frontiers | Research Topics

Artificial intelligence and big data for value-based care

Edited by
Md. Mohaimenul Islam, Ming-Chin Lin and Abeer Alsadoon

Published in
Frontiers in Medicine
Frontiers in Public Health

ISSN 1664-8714
ISBN 978-2-83251-588-4
DOI 10.3389/978-2-83251-588-4

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

Artificial intelligence and big data for value-based care

Topic editors

Md. Mohaimenul Islam — AESOP TECHNOLOGY, Taiwan

Ming-Chin Lin — Taipei Medical University, Taiwan

Abeer Alsadoon — Charles Sturt University, Australia

Citation

Islam, M. M., Lin, M.-C., Alsadoon, A., eds. (2023). *Artificial intelligence and big data for value-based care*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-83251-588-4

Table of contents

frontiers | Frontiers in Medicine

TYPE Editorial
PUBLISHED 23 January 2023
DOI 10.3389/fmed.2023.1134021

OPEN ACCESS

EDITED AND REVIEWED BY
Arch Mainous,
University of Florida, United States

*CORRESPONDENCE
Md. Mohaimenul Islam
✉ 2010rubel@gmail.com

SPECIALTY SECTION
This article was submitted to Family Medicine and Primary Care, a section of the journal Frontiers in Medicine

RECEIVED 29 December 2022
ACCEPTED 13 January 2023
PUBLISHED 23 January 2023

CITATION
Han CT, Lin M-C, Alsadoon A and Islam MM (2023) Editorial: Artificial intelligence and big data for value-based care.
Front. Med. 10:1134021.
doi: 10.3389/fmed.2023.1134021

Editorial: Artificial intelligence and big data for value-based care

Cheng Ta Han[1,2], Ming-Chin Lin[1,2,3], Abeer Alsadoon[4] and Md. Mohaimenul Islam[5]*

[1]Department of Neurosurgery, Shuang Ho Hospital, Taipei Medical University, New Taipei City, Taiwan, [2]Taipei Neuroscience Institute, Taipei Medical University, Taipei, Taiwan, [3]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan, [4]School of Computing and Mathematics, Charles Sturt University (CSU), Wagga Wagga, NSW, Australia, [5]International Center for Health Information Technology, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan

KEYWORDS

artificial intelligence, healthcare, healthcare system, value-based care, healthcare cost

Editorial on the Research Topic
Artificial intelligence and big data for value-based care

Quality care is a key component of the health care system. The gap between actual care received and ideal care quality in the existing healthcare system is enormous. Although healthcare spending is remarkably rising than any other global economy, the healthcare system is still facing immense challenges in inaccurate diagnoses, medication errors, inappropriate or unnecessary treatments, and insufficient clinical practices. The World Health Organization (WHO) report 2020 shows that global spending on healthcare has reached US$ 8.3 trillion, ∼10% of the global GDP (1). A decisive effort is necessary to move value-based care from fee-for-service to improve financial and clinical performance. Value-based care has the potential to promote better clinical outcomes without increasing costs.

A shift to value-based care from fee-for-service is not a dream because of the availability of patient data in the electronic health record (EHR) systems, standardization framework, and advanced algorithms. Clinicians can collect overwhelming amounts of patient data and utilize advanced analytical tools to make accurate predictions and actionable insights to improve overall provider performance, decrease medical errors, and reduce healthcare waste (2). Chen et al. developed an artificial intelligence (AI) system to correctly classify long-term cardiovascular outcomes in patients with normal ejection fraction. Echocardiographic data from 61,525 patients were collected to develop an AI model, which was later internally and externally validated using data from 3,810 and 5,760 patients. This AI-based system was able to stratify patients with a left ventricular end-diastolic diameter (LV-D) and predicts ECG-EF accurately with high AUCs. Nowadays, stereotactic body radiotherapy (SBRT) is considered one of the key treatment options for patients with early-stage lung cancer. It has shown a beneficial effect in improving tumor control and overall survival rate. A recent study tested the performance of the Mask R-CNN-based algorithm for evaluating the dose accuracy of a lung SBRT treatment plan with the target of a newly predicted internal target volume ($ITV_{predict}$) and the feasibility of its clinical application (Zhang et al.). The cone-beam CT (CBCT) images were collected from early-stage 45 lung cancer patients who underwent SBRT at Huadong Hospital. This AI-enable tool was able to predict the ITV volume of large tumors more accurately, which ensures the feasibility of this automated model in making an appropriate treatment plan.

Alzheimer's disease (AD) is a critical global health problem contributing to a substantial financial burden. A previous study reported that ∼6.5 million aged 65 or older are living with AD in the USA. Early identification of AD patients significantly reduces healthcare costs and improves patients' quality of life. Since AI techniques based on MRI are being used in the early diagnosis of AD, a novel deep learning radiomics (DLR) model was developed to classify cognitively normal adults at risk of AD from normal control using T1-weighted structural MRI images (Jiang et al.). A total of 417 patients were included in the study, and MRI data of those patients were used to divide patients into pre-AD (181 individuals) and control groups (236 individuals) based on a standard uptake ratio >1.18. AI model achieved state-of-the-art performance in classifying pre-AD and normal control with an accuracy of 89.85% ± 1.12%. It is now established that advanced AI algorithms have surpassed traditional statistical methods in image recognition and being extensively used in medical image analysis. In the last decade, AI-based radiomic models have made meaningful contributions to detecting chronic diseases, including lung cancer. A systematic review and meta-analysis included a total of 19 published studies to evaluate the diagnostic accuracy of AI models for lung cancer staging (Zheng et al.). The findings of AI models have the potential to improve diagnostic accuracy for lung cancer staging in terms of sensitivity, specificity, and the area under the receiver operating curve (AUROC).

AI models have tremendous potential to reduce medical errors, effectively utilize limited resources, and ultimately improve value by making accurate and effective clinical decisions. Recently, several studies internationally validated AI tools and achieved classification accuracy performance that outperformed human performance (3, 4). Transforming to value-based care from a fee-for-service will face significant challenges in achieving quality for all and will take time. But widespread adoption of value-based care would help lower healthcare costs while simultaneously improving the quality of care.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

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. World Health Organization. *Global Spending on Health 2020: Weathering the Storm*. Geneva: WHO (2020).

2. Crowson MG, Chan TC. Machine learning as a catalyst for value-based health care. *J Med Syst.* (2020) 44:1–3. doi: 10.1007/s10916-020-01607-5

3. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. *Nature.* (2020) 577:89–94. doi: 10.1038/s41586-019-1799-6

4. Birkenbihl C, Emon MA, Vrooman H, Westwood S, Lovestone S, Hofmann-Apitius M, et al. Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia-lessons for translation into clinical practice. *EPMA J.* (2020) 11:367–76. doi: 10.1007/s13167-020-00216-z

frontiers in Public Health

ORIGINAL RESEARCH
published: 22 March 2022
doi: 10.3389/fpubh.2022.860135

Dosimetric Study of Deep Learning-Guided ITV Prediction in Cone-beam CT for Lung Stereotactic Body Radiotherapy

*Shujun Zhang†, Bo Lv†, Xiangpeng Zheng, Ya Li, Weiqiang Ge, Libo Zhang, Fan Mo and Jianjian Qiu**

Department of Radiation Oncology, Huadong Hospital, Fudan University, Shanghai, China

OPEN ACCESS

Edited by:
Ming-Chin Lin,
Taipei Medical University, Taiwan

Reviewed by:
Yueh-hsun Lu,
Taipei Medical University, Taiwan
Rahul Pratap Kotian,
Gulf Medical University, United Arab Emirates

***Correspondence:**
Jianjian Qiu
qiujianjian@fudan.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Family Medicine and Primary Care, a section of the journal Frontiers in Public Health

Received: *22 January 2022*
Accepted: *21 February 2022*
Published: *22 March 2022*

Citation:
Zhang S, Lv B, Zheng X, Li Y, Ge W, Zhang L, Mo F and Qiu J (2022) Dosimetric Study of Deep Learning-Guided ITV Prediction in Cone-beam CT for Lung Stereotactic Body Radiotherapy. Front. Public Health 10:860135. doi: 10.3389/fpubh.2022.860135

Purpose: The purpose of this study was to evaluate the accuracy of a lung stereotactic body radiotherapy (SBRT) treatment plan with the target of a newly predicted internal target volume ($ITV_{predict}$) and the feasibility of its clinical application. $ITV_{predict}$ was automatically generated by our in-house deep learning model according to the cone-beam CT (CBCT) image database.

Method: A retrospective study of 45 patients who underwent SBRT was involved, and Mask R-CNN based algorithm model helped to predict the internal target volume (ITV) using the CBCT image database. The geometric accuracy of $ITV_{predict}$ was verified by the Dice Similarity Coefficient (DSC), 3D Motion Range (R_{3D}), Relative Volume Index (RVI), and Hausdorff Distance (HD). The $PTV_{predict}$ was generated by $ITV_{predict}$, which was registered and then projected on free-breath CT (FBCT) images. The PTV_{FBCT} was margined from the GTV on FBCT images gross tumor volume on free-breath CT (GTV_{FBCT}). Treatment plans with the target of Predict planning target volume on CBCT images ($PTV_{predict}$) and planning target volume on free-breath CT (PTV_{FBCT}) were respectively re-established, and the dosimetric parameters included the ratio of the volume of patients receiving at least the prescribed dose to the volume of PTV ($R_{100\%}$), the ratio of the volume of patients receiving at least 50% of the prescribed dose to the volume of PTV in the Radiation Therapy Oncology Group (RTOG) 0813 Trial ($R_{50\%}$), Gradient Index (GI), and the maximum dose 2 cm from the PTV (D_{2cm}), which were evaluated *via* $Plan_{4DCT}$, plan which based on $PTV_{predict}$ ($Plan_{predict}$), and plan which based on PTV_{FBCT} ($Plan_{FBCT}$).

Result: The geometric results showed that there existed a good correlation between $ITV_{predict}$ and ITV on the 4-dimensional CT [ITV_{4DCT}; DSC= 0.83 ±0.18]. However, the average volume of $ITV_{predict}$ was 10% less than that of ITV_{4DCT} ($p = 0.333$). No significant difference in dose coverage was found in $V_{100\%}$ for the ITV with 99.98 ± 0.04% in the ITV_{4DCT} vs. 97.56 ± 4.71% in the $ITV_{predict}$ ($p = 0.162$). Dosimetry parameters of PTV, including $R_{100\%}$, $R_{50\%}$, GI and D_{2cm} showed no statistically significant difference between each plan ($p > 0.05$).

Conclusion: Dosimetric parameters of $Plan_{predict}$ are clinically comparable to those of the original $Plan_{4DCT}$. This study confirmed that the treatment plan based on $ITV_{predict}$ produced by our model could automatically meet clinical requirements. Thus, for patients undergoing lung SBRT, the model has great potential for using CBCT images for ITV contouring which can be used in treatment planning.

Keywords: 4DCT, CBCT (cone beam computed tomography), SBRT (stereotactic body radiation therapy), deep learning, Mask R-CNN

INTRODUCTION

For patients with early-stage lung cancer, stereotactic body radiotherapy (SBRT) has become one of the primary treatment options. It has been proven to significantly improve the tumor control and overall survival rate of patients with early-stage lung cancer (1–4).

Currently, the most popular treatment method is to use four-dimensional CT (4DCT) imaging to generate the internal target volume (ITV) contour, which expresses the volume of a tumor moving throughout a patient's breathing. This ITV contour from a four-dimensional averaged (4DAVG) image is used to generate a radiation treatment plan (5).

With the 4DCT technique, image acquisition is associated with the patient's breathing curve. The limitations of 4DCT are as follows: (1) Required high patient compliance as an irregular breathing curve can reduce the image quality and affect the accuracy of tumor contouring (6); (2) Complicated and professional operation as it requires a longer time to acquire 4DCT images, which could increase the instability and randomness of the simulation (7); (3) Low popularity as it is estimated that less than half of radiotherapy centers are equipped with four-dimensional scanners (8). Overall, these limits may potentially reduce the SBRT accuracy in treatment.

Conversely, CBCT has high popularity and is conventionally equipped in a linear accelerator (9). In addition, it is mainly used to compare the anatomical landmarks from treatment planning CT images in clinical practice, which are used to determine intra/inter-fraction motion (10). CBCT rotates 360° around the patient's body and then finishes the CT image scanning within a period of time (∼1 min) which includes 10–12 respiration time phases and the motion trajectory of the tumor. The limitations of CBCT are as follows. First, poor image quality is the main factor affecting radiation oncologist determination of lung tumor volume (11). Second, due to a prominent amount of and artifacts, the dose calculation based on CBCT images may be inaccurate for treatment planning (12).

In recent years, "deep learning" has been extensively used in medical image processing. Among them, convolutional neural networks (CNNs) are the primary methods of target detection and segmentation (13–16). Mask R-CNN is a simple, flexible, commonly used framework for object instance segmentation, and is popular in medical image processing (17). Bouget et al. used the detection of mediastinal lymph nodes in CT images for lung cancer staging while enabling good instance detection (18). Zhang et al. successfully used Mask R-CNN to detect lung tumors on PET images, which has more effectively and precisely while suitably avoiding incorrect detection of tumors (19). Some previous studies used Mask R-CNN on segmentation, such as detection and classification the breast tumors on sonograms (20) and brain tumor segmentation for dynamic susceptibility contrast-enhanced perfusion imaging (21). These studies also show the great potential of Mask R-CNN in object detection and segmentation, presenting a possibility of it being used in clinical applications of medical images in the future.

Our preliminary research confirmed that the upgraded Mask R-CNN model could predict the ITV with CBCT image accuracy (22). Meanwhile, SBRT delivers high radiation doses to the tumor target in a hypo-fractionated area with a minimum dose to the tissue around the target area (19). Therefore, dosmetric research for lung SBRT is important. This study aimed to evaluate the dose accuracy of a lung SBRT treatment plan with $ITV_{predict}$ and the feasibility of its clinical application.

MATERIALS AND METHODS

Patient Data

Forty-five lung cancer patients (average age was 68 years, range: 55–86 years) who underwent SBRT at Huadong Hospital from January 2020 to July 2021 were randomly selected for this study.

Image Acquisition

Each patient's free-breath CT (FBCT) was scanned by a Siemens Somatom Definition AS® CT scanner (Siemens Somatom Sensation, Munich, Germany) with a pitch of 1.5 and slice thickness of 1 mm. 4DCT images were acquired with the additional assistance of Varian real-time position management (RPM) (Varian Medical Systems, Palo Alto, USA) using the same scanning parameters. The first treatment fraction of CBCT images were acquired in our CBCT system (100 kVp and 100 mAs, rotated at 360° with a speed of 6° per second), equipped with a VarianVitalBeam™ linear accelerator (Varian Medical Systems, Palo Alto, USA).

ITV Acquisition

The gross tumor volume (GTV) on FBCT images (GTV_{FBCT}) and the ITV on 4DAVG (ITV_{4DCT}) images were contoured by two radiation oncologists with expertise in lung tumors. Our model, using the Mask R-CNN algorithm with a convolutional block attention module (CBAM) module embedded, was used to automatically establish a newly predicted ITV ($ITV_{predict}$)

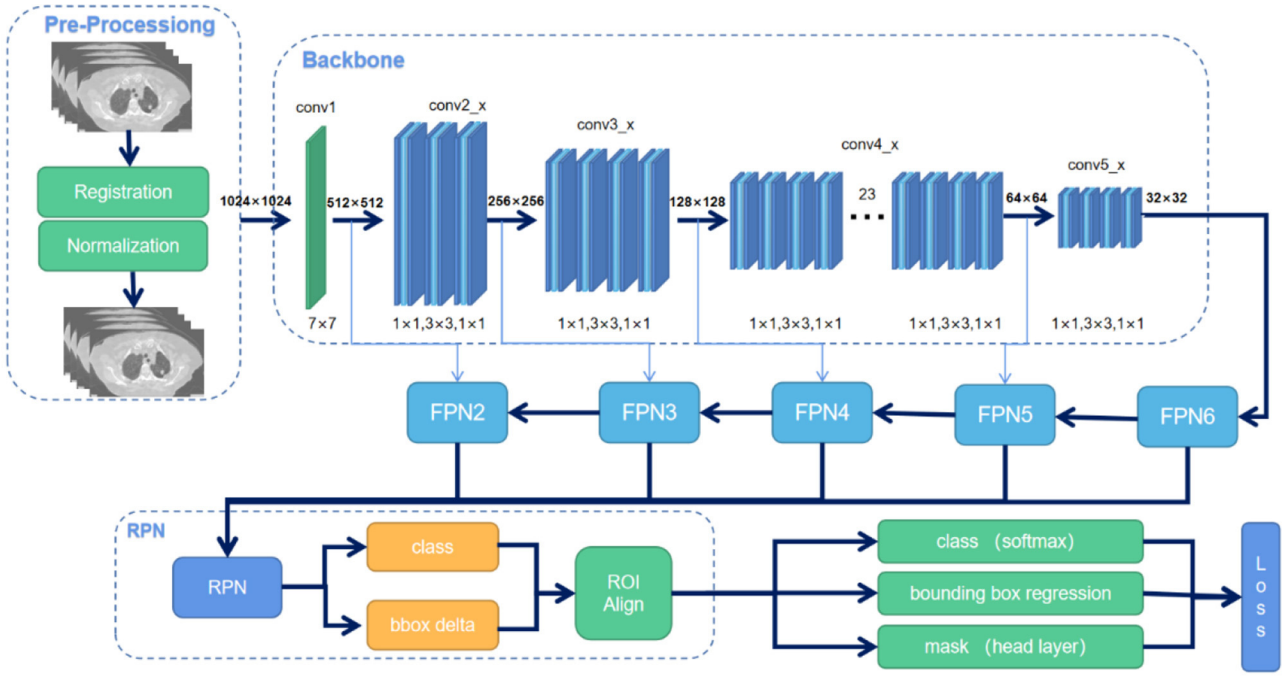

FIGURE 1 | Overview of the Mask R-CNN workflow[22]. There are two steps. First, it generates proposals for lung tumor regions based on the input cone-beam CT (CBCT) image. Second, it generates a tumor pixel-level mask based on the proposal created in the first step.

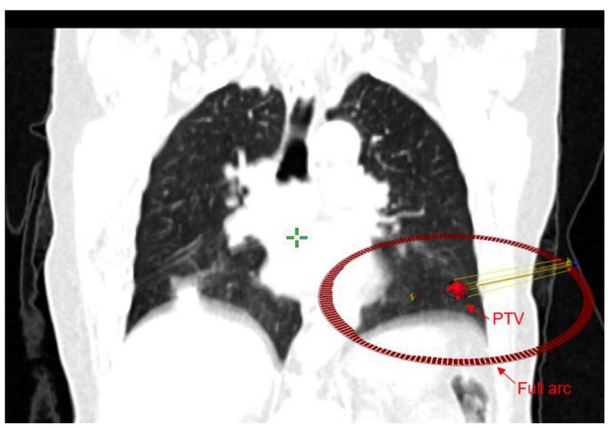

FIGURE 2 | Volumetric-modulated arc therapy (VMAT) plans with a full arc.

using the CBCT image database (18). Then, we registered the CBCT image and FBCT image and projected the $ITV_{predict}$ on the FBCT image to calculate the dose. The model workflow is shown in **Figure 1**.

Treatment Plan

PTV_{4DCT} and $PTV_{predict}$ were defined as ITV margins of 5 mm on 4DAVG images and FBCT images, respectively. PTV_{FBCT} was defined as GTV_{FBCT} with a margin of 10 mm in the craniocaudal (CC) direction and 5 mm in the left-right (LR) and anterior-posterior (AP) directions in the FBCT image.

All patient plans were replanned in the Varian Eclipse® system (version 15.5), which was generated by a full arc and was used depending on the location and anatomic relationships of the tumors and normal tissues (**Figure 2**), by our experienced medical physicists. We used a 6 MV-FFF (DR: 1400 MU/min) energy and the anisotropic analytical dose calculation algorithm (AAA) with a 2.5 mm^3 calculating grid in all plans. All patients received prescription of 60 grays (Gy) in 10 fractions (6 Gy per fraction) for over 2 weeks. The planning objectives aimed to cover the PTV with 95% of the prescribed dose in all plans.

Geometry Evaluation Parameters for the Prediction Model

The Dice Similarity Coefficient (DSC), 3D Motion Range (R_{3D}), Relative Volume Index (RVI), and Hausdorff Distance (HD) were calculated to assess the agreement between $ITV_{predict}$ and ITV_{4DCT}. All statistical tests were performed using SciPy (23) in Python.

PTV Evaluation Parameters

The volume of PTV (V_{PTV}, cm^3), mean dose (D_{mean}), the maximum dose received by 2% ($D_{2\%}$), and the minimum dose received by 98% of the evaluated PTV volume ($D_{98\%}$) were determined. The percent of the PTV receiving 100% of the prescription dose ($V_{100\%}$) and the dose covering 95% of PTV ($D_{95\%}$) were also calculated.

A steep dose gradient at the margin of the target volume is an important part of the SBRT plan to protect the normal organization. Some parameters for quantification have been reported in the literature, including $R_{100\%}$, $R_{50\%}$, the

Gradient Index (GI), and the maximum dose 2 cm from the PTV (D_{2cm}).

$R_{100\%}$

$R_{100\%}$ is the ratio of the volume of patients receiving at least the prescribed dose to the volume of PTV (9). When the value of $R_{100\%}$ is closer to 1, it means that the dose distribution has more conformity for PTV.

$$R_{100\%} = \frac{V_{100\%}}{V_{PTV}}$$

TABLE 1 | Planning objectives for critical structures.

Objectives	Parameters	Limit
Normal Lung	V 20 Gy	<10%
	V12.5 Gy	<15%
Heart	D_{max}	<32.5Gy
Trachea	D_{max}	<32.0Gy
Esophagus	D_{max}	<35.0Gy
Spinal Cord	D_{max}	<25.0Gy

TABLE 2 | Patient and tumor characteristics.

Parameter	Total
Patients ($n = 45$)	Female = 17, Male = 28
Median age in years (range)	68 (55–86)
Median ITV in cm^3 (range)	21.42 (0.7–65.7)
Tumor location ($n = 45$)	5 LUL, 13 LLL, 6 RUL, 7 RLL, 14 RML

LUL, left upper lobe; RUL, right upper lobe; LLL, left lower lobe; RLL, right lower lobe; RML, right middle lobe.

TABLE 3 | Geometry parameters of the patients.

Parameters	DSC (mean ± SD)	R_{3D} (mean ± SD)	RVI (mean ± SD)	HD (mean ± SD)
	0.83 ± 0.18	3.08 ± 2.81	1.14 ± 0.21	19.77 ± 21.59

$R_{50\%}$

$R_{50\%}$ was defined as the ratio of the volume of patients receiving at least 50% of the prescribed dose to the volume of PTV in the Radiation Therapy Oncology Group (RTOG) 0813 Trial (9). The $R_{50\%}$ is an evaluation index for damage to irradiated normal tissues (24).

$$R_{50\%} = \frac{V_{50\%}}{V_{PTV}}$$

Gradient Index (GI)

The Gradient Index (GI) is defined as the ratio of the volume of the patient receiving at least 50% of the prescription dose to the volume of the patient receiving at least 100% of the prescription dose (25). It was used to measure dose fall-off outside of the PTV. The dose falling off outside the target volume is very important in SBRT, especially as a predictor of complications (26).

$$Gradient\ Indax\ (GI) = \frac{V_{50\%}}{V_{100\%}}$$

OARs Evaluation Parameters and Treatment Efficiency Parameters

The dosimetric parameter acceptance of normal tissues (27), which refers to the RTOG 0813 Trial, is listed in **Table 1**. The ITV acquisition time was manually recorded for delineation efficiency and automatically recorded for model, and machine monitor units (Mus) were recorded for treatment efficiency.

Statistical Analysis

The statistical significance of the difference between the groups was assessed using 1-way analysis of variance (ANOVA) by SPSS software release 20.0, and the statistical significance was $p < 0.05$.

RESULTS

Clinical Characteristics

Patient and tumor characteristics of the 45 patients are described in **Table 2**.

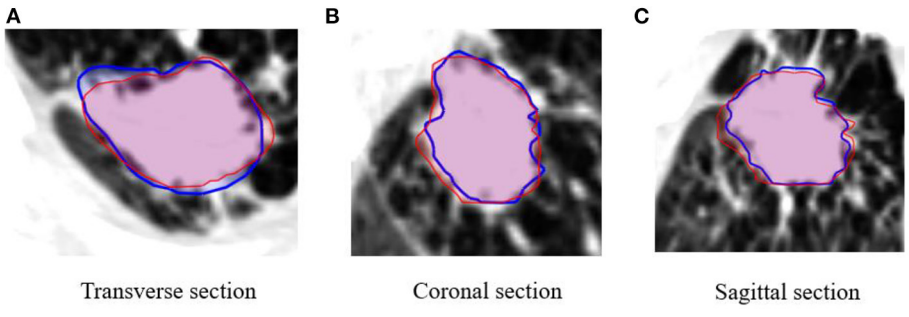

FIGURE 3 | Comparison of contours between ITV on the 4-dimensional CT (ITV_{4DCT}, red) and ITV prediction ($ITV_{predict}$, blue) in different views of patient images.

Geometry Evaluation

The DSC value between ITV_{4DCT} and $ITV_{predict}$ was 0.83 ± 0.18. The DSC value indicates that ITV_{4DCT} and $ITV_{predict}$ have a good correlation (**Table 3**). In addition, **Figure 3** shows a visual evaluation from different perspectives. $ITV_{predict}$ can outline the patient's tumor contour and is similar to the radiation oncologist contouring (ITV_{4DCT}). The visual assessment shows that the results are reasonable.

TABLE 4 | Calculated prescription dose coverage (V_{100}) and dose to 95% (D_{95}) of ITV.

Variables	ITV_{4DCT} (mean ± SD)	$ITV_{predict}$ (mean ± SD)	*P* value
Volume (cm^3)	21.41 ± 9.38	19.31 ± 8.83	0.333
$V_{100\%}$ (%)	99.98 ± 0.04	97.56 ± 4.71	0.162
$D_{95\%}$ (Gy)	63.84 ± 1.55	61.02 ± 5.56	0.207

**A significant difference existed ($p < 0.05$).*
SD, standard deviation.

ITV Evaluation

The average volumes of the ITV_{4DCT} and $ITV_{predict}$ were 21.41 ± 9.38 cm^3 and 19.31 ± 8.83 cm^3, respectively. Compared to ITV_{4DCT}, $ITV_{predict}$ reduced ITV volume by 10% on average ($p = 0.333$). No significant difference was found in ITV volume. However, no significant difference was found in $V_{100\%}$ for the ITV with 99.98 ± 0.04% in ITV_{4DCT} vs. 97.56 ± 4.71% in $ITV_{predict}$ ($p = 0.162$) (**Table 4**).

PTV Evaluation

The PTV evaluation parameters are shown in **Table 5**. The GI value of $Plan_{predict}$ (GI = 3.98 ± 0.42) was slightly lower than that of $Plan_{FBCT}$ (GI = 4.74 ± 1.01), indicating that the descending gradient of PTV was better than that of $Plan_{FBCT}$ and second to that of $Plan_{4DCT}$ (GI = 3.31 ± 0.89). The $R_{100\%}$ value for $Plan_{4DCT}$ ($R_{100\%}$ =1.05 ± 0.11) was always lower than that for $Plan_{predict}$ ($R_{100\%}$ = 1.08 ± 0.05) and $Plan_{FBCT}$ ($R_{100\%}$ = 1.12 ± 0.06), and the results show that $Plan_{4DCT}$ has the best performance and high conformability. However, there was no statistically significant difference between the plans (F = 0.141).

TABLE 5 | Dosimetric parameter comparison among $Plan_{4DCT}$, $Plan_{predict}$, and $Plan_{FBCT}$.

Variables	$Plan_{4DCT}$ (mean ± SD)	$Plan_{predict}$(mean ± SD)	$Plan_{FBCT}$ (mean ± SD)	*F*	*P*
PTV					
Volume (cm^3)	48.81 ± 38.99	44.84 ± 31.93	43.89 ± 34.05	0.051	0.952
D_{98} (Gy)	45.54 ± 10.72	51.01 ± 6.84	50.88 ± 4.33	0.384	0.685
D_2 (Gy)	77.21 ± 7.03	78.4 ± 14.89	81.93 ± 12.37	1.430	0.259
V_{100} (%)	94.40 ± 1.80	94.9 ± 0.28	94.9 ± 0.25	0.931	0.408
D_{95} (Gy)	59.76 ± 5.76	59.98 ± 3.62	59.59 ± 7.82	0.662	0.525
$R_{100\%}$	1.05 ± 0.11	1.08 ± 0.05	1.12 ± 0.06	0.141	0.872
$R_{50\%}$	3.48 ± 0.82	5.25 ± 2.01	4.45 ± 0.70	0.560	0.598
GI	3.31 ± 0.89	3.98 ± 0.42	4.74 ± 1.01	0.573	0.592
D_{2cm} (Gy)	30.27 ± 4.39	31.18 ± 3.46	33.46 ± 2.24	0.070	0.933
OARs					
Lung					
V12.5 (%)	5.83 ± 1.48	5.30 ± 0.85	5.50 ± 0.99	0.420	0.690
V20 (%)	3.05 ± 0.92	2.58 ± 0.71	2.80 ± 0.57	0.820	0.520
Heart					
D_{max} (Gy)	11.15 ± 6.58	9.84 ± 5.89	9.91 ± 6.42	0.058	0.994
Trachea					
D_{max} (Gy)	0.45 ± 0.04	0.55 ± 0.07	0.58 ± 0.06	0.681	0.791
Esophagus					
D_{max} (Gy)	9.14 ± 3.42	8.43 ± 2.14	7.01 ± 2.61	0.457	0.647
Spinal Cord					
D_{max} (Gy)	9.57 ± 2.61	8.89 ± 0.01	6.70 ± 2.45	0.926	0.431
Treatment efficiency parameters					
Generate ITV time (min)	30.28 ± 3.74	1.45 ± 0.31	24.13 ± 4.93	756	0.000*
MU	1,023.31 ± 83.61	1,059.14 ± 92.38	1,042.36 ± 97.25	0.837	0.713

**A significant difference existed ($p < 0.05$).*
SD, standard deviation.

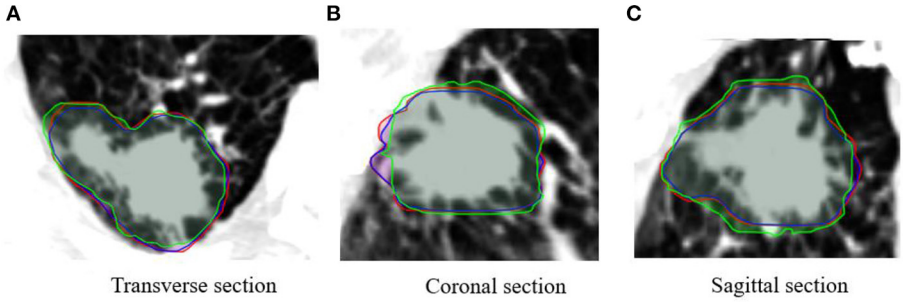

FIGURE 4 | Comparison of contours between planning target volume on four-dimensional CT (PTV_{4DCT}) (red), $PTV_{predict}$ (blue), and PTV_{FBCT} (green) in different views of patient images (projected $PTV_{predict}$ and PTV_{FBCT} on 4DCT images).

