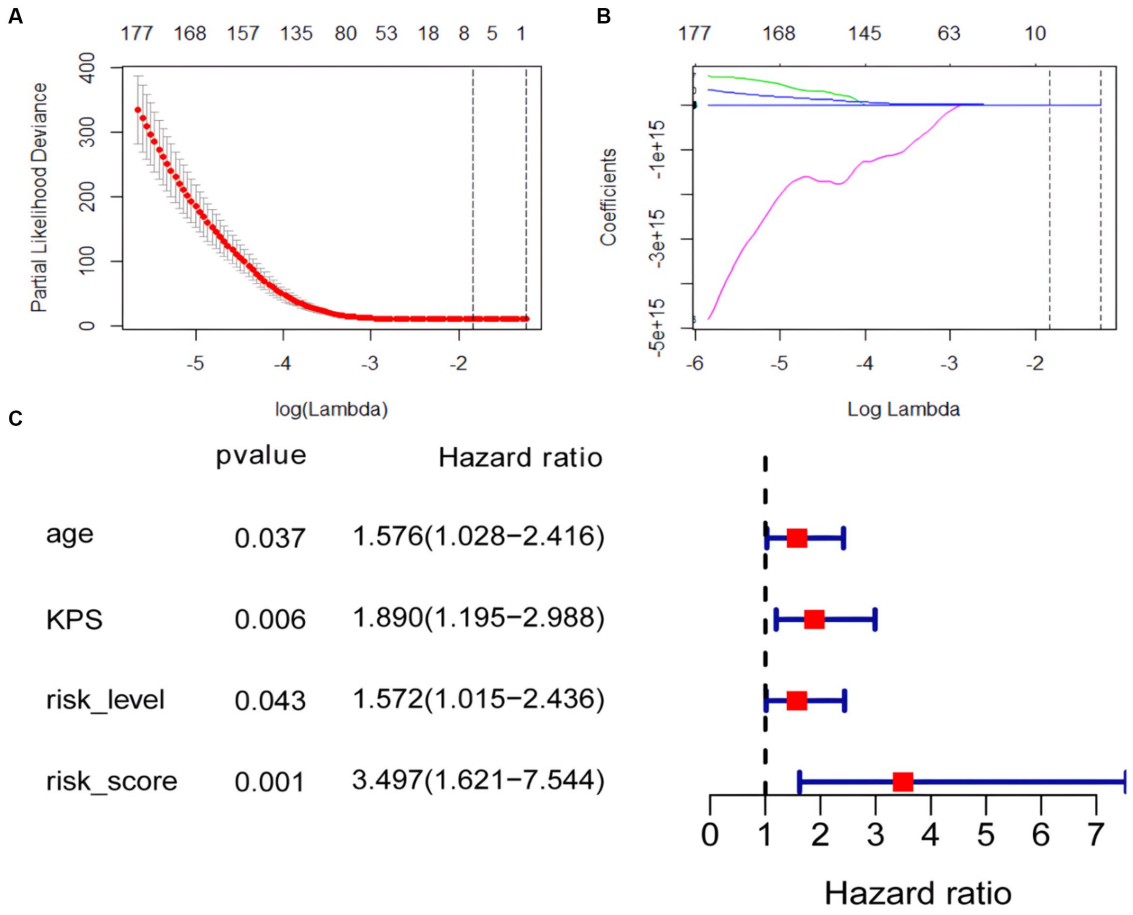


FIGURE 3
Coefficient convergence of LASSO Cox model for screening radiomics features and forest plot of univariate cox analysis. **(A)** The LASSO Cox model used tenfold cross-validation to select the optimal parameters. **(B)** The convergence of the coefficients of radiomics features under the parameters corresponding to the left figure, with each curve in the panel representing the trajectory of a feature coefficient. **(C)** Forest plot of univariate cox analysis.

TABLE 3 Hyperparameters, C-index and IBS results for the five models.

| Model | C-index | | Hyperparameters | C-index | Brier score | | | IBS |
|---|---|---|---|---|---|---|---|---|
| | Training set | Test set | | | 1-year | 3-year | 5-year | |
| CoxPH | 0.663 | 0.635 | none | 0.662 | 0.225 | 0.080 | 0.040 | 0.102 |
| Survival Tree | 0.702 | 0.655 | max_depth:5,min_samples_leaf:2,min_samples_split:12,n_estimators = 10 | 0.564 | 0.263 | 0.190 | 0.133 | 0.192 |
| RSF | 0.735 | 0.667 | max_features:sqrt,min_samples_leaf = 2,min_samples_split = 4,n_estimators = 10 | 0.642 | 0.214 | 0.143 | 0.124 | 0.152 |
| DeepSurv | 0.882 | 0.732 | num_nodes = [32,32],out_features = 1,dropout = 0.2,learning rate = 0.005 | 0.691 | 0.203 | 0.139 | 0.124 | 0.116 |
| DeepHit | 0.608 | 0.560 | num_nodes = [32,32],out_features = labtrans.out_features,dropout = 0.1,learning rate = 0.001 | 0.617 | 0.348 | 0.330 | 0.146 | 0.261 |

phenotypes and biological features and has been widely used in tumour diagnosis and prognosis evaluation (4). Studies have shown that FLAIR sequences are more advantageous in showing the extent of tumour borders and edema and that 90% of GBM recurrence occurs in the peritumoral edema area and has been shown to correlate with the prognosis of GBM (26). The FLAIR sequence was superior in showing the extent of the tumour border and edema. Some progressive patients showed no significant enhancement on the contrast scan but showed a high signal on the FLAIR sequence (27). Therefore, it is important to explore the prognostic evaluation of non-contrast FLAIR sequences in GBM. In order to construct a radiomics prognostic signature, we used the LASSO cox regression model to reduce 1792

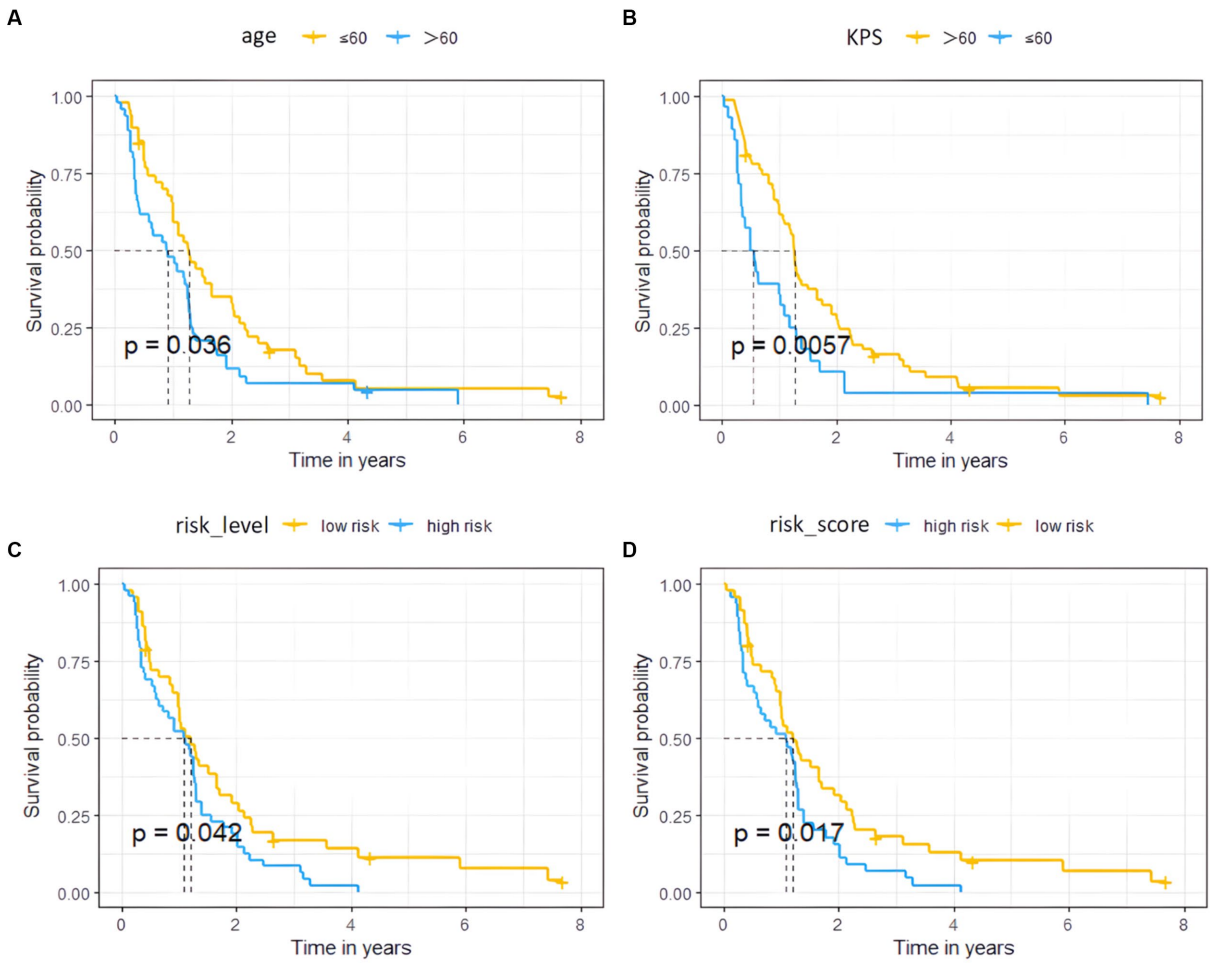


FIGURE 4
Survival curves of the high and low risk groups by univariate Cox analysis. **(A–D)** represent age, KPS, radiomics risk level, and radiomics risk score, respectively.

features to 7 potential predictive features. The results of this study showed that the seven radiomics features obtained in the FLAIR sequence were strongly associated with GBM survival, and these features indicated grey-scale heterogeneity of GBM. In addition, the radiomics risk score was shown to be an independent prognostic factor for GBM by cox univariate and multivariate analyses. The radiomics risk score was likewise the more important feature in the tree model-based feature importance ranking, suggesting that our constructed radiomics risk score can be used as a prognostic marker for GBM.

The CoxPH model is a classic approach to survival analysis and event prediction; however, the model is semi-parametric and assumes that the risk of an event is linearly related to the variables. Recently, tree-based models have received increasing attention from researchers in addressing the identification of multidimensional interactions. SurvivalTree is similar to decision trees because it is constructed by the recursive splitting of tree nodes. Compared to CoxPH, SurvivalTree is more relaxed in its requirements for survival information and does not require survival times to satisfy a specific distribution (21). RSF is a combination of random forest and SurvivalTree. The advantage of the RSF model is that it is not constrained by the assumptions of proportional risk and log-linearity (22). Also, it can prevent the overfitting problem of its algorithm through two random sampling processes (28). In our study, the SurvivalTree and the RSF model achieved a C-index of 0.70 or higher in the training set. However, as the survival tree model has fewer parameters available for adjustment and is not an integrated algorithm, it has a lower overall C-index. The IBS results for both models also showed that the RSF performed better. In addition, the AUC values for the cumulative survival times of the two models indicate a significant difference between the first and second half of the time horizon, with higher AUC values for the model in the first half of the time horizon and lower AUC values in the second half of the time horizon. Therefore, the models are most effective in predicting mid-term mortality.

Deep learning models can learn and infer higher-order nonlinear combinations between patient clinical outcomes and predictor variables in an entirely data-driven manner and have been shown to outperform standard survival analysis, with one advantage being the ability to discern complex relationships between clinical outcomes and predictor variables without prior feature selection (14). In this study, the DeepSurv model achieved the highest C-index in both the training and test set. At the same time, the overall C-index also indicated that the model was superior, suggesting that the deep learning-based

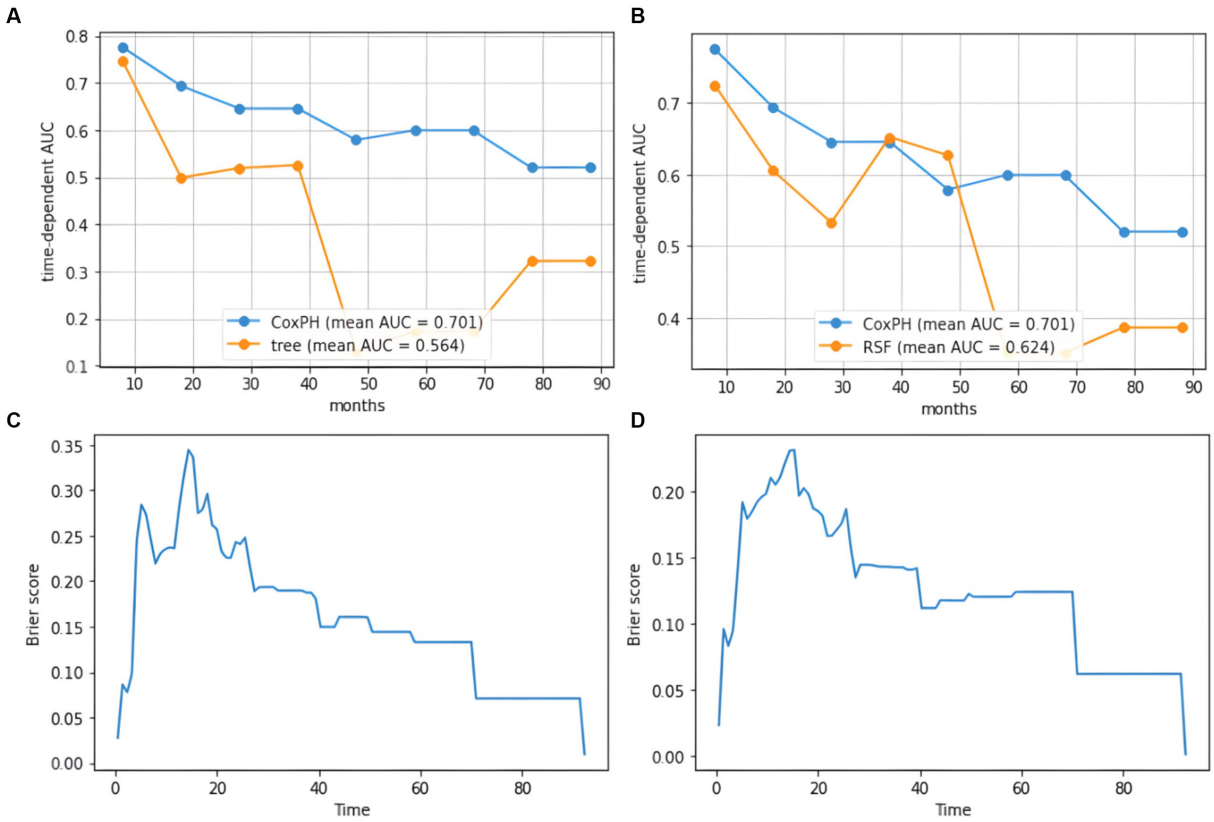


FIGURE 5
AUC plot of cumulative survival time and IBS diagram based on the tree model. **(A)** AUC results for the SurvivalTree model, **(B)** AUC results for the RSF model, **(C)** IBS results for the SurvivalTree model, **(D)** IBS results for the RSF model.

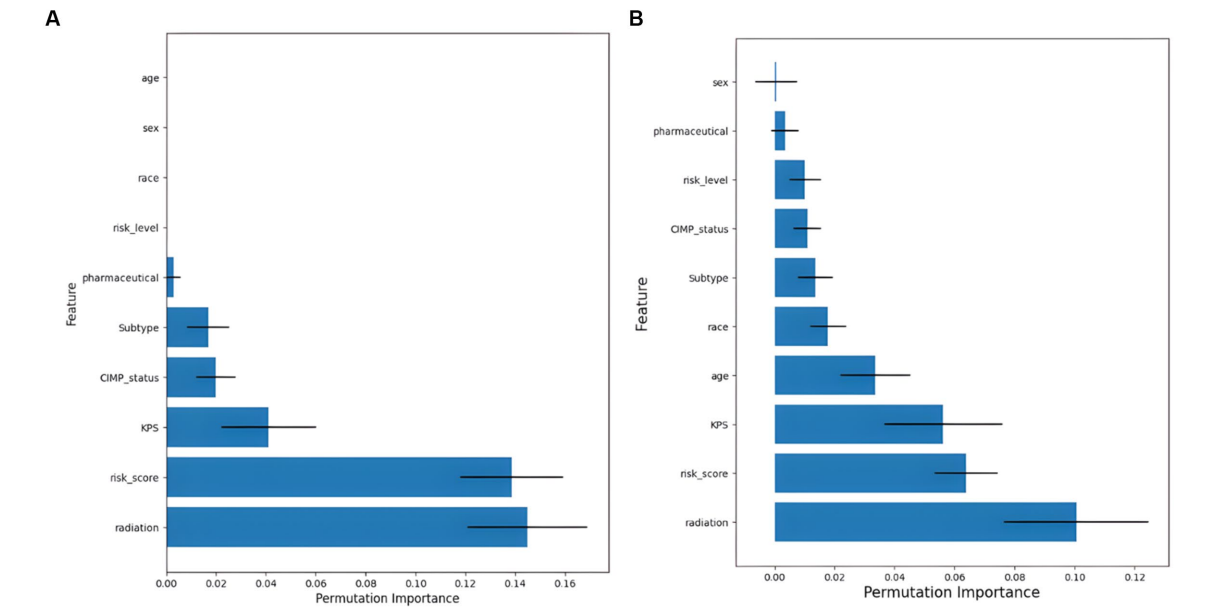


FIGURE 6
Feature importance ranking results. **(A)** Results of feature importance ranking for the SurvivalTree model, and **(B)** results of feature importance ranking for the RSF model.

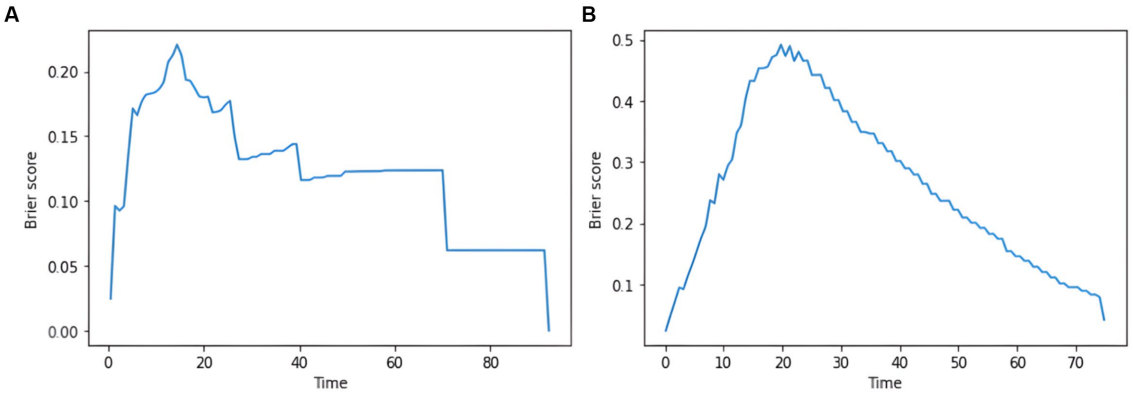


FIGURE 7
plot of the integrated brier score based on the deep learning model. **(A)** IBS results for the DeepSurv model. **(B)** IBS results for the DeepHit model.

survival model outperformed the CoxPH and RSF models in predicting GBM survival. Previous deep learning prognostic models based on clinical risk factors achieved a C-index of 0.823 and 0.700 in the training and test set, respectively (17); the present study achieved 0.882 and 0.667 in the training and test set, indicating the superior performance of the prognostic model based on radiomics features. Another deep learning model constructed in this study is DeepHit, which can directly learn the distribution of first death times and performs better in dealing with multiple competing risks (28). However, since the ending of this study is a dichotomous variable and there are no multiple competing risks, the performance of this model was not improved by hyperparameter tuning, and this model may not apply to our data structure.

There are limitations to this study. First, MRI images were collected retrospectively from the TCIA database, and the heterogeneity of different imaging parameters generated by different devices and field strengths could not be controlled. In addition, there was a relatively low number of patients in this study. Some patients also had incomplete clinical risk factors. Second, a large amount of redundant information in the sequence images leads to a considerable workload and subjectivity in manual segmentation. A more advanced approach is to use deep learning models such as CNN to learn features directly from images, which reduces the presence of subjectivity between the raters. Finally, to construct prognostic models, our study only extracted features from FLAIR images. In constructing the models, it did not make use of structural images or functional MRI techniques.

## 5 Conclusion

In conclusion, based on the TCGA and TCIA databases combined with a radiomics approach, this study confirmed that the DeepSurv model based on deep learning achieves better performance in GBM patient data compared to the CoxPH model. Based on the above-optimized model, a personalized treatment recommendation system for GBM can be developed to predict patient prognosis accurately.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary materials, further inquiries can be directed to the corresponding authors.

## Ethics statement

The study is based on the data available in the public domain to use; therefore, no ethics statement is required for this work.

## Author contributions

DZ: Data curation, Methodology, Software, Writing – original draft, Writing – review & editing. JL: Methodology, Software, Validation, Writing – original draft, Writing – review & editing. BL: Investigation, Software, Validation, Visualization, Writing – original draft. AY: Data curation, Project administration, Writing – original draft. KL: Data curation, Validation, Writing – original draft. PH: Data curation, Formal analysis, Writing – review & editing. XH: Conceptualization, Data curation, Methodology, Writing – original draft. HY: Conceptualization, Formal analysis, Investigation, Writing – original draft. AS: Writing – review & editing. GM: Funding acquisition, Supervision, Writing – review & editing, Writing – original draft. CZ: Funding acquisition, Methodology, Supervision, Writing – review & editing, Writing – original draft.