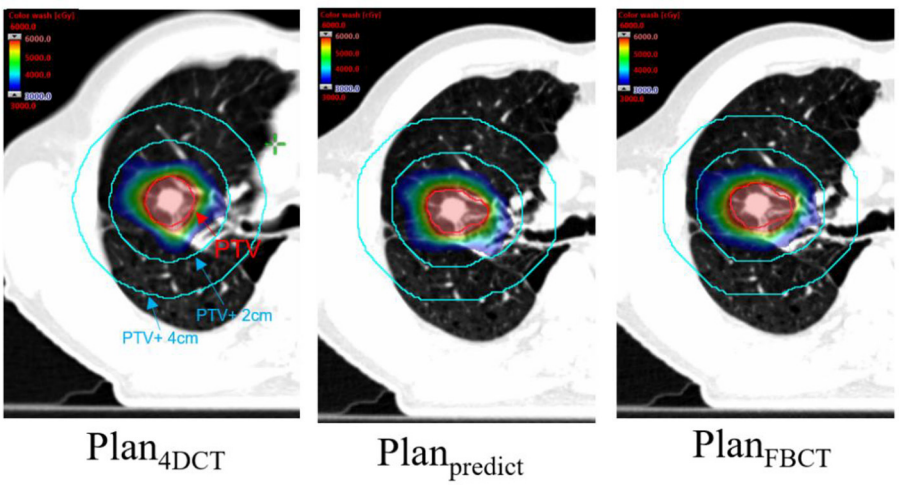

FIGURE 5 | Dose color wash map of each plan for PTV (the area of the red line is PTV, and the red and blue color wash maps represent the 100 and 50% prescription doses, respectively).

The contouring in different views is also similar for the PTV of each plan (**Figure 4**). The visual assessment shows that the results are reasonable. **Figures 5**, **6** show the dose color wash and the dose volume histogram (DVH) of the patients, which were calculated for each plan.

OARs and Treatment Efficiency Evaluation

The estimated dosimetric parameters for all plans met the criteria specified in the RTOG 0813 protocol. The results show that the V20 of $Plan_{predict}$ was better than that of $Plan_{4DCT}$ (with a reduction of 15.4%) and $Plan_{FBCT}$ (with a reduction of 7.8%), which obviously protected the lung (**Figure 7**), but no difference existed in the dosimetric parameters. Additionally, there was no statistically significant difference in each plan with the maximum dose (D_{max}) for the heart, esophagus, and spinal cord (**Table 4**).

The average times of generating ITVs were 1.45 ± 0.31 and 30.28 ± 3.74 min for automatic (by model) and manual ITVs, respectively. The model helped reduce it by 95% of the time on average ($p = 0.000$).

DISCUSSION

Geometrical Accuracy of the Prediction Model

We used geometric parameters to evaluate the similarity between ITV_{4DCT} and $ITV_{predict}$. The DSC value was 0.83 ± 0.18, showing good agreement between the $ITV_{predict}$ and ITV_{4DCT} contour. DSC > 0.7 is considered to be in good agreement with the gold standard (28, 29). This result can become the basis for follow-up research as it confirmed the feasibility of CBCT images to predict ITV. The accuracy of the ITV contours will directly affect the optimization and calculation of the DVH plan (30). The results show that the $ITV_{predict}$ volume is 10% smaller than the ITV_{4DCT} volume. Dou et al. found that 4DCT images should be used with caution for patients with highly irregular breathing. The simulation indicates that low-pitch helical 4DCT processes potentially yield large tumor motion measurement errors and overestimate tumor motion (31). However, there was no significant difference between ITV_{4DCT} and $ITV_{predict}$. This result of $ITV_{predict}$ volume reduction was acceptable.

4DCT Limitation and CBCT Potential Application

Four dimensional CT (4DCT) was acquired during the patient positioning stage and could not reflect the deviation of the tumor's respiratory movement during treatment (32, 33). Rabinowitz et al. showed that during the patient positioning stage and the treatment stage, the tumor deviation caused by respiratory motion was an average of 5.1 mm. For thoracic tumors, the tumor deviation caused by respiratory motion could reach 5.8 mm (34). Yang et al. showed that 4DCT could only collect signals of a limited number of respiratory phases, but the patient's breathing may change at any time and cannot accurately reflect the patient's tumor movement during treatment (35). The study showed that CBCT and maximum intensity projection (MIP) images are equivalent in determining the location of ITV (36). Li et al. showed that 4DCT and CBCT images can indicate variations and inter-fractional setup displacement (37). In our study, we used the first CBCT image at the time of treatment to show the respiratory motion range of a tumor during treatment.

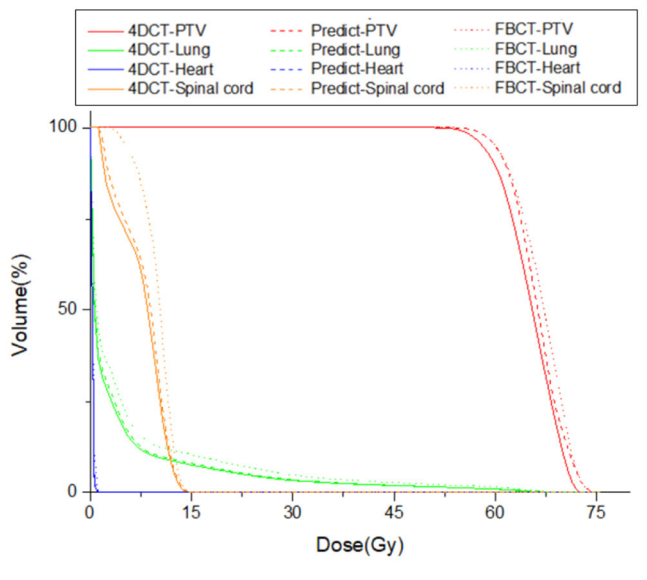

FIGURE 6 | Comparison of the dose-volume histogram (DVH) of one case under the 3 treatment plans: $Plan_{4DCT}$ (solid line), $Plan_{predict}$ (dash line), and $Plan_{FBCT}$ (dot line).

PTV Evaluation

Compared with conventional treatment, SBRT has higher requirements for the PTV dose gradient and limited dose limitation of organ at risks (OARs). The RTOG0813 protocol (26) provides guidance on the acceptable values of $R_{100\%}$, $R_{50\%}$, and D_{2cm} based on the PTV volume. The values of $R_{100\%}$ and $R_{50\%}$ in $Plan_{4DCT}$ and $Plan_{predict}$ were 1.05 ± 0.11, 1.08 ± 0.05, 3.48 ± 0.82, and 5.25 ± 2.01 respectively, which were comparable but slightly different. The result of $R_{100\%}$ indicated that $Plan_{predict}$ and $Plan_{4DCT}$ have a similar dose coverage for PTV. The $R_{50\%}$ value of $Plan_{predict}$ increased by nearly 30% compared with that of $Plan_{4DCT}$, which shows that $Plan_{4DCT}$ has a stronger ability to constrain PTV. All plans meet the RTOG0813 protocol and can be used in clinical practice.

In our study, we also researched the dosimetry of GTV on FBCT images to generate $Plan_{FBCT}$. The results show that after the RTOG0813 guide on margin from GTV, it can also meet the treatment standards and hence can be used for treatment planning. Tian et al. (38) compared the treatment planning and dose calculation of average intensity projection (AIP) and FBCT for SBRT and concluded that the dosimetric of the two datasets were similar.

OARs Evaluation

In addition, the RTOG 0813 agreement contains restrictions on each OAR, such as the lung and spinal cord. For SBRT patients, high-energy rays inevitably pass through a part of normal lung tissue, which affects lung function. Jin et al. (39) found that when V20 was >25%, the incidence of radiation pneumonia significantly increased. Our results show that the mean values of V20 in $Plan_{4DCT}$ and $Plan_{predict}$ were $3.05 \pm 0.92\%$ and $2.58 \pm 0.71\%$, respectively, indicating that $Plan_{predict}$ reduces the volume of radiation received by normal lung tissue, thereby reducing the incidence of radiation pneumonitis. The thoracic spinal cord is

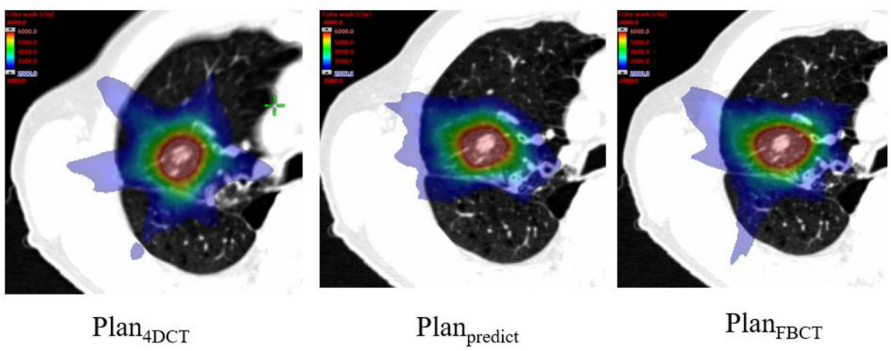

FIGURE 7 | Comparison of color-wash dose distributions of the plans (blue color wash map represents a 20-Gy dose).

more likely to be injured in patients with lung SBRT. Radiation myelitis may cause serious consequences, such as paraplegia and respiratory paralysis. The values of spinal cord D_{max} in the $Plan_{4DCT}$ and the $Plan_{predict}$ are 9.57 ± 2.61 Gy and 8.89 ± 0.01 Gy. This shows that $Plan_{predict}$ reduces the maximum dose received in the spinal cord, which can thus reduce the incidence of radiation myelitis.

Treatment Efficiency Evaluation

Currently, the model predicts the ITV volume of large tumors more accurately. For patients with lung SBRT, this model can generate ITV on CBCT images in, on average, 1.45 ± 0.31 min. In our research, the generated ITV time of $Plan_{predict}$ was significantly reduced by nearly 95% compared with that of $Plan_{4DCT}$. Hence, Using the model to input CBCT images can greatly shorten the time to collect patient images and significantly increase the efficiency of tumor delineation for physicians.

This study has some limitations. Firstly, the number of patient samples included was small. Therefore, our patient data did not represent the whole spectrum. Tumors of different sizes and different locations should be included in the future. This could increase patient data for a more uniform tumor volume distribution to ensure accuracy of results in the future.

CONCLUSION

Geometric results show that self-generated $ITV_{predict}$ has a good correlation with ITV_{4DCT}, although the $ITV_{predict}$ volume is 10% smaller than the ITV_{4DCT} volume. This work confirmed the feasibility of the clinical application of $ITV_{predict}$ to make a treatment plan. There were no significant dosimetry differences between $Plan_{4DCT}$ and $Plan_{predict}$. Thus, our model has potential application in institutions with or without 4DCT scanning technology or when patient breathing is irregular.

DATA AVAILABILITY STATEMENT

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

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

SZ and BL: conception and design. XZ and YL: acquisition of data. WG, LZ, and FM: analysis of data. SZ, BL, and JQ: writing, review and/or revision of the manuscript. All authors reviewed, read, and approved the final manuscript and authors contributed to the article, and approved the submitted version.

FUNDING

This work was partially supported by the National Natural Science Foundation of China (Grant No.11505029 and No.81472794), Shanghai Municipal Commission of Health (Grant Nos. 2018BR23 and 20184Y0099), and Shanghai Municipal Science and Technology Commission (Grant No. 18441904400).

REFERENCES

1. Fakiris AJ, McGarry RC, Yiannoutsos CT, Papiez L, Williams M, Henderson MA, et al. Stereotactic body radiation therapy for early-stage non-small-cell lung carcinoma: four-year results of a prospective phase II study. *Int J Radiat Oncol Biol Phys*. (2009) 75:677–82. doi: 10.1016/j.ijrobp.2008.11.042
2. Onishi H, Shirato H, Nagata Y, Hiraoka M, Fujino M, Gomi K, et al. Hypofractionated stereotactic radiotherapy (HypoFXSRT) for stage I non-small cell lung cancer: updated results of 257 patients in a Japanese multi-institutional study. *J Thorac Oncol*. (2007) 2(7 Suppl 3):S94–100. doi: 10.1097/JTO.0b013e318074de34
3. Uematsu M, Shioda A, Suda A, Fukui T, Ozeki Y, Hama Y, et al. Computed tomography-guided frameless stereotactic radiotherapy for stage I non-small cell lung cancer: a 5-year experience. *Int J Radiat Oncol Biol Phys*. (2001) 51:666–70. doi: 10.1016/S0360-3016(01)01703-5
4. Qiu JJ, Ge W, Zhang L, Yao Y, Zheng X. The feasibility and efficiency of volumetric modulated arc therapy-based breath control stereotactic body radiotherapy for liver tumors. *Technol Cancer Res Treat*. (2016) 15:674–82. doi: 10.1177/1533034615596273
5. Clements N, Kron T, Franich R, Dunn L, Roxby P, Aarons Y, et al. The effect of irregular breathing patterns on internal target volumes in four-dimensional CT and cone-beam CT images in the context of stereotactic lung radiotherapy. *Med Phys*. (2013) 40:021904. doi: 10.1118/1.4773310
6. Rietzel E, Pan T, Chen GT. Four-dimensional computed tomography: image formation and clinical protocol. *Med Phys*. (2005) 32:874–89. doi: 10.1118/1.1869852
7. Watkins WT, Li R, Lewis J, Park JC, Sandhu A, Jiang SB, et al. Patient-specific motion artifacts in 4DCT. *Med Phys*. (2010) 37:2855–61. doi: 10.1118/1.3432615
8. de Oliveira Duarte S, Rancoule C, He MY, Baury M, Sotton S, Vallard A, et al. Use of 4D-CT for radiotherapy planning and reality in France: data from a national survey. *Cancer Radiother*. (2019) 23:395–400. doi: 10.1016/j.canrad.2019.02.006
9. Feuvret L, Noël G, Mazeron JJ, Bey P. Conformity index: a review. *Int J Radiat Oncol Biol Phys*. (2006) 64:333–42. doi: 10.1016/j.ijrobp.2005.09.028
10. Barney BM, Lee RJ, Handrahan D, Welsh KT, Cook JT, Sause WT. Image-guided radiotherapy (IGRT) for prostate cancer comparing kV imaging of fiducial markers with cone beam computed tomography (CBCT). *Int J Radiat Oncol Biol Phys*. (2011) 80:301–5. doi: 10.1016/j.ijrobp.2010.06.007
11. Fave X, Mackin D, Yang J, Zhang J, Fried D, Balter P, et al. Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer? *Med Phys*. (2015) 42:6784–97. doi: 10.1118/1.4934826
12. Zhang H, Ouyang L, Ma J, Huang J, Chen W, Wang J. Noise correlation in CBCT projection data and its application for noise reduction in low-dose CBCT. *Med Phys*. (2014) 41:031906. doi: 10.1118/1.4865782
13. Wang C, Hunt M, Zhang L, Rimner A, Yorke E, Lovelock M, et al. Technical Note: 3D localization of lung tumors on cone beam CT projections *via* a convolutional recurrent neural network. *Med Phys*. (2020) 47:1161–6. doi: 10.1002/mp.14007

14. Cao H, Liu H, Song E, Ma G, Xu X, Jin R, et al. A two-stage convolutional neural networks for lung nodule detection. *IEEE J Biomed Health Inform*. (2020) 24:2006–15. doi: 10.1109/JBHI.2019.2963720
15. Gu Y, Lu X, Yang L, Zhang B, Yu D, Zhao Y, et al. Automatic lung nodule detection using a 3D deep convolutional neural network combined with a multi-scale prediction strategy in chest CTs. *Comput Biol Med*. (2018) 103:220–31. doi: 10.1016/j.compbiomed.2018.10.011
16. Liu C, Hu SC, Wang C, Lafata K, Yin FF. Automatic detection of pulmonary nodules on CT images with YOLOv3: development and evaluation using simulated and patient data. *Quant Imaging Med Surg*. (2020) 10:1917–29. doi: 10.21037/qims-19-883
17. He K, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. *IEEE Trans Pattern Anal Mach Intell*. (2020) 42:386–97. doi: 10.1109/TPAMI.2018.2844175
18. Bouget D, Jørgensen A, Kiss G, Leira HO, Langø T. Semantic segmentation and detection of mediastinal lymph nodes and anatomical structures in CT data for lung cancer staging. *Int J Comput Assist Radiol Surg*. (2019) 14:977–86. doi: 10.1007/s11548-019-01948-8
19. Zhang J, Jiang D, Su H, Dai Z, Dai J, Liu H, et al. Dosimetric comparison of different algorithms in stereotactic body radiation therapy (SBRT) plan for non-small cell lung cancer (NSCLC). *Onco Targets Ther*. (2019) 12:6385–91. doi: 10.2147/OTT.S201473
20. Chiao JY, Chen KY, Liao KY, Hsieh PH, Zhang G, Huang TC. Detection and classification the breast tumors using mask R-CNN on sonograms. *Medicine*. (2019) 98:e15200. doi: 10.1097/MD.0000000000015200
21. Jeong J, Lei Y, Kahn S, Liu T, Curran WJ, Shu HK, et al. Brain tumor segmentation using 3D Mask R-CNN for dynamic susceptibility contrast enhanced perfusion imaging. *Phys Med Biol*. (2020) 65:185009. doi: 10.1088/1361-6560/aba6d4
22. Zhen Li, Shujun Zhang, Libo Zhang, Ya Li, Xiangpeng Zheng, Jie Fu, et al. Deep learning-based Internal Target Volume (ITV) prediction using cone-beam CT Images in lung stereotactic body radiotherapy. *Technol Cancer Res Treat*. (2022) 21:1–10. doi: 10.1177/15330338211073380
23. Jones E, Oliphant T, Peterson P. SciPy: Open Source Scientific Tools for Python. (2001).
24. Desai DD, Cordrey IL, Johnson EL. A physically meaningful relationship between R50% and PTV surface area in lung SBRT. *J Appl Clin Med Phys*. (2020) 21:47–56. doi: 10.1002/acm2.12964
25. Paddick I, Lippitz B. A simple dose gradient measurement tool to complement the conformity index. *J Neurosurg*. (2006) 105(Suppl):194–201. doi: 10.3171/sup.2006.105.7.194
26. Menon SV, Paramu R, Bhasi S, Nair RK. Evaluation of plan quality metrics in stereotactic radiosurgery/radiotherapy in the treatment plans of arteriovenous malformations. *J Med Phys*. (2018) 43:214–20. doi: 10.4103/jmp.JMP_25_18
27. Bezjak A, Paulus R, Gaspar LE, Timmerman RD, Straube WL, Ryan WF, et al. Safety and efficacy of a five-fraction stereotactic body radiotherapy schedule for centrally located non-small-cell lung cancer: NRG oncology/RTOG 0813 trial. *J Clin Oncol*. (2019) 37:1316–25. doi: 10.1200/JCO.18.00622
28. Gaede S, Olsthoorn J, Louie AV, Palma D, Yu E, Yaremko B, et al. An evaluation of an automated 4D-CT contour propagation tool to define an internal gross tumour volume for lung cancer radiotherapy. *Radiother Oncol*. (2011) 101:322–8. doi: 10.1016/j.radonc.2011.08.036
29. Eldesoky AR, Yates ES, Nyeng TB, Thomsen MS, Nielsen HM, Poortmans P, et al. Internal and external validation of an ESTRO delineation guideline - dependent automated segmentation tool for loco-regional radiation therapy of early breast cancer. *Radiother Oncol*. (2016) 121:424–30. doi: 10.1016/j.radonc.2016.09.005
30. Cao M, Stiehl B, Yu VY, Sheng K, Kishan AU, Chin RK, et al. Analysis of geometric performance and dosimetric impact of using automatic contour segmentation for radiotherapy planning. *Front Oncol*. (2020) 10:1762. doi: 10.3389/fonc.2020.01762
31. Dou TH, Thomas DH, O'Connell D, Bradley JD, Lamb JM, Low DA. Technical note: simulation of 4DCT tumor motion measurement errors. *Med Phys*. (2015) 42:6084–9. doi: 10.1118/1.4931416
32. Shah AP, Kupelian PA, Waghorn BJ, Willoughby TR, Rineer JM, Mañon RR, et al. Real-time tumor tracking in the lung using an electromagnetic tracking system. *Int J Radiat Oncol Biol Phys*. (2013) 86:477–83. doi: 10.1016/j.ijrobp.2012.12.030
33. Purdie TG, Moseley DJ, Bissonnette JP, Sharpe MB, Franks K, Bezjak A, et al. Respiration correlated cone-beam computed tomography and 4DCT for evaluating target motion in stereotactic lung radiation therapy. *Acta Oncol*. (2006) 45:915–22. doi: 10.1080/02841860600907345
34. Rabinowitz I, Broomberg J, Goitein M, McCarthy K, Leong J. Accuracy of radiation field alignment in clinical practice. *Int J Radiat Oncol Biol Phys*. (1985) 11:1857–67. doi: 10.1016/0360-3016(85)90046-X
35. Yang M, Timmerman R. Stereotactic ablative radiotherapy uncertainties: delineation, setup and motion. *Semin Radiat Oncol*. (2018) 28:207–17. doi: 10.1016/j.semradonc.2018.02.006
36. Wang L, Chen X, Lin MH, Xue J, Lin T, Fan J, et al. Evaluation of the cone beam CT for internal target volume localization in lung stereotactic radiotherapy in comparison with 4D MIP images. *Med Phys*. (2013) 40:111709. doi: 10.1118/1.4823785
37. Li Y, Ma JL, Chen X, Tang FW, Zhang XZ. 4DCT and CBCT based PTV margin in Stereotactic Body Radiotherapy(SBRT) of non-small cell lung tumor adhered to chest wall or diaphragm. *Radiat Oncol*. (2016) 11:152. doi: 10.1186/s13014-016-0724-5
38. Tian Y, Wang Z, Ge H, Zhang T, Cai J, Kelsey C, et al. Dosimetric comparison of treatment plans based on free breathing, maximum, and average intensity projection CTs for lung cancer SBRT. *Med Phys*. (2012) 39:2754–60. doi: 10.1118/1.4705353
39. Jin H, Tucker SL, Liu HH, Wei X, Yom SS, Wang S, et al. Dose-volume thresholds and smoking status for the risk of treatment-related pneumonitis in inoperable non-small cell lung cancer treated with definitive radiotherapy. *Radiother Oncol*. (2009) 91:427–32. doi: 10.1016/j.radonc.2008.09.009

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.

frontiers | Frontiers in Medicine

ORIGINAL RESEARCH
published: 11 April 2022
doi: 10.3389/fmed.2022.870523

Artificial Intelligence-Enabled Electrocardiogram Predicted Left Ventricle Diameter as an Independent Risk Factor of Long-Term Cardiovascular Outcome in Patients With Normal Ejection Fraction

*Hung-Yi Chen[1], Chin-Sheng Lin[2], Wen-Hui Fang[3,4,5], Chia-Cheng Lee[6,7], Ching-Liang Ho[8], Chih-Hung Wang[9,10] and Chin Lin[5,11]**

[1] Department of Internal Medicine, National Defense Medical Center, Tri-Service General Hospital, Taipei, Taiwan, [2] Division of Cardiology, Department of Internal Medicine, National Defense Medical Center, Tri-Service General Hospital, Taipei, Taiwan, [3] Department of Family and Community Medicine, National Defense Medical Center, Tri-Service General Hospital, Taipei, Taiwan, [4] Department of Internal Medicine, National Defense Medical Center, Tri-Service General Hospital, Taipei, Taiwan, [5] Artificial Intelligence of Things Center, National Defense Medical Center, Tri-Service General Hospital, Taipei, Taiwan, [6] Medical Informatics Office, National Defense Medical Center, Tri-Service General Hospital, Taipei, Taiwan, [7] Division of Colorectal Surgery, Department of Surgery, National Defense Medical Center, Tri-Service General Hospital, Taipei, Taiwan, [8] Division of Hematology and Oncology, Department of Internal Medicine, National Defense Medical Center, Tri-Service General Hospital, Taipei, Taiwan, [9] Department of Otolaryngology-Head and Neck Surgery, National Defense Medical Center, Tri-Service General Hospital, Taipei, Taiwan, [10] National Defense Medical Center, Graduate Institute of Medical Sciences, Taipei, Taiwan, [11] Medical Technology Education Center, National Defense Medical Center, School of Medicine, Taipei, Taiwan

OPEN ACCESS

Edited by:
Ming-Chin Lin,
Taipei Medical University, Taiwan

Reviewed by:
Chia-Ter Chao,
National Taiwan University Hospital, Taiwan
Belal Alsinglawi,
Western Sydney University, Australia

***Correspondence:**
Chin Lin
xup6fup0629@gmail.com

Specialty section:
This article was submitted to Family Medicine and Primary Care, a section of the journal Frontiers in Medicine

Received: 07 February 2022
Accepted: 10 March 2022
Published: 11 April 2022

Citation:
Chen H-Y, Lin C-S, Fang W-H, Lee C-C, Ho C-L, Wang C-H and Lin C (2022) Artificial Intelligence-Enabled Electrocardiogram Predicted Left Ventricle Diameter as an Independent Risk Factor of Long-Term Cardiovascular Outcome in Patients With Normal Ejection Fraction. Front. Med. 9:870523. doi: 10.3389/fmed.2022.870523

Background: Heart failure (HF) is a global disease with increasing prevalence in an aging society. However, the survival rate is poor despite the patient receiving standard treatment. Early identification of patients with a high risk of HF is important but challenging. Left ventricular end-diastolic diameter (LV-D) increase was an independent risk factor of HF and adverse cardiovascular (CV) outcomes. In this study, we aimed to develop an artificial intelligence (AI) enabled electrocardiogram (ECG) system to detect LV-D increase early.

Objective: We developed a deep learning model (DLM) to predict left ventricular end-diastolic and end-systolic diameter (LV-D and LV-S) with internal and external validations and investigated the relationship between ECG-LV-D and echocardiographic LV-D and explored the contributions of ECG-LV-D on future CV outcomes.

Methods: Electrocardiograms and corresponding echocardiography data within 7 days were collected and paired for DLM training with 99,692 ECGs in the development set and 20,197 ECGs in the tuning set. The other 7,551 and 11,644 ECGs were collected from two different hospitals to validate the DLM performance in internal and external validation sets. We analyzed the association and prediction ability of ECG-LVD for CV

outcomes, including left ventricular (LV) dysfunction, CV mortality, acute myocardial infarction (AMI), and coronary artery disease (CAD).

Results: The mean absolute errors (MAE) of ECG-LV-D were 5.25/5.29, and the area under the receiver operating characteristic (ROC) curves (AUCs) were 0.8297/0.8072 and 0.9295/0.9148 for the detection of mild ($56 \leqq$ LV-D < 65 mm) and severe (LV-D $\geqq 65$ mm) LV-D dilation in internal/external validation sets, respectively. Patients with normal ejection fraction (EF) who were identified as high ECHO-LV-D had the higher hazard ratios (*HR*s) of developing new onset LV dysfunction [*HR*: 2.34, 95% conference interval (*CI*): 1.78–3.08], CV mortality (*HR* 2.30, 95% *CI* 1.05–5.05), new-onset AMI (*HR* 2.12, 95% *CI* 1.36–3.29), and CAD (*HR* 1.59, 95% *CI* 1.26–2.00) in the internal validation set. In addition, the ECG-LV-D presents a 1.88-fold risk (95% *CI* 1.47–2.39) on new-onset LV dysfunction in the external validation set.

Conclusion: The ECG-LV-D not only identifies high-risk patients with normal EF but also serves as an independent risk factor of long-term CV outcomes.

Keywords: artificial intelligence, electrocardiogram, deep learning, heart failure, ejection fraction, left ventricular end-diastolic diameter, cardiovascular outcome

INTRODUCTION

Heart failure (HF) is a common clinical entity with increasing prevalence in an aging society, which affects 5.7 million patients and more than 870,000 new cases are diagnosed in the United States every year (1). In developed countries, about 2% of the population lives with HF (1, 2). The American Heart Association forecasted that total costs associated with HF were at $20.9 billion in 2012 and are projected to rise to $53.1 billion by 2030 (3). Currently, HF is classified as reduced ejection fraction (HFrEF), mildly reduced ejection fraction (HFmrEF), and preserved ejection fraction (HFpEF) based on different ejection fraction (EF) levels (4). Multiple modality treatment for the patients with HF, such as renin-angiotensin system inhibition, beta-blocker, and aldosterone antagonist, is evidence-based and recommended in guidelines (4, 5). However, even with treatment, the HF survival rate remains poor globally and the mortality ranged from 17 to 45% in a year among the patients who were admitted to a hospital because of HF (1, 2, 6). Such evidence points out the significant problem of HF in aged society. Early identification of those patients who are at risk to develop HF and adequate risk reduction helps to improve the quality of life, reduce hospitalization, and promote survival outcomes.

In patients with HF, there were several important parameters for the assessment of cardiac functional and structural changes. As EF was the ratio of blood leaving heart each time it contracts, the left ventricular end-diastolic diameter (LV-D) and end-systolic diameter (LV-S) influenced the value of ECHO-EF. The principal ECG changes in patients with increased LV-D and LV-S in LV hypertrophy include augmented QRS amplitude, prolonged QRS conduction time, changes in instantaneous and mean QRS vectors, ST depression and/or T-wave inversion, and P-wave abnormalities, such as left atrial enlargement (7, 8). VF frequency was consistently lower in patients with an increased LV diameter (9). However, these ECG changes were neither sensitive nor specific for increased LV-D or LV-S detection. The EF serves as an indicator for cardiac contractility and a significant predictor of survival (10–13). Previous studies presented that LV-D increase was an independent risk factor of cardiovascular outcomes (14, 15), ventricular arrhythmia inducibility (16), and mortality (17, 18). By the investigation of 1,138 patients with HFrEF and sinus rhythm, Ito et al. proposed strong association between LV diameters and cardiovascular (CV) outcomes, which is independent of ECHO-EF (14). Moreover, in a combination with QRS duration, the LV-D could be applied to identify the patients at risk for tachyarrhythmias. Makaryus et al. revealed myocardial infarction with scar formation or cardiomyopathy with disordered ventricular excitation accounts for the ventricular arrhythmia and poor prognosis in patients with dilated LV-D (16). In patients with mitral regurgitation, the LV-S increase is independently associated with increased mortality even under medical management (19). All the results highlight the significance of EF, LV-D, and LV-S in patients with HF.

Artificial intelligence-based ECG (AI-ECG) has expanded to multiple applications and achieved human-level performance, effectively detecting cardiac diseases with large annotated ECG datasets, including echocardiogram predictions (20, 21), arrhythmia detection (22), dyskalemia and its cause (23–25), glycated hemoglobin (26), digoxin toxicity (27), aortic dissection (28), pneumothorax (29), and myocardial infarction (30–32). Importantly, previous studies revealed significant correlation and predictability between ECG-predicted EF (ECG-EF) and echocardiographic EF (ECHO-EF). This study not only revealed the diagnostic value of ECG on HF but also further identified a new subtype of HF, which has normal ECHO-EF but lower ECG-EF and a high risk of future LV dysfunction (20). Meanwhile, age estimated from ECG (ECG-age) is also a measure of cardiovascular health, and the difference between the ECG-age and the chronological age can be used as a marker of the risk of

deaths even in different cohorts (33). The new concept of disease previvor was proposed as individuals who are healthy but have a markedly increased predisposition to develop the disease (34, 35).

However, the discrepancy between ECG-EF and ECHO-EF was not fully interpreted. ECHO-EF is evaluated regularly in echocardiography, similar to other cardiac structure measurements, such as LV-D, LV-S, interventricular, and posterior wall thickness. With the aid of AI-ECG, we hypothesized that AI-ECG predicting LV-D (ECG-LV-D) may provide additional information on CV outcomes in patients with initially normal ECHO-EF, who are recognized as low ECG-EF. Therefore, the aim of this study is to build a deep learning model (DLM) to predict LV-D and LV-S and verify the accuracy by echocardiography in two independent hospitals. Finally, we tried to apply ECG-LV-D in different clinical scenarios and acquire additional information on the prediction of future CV diseases.

MATERIALS AND METHODS

Data Source and Population

This multicenter retrospective study was ethically approved by the institutional review board of Tri-Service General Hospital, Taipei, Taiwan (IRB NO. C202105049). The electronic medical records (EMRs) of our hospital included digital ECG signals, echocardiography images, hospital courses records, and future outcomes between 1 January 2010 and 31 September 2021. We identified patients who had at least one pair of 12-lead ECG and transthoracic echocardiography (TTE) records within 7 days. Subjects with inadequate ECG or echocardiographic information were excluded, such as noise interference, leads dislodge or dislocation, data loss of heart rate, EF, LV-D, or LV-S. The remaining ECGs were annotated by TTE information collected in this study. Finally, there were 75,942 patients in NeiHu General Hospital at NeiHu District (hospital A), an academic medical center in our hospital system, and 11,633 patients in Tingzhou Branch Hospital at Zhongzheng District (hospital B), a community hospital (**Figure 1**).

We divided ECGs into development, tuning, internal validation, and external validation sets by different dates and hospitals. For DLM training, there were 99,692 ECGs from 60,790 patients included in development set and 20,197 ECGs from 7,601 patients were included in tuning set. We only used the first records in the validation step for the patients with multiple ECG-TTE pairs, and the internal and external validation sets included 7,551 ECGs before 31 December 2015 in hospital A and 11,644 ECGs in hospital B. No repeated patients were recruited into more than one group.

Observational Variables

The ECGs were acquired at a sampling rate of 500 Hz with a 10-s period using a Philips 12-lead ECG machine (PH080A, Philips Medical Systems, 3000 Minuteman Road Andover, MA 01810 United States). Comprehensive 2D ECG and quantitative data were recorded at the time of the acquisition in a Philips image system for all patients. The LV parameters included EF, LV-D, and LV-S, which were routinely acquired by experienced cardiologists or technicians using standardized methods. The EF was assessed using the Simpson method, M-mode, and the reported visually estimated EF. LV dimensions and wall thickness were measured by M mode under para-sternal long axis view and recorded by millimeter. The cut-off values of EF are 50 and 35% as mild and severe LV dysfunction, which are comparable criteria described in previous studies (20, 34, 36, 37). We selected LV-D and LV-S as they can be measured more easily and are more reproducible than other indices. Patients were divided into three groups according to LV-D at initial echocardiography: $\leq$ 56 mm (normal), 56 < LV-D $\leq$ 65 mm (mild increase), and > 65 mm (severe increase). The criteria for LV-S were $\leq$ 38 mm (normal), 38 < LV-S $\leq$ 45 mm (mild increase), and > 45 mm (severe increase). These diameters were determined according to the reference values for LV size from studies based on ethnic-appropriate population datasets (18, 38–43).

The demographic characteristics were obtained in our EMRs and disease history before the index date of ECG was collected using the corresponding code of International Classification of Disease, Ninth Revision and Tenth Revision (ICD-9 and ICD-10, respectively), as described previously (24, 26, 32, 44). The remaining echocardiographic parameters, such as interventricular septum (IVS) diameter, left ventricular posterior wall (LVPW) diameter, left atrium (LA) size, aortic root (AO) diameter, right ventricular (RV) diameter, pulmonary artery systolic pressure (PASP), and pericardial effusion (PE), were also collected in this study.

According to the promising ability of disease previvor identification by AI-ECG, we analyzed the correlation between ECG-LV-D increased and new-onset LV dysfunction, defined as ECHO-EF $\leq$ 35. Moreover, patients' data were censored at the last known TTE examination to limit bias from incomplete records. In addition to LV dysfunction, we followed and analyzed other three CV outcomes, including CV mortality, new-onset acute myocardial infarction (AMI), and new-onset coronary artery disease (CAD). CV mortality included arrhythmia-related death, acute coronary syndrome-related death, stroke death, and HF-related death. These outcomes were censored at the patient's last known hospital alive encounter without corresponding events to limit bias from incomplete records. The end of follow-up in this study was 30 September 2021 for all the above outcomes.

The Implementation of the Deep Learning Model

The ECG-based EF, LV-D, and LV-S were, respectively, considered as function score and structure status of the heart, both estimated by DLMs. The ECG12Net architecture with 82 convolutional layers and an attention mechanism was used for estimation and the technology details, such as model architecture, data augmentation, and model visualization, were described previously (24). We used an oversampling process to adequately recognize extreme EF, LV-D, and LV-S values. The process was based on weights computed based on the prevalence of 20 equidistant intervals in the development set. The output of these DLMs was a continuous estimation value of actual EF, LV-D,

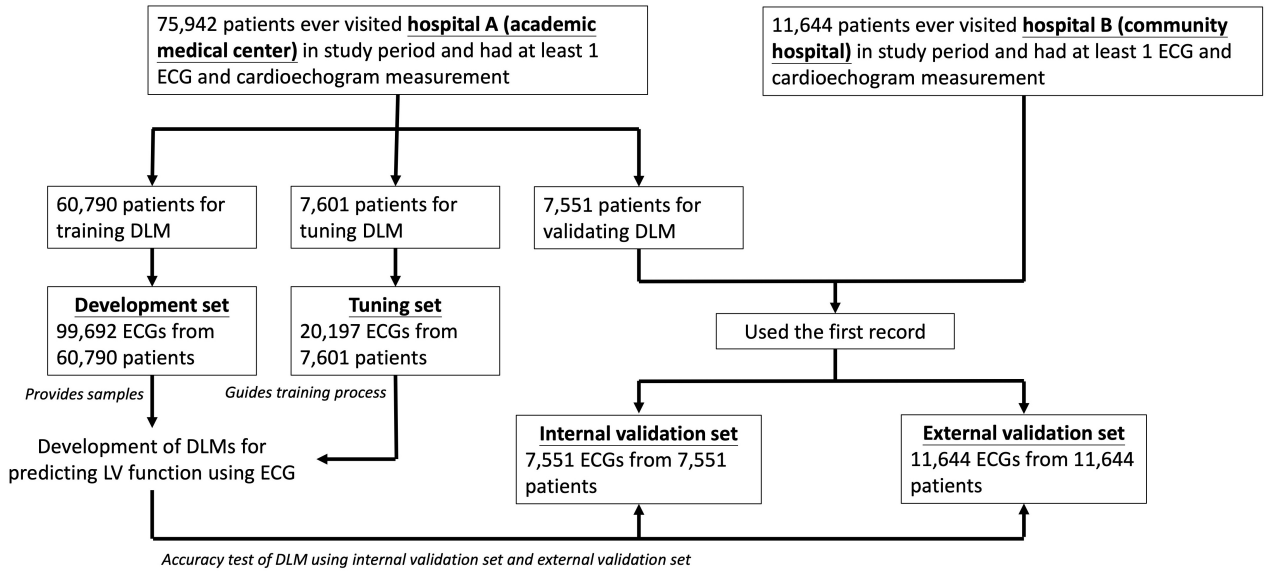

FIGURE 1 | Development, tuning, internal validation, and external validation sets generation and the ECG labeling of echocardiogram. Schematic of the dataset creation and analysis strategy, which was devised to assure a robust and reliable dataset for training, validating, and testing of the network. Once a patient's data were placed in one of the datasets, that individual's data were used only in that set, avoiding "cross-contamination" among the training, validation, and test datasets. The details of the flowchart and how each of the datasets was used are described in "Materials and Methods" section.

and LV-S, which was called ECG-EF, ECG-LV-D, and ECG-LV-S, respectively.

Statistical Analysis and Model Performance Assessment

Patient characteristics are presented as numbers of patients, population percentages, means, and standard deviations (SDs), with the significance level set as $p < 0.05$. We used scatter plots to describe the predicted value by ECG voltage-time traces compared with actual EF and left ventricular diameters (LV-D/LV-S). The accuracy of DLMs was evaluated by mean difference (Diff), Pearson's correlation coefficients (r), and mean absolute errors (MAEs), calculated in both internal and external validation sets. The diagnostic value of DLMs was measured with the receiver operating characteristic (ROC) curve and the area under the curve (AUC). The tuning set was used to decide the operating point based on the maximum of Yunden's index, which was calculated for the corresponding sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) in both validation sets. To identify the underlying subtype of patients with no correspondence between ECG-EF and ECHO-EF, the proportion of patients with larger ECHO/ECG-LV-D were explored in diverse ECHO/ECG-EF groups for the disease previvors of future LV dysfunction.

The relationship between traditional ECG features and AI-ECG-based ECG-LV-D was also analyzed. We showed the importance rank of different traditional ECG features, including 31 diagnostic pattern classes and 8 continuous ECG measurements based on an automatic Philips analysis system. These features were used to train an eXtreme gradient boosting (XGB) model to predict ECG-LV-D. To identify the most important ECG features in this analysis, the stepwise program was used and the p-value to enter and to remove were 0.05 and 0.15, respectively.

To investigate the long-term incidence of developing new-onset LV dysfunction, we plotted Kaplan–Meier curves of patients with an initially normal EF (EF > 50%), stratified by ECG-EF, left ventricle (end-diastole) diameter (LV-D), and ECG-LV-D. Multivariable Cox proportional hazard models were used to evaluate the predictive ability of ECG-EF, ECHO-LV-D, and ECG-LV-D adjusted by gender and age on all outcomes of interested, presenting in hazard ratios (*HR*s) and 95% conference intervals (95% *CI*s). We assessed the risk of adverse CV outcomes in patients with different ECG-EF/ECG-LV-D using a Cox proportional hazard model after adjusting by gender and age and demonstrated the risk matrixes of different outcomes with *HR*s and the concordance statistic (C-index), which were used to quantify their contributions. All the statistical analyses were conducted in R software, version 3.4.4.

RESULTS

The baseline characteristics of patients, including disease histories and echocardiographic data are presented in **Table 1** for the development, tuning, internal validation, and external validation sets. In internal and external validation sets, 3,810 (50.5%) and 5,760 (49.5%) patients were men, and mean age was 63.4 and 65.7 years, respectively. According to disease history, there were 2,248 (29.8%) and 3,612 (31.0%) patients with diabetes mellitus (DM), 3,938 (52.2%) and 6,435 (55.3%) with hypertension (HTN), 3,125 (41.4%) and 5,176 (44.5%) with hyperlipidemia (HLP), 245 (3.2%) and 270 (2.4%) with AMI,

TABLE 1 | Baseline characteristics.

	Development	Tuning	Internal validation	External validation
Demography				
Sex (male)	50,925 (53.6%)	10,600 (52.5%)	3,810 (50.5%)	5,760 (49.5%)
Age (years)	63.8 ± 17.4	68.0 ± 16.3	63.4 ± 16.6	65.7 ± 18.1
BMI (kg/m^2)	24.6 ± 4.4	24.3 ± 4.4	24.5 ± 4.4	24.5 ± 4.3
Disease history				
DM	22,471 (23.6%)	7,211 (35.7%)	2,248 (29.8%)	3,612 (31.0%)
HTN	38,268 (40.3%)	11,778 (58.3%)	3,938 (52.2%)	6,435 (55.3%)
HLP	28,542 (30.0%)	9,088 (45.0%)	3,125 (41.4%)	5,176 (44.5%)
CKD	22,821 (24.0%)	8,820 (43.7%)	1,848 (24.5%)	2,896 (24.9%)
AMI	6,062 (6.4%)	2,099 (10.4%)	245 (3.2%)	279 (2.4%)
STK	13,055 (13.7%)	4,548 (22.5%)	1,274 (16.9%)	2,169 (18.6%)
CAD	26,382 (27.8%)	8,285 (41.0%)	2,358 (31.2%)	3,630 (31.2%)
HF	12,488 (13.1%)	4,777 (23.7%)	957 (12.7%)	1,484 (12.7%)
Afib	6,429 (6.8%)	2,570 (12.7%)	501 (6.6%)	754 (6.5%)
COPD	11,874 (12.5%)	4,372 (21.6%)	1,502 (19.9%)	2,758 (23.7%)
Echocardiography data				
EF (%)	63.6 ± 12.6	61.1 ± 14.2	65.3 ± 11.4	65.5 ± 10.8
LV-D (mm)	47.5 ± 7.1	47.9 ± 7.8	47.3 ± 7.1	47.1 ± 6.8
LV-S (mm)	30.3 ± 6.9	31.2 ± 7.8	29.8 ± 6.7	29.6 ± 6.3
IVS (mm)	11.2 ± 2.6	11.5 ± 2.6	11.2 ± 2.6	11.1 ± 2.6
LVPW (mm)	9.3 ± 1.7	9.5 ± 1.8	9.3 ± 1.7	9.1 ± 1.7
LA (mm)	38.4 ± 7.5	39.6 ± 8.0	38.6 ± 7.6	38.7 ± 7.3
AO (mm)	32.7 ± 4.4	33.1 ± 4.4	32.9 ± 4.5	32.8 ± 4.3
RV (mm)	23.7 ± 4.9	24.2 ± 5.1	24.1 ± 5.0	24.0 ± 5.0
PASP (mmHg)	33.3 ± 11.1	34.8 ± 12.4	32.2 ± 10.4	33.0 ± 10.7
PE (mm)	0.5 ± 2.1	0.6 ± 2.1	0.3 ± 1.8	0.4 ± 1.7

BMI, body mass index; DM, diabetes mellitus; HTN, hypertension; HLP, hyperlipidemia; CKD, chronic kidney disease; AMI, acute myocardial infraction; STK, stroke; CAD, coronary artery disease; HF, heart failure; Afib, atrial fibrillation; COPD, chronic obstructive pulmonary disease; EF, ejection fraction; LV-D, left ventricle (end-diastole); LV-S, left ventricle (end-systole); IVS, Inter-ventricular septum; LVPW, left ventricular posterior wall; LA, left atrium; AO, aortic root; RV, right ventricle; PASP, pulmonary artery systolic pressure; PE, pericardial effusion.

2,358 (31.2%) and 3,630 (31.2%) with CAD, and 957 (12.7%) and 1,484 (12.7%) with HF. The echocardiographic characteristics are similar between internal and external validation sets, such as EF (65.3%/65.5%), LV-D (47.4 mm/47.1 mm), and LV-S (29.8 mm/29.6 mm).

Figure 2 demonstrated the accuracy of DLMs with the scatter plots of ECG-based LV parameters compared to actual ones. The ECG-EF showed a high correlation with the Diff of 1.23 ± 10.52/1.21 ± 10.45, Pearson's correlation coefficients (*r*) of 0.59/0.56, and MAEs of 7.95/7.91 in the internal/external validation set, respectively. Meanwhile, the similar correlation was observed in our analysis of ECG-LV-D and ECG-LV-S, with Diff of 0.03 ± 6.75/0.86 ± 6.27, *r* of 0.53/0.59, and MAE of 5.26/4.83 in the internal validation set, and Diff of 0.06 ± 6.81/0.78 ± 6.40, *r* of 0.49/0.53, and MAE of 5.29/4.93 in the external validation set.

The ROC curve analysis was used to test the diagnostic value of AI-enabled ECG parameters (**Figure 3**). The AUCs of ECG-EF for mild/severe reduced EF in the internal validation set were 0.8793/0.9618, with a percentage of sensitivity of 69.6/86.8, specificity of 89.1/92.5, PPV of 42.4/28.3, and NPV of 96.2/99.5. Meanwhile, the AUCs of ECG-LV-D for detecting mild/severe increased ECHO-LV-D were 0.8297/0.9295 with the percentage of sensitivity of 66.6/80.2, specificity of 82.2/88.1, PPV of 27.4/9.5, and NPV of 96.1/99.7, and the AUCs of ECG-LV-S were 0.8821/0.9471 with the percentage of sensitivity of 70.1/87.0, specificity of 88.3/89.6, PPV of 35.2/20.9, and NPV of 97.0/99.5. The external validation analysis validated the generalization ability of DLMs in a heterogeneous population (AUC = 0.8816/0.9447 in ECG-EF, 0.8072/0.9148 in ECG-LV-D, and 0.8485/0.9363 in ECG-LV-S). These results revealed the possibility to detect abnormal EF/LV-D/LV-S *via* ECG accurately.

Subgroup analysis was stratified by the different clinical settings and comorbidities in **Figure 4**. DLM performed better in patients from the out-patient department (OPD) than those from the emergency room (ER) or the inpatient department (IPD). Compared to patients without comorbidities, ECG-EF, ECG-LV-D, and ECG-LV-S had lower AUC in patients with comorbidities, especially in patients with a history of AMI. These comorbidities may be potential confounding factors for new-onset LV dysfunction (45, 46). In other words, electrical abnormalities induced by comorbidities may cause ECG changes that interfere with the performance of our DLM.

In the previous study, we noticed patients with low ECG-EF and normal ECHO-EF had a higher incidence of future LV dysfunction. We hypothesized that the disease previvor was

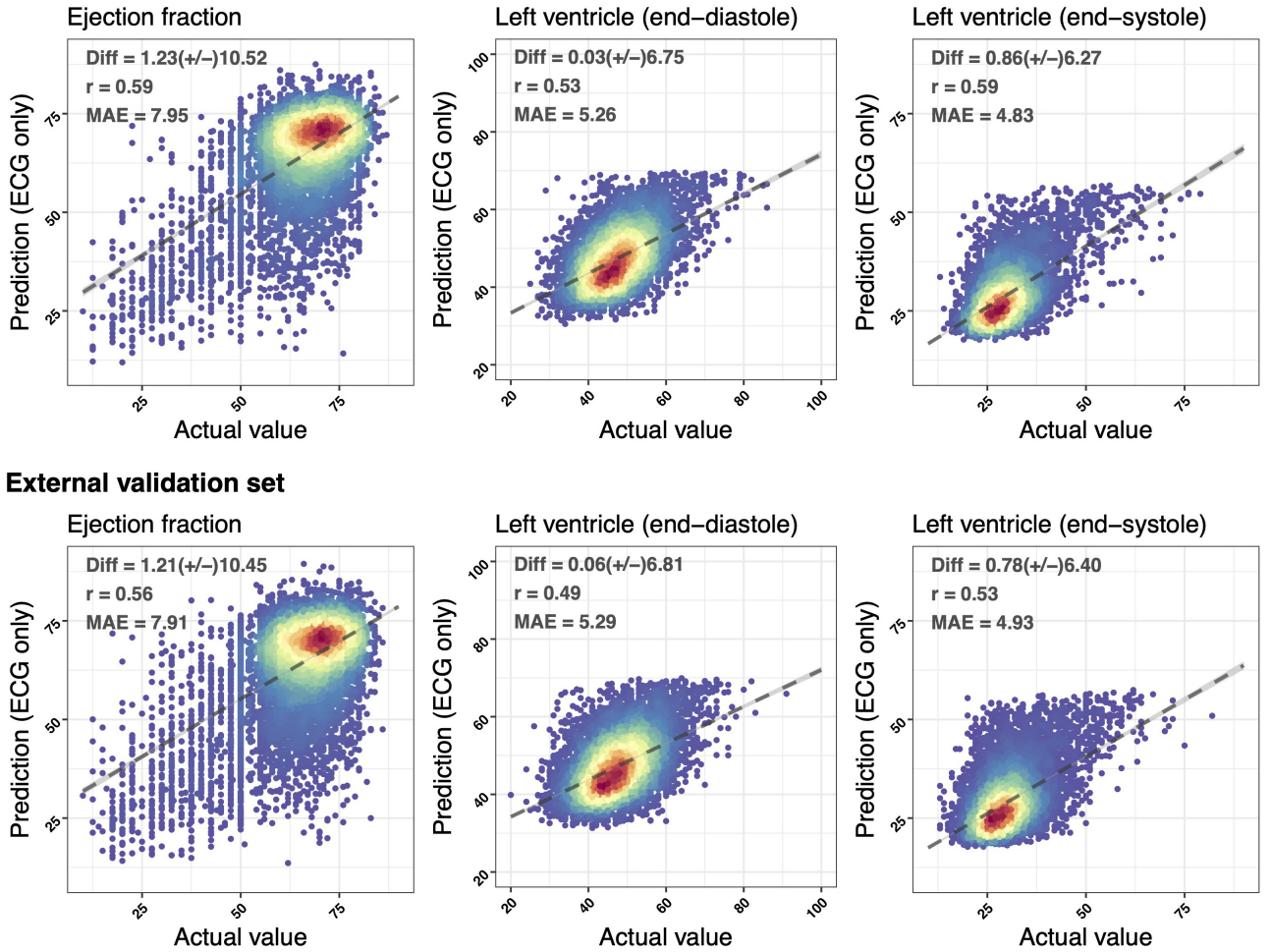

FIGURE 2 | Scatter plots of predicted value *via* ECG voltage–time traces only compared with the actual ejection fraction (EF) and left ventricle (end-diastole/end-systole) (LV-D/LV-S). The *x*-axis indicates the actual value and the *y*-axis presents the ECG-predictions. Red points represent the highest density, followed by yellow, green light blue, and dark blue. We presented the mean difference (Diff), Pearson's correlation coefficients (*r*), and mean absolute errors (MAEs) to demonstrate the accuracy of a deep learning model (DLM). The black lines with 95% conference intervals (*CI*s) are fitted *via* simple linear regression.

associated with obscure structural abnormalities, which could be detected by ECG-LV-D before actual LV dilation. Our DLM exhibited similar performance in predicting the size of LV-D and LV-S (**Figure 3**). Due to the similar clinical meaning of LV-S and LV-D in association with EF, we applied LV-D for further analysis. **Figure 5** presents the scatter plots of predicted and actual EF correlated with LV-D. Initially, we applied ECHO-LV-D in the internal validation set but only 15.2% of patients with low ECG-EF and normal ECHO-EF were identified as the mild increase (>56 mm) in the internal validation set, however, the percentage increased to 65.8% in ECG-LV-D application group. In the external validation set, the percentage increased from 20.0 to 61.1% similarly. These results may reveal the importance of ECG-LV-D on previvors detection.

Figure 6 demonstrated the relationship between known ECG features and ECG-LV-D. Our DLM identified those patients with increased ECG-LV-D were associated with the ECG features of ischemia/infarction, atrial fibrillation, tachycardia, left ventricular hypertrophy, widening QRS duration, prolonged PR interval, prolonged QT interval, augmented QRS amplitude, higher T-wave axis, lower RS wave axis, and lower P-wave axis compared to the ECG of normal patients. The explainable variation of known ECG features for DLM-based ECG-LV-D was 41.89 and 37.28% in the internal and external validation sets, respectively, which suggested that DLM could extract more than 50% additional information from raw ECGs.

In **Figure 7**, a long-term incidence of developing a new-onset LV dysfunction in the patient with initially normal EF was presented. We stratified by ECG-EF, ECHO-LV-D, and ECG-LV-D and defined normal patient groups as reference. There were 6,083 patients and 9,281 patients at risk cases and the cumulative incidence rates in the low ECG-EF (false positive) group were percentages of 32.0/44.4/44.4 and 31.7/36.0/52.0 at 2/4/6 years in the internal and external validation sets, respectively, with

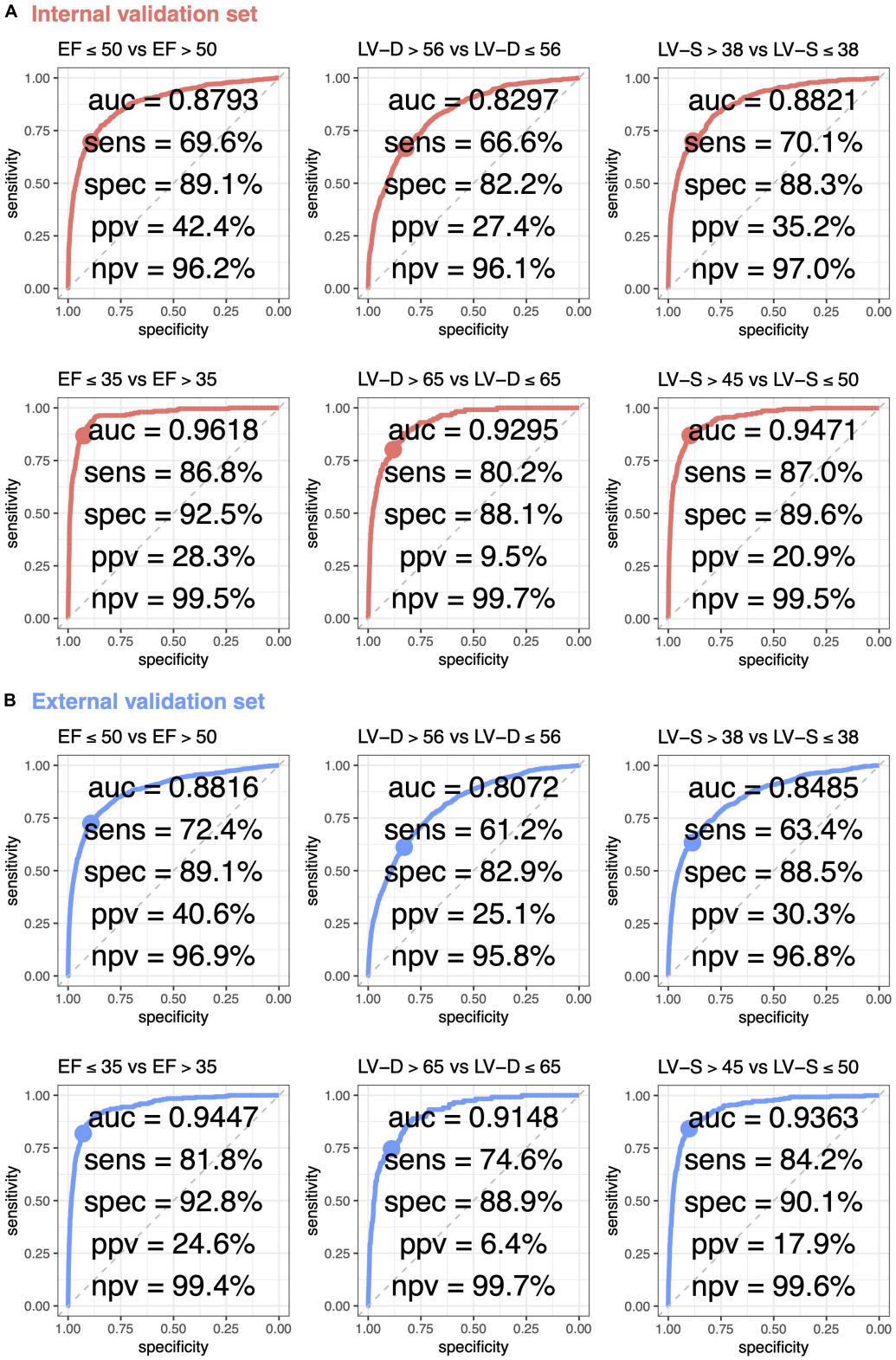

FIGURE 3 | Receiver operating characteristic (ROC) curve analysis for mild to severe left ventricle abnormality from deep learning model based ECG voltage–time traces. The ROC curve (*x*-axis = specificity and *y*-axis = sensitivity) and area under ROC curve (AUC) were calculated using the internal validation set **(A)** and external validation set **(B)**. The operating point was selected based on the maximum of Yunden's index in tuning set, which was used for calculating the corresponding sensitivities and specificities in two validation sets.

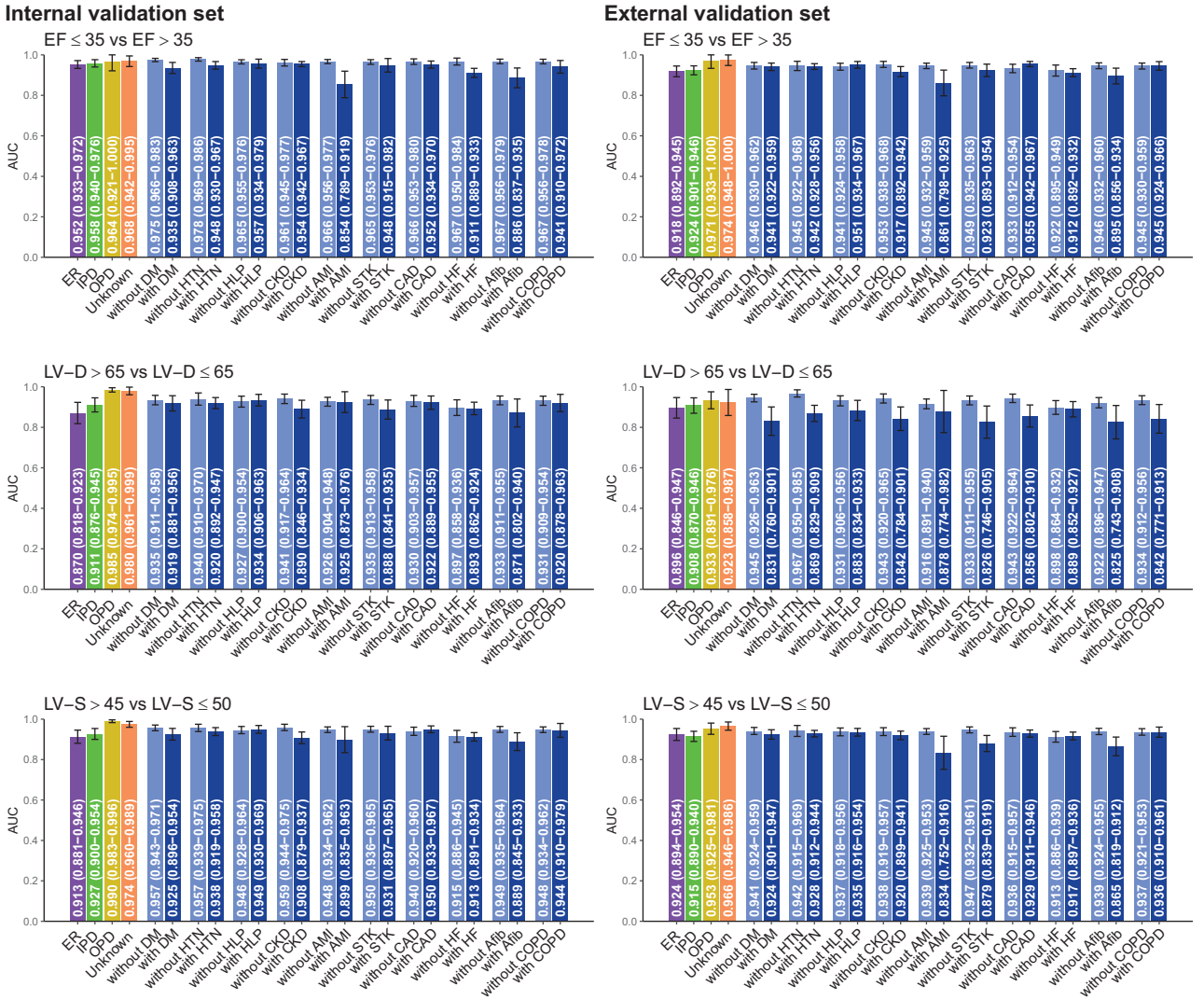

FIGURE 4 | Stratified analysis for model performance for predicting electrocardiogram-based ejection fraction (ECG-EF), left ventricular end-diastolic diameter (LV-D), and left ventricular end-systolic diameter (LV-S). The analyses were stratified by the disease histories or data source. The AUC and 95% *CI*s were presented based on LV EFs and diameters. ER, Emergency room; IPD, inpatient department; OPD, outpatient department; DM, diabetes mellitus; HTN, hypertension; HLP, hyperlipidemia; CKD, chronic kidney disease; AMI, acute myocardial infarction; STK, stroke; CAD, coronary artery disease; HF, heart failure; Afib, atrial fibrillation; COPD, chronic obstructive pulmonary disease (COPD).

corresponding significant gender-age adjusted *HR*s (95% *CI*) of 5.91 (3.58–9.78) and 5.63 (3.55–8.93). The C-index analyses also show the significant prognostic value on new onset LV dysfunction of 0.774 (95% *CI*: 0.753–0.7950) and 0.791 (95% *CI*: 0.773–0.808), which emphasized the importance of ECG-EF. In the analyses of ECHO-LV-D and ECG-LV-D, the significant gender-age adjusted *HR*s demonstrated the contributions on new-onset LV dysfunction in both validation sets. The *HR*s of severe/mild ECG-LV-D increase was 7.30 (95% *CI* 3.61–14.77)/3.12 (95% *CI* 2.41–4.03) in the internal validation set and 5.51 (95% *CI* 2.85–10.66)/2.65 (95% *CI* 2.11–3.33) in the external validation set. The C-indexes were higher in ECG-LV-D (0.750, 95% *CI* 0.727–0.772) than in ECHO-LV-D (0.723, 95% *CI* 0.699–0.747) in internal validation set, which was consistent in external validation set [0.750 (95% *CI*: 0.730–0.769) vs. 0.737 (95% *CI* 0.718–0.757)]. It suggested that ECG-LV-D may be a better differential indicator than ECHO-LV-D, which supplements the ECG-EF to identify patients at the risk of LV dysfunction in future.