## 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 National Natural Science Foundation of China (No. 61976110, 81971585 and 82271953), Guangzhou Science and Technology Planning Project (No. 202103010001).

## Acknowledgments

The authors thank Zeshan Yao for English editing. The authors thank the TCGA platform (https://www.cancer.gov/tcga/) and the TCIA platform (https://www.cancerimagingarchive.net/) for making their data sets publicly available.

## 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/fmed.2023.1271687/full#supplementary-material

## References

1. Ostrom QT, Patil N, Cioffi G, Waite K, Kruchko C, Barnholtz-Sloan JS. Cbtrus statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2013–2017. *Neuro-Oncology*. (2020) 22:iv1–iv96. doi: 10.1093/neuonc/noaa200

2. Ferguson SD, Hodges TR, Majd NK, Alfaro-Munoz K, al-Holou WN, Suki D, et al. A validated integrated clinical and molecular glioblastoma Long-term survival-predictive nomogram. *Neurooncol Advances*. (2020) 3:vdaa146. doi: 10.1093/noajnl/vdaa146

3. Senders JT, Staples P, Mehrtash A, Cote DJ, Taphoorn MJB, Reardon DA, et al. An online calculator for the prediction of survival in glioblastoma patients using classical statistics and machine learning. *Neurosurgery*. (2020) 86:E184–92. doi: 10.1093/neuros/nyz403

4. 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

5. Lao J, Chen Y, Li Z-C, Li Q, Zhang J, Liu J, et al. A deep learning-based Radiomics model for prediction of survival in glioblastoma Multiforme. *Sci Rep*. (2017) 7:10353. doi: 10.1038/s41598-017-10649-8

6. Ammari S, Sallé de Chou R, Balleyguier C, Chouzenoux E, Touat M, Quillent A, et al. A predictive clinical-Radiomics nomogram for survival prediction of glioblastoma using Mri. *Diagnostics*. (2021) 11:2043. doi: 10.3390/diagnostics11112043

7. Zhang X, Lu H, Tian Q, Feng N, Yin L, Xu X, et al. A Radiomics nomogram based on multiparametric Mri might stratify glioblastoma patients according to survival. *Eur Radiol*. (2019) 29:5528–38. doi: 10.1007/s00330-019-06069-z

8. Bathla G, Priya S, Liu Y, Ward C, le NH, Soni N, et al. Radiomics-based differentiation between glioblastoma and primary central nervous system lymphoma: a comparison of diagnostic performance across different Mri sequences and machine learning techniques. *Eur Radiol*. (2021) 31:8703–13. doi: 10.1007/s00330-021-07845-6

9. Waljee AK, Higgins PDR. Machine learning in medicine: a primer for physicians. *Am J Gastroenterol*. (2010) 105:1224–6. doi: 10.1038/ajg.2010.173

10. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. *Nat Med*. (2019) 25:24–9. doi: 10.1038/s41591-018-0316-z

11. Xia W, Hu B, Li H, Shi W, Tang Y, Yu Y, et al. Deep learning for automatic differential diagnosis of primary central nervous system lymphoma and glioblastoma: multi-parametric magnetic resonance imaging based convolutional neural network model. *J Magn Reson Imaging*. (2021) 54:880–7. doi: 10.1002/jmri.27592

12. González G, Ash SY, Vegas-Sánchez-Ferrero G, Onieva Onieva J, Rahaghi FN, Ross JC, et al. Disease staging and prognosis in smokers using deep learning in chest computed tomography. *Am J Resp Crit Care Med*. (2018) 197:193–203. doi: 10.1164/rccm.201705-0860oc

13. Kim DW, Lee S, Kwon S, Nam W, Cha I-H, Kim HJ. Deep learning-based survival prediction of Oral Cancer patients. *Sci Rep*. (2019) 9:6994. doi: 10.1038/s41598-019-43372-7

14. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Deepsurv KY. Personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Med Res Methodol*. (2018) 18:24. doi: 10.1186/s12874-018-0482-1

15. Kim YJ, Lee H-J, Kim KG, Lee SH. The effect of Ct scan parameters on the measurement of Ct Radiomic features: a lung nodule phantom study. *Comput Math Method Med*. (2019) 2019:1–12. doi: 10.1155/2019/8790694

16. Liu K, Xia W, Qiang M, Chen X, Liu J, Guo X, et al. Deep learning pathological microscopic features in endemic nasopharyngeal Cancer: prognostic value and Protentional role for individual induction chemotherapy. *Cancer Med*. (2019) 9:1298–306. doi: 10.1002/cam4.2802

17. Moradmand H, Aghamiri SMR, Ghaderi R, Emami H. The role of deep learning-based survival model in improving survival prediction of patients with glioblastoma. *Cancer Med*. (2021) 10:7048–59. doi: 10.1002/cam4.4230

18. Koo TK, Li MY. A guideline of selecting and reporting Intraclass correlation coefficients for reliability research. *J Chiropr Med*. (2016) 15:155–63. doi: 10.1016/j.jcm.2016.02.012

19. Osadebey ME, Pedersen M, Arnold DL, Wendel-Mitoraj KE. Blind blur assessment of Mri images using parallel multiscale difference of Gaussian filters. *Biomed Eng Online*. (2018) 17:76. doi: 10.1186/s12938-018-0514-4

20. Guang D. An entropy interpretation of the logarithmic image processing model with application to contrast enhancement. *IEEE Trans Image Process*. (2009) 18:1135–40. doi: 10.1109/tip.2009.2016796

21. Nunn ME, Fan J, Su X, Levine RA, Lee H-J, McGuire MK. Development of prognostic indicators using classification and regression trees for survival. *Periodontol*. (2011) 58:134–42. doi: 10.1111/j.1600-0757.2011.00421.x

22. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. (2008) 2:841–60. doi: 10.1214/08-aoas169

23. Lee C, Yoon J, Mvd S. Dynamic-Deephit: a deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Trans Biomed Eng*. (2020) 67:122–33. doi: 10.1109/tbme.2019.2909027

24. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Stat Med*. (2007) 26:5512–28. doi: 10.1002/sim.3148

25. Pasquini L, Napolitano A, Tagliente E, Dellepiane F, Lucignani M, Vidiri A, et al. Deep learning can differentiate Idh-mutant from Idh-wild Gbm. *J Pers Med*. (2021) 11:290. doi: 10.3390/jpm11040290

26. Lemée J-M, Clavreul A, Menei P. Intratumoral heterogeneity in glioblastoma: Don't forget the Peritumoral brain zone. *Neuro-Oncology*. (2015) 17:1322–32. doi: 10.1093/neuonc/nov119

27. Grossman R, Shimony N, Shir D, Gonen T, Sitt R, Kimchi TJ, et al. Dynamics of Flair volume changes in glioblastoma and prediction of survival. *Ann Surg Oncol*. (2016) 24:794–800. doi: 10.1245/s10434-016-5635-z

28. Lee C, Light A, Saveliev ES, van der Schaar M, Gnanapragasam VJ. Developing machine learning algorithms for dynamic estimation of progression during active surveillance for prostate Cancer. *NPJ Digit Med*. (2022) 5:110. doi: 10.1038/s41746-022-00659-w

frontiers | Frontiers in Medicine

TYPE Systematic Review
PUBLISHED 18 December 2023
DOI 10.3389/fmed.2023.1240360



OPEN ACCESS

EDITED BY
Giorgio Treglia,
Ente Ospedaliero Cantonale (EOC), Switzerland

REVIEWED BY
Jiong Wu,
University of Pennsylvania, United States
Jian Shen,
Beijing Institute of Technology, China
Mariana Bento,
University of Calgary, Canada

*CORRESPONDENCE
Serestina Viriri
✉ viriris@ukzn.ac.za

RECEIVED 14 June 2023
ACCEPTED 01 November 2023
PUBLISHED 18 December 2023

CITATION
Mhlanga ST and Viriri S (2023) Deep learning techniques for isointense infant brain tissue segmentation: a systematic literature review. *Front. Med.* 10:1240360.
doi: 10.3389/fmed.2023.1240360



# Deep learning techniques for isointense infant brain tissue segmentation: a systematic literature review

Sandile Thamie Mhlanga and Serestina Viriri*

School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa

**Introduction:** To improve comprehension of initial brain growth in wellness along with sickness, it is essential to precisely segment child brain magnetic resonance imaging (MRI) into white matter (WM) and gray matter (GM), along with cerebrospinal fluid (CSF). Nonetheless, in the isointense phase (6-8 months of age), the inborn myelination and development activities, WM along with GM display alike stages of intensity in both T1-weighted and T2-weighted MRI, making tissue segmentation extremely difficult.

**Methods:** The comprehensive review of studies related to isointense brain MRI segmentation approaches is highlighted in this publication. The main aim and contribution of this study is to aid researchers by providing a thorough review to make their search for isointense brain MRI segmentation easier. The systematic literature review is performed from four points of reference: (1) review of studies concerning isointense brain MRI segmentation; (2) research contribution and future works and limitations; (3) frequently applied evaluation metrics and datasets; (4) findings of this studies.

**Results and discussion:** The systemic review is performed on studies that were published in the period of 2012 to 2022. A total of 19 primary studies of isointense brain MRI segmentation were selected to report the research question stated in this review.

KEYWORDS

**isointense infant brain, segmentation, deep learning, convolutional neural networks, magnetic resonance imaging**

## 1 Introduction

In brain research, the precise separation of infant brain tissues into non-overlapping regions such as white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF) is crucial for determining how the normal and abnormal development of the developing brain (1–3). The first year of life is the most dynamic period in the development of the human brain, with fast tissue growth and the emergence of a vast array of cognitive and physical abilities (4, 5). Major brain diseases that are difficult to treat, such as attention deficit hyperactivity disorder (ADHD), baby autism, bipolar affective disorder, and schizophrenia, may show up in the patient's developing brain tissue (6). Therefore, it is important that brain structures are adequately segmented in new-born images. The aim of precise brain tissue image segmentation is to provide crucial information for clinical diagnostics, treatment assessments, analysing brain changes, enabling clinical preparations together with presenting image-guided interventions (7–9).

Thus far, magnetic resonance imaging (MRI) is the predominant technique for imaging baby brain, specifically T1-weighted and T2-weight MRI, because it is safe, non-invasive and attains

non-intrusive cross-sectional views of the brain in multiple contrast without ionizing radiation (10, 11). Compared to automated segmentation, manual segmentation is tremendously arduous and time-consuming assignment which compels a comprehensive expertise base of brain structure and impossible at large scale. In addition, manual segmentation experiences small reproducibility, which is highly inclined to errors due to inter or inter-operator unpredictability (7, 8, 12, 13). Therefore, precise and automatic segmentation methods are highly needed.

Infant brain MRI segmentation is recognized to be far more challenging than adult brain segmentation (5), due to ongoing white matter myelination, significant partial volume effects, decreased tissue contrast (14), increased noise, and infant brain pictures (14, 15). In actuality, as depicted in Figure 1, there are three distinct phases in the first-year brain MRI (16). Gray matter exhibits a higher signal strength than white matter in T1-weighted images during (1) the infancy phase (5 months). The gray matter has the lowest signal differentiation with the white matter in both T1 and T2 imaging during the second isointense phase (6–9 months), in which the signal intensity of white matter is growing during development due to myelination and maturation process. The final stage is the early adult-like stage (9 months), where the distribution of gray matter intensity in T1 images is significantly lower than that of white matter, resembling the pattern of tissue contrast in adult MRI (5, 16).

Furthermore, the intensity distributions of the voxels in the gray and white matter continue to heavily overlap in the isointense stage, particularly in the cortical areas, in this way driving to the least tissue differentiation and making the primary challenging for tissue segmentation, in relationship to pictures on previous stages of brain development (5, 16–18). Numerous efforts have been made in the past few years to segment the baby brain using MRI (4, 6, 11, 19–28).

Despite having an array of infant brain segmentation models, to determine which segmentation techniques are most frequently employed and in what combinations, there is a need to assess the body of literature as a whole using a systematic literature review paper. By doing this, the restrictions on personal searches for isointense brain MRI segmentation models would be lessened. What are the current isointense brain MRI segmentation algorithms, and what are the application challenges? Is the main research question leading this systematic literature review (SLR). As a result, the study's goal is to examine isointense brain MRI segmentation models utilizing a literature review.

## 2 Literature review

As of late, deep learning techniques centred around convolutional neural networks (CNNs) have demonstrated exceptional execution around a range of computer visualization and photograph evaluation usages in the clinical space (16, 17, 29–32). CNNs have accomplished advanced outcomes in numerous brain segmentation tribulations (7, 8, 12, 33–36), including the subdivision of 6-moths old brain MRI (1, 11, 21, 22, 24, 25, 32, 37, 38).

Some researchers have refined many recognized CNNs, for example U-Net (36, 38, 39) and the DenseNet (11, 21, 24, 34), for brain MRI division on 6-months-old child (1, 40, 41). These methods improve the viable conveyance and combination of the semantic data in a multimodal characteristics and have accomplished enhanced functioning contrasted with common machine learning techniques (16, 17). Nevertheless, inadequacies however occur in the present CNN-based division techniques for child brain for example, previous models focus on enhancing network architecture for example modality blend (41) and interlayer links (37, 42), which requires seasoned expertise experience for network designing and the training turn out to be more challenging as the network amplifies the depth (21). Furthermore, hardware requirements for computing and memory escalates drastically as the depth increase (21). Combination of these methods for improved performance is also problematic due to the inconsistence network layouts, tedious hyper-parameter

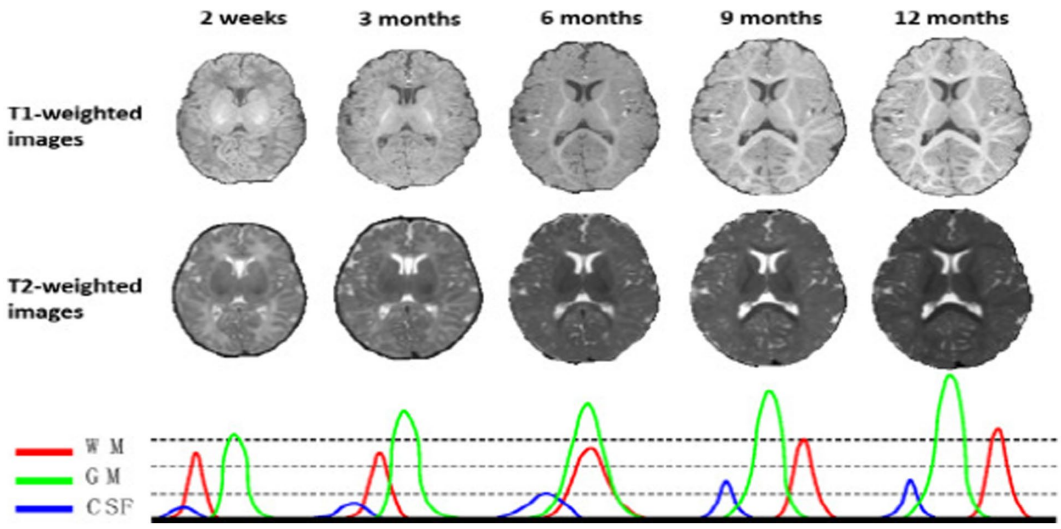


FIGURE 1
T1 and T2 weighted MRI images of a baby taken at various ages—2 weeks, 3, 6, 9, and 12 months. The MR images of infants around 6 months old (i.e., the isointense phase) show the lowest tissue contrast, indicating the most difficult tissue segmentation. The bottom row displays the equivalent tissue intensity distributions from T1w MR images, where the WM and GM intensities are heavily overlapping during the isointense period. Reprinted with permission from IEEE, Copyright © 2019 IEEE (16).

TABLE 1 Keywords used in this research.

| Automatic Image Segmentation construct | AND | Group of participants' construct | OR | Characteristic of interest construct |
|---|---|---|---|---|
| "Automatic segmentation" OR | | "Isointense" OR | | "brain MRI" OR |
| "Image segmentation" OR | | "6-months" OR | | "brain MRI tissues" OR |
| "Brain tissue segmentation" OR | | "Infant" OR | | "white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF)" OR |
| "Segmentation" | | "Neonatal" OR | | "MRI brain tissues" |

alteration and extreme graphic processing unit (GPU) memory utilization (17).

# 3 Methodology

The process of finding and critically evaluating pertinent research, as well as gathering and analysing data from this research, is known as a systematic literature review, or SRL (43). A systematic review's objective is to locate all empirical data that satisfies the inclusion criteria and provides an answer to a particular research question (43, 44). Additionally, it takes time to separate the known from the unknown. That is a crucial justification for conducting SLRs in accordance with a set of clear-cut methodological stages (45). This study established a systematic literature review (SLR) on the segmentation of isointense brain MRI using the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA). PRISMA is a well-known systematic review methodology that has been used in a variety of research domains, including the medical field (46), business (47) and safety mining (45). Because of its 27 evidence-based checklist and four-phase analysis, PRISMA is acceptable in the research area even if it is not a quality assessment approach. This allows systematic literature reviews (SLRs) to be clear and transparent (43, 48). Identification, screening, eligibility, and data abstraction and analysis are the four core PRISMA phases. This systematic review was conducted from 1 August 2022 to 31 December 2022.

## 3.1 Research question

This study assesses segmentation results of isointense brain MRI studies that have been conducted in the past. For the purpose of describing the systematic literature review, the following four research questions have been developed.

- [RQ-1] What techniques have been used for isointense brain MRI segmentation in neurosciences?
  - [RQ-1a] What are the existing isointense brain MRI segmentation machine learning algorithm?
  - [RQ-1b] What evaluation metrics have been used to measure accuracy of the techniques?
- [RQ-2] What are the characteristic of the dataset used in neurosciences for isointense brain MRI segmentation?
- [RQ-3] What are the findings of isointense brain MRI segmentation in this study?
- [RQ-4] What are the future works and limitations to ease the other researchers search for isointense brain MRI segmentation?

## 3.2 PRISMA phases

### 3.2.1 Identification

The identification stage is the first step in the systematic literature review (SLR) process. The study question and goals are clearly defined at this point. A widespread search study was executed using Web of Science (WoS) and Scopus. All significant publishers, including Science Direct, Emerald, Taylor & Francis, Springer Links, IEEE, and Willey, are included in the Scopus integrated database. Due to its high calibre indexing information, many academics have regarded the Scopus database as a trustworthy resource for SLR. All appropriate peer-reviewed articles published between 2012 and December 31, 2022, are included in the search. When looking for pertinent publications, use terms like "*automatic isointense MRI brain segmentation*," "*Image segmentation 6-month brain MRI*," "Infant *brain tissue segmentation*," and "*Segmentation neonatal brain MRI*." The Boolean operators are combined with various keywords to enlarge the search range 634 articles were obtained as a consequence of this method from the combined Scopus and Web of Science databases (Table 1).

### 3.2.2 Screening

The subsequent stage is the screening procedure, in which articles are included or excluded based on standards set by the writers. Tables 2–4 provide specifics regarding inclusion and exclusion. Following the identifying procedure, 634 articles needed to be screened. Duplications were identified and removed, and 580 for the title and abstract screening, articles were found. Relevant articles were forwarded to the candidate data. After reviewing all available literature, the candidate data set was reviewed, and the inclusion and exclusion criteria were used to populate the chosen data. The screening stage produced 167 publications that were only focused on isointense brain MRI segmentation and were published between January 2012 and December 31, 2022. Journals that published systematic reviews, review papers, proceedings from conferences, book chapters, book series, and novels were not included. The goal is to concentrate on legitimate isointense brain MRI segmentation research.

### 3.2.3 Eligibility

The third phase is the eligibility procedure, in which articles are included or eliminated according to the precise standards set forth by the writers. Manual screening of literature with a focus on the segmentation of isointense brain MRI and the inclusion and exclusion criteria from previous screening processes. The review was able to collect 19 carefully chosen articles on isointense brain MRI segmentation.