Figure 8 shows the risk matrixes of different ECG-EF and ECG-LV-D on adverse events in patients with normal ECHO-EF. The patients with increased ECG-LV-D were more susceptible to adverse CV outcomes. Combining ECG-EF and ECG-LV-D, the gender-age-adjusted *HR*s increased to 4.60 (95% *CI* 3.17–6.68), 4.31 (95% *CI* 1.68–11.07), 4.80 (95% *CI* 2.78–8.28), and 2.23 (95% *CI* 1.65–3.31) on new-onset LV dysfunction, CV mortality, new-onset AMI, and CAD, respectively. Moreover, the ECG-LV-D independently provided the ability of risk stratification

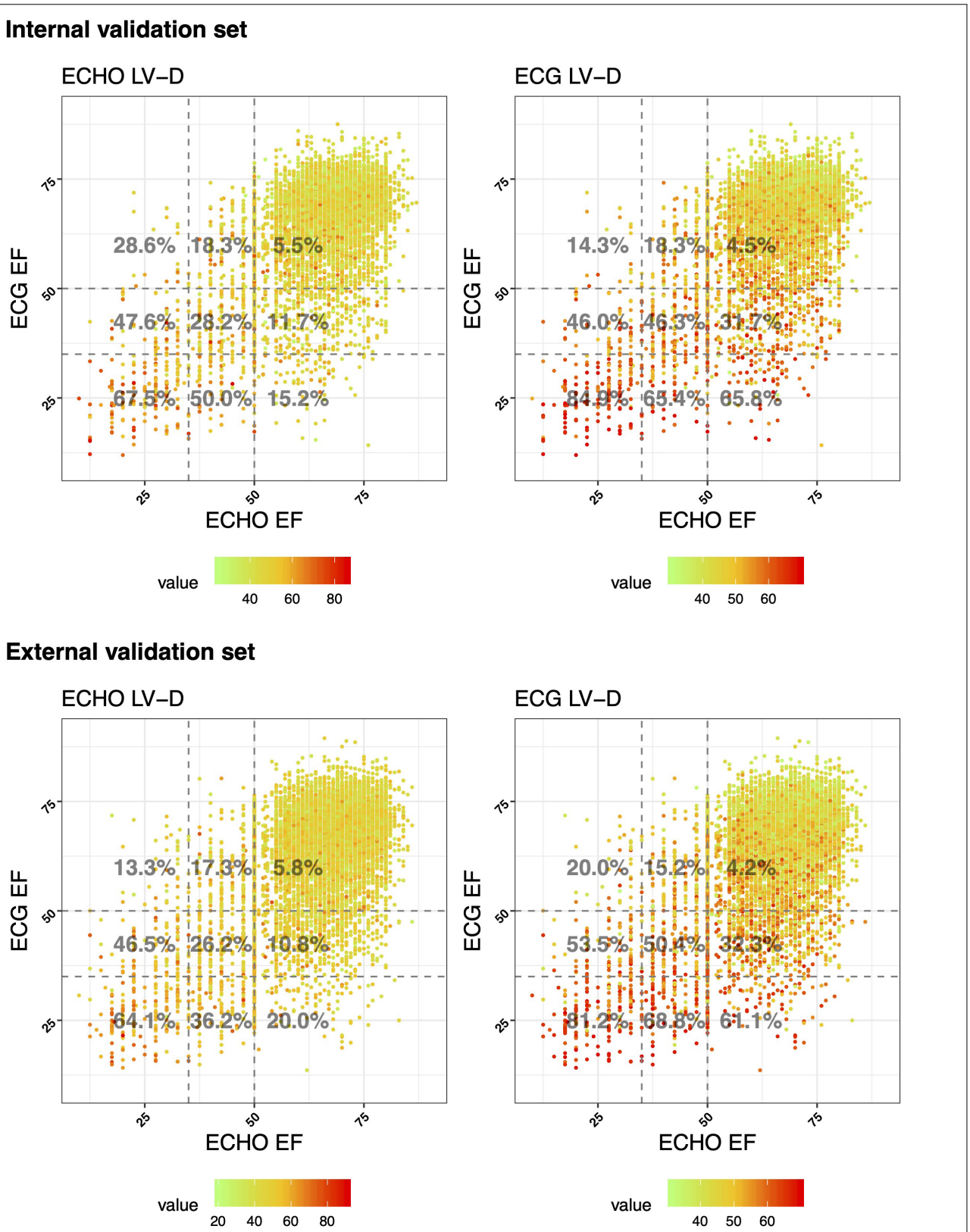

FIGURE 5 | Scatter plots of predicted and actual EF correlated with LV-D. The *x*-axis indicates the actual EF and the *y*-axis presents the ECG-EF. Green to red points represent the small and large predicted and actual LV-D, respectively. The percentages were the proportion of people with an ECHO/ECG LV-D > 56 mm in each ECHO and ECG EF group.

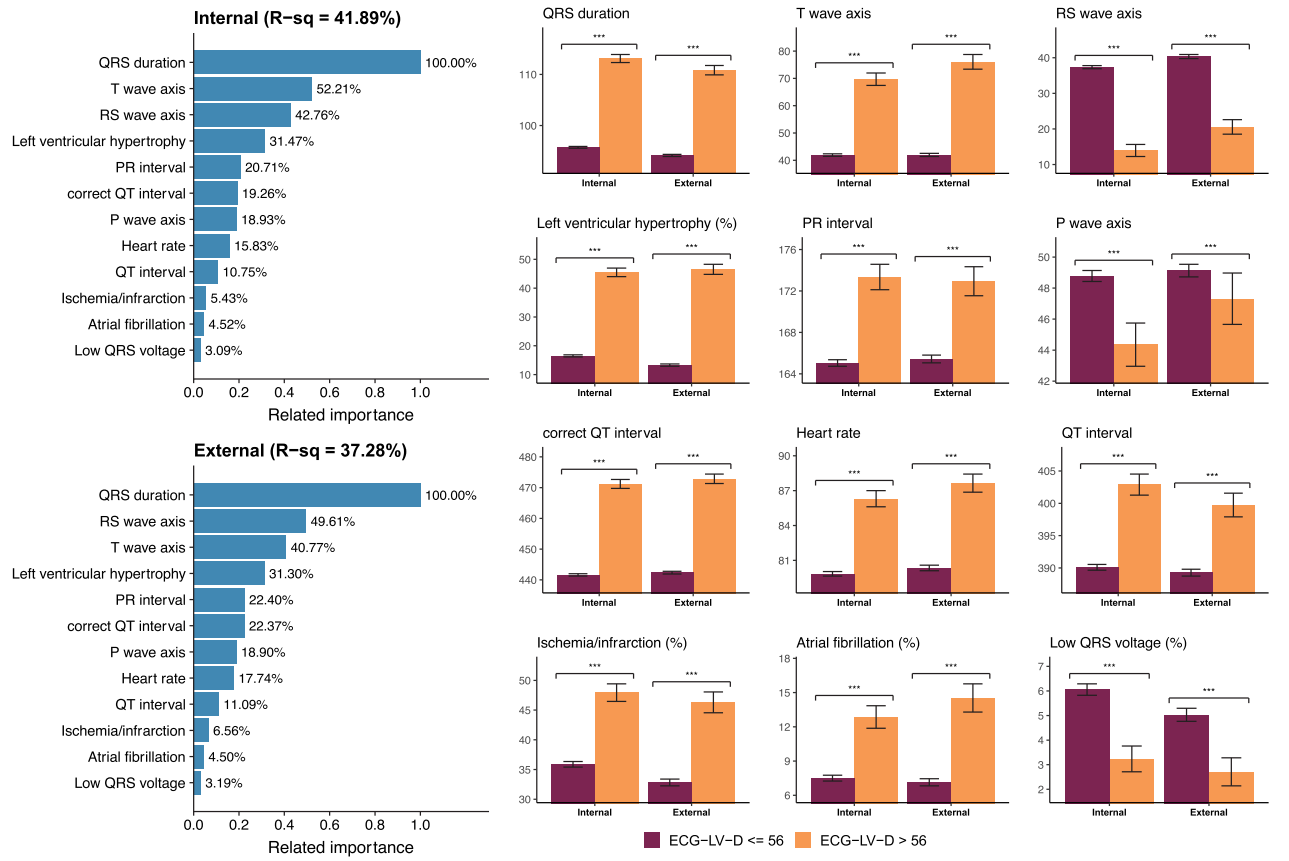

FIGURE 6 | Relationship between most important ECG features and ECG estimated left ventricle (end-diastole) (ECG-LV-D). The related importance is based on the information gain of eXtreme gradient boosting (XGB) model, and the R-square (R-sq) is the coefficient of determination to use selected ECG features for predicting ECG-LV-D. The AI-ECG predictions were classified as ECG-normal (ECG-LV-D $\leq$ 56) and ECG-abnormal (ECG-LV-D $>$ 56). The analyses are conducted both in internal and external validation sets (***$p < 0.001$).

on new-onset LV dysfunction (*HR* 2.34, 95% *CI* 1.78–3.08), CV mortality (*HR* 2.30, 95% *CI* 1.05–5.05), new-onset AMI (*HR* 2.12, 95% *CI* 1.36–3.29), and CAD (*HR* 1.59, 95% *CI* 1.26–2.00) in the internal validation set, and achieved similar trends with 1.88-fold-risk (95% *CI* 1.47–2.39) of new-onset LV dysfunction in the external validation set. In the consideration of confounding bias, we further adjusted more potential confounding factors, such as comorbidities. Our data indicated that the trend of results was similar with results adjusted by gender, age, and comorbidities (**Supplementary Figure 1**), which emphasized the importance and independency of ECG-EF and ECG-LV-D on early identification of HF risk.

DISCUSSION

In this study, we reported an AI-ECG DLM including more than 110,000 pairs of ECG and echocardiographic data and analyzed the longitudinal data, such as EF reduction, mortality, and adverse CV outcomes. Our DLM predicts ECG-EF accurately with the high AUCs of 0.9618/0.9447 for reduced EF detection (EF $\leq$ 35%) in the internal/external validation set, respectively. The high correlation between ECHO-EF and ECG-EF suggested the latter is a potential diagnostic tool. Severe/mild ECG-LV-D increase with the AUCs of 0.9295/0.8297 and 0.9148/0.8072 in internal/external validation set, which exhibited its valuable diagnostic power in patients with normal ECHO-EF. Moreover, we found a higher prevalence of ECG-LV-D increase in patients with low ECG-EF. Of these false positive patients, gender and age-adjusted *HR*s of future LV dysfunction were significantly high, suggesting that the DLM identified high-risk patients. Most importantly, the ECG-LV-D additionally contributes to predicting future LV dysfunction, which may provide the information of prognosis independently. The HRs of adverse CV outcomes increased significantly in patients identified as high ECG-LV-D and low ECG-EF compared with those with normal ECG-LV-D and ECG-EF. This is the first research to describe AI-enabled ECG-LV-D, which was demonstrated with high accuracy for the prediction of future LV dysfunction in patients with initially normal ECHO-EF.

Heart failure is an increasing problem affecting more than 30 million people globally. In these patients, asymptomatic LV dysfunction (ALVD, EF $<$ 50%) patients are difficult to diagnose, who account for 7.9–23% of population (4, 5). Patients

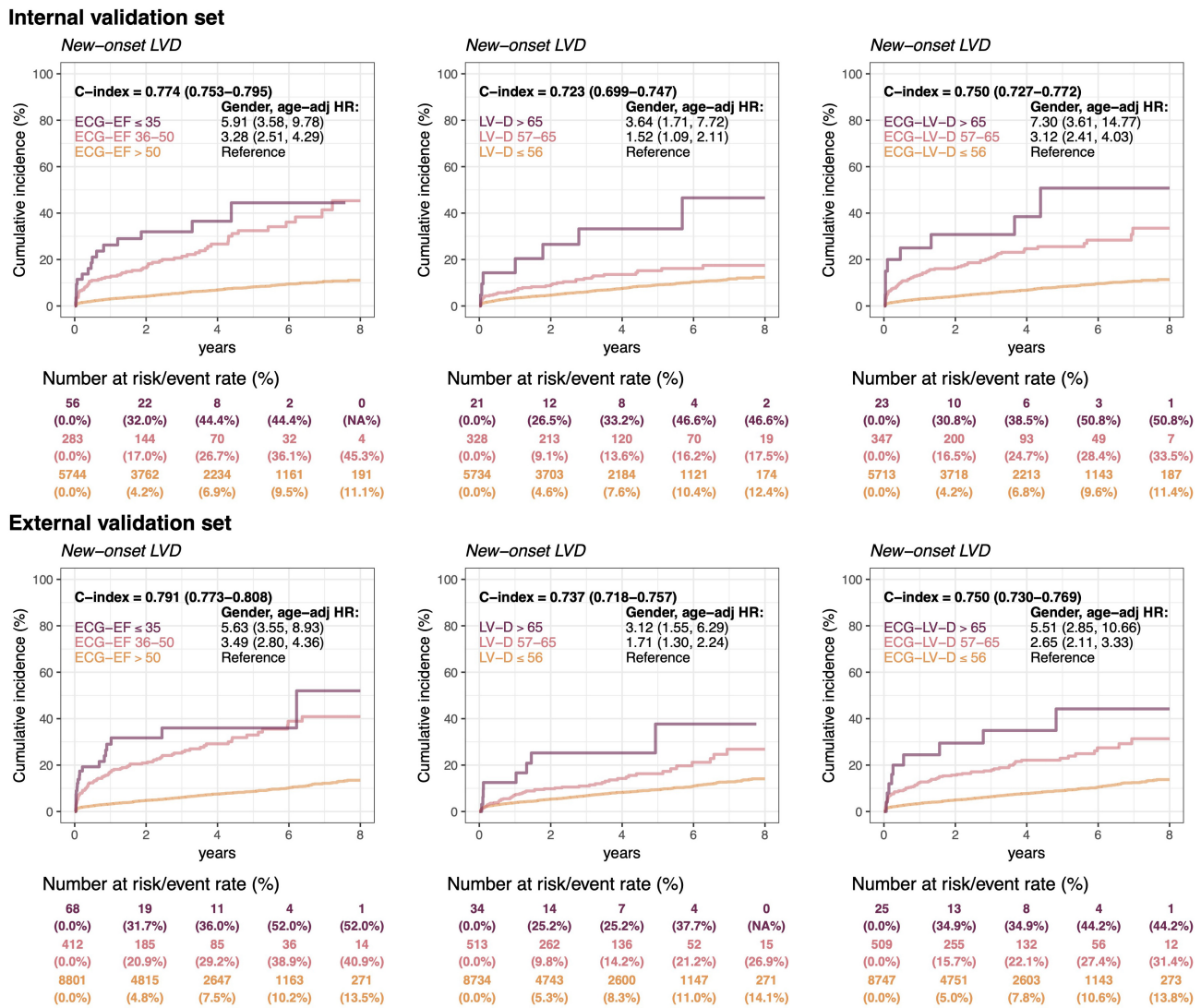

FIGURE 7 | Long-term incidence of developing a new-onset left ventricular dysfunction (LVD, EF ≤ 35%) in patients with an initially normal EF (EF > 50%) stratified by ECG-EF, LV-D, and ECG-LV-D. The C-index is calculated based on the continuous value combined with sex and age. The analyses are conducted both in internal and external validation sets. The table shows the at-risk population and cumulative risk for the given time intervals in each risk stratification.

with ALVD were associated with the reduced quality of life, increased hospitalization, morbidity, and mortality (47, 48). Although current evidence highlights the significance of ALVD and emphasized the early intervention to these patients, there is no effective tool to screen patients with ALVD (49–53). In previous studies for LV dysfunction detection, Kwon et al. proved that the DLM outperformed other machine-learning methods (54). Even with different sex, age, and body mass index, Attia et al. and Cho et al. have demonstrated ECG-EF performance stability and robustness in internal and external validation sets (36, 55, 56). Our DLM exhibits excellent predictive performance in ECG-EF and ECG-LV-D. The concept of ECG-LV-D is proposed to expand the application of ECG-EF and tried to explain the discrepancy between ECG-EF and ECHO-EF. ECG-LV-D is thought to be a structural indicator with subtle electrical signal changes which provides critical information that helps to early identify those patients who are at risk to develop LV dysfunction. In combination with ECG-EF, the diagnostic power significantly enhanced, which could be applied for large-scale screening and for patients with asymptomatic HF to improve their CV outcomes.

There are several ECG changes in LV-D increase. In dilated cardiomyopathy (DCM), about 80% of patients had ECG abnormalities, including LV hypertrophy, left/right atrial enlargement, left/right bundle branch block, abnormal Q wave, atrial fibrillation, first-degree atrial-ventricular block, and T-wave inversion in inferior and anterolateral leads (57). Merlo et al. demonstrated that LV hypertrophy, increased heart rate, and anterior T-wave inversion predicted death or heart transplantation in patients with DCM and ECHO-EF < 50%

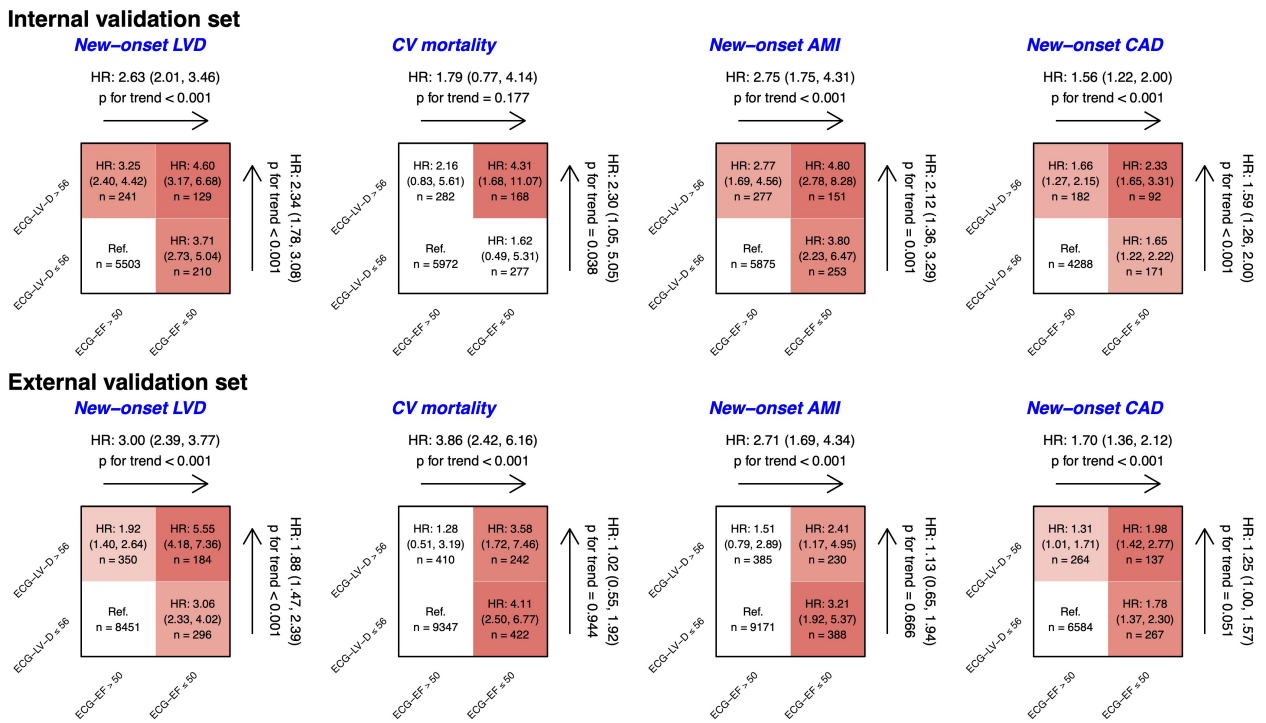

FIGURE 8 | Risk matrixes of different predicted ejection fraction (ECG-EF) and left ventricle (end-diastole) (ECG-LV-D) groups on adverse events in patients with an initially normal EF (EF > 50%). The hazard ratios (*HRs*) are based on the Cox proportional hazard model adjusted by gender and age. The color gradient represents the risk of corresponding group, and the non-significant results are defined as white.

(58). Previous studies proposed that delayed LV conduction with QRS prolongation (≥ 120 ms) was associated with restrictive LV filling, more advanced myocardial disease, worse LV function, poorer prognosis, and a higher all-cause mortality rate (59, 60). We found that our DLM was strongly correlated with prolonged QRS duration, which partly explained why the patients with high ECG-LV-D had a higher risk of LV dysfunction compared to patients with normal ECG-LV-D. Meanwhile, the possible mechanisms underlying the interference of DLM performance among patients with AMI could be myocardial scarring, which may affect electrical vectors, create regions of slowed conduction, and re-entrant circuits supporting sustained ventricular tachycardia (61, 62). Along with ECG-EF, the ECG-LV-D performed significantly better prediction capacity on new-onset LV dysfunction, CV mortality, new-onset AMI, and CAD compared to ECG-EF alone in the internal validation set. However, in the external validation set, in which the data from mild disease patients in community hospital, only the prediction of LV dysfunction could be significantly enhanced. Possible reasons underlying the inconsistency include different patient population and disease severities. Considering the better performance of our DLM in patients with less comorbidities from OPD than those from ER or IPD, our DLM could be more suitable for community screening than for hospitalized patients. Further large-scale studies are needed to confirm the combination effects of ECG-LV-D and ECG-EF.

The clinical application of AI-ECG is a worldwide tendency and developed rapidly. As the AI-ECG could predict the disease development in healthy individuals without abnormal imaging findings or symptoms, the concept of previvors was proposed recently. With apparent false positive AI-ECG findings, patients with a higher risk of many diseases, such as LV dysfunction (20), future atrial fibrillation (63), hyperkalemia (64), and elder heart age (44), could receive preventive interventions or medical surveillance early.

The importance and clinical significance of our ECG-LV-D should be emphasized. Both ECG-EF and ECG-LV-D are promising screening tools for patients who had a high risk of future LV dysfunction. The advantage of timely HF risk identification is evident to prevent adverse CV events and reduce medical costs. Moreover, from a large community-based study of sudden cardiac death (SCD), LV-D may contribute to the risk of SCD independent of the EF (41). The ECG-EF and ECG-LV-D models could be applied for risk stratification in patients with HF, especially those with stage A or B HF (65). Importantly, the wearable devices with ECG-EF and ECG-LV-D algorithms would provide timely conditions and beneficial effects for high-risk patients. Finally, considering that ECG is widely used and is a standardized examination in a rural or remote hospital, the AI-ECG could analyze and alert physician automatically and immediately among these areas. Further community-based studies of ECG-LV-D application are necessary to validate clinical benefits on HF patient care.

There are some limitations to this study. First, this study was a retrospective study. Although ECG/ECHO pairs were collected and the DLM was validated, the accuracy in different hospital settings and prospective studies are necessary to generalize the application of ECG-LV-D and promote treatment strategy. Second, the clinical impact of treatment is needed to verify. The actual benefit of ECG-LV-D import to clinical practice is not clear now. Investigation of clinical benefits including accidental HF detection, time reduction, prognosis management, and outcomes evaluation should be conducted. Third, the best application of AI-ECG is to screen asymptomatic patients with HF, but the relationship between abnormal ECG and HF symptoms was unclear. Future study should conduct a large-scale community screening to validate the benefit in asymptomatic patients with HF. Fourth, AI-ECG performed worse in patients with more comorbidities, especially in patients with a history of AMI. Interestingly, even after the adjustment of all the confounding factors, our models of ECG-EF and ECG-LV-D still provide significant predictive power for newly onset LV dysfunction. Finally, the DLM design is an uninterpretable set of methods, such as a black box, and full interpretability will be a focus of future work.

In conclusion, our AI-ECG DLM could identify patients with high ECG-LV-D and predict future LV dysfunction. ECG-LV-D serves as an independent risk factor of long-term CV outcomes in patients with normal ECHO-EF and low ECG-EF. The combination of ECG-EF and ECG-LV-D provides significantly synergistic diagnostic power to predict patients with future LV dysfunction. Although further studies are needed, our ECG-LV-D could be used as a screening tool for patients with normal EF but with high cardiovascular risk to initiate appropriate treatment in time.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

CL, C-SL, and W-HF contributed to conception and design of the study. CL, C-HW, and C-CL organized the database. CL and C-SL performed the statistical analysis and wrote sections of the manuscript. H-YC and C-SL wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This study was supported by funding from the Ministry of Science and Technology, Taiwan (MOST110-2314-B-016-010-MY3 to CL and MOST110-2321-B-016-002 to C-HW), the Tri-Service General Hospital, Taiwan (TSGH-B-111020 to C-LH), and the Cheng Hsin General Hospital, Taiwan (CHNDMC-111-07 to CL).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.870523/full#supplementary-material

REFERENCES

1. Ziaeian B, Fonarow GC. Epidemiology and aetiology of heart failure. *Nat Rev Cardiol.* (2016) 13:368–78. doi: 10.1038/nrcardio.2016.25
2. Ponikowski P, Anker SD, AlHabib KF, Cowie MR, Force TL, Hu S, et al. Heart failure: preventing disease and death worldwide. *ESC Heart Fail.* (2014) 1:4–25. doi: 10.1002/ehf2.12005
3. Heidenreich PA, Albert NM, Allen LA, Bluemke DA, Butler J, Fonarow GC, et al. Forecasting the impact of heart failure in the united states: a policy statement from the American heart association. *Circ Heart Fail.* (2013) 6:606–19. doi: 10.1161/HHF.0b013e318291329a
4. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE Jr, Colvin MM, et al. 2017 Acc/Aha/Hfsa focused update of the 2013 Accf/Aha guideline for the management of heart failure: a report of the american college of cardiology/american heart association task force on clinical practice guidelines and the heart failure society of America. *J Am Coll Cardiol.* (2017) 70:776–803. doi: 10.1016/j.jacc.2017.04.025
5. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE Jr, Drazner MH, et al. 2013 Accf/Aha guideline for the management of heart failure: a report of the American college of cardiology foundation/American heart association task force on practice guidelines. *J Am Coll Cardiol.* (2013) 62:e147–239. doi: 10.1016/j.jacc.2013.05.019
6. Ambrosy AP, Fonarow GC, Butler J, Chioncel O, Greene SJ, Vaduganathan M, et al. The global health and economic burden of hospitalizations for heart failure: lessons learned from hospitalized heart failure registries. *J Am Coll Cardiol.* (2014) 63:1123–33. doi: 10.1016/j.jacc.2013.11.053
7. Budhwani N, Patel S, Dwyer EM. Electrocardiographic diagnosis of left ventricular hypertrophy: the effect of left ventricular wall thickness, size, and mass on the specific criteria for left ventricular hypertrophy. *Am Heart J.* (2005) 149:709–14. doi: 10.1016/j.ahj.2004.07.040
8. Hancock EW, Deal BJ, Mirvis DM, Okin P, Kligfield P, Gettes LS. Aha/Accf/Hrs recommendations for the standardization and interpretation of the electrocardiogram. *Circulation.* (2009) 119:e251–61. doi: 10.1161/CIRCULATIONAHA.108.191097
9. Bonnes JL, Thannhauser J, Nas J, Westra SW, Jansen RMG, Meinsma G, et al. Ventricular fibrillation waveform characteristics of the surface ECG: impact of the left ventricular diameter and mass. *Resuscitation.* (2017) 115:82–9. doi: 10.1016/j.resuscitation.2017.03.029
10. Moyssakis I, Moschos N, Triposkiadis F, Hallaq Y, Pantazopoulos N, Aessopos A, et al. Left ventricular end-systolic stress/diameter relation as a contractility index and as a predictor of survival. Independence of preload after normalization for end-diastolic diameter. *Heart Vessels.* (2005) 20:191–8. doi: 10.1007/s00380-005-0832-x
11. Katz AM. Cardiomyopathy of overload. A major determinant of prognosis in congestive heart failure. *N Engl J Med.* (1990) 322:100–10. doi: 10.1056/NEJM199001113220206
12. Rihal CS, Nishimura RA, Hatle LK, Bailey KR, Tajik AJ. Systolic and diastolic dysfunction in patients with clinical diagnosis of dilated cardiomyopathy. Relation to symptoms and prognosis. *Circulation.* (1994) 90:2772–9. doi: 10.1161/01.cir.90.6.2772
13. Dec GW, Fuster V. Idiopathic dilated cardiomyopathy. *N Engl J Med.* (1994) 331:1564–75. doi: 10.1056/nejm199412083312307

14. Ito K, Li S, Homma S, Thompson JLP, Buchsbaum R, Matsumoto K, et al. Left ventricular dimensions and cardiovascular outcomes in systolic heart failure: the Warcef trial. *ESC Heart Fail.* (2021) 8:4997–5009. doi: 10.1002/ehf2.13560
15. Vasan RS, Larson MG, Benjamin EJ, Evans JC, Levy D. Left ventricular dilatation and the risk of congestive heart failure in people without myocardial infarction. *N Engl J Med.* (1997) 336:1350–5. doi: 10.1056/nejm199705083361903
16. Makaryus AN, Catanzaro JN, Hametz CD, Jadonath RL. Clinical investigation: utility of left ventricular end diastolic diameter in the prediction of susceptibility to ventricular tachyarrhythmias. *Int J Cardiol.* (2007) 120:399–403. doi: 10.1016/j.ijcard.2006.10.030
17. Inoue T, Ogawa T, Iwabuchi Y, Otsuka K, Nitta K. Left ventricular end-diastolic diameter is an independent predictor of mortality in hemodialysis patients. *Ther Apher Dial.* (2012) 16:134–41. doi: 10.1111/j.1744-9987.2011.01048.x
18. Segawa K, Sugawara N, Maruo K, Kimura K, Komaki H, Takahashi Y, et al. Left ventricular end-diastolic diameter and cardiac mortality in duchenne muscular dystrophy. *Neuropsychiatr Dis Treat.* (2020) 16:171–8. doi: 10.2147/NDT.S235166
19. Tribouilloy C, Grigioni F, Avierinos JF, Barbieri A, Rusinaru D, Szymanski C, et al. Survival implication of left ventricular end-systolic diameter in mitral regurgitation due to flail leaflets: a long-term follow-up multicenter study. *J Am Coll Cardiol.* (2009) 54:1961–8. doi: 10.1016/j.jacc.2009.06.047
20. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med.* (2019) 25:70–4. doi: 10.1038/s41591-018-0240-2
21. Lou YS, Lin CS, Fang WH, Lee CC, Ho CL, Wang CH, et al. Artificial intelligence-enabled electrocardiogram estimates left atrium enlargement as a predictor of future cardiovascular disease. *J Pers Med.* (2022) 12:315. doi: 10.3390/jpm12020315
22. 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
23. Galloway CD, Valys AV, Shreibati JB, Treiman DL, Petterson FL, Gundotra VP, et al. Development and validation of a deep-learning model to screen for hyperkalemia from the electrocardiogram. *JAMA Cardiol.* (2019) 4:428–36. doi: 10.1001/jamacardio.2019.0640
24. Lin CS, Lin C, Fang WH, Hsu CJ, Chen SJ, Huang KH, et al. A deep-learning algorithm (Ecg12net) for detecting hypokalemia and hyperkalemia by electrocardiography: algorithm development. *JMIR Med Inform.* (2020) 8:e15931. doi: 10.2196/15931
25. Lin C, Lin C-S, Lee D-J, Lee C-C, Chen S-J, Tsai S-H, et al. Artificial intelligence assisted electrocardiography for early diagnosis of thyrotoxic periodic paralysis. *J Endocr Soc.* (2021) 5:bvab120. doi: 10.1210/jendso/bvab120
26. Lin CS, Lee YT, Fang WH, Lou YS, Kuo FC, Lee CC, et al. Deep learning algorithm for management of diabetes mellitus *via* electrocardiogram-based glycated hemoglobin (ECG-HbA1c): a retrospective cohort study. *J Pers Med.* (2021) 11:725. doi: 10.3390/jpm11080725
27. Chang D-W, Lin C-S, Tsao T-P, Lee C-C, Chen J-T, Tsai C-S, et al. Detecting digoxin toxicity by artificial intelligence-assisted electrocardiography. *Int J Environ Res Public Health.* (2021) 18:3839. doi: 10.3390/ijerph18073839
28. Liu WT, Lin CS, Tsao TP, Lee CC, Cheng CC, Chen JT, et al. A deep-learning algorithm-enhanced system integrating electrocardiograms and chest X-rays for diagnosing aortic dissection. *Can J Cardiol.* (2022) 38:160–8. doi: 10.1016/j.cjca.2021.09.028
29. Lee CC, Lin CS, Tsai CS, Tsao TP, Cheng CC, Liou JT, et al. A deep learning-based system capable of detecting pneumothorax *via* electrocardiogram. *Eur J Trauma Emerg Surg.* (2022). doi: 10.1007/s00068-022-01904-3 [Epub ahead of print].
30. Cho Y, Kwon JM, Kim KH, Medina-Inojosa JR, Jeon KH, Cho S, et al. Artificial intelligence algorithm for detecting myocardial infarction using six-lead electrocardiography. *Sci Rep.* (2020) 10:20495. doi: 10.1038/s41598-020-77599-6
31. Liu WC, Lin C, Lin CS, Tsai MC, Chen SJ, Tsai SH, et al. An artificial intelligence-based alarm strategy facilitates management of acute myocardial infarction. *J Pers Med.* (2021) 11:1149. doi: 10.3390/jpm11111149
32. Liu WC, Lin CS, Tsai CS, Tsao TP, Cheng CC, Liou JT, et al. A deep-learning algorithm for detecting acute myocardial infarction. *Eurointervention.* (2021) 17:765–73. doi: 10.4244/eij-d-20-01155
33. Lima EM, Ribeiro AH, Paixão GMM, Ribeiro MH, Pinto-Filho MM, Gomes PR, et al. Deep neural network-estimated electrocardiographic age as a mortality predictor. *Nat Commun.* (2021) 12:5117. doi: 10.1038/s41467-021-25351-7
34. Attia ZI, Harmon DM, Behr ER, Friedman PA. Application of artificial intelligence to the electrocardiogram. *Eur Heart J.* (2021) 42:4717–30. doi: 10.1093/eurheartj/ehab649
35. Vardas PE, Asselbergs FW, van Smeden M, Friedman P. The year in cardiovascular medicine 2021: digital health and innovation. *Eur Heart J.* (2022) 43:271–9. doi: 10.1093/eurheartj/ehab874
36. Attia IZ, Tseng AS, Benavente ED, Medina-Inojosa JR, Clark TG, Malyutina S, et al. External validation of a deep learning electrocardiogram algorithm to detect ventricular dysfunction. *Int J Cardiol.* (2021) 329:130–5. doi: 10.1016/j.ijcard.2020.12.065
37. Vaid A, Johnson KW, Badgeley MA, Somani SS, Bicak M, Landi I, et al. Using deep-learning algorithms to simultaneously identify right and left ventricular dysfunction from the electrocardiogram. *JACC Cardiovasc Imaging.* (2022) 15:395–410. doi: 10.1016/j.jcmg.2021.08.004
38. Nagueh SF, Smiseth OA, Appleton CP, Byrd BF, Dokainish H, Edvardsen T, et al. Recommendations for the evaluation of left ventricular diastolic function by echocardiography: an update from the American society of echocardiography and the European association of cardiovascular imaging. *J Am Soc Echocardiogr.* (2016) 29:277–314. doi: 10.1016/j.echo.2016.01.011
39. 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. *Eur Heart J Cardiovasc Imaging.* (2015) 16:233–71. doi: 10.1093/ehjci/jev014
40. Poppe KK, Doughty RN, Gardin JM, Hobbs FDR, McMurray JJV, Nagueh SF, et al. Ethnic-specific normative reference values for echocardiographic La and Lv Size, Lv mass, and systolic function: the echonormal study. *JACC Cardiovasc Imaging.* (2015) 8:656–65. doi: 10.1016/j.jcmg.2015.02.014
41. Narayanan K, Reinier K, Teodorescu C, Uy-Evanado A, Aleong R, Chugh H, et al. Left ventricular diameter and risk stratification for sudden cardiac death. *J Am Heart Assoc.* (2014) 3:e001193. doi: 10.1161/jaha.114.001193
42. 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
43. Seko Y, Kato T, Morita Y, Yamaji Y, Haruna Y, Izumi T, et al. Age- and body size-adjusted left ventricular end-diastolic dimension in a Japanese hospital-based population. *Circ J.* (2019) 83:604–13. doi: 10.1253/circj.CJ-18-1095
44. Chang C-H, Lin C-S, Luo Y-S, Lee Y-T, Lin C. Electrocardiogram-based heart age estimation by a deep learning model provides more information on the incidence of cardiovascular disorders. *Front Cardiovasc Med.* (2022) 9:754909. doi: 10.3389/fcvm.2022.754909
45. Kuznetsova T, Herbots L, Jin Y, Stolarz-Skrzypek K, Staessen JA. Systolic and diastolic left ventricular dysfunction: from risk factors to overt heart failure. *Expert Rev Cardiovasc Ther.* (2010) 8:251–8. doi: 10.1586/erc.10.3
46. Zheng C, Chen Z, Zhang L, Wang X, Dong Y, Wang J, et al. Metabolic risk factors and left ventricular diastolic function in middle-aged Chinese living in the Tibetan plateau. *J Am Heart Assoc.* (2019) 8:e010454. doi: 10.1161/JAHA.118.010454
47. Sara JD, Toya T, Taher R, Lerman A, Gersh B, Anavekar NS. Asymptomatic left ventricle systolic dysfunction. *Eur Cardiol.* (2020) 15:e13. doi: 10.15420/ecr.2019.14
48. Goldberg LR, Jessup M. Stage B heart failure: management of asymptomatic left ventricular systolic dysfunction. *Circulation.* (2006) 113:2851–60. doi: 10.1161/CIRCULATIONAHA.105.600437
49. Yusuf S, Pitt B, Davis CE, Hood WB Jr, Cohn JN. Effect of enalapril on mortality and the development of heart failure in asymptomatic patients with

reduced left ventricular ejection fractions. *N Engl J Med.* (1992) 327:685–91. doi: 10.1056/nejm199209033271003
50. Jong P, Yusuf S, Rousseau MF, Ahn SA, Bangdiwala SI. Effect of Enalapril on 12-year survival and life expectancy in patients with left ventricular systolic dysfunction: a follow-up study. *Lancet.* (2003) 361:1843–8. doi: 10.1016/S0140-6736(03)13501-5
51. Køber L, Torp-Pedersen C, Carlsen JE, Bagger H, Eliasen P, Lyngborg K, et al. A clinical trial of the angiotensin-converting-enzyme inhibitor trandolapril in patients with left ventricular dysfunction after myocardial infarction. trandolapril cardiac evaluation (trace) study group. *N Engl J Med.* (1995) 333:1670–6. doi: 10.1056/nejm199512213332503
52. Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, et al. Heart disease and stroke statistics–2015 update: a report from the American heart association. *Circulation.* (2015) 131:e29–322. doi: 10.1161/cir.0000000000000152
53. Pfeffer MA, Braunwald E, Moyé LA, Basta L, Brown EJ Jr, Cuddy TE, et al. Effect of captopril on mortality and morbidity in patients with left ventricular dysfunction after myocardial infarction. Results of the survival and ventricular enlargement trial. the save investigators. *N Engl J Med.* (1992) 327:669–77. doi: 10.1056/nejm199209033271001
54. Kwon JM, Kim KH, Jeon KH, Kim HM, Kim MJ, Lim SM, et al. Development and validation of deep-learning algorithm for electrocardiography-based heart failure identification. *Korean Circ J.* (2019) 49:629–39. doi: 10.4070/kcj.2018.0446
55. Attia ZI, Kapa S, Yao X, Lopez-Jimenez F, Mohan TL, Pellikka PA, et al. Prospective validation of a deep learning electrocardiogram algorithm for the detection of left ventricular systolic dysfunction. *J Cardiovasc Electrophysiol.* (2019) 30:668–74. doi: 10.1111/jce.13889
56. Cho J, Lee B, Kwon JM, Lee Y, Park H, Oh BH, et al. Artificial intelligence algorithm for screening heart failure with reduced ejection fraction using electrocardiography. *ASAIO J.* (2021) 67:314–21. doi: 10.1097/MAT.0000000000001218
57. Finocchiaro G, Merlo M, Sheikh N, De Angelis G, Papadakis M, Olivotto I, et al. The electrocardiogram in the diagnosis and management of patients with dilated cardiomyopathy. *Eur J Heart Fail.* (2020) 22:1097–107. doi: 10.1002/ejhf.1815
58. Merlo M, Zaffalon D, Stolfo D, Altinier A, Barbati G, Zecchin M, et al. ECG in dilated cardiomyopathy: specific findings and long-term prognostic significance. *J Cardiovasc Med (Hagerstown).* (2019) 20:450–8. doi: 10.2459/JCM.0000000000000804
59. Erdogan T, Durakoglugil ME, Cicek Y, Cetin M, Duman H, Satiroglu O, et al. Prolonged QRS duration on surface electrocardiogram is associated with left ventricular restrictive filling pattern. *Interv Med Appl Sci.* (2017) 9:9–14. doi: 10.1556/1646.9.2017.1.05
60. Kashani A, Barold SS. Significance of QRS complex duration in patients with heart failure. *J Am Coll Cardiol.* (2005) 46:2183–92. doi: 10.1016/j.jacc.2005.01.071
61. Nable JV, Brady W. The evolution of electrocardiographic changes in St-segment elevation myocardial infarction. *Am J Emerg Med.* (2009) 27:734–46. doi: 10.1016/j.ajem.2008.05.025
62. Strauss DG, Selvester RH, Lima JA, Arheden H, Miller JM, Gerstenblith G, et al. ECG quantification of myocardial scar in cardiomyopathy patients with or without conduction defects: correlation with cardiac magnetic resonance and arrhythmogenesis. *Circ Arrhythm Electrophysiol.* (2008) 1:327–36. doi: 10.1161/circep.108.798660
63. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet.* (2019) 394:861–7. doi: 10.1016/s0140-6736(19)31721-0
64. Lin C, Chau T, Lin C-S, Shang H-S, Fang W-H, Lee D-J, et al. Point-of-care artificial intelligence-enabled ECG for dyskalemia: a retrospective cohort analysis for accuracy and outcome prediction. *NPJ Digit Med.* (2022) 5:8. doi: 10.1038/s41746-021-00550-0
65. Bozkurt B, Coats AJ, Tsutsui H, Abdelhamid M, Adamopoulos S, Albert N, et al. Universal definition and classification of heart failure: a report of the heart failure society of America, heart failure association of the European society of cardiology, Japanese heart failure society and writing committee of the universal definition of heart failure. *J Card Fail.* (2021). doi: 10.1016/j.cardfail.2021.01.022 [Epub ahead of print].

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.

frontiers | Frontiers in Medicine

ORIGINAL RESEARCH
published: 21 April 2022
doi: 10.3389/fmed.2022.894726

Using Deep Learning Radiomics to Distinguish Cognitively Normal Adults at Risk of Alzheimer's Disease From Normal Control: An Exploratory Study Based on Structural MRI

Jiehui Jiang[1,2], Jieming Zhang[3], Zhuoyuan Li[3], Lanlan Li[3], Bingcang Huang[1] and Alzheimer's Disease Neuroimaging Initiative[†]*

[1] Department of Radiology, Gongli Hospital, School of Medicine, Shanghai University, Shanghai, China, [2] School of Life Sciences, Institute of Biomedical Engineering, Shanghai University, Shanghai, China, [3] School of Communication and Information Engineering, Shanghai University, Shanghai, China

Objectives: We proposed a novel deep learning radiomics (DLR) method to distinguish cognitively normal adults at risk of Alzheimer's disease (AD) from normal control based on T1-weighted structural MRI images.

Methods: In this study, we selected MRI data from the Alzheimer's Disease Neuroimaging Initiative Database (ADNI), which included 417 cognitively normal adults. These subjects were divided into 181 individuals at risk of Alzheimer's disease (preAD group) and 236 normal control individuals (NC group) according to standard uptake ratio >1.18 calculated by amyloid Positron Emission Tomography (PET). We further divided the preaAD group into APOE+ and APOE− subgroups according to whether APOE ε4 was positive or not. All data sets were divided into one training/validation group and one independent test group. The proposed DLR method included three steps: (1) the pre-training of basic deep learning (DL) models, (2) the extraction, selection and fusion of DLR features, and (3) classification. The support vector machine (SVM) was used as the classifier. In the comparative experiments, we compared our proposed DLR method with three existing models: hippocampal model, clinical model, and traditional radiomics model. Ten-fold cross-validation was performed with 100 time repetitions.

Results: The DLR method achieved the best classification performance between preAD and NC than other models with an accuracy of 89.85% ± 1.12%. In comparison, the accuracies of the other three models were 72.44% ± 1.37%, 82.00% ± 4.09% and 79.65% ± 2.21%. In addition, the DLR model also showed the best classification performance (85.45% ± 9.04% and 92.80% ± 2.61%) in the subgroup experiment.

Conclusion: The results showed that the DLR method provided a potentially clinical value to distinguish preAD from NC.

Keywords: deep learning radiomic, Alzheimer's disease, magnetic resonance imaging, support vector machine, artificial intelligence

Edited by:
Md.Mohaimenul Islam,
Aesop Technology, Taiwan

Reviewed by:
Chuantao Zuo,
Fudan University, China
Jimin Hong,
University of Bern, Switzerland

****Correspondence:***
Bingcang Huang
hbc9209@sina.com

[†]Data used in preparation of this manuscript were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Specialty section:
This article was submitted to Family Medicine and Primary Care, a section of the journal Frontiers in Medicine

Received: *12 March 2022*
Accepted: *28 March 2022*
Published: *21 April 2022*

Citation:
Jiang J, Zhang J, Li Z, Li L, Huang B and Alzheimer's Disease Neuroimaging Initiative (2022) Using Deep Learning Radiomics to Distinguish Cognitively Normal Adults at Risk of Alzheimer's Disease From Normal Control: An Exploratory Study Based on Structural MRI. Front. Med. 9:894726. doi: 10.3389/fmed.2022.894726

INTRODUCTION

Alzheimer's disease (AD) is a neurodegenerative disease characterized by progressive cognitive decline (1). Due to the irreversibility of AD, it is critical to identify AD patients at an ultra-early stage. According to the latest A-T-N diagnosis criteria (2–4), individuals who showed obvious brain amyloid beta (Aβ+) deposition have entered the Alzheimer's continuum and represented a high-risk of AD. This population could be defined as the preclinical AD group (PreAD) (5, 6).

So far, structural resonance imaging (MRI) have been widely used in the diagnosis of AD (7–11). For instance, previous studies have shown that patients with mild cognitive impairment(MCI) had increased hippocampal atrophy compared to normal control (NC) subjects (12). The atrophy of hippocampal and entorhinal cortex could also be used as an index to predict the conversion from MCI to AD (13).

Currently, artificial intelligence (AI) techniques based on MRI have frequently been used in the early diagnosis of AD. One typical AI application is radiomics. For example, Zhao et al. investigated hippocampal texture radiomics features as effective MRI biomarkers for AD and achieved an accuracy of 87.4% to distinguish AD and normal controls (NC) (14). Zhou and Shu et al. utilized MRI radiomics features to predict development of MCI to AD and achieved the accuracy of 78.4 and 80.7%, respectively (15, 16). Notably, Li et al. conducted an exploratory study to diagnosis preAD from NC based on radiomics multi-parameter MRI and obtained an average accuracy of 83.7% [T. (17)]. Although the feasibility of traditional radiomics methods has been proven, these methods could not be widely applied because of obvious shortcomings, such as manual extraction of regions of interest (ROIs) and hand-coding, which usually require complex manual operations. Therefore, an alternative method is required.

The deep learning radiomics (DLR) method may be the alternative (18, 19). This technique was able to mine the high dimension features of medical images automatically, and effectively address the shortage of hand-coding by radiomics. Recently, DLR has been used in brain tumor-related research and AD diagnosis (20, 21). For example, previous studies achieved good predictive performance of preoperative meningioma with an accuracy of 92.6% (22). Wang et al. extracted MRI-based DLR features to predict the prognosis of high-grade glioma (23). For AD diagonosis, early DLR-based methods always focused on pre-determined regions of interest prior to deep training, which may hamper diagnostic performance. For example, Khvostikov et al. and Li et al. trained DLR models based on pre-extracted hippocampal regions of MRI and other multimodal neuroimaging data (24, 25). Apart from the above, Basaia et al. used a single cross-sectional MRI scan and deep neural networks to automatically classify AD and MCI, with high accuracies of 98.2% between AD and NC, and of 74.9% from MCI to AD progression (26). Lee et al. also used DLR method for AD classification and achieved the accuracies of 95.35% and 98.74% on different datasets (27). However, there is no existing DLR model for preAD detection.

Therefore, in this study we hypothesized that the DLR method was useful in the diagnosis of PreAD. Considering the hippocampus volume has not been shrunk in AD early stage, we used MRI images of the whole brain for DLR classification. In addition, we also hypothesized that the DLR technique could achieve high classification accuracy in detecting subgroups of preAD from NC, such as APOE ε4+ individuals.

MATERIALS AND METHODS

Figure 1 showed the overall framework of this study, which consisted of six steps: (1) enrolled subjects; (2) imaging preprocessing, including segmentation, normalization and smoothing; (3) basic deep learning (DL) model pre-training, in this step several DL models were pre-trained in order to get the best one for DLR feature extraction; (4) feature extraction and fusion; (5) classification; (6) comparative experiments.

Subjects

The data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database.[1] ADNI is a longitudinal, multicenter study to develop clinical, imaging, genetic and biochemical biomarkers for early detection and tracking of AD. The latest information is available at http://adni.loni.usc.edu/about/.

In this study, we collected 236 NC and 181 preAD data. Demographic data included age, sex, gender, education, neuropsychological assessment tests [Dementia Rating Scale (CDRSB) and Mini-Mental State Examination (MMSE)], Apolipoprotein E (APOE) ε4 and imaging information. T1-MRI and amyloid positron emission tomography (PET) images were selected for all subjects. The preAD group was defined as who standard uptake value ratio (SUVR) value of amyloid PET was >1.18 in whole cerebral cortex (17, 28). Whole cerebellum was used for reference when deriving SUVR. In addition, to validate our proposed DLR model, we enrolled 12 preAD individuals who converted into the MCI state. We selected MRI images in both baseline and MCI stages.

All subjects were divided into two groups, one training/validation group and one independent test group. The training/validation group was from ADNI 1, ADNI 2 and ADNI 3, including 212 NC and 162 preAD subjects. The test group was from ADNI Go, including 24 NC subjects and 19 preAD subjects.

Images Acquisition and Preprocessing

The image acquisition process was described in the ADNI website at http://adni.loni.usc.edu/about/. All MRI data have been evaluated by quality control (QC) at the Mayo Clinic Aging and Dementia Imaging Research Laboratory. The SUVR values of Amyloid PET were downloaded from the ADNI website directly.

The preprocessing of MRI images was performed by statistical parametric mapping (SPM12) software[2] on MATLAB 2016b

[1] https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

[2] https://www.fil.ion.ucl.ac.uk/spm/software/spm12/

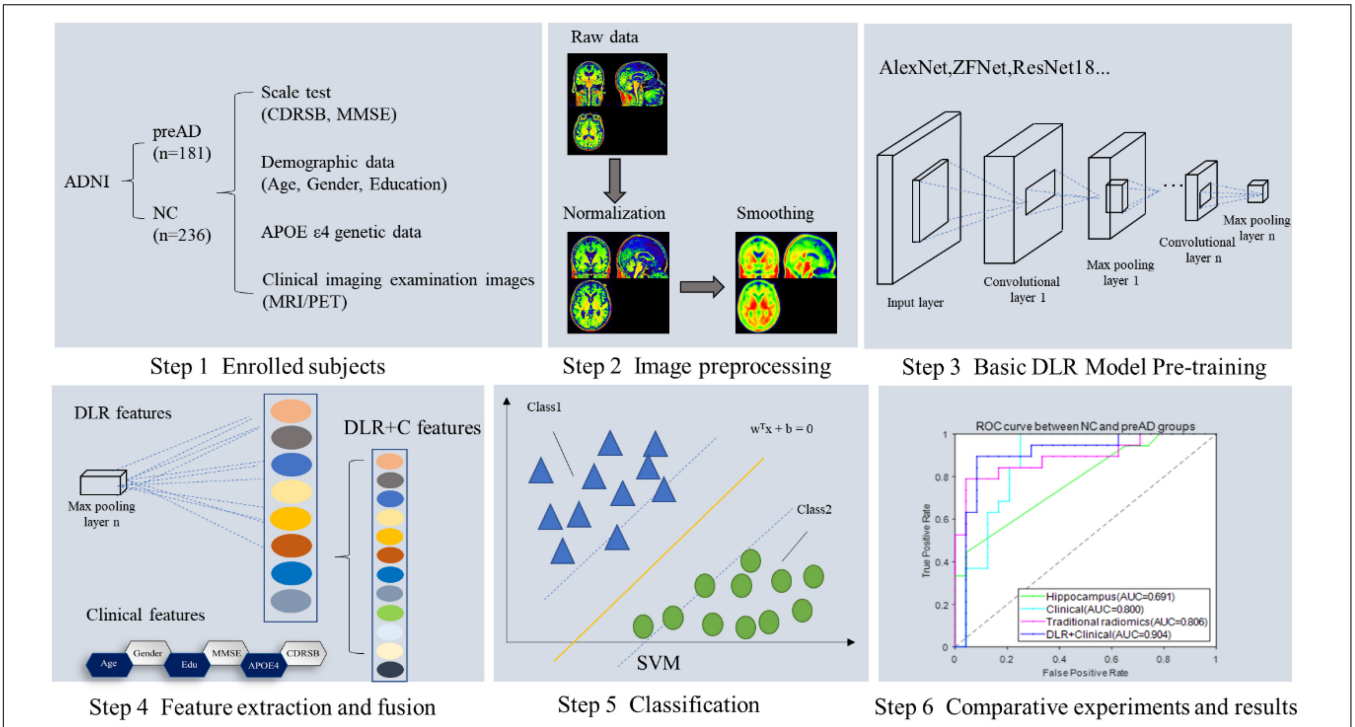

FIGURE 1 | The framework of this study.

platform.[3] First, MRI images were segmented into probabilistic gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF); Then, each GM image was normalized into the Montreal Neurological Institute (MNI) space by diffeomorphic anatomical registration *via* exponentiated lie algebra, and smoothed using an 8-mm Gaussian-smoothing kernel. As a result, each image has a spatial resolution of 91 × 109 × 91 with a voxel size of 2 mm × 2 mm × 2 mm; Finally, in order to adapt and speed up the training of the deep learning model, 3D images were sliced from the axial direction into 91 single-channel images with the size of 91 × 109 to tile 2D images, and then resized into 224 × 224 for normalizing. Each 3D MRI image was tiled into a group of 2D images and resized into 224 × 224 pixels for subsequent DL model training.

The Proposed Deep Learning Radiomics Method

Figure 2 illustrates our proposed DLR method. The method consisted of three parts: (1) Basic DL model pre-training. We used six Convolutional Neural Networks (CNN) networks as candidate DL models and pre-trained them, respectively. After training, we selected one as the final DL model to obtain the DLR features according to the classification results. (2) Feature fusion. To obtain DLR features, we obtained DLR feature maps from the last convolutional layer of the final selected DL model, and extracted the maximum value of each feature map through global max pooling. These extracted features were defined as DLR features and combined with clinical features (sex, education, etc.) as input data for classification. (3) Classification. Based on the above features, the support vector machine (SVM) was used as the classifier to distinguish preAD from NC.

Training for Candidate Deep Learning Models

Six CNN models, including AlexNet, ZFNet, ResNet18, ResNet34, InceptionV3, and Xception, were applied in the training step to define the best training model.

- AlexNet: it is the first CNN network architecture that uses ReLU as the activation function, and uses interleaving pooling technology in CNN (29).
- ZF-Net: it is fine-tuned on the basis of AlexNet. It uses deconvolution to visually analyze the intermediate feature map of CNN and improves model performance by analyzing feature behavior (30).
- Inceptionv3: it improves the CNN model by using convolution decomposition and regularization (31).
- Xception: it improves Inception V3 by using depth wise separable convolution to replace the Inception module (32, 33).
- ResNet: it introduces new network features based on the previous traditional CNN network (34). Several ResNet subtypes were proposed according to different numbers of hidden layers, such as ResNet18, ResNet34, ResNet101, and so on.

As an example, **Figure 3** showed the network structure of the ResNet34 model.

[3]https://www.mathworks.com/products/matlab.html

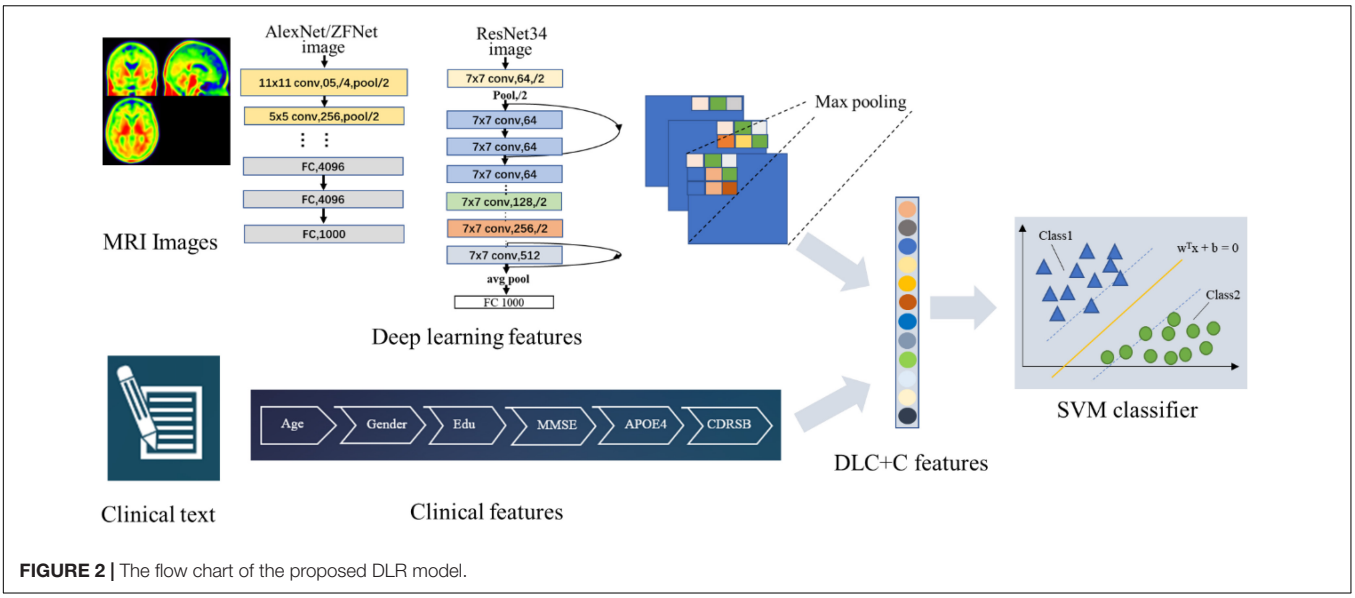

FIGURE 2 | The flow chart of the proposed DLR model.

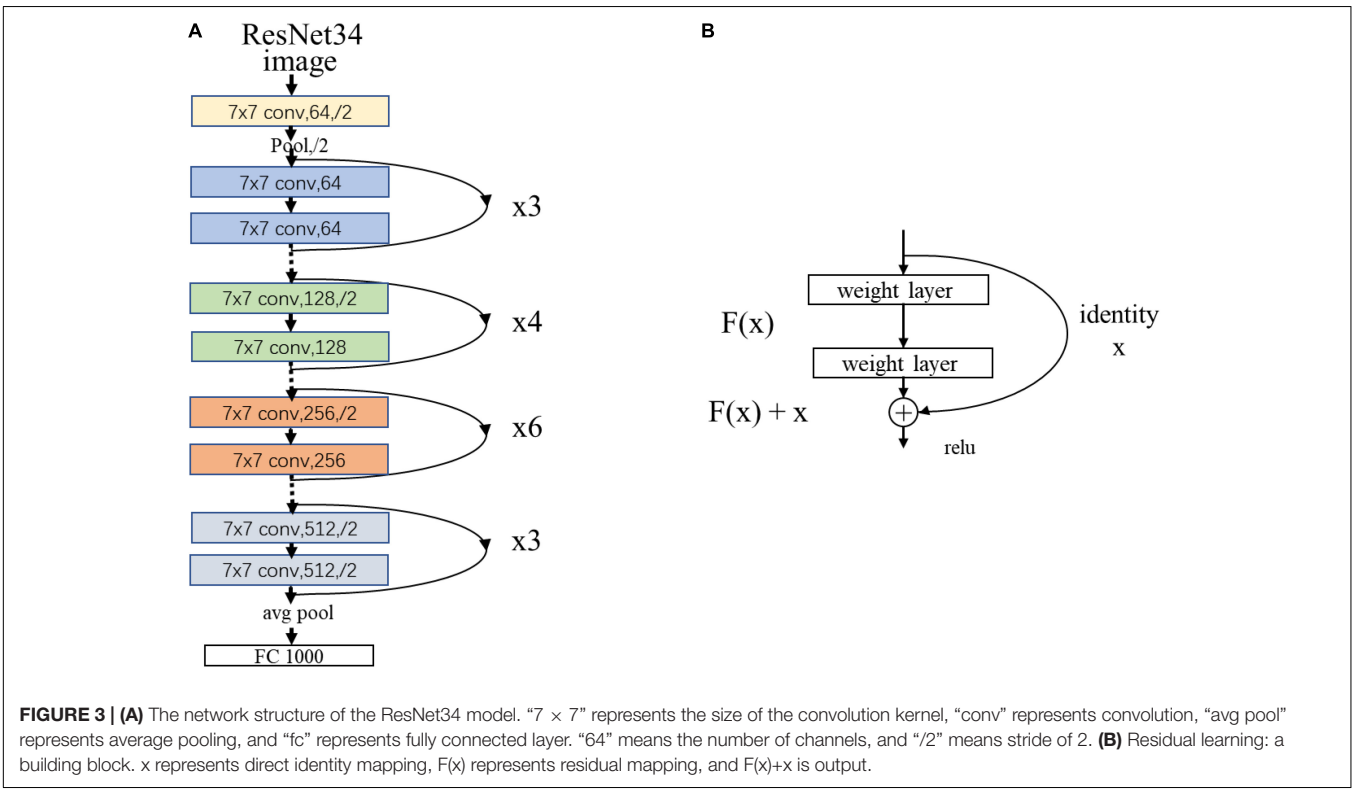

FIGURE 3 | (A) The network structure of the ResNet34 model. "7 × 7" represents the size of the convolution kernel, "conv" represents convolution, "avg pool" represents average pooling, and "fc" represents fully connected layer. "64" means the number of channels, and "/2" means stride of 2. **(B)** Residual learning: a building block. x represents direct identity mapping, F(x) represents residual mapping, and F(x)+x is output.

During the raining step, the selected six models were trained in the training/validation group and tested in the test group. Guided by the test results, we optimized the DL model by tuning hyper parameters.

Classifier

We combined DLR features and clinical information (gender, education, age, etc.) as input data for classification. SVM was used as the classifier. As a classic supervised learning method, SVM has been widely used in statistical classification and regression analysis due to its ability to map vectors to a higher dimensional space that creates a maximum margin hyperplane to achieve high classification performance (35). In this study, we used the linear kernel function in SVM to detect classification reliability and generalization ability.

Comparative Experiments

To demonstrate the superiority of our proposed DLR method, we compared our model and three existing models in comparative

TABLE 1 | Demographic information for subjects.