TABLE 2 Literature inclusion criteria.

| Number | Criteria | Inclusion |
|---|---|---|
| 1 | Primary Source | Literature describes data collected and analysed by the authors and not based on the other research conclusion |
| 2 | Relevant topic | Literature directly references isointense infant brain image segmentation and provide analysis of the proposed models and the metrics used to evaluate the models |
| 3 | Publication timeline | January 2012 – December 2022 |
| 4 | Review quality | Literature is published in a peer-reviewed journal |
| 5 | Dataset used | Studies that use iSeg-2017 and iSeg-2019 dataset. |
| 6 | Data quality | Literature must show data sources are numerous enough, qualified enough and representative enough to avoid bias in qualitative literature. |

TABLE 3 Literature exclusion criteria.

| Number | Criteria | Exclusion |
|---|---|---|
| 1 | Secondary Source | Article is a secondary source. Secondary data can distort this analysis by presenting a single model with multiple results. |
| 2 | Irrelevant studies | Literature that does not reference infant brain image segmentation, specifically isointense (6–8 months) |
| 3 | Publication timeline | 2011 and before |
| 4 | Document type | Journals (systematic review), review papers, conference proceedings, dissertations, these, white papers, incomplete bibliographic records, industry reports, others on the basis of relevance, chapters in a book, book series, books |
| 5.1 | Unavailability | Literature is not available as a full-text article in the selected data source. |
| 5.2 | | Literature not available in research data source at the time of data collection. |
| 6.1 | Inadmissible quality | Literature is not published in a peer-reviewed journal. |
| 6.2 | | Literature does not adequately or completely its methodology such that it cannot determined how the model was created and evaluated. |
| 6.3 | | Literature were T1-weighted and T2-weight MRI are not used. |
| 6.4 | | Literature were fetal MRI imaged was used. (0–5 months). |
| 6.5 | | Literature were not all 3 tissues (WM, GM and CSF) are segmented. |
| 7 | Language | Literature is not in English |
| 8 | Duplication | Literature is a duplicate of other literature in the study. |

TABLE 4 Quality assessment checklist adopted from Kitchenham et al. (49) as cited by Usman et al. (50).

| NO# | Question | Score |
|---|---|---|
| 1 | Are the research aim clearly specified? | Y\|N\|P |
| 2 | Was the study designed to achieve these aims? | Y\|N\|P |
| 3 | Are the segmentation techniques clearly described? | Y\|N\|P |
| 4 | Are the evaluation metrics used adequately described | Y\|N\|P |
| 5 | Are all research question answered adequately? | Y\|N\|P |
| 6 | Are negative (if any) presented? | Y\|N\|P |
| 7 | Are datasets considered by the study? | Y\|N\|P |
| 8 | Is the purpose of data analysis clear? | Y\|N\|P |
| 9 | Do the researcher discuss any problems with validity/reliability of the results | Y\|N\|P |
| 10 | How clear are the links between data interpretation and conclusion? | Y\|N\|P |
| 11 | Are finding based on multiple projects | Y\|N\|P |
| 12 | Are statistical techniques are used to analyse data adequately? | Y\|N\|P |
| 13 | Are data collection method adequately described? | Y\|N\|P |

### 3.2.4 Data abstraction and analysis

Data abstraction and analysis come last. The remaining publications were assessed, examined, and analysed, and 19 were chosen for in-depth discussion in this paper (see Table 5). Reviews were based on particular studies that addressed the study's research issue and purpose. Then, by reviewing the article's title, abstract, and full text, the studies were extracted to find pertinent themes for the current study. Figure 2 depicts a synopsis of the SLR procedure. In this

TABLE 5 Summary of the 19 selected studies using PRISMA approach for isointense brain tissue segmentation.

| Authors | Techniques | Modality | Development stage at scan | | |
|---|---|---|---|---|---|
| | | | Infantile | Isointense | Early-Adult |
| (15) | - | T1, T2 | | ✓ | |
| (20) | K- Nearest Neighbour | T1, T2 | ✓ | ✓ | |
| (5) | Multi-Atlas | T1, T2, FA | ✓ | ✓ | ✓ |
| (18) | Random Forest | T1, T2, FA | ✓ | | |
| (27) | 2D CNN | T1, T2, FA | | ✓ | |
| (25) | SVM | T1, T2 | | ✓ | |
| (51) | Random Forest | T1, T2 | | ✓ | |
| (2) | 3D CNN | T1, T2 | | ✓ | |
| (21) | 3D CNN | T1, T2 | | ✓ | |
| (24) | 3D CNN | T1, T2 | | ✓ | ✓ |
| (34) | 3D CNN | T1, T2 | | ✓ | |
| (6) | FCN | T1, T2 | | ✓ | |
| (52) | 3D CNN | T1, T2 | | ✓ | |
| (42) | 3D CNN | T1, T2 | | ✓ | |
| (53) | CNN | T1, T2 | ✓ | ✓ | |
| (54) | 2D CNN | T1, T2 | | ✓ | |
| (28) | 3D FCN | T1, T2 | | ✓ | ✓ |
| (55) | 3D CNN | T1, T2 | | ✓ | ✓ |
| (56) | GAN | T1, T2 | | ✓ | ✓ |

study, quality assessment was based on the checklist suggested and provided by Kitchenham et al. (49) as cited by Usman et al. (50). A three-point scale was used in this study which is Yes/ NO/ Partial. Yes (Y), represented 1, Partial represented 0.5, and NO represented 0, This study used first quartile as the cut-off point which is 3.25. If a study scored less than 3.25 it would be removed from the primary studies.

The scoring process was Y = 1, P = 0.5, N = 0.

# 4 Results

Below, you will find a review of these 19 studies, with four categories of the methodologies that were examined. Knowledge-driven segmentation methods are covered in the first section. These methods are based on the use of advanced knowledge of brain morphology, including information on the relative position, connection, and structure of brain tissue. The second section presents an approach atlas-based and patch-driven approach. Methods that primarily rely on propagated atlas labels, registration techniques for the best atlas alignment, and various label fusion techniques for multi-atlas methods are all examples of atlas-based approaches. The third section presents machine learning methods such as random forest, k-nearest (*k*NN) neighbour and support vector machine (SVM). When a multi-class classifier is used to create a brain tissue probability map for each tissue type (i.e., WM, GM, CSF), these supervised algorithms are intrinsically well suited for multi-class challenges. Convolution neural network-based deep learning techniques are covered in the final section. In a variety of computer vision applications, including the segmentation of infant brain MRI, CNN has displayed exceptional achievements (42, 57).

## 4.1 Knowledge-based approach

By incorporating knowledge of tissue connectivity, structure, and relative placements (15), offer a brain MRI segmentation technique that is based on general and widely acknowledged knowledge of neonatal brain morphology. The authors, for instance, utilised knowledge that the extra-ventricular CSF surrounds the cerebral gray matter, which is itself surrounded by the cortical white matter. The outline in Figure 3 summarizes the segmentation algorithm's five steps. The procedures involve removing the brain's intracranial cavity and hemispheres, detecting subcortical gray matter, separating cortical gray matter, unmyelinated white matter, and CSF, segmenting the cerebellum and brain stem, and detecting unmyelinated white matter (15). An infant's brain's T1 and T2 MR scans served as the algorithm's input data.

## 4.2 Atlas-based and patch-driven approach

The authors provide a basic framework for isointense new-born brain MRI segmentation that uses sparse representation to combine the information from multiple imaging modalities (5). The authors initially create a library made up of a collection of multi-modality images from the training subjects and the ground-truth segmentations that match to those images. T1 and T2 images as well as fractional anisotropy (FA) images make up multi-modality. The training library patches provide a sparse representation of each patch needed to segment a brain image. The generated sparse coefficients are then used to obtain the first segmentation. The initial segmentation will be further considered in light of the patch similarities between the

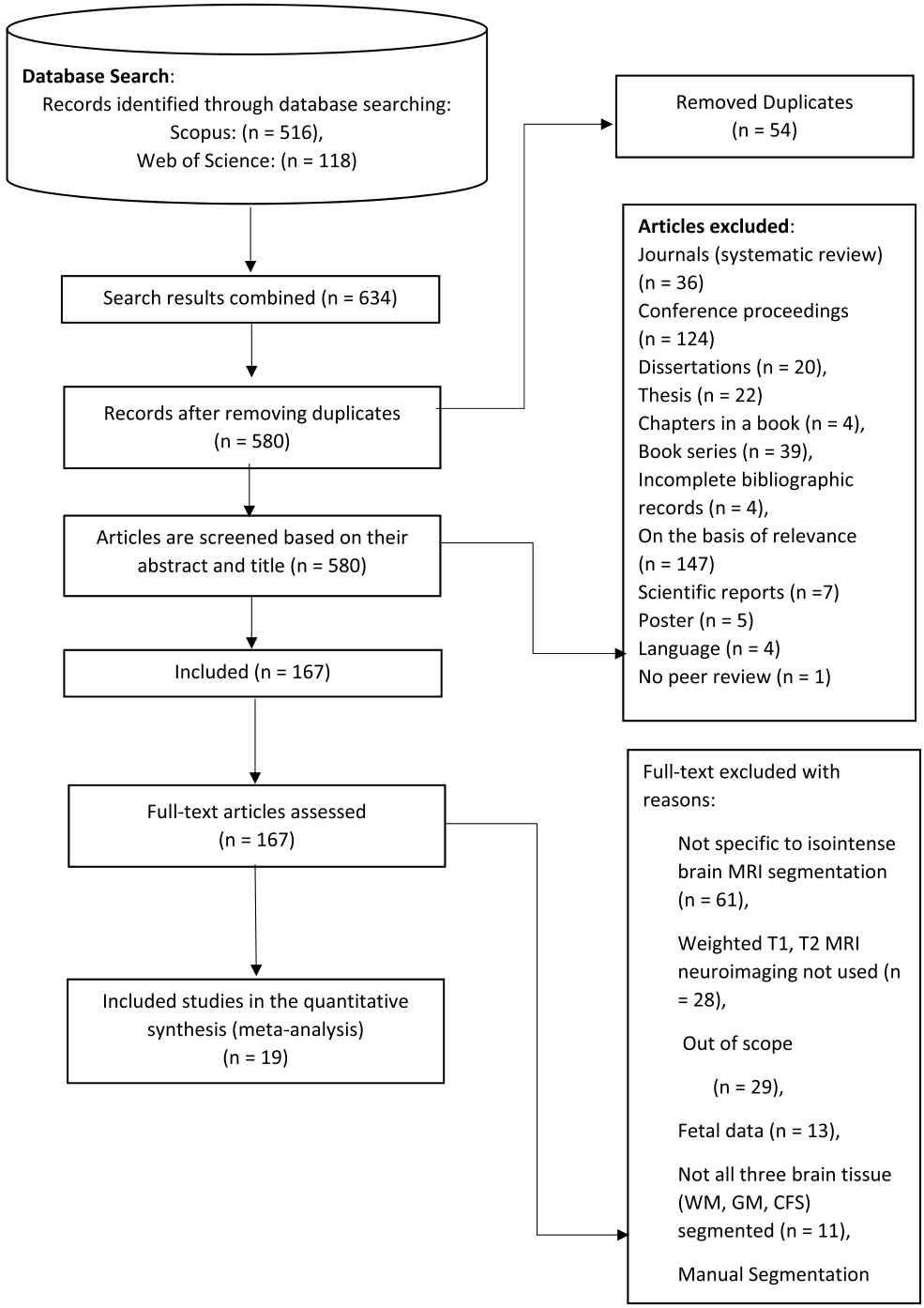


FIGURE 2
The step of PRISMA for the systematic literature review. Adapted with permission from Liberati et al. (43), licensed under CC BY.

segmented testing picture and the manual segmentation (ground-truth) in the library images in order to enforce the anatomical correctness of the segmentation (5). Figure 4 illustrates the tissue probability maps calculated using the suggested approach.

## 4.3 Machine learning approaches

A segmentation technique based on supervised pixel categorization is suggested by Anbeek et al. (20). Both spatial and intensity characteristics were provided for each voxel. Each brain voxel was classified into one of the eight tissue classes using the k-nearest neighbour (kNN) classifier based on these characteristics. A preterm cohort of 108 infants' T1- and T2-weighted MR images were obtained at term equivalent age. The brainstem, cerebellum, cortical and central grey matter, unmyelinated and myelinated white matter, cerebrospinal fluid in the ventricles and in the extra cerebral space were all segmented into eight classes using an automatic probabilistic segmentation method. Using leave-one-out tests on seven photos for which a reference standard had been

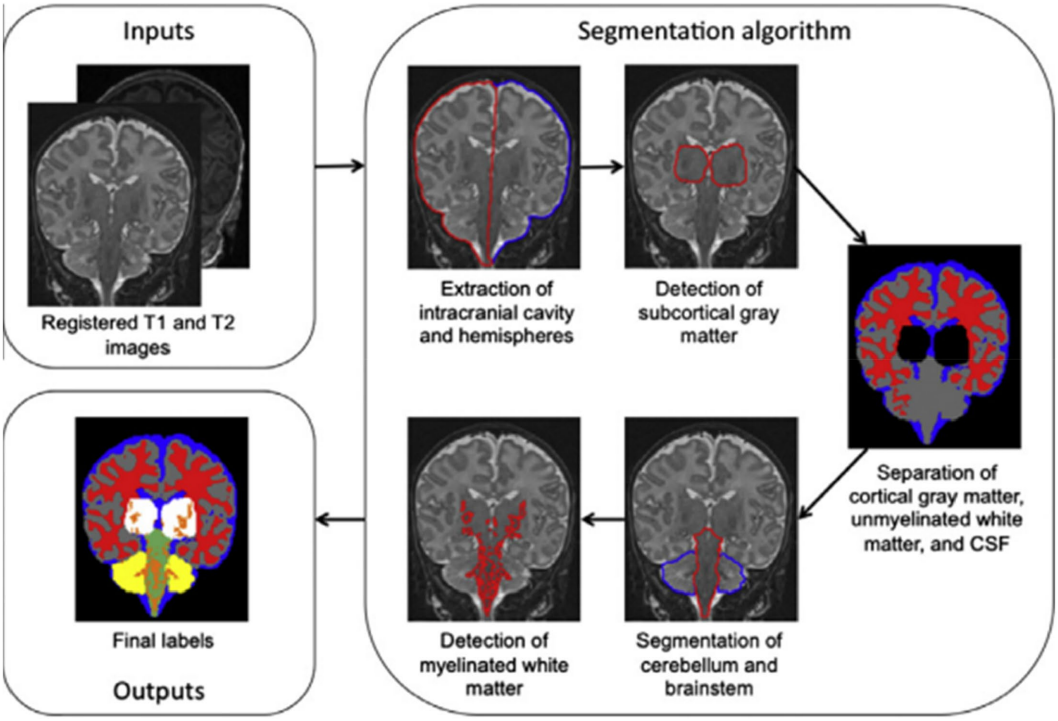


FIGURE 3
Outline of the segmentation algorithm. Reprinted with permission from Elsevier, Copyright © 2012 Elsevier (15).

manually established by a subject matter expert, the approach was trained and evaluated (20). The approach was then used on the remaining 101 scans, and the segmentations that resulted were assessed visually by three specialists. The volumes of the eight groups of segmented tissue were then calculated for each subject (20).

A strategy based on learning, employing random forest classifier for infant brain MRI segmentation is proposed by Sanroma et al. (25), Wang et al. (51), and Wang et al. (18). The authors propose a novel learning-based multi-source integration architecture for segmentation (18), where the tissue segmentation challenge is formulated as a tissue categorization challenge. In particular, tissue probability maps for each tissue type can be produced via voxel-wise classification using the random forest classifier, which is naturally suited for multi-class situations. In order to completely capture both local and contextual picture information, a large amount of training data with high data dimensions can be handled by random forest. This allows for the exploration of a huge number of image features. Additionally, an anatomy-guided tissue segmentation for 6-month-old new-born brain MRIs with autism risk was presented by Wang et al. (51). Intensity images' 3D Harr-like feature extract is input to a random forest classifier, which outputs a class classification. Figure 5 shows a training flowchart for a series of classifiers for WM versus GM. A combination of strategies is presented by Sanroma et al. (25) for infant brain MRI segmentation. The standard approaches include support vector machine (SVM) and multi-atlas joint label fusion, which serve as examples of registration-based methods. A collection of several annotated photos is necessary for both registration and learning-based approaches.

## 4.4 Deep learning methods

As of late, deep learning techniques centred around convolutional neural networks (CNNs) have demonstrated exceptional execution around a range of computer visualization and photograph evaluation usages in the clinical space (30, 31, 35, 36, 39). Convolutional neural networks were used in the majority of the publications found through the systematic literature review study using the PRISMA approach; 12 out of the 19 articles used CNNs.

### 4.4.1 Deep fully convolutional neural networks

Deep convolutional neural networks (CNN) are suggested for multi-modality MRI segmentation of isointense brain tissue (27). According to Figure 6, the authors created CNN architectures with three input feature maps for $13 \times 13$ T1, T2, and FA image patches. There are three convolutional layers and one fully connected layer used. Local response normalization and softmax layers were also used in this network.

It is advised that more research be done on deep convolutional neural networks and suggestive annotations for new-born brain MRI segmentation (42). This study uses an ensemble of semi-dense fully convolution neural networks with T1- and T2-weighted MRI as the input to examine the issue. The study shows that there is a strong correlation between segmentation mistakes and ensemble agreement. The approach thus offers measurements that can direct local user corrections. The performance of deep architectures was also examined by the authors in relation to the effects that early or late fusion of various image modalities might have (42).

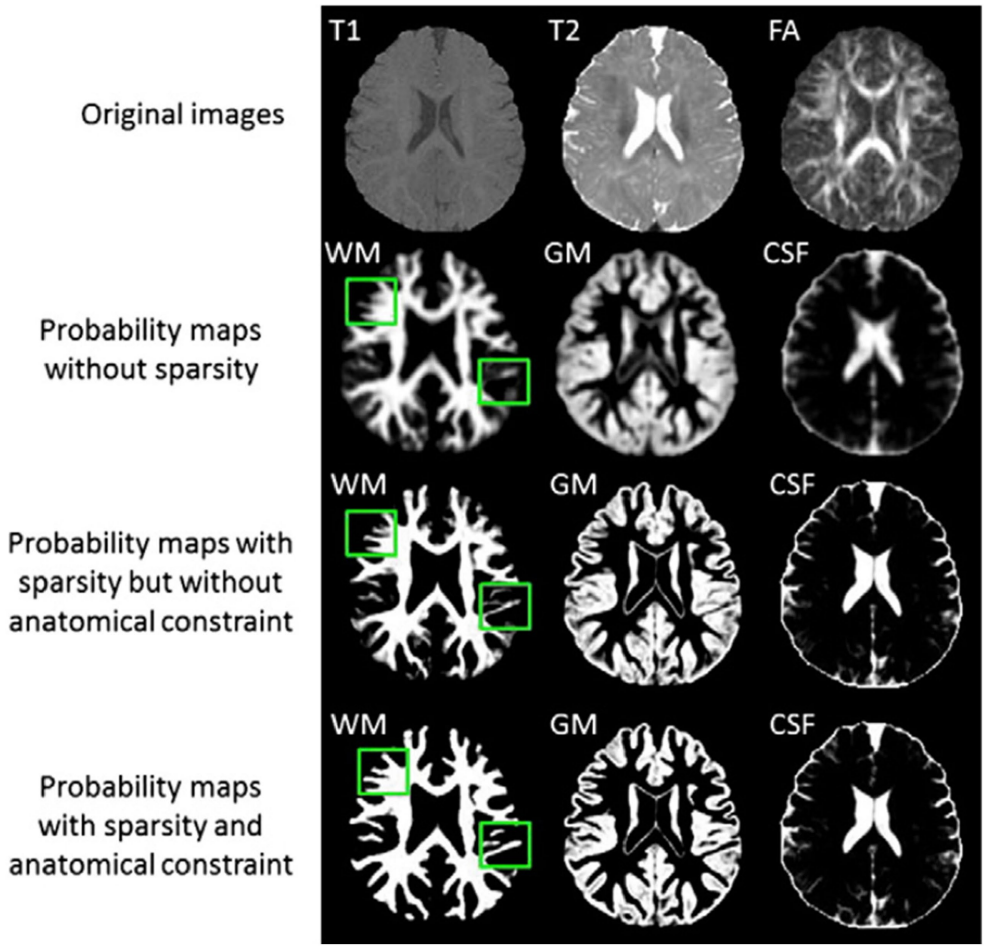


FIGURE 4
Tissue probability maps calculated using the suggested approach without and with the anatomical restriction, as well as with and without the sparse constraint. Reprinted with permission from Elsevier, Copyright © 2014 Elsevier (5).

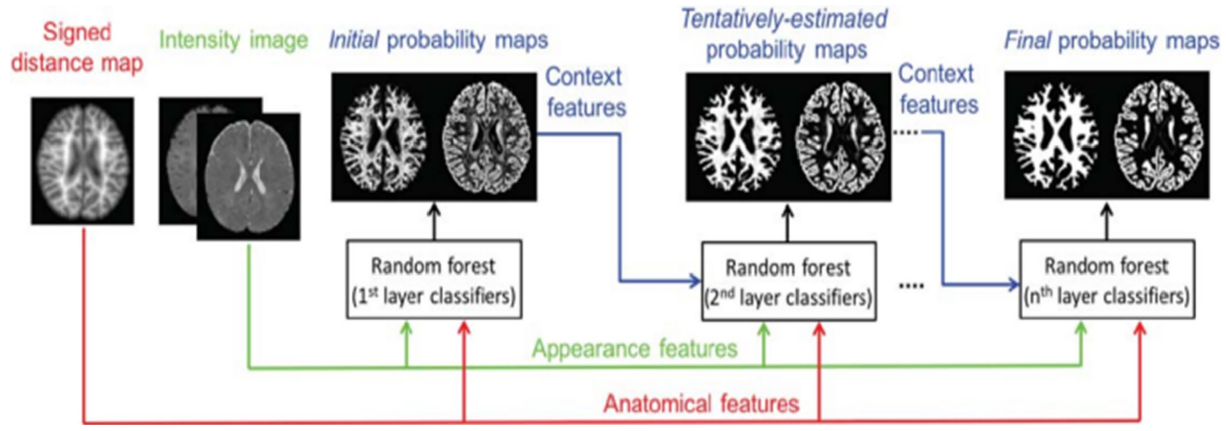


FIGURE 5
Training flowchart for a series of classifiers for WM versus GM. Reprinted with permission from Wiley, Copyright © 2018 Wiley (51).