	Training/validation groups				Test groups				Longitudinal data
	preAD	APOE+	APOE−	NC	preAD	APOE+	APOE−	Baseline	MCI
N	162	70	92	212	19	9	10	16	16
Gender(M/F)	68/94	36/34	32/60	103/109	5/14	3/6	2/8	9/7	9/7
Age(years)	76.3 ± 5.4	75.3 ± 6.3	76.9 ± 4.5	71.8 ± 5.7[b]	75.3 ± 5.1	74.7 ± 6.9	75.9 ± 3.4	71.5 ± 5.8[b]	80.8 ± 5.4
EDU	15.4 ± 3.0	14.9 ± 3.5	15.8 ± 2.5	16.7 ± 2.5[b]	15.4 ± 2.1	16.0 ± 2.4	14.8 ± 1.7	16.13 ± 2.4	16.13 ± 2.4
MMSE	28.7 ± 1.6	28.5 ± 1.6	28.8 ± 1.6	29.1 ± 1.3[b]	28.7 ± 1.3	28.8 ± 1.1	28.6 ± 1.6	29.2 ± 0.9	27.43 ± 2.0
CDRSB	0.3 ± 0.7	0.3 ± 0.8	0.3 ± 0.7	0.2 ± 0.4[b]	0.3 ± 0.9	0.5 ± 1.2	0.1 ± 0.2	0.1 ± 0.2	1.63 ± 0.9
APOE ε4 positive rate	70/162	N/A	N/A	34/212	9/19	N/A	N/A	3/13	3/13

All data except APOEε4 positive rate were presented as mean ± standard deviation. EDU, education; MMSE, Mini-mental State Examination; CDRSB, clinical dementia rating sum of boxes.
[a]Age, Education, MMSE and CDRSB performed a two-sample t-test between NC and preAD groups; Gender performed a Chi-square test between NC and preAD groups.
[b]Means that there was a significant difference ($p < 0.05$) between the preAD group and the NC group in the training/validation group and test group with two-sample t-tests.

TABLE 2 | Classification performance of different DL models in the pre-training step.

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
Training/Validation Groups				
AlexNet	96.28 ± 3.24	94.86 ± 5.88	97.38 ± 2.46	0.962 ± 0.04
ZF-Net	98.18 ± 1.88	97.55 ± 3.50	98.83 ± 1.98	0.980 ± 0.02
ResNet18	95.68 ± 2.66	94.49 ± 4.93	96.58 ± 3.05	0.962 ± 0.03
ResNet34	**96.29 ± 2.54**	**96.62 ± 2.26**	**96.02 ± 3.58**	0.964 ± 0.02
InceptionV3	97.63 ± 2.43	95.91 ± 4.99	98.95 ± 1.35	0.976 ± 0.01
Xception	97.02 ± 3.84	97.62 ± 3.62	96.54 ± 5.15	0.973 ± 0.03
Test Groups				
AlexNet	87.91 ± 3.06	78.95 ± 4.30	95.00 ± 3.83	0.869 ± 0.03
ZF-Net	87.91 ± 2.40	79.47 ± 3.88	94.58 ± 2.01	0.870 ± 0.03
ResNet18	87.67 ± 1.91	84.21 ± 3.50	90.41 ± 2.01	0.872 ± 0.02
ResNet34	**89.53 ± 2.51**	**87.89 ± 2.54**	**90.83 ± 5.12**	0.893 ± 0.03
InceptionV3	84.88 ± 2.26	84.21 ± 3.51	85.42 ± 4.05	0.848 ± 0.03
Xception	88.84 ± 2.14	88.40 ± 3.30	89.17 ± 4.48	0.886 ± 0.04

The bold values indicate classification results of the optimal model ResNet34 for Base DLR Model Selection.

experiments, including: (1) Clinical model: clinical characteristics included demographic data, neuropsychological cognitive assessment results, and APOE ε4 genotyping characteristics of all subjects. (2) Hippocampal model: the hippocampal volumes were used as inputs for the classification; (3) Traditional radiomics model: traditional radiomics features of were extracted for the classification. In this experiment, we extracted features by using the radiomics tool developed by Vallieres et al.[4] We used brain DMN regions as ROIs and performed texture analysis on each input ROI using the "Texture Toolbox" in the Radiomics Toolbox. Feature extraction steps included wavelet bandpass filtering, isotropic resampling, Lloyd–Max quantization and feature calculation. The detailed extraction process of the radiomics features were described in the previous studies (36, 37).

Three comparative experiments were employed in this study: (1) NC vs. preAD; (2) NC vs. preAD APOE+; and (3) NC vs. preAD APOE−. Ten-fold cross-validation was performed with 100 time repetitions. We calculated accuracy, sensitivity, and specificity to evaluate the classification results. The mathematical expressions of the three indicators were as follows:

$$\text{Accuracy} = \frac{TP}{TP + FP}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{FP + TN}$$

Longitudinal Study

The 12 individuals with longitudinal data were used to validate the proposed DLR model. Firstly, we calculated the probability value of SVM classifier, and defined it as the decision score; then we compared the decision scores in both baseline and MCI states in 12 individuals.

Statistical Analysis

In this study, we used two-sample *t*-tests or chi-square tests to compare demographic and clinical characteristics between the

[4]https://github.com/mvallieres/radiomics

TABLE 3 | The classification results of preAD vs. NC.

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)
Training/Validation Groups			
Hippocampal model	76.20 ± 6.05	44.72 ± 10.58	99.05 ± 2.27
Traditional radiomics model	77.01 ± 7.77	62.61 ± 10.31	87.73 ± 9.50
Clinical model	85.66 ± 5.24	83.31 ± 9.56	87.70 ± 7.65
DLR model	**99.40 ± 3.23**	**99.00 ± 4.00**	**99.56 ± 1.65**
Test Groups			
Hippocampal model	72.44 ± 1.46	42.68 ± 2.93	96.09 ± 1.31
Traditional radiomics model	82.00 ± 4.09	68.59 ± 8.35	92.62 ± 4.58
Clinical model	79.65 ± 2.21	82.75 ± 4.24	77.20 ± 2.61
DLR method	**89.85 ± 1.12**	**94.74 ± 0. 10**	**85.98 ± 2.01**

Bold values represent the classification performance of our proposed model.

NC and preAD groups and between the APOE+ and APOE− subgroups. All statistical analyses were performed using SPSS version 22.0 software (SPSS Inc., Chicago, IL, United States) and performed in Matlab2019b (Mathworks Inc., Sherborn, MA, United States). A p-value < 0.05 was considered to be significantly different.

RESULTS

Demographic Information

The results of demographic data were shown in **Table 1**. There was a significant difference in age and years of education between the preAD group and the NC group in the training/validation group ($p < 0.001$), and there was a difference in CDRSB and MMSE (CDRSB: $p = 0.006$, MMSE: $p = 0.003$), while no difference in gender between the two groups. There was no significant difference in gender, education level, CDRSB and MMSE between the preAD group and the NC group in the test group, whereas there was a difference in age ($p = 0.03$).

Pre-training for Candidate Deep Learning Models

Table 2 summarized the classification performance of six candidate DL models, including classification accuracy, sensitivity, and specificity. By comparing the results of the two groups, ResNet34 was selected to be the best model. Therefore, we chose the pre-training ResNet34 model and extracted DLR features for the next step.

Comparative Experiments

Normal Control vs. Preclinical Alzheimer's Disease Group

Table 3 showed the classification results of the four models between NC and preAD groups. Among the four models, the DLR model showed the best classification performance in the test group, with the accuracy of 89.85% ± 1.12%, sensitivity of 94.74% ± 0.1%, and specificity of 85.98% ± 2.01%. The performance of the hippocampal model, traditional radiomics model, and clinical model were all significantly lower than DLR model, with the accuracies of 72.44% ± 1.37%, 82.00% ± 4.09% and 79.65% ± 2.21%, sensitivities of 42.68% ± 2.93%, 68.59% ± 8.35% and 82.754% ± 4.24%, specificities of 96.09% ± 1.31%, 92.62% ± 4.58% and 77.20% ± 2.61%, respectively.

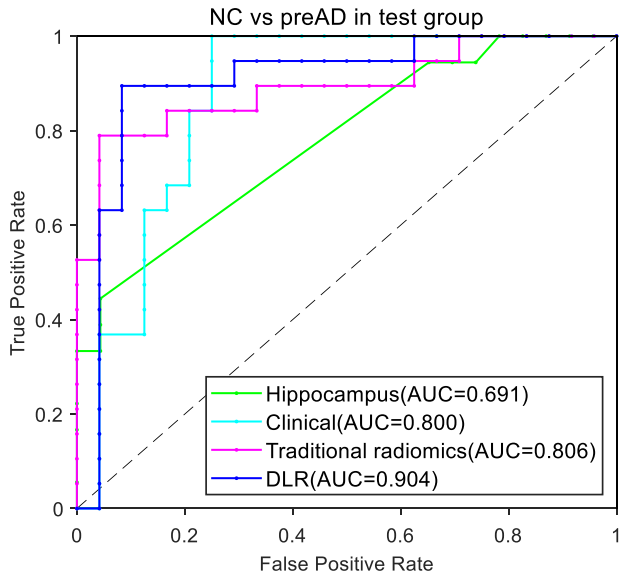

FIGURE 4 | ROC curves of the four models between NC and preAD groups.

TABLE 4 | The classification results of NC vs. preAD APOE+.

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)
Training/Validation Groups			
Hippocampal model	76.90 ± 11.62	49.78 ± 26.80	99.37 ± 5.44
Traditional radiomics model	71.11 ± 10.60	54.42 ± 16.69	83.66 ± 9.33
Clinical model	71.11 ± 10.98	50.80 ± 15.66	84.94 ± 12.35
DLR model	**99.94 ± 0.59**	**99.95 ± 0.01**	**99.88 ± 3.72**
Test Groups			
Hippocampal model	69.00 ± 6.84	30.71 ± 16.94	96.84 ± 15.91
Traditional radiomics model	78.87 ± 5.00	54.42 ± 16.21	83.66 ± 6.72
Clinical model	71.39 ± 4.65	32.84 ± 13.65	96.17 ± 4.37
DLR model	**92.80 ± 2.61**	**88.89 ± 0.01**	**94.47 ± 3.72**

Bold values represent the classification performance of our proposed model.

TABLE 5 | The classification results of NC vs. preAD APOE−.

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)
Training/Validation Groups			
Hippocampal model	76.88 ± 12.86	75.46 ± 23.37	77.83 ± 12.60
Traditional radiomics model	73.50 ± 9.44	73.20 ± 12.45	72.51 ± 12.35
Clinical model	70.28 ± 9.69	60.20 ± 16.61	79.22 ± 12.05
DLR model	**95.74 ± 11.85**	**89.60 ± 10.54**	**98.03 ± 10.76**
Test Groups			
Hippocampal model	63.36 ± 7.42	75.82 ± 24.86	50.90 ± 21.02
Traditional radiomics model	83.87 ± 3.04	78.00 ± 11.35	86.67 ± 6.66
Clinical model	70.10 ± 3.50	62.03 ± 7.93	73.95 ± 7.27
DLR model	**85.45 ± 9.04**	**90.40 ± 9.47**	**83.10 ± 11.66**

Bold values represent the classification performance of our proposed model.

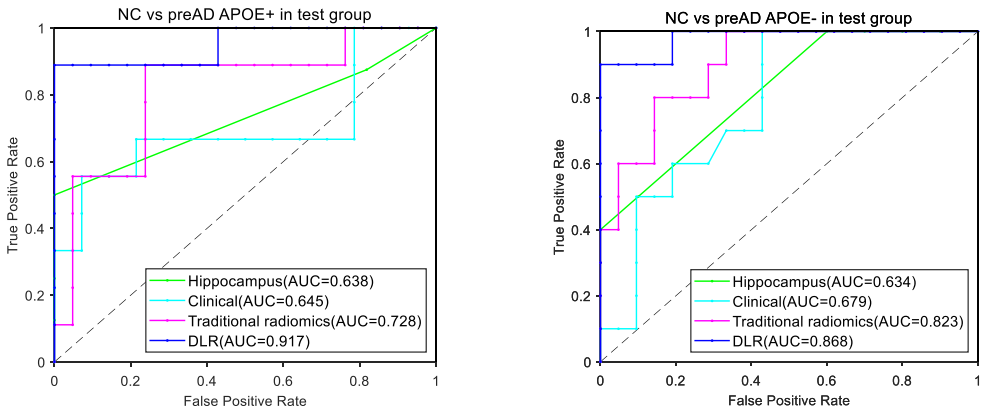

FIGURE 5 | ROC curves of the four models between NC and preAD APOE+ groups **(left)** and between NC and preAD APOE**–** groups **(right)**.

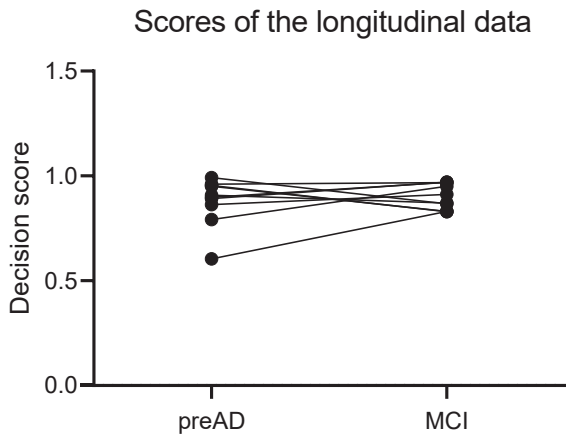

FIGURE 6 | The scores of the longitudinal data in preAD stage and MCI stage.

Figure 4 presented the ROC curves of the four models. The mean AUCs (± SD) for the hippocampal model, traditional radiomics model, clinical model, and DLR model in were 0.691 ± 0.012, 0.806 ± 0.013, 0.800 ± 0.021 and 0.904 ± 0.014, respectively.

Normal Control vs. Preclinical Alzheimer's Disease Subgroups

Table 4 showed the classification results between NC and preAD APOE+ groups. The accuracy, sensitivity and specificity of the DLR model in the test group were 92.80% ± 2.61%, 88.89% ± 0.01%, and 94.47% ± 3.72%. The performance of the hippocampal model, traditional radiomics model, and clinical model were all significantly lower than our proposed model, with the accuracies of 72.44%.

Table 5 showed the classification results between NC and preAD APOE− groups. The accuracy, sensitivity and specificity of the DLR model in the test group were 85.45 ± 9.04%, 90.40% ± 9.47%, and 83.10% ± 11.66%. The performance of the hippocampal model, traditional radiomics model, and clinical model were all significantly lower than our proposed model, with the accuracies of 63.36% ± 7.42%, 83.87% ± 3.04%, and 70.10% ± 3.50%. In **Tables 3**, **4**, the bold values represented the classification performance of the our proposed method.

Figure 5 showed the ROC curves of the four models. The mean AUCs (± SD) for the hippocampal model, traditional radiomics model, clinical model and the best DLR model between NC and preAD APOE+ were 0.638 ± 0.061, 0.728 ± 0.024, 0.645 ± 0.041 and 0.917 ± 0.010, and between NC and preAD APOE− were 0.634 ± 0.075, 0.823 ± 0.041, 0.679 ± 0.042, and 0.868 ± 0.011, respectively.

Longitudinal Study

Figure 6 showed the results of the longitudinal study. The decision scores had a slight upward trend from the PreAD baseline to the MCI stage. The results showed that our model also had a great prediction performance.

DISCUSSION

Currently, DLR is the hot spot and focus of current imaging development. In view of its superiority in disease diagnosis, DLR methods have been successfully applied in tumor genotype prediction, preoperative analysis, prognosis evaluation, and cancer diagnosis, etc., but DLR research for neurological diseases remained lacking. In this study, we proposed a DLR model to distinguish cognitively normal adults at risk of Alzheimer's disease from normal control based on T1-weighted structural MRI images. Compared with other traditional models, such as hippocampal model, clinical model or traditional radiomics model, our proposed DLR model achieved best classification results.

In the comparative experiments, the DLR method achieved the highest accuracy in both training/validation group (99.40% ± 3.23%) and separate test group (89.85% ± 1.12%). Therefore, we proved the robustness of the DLR model.

Currently, several studies have investigated the classification between preAD and NC by using machine learning or traditional quantitative methods. For example, Ding et al. distinguished preAD from NC by investigating the coupling relationship

between glucose and oxygen metabolism from hybrid PET/MRI, with an AUC of 0.787 (38). Li et al. used a voxel-based SSM/PCA method to analyze fluorodeoxyglucose-PET (FDG-PET) images with AUC of 0.815 (39), Li et al. conducted an exploratory study for identifying preAD based on radiomics analysis of MRI and obtained an average accuracy of 83.7% [T. (17)]. In comparison to previous studies, our DLR model achieved the best classification results. The reason can be explained as following: (1) the DLR method can directly extract high-throughput image features from CNN. Since it does not involve additional feature extraction operations, it will not bring additional errors; (2) the results of traditional methods were usually influenced by individual factors and imaging machine parameters; while the DLR method combined DLR image features and clinical information, which partly solved the problems of individual heterogeneity.

To demonstrate the robustness of the proposed DLR model, we performed experiments in the APOE ε4 subgroup analysis. Notably, cerebral amyloid deposition is also affected by the ApoE ε4 genotype (40). Higher levels of amyloid accumulation were observed in SCD subjects with ApoE ε4 carriers than noncarriers (41, 42). Therefore, we proposed to add ApoE ε4 genotype features to further validate the accuracy of the model. Notably, the DLR model achieved better classification results between NC vs. preAD APOE+ (92.80% ± 2.61%) than the two other experiments (89.85% ± 1.12% and 85.45% ± 9.04%). The high sensitivity (88.89% ± 0.01%) and specificity (94.47% ± 3.72%) results also showed that the DLR model was very powerful in identifying cognitively normal adults at risk of Alzheimer's disease.

Although the DLR method could distinguish preAD from NC, it still had some limitations. First, more data was still needed to verify the generality and robustness of the proposed method. In this study, subjects were collected only from the ADNI database. Whether our model was powerful for other racial populations need further exploration. Secondly, we only compared six DL models. Although the Resnet34 model achieved good classification performance, it was unknown whether other DL models beyond the six were more suitable. In addition, we used the whole brain MRI image to train the DLR models in this study. However, future studies were required to explore whether DLR models based on the hippocampus or entorhinal cortex instead were more effective. Furthermore, in this study, 2D DLR models were employed. However, whether 3D DLR models could achieve better classification performances need further exploration. Finally, the proposed DLR model was based on T1-MRI images. It may be possible to improve the classification performance of DLR by combining other imaging modals, such as FDG PET, amyloid PET and tau PET images.

CONCLUSION

We proposed a DLR method based on T1-MRI images to discriminate preAD and NC. The results demonstrated that our proposed DLR method could improve diagnostic performance. The DLR method had potentials for clinical applications in the future.

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

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

JJ conceived and designed the experiments, analyzed and interpreted the data, and wrote the manuscript. JZ performed the experiments, interpreted the data, and wrote the manuscript. ZL and LL performed the experiments and wrote the manuscript. BH conceived and designed the experiments, provided research funding, analyzed and interpreted the data, and reviewed the manuscript. All authors read and approved the final version of the article for publication.

FUNDING

This article was supported by grants from the Shanghai Pudong New Area Health System Leading Personnel Training Program (PWRl2017-04) and Discipline Construction of Pudong New Area Health Committee (PWGw2020-01).

ACKNOWLEDGMENTS

Data collection and sharing for this project were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI; National Institutes of Health Grant U01 AG024904) and DODADNI (Department of Defense, award number W81XWH-12–2-0012). ADNI is funded by the National Institute of Aging and the National Institute of Biomedical Imaging and Bioengineering and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica Inc; Biogen; Bristol-Myers Squibb Company; CereSpir Inc.; Eisai Inc.; Elan Pharmaceuticals Inc; Eli Lilly and Company; EuroImmun; F.Hoffmann-La Roche Ltd. and its affiliated company Genentech Inc; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC; NeuroRx Research; Neu-rotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada.

REFERENCES

1. Laurent C, Buée L, Blum D. Tau and neuroinflammation: what impact for alzheimer's disease and tauopathies? *Biomed J.* (2018) 41:21–33. doi: 10.1016/j.bj.2018.01.003
2. Dubois B, Hampel H, Feldman H, Scheltens P, Aisen P, Andrieu S, et al. Preclinical alzheimer's disease: definition, natural history, and diagnostic criteria. *Alzheimers Dement.* (2016) 12:292–323. doi: 10.1016/j.jalz.2016.02.002
3. Jack C, Bennett D, Blennow K, Carrillo M, Dunn B, Haeberlein S, et al. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimers & Dementia* (2018) 14:535–62. doi: 10.1016/j.jalz.2018.02.018
4. Li T-R, Yang Q, Hu X, Han Y. Biomarkers and tools for predicting alzheimer's disease at the preclinical stage. *Curr Neuropharmacol.* (2022) 20:713–37. doi: 10.2174/1570159X19666210524153901
5. Alzheimer's Association. 2017 Alzheimer's disease facts and figures. *Alzheimers Dement.* (2017) 13:325–73. doi: 10.1016/j.jalz.2017.02.001
6. Jessen F, Amariglio RE, Buckley RF, van der Flier WM, Han Y, Molinuevo JL, et al. The characterisation of subjective cognitive decline. *Lancet Neurol.* (2020) 19:271–8. doi: 10.1016/S1474-4422(19)30368-0
7. De Santi S, de Leon M, Rusinek H, Convit A, Tarshish C, Roche A, et al. Hippocampal formation glucose metabolism and volume losses in MCI and AD. *Neurobiol Aging.* (2001) 22:529–39. doi: 10.1016/S0197-4580(01)00230-5
8. Gyasi Y, Pang Y, Li X, Gu J, Cheng X, Liu J, et al. Biological applications of near infrared fluorescence dye probes in monitoring alzheimer's disease. *Eur J Med Chem.* (2020) 187:111982. doi: 10.1016/j.ejmech.2019.111982
9. Jagust W. Imaging the evolution and pathophysiology of alzheimer disease. *Nat Rev Neurosci.* (2018) 19:687–700. doi: 10.1038/s41583-018-0067-3
10. Johnson K, Fox N, Sperling R, Klunk W. Brain imaging in alzheimer disease. *Cold Spring Harb Perspect Med.* (2012) 2:a006213. doi: 10.1101/cshperspect.a006213
11. Suk H, Lee S, Shen D, Alzheimer's Disease Neuroimaging Initiative. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *Neuroimage.* (2014) 101:569–82. doi: 10.1016/j.neuroimage.2014.06.077
12. Dickerson B, Salat D, Greve D, Chua E, Rand-Giovannetti E, Rentz D, et al. Increased hippocampal activation in mild cognitive impairment compared to normal aging and AD. *Neurology.* (2005) 65:404–11. doi: 10.1212/01.wnl.0000171450.97464.49
13. Devanand D, Pradhaban G, Liu X, Khandji A, De Santi S, Segal S, et al. Hippocampal and entorhinal atrophy in mild cognitive impairment—prediction of alzheimer disease. *Neurology.* (2007) 68:828–36. doi: 10.1212/01.wnl.0000256697.20968.d7
14. Zhao K, Ding Y, Wang P, Dou X, Zhou B, Yao H, et al. Early classification of alzheimer's disease using hippocampal texture from structural MRI. In: Krol A, Gimi B editors. *Conference of the SPIE Medical Imaging*. (Vol. 10137), Orlando, FA: (2017). doi: 10.1117/12.2254198
15. Shu Z-Y, Mao D-W, Xu Y, Shao Y, Pang P-P, Gong X-Y. Prediction of the progression from mild cognitive impairment to Alzheimer's disease using a radiomics-integrated model. *Ther Adv Neurol Disord.* (2021) 14:17562864211029552. doi: 10.1177/17562864211029551
16. Zhou H, Jiang J, Lu J, Wang M, Zhang H, Zuo C, et al. Dual-model radiomic biomarkers predict development of mild cognitive impairment progression to alzheimer's disease. *Front Neurosci.* (2019) 12:1045. doi: 10.3389/fnins.2018.01045
17. Li T, Wu Y, Jiang J, Lin H, Han C, Jiang J, et al. Radiomics analysis of magnetic resonance imaging facilitates the identification of preclinical alzheimer's disease: an exploratory study. *Front Cell Dev Biol.* (2020) 8:605734. doi: 10.3389/fcell.2020.605734
18. Liu Z, Wang S, Dong D, Wei J, Fang C, Zhou X, et al. The applications of radiomics in precision diagnosis and treatment of oncology: opportunities and challenges. *Theranostics.* (2019) 9:1303–22. doi: 10.7150/thno.30309
19. Suzuki K. Overview of deep learning in medical imaging. *Radiol Phys Technol.* (2017) 10:257–73. doi: 10.1007/s12194-017-0406-5
20. Li Y, Wei D, Liu X, Fan X, Wang K, Li S, et al. Molecular subtyping of diffuse gliomas using magnetic resonance imaging: comparison and correlation between radiomics and deep learning. *Eur Radiol.* (2022) 32:747–58. doi: 10.1007/s00330-021-08237-6
21. Park J, Kickingereder P, Kim H. Radiomics and deep learning from research to clinical workflow: neuro-oncologic imaging. *Korean J Radiol.* (2020) 21:1126–37. doi: 10.3348/kjr.2019.0847
22. Yang L, Xu P, Zhang Y, Cui N, Wang M, Peng M, et al. A deep learning radiomics model may help to improve the prediction performance of preoperative grading in meningioma. *Neuroradiology.* (2022). doi: 10.1007/s00234-022-02894-0
23. Wang Y, Shao Q, Luo S, Fu R. Development of a nomograph integrating radiomics and deep features based on MRI to predict the prognosis of high grade Gliomas. *Mathe Biosci Eng.* (2021) 18:8084–95. doi: 10.3934/mbe.2021401
24. Khvostikov A, Aderghal K, Benois-Pineau J, Krylov A, Catheline G. 3D CNN-based classification using sMRI and MD-DTI images for alzheimer disease studies. *arXiv* [Preprint] (2018). doi: 10.48550/ARXIV.1801.05968
25. Li H, Habes M, Fan Y. Deep ordinal ranking for multi-category diagnosis of alzheimer's disease using hippocampal MRI data. *arXiv* [Preprint] (2017). doi: 10.48550/ARXIV.1709.01599
26. Basaia S, Agosta F, Wagner L, Canu E, Magnani G, Santangelo R, et al. Automated classification of alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *Neuroimage Clin.* (2019) 21:101645. doi: 10.1016/j.nicl.2018.101645
27. Lee B, Ellahi W, Choi J. Using deep CNN with data permutation scheme for classification of alzheimer's disease in structural magnetic resonance imaging (sMRI). *Ieice Trans Inform Syst.* (2019) E102D:1384–95. doi: 10.1587/transinf.2018EDP7393
28. Fakhry-Darian D, Patel NH, Khan S, Barwick T, Win Z. Optimisation and usefulness of quantitative analysis of 18 F-florbetapir pet. *Br J Radiol.* (2019) 92:20181020. doi: 10.1259/bjr.20181020
29. Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. *Commun Acm.* (2017) 60:84–90. doi: 10.1145/3065386
30. Zeiler M, Fergus R. Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T editors. *Proceedings of the Computer Vision, ECCV 2014 - 13th European Conference*. (Vol. 8689), Cham: (2014). p. 818–33. doi: 10.1007/978-3-319-10590-1_53
31. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z, Ieee. Rethinking the inception architecture for computer vision. *Conference of the Computer Vision and Pattern Recognition 2016*. Las Vegas, NV: (2016). p. 2818–26. doi: 10.1109/CVPR.2016.308
32. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Bach F, Blei D editors. *Proceedings of the 32nd International Conference on International Conference on Machine Learning*. (Vol. Vol. 37), Stroudsburg, PA: (2015). p. 448–56.
33. 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 (CVPR)*. Boston, MA: (2015). p. 1–9. doi: 10.1109/cvpr.2015.7298594
34. He K, Zhang X, Ren S, Sun J, Ieee. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV: (2016). p. 770–8. doi: 10.1109/CVPR.2016.90
35. Amari S, Wu S. Improving support vector machine classifiers by modifying kernel functions. *Neural Netw.* (1999) 12:783–9. doi: 10.1016/S0893-6080(99)00032-5
36. Gillies R, Kinahan P, Hricak H. Radiomics: images are more than pictures. they are data. *Radiology.* (2016) 278:563–77. doi: 10.1148/radiol.2015151169
37. Kumar V, Gu Y, Basu S, Berglund A, Eschrich S, Schabath M, et al. Radiomics: the process and the challenges. *Magn Reson Imaging.* (2012) 30:1234–48. doi: 10.1016/j.mri.2012.06.010
38. Ding C, Du W, Zhang Q, Wang L, Han Y, Jiang J. Coupling relationship between glucose and oxygen metabolisms to differentiate preclinical alzheimer's disease and normal individuals. *Hum Brain Mapp.* (2021) 42:5051–62. doi: 10.1002/hbm.25599
39. Li T, Dong Q, Jiang X, Kang G, Li X, Xie Y, et al. Exploring brain glucose metabolic patterns in cognitively normal adults at risk

of alzheimer's disease: a cross-validation study with Chinese and ADNI cohorts. *Neuroimage-Clin.* (2022) 33:102900. doi: 10.1016/j.nicl.2021.102900

40. Moreno-Grau S, Rodriguez-Gomez O, Sanabria A, Perez-Cordon A, Sanchez-Ruiz D, Abdelnour C, et al. Exploring APOE genotype effects on alzheimer's disease risk and amyloid beta burden in individuals with subjective cognitive decline: The fundacioace healthy brain initiative (FACEHBI) study baseline results. *Alzheimers Dement.* (2018) 14:634–43. doi: 10.1016/j.jalz.2017.10.005
41. Risacher S, Kim S, Nho K, Foroud T, Shen L, Petersen R, et al. APOE effect on alzheimer's disease biomarkers in older adults with significant memory concern. *Alzheimers Dement.* (2015) 11:1417–29. doi: 10.1016/j.jalz.2015.03.003
42. Yi D, Lee D, Sohn B, Choe Y, Seo E, Byun M, et al. Beta-amyloid associated differential effects of apoe epsilon 4 on brain metabolism in cognitively normal elderly. *Am J Geriatr Psychiatry.* (2014) 22:961–70. doi: 10.1016/j.jagp.2013.12.173

Conflict of Interest: The authors declare that this study was conducted without any commercial or financial relationships that could be construed as potential conflicts of interest.

frontiers | Frontiers in Medicine

ORIGINAL RESEARCH
published: 25 April 2022
doi: 10.3389/fmed.2022.872824

The Long-Term Outcome in a Cohort of 52 Patients With Symptomatic Intramedullary Spinal Cavernous Hemangioma After Microsurgery and Emergency Rescue Surgery

Yu Duan[1], *Renling Mao*[1], *Xuanfeng Qin*[2], *Yujun Liao*[2], *Jian Li*[1] *and Gong Chen*[2]*

[1] *Department of Neurosurgery, Huadong Hospital, Fudan University, Shanghai, China,* [2] *Department of Neurosurgery, Huashan Hospital, Fudan University, Shanghai, China*

OPEN ACCESS

Edited by:
Ming-Chin Lin,
Taipei Medical University, Taiwan

Reviewed by:
Mirza Pojskic,
University Hospital of Giessen and Marburg, Germany
Yao-Chin Wang,
Taipei Medical University, Taiwan

***Correspondence:**
Gong Chen
onlycg123@163.com

Specialty section:
This article was submitted to Family Medicine and Primary Care, a section of the journal Frontiers in Medicine

Received: *10 February 2022*
Accepted: *14 March 2022*
Published: *25 April 2022*

Citation:
Duan Y, Mao R, Qin X, Liao Y, Li J and Chen G (2022) The Long-Term Outcome in a Cohort of 52 Patients With Symptomatic Intramedullary Spinal Cavernous Hemangioma After Microsurgery and Emergency Rescue Surgery. Front. Med. 9:872824. doi: 10.3389/fmed.2022.872824

Background: Surgery is the mainstay treatment for patients with symptomatic intramedullary spinal cavernous hemangioma (ISCH), however the time of surgical intervention remains controversial. In this study, we proposed emergency rescue surgery (ERS) for patients in deteriorative type. The prognostic factors of patients with ISCH after microsurgery and the clinical effect of ERS were analyzed.

Methods: From January 2013 to November 2019, 52 patients with symptomatic ISCH treated by microsurgical treatment were collected, ranging in age from 17 to 66 years old (mean: 45.8 ± 13.5 years). The course of the disease ranged from 2 days to 20 years. Of 52 lesions, 17 lesions were in the cervical segment, 25 in the thoracic segment, and 10 in the lumbosacral segment; while seven cases were at the ventral surface, 25 cases at the dorsal surface, and 20 cases at the central spinal cord. The sagittal diameter ranged from 1 to 58 mm (median: 17.3 mm). The transverse diameter ratio ranged from 20 to 80% (median: 50.7%). Thirty-two patients were diagnosed as deteriorative type and 22 were treated by ERS.

Results: At 12 months after surgery, all patients were followed up, and no residual or recurrence was found in all patients. Twenty-five patients (48.1%) showed spinal cord functional improvement after surgery; 25 (48.1%) had no functional change; 2 (3.8%) got worse. For deteriorative patients, ERS group had a significantly higher improvement rate than the non-ERS group ($\chi^2 = 5.393$, $P = 0.02$); For all 52 patients, the factors as a lesion at the ventral surface ($Z = 10.453$, $P = 0.015$), or lumbosacral segment ($\chi^2 = 9.259$, $P = 0.010$) and longer course of disease ($Z = -2.021$, $P = 0.043$) were potential risks in functional recovery in univariate analysis; and in multiple-factor analysis, the lesion at the lumbosacral segment (OR = 4.004, 95% CI: 1.341∼11.961, $P = 0.013$) was the independent risk factors for the functional recovery.

Conclusions: Microsurgical resection is safe and effective for symptomatic ISCH. The ERS is an effective way to improve deteriorative patients' spinal cord function at long-term follow-up. The lesion at the lumbosacral segment is one of the poor prognostic factors.

Keywords: intramedullary spinal cavernous hemangioma, spinal cord, spinal cavernous hemangioma, central nervous system, prognosis, microsurgery, salvage therapy

INTRODUCTION

Intramedullary spinal cavernous hemangioma (ISCH) is an uncommon spinal vascular disease, accounting for 5–15% of spinal vascular malformations (1–3). The natural history of symptomatic ISCH is not completely understood, although most ISCH has a benign clinical course, the annual rate for a first hemorrhage could be up to 4.5% per year and the annual rate for recurrent hemorrhage would be up to 66% (4).

For asymptomatic or small (1–3 mm) ISCH, conservative treatment might be optimal due to surgery-related complications (5, 6), and the patients with hemorrhagic cavernomas should consider surgical intervention, which prevents recurrent hemorrhage and further neurologic deterioration (7–11). The duration means the time from the onset of symptoms to surgery, which varies greatly at different centers, from several hours to several decades (12, 13). Some surgeons believe it is best to allow the neurological symptoms to plateau, to prevent further damage to viable tissue (14); while others believe the risk of rebleeding is too high to wait (15, 16). With the advancement of surgical skills and the continuous accumulation of experience, surgical excision is more active, and duration has been constantly shortening, and the patients with the shorter duration of presurgical symptoms ($\leq$ 3 months) have better clinical outcomes. However, the most surgeon still do not take operation timing seriously, and there are still few studies about the relationship between operation timing and clinical prognosis.

We believe that if surgical resection and laminectomy are performed as soon as possible, it will effectively alleviate spinal edema and avoid deterioration of spinal function. In 2013, according to our new clinical classification, we proposed emergency rescue surgery (ERS) for treating patients in deteriorative type. The present study was conducted to evaluate long-term outcomes in a cohort of 52 patients with symptomatic ISCH after microsurgery and to study the clinical effects of deteriorative patients with ERS.

MATERIALS AND METHODS

Patients and Study Design

From January 2013 to January 2019, the patients with symptomatic ISCH in two neurosurgery centers were analyzed. Inclusion criteria: 1. Diagnosed by ISCH surgically and pathologically; and 2. Over 14 years old. Exclusion criteria: 1. Recurrence of ISCH after surgery; 2. Multiple ISCHs with brain function impaired; 3. Extramedullary (roots) lesions; 4. After radiotherapy (e.g., gamma knife); and 5. Loss to follow-up. After the exclusion of 16 cases, 52 patients (28 males and 24 females) were included. The onset age ranged from 15 to 80 years (mean: 45.8 $\pm$ 14 years) and the course of disease ranged from 2 days to 20 years (median: 12 days). Twenty-eight patients suffered from muscle weakness or dyskinesia, 34 patients suffered from paresthesia (20 cases felt pain, and 14 felt numb); 33 patients suffered from bowel and/or bladder dysfunction. Nine patients (17.3%) had multiple intracranial or intramedullary ISCHs, and 6 patients (11.5%) had a familial history of ISCH.

Preoperative MRI or DSA

All patients were examined by MRI scan and enhancement. Spinal angiography would be performed to exclude other types of vascular malformations if necessary. There were 17 lesions in the cervical segment, 25 lesions in the thoracic segment, and 10 lesions in the lumbosacral segment. Sagittal length: 1$\sim$62 mm, with an average of 15.8 $\pm$ 9.8 mm. Horizontal transverse diameter ratio (maximum diameter at the horizontal position of the lesion/spinal cord diameter of the lesion): 18%$\sim$80% (49.4% $\pm$ 16.8); Horizontal position: seven cases were in ventral surface, 25 cases on the dorsal surface, and 20 cases in center.

Clinical Course Classification

The new clinical classification was based on Ogilvy types (15). In this study, four subtypes (A1, B1, B2, B3, and B4) were divided into the acute course (Type A) and chronic course (Type B). Type A: acute onset of symptoms with rapid decline; Type B1: repeating deterioration of neurological decline with acute onset; Type B2: acute onset of mild symptoms with subsequent gradual decline lasting weeks to months; Type B3: discrete episodes of neurological deterioration with varying degrees of recovery between episodes. Types A and B1 are defined as acute and chronic deteriorative types, respectively. Types B2 and B3 are defined as chronic repetitive types (**Figure 1**).

Emergency Rescue Surgery (ERS)

The patients with deteriorative types were suggested by ERS (**Figure 1**). The ERS was defined as: time interval between the day of the first acute onset to the day of operation is $<$3 days for patients with Type A, and the time interval between the last acute onset to the day of operation is $<$7 days for patients with Type B1.

Surgical Key Point

According to location, different approaches were adopted. If the lesion was visible on the surface, it could be removed directly (**Figure 2**). If the lesion was close to the center and in deep, the posterior midline approach was adopted (**Figure 3**). Somatosensory-evoked potentials (SEPs) and motor-evoked potentials (MEPs) were monitored during surgery.

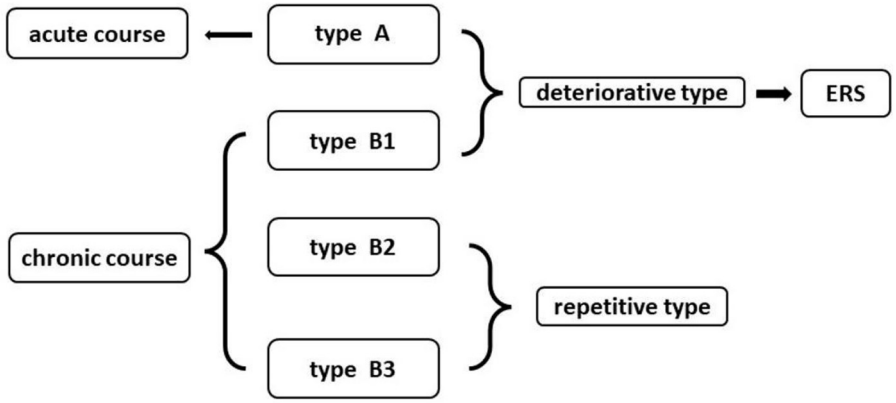

FIGURE 1 | The new clinical classification of intramedullary spinal cavernous hemangioma (ISCH) and its relationship with emergency rescue surgery (ERS).

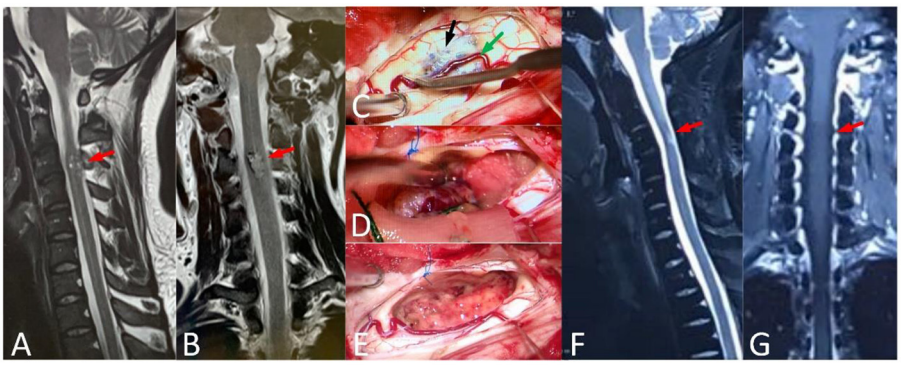

FIGURE 2 | A 37-year-old male, with numbness of limbs and trunk for 5 years, and standing difficulty for 5 days (mALs = 13 points, severe disorder, Type A), was treated by ERS. **(A,B)** Preoperative spinal MRI examination revealed ISCH in C3 (red arrow). **(C–E)** The lesion was visible on spinal cord surface (black arrow) and the artery (green arrow) was protected carefully. **(F,G)** MRI re-examination showed no residual or recurrence in the operative area at 4 years after surgery and the mAS was 7 points at the last follow-up.

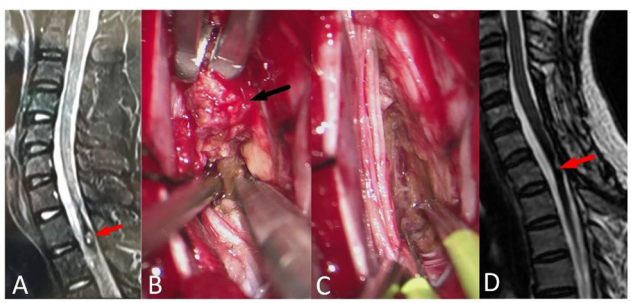

FIGURE 3 | A 15-year-old female with walking difficulty for 2 months and aggravating for 5 days (mALs = 12 points, severe disorder, Type B1), was treated by ERS. **(A)** Preoperative spinal MRI examination revealed ISCH in T1-T2 (red arrow). **(B,C)** The lesion was resected in pieces (black arrow) by posterior central approach. **(D)** MRI reexamination showed no residual or recurrence at 2 years after surgery and the mAS was 8 points at the last follow-up.

Neurofunctional Assessment and Follow-Up

The patient's spinal cord function was evaluated by modified Aminoff-Logue scale (mAL)-excellent: normal or normal, $\leq$2 points; mild disorder, 3 $\sim$ 5 points; moderate disorder, 6$\sim$8; severe disorder, $\geq$9 points. The improvement or degradation of neurological function was defined when the score was changed at least one grade and in the excellent group, the improvement was defined by patients' subjective feeling or limb muscle strength improving. Clinical follow-up was conducted after 12 months and every one year after surgery, including MRI examination and mALs assessment.

Statistical Analysis

The SPSS23.0 statistical software package was used to process data. Fisher's exact test was used for the comparison of the rates between groups. The normal distribution measurement

data (showed by mean and standard deviation) were tested by the *t*-test and the non-normal distribution measurement data (showed by median and quartile spacing) were tested by the rank-sum test (Mann-Whitney *U*-test). The improvement of factors at the neurological function was analyzed by multivariate logistic regression analysis. The $P < 0.05$ was considered statistically significant.

RESULTS

Postoperative Imaging

Patients have received an MRI plain scan and enhanced at follow-up, and no residual or recurrence of ISCH was found in the operative area at all 52 patients at the last follow-up.

ERS

Among the 32 patients with deteriorative type, the average time interval between the day of onset to the day of surgical intervention was 12.2 ± 17.9 days (2∼90 days), 22 patients were treated by ERS, and 10 patients did not receive ERS. In ERS group, 16 (68.8%, 22/32) patients showed neurofunctional improvement at long-term follow-up, and in the non-ERS group, only 2 (20%, 2/10) patients showed improvement (**Table 1**). There was a significant difference in improvement rate between the two groups ($\chi^2 = 7.767$, $P = 0.005$).

The Prognostic Factors After Microsurgical Intervention

Of 52 patients, eight patients (15.4%, 8/52) showed a decrease in mAL score after surgery, but most were transient and six had recovered to the preoperative state within two months after surgery. At 12 months after surgery, 48.1% (25/52) patients showed improvement, 48.1% (25/52) had no changed and 3.8% (2/52) got worse at mAL score. Lesion in ventral surface ($Z = 10.453$, $P = 0.015$), at lumbosacral segment ($\chi^2 = 9.259$, $P = 0.010$) and longer course of disease ($Z = -2.021$, $P = 0.043$) were potential poor factors in functional recovery in univariate analysis. In multiple-factor analysis, the lesion at lumbosacral segment (OR = 4.004, 95% CI: 1.341∼11.961, $P = 0.013$) were the independent risk factors for the functional recovery (**Table 2**).

Two Cases With Postoperative Aggravation

Case 1-A 63 years old male suffering from the sudden loss of muscle strength and bowel and urine dysfunction for 12 days. The lesion was located at T11∼T12 by MRI test. The patient was judged to Type A by our new clinical classification and evaluated at 12 points by mAL scale before surgery. During the surgery, the amplitude is permanently <20% by electrophysiologic monitoring. After 1 month of surgery, spinal cord function decreased to 13 points and had not improved at the last follow-up.

TABLE 1 | Clinical data between ERS group and non-ERS group.

	ERS group ($n = 22$)	Non-ERS group ($n = 10$)	Test value	*P* value
Age (mean ± SD, years)	46.4 ± 14.8	47.2 ± 9.1	2.362[a]	0.878
Male (*n*,%)	10 (46.7%)	8 (80.0%)	3.334[b]	0.068
Deteriorative type				
Acute deteriorative type (*n*,%)	16	6	0.518[b]	0.472
Chronic deteriorative type (*n*,%)	6	4		
Segment			1.715[b]	0.424
Cervical segment	7	2		
Thoracic segment	10	7		
Lumbosacral segment	5	1		
Horizontal position			4.368[b]	0.113
Dorsal surface	12	8		
Center	6	3		
Ventral surface	0	3		
Transverse diameter ratio (%)	49% (40%-55.3%)	50% (45.3%-64%)	−0.636[c]	0.535
Sagittal length (mm)	12 (11-20)	15.0 (12.8-17.0)	−0.986[c]	0.324
Family history (*n*,%)	1 (6.7%)	1 (8.3%)	0.027[b]	0.869
Multiple lesions (*n*,%)	1 (6.7%)	1 (8.3%)	0.027[b]	0.869
Preoperative mALs	8 (3-12)	3 (2.8-10.3)	−1.555[c]	0.120
Onset to operation (d)	3 (3-6)	15 (11-37)	−4.606[c]	<0.001
Prognosis			7.767[b]	0.005
Improvement	16	2		
No change	5	7		
Deterioration	1	1		

[a]T value, [b]χ^2 value; [c]Z value. The normal distribution measurement data were showed by mean and standard deviation and the non-normal distribution measurement data were showed by median and quartile spacing.

TABLE 2 | Univariate analysis and multi-factor regression analysis of spinal cord function recovery after operation in 52 patients with ISCH.

Variable	Univariate analysis				Multi-factor logistic regression anlysis	
	Outcome		Test value	*P value*	Odds ratio (95% confidence interval)	*P value*
	Improve (23)	No-Improve (25)				
Age (mean ± SD, years)	41.9 ± 15.7	49.4 ± 12.7	−1.999[a]	0.051		
Male (*n*, %)	11 (44.0%)	17 (63.0%)	1.878[b]	0.171		
Clinical presentation			2.720[b]	0.437		
Type A	13	9				
Type B1	5	5				
Type B2	4	9				
Type B3	3	4				
Segment			9.259[b]	0.010	4.004 (1.341∼11.961)	0.013
Cervical segment	12	5				
Thoracic segment	12	13				
Lumbosacral segment	1	9				
Horizontal position			6.264[b]	0.044	1.457 (0.805∼2.636)	0.213
Dorsal surface	16	9				
Center	8	12				
Ventral surface	1	6				
Transverse diameter ratio (%)	50% (40–57%)	56% (30–66%)	−0.018[c]	0.985		
Sagittal length (mm)	15 (11–19.5)	14 (11.8–30)	−0.350[c]	0.727		
Family history (*n*,%)	3 (12.0%)	1 (7.1%)	0.23[b]	0.632		
Multiple lesions (*n*,%)	3 (12.0%)	2 (14.3%)	0.042[b]	0.838		
Onset to operation (d)	7 (3–25)	10 (5–80)	−2.021[c]	0.043	1.001 (0.998∼1.004)	0.409
Preoperative mALS	4 (3–12)	3 (2–12)	−0.080[c]	0.936		

[a]T value, [b]χ^2 value; [c]Z value.

The normal distribution measurement data were showed by mean and standard deviation and the non-normal distribution measurement data were showed by median and quartile spacing.

Case 2-A 50 years old male suffering from episodic and recurrent pain in both lower limbs for 1 year. The lesion was located at T12 by MRI test and was evaluated at eight points by mAL scale before surgery and judged to Type B3 by our new clinical classification. During the surgery, the amplitude of electrophysiologic monitoring had no abnormity. After 1 month of surgery, spinal cord function decreased to 11 points and had improved at 10 points 1 year after surgery.

DISCUSSION

Surgery is the mainstay treatment for ISCH (18), which can eliminate the risk of subsequent hemorrhage (19), and prevent further neurological decline (20). However, the timing for surgery has been argued for decades (20). In this study, we firstly proposed ERS intervention for ISCH patients with clinical progression and further confirmed most patients could benefit from ERS compared with non-ERS. As we believe, a wide range of symptoms to either an acute hemorrhage forming a space-occupying lesion, or by edema can lead to a progressive or acute decline in neurological function (3, 21, 22), and choose to evacuate the clot early to relieve compression (17, 23); whereas, another surgeon still believed the timing should be postponed for several weeks because it would help resolve the hematoma, diminishing spinal cord swelling, and creating a discrete border on the lesion itself (12).

According to previous literature, the median duration of primary symptoms to referral was 6.5 months (24), the mean duration from primary symptoms to subsequent hemorrhage or deteriorative symptoms was 1.42 years and the mean duration from primary symptoms to surgery was 2.1 years (25). It means those patients with deteriorative symptoms may not be treated by microsurgery timely at most neurosurgical centers. For meta-analysis research, earlier timing for surgery was beneficial for neurological function (26) and Zhang reported that most pediatric patients presented with acute symptoms and they can benefit from surgery at the acute phase of neurological deterioration (27). In this study, the longer course of the disease was also one of the potential negative factors for recovery. From 2013, the ERS system for patients with symptomatic ISCH has been built at our two neurosurgical centers, cooperating with departments of emergency, imaging, and surgery, specifically for deteriorative patients (A and B1). In this study, 16 patients (68.8%) in ERS group showed neurofunctional improvement, and the rate at the non-ERS group was only 20%, which verified ERS could be beneficial for recovery of deteriorative type patients.

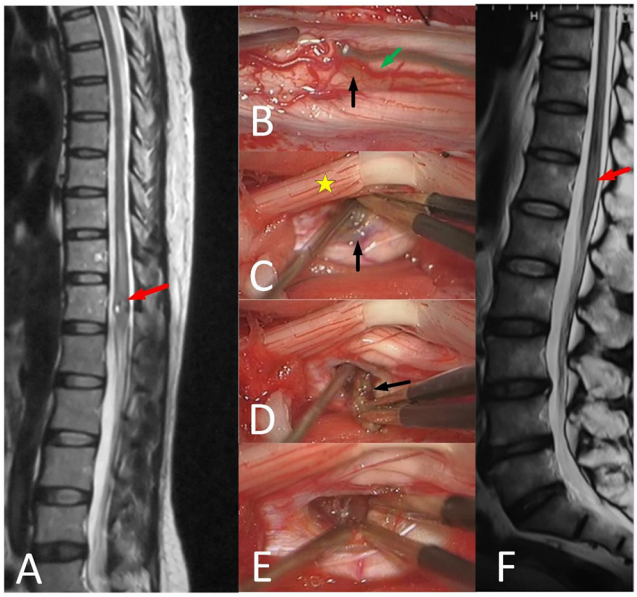

FIGURE 4 | 51-year-old female with unsteady walking for 6 months and urinating difficulty for 3 days (mALs = 11 points, severe disorder, type B1), was treated by ERS. **(A)** Preoperative spinal MRI examination revealed ISCH in ventral conus (red arrow). **(B–E)** A The brown lesion on the spinal cord surface (black arrow) could be seen after pushing posterior nerve roots (yellow star) aside and the artery (green arrow) was protected carefully. **(F)** MRI reexamination showed no residual or recurrence at 3 years after surgery and the mAS was 10 points at the last follow-up.

There were also 10 patients with deterioration who not receive ERS, seven patients were delayed at the referral process, and the other three patients were indecisive about ERS and treated by microsurgery at a routine time. As we think, the treatment system of ERS needs constantly perfecting, such as letting more primary hospitals and patients with symptoms understand the clinical characteristics and treatment of ISCH.

Many worsening predictors after resection have been reported, such as poor preoperative function, thoracolumbar-level lesions, and the depth of lesions (28, 29). In our study, lesions in the ventral surface, at a lumbosacral segment, and a longer course of disease were potential predictors for poor functional recovery. For intramedullary ventrolateral deep lesion, Ren adopted a new surgical approach, the dorsal root entry zone myelotomy (DREZ), and showed that of 10 patients, two (20%) patients improved and eight (80%) patients were stable after the new approach (30). Ginalis reported a multi-segment, hemorrhagic intramedullary cavernous malformation from C7 to T3 was resected through a lateral myelotomy approach at the site of superficial hemorrhage (31). As we believe, the reason for DREZ or a lateral myelotomy approach being chosen, is because the corridor is the closest way into the lesion. Westphal reported 500 cases of intramedullary lesions, including ependymomas, astrocytoma, vascular pathologies, indicating that safe and complete removal can be achieved by posterior midline approach (32). The posterior midline approach for deeper lesions and the direct approach for superficial lesions are our two conventional approaches: 1. It needs to be emphasized that 1 blood vessels on the surface of the spinal cord should be carefully protected during the operation (**Figure 4**); 2. Avoid pulling and twisting the spinal cord; and 3. Try not to use bipolar coagulation, and if necessary, keep its energy to the minimum.

In this study, electrophysiologic monitoring was performed in all patients, including motor-evoked potentials (MEPs) and somatosensory-evoked potentials (SEPs). As our plan, if the amplitude is <50%, the operation should be suspended (33). Compared with Li's result, of the 52 patients with ISCH under electrophysiologic monitoring, 17 patients showed permanent changes, two had long-term residual neurologic deficits (34). In our cohort, ten patients showed transient amplitude decline, and the other two patients showed permanent changes according to electrophysiologic monitoring. During 1 week after surgery, eight patients (15.4%, 8/52) showed a decrease in function, and functional impairments included hypoesthesia in six patients, sphincter dysfunction in two patients, and decreased muscle strength in two patients. Most of the functional impairments were transient, six had recovered to the preoperative state within two months after surgery, and two with lesions at the lumbosacral segment. At 12 months after surgery, only two patients got worse at sphincter dysfunction or decreased muscle strength than the status before surgery, whose lesions were all located at the lumbosacral segment, which meant that recovery is more difficult in those patients with lesions lumbosacral segment.

CONCLUSIONS

Microsurgical resection is safe and effective for symptomatic ISCH; however, lumbosacral lesions had a poor prognosis. The patients with a deteriorative type would receive a better prognosis at long-term follow-up if treated by ERS.

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 Huashan Hospital and Huadong Hospital, affiliated to Fudan University. Written informed consent was obtained from the individual(s) or their legal guardian/next of kin to participate in this study and for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

GC proposed ERS theory and performed the operations. YD wrote the article and analyzed the data. RM, XQ, and YL assisted to finish part of the operations and collected the data. JL was responsible for intraoperative neurophysiological monitoring. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by Project of Shanghai Science and Technology Committee (18411962400).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.872824/full#supplementary-material

REFERENCES

1. Gross BA, Du R, Popp AJ, Day AL. Intramedullary spinal cord cavernous malformations. *Neurosurg Focus.* (2010) 29:E14. doi: 10.3171/2010.6.FOCUS10144
2. McCormick PC, Michelsen WJ, Post KD, Carmel PW, Stein BM. Cavernous malformations of the spinal cord. *Neurosurgery.* (1988) 23:459–63. doi: 10.1097/00006123-198810000-00009
3. Sandalcioglu IE, Wiedemayer H, Gasser T, Asgari S, Engelhorn TSD. Intramedullary spinal cord cavernous malformations: clinical features and risk of hemorrhage. *Neurosurg Rev.* (2003) 26:253–6. doi: 10.1007/s10143-003-0260-2
4. Brotchi J, Noterman J, Baleriaux D. Surgery of intramedullary spinal cord tumours. *Acta Neurochir.* (1992) 116:176–8. doi: 10.1007/BF01540873
5. Anson JA, Spetzler RF. Surgical resection of intramedullary spinal cord cavernous malformations. *J Neurosurg.* (1993) 78:446–51. doi: 10.3171/jns.1993.78.3.0446
6. Zakirov BA, Konovalov NA, Belousova OB, Kaprovoy SV. Surgical treatment of spinal cord cavernous malformations. *Zh Vopr Neirokhir Im N N Burdenko.* (2021) 85:104–13. doi: 10.17116/neiro202185031104
7. Fotakopoulos G, Kivelev J, Andrade-Barazarte H, Tjahjadi M, Goehre F, Hernesniemi J. Outcome in patients with spinal cavernomas presenting with symptoms due to mass effect and/or hemorrhage: conservative versus surgical management: meta-analysis of direct comparison of approach-related complications. *World Neurosurg.* (2021) 152:6–18. doi: 10.1016/j.wneu.2021.05.094
8. Canavero S, Pagni CA, Duca S, Bradac GB. Spinal intramedullary cavernous angiomas: a literature meta-analysis. *Surg Neurol.* (1994) 41:381–8. doi: 10.1016/0090-3019(94)90031-0
9. Eicker S, Turowski B, Steiger HJ. Hänggi D. [Diagnostic work-up and therapy of spinal vascular malformations: an update]. *Nervenarzt.* (2010) 81:719–26. doi: 10.1007/s00115-010-2971-2
10. Spetzger U, Gilsbach JM, Bertalanffy H. Cavernous angiomas of the spinal cord clinical presentation, surgical strategy, and postoperative results. *Acta Neurochir.* (1995) 134:200– 6. doi: 10.1007/BF01417690
11. Zevgaridis D, Medele RJ, Hamburger C, Steiger HJ, Reulen HJ. Cavernous haemangiomas of the spinal cord. A review of 117 cases. *Acta Neurochir (Wien).* (1999) 141:237–45. doi: 10.1007/s007010050293
12. Choi GH, Kim KN, Lee S, Ji GY, Oh JK, Kim TY, et al. The clinical features and surgical outcomes of patients with intramedullary spinal cord cavernous malformations. *Acta Neurochir (Wien).* (2011) 153:1677–84. doi: 10.1007/s00701-011-1016-3
13. Lu DC, Lawton MT. Clinical presentation and surgical management of intramedullary spinal cord cavernous malformations. *Neurosurg Focus.* (2010) 29:E12. doi: 10.3171/2010.6.FOCUS10139
14. O'Phelan KH. Emergency neurologic life support: spinal cord compression. *Neurocrit Care.* (2017) 27:144–51. doi: 10.1007/s12028-017-0459-7
15. Ogilvy CS, Louis DN, Ojemann RG. Intramedullary cavernous angiomas of the spinal cord: clinical presentation, pathological features, and surgical management. *Neurosurgery.* (1992) 31:219–30. doi: 10.1097/00006123-199208000-00007
16. Mitha AP, Turner JD, Abla AA, Vishteh AG, Spetzler RF. Outcomes following resection of intramedullary spinal cord cavernous malformations: a 25-year experience. *J Neurosurg Spine.* (2011) 14:605–11. doi: 10.3171/2011.1.SPINE10454
17. Badhiwala JH, Farrokhyar F, Alhazzani W, Yarascavitch B, Aref M, Algird A. et al. Surgical outcomes and natural history of intramedullary spinal cord cavernous malformations: a single-center series and meta-analysis of individual patient data: clinic article. *J Neurosurg Spine.* (2014) 21:662–76. doi: 10.3171/2014.6.SPINE13949
18. Mitha AP, Turner JD, Spetzler RF. Surgical approaches to intramedullary cavernous malformations of the spinal cord. *Neurosurgery.* (2011) 68:317–24. doi: 10.1227/NEU.0b013e3182138d6c
19. Zhang L, Yang W, Jia W, Kong D, Yang J, Wang G. et al. Comparison of outcome between surgical and conservative management of symptomatic spinal cord cavernous malformations. *Neurosurgery.* (2016) 78:552–61. doi: 10.1227/NEU.0000000000001075
20. Azad TD, Veeravagu A, Li A, Zhang M, Madhugiri V, Steinberg GK. Long-term effectiveness of gross-total resection for symptomatic spinal cord cavernous malformations. *Neurosurgery.* (2018) 83:1201–8. doi: 10.1093/neuros/nyx610
21. Ungeheuer D, Stachura K, Moskala M. Intramedullary spinal cord cavernous malformations-clinical presentation and optimal management. *Przegl Lek.* (2015) 72:662–4.
22. Panda A, Diehn FE, Kim DK, Bydon M, Goyal A, Benson JC. et al. Spinal cord cavernous malformations: MRI commonly shows adjacent intramedullary hemorrhage. *J Neuroimaging.* (2020) 30:690–6. doi: 10.1111/jon.12738
23. Tong X, Deng X, Li H, Fu Z, Xu Y. Clinical presentation and surgical outcome of intramedullary spinal cord cavernous malformations. *J Neurosurg Spine.* (2012) 16:308–14. doi: 10.3171/2011.11.SPINE11536
24. Steiger HJ, Turowski B, Hanggi D. Prognostic factors for the outcome of surgical and conservative treatment of symptomatic spinal cord cavernous malformations: a review of a series of 20 patients. *Neurosurg Focus.* (2010) 29:E13. doi: 10.3171/2010.6.FOCUS10123
25. Ohnishi YI, Nakajima N, Takenaka T, Fujiwara S, Miura S, Terada E. et al. Conservative and surgical management of spinal cord cavernous malformations. *World Neurosurg X.* (2020) 5:100066. doi: 10.1016/j.wnsx.2019.100066
26. Nagoshi N, Tsuji O, Nakashima D, Takeuchi A, Kameyama K, Okada E. et al. Clinical outcomes and prognostic factors for cavernous hemangiomas of the spinal cord: a retrospective cohort study. *J Neurosurg Spine.* (2019) 31:271–8. doi: 10.3171/2019.1.SPINE18854
27. Zhang L, Qiao G, Yang W, Shang A, Yu X. Clinical features and long-term outcomes of pediatric spinal cord cavernous malformation-a report of 18 cases and literature review. *Childs Nerv Syst.* (2021) 37:235–42. doi: 10.1007/s00381-020-04700-9
28. Ren J, Hong T, He C, Li X, Ma Y, Yu J. et al. Surgical approaches and long-term outcomes of intramedullary spinal cord cavernous malformations: a single-center consecutive series of 219 patients. *J Neurosurg Spine.* (2019) 31:123–32. doi: 10.3171/2018.12.SPINE181263
29. Liang JT, Bao YH, Zhang HQ, Huo LR, Wang ZY, Ling F. Management and prognosis of symptomatic patients with intramedullary spinal cord cavernoma: clinical article. *J Neurosurg Spine.* (2011) 15:447–56. doi: 10.3171/2011.5.SPINE10735
30. Ren J, He C, Hong T, Li X, Ma Y, Yu J, et al. Anterior to Dorsal Root Entry Zone Myelotomy (ADREZotomy): a new surgical approach for the treatment of ventrolateral deep intramedullary spinal cord cavernous malformations. *Spine (Phila Pa 1976).* (2018) 43:E1024–32. doi: 10.1097/BRS.0000000000002607
31. Ginalis EE, Herschman Y, Patel NV, Jumah F, Xiong Z, Hanft SJ. Lateral myelotomy for resection of a ruptured intramedullary cervico-thoracic cavernous malformation. *Oper Neurosurg (Hagerstown).* (2021) 20:E317–21. doi: 10.1093/ons/opaa417