A fuzzy-informed deep learning segmentation guided network by pertinent principles, as well as building blocks to learn multimodal information from MRI images, are also proposed by Ding et al. (55). Figure 7 shows the architecture, which consists of three primary processing steps: deep supervision, fuzzy-enabled multi-scale learning, and image refinement. A volumetric fuzzy pooling layer applies fuzzification, accumulation, and de-fuzzification to the neighbourhoods of adjacency feature maps to mimic the local

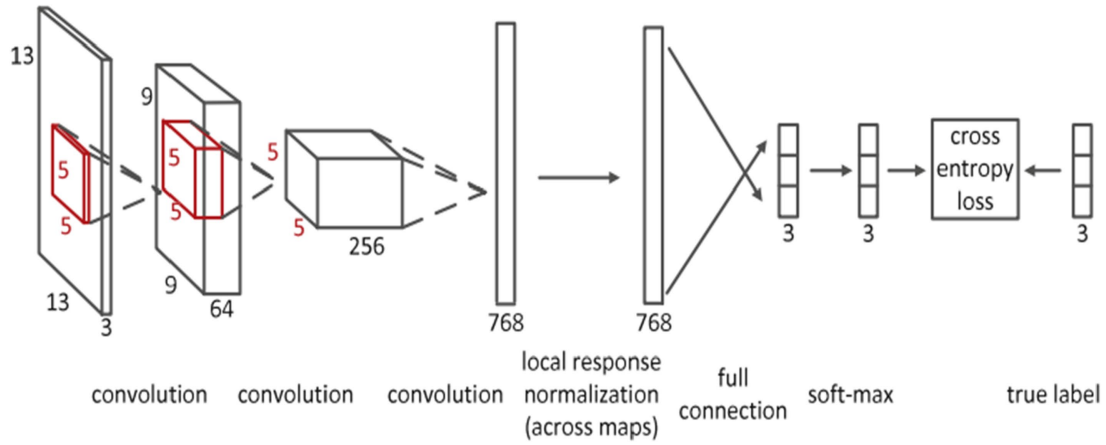


FIGURE 6
Convolutional neural network's detailed architecture using inputs in patches that are 13 by 13 in size. Reprinted with permission from Elsevier, Copyright © 2015 Elsevier (27).

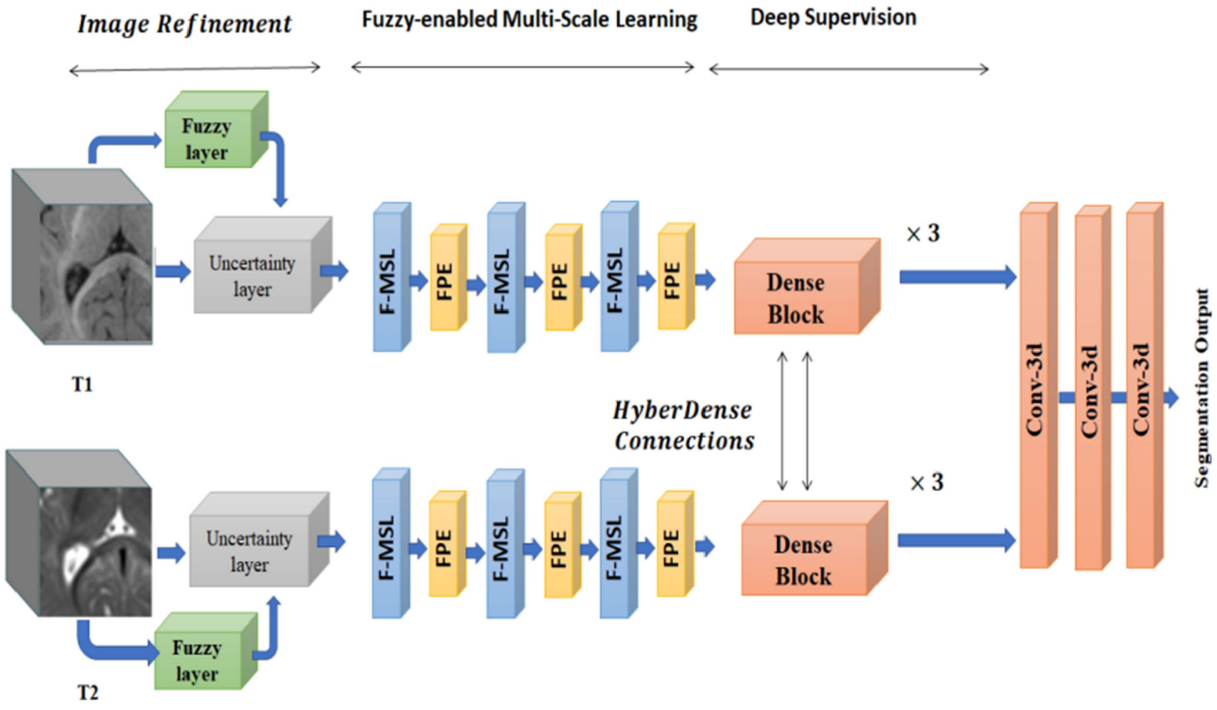


FIGURE 7
The structure of the fuzzy-guided framework that has been presented for multimodal brain MRI segmentation. Reprinted with permission from IEEE, Copyright © 2022 IEEE (55).

fuzziness of the volumetric convolutional maps. To enable the extraction of brain characteristics in various receptive fields, the fuzzy-enabled multiscale feature learning module is designed using the VFP layer. A fuzzified multichannel dense model for multimodal segmentation has also been introduced.

A powerful 2D convolutional network called Rubik-Net uses the bottleneck structure and residual connections to improve information transfer while requiring fewer network parameters. On the iSeg2017, iSeg2019, and BrainWeb datasets, the Rubik-Net demonstrated good results in terms of segmentation accuracy (54).

### 4.4.2 Hyper densely connected CNNs

Hyper-densely connected CNNs have been employed by Basnet et al. (53), Bui et al. (21), Dolz et al. (2), Hashemi et al. (34), and Qamar et al. (24) in isointense infant brain MRI segmentation. The idea of dense connection is extended to multi-modal segmentation problems by a 3D fully convolution neural network developed by Dolz et al. (2). Each image modality has a path, and dense connections can be shown in Figure 8 for both airings of layers that are on the same path as one another as well as layers that are on distinct paths.

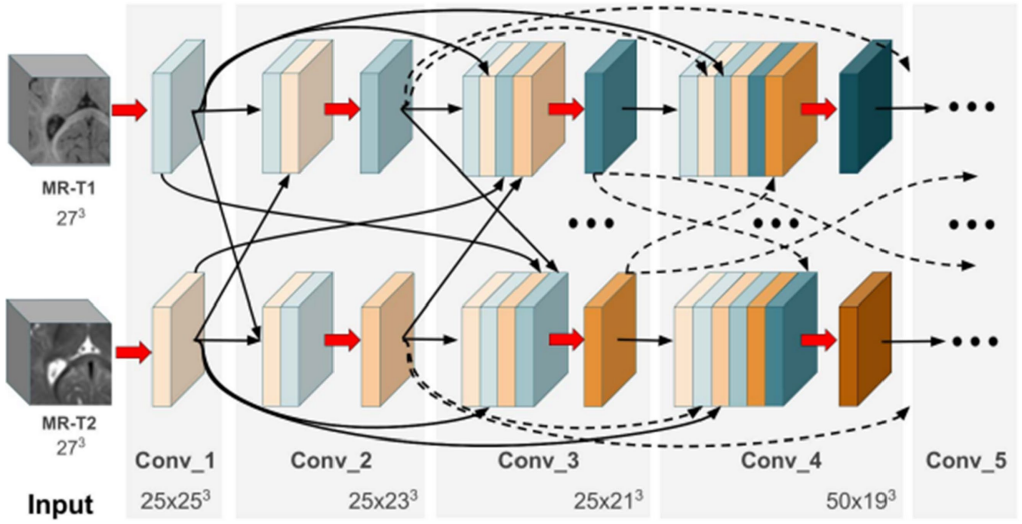


FIGURE 8
In the case of two picture modalities, a portion of the proposed HyperDenseNet. Each area of gray stands for a convolutional block. Black arrows denote dense connections between feature maps, while red arrows represent convolutions. Reprinted with permission from IEEE, Copyright © 2019 IEEE (2).

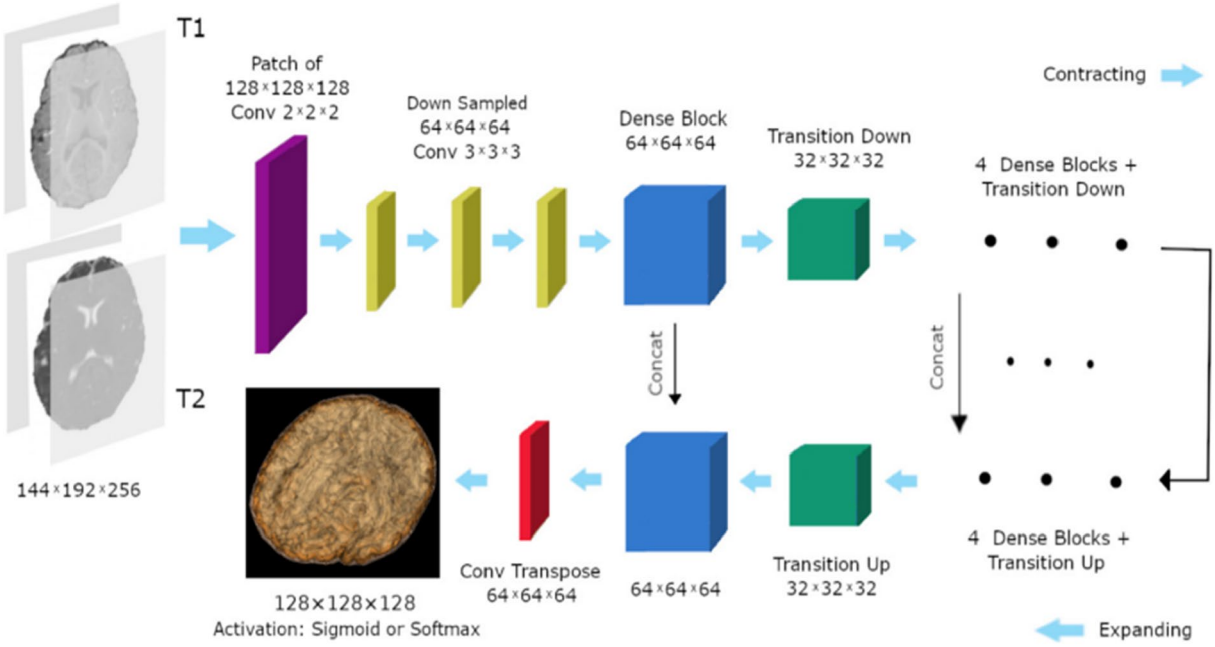


FIGURE 9
The study's 3D FC-DenseNet architecture uses a 222 convolution with stride 2 (purple) to downscale the input patch from 128 × 128 × 128 to 64 × 64 × 64 in the first layer. The patch is upsampled from 64 × 64 × 64 to 128 × 128 × 128 using a 222 convolution transpose with stride 2 (red) before the activation layer. With the help of this deep architecture, we were able to overcome memory size restrictions with big input patches, retain a wide field of vision, and add five skip connections to enhance the flow of local and global feature data. Reprinted with permission from, licensed under CC BY-4.0 (34).

A deep densely connected network called 3D FC-DenseNet has been suggested by Hashemi et al. (34). Due to its early downsampling and late upsampling layers, the network in Figure 9 has eight times the usual patch sizes (128 × 128 × 128 vs. 64 × 64 × 64), more depth, skip connections, and parameters than its predecessors.

"Deeper is the better" concepts plays an important role in deep learning architecture (24). A hyper-densely connected convolution neural networks for segmentation of infant brain MRI is presented by Qamar et al. (24). The suggested model offers close connections between layers to enhance the network's flow information performance. The algorithm employs T1 and T2 as input. On the other hand (21), carefully designed a fully convolutional densely connected network with skip connections, allowing for the direct combination of data from various densities of dense blocks to produce extremely precise segmentation results.

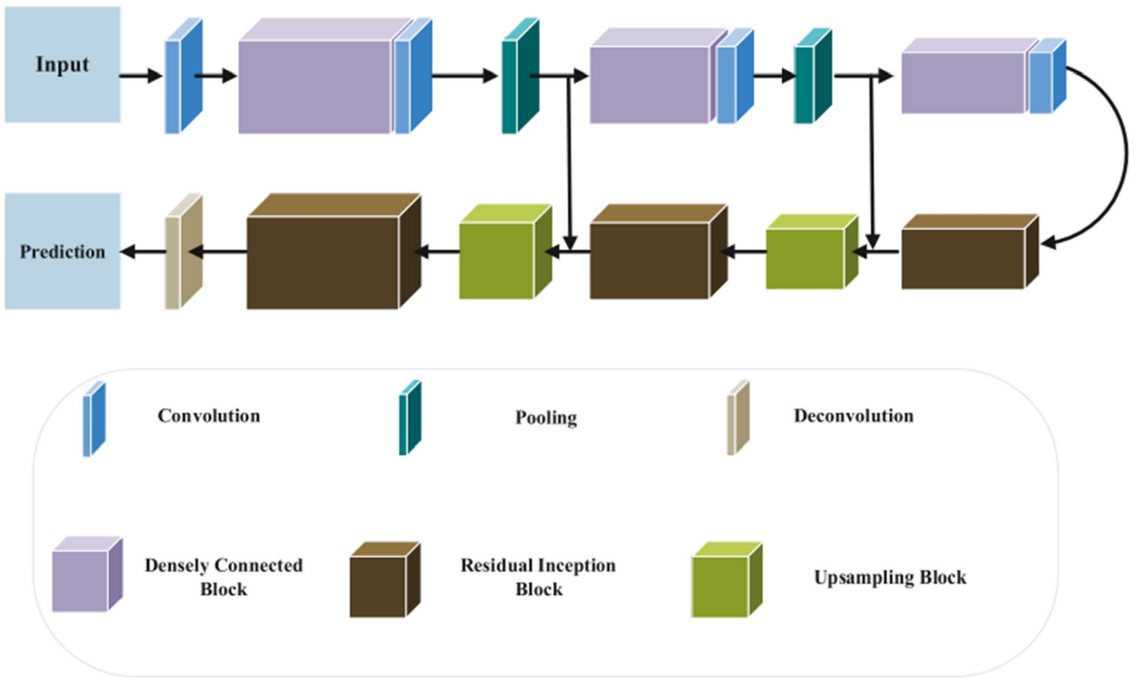


FIGURE 10
Segmentation of 3D MRI brain images using a suggested network design. In the suggested approach, DenseNet and Inception-ResNet are used concurrently. Reprinted with permission from Elsevier, Copyright © 2020 Elsevier (52).

### 4.4.3 Generative adversarial networks

A network known as a "generative adversarial network" (GAN) is made up of two networks: a generator (G) that creates a false image from a noise vector and a discriminator (D) that determines the difference between produced and real data (56). It is advised to use a multi-stage Generative Adversarial Network for image segmentation (56). The model creates a rough contour of the background and brain tissues in the first stage. The model then creates a more detailed contour for the white matter, gray matter, and cerebrospinal fluid in the subsequent stage. The performed fusion of the *coarse* and *refined* outliners.

### 4.4.4 UNet architecture

The UNet model is one of the most popular convolution neural networks (CNN) that have been successfully used to medical imaging tasks (38, 52, 53). Convolutional, pooling, and up-sampling layers make up the UNet model (52). An architecture for segmenting the baby brain is shown in Figure 10. The network has two paths: a downsampling encoder path and an upsampling decoder path. Reduced feature map resolution and increased receptive field are the goals of downsampling in the encoder path. The residual inception and upsampling blocks make up the up-sampling procedure in the decoder pipeline. Particularly, local features are found in the shallower layers, whereas global features are found in the deeper layers. For new-born brain segmentation, the concatenation of the several levels of upsampling feature maps enables the capture of multiple contextual information. To classify the concatenated features into the target classes (WG, GM, CSF), a classifier is made up of a Conv ($1 \times 1 \times 1$). The brain probability maps that were produced using the Softmax classifier (52).

On the other hand (53), proposed In order to partition the brain tissues into the three categories of white matter, gray matter, and cerebrospinal fluid, a novel 3D CNN architecture that is based on the U-Net structure is described. The basic idea behind the proposed method is to use residual skip-connections and densely connected convolutional layers, as shown in Figure 11, to reduce the number of parameters in the network, improve gradient flow, and increase representation capacity. In addition, the suggested network is trained using the loss functions, cross-entropy, dice similarity, and a combination of the two.

In addition, Triple Residual Multiscale Fully Convolutional Network, a deep network design based on U-Net, is suggested by Chen et al. (6). The model is composed of encoder and decoder process. Encoder procedure comprises: tradition 2D convolution, max-pooling and residual block while the decoder procedure comprises deconvolution, residuals multiscale block, concatenate block and traditional 2D convolution. Furthermore, APR-Net, a new 3D fully convolutional neural network for segmenting brain tissue, is presented by Zhuang et al. (28). The model is made up of several encoded streams and one decoded stream, three primary components make up APRNet: Multi-modal cross-dimension attention modules, 3D anisotropic pyramidal convolutional reversible residual sequence modules, and the core of the APRNet.

The common evaluation metrics that were applied to the 19 studies that were obtained for this analysis utilizing the PRISMA approach are detailed in the section that follows.

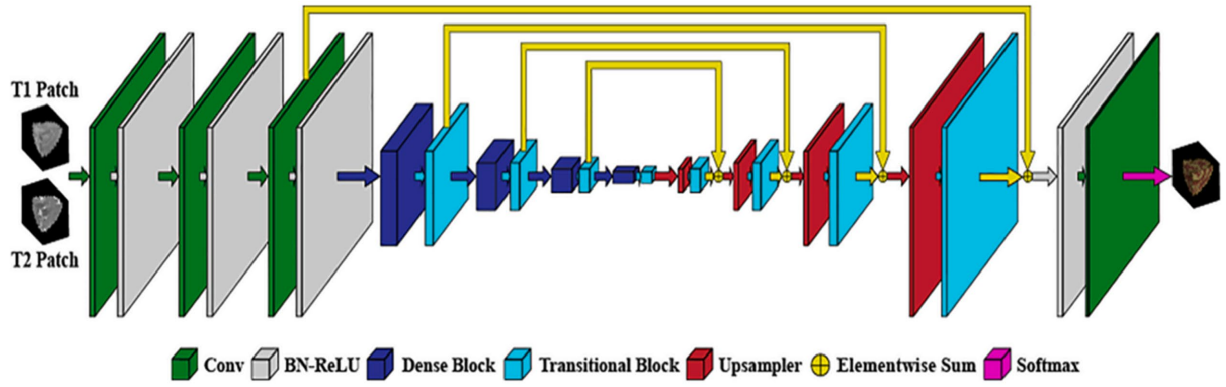


FIGURE 11
Architecture of the proposed network. Reprinted with permission from Elsevier, Copyright © 2021 Elsevier (53).

TABLE 6 A list of evaluation metrics employed by the 19 selected articles using PRISMA approach.

| Authors | Evaluation Metrics | Dataset | DSC | WM | | | GM | | | CSF | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | MHD | ASD | DSC | MHD | ASD | DSC | MHD | ASD |
| (15) | Dice | | 0.94 | | | 0.92 | | | 0.84 | | |
| (20) | Dice | | 0.47 | | | 0.91 | | | 0.75 | | |
| (5) | Dice | | 0.89 | | | 0.87 | | | | | |
| (18) | Dice, MHD | NeoBrain12 | 0.86 | | | 0.88 | | | 0.92 | | |
| (27) | Dice, MHD | | 0.86 | 0.28 | | 0.85 | 0.24 | | 0.83 | 0.44 | |
| (25) | Dice | iSeg2017 | 0.97 | | | 0.90 | | | 0.95 | | |
| (51) | Dice, MHD | NDAR | 0.89 | 0.28 | | 0.90 | 0.24 | | 0.92 | 0.43 | |
| (2) | Dice, MHD | iSeg2017, MRBrainS13 | 0.89 | 1.78 | 6.03 | 0.86 | 1.34 | 6.19 | 0.83 | 2.26 | 7.31 |
| (21) | Dice, MHD, ASD | iSeg2017 | 0.91 | 5.92 | 0.39 | 0.91 | 5.75 | 0.34 | 0.94 | 13.64 | 0.13 |
| (24) | Dice, MHD, ASD | iSeg2017 | 0.90 | 6.88 | 0.39 | 0.92 | 5.63 | 0.31 | 0.96 | 9.00 | 0.11 |
| (34) | Dice, MHD, ASD | iSeg2017 | 0.90 | 7.1 | 0.36 | 0.92 | 9.55 | 0.31 | 0.96 | 8.85 | 0.11 |
| (6) | | iSeg2017 | | | | | | | | | |
| (52) | Dice, MHD, ASD | iSeg2017 | 0.91 | 6.56 | 0.37 | 0.92 | 5.75 | 0.31 | 0.96 | 9.23 | 0.13 |
| (42) | Dice, MHD, ASD | iSeg2017 | 0.90 | 7.45 | 0.41 | 0.92 | 6.06 | 0.34 | 0.96 | 9.13 | 0.12 |
| (53) | Dice, MHD, ASD | iSeg2017, IBSR18 | 0.90 | 6.77 | 0.39 | 0.91 | 5.94 | 0.32 | 0.95 | 9.20 | 0.11 |
| (54) | Dice, MHD, ASD | iSeg2017, iSeg2019, IBSR, BrainWeb | 0.86 | 8.92 | 0.53 | 0.81 | 8.17 | 0.53 | 0.82 | 11.6 | 0.53 |
| (28) | Dice, MHD, ASD | iSeg2017, MRBrainS13 | 0.91 | 6.22 | 0.35 | 0.92 | 6.41 | 0.32 | 0.95 | 9.13 | 0.12 |
| (55) | Dice, MHD, ASD | iSeg2017 | 0.92 | 6.21 | 0.29 | 0.93 | 5.24 | 0.28 | 0.96 | 7.66 | 0.09 |
| (56) | Dice | iSeg2017, MRBrainS13 | 0.88 | | | 0.93 | | | 0.93 | | |

## 5 Evaluation metrics

To assess the accurateness of an automatic segmentation algorithm: Dice Similarity Coefficient (DSC) (58, 59), Modified Hausdorff distance (MHD), where the 95-th percentile of all Euclidean distance is utilized, along with Average Surface Distance (ASD). The initial method computes intensity of overlap amongst the segmented area together with the ground truth, while the additional two techniques estimate the border distances (2, 21).

19 out of 21 of the articles obtained from the PRISMA approach employed one or more of the evaluation metrics (DSC, MHD, and ASD). Table 6 presents a list of all 19 studies and the metrics applied to assess the results of an segmentation algorithm.

In addition, Dice Similarity Coefficient, Modified Hausdorff Distance, Average Surface Distance metrics were also employed by iSeg-2017 organizers to assess the accurateness of the contesting segmentation techniques (16, 17):

To measure the intersection amongst separations, outcome X together with ground truth Y, the Dice Similarity Coefficient is characterised as tails:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \quad (1)$$

where $X$ and $Y$ represent two segmentation labels created physically and computationally, correspondingly, $|X|$ represents the

amount of optimistic portions in the binary segmentation $X$, and $|X \cap Y|$ is the amount of common optimistic elements by $X$ together with Y. A bigger DICE reveals a greater intersection among the physical and projecting division regions. The threshold should not be greater than 1 (16, 17).

Allow $R$ along with $S$ be the series of voxels within the physical and predicative segmentation limit, correspondingly. A modified Hausdorff distance (MHD) is described as follows:

$$MHD(R,S) = \max\{h(R,S), h(S,R)\} \quad (2)$$

where $h(R,S) = \frac{1}{N_c}\sum_{r\in r} d(r,S)$ and $d(r,S) = \min_{r\in R} \| r - s \|$ with $\| . \|$ representing the Euclidean distance. A lesser MHD coefficient implies bigger resemblance between manual and predictive segmentation contours (7, 60). The maximum MDH from set X to set Y is a max function defined as 95%.

The third computation metric is the average surface distance (ASD), termed as:

$$ASD(C,D) = \frac{1}{2}\left( \frac{\sum_{v_i \in S_C} \min_{v_j \in S_D} \|v_i - v_j\|}{\sum_{v_i \in S_C} 1} + \frac{\sum_{v_j \in S_D} \min_{v_i \in S_C} \|v_i - v_j\|}{\sum_{v_j \in S_D} 1} \right) \quad (3)$$

where $S_C$ and $S_D$ signify the outside meshes of C and D, correspondingly. A lesser ASD number implies superior segmentation accuracy (17).

The performance comparison of this study was done using DCS, MHD and ASD, comparing it with previous studies (21, 24, 34, 42, 52). This shows the room of improvement or lack of improvement of our study using different evaluation metrics. The evaluation metric employed are DSC, MHD, and ASD for white matter (WM). The most favourable results of DCS was which was highest was 0.97, achieved by Sanroma et al. (25) followed by Gui et al. (15) which obtained DSC value of 0.94. Other authors have results less than 0.94. Regarding MHD results, the most optimal results were obtained by Luan et al. (54), which identified a value of 8.92 (11) followed and obtained the results of 6.03.

In addition to that; DSC, MHD, and ASD were computed to identify gray Matter (GM). The most accuracy results were obtained for DSC are 0.93 (55, 56), MHD of 9.55 was obtained by Hashemi et al. (34). For ASD (11), achieved a value of 6.19. Furthermore, CSF accuracy was measured using DSC, MHD, and ASD. Pertaining DCS, the most favourable accuracy was 0.96 supported by Hashemi et al. (34), Qamar et al. (24), Qamar et al. (52), Dolz et al. (42), and Ding et al. (55). The most accuracy value of the metric MHD was 13.64 which was supported by Bui et al. (21). The most favourable metric value for ASD was 7.31 which was supported by Dolz et al. (11).

The most promising algorithm is supported by Dolz et al. (11). Their study was produced most accuracy when using WM, GM, and CSF. Interestingly, no strategy had a statistically significant superior performance than all other methods for segmentation of WM, GM, and CSF across any parameter. For example (25), obtained the highest median in terms of DCS for white matter (WM). Nonetheless, the differences between their findings and those of (15) are not statistically significant. Furthermore, Dolz et al. (11) has the highest ASD values for both WM, GM, and CSF, but one of the lowest MDH medians for WM, GM, and CSF. As a result, there is no discernible, statistically significant difference with any other methods.

The following dataset were used by in the 19 studies selected using the PRISMA.

iSeg-2017 dataset is a publicly available to the research community[1] consisting of 10 infant subjects (5 females and 5 male) with manual labels were provided for training and 13 infant subjects (7 females and 6 male) were provided for testing. However, manual labels for testing subjects are not provided (16). In addition, iSeg-2019 challenge was done with the aim of promoting automatic segmentation algorithms on infant brain MRI from multiple sites, MR images from four different sites as training, validation, and testing datasets, respectively are available from https://iseg2019.web.unc.edu/.

Three separate image sets of premature babies are included in the NeoBrainS12 data set: (i) axial scans taken at 40 weeks corrected gestational age; (ii) coronal scans taken at 30 weeks corrected gestational age; and (iii) coronal scans taken at 40 weeks corrected gestational age. At the neonatal critical care unit of the University Medical Center Utrecht in the Netherlands, all scans were performed as part of routine clinical procedures. You can get the remaining photos from the first two sets along with the appropriate manual annotations from the NeoBrainS12 website at http://www.miccai2012.org and use them as training data (61).

MRBrainS13 challenge workshop at the Medical Image Computing and Computer Assisted Intervention (MICCAI) conference provided dataset consisting of 20 subjects (mean age ± SD = 71 ± 4 years, 10 males, 10 female) were selected from an ongoing Computational Intelligence and Neuroscience 3 cohort study of older (65–80 years of age) functionally independent individuals without a history of invalidating stroke or other brain diseases. This dataset is publicly available from http://www.miccai2013.org (62).

Along with magnetic resonance brain image data, the Internet Brain Segmentation Repository (IBSR) offers manually guided expert segmentation results. Its goal is to promote the analysis and advancement of segmentation techniques https://www.nitrc.org/projects/ibsr.

Through data sharing, data harmonization, and the publication of study findings, the National Database for Autism study (NDAR), a research data repository supported by the National Institutes of Health (NIH), seeks to further the understanding of autism spectrum disorders (ASD). In addition, NDAR acts as a platform for the scientific community and a gateway to numerous additional research repositories, enabling data aggregation and secondary analysis. Dataset can be accessed from https://www.re3data.org/repository/r3d100010717

## 6 Findings and limitation of the presented frameworks

The findings of this study and drawback of the concerned frameworks on isointense brain MRI segmentation can be seen in Table 6.

1 http://iseg2017.web.unc.edu

## 6.1 Findings

Deep learning methods are popular in isointense brain MRI segmentation, specifically convolution neural networks. An interesting discovery is that 13 of the 19 studies obtained using PRISMA employed convolution neural networks. In addition, Dice similarity coefficient (DCS) was the most frequently used evaluation metrics, where 17 out of the 19 studies used DCS. Modified Hausdorff Distance (MHD) was also employed in 13 studies out of 19, while Average Surface Distance (ASD) was the least utilized evaluation metrics, where nine studies out of the 19 used it. Furthermore, the most commonly used dataset for training and testing was from MICCAI iSEG-2017 Grand Challenge on 6-month infant brain MRI segmentation as illustrated in Table 6. iSEG-2017 dataset is a publicly available to the research community[2] consisting of 10 infant subjects (5 females and 5 male) with manual labels were provided for training and 13 infant subjects (7 females and 6 male) were provided for testing. However, manual labels for testing subjects are not provided.

## 6.2 Limitation of the presented frameworks

Limitations presented from the assessed frameworks included the omission of ensemble to improve the evaluation metrics. Another studies used Dice similarity coefficient (DCS) and did not compare it with Modified Hausdorff Distance (MHD) and Average Surface Distance (ASD) to provide better results. On the other hand, some authors applied DCS and MHD and did not compare it with ASD to provide better results. Wilcoxon signed-rank test with all-against-all was used to see whether any study performs noticeably better than the others in terms of DCS, MHD, and ASD. Surprisingly, no study was able to partition WM, GM, and CSF across all parameters (DCS, MHD, and ASD) with a substantial statistically significant performance advantage over all other studies. In order to detect the significant difference, ensemble techniques must be employed in conjunction with CNN, and the segmentation error can decrease in order to improve the model. With minimal user interaction, this idea has the potential to deliver expert-level performance.

Most researchers do not focus on improving the accuracy of the model, reducing the amount of Rubik convolutional calculations, and using multi-axis information more efficiently (54). While other avoid image processing due to the lack of datasets (56). Researchers are lacking to integrate different deep fuzzy structures to model data ambiguity and further explore training of deep fuzzy models using incremental and reinforcement learning. In addition, comparison of the research and other study to evaluate performance of proposed architectures using other challenges to take advantage of multi-modal data was lacking in their studies (24). A large amount of researchers have focused on image recognition and classification, there is a lack of CNNs focusing on semantic image segmentation (11). Some emerging research approach such as FCNN minimize redundant convolution results in computation being more efficient. Also few researchers have focused on 3D CNN-ensemble learning strategy used to improve performance (42). To overcome the challenges, single non-linear convolutional can be used. Lastly, this study considered paper published between 1st of January 2012 and 31st of December, 2022.

# 7 Limitation and future work

The limitation of this study come from fact that number of images in iSEG-2017 dataset is not enormous, it consists of only 10 (T1-weighted and T2-weighted MRI) for training and 13 (T1 and T2 MRI) for testing. In addition, the ground truth labels for the test instance are not available. In this study, both T1-weight and T2-weight MRI are studied. In future, only T1-weight or T2-weight MRI will be considered. In addition, accurate segmentation of child brain MRI is extremely difficult than grown-up brain segmentation, because of low tissue differentiate, excessive noise, continuing WM Mylenium, and uncompromising incomplete volume effects which makes tissues to remain miscategorised together with diminishing the exactness of the segmentation algorithm (14, 16, 63).

Most of the CNN models, experiments were performed on computational servers or CPU with a graphic processing unit (GPU) memory. Furthermore, similar article written by same authors were treated as separate paper based on different ideas of contribution (5, 18). Most dataset are already cleaned as secondary dataset, as a result, they contain lots of errors which can be minimized by re-cleaning the dataset. In the future, data augmentation could be applied to possible improve the results, by amplifying the size of the dataset. Furthermore, other evaluation metrics could be utilized such Jaccard index which is also common for the evaluating of image segmentation tasks. The same algorithms selected in this study can be applied to adult brain MRI segmentation.

# 8 Conclusion

This systematic review investigates isointense brain MRI segmentation. An extensive literature search for relevant studies published in the period of 2012 to 2022 and finally identified 19 primary studies that are pertaining to the four research questions (RQs) raised in this review. A summarized research approach of the existing literature along with the research contribution, evaluation metrics, datasets, finding and future recommendations to study isointense brain MRI segmentation models are described. The principle findings of this review are summarized as follows:

- [RQ-1] The detailed review of infant brain MRI segmentation techniques and deep learning techniques has been deliberated in Section 4 and Sub-Section D of Section 4, respectively. The summarized review is examined in Table 5.
- [RQ-2] Section 5 of this study reviews datasets. Table 6 presents the evaluation metrics and the most frequently used dataset for isointense brain MRI segmentation.
- [RQ-3] It has been observed that deep learning techniques are popular in isointense brain MRI segmentation. Thirteen out of the nineteen studies used convolutional neural network and Dice Similarity Coefficient is also the most used evaluation metric from the presented frameworks.
- [RQ-4] Future works and limitations from researcher play a vital role to explore further research in a relevant domain. To answer this RQ, the limitations and future works of deep learning technique and

2 See footnote 1.

evaluation metrics is discussed in Section 6 and 8, respectively. It was found that most studies recommended the use of data augmentation to amplify the size of the dataset, which could possibly improve the results.

## Data availability statement

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

## Author contributions

SM and SV contributed on literature review, SM and SV defined the research problem, SM and SV designed and implemented a framework, SM and SV analysed and computed the results. All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

## Funding

The University of Johannesburg provides the funding under the University Capacity Development Programme (UCDP), and University Staff Development Programme (USDP).

## 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. Bui T.D., Shin J., Moon T. (2017). *3D densely convolutional networks for volumetric segmentation*. arXiv [preprint].

2. Dolz J, Gopinath K, Yuan J, Lombaert H, Desrosiers C, Ben Ayed I. HyperDense-net: a hyper-densely connected CNN for multi-modal image segmentation. *IEEE Trans Med Imaging*. (2019) 38:1116–26. doi: 10.1109/TMI.2018.2878669

3. Lei Z., Qi L., Wei Y., Zhou Y., Qi W. (2019). *Infant brain MRI segmentation with dilated convolution pyramid downsampling and self-attention. arXiv* [preprint]. doi: 10.48550/arXiv.1912.12570

4. Kumar S., Conjeti S., Roy A.G., Wachinger C., Navab N. (2018). "InfiNet: Fully convolutional networks for infant brain MRI segmentation." in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. Washington, DC: IEEE. pp. 145–148. doi: 10.1109/ISBI.2018.8363542

5. Wang L, Shi F, Gao Y, Li G, Gilmore JH, Lin W, et al. Integration of sparse multi-modality representation and anatomical constraint for isointense infant brain MR image segmentation. *NeuroImage*. (2014) 89:152–64. doi: 10.1016/j.neuroimage.2013.11.040

6. Chen Y, Qin Y, Jin Z, Fan Z, Cai M. A triple residual multiscale fully convolutional network model for multimodal infant brain MRI segmentation. *KSII Trans Internet Inform Syst*. (2020) 14:962–75. doi: 10.3837/tiis.2020.03.003

7. Chen H, Dou Q, Yu L, Qin J, Heng P-A. VoxResNet: deep voxelwise residual networks for brain segmentation from 3D MR images. *NeuroImage*. (2018) 170:446–55. doi: 10.1016/j.neuroimage.2017.04.041

8. Moeskops P, Viergever MA, Mendrik AM, de Vries LS, Benders MJNL, Isgum I. Automatic segmentation of MR brain images with a convolutional neural network. *IEEE Trans Med Imaging*. (2016) 35:1252–61. doi: 10.1109/TMI.2016.2548501

9. Wang L, Shi F, Li G, Gao Y, Lin W, Gilmore JH, et al. Segmentation of neonatal brain MR images using patch-driven level sets. *NeuroImage*. (2014) 84:141–58. doi: 10.1016/j.neuroimage.2013.08.008

10. Devi CN, Chandrasekharan A, Sundararaman VK, Alex ZC. Neonatal brain MRI segmentation: a review. *Comput Biol Med*. (2015) 64:163–78. doi: 10.1016/j.compbiomed.2015.06.016

11. Dolz J., Ayed I.B., Yuan J., Desrosiers C. *Isointense infant brain segmentation with a hyper-dense connected convolutional neural network*. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). (2019). 616–20. doi: 10.1109/ISBI.2018.8363651

12. Chen J, Sun Y, Fang Z, Lin W, Li G, Wang L. Harmonized neonatal brain MR image segmentation model for cross-site datasets. *Biomed Signal Process Cont*. (2021) 69:102810. doi: 10.1016/j.bspc.2021.102810

13. Wu J, Tang X. Brain segmentation based on multi-atlas and diffeomorphism guided 3D fully convolutional network ensembles. *Pattern Recogn*. (2021) 115:107904. doi: 10.1016/j.patcog.2021.107904

14. Weisenfeld NI, Warfield SK. Automatic segmentation of newborn brain MRI. *NeuroImage*. (2009) 47:564–72. doi: 10.1016/j.neuroimage.2009.04.068

15. Gui L, Lisowski R, Faundez T, Hüppi PS, Lazeyras F, Kocher M. Morphology-driven automatic segmentation of MR images of the neonatal brain. *Med Image Anal*. (2012) 16:1565–79. doi: 10.1016/j.media.2012.07.006

16. Wang L, Nie D, Li G, Puybareau E, Dolz J, Zhang Q, et al. Benchmark on automatic six-month-old infant brain segmentation algorithms: the iSeg-2017 challenge. *IEEE Trans Med Imaging*. (2019) 38:2219–30. doi: 10.1109/TMI.2019.2901712

17. Sun Y, Gao K, Wu Z, Li G, Zong X, Lei Z, et al. Multi-site infant brain segmentation algorithms: the iSeg-2019 challenge. *IEEE Trans Med Imaging*. (2021) 40:1363–76. doi: 10.1109/TMI.2021.3055428

18. Wang L, Gao Y, Shi F, Li G, Gilmore JH, Lin W, et al. LINKS: learning-based multi-source IntegratioN frameworK for segmentation of infant brain images. *NeuroImage*. (2015) 108:160–72. doi: 10.1016/j.neuroimage.2014.12.042

19. Alghamdi NS, Taher F, Kandil H, Sharafeldeen A, Elnakib A, Soliman A, et al. Segmentation of infant brain using nonnegative matrix factorization. *Appl Sci*. (2022) 12:5377. doi: 10.3390/app12115377

20. Anbeek P, Išgum I, van Kooij BJM, Mol CP, Kersbergen KJ, Groenendaal F, et al. Automatic segmentation of eight tissue classes in neonatal brain MRI. *PLoS One*. (2013) 8:e81895. doi: 10.1371/journal.pone.0081895

21. Bui TD, Shin J, Moon T. Skip-connected 3D dense net for volumetric infant brain MRI segmentation. *Biomed Signal Process Cont*. (2019) 54:101613. doi: 10.1016/j.bspc.2019.101613

22. Moeskops P., Pluim J.P.W. (2017). *Isointense infant brain MRI segmentation with a dilated convolutional neural network. arXiv* [preprint]. doi: 10.48550/arXiv.1708.02757

23. Prastawa M, Gilmore JH, Lin W, Gerig G. Automatic segmentation of MR images of the developing newborn brain. *Med Image Anal*. (2005) 9:457–66. doi: 10.1016/j.media.2005.05.007

24. Qamar S, Jin H, Zheng R, Ahmad P. Multi stream 3D hyper-densely connected network for multi modality isointense infant brain MRI segmentation. *Multimed Tools Appl*. (2019) 78:25807–28. doi: 10.1007/s11042-019-07829-1

25. Sanroma G, Benkarim OM, Piella G, Lekadir K, Hahner N, Eixarch E, et al. Learning to combine complementary segmentation methods for fetal and 6-month infant brain MRI segmentation. *Comput Med Imaging Graph*. (2018) 69:52–9. doi: 10.1016/j.compmedimag.2018.08.007

26. Weisenfeld N.I., Mewes A.U.J., Warfield S.K. (2006). "Segmentation of Newborn Brain MRI." in *3rd IEEE International Symposium on Biomedical Imaging: Macro to Nano, 3rd IEEE International Symposium on Biomedical Imaging: Macro to Nano, 2006*. Arlington, Virginia, USA: IEEE. pp. 766–769. doi: 10.1109/ISBI.2006.1625029

27. Zhang W, Li R, Deng H, Wang L, Lin W, Ji S, et al. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *NeuroImage*. (2015) 108:214–24. doi: 10.1016/j.neuroimage.2014.12.061

28. Zhuang Y, Liu H, Song E, Ma G, Xu X, Hung C-C. APRNet: a 3D anisotropic pyramidal reversible network with multi-modal cross-dimension attention for brain tissue segmentation in MR images. *IEEE J Biomed Health Inform*. (2022) 26:749–61. doi: 10.1109/JBHI.2021.3093932

29. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-net: learning dense volumetric segmentation from sparse annotation In: S Ourselin, L Joskowicz, MR Sabuncu, G Unal and W Wells, editors. *Medical image computing and computer-assisted intervention – MICCAI 2016. Lecture notes in computer science*. Cham: Springer International Publishing (2016). 424–32. doi: 10.1007/978-3-319-46723-8_49

30. Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal*. (2017) 36:61–78. doi: 10.1016/j.media.2016.10.004

31. Lee B, Yamanakkanavar N, Choi JY. Automatic segmentation of brain MRI using a novel patch-wise U-net deep architecture. *PLoS One*. (2020) 15:e0236493. doi: 10.1371/journal.pone.0236493

32. Milletari F., Navab N., Ahmadi S.-A. (2016). *V-net: Fully convolutional neural networks for volumetric medical image segmentation*. In 2016 fourth international conference on 3D vision (3DV), 565–71. doi: 10.1109/3DV.2016.79

33. De Brebisson A., Montana G. (2015). *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Boston, MA, USA: IEEE. pp. 20–28.

34. Hashemi S.R., Prabhu S.P., Warfield S.K., Gholipour A. (2019). *Exclusive independent probability estimation using deep 3D fully convolutional DenseNets: Application to IsoIntense infant brain MRI segmentation. Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning, PMLR 102:260-272*. Available at: https://proceedings.mlr.press/v102/hashemi19a.html

35. Milletari F, Ahmadi S-A, Kroll C, Plate A, Rozanski V, Maiostre J, et al. Hough-CNN: deep learning for segmentation of deep brain regions in MRI and ultrasound. *Comput Vis Image Underst*. (2017) 164:92–102. doi: 10.1016/j.cviu.2017.04.002

36. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation In: N Navab, J Hornegger, WM Wells and AF Frangi, editors. *Medical image computing and computer-assisted intervention – MICCAI 2015. Lecture notes in computer science*. Cham: Springer International Publishing (2015). 234–41. doi: 10.1007/978-3-319-24574-4_28

37. Dolz J., Ayed I.B., Yuan J., Desrosiers C. (2018). *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. Washington, DC: IEEE. pp. 616–620. doi: 10.1109/ISBI.2018.8363651

38. Fonov V.S., Doyle A., Evans A.C., Collins D.L. (2018). NeuroMTL iSEG challenge methods. *bioRxiv* [preprints]. doi: 10.1101/278465

39. Khalili N., Lessmann N., Turk E., Claessens N., Heus R.De, Kolk T., et al. (2019) Automatic brain tissue segmentation in fetal MRI using convolutional neural networks. *Magn Reson Imaging* 64, 77–89. doi: 10.1016/j.mri.2019.05.020

40. Nie D., Wang L., Gao Y., Shen D. (2016). *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. Prague: IEEE. pp. 1342–1345. doi: 10.1109/ISBI.2016.7493515

41. Zeng G., Zheng G. (2018). "Multi-stream 3D FCN with multi-scale deep supervision for multi-modality isointense infant brain MR image segmentation." in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. Washington, DC: IEEE. pp. 136–140. doi: 10.1109/ISBI.2018.8363540

42. Dolz J, Desrosiers C, Wang L, Yuan J, Shen D, Ben Ayed I. Deep CNN ensembles and suggestive annotations for infant brain MRI segmentation. *Comput Med Imaging Graph*. (2020) 79:101660. doi: 10.1016/j.compmedimag.2019.101660

43. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and Meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med*. (2009) 6:e1000100. doi: 10.1371/journal.pmed.1000100

44. Snyder H. Literature review as a research methodology: an overview and guidelines. *J Bus Res*. (2019) 104:333–9. doi: 10.1016/j.jbusres.2019.07.039

45. Ismail SN, Ramli A, Aziz HA. Influencing factors on safety culture in mining industry: a systematic literature review approach. *Res Policy*. (2021) 74:102250. doi: 10.1016/j.resourpol.2021.102250

46. Tavse S, Varadarajan V, Bachute M, Gite S, Kotecha K. A systematic literature review on applications of GAN-synthesized images for brain MRI. *Fut Internet*. (2022) 14:351. doi: 10.3390/fi14120351

47. Sharma R, De Sousa L, Jabbour AB, Jain V, Shishodia A. The role of digital technologies to unleash a green recovery: pathways and pitfalls to achieve the European green Deal. *J Enterp Inf Manag*. (2022) 35:266–94. doi: 10.1108/JEIM-07-2021-0293

48. Ahsan MM, Siddique Z. Machine learning-based heart disease diagnosis: a systematic literature review. *Artif Intell Med*. (2022) 128:102289. doi: 10.1016/j.artmed.2022.102289

49. Kitchenham BA, Mendes E, Travassos GH. Cross versus within-company cost estimation studies: a systematic review. *IEEE Trans Softw Eng*. (2007) 33:316–29. doi: 10.1109/TSE.2007.1001

50. Usman M., Mendes E., Weidt F., Britto R. (2014). "Effort estimation in agile software development: a systematic literature review." in *Proceedings of the 10th International Conference on Predictive Models in Software Engineering. PROMISE '14: The 10th International Conference on Predictive Models in Software Engineering*. Turin Italy: ACM. pp. 82–91. doi: 10.1145/2639490.2639503

51. Wang L, Li G, Adeli E, Liu M, Wu Z, Meng Y, et al. Anatomy-guided joint tissue segmentation and topological correction for 6-month infant brain MRI with risk of autism. *Hum Brain Mapp*. (2018) 39:2609–23. doi: 10.1002/hbm.24027

52. Qamar S, Jin H, Zheng R, Ahmad P, Usama M. A variant form of 3D-UNet for infant brain segmentation. *Futur Gener Comput Syst*. (2020) 108:613–23. doi: 10.1016/j.future.2019.11.021

53. Basnet R, Ahmad MO, Swamy MNS. A deep dense residual network with reduced parameters for volumetric brain tissue segmentation from MR images. *Biomed Signal Process Cont*. (2021) 70:103063. doi: 10.1016/j.bspc.2021.103063

54. Luan X, Zheng X, Li W, Liu L, Shu Y, Guo Y. Rubik-net: learning spatial information via rotation-driven convolutions for brain segmentation. *IEEE J Biomed Health Inform*. (2022) 26:289–300. doi: 10.1109/JBHI.2021.3095846

55. Ding W, Abdel-Basset M, Hawash H, Pedrycz W. Multimodal infant brain segmentation by fuzzy-informed deep learning. *IEEE Trans Fuzzy Syst*. (2022) 30:1088–101. doi: 10.1109/TFUZZ.2021.3052461

56. Khaled A, Han J-J, Ghaleb TA. Multi-model medical image segmentation using multi-stage generative adversarial networks. *IEEE Access*. (2022) 10:28590–9. doi: 10.1109/ACCESS.2022.3158342

57. Makropoulos A, Gousias IS, Ledig C, Aljabar P, Serag A, Hajnal JV, et al. Automatic whole brain MRI segmentation of the developing neonatal brain. *IEEE Trans Med Imaging*. (2014) 33:1818–31. doi: 10.1109/TMI.2014.2322280

58. Dice LR. Measures of the amount of ecologic association between species. *Ecology*. (1945) 26:297–302. doi: 10.2307/1932409

59. Moeskops P, Benders MJNL, Chiţă SM, Kersbergen KJ, Groenendaal F, de Vries LS, et al. Automatic segmentation of MR brain images of preterm infants using supervised classification. *NeuroImage*. (2015) 118:628–41. doi: 10.1016/j.neuroimage.2015.06.007

60. Çelik G. *iSeg-WNet: Volumetric segmentation of infant brain MRI images*. Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi. (2022). 38:508–18. Available at: https://dergipark.org.tr/en/pub/erciyesfen/issue/74713/1099510

61. Išgum I, Benders MJNL, Avants B, Cardoso MJ, Counsell SJ, Gomez EF, et al. Evaluation of automatic neonatal brain segmentation algorithms: the NeoBrainS12 challenge. *Med Image Anal*. (2015) 20:135–51. doi: 10.1016/j.media.2014.11.001

62. Mendrik AM, Vincken KL, Kuijf HJ, Breeuwer M, Bouvy WH, De Bresser J, et al. MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans. *Comput Intell Neurosci*. (2015) 2015:1–16. doi: 10.1155/2015/813696

63. Xue H, Srinivasan L, Jiang S, Rutherford M, Edwards AD, Rueckert D, et al. "Automatic segmentation and reconstruction of the cortex from neonatal MRI,"in *Neuroimage*. (2007) 38:461–77. doi: 10.1016/j.neuroimage.2007.07.030

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 03 April 2024
DOI 10.3389/fmed.2024.1354070



OPEN ACCESS

EDITED BY
Gongning Luo,
Harbin Institute of Technology, China

REVIEWED BY
Zhongyi Han,
King Abdullah University of Science and Technology, Saudi Arabia
Le Zhang,
University of Oxford, United Kingdom

*CORRESPONDENCE
Jinming Duan
✉ j.duan@bham.ac.uk

RECEIVED 11 December 2023
ACCEPTED 18 March 2024
PUBLISHED 03 April 2024

CITATION
Zhang Y, Liu B, Bunting KV, Brind D, Thorley A, Karwath A, Lu W, Zhou D, Wang X, Mobley AR, Tica O, Gkoutos GV, Kotecha D and Duan J (2024) Development of automated neural network prediction for echocardiographic left ventricular ejection fraction.
*Front. Med.* 11:1354070.
doi: 10.3389/fmed.2024.1354070



# Development of automated neural network prediction for echocardiographic left ventricular ejection fraction

Yuting Zhang[1], Boyang Liu[2], Karina V. Bunting[3,4], David Brind[5,6,7], Alexander Thorley[1], Andreas Karwath[5,7], Wenqi Lu[8], Diwei Zhou[9], Xiaoxia Wang[3,4,6], Alastair R. Mobley[3,4], Otilia Tica[3], Georgios V. Gkoutos[5,6,7], Dipak Kotecha[3,4,6] and Jinming Duan[1]* on behalf of the card*AI*c group

[1]School of Computer Science, University of Birmingham, Edgbaston, Birmingham, United Kingdom, [2]Manchester University NHS Foundation Trust, Manchester, United Kingdom, [3]Institute of Cardiovascular Sciences, University of Birmingham, Edgbaston, Birmingham, United Kingdom, [4]NIHR Birmingham Biomedical Research Centre and West Midlands NHS Secure Data Environment, University Hospitals Birmingham NHS Foundation Trust, Birmingham, United Kingdom, [5]Institute of Cancer and Genomic Sciences, University of Birmingham, Edgbaston, Birmingham, United Kingdom, [6]Health Data Research UK Midlands, University Hospitals Birmingham NHS Foundation Trust, Birmingham, United Kingdom, [7]Centre for Health Data Science, University of Birmingham, Edgbaston, Birmingham, United Kingdom, [8]Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, United Kingdom, [9]Department of Mathematical Sciences, Loughborough University, Loughborough, United Kingdom

**Introduction:** The echocardiographic measurement of left ventricular ejection fraction (LVEF) is fundamental to the diagnosis and classification of patients with heart failure (HF).

**Methods:** This paper aimed to quantify LVEF automatically and accurately with the proposed pipeline method based on deep neural networks and ensemble learning. Within the pipeline, an Atrous Convolutional Neural Network (ACNN) was first trained to segment the left ventricle (LV), before employing the area-length formulation based on the ellipsoid single-plane model to calculate LVEF values. This formulation required inputs of LV area, derived from segmentation using an improved Jeffrey's method, as well as LV length, derived from a novel ensemble learning model. To further improve the pipeline's accuracy, an automated peak detection algorithm was used to identify end-diastolic and end-systolic frames, avoiding issues with human error. Subsequently, single-beat LVEF values were averaged across all cardiac cycles to obtain the final LVEF.

**Results:** This method was developed and internally validated in an open-source dataset containing 10,030 echocardiograms. The Pearson's correlation coefficient was 0.83 for LVEF prediction compared to expert human analysis ($p < 0.001$), with a subsequent area under the receiver operator curve (AUROC) of 0.98 (95% confidence interval 0.97 to 0.99) for categorisation of HF with reduced ejection (HFrEF; LVEF<40%). In an external dataset with 200 echocardiograms, this method achieved an AUC of 0.90 (95% confidence interval 0.88 to 0.91) for HFrEF assessment.

**Conclusion:** The automated neural network-based calculation of LVEF is comparable to expert clinicians performing time-consuming, frame-by-frame manual evaluations of cardiac systolic function.

KEYWORDS
artificial intelligence, echocardiogram, ejection fraction, heart failure, atrial fibrillation

# 1 Introduction

Heart failure (HF) is a common and increasingly prevalent condition that results in profound burdens on patients, healthcare services, and society (1). It is not a single pathological diagnosis but rather a clinical syndrome consisting of cardinal symptoms, typical signs on clinical examination, and evidence of impairment of either systolic or diastolic function on cardiac imaging (2). HF is divided into distinct phenotypes based primarily on the measurement of systolic left ventricular ejection fraction (LVEF): HF with reduced LVEF (HFrEF, LVEF<40%); HF with mildly reduced ejection fraction (HFmrEF, LVEF 40–49%); and HF with preserved ejection fraction (HFpEF, LVEF >= 50%) (2, 3).

Echocardiography is one of the most widely used diagnostic techniques in cardiology and is the first-line imaging modality for suspected cardiac pathology due to its availability and portability. The standard method to quantify LVEF using echocardiography as per recommendations from the American Society of Echocardiography (ASE) and the European Association of Cardiovascular Imaging (EACVI) is to first calculate left ventricular end-diastolic volumes (LVEDVs) and end-systolic volumes (LVESVs) using Simpson's biplane method of multiple discs (4, 5). Practically, this method requires sonographers or cardiologists to visually identify LVED and LVES frames from a given cine video, which is both time-consuming and prone to error. There is significant intra- and inter-observer variability in LVEF quantification as a result of poor image quality (the endocardial border is often not well seen) and variable cardiac cycle lengths, for example, due to arrhythmias such as atrial fibrillation (AF) (6, 7). To ensure reproducible measurements of LVEF are obtained, it is recommended to average three cardiac cycles for patients in sinus rhythm and 5 to 10 cardiac cycles in AF. These recommendations require substantial training, are rarely followed in clinical practice, and are based on consensus opinion only; the available data show that even best practice is time-consuming and poorly reproducible (4, 8).

To make the calculation of LVEF more efficient and accurate, this paper makes four novel contributions: (1) proposing a new pipeline method to provide comprehensive, transparent details on the calculation of LVEF, which might be more acceptable to clinicians and cardiologists (9); (2) following the recommendation by the ASE and EACVI to average LVEF values across all automatically identified cycles for each apical 4 chamber (A4C) echocardiogram; (3) visualising the LV across the full cardiac cycle in a given echocardiogram, which is useful as an instantaneous summary of beat-to-beat volumetric differences, including the impact of arrhythmias such as AF (10, 11); and (4) the capacity to predict highly accurate LVEF values at scale without relying on manual approaches that have high workforce requirements.

# 2 Methods

This project used an overall framework of transparency, as developed by the card*AI*c group (Application of Artificial Intelligence to Routine Healthcare Data to Benefit Patients with Cardiovascular Disease) and the BigData@Heart Consortium (1). Reporting follows the DECIDE-AI approach for clinical evaluation of decision support systems driven by artificial intelligence (see supplementary file for DECIDE-AI checklist) (12, 13).

## 2.1 Datasets

Two open datasets were used in this project, and both of them have obtained ethical approval (13, 14). One is the Stanford dataset with 10,030 A4C 2D grey-scale echocardiogram videos, each of which represented a unique individual who underwent echocardiogram between 2006 and 2018 as part of clinical care; another one is the CAMUS dataset with 450 A4C view sequences, acquired with different ultrasound scanners at the University Hospital of St Etienne (France). For both datasets, labels for each video included the location of the left ventricle (LV) endocardium (Figures 1A,D), LVEF, LVESV, and LVEDV, which were given by cardiologist experts in the standard clinical workflow. Note that the estimation of LV ejection fraction values was based on Simpson's biplane method of discs. For the Stanford dataset, the LV endocardium in ED or ES frames was marked with 42 coordinates, as shown in Figure 1A. More details were supplied in Appendix B of the Supplementary file.

## 2.2 AI system

### 2.2.1 Methodology

In this article, the proposed pipeline consisted of three steps to assess patients with HFrEF using their corresponding echocardiogram cine in the A4C view (Figure 2A). First, an atrous convolutional neural network (ACNN) was used to segment the LV in each frame of a given video. Based on the segmentation mask, information as shown in Figure 1C, including LV area, LV width, and LV height, was extracted. In addition to segmentation, all ED and ES frames were identified in each video for further beat-to-beat analysis. Second, with the results computed from step 1, an ensemble learning model was developed to predict the LV length, which was then combined with the LV area to compute LV volumes at ED and ES frames. Based on these LV volumes, the final LVEF was computed (see formulas below). Next, whether a patient has HFrEF was determined based on the LVEF value from Step 2, defined as LVEF <40% (2, 3). In addition, a beat-to-beat visualiser was provided based on segmentation results to provide an instantaneous summary of beat-to-beat volumetric differences as a result of the heart rhythm.

### 2.2.2 Inputs and outputs

The segmentation model required frames or arrays as input, as shown in Figure 2B, with a size of 112×112. Therefore, data preprocessing was carried out before training the pipeline, as described in Appendix B. This pipeline could generate two kinds of outputs, as shown in Figure 2C. One was the segmentation results, which would be displayed in video format for cardiologists to visualise

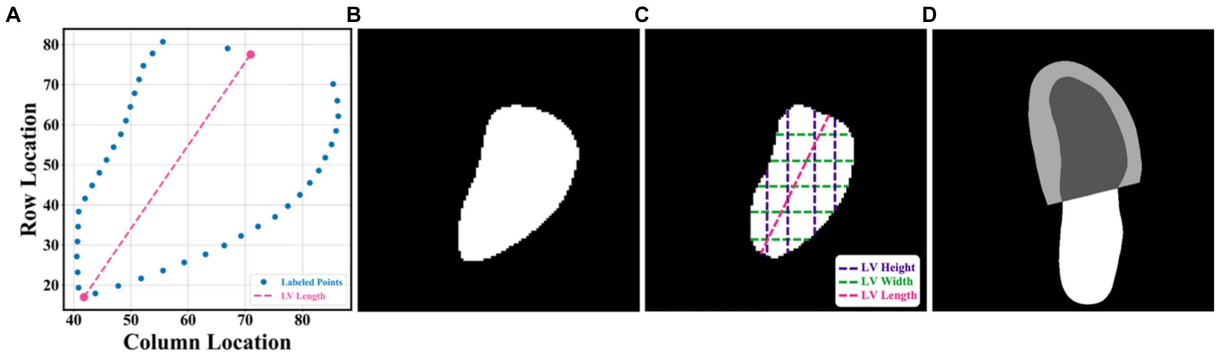


FIGURE 1
**(A–C)** were from the Stanford dataset; **(D)** the CAMUS dataset. **(A)** human-labelled coordinate points in one frame. A Euclidean distance between two pink points was the LV length; **(B)** mask generated from these coordinate points, which was used for training our segmentation network; **(C)** LV area, LV widths, LV heights, and LV length; and **(D)** annotations included information including the left ventricle endocardium, the left ventricle myocardium, and the left atrium.

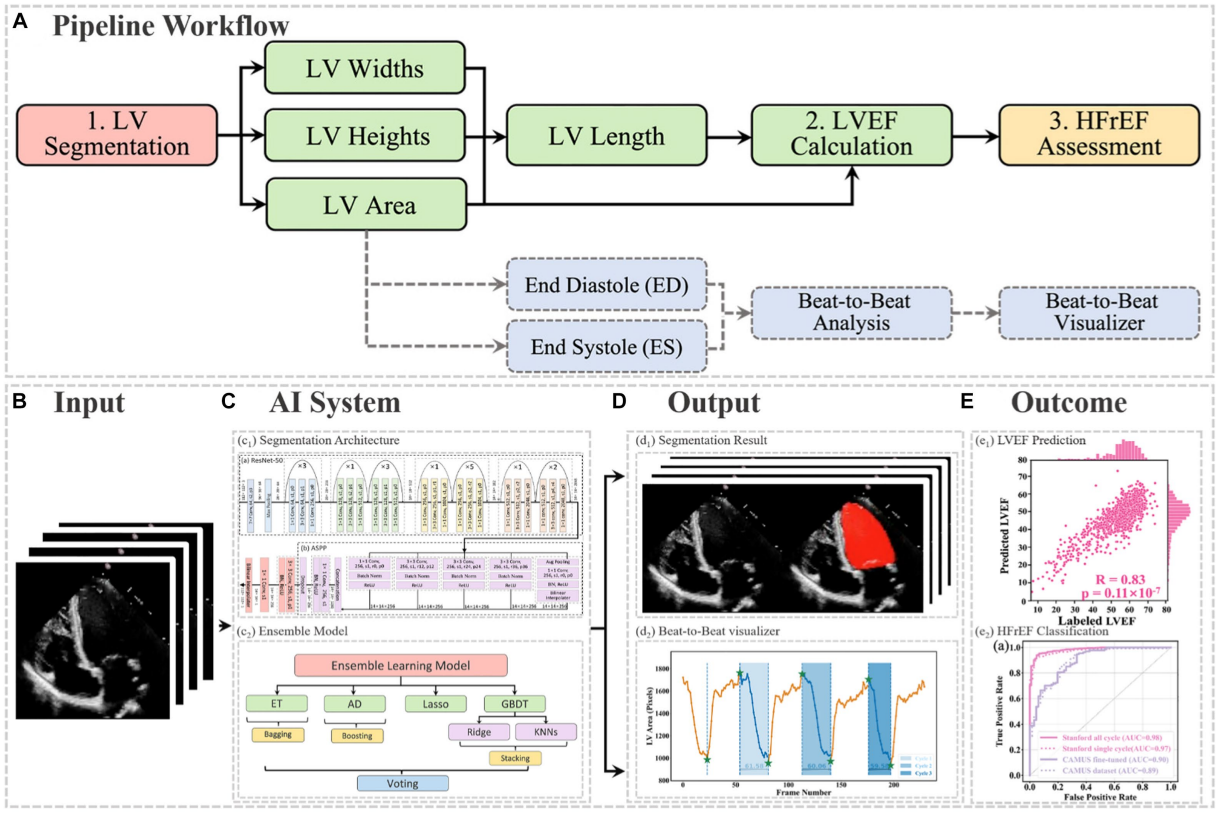


FIGURE 2
**(A)** Flowchart of the pipeline. There were three main steps, including LV segmentation, LVEF calculation, and HFrEF assessment. The area information from segmentation could also be used for ED and ES identification, beat-to-beat analysis of the heart, as well as visualising changes in volume (for example, due to an arrhythmia such as atrial fibrillation). ED = end diastole; ES = end systole; HFrEF = heart failure with reduced LVEF; LV = left ventricle; LVEF = left ventricular ejection fraction. **(B)** Input of the pipeline. **(C)** Proposed AI system. **(D)** Output information, including the segmentation result and the beat-to-beat visualiser. The calculated LVEF values are presented in this visualiser, along with the results of the HF phenotype classification. **(E)** outcome.

LV areas across the full cardiac cycle in a given echocardiogram. Another one was the beat-to-beat visualiser, which could be used to visualise the heartbeats and present the LVEF values for each cardiac cycle, along with their average for all cycles. Moreover, based on the LVEF from the all-cycle method, the result of the HF phenotype classification was presented in the visualiser.

## 2.3 Implementation

In the proposed pipeline, the ellipsoid single-plane model (area-length method) was used to calculate LVEF (15), which was defined in Eq. 1.

$$v = \frac{8}{3\pi} \times \frac{A^2}{L} \quad (1)$$

In Eq. 1, A denoted the LV area, L represented the LV length (the distance from the apex to the midpoint of the annular plane), and V stood for the volume of LV. With this equation, it was possible to compute the end-diastolic volume (EDV) and end-systolic volume (ESV) of the LV, based on which LVEF is calculated as follows:

$$LVEF = \frac{1}{N}\sum_{i=1}^{N} \frac{ESV_i - EDV_i}{EDV_i} \times 100\% \quad (2)$$

Note that information from all cardiac cycles was used and that N here was the available number of cardiac cycles in a video.

### 2.3.1 LV area

In this project, a segmentation network, shown in Figure 3, was used to detect the LV contours first, and then LV areas at ED or ES phases were computed fairly easily by counting the number of pixels from a corresponding binary mask predicted from the trained segmentation model. The proposed network combined ResNet-50, atrous convolutions, and atrous spatial pyramid pooling (ASPP) to extract feature maps and capture long-range context information in the image (16, 17). It was trained first on the training set of the Stanford dataset, and the built-in hyperparameters were tuned on its validation set. After the network had been trained, it was directly deployed to segment all frames in each video in the test set of the Stanford dataset and then to present the trained model performance by calculating the DSC between predicted masks and labelled masks at given ED and ES only. In addition, this trained model was fine-tuned in the training and validation sets of the CAMUS dataset and evaluated in its testing dataset. More details about the architecture, settings, and training procedure of the model are provided in Appendix C.

### 2.3.2 LV length

LV length was defined as the Euclidean distance from the midpoint of the annular plane to the apex in the apical four-chamber view (18). Given that there is a correlation between the width, the height, and the area of the polygon (representing the LV shape), as shown in Figure 1C, a regression model based on ensemble learning (Figure 4) was developed to predict the LV length, which consists of four base regression models including Extra Trees (ETs) (19), Adaboosting (AD) (20), Lasso (21), and a stacking algorithm combining Ridge (22), K-nearest neighbours (KNNs) (23), and Gradient Boosting Decision Tree (GBDT) (24). This ensemble model was trained using the validation set of the Stanford dataset, and its accuracy was reported on both the validation and test sets of the Stanford dataset. The k-fold cross-validation (25) and the $R^2$ score (26) were used to evaluate the proposed model compared with other regression models. The analysis of variance (ANOVA) test was conducted to prove a significant difference between the proposed model and other comparative models (27). In addition, Pearson's correlation coefficient ($r_{corr}$) and *p*-value were used to show the trained model's performance on the test set of the Stanford dataset (28). More details are supplied in Appendix D.

### 2.3.3 ED and ES identification

To detect all ED and ES phases in a given video, the peak detection algorithm was used, taking as input the LV areas across all cardiac cycles in the video. The frame with the biggest LV area represents the ED phase, whilst the frame with the smallest LV area represents the

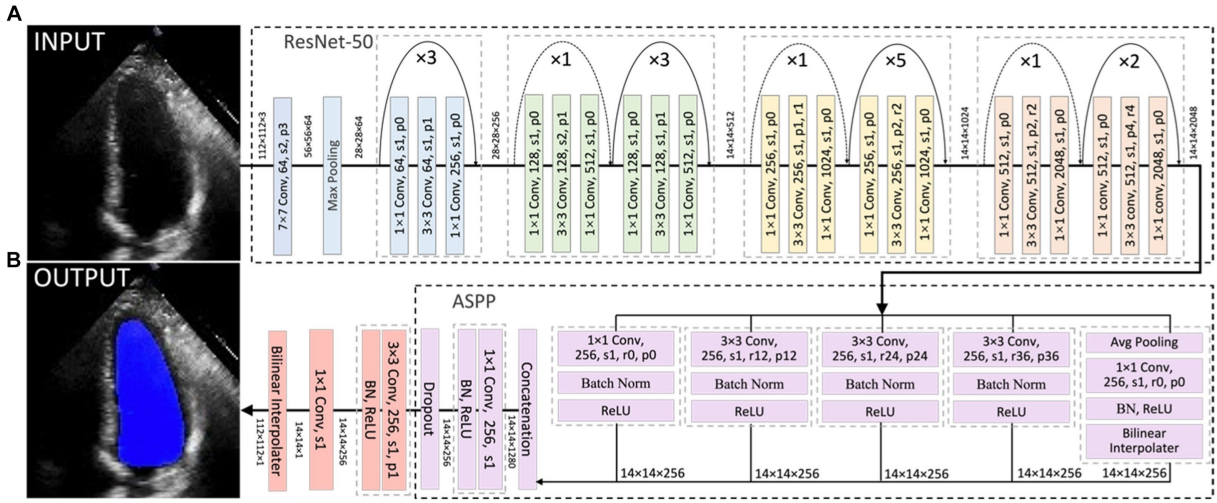


FIGURE 3
Overall segmentation architecture. The segmentation network combined ResNet-50 **(A)**, atrous convolutions, and atrous spatial pyramid pooling (ASPP) **(B)** to resample features at different scales and to capture multi-scale information. As an example, p0, r2, and s1 in the figure denote padding = 0, atrous convolution with rate = 2, and stride = 1, respectively.

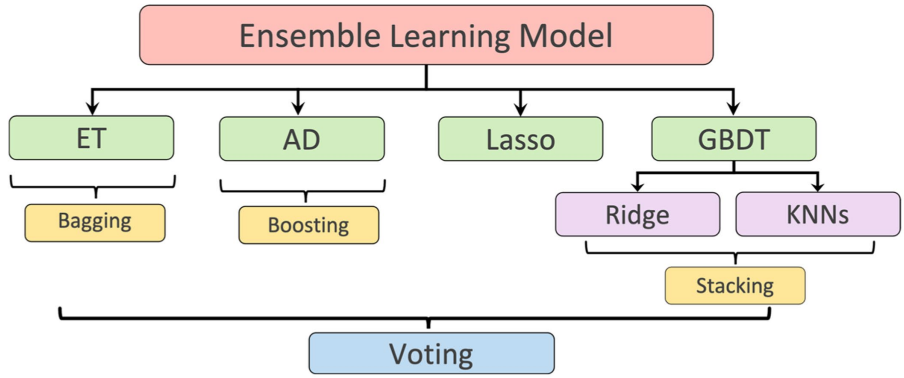


FIGURE 4
Ensemble learning model: including Extra Tree (ET), AdaBoosting (AD), Lasso, and a stacking algorithm combining Ridge, K-nearest neighbours (KNNs), and Gradient Boosting Decision Tree (GBDT). The predicted LV lengths from these regressors were finally ensembled by a voting mechanism.

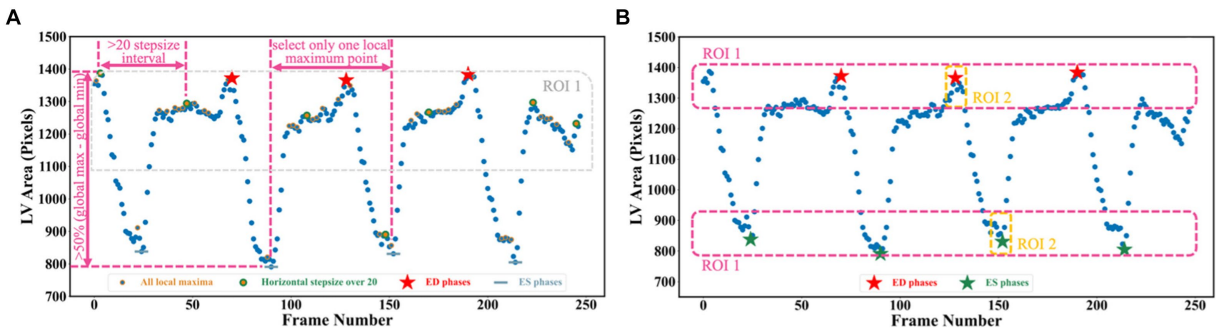


FIGURE 5
**(A)** Three scenarios are used for selecting true peaks, which are identified as ED and ES phases. **(B)** Improved Jeffrey's method used to fine-tune LV areas computed from segmentation. Here, three parts were averaged to compute the final LV areas at ED or ES.

ES phase. For each echocardiogram video, there are often multiple cardiac cycles. In order to identify all cardiac cycles, two parameters were defined for this algorithm. The first one was the horizontal stepsize, which was set to 20 to ensure the effective capture of all cardiac cycles (Figure 5A). Another parameter was the prominence value, which was set to be higher than 50% of the global maximum minus the global minimum to assume the true peaks were located within half of the range between the maximum and minimum values (ROI 1 in Figure 5A). Appendix E of the Supplementary file explains the parameter settings.

## 2.4 Outcomes

The main objective of this project was to determine LVEF, which is a measurement of LV systolic function utilised for HF phenotype classification. As a secondary outcome, this project conducted a classification task based on LVEF <40%, as previously calculated, using all cardiac cycles to detect HFrEF samples from the test sets of both the Stanford and CAMUS datasets (2, 3). In addition, with the computed LV areas and the identified ED as well as ES phases in Section 3.3, the beat-to-beat visualiser could be plotted with a 1D curve, where on the vertical axis it showed LV areas whilst on the horizontal axis it displayed frame numbers. This curve could be used to visualise the heartbeats and carry out the beat-to-beat analysis of the heart.

### 2.4.1 Safety and errors

Though the proposed segmentation network was quite accurate (0.922 dice similarity coefficient on the test set), there were still errors in deriving the LV area due to noise. This may affect the accuracy of the LVEF, which could result in the misclassification of HF and lead to the implementation of inappropriate treatment approaches (29). To further improve the performance, one method inspired by Jeffrey's method was proposed to fine-tune the network prediction (30). Instead of directly selecting the 90th and 10th percentiles of the left ventricular areas to serve as LVED and LVES areas, the improved Jeffrey's method also required averaging the top 10% ROI 1 and the top 10% ROI 2 in Figure 5B.

Using LV area at ED as an example, the improved Jeffrey's method consisted of four steps: (1) computing the LV area at ED at a specific (e.g., second) cardiac cycle (indicated by the second red pentagram in Figure 5B); (2) computing the LV areas for each frame and sorting them according to the calculated LV areas in descending order, then selecting the top 10% of this sorted sequence (as indicated by top ROI 1 in Figure 5B); (3) sorting the frames between ED and ES within that specific (e.g., second) cardiac cycle (indicated by top ROI 2 in

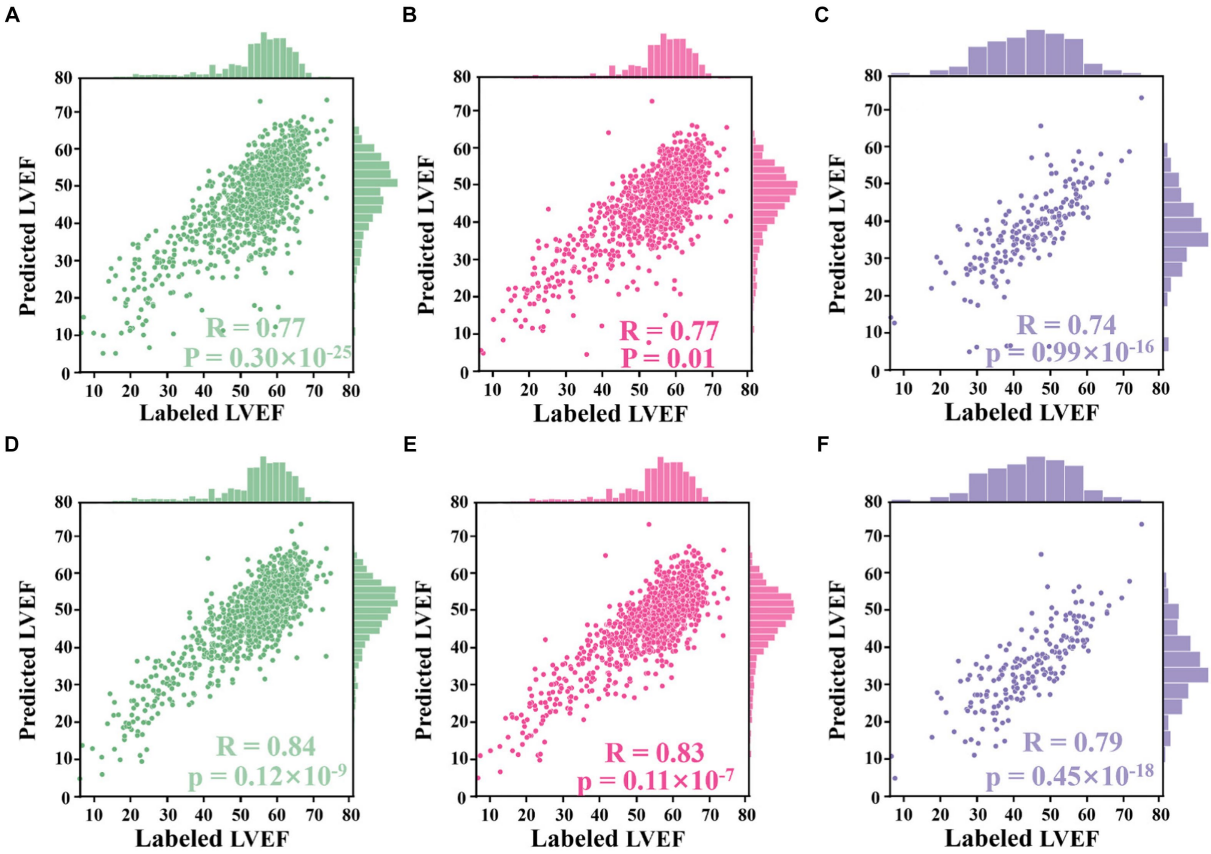


FIGURE 6
Correlation plots. **(A–D)** Results from the Stanford dataset, whilst **(E,F)** from the CAMUS dataset. **(A)** Correlation between LVEF values derived from segmentation results directly and those labelled by an experienced clinician. **(B)** Correlation between LVEF values derived from the proposed Jeffrey's method and those labelled by the clinician. **(C)** Correlation between LVEF values computed from a single cardiac cycle and labelled LVEF values. **(D)** Correlation between LVEF values computed from all cardiac cycles and labelled LVEF values. **(E)** Correlation between LVEF values derived from fine-tuned segmentation results and labelled LVEF values. **(F)** Correlation between LVEF values derived from the improved Jeffrey's method and labelled LVEF values.

Figure 5B), and then selecting the top 10% of LV areas; and (4) averaging all these selected areas to compute the final area of LV at ED. For the LV area at ES, a similar method was used, but using descending order for sorting. The improved Jeffrey's method was able to exclude outliers from segmentation effectively and thus improve the accuracy of predicted LVEF significantly, as shown in Figures 6A,B.

### 2.4.2 Analysis methods

To evaluate the accuracy of computed LVEF, Pearson's correlation coefficient ($r_{corr}$) was used to show the correlation between calculated LVEF values and those provided in the respective test set (28). Additionally, the *p*-value was used to measure whether the observed correlation coefficient is statistically significant. Furthermore, student's *t*-test was used to determine whether there was a significant difference between the results from the one-cycle method and those from the all-cycle method. In order to evaluate the HFrEF classification, ROC curves with respective AUC values were plotted to compare the predictions with benchmark methods, which can assess the performance and discriminative ability of the classification model (31, 32). The confusion matrix was also used to visualise the performance of the proposed algorithm, showing how well the model was performing in terms of correctly predicting the target variable (33, 34). This is particularly important because false negatives can lead to missed diagnoses or delayed treatment, highlighting their significance in medical decision-making. The confidence intervals were calculated by generating 100 bootstrapped samples and obtaining 95 percentile ranges for each prediction, aiming to estimate the level of uncertainty associated with the model's predictions.

## 3 Results

The proposed pipeline was trained and validated using the Stanford dataset (7,465 and 1,288 patients, respectively). The final analysis included 1,270 patients, of whom 8% (106) had LVEF <40%. Iteration and external validation used the CAMUS dataset of 200 patients, of which 66 (33%) had LVEF <40%, 62 (31%) were women, and the average age was 64.9 years. Image quality for echocardiography in the CAMUS dataset was reported as good in 113 patients (57%), adequate in 65 patients (32%), and poor in 22 patients (11%). Further details on patient characteristics are summarised in Supplementary Table S1 in Appendix B.

## 3.1 Accuracy of automated LVEF calculation

The automated method to compute LVEF given in formulation (2) was assessed in three experiments based on the segmentation network and LV length model that were trained and elaborated upon in Appendices C,D of the Supplementary file.

### 3.1.1 Experiment 1

The alternative hypothesis was that Jeffrey's method proposed in Section 3.4 could improve the performance of computing LVEF. For this, the ED and ES frames provided in the test set of the Stanford dataset were used. For each sample in the test set, LV lengths were predicted by the proposed voting ensemble learning model already trained in Section 4.2. LV areas were predicted by two methods: one was to deploy the trained network to segment their ED and ES frames and then count the number of pixels in the segmentation masks, and the other was the improved Jeffrey's method. As shown in Figures 6A,B, these two sub-figures showed that the LVEF values derived from segmentation directly had a $r_{corr}$ value of 0.77 ($p$-value <0.0001, 95% CI 0.74 to 0.80) with respect to these LVEF values provided in the test set. The correlation could be boosted to 0.84 ($p$-value <0.0001, 95% CI 0.82 to 0.86) when using the improved Jeffrey's method to compute LV areas. This experiment showed that it was necessary to fine-tune LV areas after segmentation using the proposed Jeffrey's method, which improves the accuracy of the resulting LVEF with a t-value less than 0.0001.

### 3.1.2 Experiment 2

The alternative hypothesis was that LVEF computed by averaging across all cardiac cycles (i.e., our Eq. 2 where $N > 1$) was more accurate than that from only a single cardiac cycle (i.e., the Eq. 2 where $N = 1$), where the reference was human estimates of LVEF. First, the proposed peak detection algorithm was used to identify all ED and ES phases in a given echocardiogram video from the test set of the Stanford dataset. For the former method, the first paired ED and ES frames were selected as a cycle and then computed LVEF. For the latter method, all identified cycles were used to compute an averaged LVEF value using (2) for this video (85% of the videos contain more than three cardiac cycles). As shown in Figures 6C,D, the LVEF values derived from single cycles ($r_{corr} = 0.77$, $p$-value = 0.01, 95% CI 0.75 to 0.80) were less accurate than those derived from all cycles ($r_{corr} = 0.83$, $p$-value <0.0001, 95% CI 0.81 to 0.85), when referring to these LVEF values provided in the test set ($t$-value <0.0001). Furthermore, if the second cycle was selected to compute LVEF, their respective $r_{corr}$ value could be boosted to 0.78 ($p$-value <0.0001, 95% CI 0.78 to 0.80), still inferior to the proposed all-cycle method ($t$-value <0.0001).

### 3.1.3 Experiment 3

The alternative hypothesis was that the performance of the model would be retained in an external dataset (the test set of the CAMUS dataset). To predict LV areas, the segmentation network trained from the Stanford dataset was fine-tuned on the training set of the CAMUS dataset, and then it was deployed on the test set of CAMUS. To predict LV lengths, the voting ensemble learning model trained from the Stanford dataset was deployed directly on the test set of CAMUS. As shown in Figures 6E,F, it could be seen that the $r_{corr}$ value was improved from 0.74 (p-value <0.0001, 95% CI 0.68 to 0.78) to 0.79 ($p$-value <0.0001, 95% CI 0.74 to 0.84) before and after applying for the proposed Jeffrey's method.

## 3.2 Classification of patients with HFrEF

Current HFrEF terminology was used as guidance to detect HFrEF samples from the test sets of both the Stanford and CAMUS datasets based on their LVEF predicted in Section 4.3.1. ROC curves were plotted, and their AUC values were computed in Figure 7A. Amongst these curves (see the plot legend), the first two were obtained on the Stanford dataset, and the last two on the CAMUS. The proposed

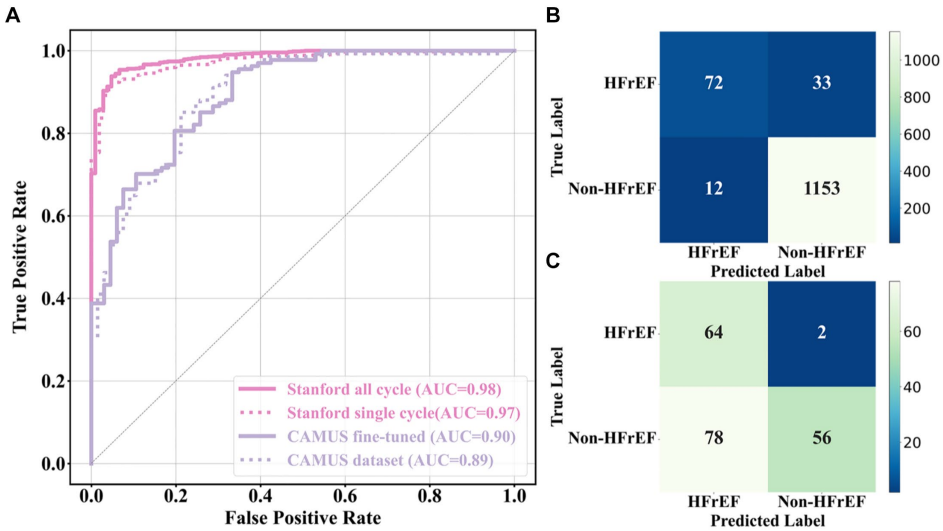


FIGURE 7
HFrEF assessment results. **(A)** ROC curves of different methods, each having an AUC value. **(B)** and **(C)** Confusion matrices computed from the Stanford and CAMUS datasets, respectively.

TABLE 1 HFrEF assessment results using AUC values with a confidence interval of 95%.

| | Stanford | CAMUS |
|---|---|---|
| Single cycle | 0.97 (0.96–0.98) | 0.89 (0.87–0.91) |
| Average cycle | 0.98 (0.97–0.99) | 0.90 (0.88–0.91) |

all-cycle method achieved an AUC value of 0.98 (95% confidence interval: 0.97 to 0.99) in the internal validation (Stanford dataset). On external validation using the CAMUS dataset, the AUC was 0.90 (95% confidence interval 0.88 to 0.91), as shown in Table 1.

In addition, the confusion metric was presented to further evaluate the accuracy of the proposed methods. Figures 7B,C show the results from the test sets of the Stanford dataset and CAMUS, respectively. For the Stanford dataset, there were 1,270 samples in its test set, of which 97% were classified correctly. There were 12 that were not HFrEF samples, but the classifier classified them as HFrEF. There were 33 HFrEF samples, but the classifier classified them as non-HFrEF. With regards to the confusion metric for CAMUS, the proposed method predicted 78 non-HFrEF as HFrEF patients, but only two with HFrEF were mistaken as non-HFrEF.

## 3.3 Beat-to-beat visualiser

A beat-to-beat visualiser was provided as the output for diagnostic purposes, in addition to the quantitative results given in the previous sections. Based on the computed LV areas and the identified ED as well as ES phases, two beat-to-beat visualisers are presented in Figures 8A,B, which were used to provide an overview of LV volumes across all cardiac cycles and provide an instantaneous summary of beat-to-beat volumetric differences as a result of sinus or pathological arrhythmias. In Figure 8A, there was a similar gap between the ED and ES frames, which was the sample with a normal sinus rhythm in heartbeats. Figure 8B was a sample marked as a patient with AF by the dataset publisher. This figure showed that the sample had irregular heartbeats, and the gap between the ED and ES frames varied across all cardiac cycles. These examples provided a visualisation of hearts having different conditions.

# 4 Discussion

This project proposed a novel pipeline method to assess cardiac function that achieved state-of-the-art results. It involved training a weakly supervised algorithm to identify the LV using expert tracings, followed by using an ellipsoid single-plane model to determine LVEF values. This pipeline outperformed previous attempts that relied on segmentation-based deep learning methods (30). Furthermore, its performance in predicting the LVEF values was robust when applied to an external dataset of echocardiogram sequences from an independent medical centre. As a result, this pipeline could have the potential to assist clinicians in achieving a more precise and reproducible assessment of cardiac function and could have the capability to identify subtle changes in LVEF beyond the precision of human readers.

One difference between the proposed pipeline and human evaluation was the method of calculating LVEF, where the pipeline was based on beat-to-beat evaluation across numerous cardiac cycles, whilst the typical clinical approach is to take just one representative beat. The process of tracing three or five beats is not commonly performed in routine practice due to the labour-intensive and time-consuming nature of the task. By automating the segmentation task, the proposed pipeline reduced the labour involved in assessing cardiac function and allowed for more frequent and accurate evaluations.

Two examples from the test set of the Stanford dataset are presented in Figures 8C,D to further explain the reason for using the all-cycle method. As can be seen, there were three cardiac cycles in Figure 8C, with three LVEF values being 63.53, 62.86, and 63.50%, respectively. In this case, calculating LVEF from any cycle would not make a significant difference. In Figure 8D, there were also three cycles, with the corresponding LVEF values being 53.68, 51.28, and 45.30%, respectively. If using the third cycle to compute LVEF, it would end up identifying this sample with HFmrEF, which would result in a true negative classification. Using the all-cycle method, the LVEF value was 50.09%, with which it was able to classify this sample correctly as HFpEF. Therefore, some recent studies based on only single-cycle information rather than all-cycle information might lead to reduced reliability and accuracy in diagnosing patients with systolic HF (14, 30, 35–37).

Another difference was that the pipeline relied on the machine to identify LV contours and ED as well as ES frames, which had the capability of computing LVEF more accurately. For example, in Figure 8D, with pink ED and ES, the LVEF value is 46.98% (HFmrEF), whilst with the corresponding green ED and ES, the LVEF value is 51.28% (HFpEF). According to the Stanford dataset publisher, this sample should have an LVEF value above 50% (38). Clearly, this method computed a correct LVEF, proving the effectiveness of the proposed peak detection algorithm, whilst labelling ED and ES incorrectly would result in an incorrect LVEF. This means the ground truth LVEF values used to train the network may already be inaccurate for some regression methods due to the fact that the selection of ED and ES frames might be incorrect and that only one cycle was used to calculate LVEF in practice rather than using three or five consecutive cardiac cycles as per the ASE recommendation. Therefore, if some regression methods used these incorrect labels to train models, their prediction and evaluation accuracy could be degraded and biased (38–41). However, the automated methods in this study had no such issues and therefore were better than direct regression methods.

One limitation of the validation was the relatively small sample size of the CAMUS dataset (only 200 samples were used for fine-tuning the network). However, the results of the LVEF were still robustly accurate when applying this learned model to the CAMUS dataset originating from a different site and time interval. Another limitation was the inability to use Simpson's biplane method (measurement of LVEF using both A4C and apical 2-chamber views), as recommended by ASE and EACVI, due to the Stanford Echo-Dynamic dataset only providing A4C views (15, 42). Instead, the area-length formulation was used based on the ellipsoid single-plane model, which still showed an excellent correlation with human-labelled LVEF calculated with Simpson's biplane ($r = 0.99$; $p < 0.0001$; mean absolute error 4.4%). Furthermore, the proposed approach could easily be modified to take into account the biplane method of LVEF calculation, with LV areas for both views derived from two separate segmentation methods (ACNN and the improved Jeffrey's

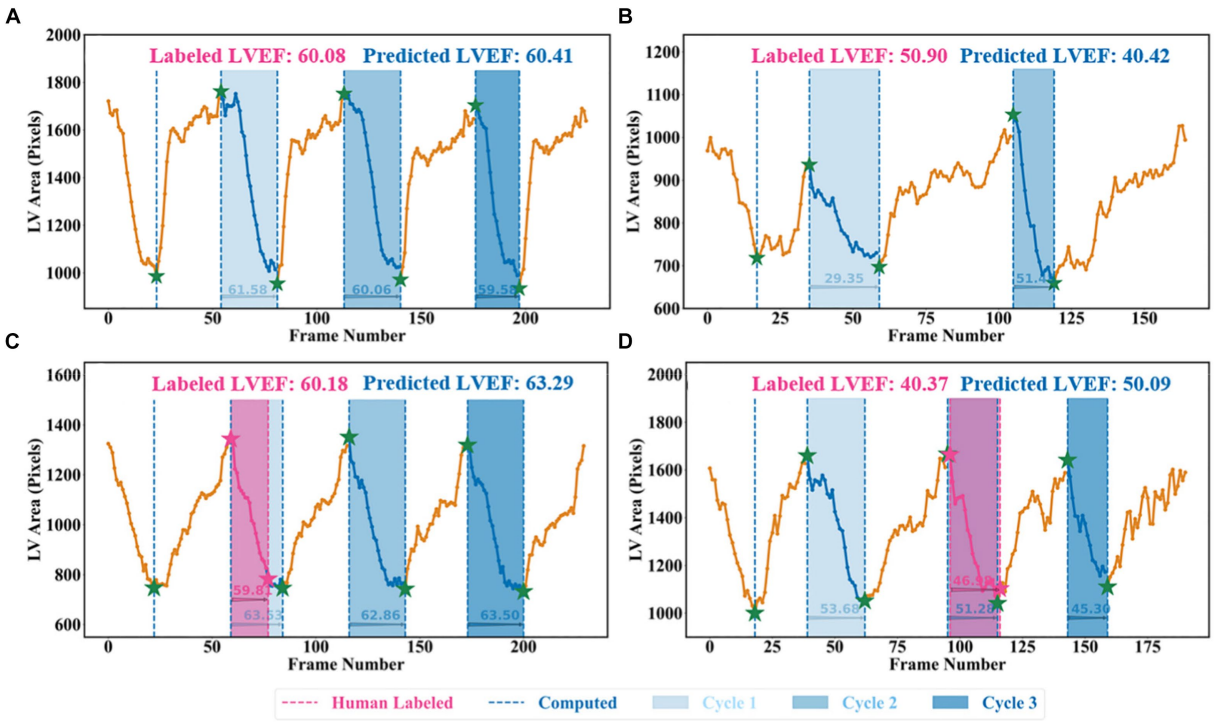


FIGURE 8
Beat-to-beat analysis. **(A)** and **(C)** Two samples with normal sinus rhythm. **(B)** Patient with atrial fibrillation. **(C)** and **(D)** Human-labelled ED and ES were not exactly at peak or bottom positions.

method), whilst LV length could be derived from the novel ensemble learning model.

## 5 Conclusion

In this project, a new pipeline method was proposed to assess cardiac function based on only Apical 4 chamber cines, which could not only provide quantitative results, such as LVEF, but also present left ventricular contours and beat-to-beat visualisers for cardiologists to visually view the samples whilst making diagnoses. Additionally, the study highlighted the importance of following the ASE and EACVI recommendations of averaging three or five cycles to obtain a more precise assessment.

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://stanfordaimi.azurewebsites.net/datasets/834e1cd1-92f7-4268-9daa-d359198b310a.

## Author contributions

YZ: Conceptualization, Methodology, Visualization, Writing - original draft, Writing – review & editing. BL: Writing – review & editing. KB: Writing – review & editing. DB: Writing – review & editing. AT: Writing – review & editing. AK: Writing – review & editing. WL: Writing – review & editing. DZ: Writing – review & editing. XW: Writing – review & editing. AM: Writing – review & editing. OT: Writing – review & editing. GG: Writing – review & editing. DK: Writing – review & editing. JD: Writing – review & editing, Supervision.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The card*AI*c team at the University of Birmingham and University Hospitals Birmingham NHS Foundation Trust have received support and funding from the NIHR Birmingham Biomedical Research Centre (NIHR203326), MRC Health Data Research UK (HDRUK/CFC/01), NHS Data for R&D Subnational Secure Data Environment Programme (West Midlands), the British Heart Foundation University of Birmingham Accelerator (AA/18/2/34218), and the Korea Cardiovascular Bioresearch Foundation (CHORUS Seoul 2022).

## 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.

## Author disclaimer

The opinions expressed in this paper are those of the authors and do not represent any of the listed organisations; none of the organisations had any role in the design or conduct of the study (including collection, analysis, and interpretation of the data) or any involvement in the preparation, review, or approval of the study.

## Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2024.1354070/full#supplementary-material

## References


1. Savarese G, Becher PM, Lund LH, Seferovic P, Rosano GMC, Coats AJS. Global burden of heart failure: a comprehensive and updated review of epidemiology. *Cardiovasc Res*. (2023) 118:3272–87. doi: 10.1093/cvr/cvac013

2. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Bohm M, et al. 2021 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure. *Eur Heart J*. (2021) 42:3599–726. doi: 10.1093/eurheartj/ehab368

3. Cleland JGF, Bunting KV, Flather MD, Altman DG, Holmes J, Coats AJS, et al. Beta-blockers for heart failure with reduced, mid-range, and preserved ejection fraction: an individual patient-level analysis of double-blind randomized trials. *Eur Heart J*. (2018) 39:26–35. doi: 10.1093/eurheartj/ehx564

4. Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. *J Am Soc Echocardiogr*. (2015) 28:e14:1–39.e14. doi: 10.1016/j.echo.2014.10.003

5. Lang RM, Bierig M, Devereux RB, Flachskampf FA, Foster E, Pellikka PA, et al. Recommendations for chamber quantification: a report from the American Society of Echocardiography's guidelines and standards committee and the chamber quantification writing group, developed in conjunction with the European Association of Echocardiography, a branch of the European Society of Cardiology. *J Am Soc Echocardiogr*. (2005) 18:1440–63. doi: 10.1016/j.echo.2005.10.005

6. Myhr KA, Pedersen FHG, Kristensen CB, Visby L, Hassager C, Mogelvang R. Semi-automated estimation of left ventricular ejection fraction by two-dimensional and three-dimensional echocardiography is feasible, time-efficient, and reproducible. *Echocardiography*. (2018) 35:1795–805. doi: 10.1111/echo.14112

7. Phad N, de Waal K. Left ventricular ejection fraction using manual and semi-automated biplane method of discs in very preterm infants. *Echocardiography*. (2020) 37:1265–71. doi: 10.1111/echo.14784

8. Bunting KV, Gill SK, Sitch A, Mehta S, O'Connor K, Lip GY, et al. Improving the diagnosis of heart failure in patients with atrial fibrillation. *Heart*. (2021) 107:902–8. doi: 10.1136/heartjnl-2020-318557

9. Moal O, Roger E, Lamouroux A, Younes C, Bonnet G, Moal B, et al. Explicit and automatic ejection fraction assessment on 2D cardiac ultrasound with a deep learning-based approach. *Comput Biol Med*. (2022) 146:105637. doi: 10.1016/j.compbiomed.2022.105637

10. Sartipy U, Dahlstrom U, Fu M, Lund LH. Atrial fibrillation in heart failure with preserved, mid-range, and reduced ejection fraction. *JACC Heart Fail*. (2017) 5:565–74. doi: 10.1016/j.jchf.2017.05.001

11. Taniguchi N, Miyasaka Y, Suwa Y, Harada S, Nakai E, Shiojima I. Heart failure in atrial fibrillation - an update on clinical and echocardiographic implications. *Circ J*. (2020) 84:1212–7. doi: 10.1253/circj.CJ-20-0258

12. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ*. (2022) 377:e070904. doi: 10.1136/bmj-2022-070904

13. Ouyang D, He B, Ghorbani A, Lungren MP, Ashley EA, Liang DH, et al. Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning In: *NeurIPS ML4H workshop*. Vancouver, BC, Canada: NeurIPS ML4H workshop (2019)

14. Leclerc S, Smistad E, Pedrosa J, Ostvik A, Cervenansky F, Espinosa F, et al. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. *IEEE Trans Med Imaging*. (2019) 38:2198–210. doi: 10.1109/TMI.2019.2900516

15. BAC RBS, Mayers DL, Martin RP. Two-dimensional echocardiographic measurement of left ventricular ejection fraction: prospective analysis of what constitutes an adequate determination. *Am Heart J*. (1982) 104:136–44. doi: 10.1016/0002-8703(82)90651-2

16. Chen LC, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. *arXiv Preprint*. (2017) arXiv:1706.05587. doi: 10.48550/arXiv.1706.05587

17. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. *arXiv Preprint*. (2015) arXiv:1511.07122. doi: 10.48550/arXiv.1511.07122

18. Smistad E., Østvik A., Salte I.M., Leclerc S., Bernard O., Lovstakken L. Fully automatic real-time ejection fraction and MAPSE measurements in 2D echocardiography using deep neural networksC. (2018) IEEE International Ultrasonics Symposium (IUS). 1–4

19. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn*. (2006) 63:3–42. doi: 10.1007/s10994-006-6226-1

20. Hastie T, Rosset S, Zhu J, Zou H. Multi-class adaboost. *Stat Interface*. (2009) 2:349–60. doi: 10.4310/SII.2009.v2.n3.a8

21. Ranstam J, Cook JA. LASSO regression. *Br J Surg*. (2018) 105:1348. doi: 10.1002/bjs.10895

22. Pereira JM, Basto M, Silva AF. The logistic Lasso and ridge regression in predicting corporate failure. *Proc Econ Finance*. (2016) 39:634–41. doi: 10.1016/S2212-5671(16)30310-0

23. Shakhnarovich G., Darrell T., Indyk P.. Nearest-neighbor methods in learning and vision. IEEE Trans Neural Networks, (2008) 19:377.

24. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. (2001) 29:1189–232. doi: 10.1214/aos/1013203451

25. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. IEEE Conference on Computer Vision and Pattern Recognition (1995);14:1137–1145.

26. Nakagawa S, Johnson PCD, Schielzeth H. The coefficient of determination R(2) and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J R Soc Interface*. (2017) 14:20170213. doi: 10.1098/rsif.2017.0213

27. Jonathan Long ES, Darrell Trevor. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition. (2015):3431–3440.

28. Dokeroglu T, Deniz A, Kiziloz HE. A comprehensive survey on recent metaheuristics for feature selectionJ. *Neurocomputing*. (2022) 494:269–96. doi: 10.1016/j.neucom.2022.04.083

29. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, et al. 2016 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure: the task force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC)developed with the special contribution of the heart failure association (HFA) of the ESC. *Eur Heart J*. (2016) 37:2129–200. doi: 10.1093/eurheartj/ehw128

30. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, et al. Fully automated echocardiogram interpretation in clinical practice. *Circulation*. (2018) 138:1623–35. doi: 10.1161/CIRCULATIONAHA.118.034338

31. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. (1983) 148:839–43. doi: 10.1148/radiology.148.3.6878708

32. Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett*. (2006) 27:861–74. doi: 10.1016/j.patrec.2005.10.010

33. Stehman SV. Selecting and interpreting measures of thematic classification accuracy. *Remote Sens Environ*. (1997) 62:77–89. doi: 10.1016/S0034-4257(97)00083-7

34. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *J Mach Learn Technol*. (2011) 2:37–63. doi: 10.48550/arXiv.2010.16061

35. Dong S, Luo G, Sun G, Wang K, Zhang HA. Left ventricular segmentation method on 3D echocardiography using deep learning and Snake. 2016 Computing in Cardiology Conference (CinC) (2016)

36. Smistad E, Ostvik A, Salte IM, Melichova D, Nguyen TM, Haugaa K, et al. Real-time automatic ejection fraction and foreshortening detection using deep learning. *IEEE Trans Ultrason Ferroelectr Freq Control*. (2020) 67:2595–604. doi: 10.1109/TUFFC.2020.2981037

37. Thavendiranathan P, Liu S, Verhaert D, Calleja A, Nitinunu A, Van Houten T, et al. Feasibility, accuracy, and reproducibility of real-time full-volume 3D transthoracic echocardiography to measure LV volumes and systolic function: a fully automated

endocardial contouring algorithm in sinus rhythm and atrial fibrillation. *JACC Cardiovasc Imaging*. (2012) 5:239–51. doi: 10.1016/j.jcmg.2011.12.012

38. Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP, et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*. (2020) 580:252–6. doi: 10.1038/s41586-020-2145-8

39. Ghorbani A, Ouyang D, Abid A, He B, Chen JH, Harrington RA, et al. Deep learning interpretation of echocardiograms. *NPJ Digit Med*. (2020) 3:10. doi: 10.1038/s41746-019-0216-8

40. Wenhao Jiang KHL, Liu Z, Fan Y, Kwok K-W, Lee AP-W. Deep learning algorithms to automate left ventricular ejection fraction assessments on 2-dimensional echocardiography. *J Am Coll Cardiol*. (2019) 73:1610. doi: 10.1016/S0735-1097(19)32216-8

41. Kusunose K, Haga A, Yamaguchi N, Abe T, Fukuda D, Yamada H, et al. Deep learning for assessment of left ventricular ejection fraction from echocardiographic images. *J Am Soc Echocardiogr*. (2020) 33:e1:632–635.e1. doi: 10.1016/j.echo.2020.01.009

42. Fonarow GC, Hsu JJ. Left ventricular ejection fraction: what is "Normal"? *JACC Heart Fail*. (2016) 4:511–3. doi: 10.1016/j.jchf.2016.03.021

# Frontiers in Medicine

**Translating medical research and innovation into improved patient care**

A multidisciplinary journal which advances our medical knowledge. It supports the translation of scientific advances into new therapies and diagnostic tools that will improve patient care.

## 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