32. Westphal M, Mende KC, Eicker SO. Refining the treatment of spinal cord lesions: experience from 500 cases. *Neurosurg Focus.* (2021) 50:E22. doi: 10.3171/2021.2.FOCUS201107
33. Sala F, Bricolo A, Faccioli F, Lanteri P, Gerosa M. Surgery for intramedullary spinal cord tumors: the role of intraoperative (neurophysiological) monitoring. *Eur Spine J.* (2007) 16:S130–9. doi: 10.1007/s00586-007-0423-x
34. Li X, Zhang HQ, Ling F, He C, Ren J. Differences in the electrophysiological monitoring results of spinal cord arteriovenous and intramedullary spinal cord cavernous malformations. *World Neurosurg.* (2019) 122:e315–24. doi: 10.1016/j.wneu.2018.10.032

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.

frontiers | Frontiers in Medicine

ORIGINAL RESEARCH
published: 03 June 2022
doi: 10.3389/fmed.2022.893208

Using Artificial Intelligence to Establish Chest X-Ray Image Recognition Model to Assist Crucial Diagnosis in Elder Patients With Dyspnea

Liu Liong-Rung[1,2,3], Chiu Hung-Wen[3,4], Huang Ming-Yuan[1,2], Huang Shu-Tien[1,2,3†], Tsai Ming-Feng[2,3,5†], Chang Chia-Yu[3†] and Chang Kuo-Song[1]*

[1] Department of Emergency Medicine, Mackay Memorial Hospital, Taipei, Taiwan, [2] Department of Medicine, Mackay Medical College, New Taipei City, Taiwan, [3] Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan, [4] Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei, Taiwan, [5] Division of Plastic Surgery, Department of Surgery, Mackay Memorial Hospital, Taipei, Taiwan

OPEN ACCESS

Edited by:
Abeer Alsadoon, Charles Sturt University, Australia

Reviewed by:
Muhammad Fazal Ijaz, Sejong University, South Korea
Hosna Salmani, Iran University of Medical Sciences, Iran
Woon-Man Kung, Chinese Culture University, Taiwan

***Correspondence:**
Chiu Hung-Wen
hwchiu@tmu.edu.tw

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Family Medicine and Primary Care, a section of the journal Frontiers in Medicine

Received: *10 March 2022*
Accepted: *09 May 2022*
Published: *03 June 2022*

Citation:
Liong-Rung L, Hung-Wen C, Ming-Yuan H, Shu-Tien H, Ming-Feng T, Chia-Yu C and Kuo-Song C (2022) Using Artificial Intelligence to Establish Chest X-Ray Image Recognition Model to Assist Crucial Diagnosis in Elder Patients With Dyspnea. Front. Med. 9:893208. doi: 10.3389/fmed.2022.893208

Pneumonia and pulmonary edema are the most common causes of acute respiratory failure in emergency and intensive care. Airway maintenance and heart function preservation are two foundations for resuscitation. Laboratory examinations have been utilized for clinicians to early differentiate pneumonia and pulmonary edema; however, none can provide results as prompt as radiology examinations, such as portable chest X-ray (CXR), which can quickly deliver results without mobilizing patients. However, similar features between pneumonia and pulmonary edema are found in CXR. It remains challenging for Emergency Department (ED) physicians to make immediate decisions as radiologists cannot be on-site all the time and provide support. Thus, Accurate interpretation of images remains challenging in the emergency setting. References have shown that deep convolutional neural networks (CNN) have a high sensitivity in CXR readings. In this retrospective study, we collected the CXR images of patients over 65 hospitalized with pneumonia or pulmonary edema diagnosis between 2016 and 2020. After using the ICD-10 codes to select qualified patient records and removing the duplicated ones, we used keywords to label the image reports found in the electronic medical record (EMR) system. After that, we categorized their CXR images into five categories: positive correlation, negative correlation, no correlation, low correlation, and high correlation. Subcategorization was also performed to better differentiate characteristics. We applied six experiments includes the crop interference and non-interference categories by GoogLeNet and applied three times of validations. In our best model, the F1 scores for pneumonia and pulmonary edema are 0.835 and 0.829, respectively; accuracy rate: 83.2%, Recall rate: 83.2%, positive predictive value: 83.3%, and F1 Score: 0.832. After the validation, the best accuracy rate of our model can reach up to 73%. The model has a high negative predictive value of excluding pulmonary edema, meaning the CXR shows no sign of pulmonary edema. At the time, there was

a high positive predictive value in pneumonia. In that way, we could use it as a clinical decision support (CDS) system to rule out pulmonary edema and rule in pneumonia contributing to the critical care of the elderly.

Keywords: computer-aided detection (CAD), artificial intelligence, geriatrics medicine, critical care medicine, chest X-ray (CXR)

INTRODUCTION

Chest X-ray (CXR) is one of the most commonly used clinical imaging examinations in the medical field due to its adequate image resolution and standardized sampling techniques (1). Before admission to an outpatient clinic or emergency department, patients usually undergo at least one routine CXR, which is rapid and has high diagnostic value for patients displaying symptoms of dyspnea (2). The appearance of pneumonia (PN) on CXR films is inconsistent, and some lung field characteristics, such as infiltration, are similar to pulmonary edema (PE), which is also one of the most severe respiratory diseases. These features were difficult to obtain features with mathematical definitions and traditional image processing methods on CXR. Previous studies suggested CXR performed usually could not be timely interpreted by radiologists to generate proved reports to assist clinicians to make proper diagnosis (3, 4). Even in medical centers of Taiwan, CXR image report generated by radiologist is not as timely as clinical required. Thus, the correct early-stage interpretation of received images is a substantial clinical challenge in emergency and intensive care units.

Although pneumonia and pulmonary edema share some similar characteristics on X-ray films, the main problem in pneumonia is the inflammation of lung parenchyma or interstitium, whereas that in edema is the abnormal accumulation of fluid in the extravascular space of the lung; thus, the pathophysiology and treatment of these diseases are completely different. Pneumonia treatment involves controlling lung infection and relieving inflammation, whereas edema treatment prioritizes the elimination of pulmonary fluid. Appropriate treatment after diagnosis can reduce the duration of hospitalization and may save lives by avoiding respiratory failure; thus, accurately distinguishing these diseases is key for improving patient outcomes (5). In particular, for patients in extreme age groups, namely children and older adults aged 65 years or above, early diagnosis is significantly correlated with mortality rate (6).

AI approach from machine learning to deep learning contributes to comprehensive healthcare in many ways, such as: symptoms detection, disease classification. Not only has the opportunity to improve the diagnosis and helping decision-making, but also has the potential reduce the cost of medical care (7). Deep learning can be used to identify and derive meaning from image features. Its performance in image recognition tasks has been confirmed in previous studies; deep learning has performance superior to conventional machine learning in the medical filed (8), and can be used in computer-aided detection (CAD) (9). Recently, deep learning has been applied for clinical decision-making assistance for the diagnosis of various diseases, because of it is efficient to deal with unstructured and ambiguous data (10), including diabetic retinopathy, macular edema (11, 12), skin cancer (13), and breast cancer (14). CXR is one of the most commonly used examinations in hospitals, and numerous CXR images can be easily obtained. However, laboratory findings are always more trustworthy than diagnosis based on image features alone, which often challenge early diagnosis. Deep learning models would helped in recognize complex patterns precisely (15). Many papers have used deep learning to help identified chest lesions such as pneumonia, pneumothorax, etc. (16, 17), Furthermore, the specific pattern of pneumonia caused by Covid-19 could also be recognized by deep learning method (18). Deep convolutional neural networks (CNNs) have exceptional performance in image classification. In 2012, CNNs demonstrated excellent image recognition performance in the ImageNet Large-Scale Visual Recognition Competition (ILSVRC) classification task challenge (19). CNNs have a multilayer neural network structure with strong fault tolerance, self-learning, and parallel processing capabilities. In CNN learning, suitable features can be selected as inputs without additional manual processing, the features can be automatically analyzed from the original image data, and feature classification can be learned. CNNs use convolutional layers to extract features and use pooling (max or average) layers to generalize features. The set of the various filters they used for Convolutional Layers extract different sets of features. The biggest advantage of Deep Learning is that we do not need to manually extract features from the image. The network learns to extract features while training. Thus, CNN learning considerably reduces manual preprocessing, facilitating the learning and classification of optimal visual features. Compared with the general feedforward network, the local connection method of the CNN greatly reduces the network parameters. Many CNNs have been developed, such as AlexNet (20), GoogLeNet (21), ResNet (22), and VGGNet (23).

Numerous studies have verified that CNNs for lung disease identification can produce diagnosis results with accuracy meeting that of radiologists, such as ChestX-ray14, which is used public datasets of National Institutes of Health (24), and the CheXNeXt, which is based on the DenseNet (25). However, research for critical cases or cases in older adults were not mentioned in previous studies. The standard CXR uses the posterior-anterior view (PA view) and is performed with the patient standing. The PA view is optimal for image interpretation and for analysis of the mediastinal space and lungs and can be used for accurate heart size assessment (26). For patients with severe illness who are bedridden or unable to stand, the anterior-posterior view (AP view) or portable

CXR are alternate methods. Because the heart is located further away from the film, the AP view may cause the ratio of the heart to the mediastinal space to be enlarged by 15–20%, affecting the clinician's judgment of the sizes of the heart, blood vessels, and lymph vessels in the anterior mediastinal space. Moreover, factors such as enlarged mediastinal space, elevated diaphragm, skin folds, and incomplete opening of the scapula in the AP view can affect the physician's interpretation and increase the possibility of errors (27). Furthermore, patients with critical illnesses are often have life support instruments such as endotracheal tube or vitals monitoring equipments attached to their body likes electrocardiograph wires. Given CNNs' high potential for identifying tiny particles or objects in images (28), studies have not individually discussed the interference caused by these instruments or have even excluded this group of patients. Although these images are the most challenging for machine learning, the capacity for interpreting them in clinical practice is urgently needed.

In this retrospective study, we discussed approaches of distinguishing between pneumonia and pulmonary edema on radiograph. GoogleNet transfer learning was used to analyze the performance of machine learning in distinguishing between PN and PE in the chest radiograph of patients aged 65 years or older who were admitted to the emergency department in Mackay Memorial Hospital. Moreover, we explored the effects of instrument interference, image cropping, and text labels on the capacity of machine learning to classify images. The objective of this study was to establish a CAD model for early diagnosis aimed at patients with critical illnesses.

The main contributions of this study are as follows:

- We provide CAD tools for critically ill elderly who urgently need assistance in image interpretation.
- We demonstrate the interferences such as life-supporting catheters? instruments affect the machine learning outcomes.
- The performance of machine learning in chest X-ray is consistent with the radiologists, when the EMR have more clear features such as pneumonia and edema, the better results are trained on these images.

MATERIALS AND METHODS

This study was approved by the Institutional Review Board of Mackay Memorial Hospital. The *International Classification of Disease, Tenth Revision* (ICD-10) hospital discharge codes collected in one medical center in Taiwan (Mackay Memorial Hospital) since 2015 to 2020 for patients aged 65 years and older who were admitted to the hospital through the emergency department. The number of CXR images from patients with PN (ICD-10: J18) and PE (ICD-10: J81) were 45,781 and 43,674, respectively. Moreover, the electronic medical records (EMR) compiled by radiologists were labeled using keywords and subsequently analyzed by two emergency physicians with more than 15 years of experience. A plastic surgeon assisted with image classification and training. The experiment was divided into six steps comprising tasks including preprocessing, text labeling, and machine learning (**Figure 1**).

Data Acquisition

The CXR images were downloaded through the picture archiving and communication system (PACS), after which deidentification and annotation removal were performed. Moreover, 800 CXR images of patients without lung disease at admission were collected and similarly subjected to deidentification and annotation removal for joint training of the proposed CAD model with the CXR images. The training image format was JPG, and the image conversion size was $224 \times 224 \times 3$ pixels.

Quality Assurance

To exclude repeated cases and ensure the quality of machine learning, a pretraining process using image numbers and text labels was performed before training the CNN (**Figure 2**).

First, during deidentification, we discovered that data were repeated in the PN and PE sets; 16,762 images were present in the two disease lists, indicating that both J18 and J81 were included in the ICD-10 codes of these cases. These duplicated images were not errors in case collection; instead, they represented older adults with multiple comorbidities. For example, many patients with severe pulmonary edema (e.g., acute decompensated heart failure) were complicated by pneumonia due to respiratory

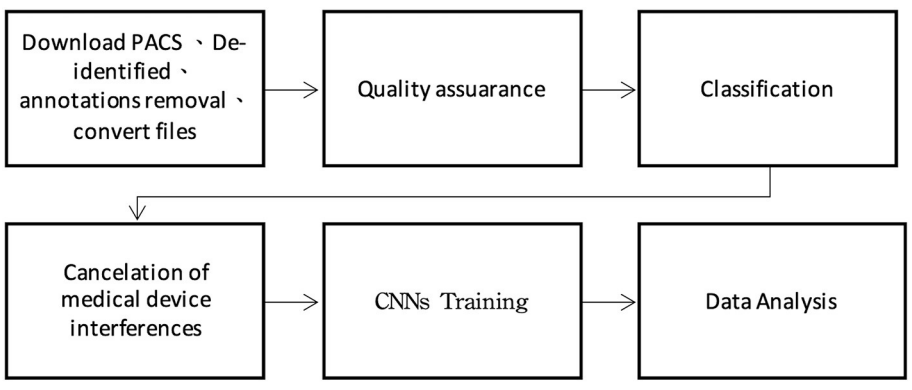

FIGURE 1 | Data processing procedure.

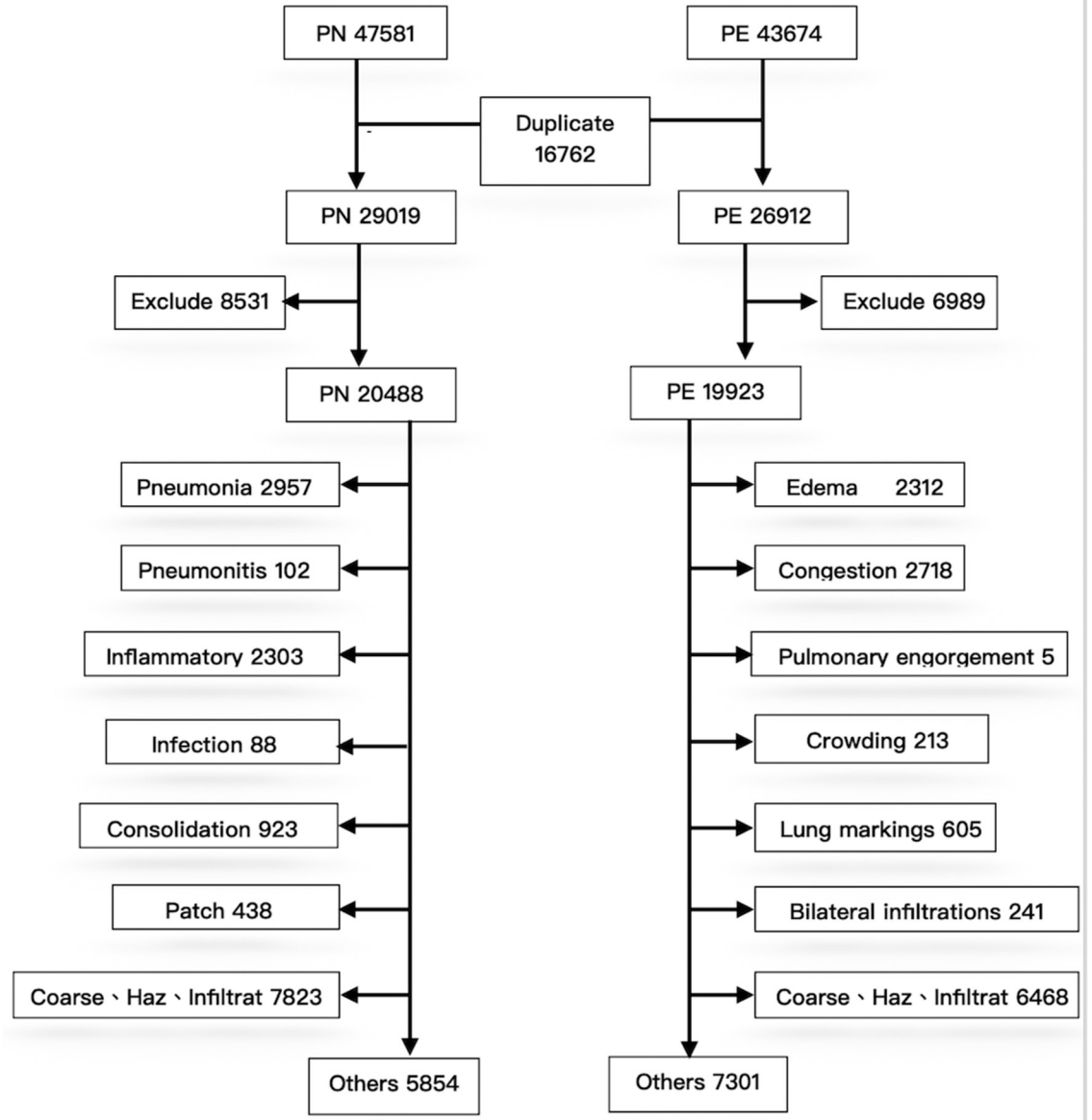

FIGURE 2 | Classification process for image data. PN, pneumonia; PE, pulmonary edema.

tract infection after hospitalization. Conversely, patients with pneumonia may also develop multiple organ failure after hospitalization, leading to edema (e.g., heart or renal failure). Therefore, duplicated images of the two diseases are expected and reasonable in the collection of clinical cases. Accordingly, the 16,272 repeated cases were excluded; otherwise, they could not be classified during CNN training. After exclusion of the repeated cases, PN and PE each had 29,019 and 26,912 images. Because the data were obtained directly from the PACS system, some erroneous data might be included. After reconducting a query of reports using keywords to exclude irrelevant cases, the PN and PE data sets had 20,488 and 19,923 images, respectively.

Experimental Design for Image Classification

The effects of CNN on the interpretation results under different conditions were investigated with six experiments as follows.

Experiment 1: First, we tested the CNN's capability to identify diseases and its capacity to distinguish between PN and PE with correct ICD-10 diagnoses. A total of 2,000 files were randomly sampled from the 20,488 PN and 19,923 PE images and were combined with the 800 images of patients without lung disease for transfer learning. The training model was named G_random.

Experiment 2: Because the images collected in this study were those of patients with critical illnesses, more than half of the images contained extracorporeal life support instruments or tubes. To determine the degree of interference of this equipment on machine learning, the 2,000 PN and PE images were further divided into images with and without interference; these sets were independently used to train the machine learning model. For images with interference, we randomly sampled 1,000 files from the two disease data sets for training. This training model was named G_int.

Experiment 3: PN and PE were confirmed to contain only 650 and 480 images without interference, respectively. We named this training model G_NCC and determined whether superior results were obtained for the images without interference.

Experiment 4: To improve the training model, the images with interference were processed using image cropping. In **Figure 3**, the oxygen supply mask (indicated by a white arrow in the image) was cropped to produce a clearer lung field. Finally, we processed 1,100 PN and 670 PE images; this training model was named G_clean.

Experiment 5: To further determine whether EMR labels precisely would produce superior results in training, 2,000 each image which obviously clamed pneumonia and edema were collected and separated into two categories; this model was named G_DC2.

Experiment 6: The data from Experiment 5 were combined with 800 normal CXR images for joint training; this model was called G_DC. Experiments 5 and 6 were performed to compare whether machine learning was affected by including the comparatively easily identifiable normal CXR images.

Model Training

The built-in neural network toolbox of MATLAB R2020b (The MathWorks, Natick, MA, USA) on Windows 10 (Microsoft, Redmont, WA, USA) was used for the experiments. The computer had a GeForce RTX 2060 (Nvidia, Santa Clara, CA, USA) graphics processing unit, and the training image format was 24-bit JPG.

The transfer learning used the GoogLeNet Inception V4 architecture. GoogLeNet is a type of convolutional neural network based on the Inception architecture (29). It utilizes Inception modules, which allow the network to choose between multiple convolutional filter sizes in each block. An Inception network stacks these modules on top of each other, with occasional max-pooling layers with stride 2 to halve the resolution of the grid. The GoogLeNet we used in this study 22 layers deep and have an image input size of 224-by-224.The data were trained through multilayer calculations, and the composition of each layer was automatically learned from the data set. A key feature was the Inception module, which was regarded as a milestone in the history of CNN development in a previous study (30). Because this module replaces the fully connected structure with sparse connections for the input images, performs multiple convolution operations or pooling operations, and splices all of the output results into an extremely deep feature map, the module reduces the computational burden of including numerous parameters as well as the problem of overfitting. The performance of the current iteration of Inception, Inception-v4, was verified in the 2015 ILSVRC challenge; it has superior image recognition capabilities due to its use of residual Inception networks.The training environment settings were as follows: minimal batch size = 20, maximum epochs = 50, pixel range= [−3, 3], Rotation Range = [−15, 15], and training/validation ratio = 70:30.

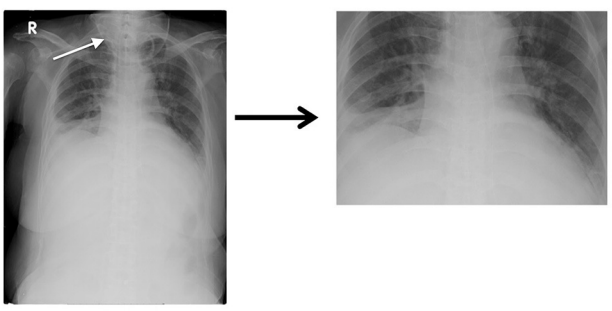

FIGURE 3 | Comparison of images before and after manual cropping.

Model Performance Evaluation

The built-in neural network toolbox in MATLAB R2020b was used to draw the receiving operating characteristic (ROC) curve and produce a confusion matrix. The recall, precision, F1 score, and accuracy of each model were then calculated. Recall, precision, and F1 Score are frequently used for analyzing model performance. A high F1 score indicates higher precision and recall for disease decision-making, and the results of the aforementioned transfer learning models were analyzed using these indicators.

RESULTS

Table 1 presents the model performance evaluation results for all six experiments. The G_DC model that used images clearly identified as having PN or PE had the highest accuracy and F1 score. The F1 score, and accuracy of the G_DC model (F1 score = 0.882, validation accuracy = 86.4%) were significantly superior to those of the G_random model that was trained using only ICD-10 codes (F1 score = 0.82; validation accuracy = 79.1%).

In addition, the G_int model that was trained solely using images with interference had the worst results with an accuracy of only 73%. Both the G_NCC model, which trained on images without interference from the beginning, and the G_clean model, which trained on cropped images, did not have significant improvements in their validation accuracy or F1 scores. In addition, the G_clean model had a significantly increase for recall of PN from 78.3 to 90.5%; however, its PE recall declined from 79.2 to 56.4%. No significant change was observed in the precision for the two diseases (PN: 76.7 to 77.1%; PE: 77.4 to 78.6%).

TABLE 1 | Accuracy, recall, precision, and F1 score results for the six experiments.

	Recall	Precision	Accuracy	F1 score
G_random	81.3%	82.8%	79.1%	0.82
G_int	73.2%	73.4%	73.4%	0.733
G_clean	73.5%	77.9%	77.5%	0.756
G_NCC	74%	73.6%	74.1%	0.738
G_DC	87.7%	88.7%	86.4%	0.882
G_DC2	83.2%	83.3%	83.2%	0.832

**G_random: Randomly selected from the PN and PE category and combined with normal CXR.*
G_int: Images with interferences from G_random.
G_clean: Images with interferences form G_random cropped manually.
G_NCC: Images without interferences from G_random.
G_DC: G_DC2 combined with 800 normal CXR images.
G_DC2: Images labeled pneumonia and edema.

The results of the G_random and G_DC models, the training of which incorporated normal CXR images, revealed that normal CXR resulted in an optimal area under the ROC (AUC; **Figures 4**, **5**), indicating that normal CXR images are easier to identify.

Based on the aforementioned results, we believed that, rather than medical interferences, images used for training with more precise description from EMR were the decisive reason that affected machine learning performance; such images proved to be the main factor for improving machine learning performance.

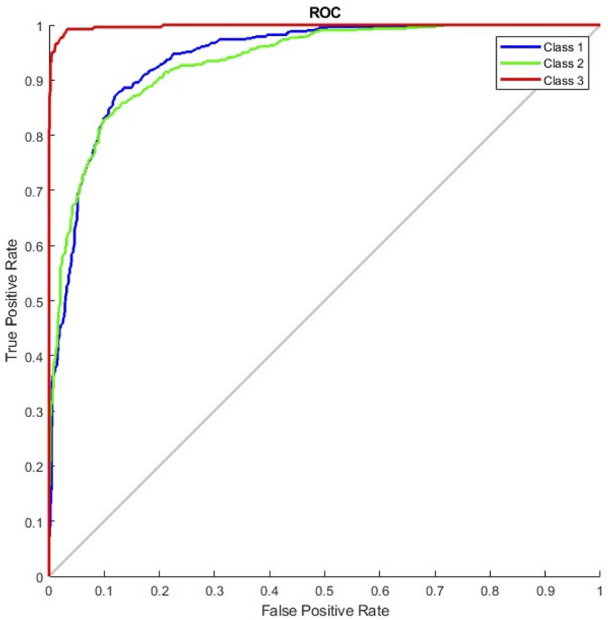

FIGURE 4 | Receiver operating characteristic curve of the G_random model (class 1, pulmonary edema; class 2, pneumonia; class 3, normal).

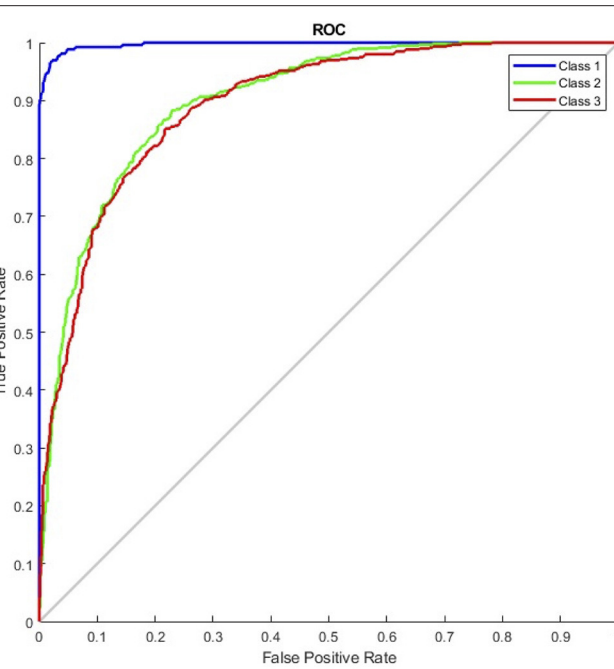

FIGURE 5 | Receiver operating characteristic curve of the G_DC model (class 1, normal; class 2, pulmonary edema; class 3, pneumonia).

DISCUSSION

Effects of Incorporating Normal CXR

In the experiments, normal CXR images of patients without PN or PE had an F1 score over 95%. Machine learning had superior performance among normal CXR than radiographs with lesions. Similar results also demonstrated by Cicero et al. (21), the positive predictive value of the normal category reached 90%, whereas those for the consolidation and for edema were merely 23 and 43%, respectively. Thus, training with normal CXR images could raise overall model accuracy by increasing both the true positive and true negative values.

Comorbidities

For cases that might be diagnosed with PN and PE simultaneously, we used two steps for preprocessing: (1) excluding 16,762 repeated files on ICD-10 diagnosis; and (2) excluding images based on the imaging reports. In machine learning, feature selection is considered a critical step in data preprocessing. When we directly use raw data such as ICD code for classification, we sometimes observe that learning algorithms perform poorly (31). These images were excluded because our experiments did not aim to identify comorbidities and the presence of two or more diseases in one radiograph would reduce the machine learning performance (32).

Interferences

There is a significant difference influence between machine learning and physician interpretation for medical devices and life support equipment. Those *in vitro* instruments do not affect physicians reading images, whereas machine learning can detect even subtle features that would not normally be detected

(33), affecting the learning outcome. In our study, the G_int model that all images with medical equipment for training had a significant decrease in its predictive performance; its accuracy was reduced from 79.1 to 73.4%, and its F1 score was reduced from 0.82 to 0.733 (**Table 1**). We performed image processing using cropping but did not obtain a more favorable result. As we known, machine learning is more efficient in distinction of localized lesions rather than lesions with global symmetrical patterns (25). Therefore, pneumonia which sometimes shows unilateral consolidation is easier to be identified than pulmonary edema which is bilateral symmetrical pictures.

The cropped images differed substantially from the original images; this may explain why the training performance was worse than as expected. Moreover, the ratio of the lung field to the lesions might have changed after cropping, causing local consolidation, which were originally easily identified by the models, to exhibit features that more closely resembled diffusion. In addition, for patients in critical condition, PE images almost always contained one or more medical instruments or life support tubes, leading to the exclusion of many images that could not be fully cropped in training. Only 1,100 and 670 PN and PE training images, respectively, were retained after cropping. Moreover, we discovered that the recall of PN was significantly higher than that of PE (90.5 vs. 56.4%). The number of images in training sets must be balanced because an unbalanced number of training images causes learning to be biased toward image types that the model had more exposure to (34). Thus, insufficient datasets and unbalanced training sets might also have affected the performance.

Model Comparison

There were many previous studies used CNN as a chest X-ray CAD tools. Some models were published based on public institutions datasets such as ChestX-ray14 which built by The National Institutes of Health (35). Cicero et al. used GoogLeNet in 2017 to construct a model that resolves a total of about 35,000 images. It includes normal chest plain films and other five features: Pleural effusion, Cardiomegaly, Consolidation, Pulmonary edema and Pneumothorax. It is found that normal chest plain films had the best recognition, which both sensitivity and specificity can reach above 91%. CheXNeXt used ChestX-ray14 datasets compared with radiologists for identification of 14 chest X-ray features in 2018. Results showed that CheXNeXt performed as well as radiologists on 10 features (no statistically significant difference in AUC) and it was superior than expert on atelectasis. Not as good as radiologists on three characteristics (cardiomegaly, emphysema, emphysema). We compared the performance of pneumonia and pulmonary edema in G_DC and G_DC2 with above literature models. In our experiments both pneumonia and pulmonary edema have higher sensitivity, PPV and F1 score (**Table 2**).

TABLE 2 | Experiments 5 and Experiments 6 compares with previous study.

	Cicero et al. (21)	CheXNeXt	G_DC	G_DC2
PE sensitivity	0.82	0.682	0.868	0.834
PE PPV	0.43	0.662	0.83	0.825
PE F1 score	0.564	0.672	0.849	0.829
PN sensitivity	0.74	0.650	0.832	0.83
PN PPV	0.23	0.377	0.852	0.84
PN F1 score	0.351	0.477	0.842	0.835

Limitations

It has been demonstrated that medical history and laboratory tests would improve radiologist interpretations (36). In this study, we did not combine patients' history and clinical data together for thorough analysis which might provide important part in clinical CAD tool. In addition, due to the limitations of deep learning, our tools currently cannot articulate the eigenvalues by which to classify images. Data preprocessing and text labeling both revealed that PN and PE are related to many diseases and share mutual comorbidities. To maintain a simple training environment during data processing, cases with shared comorbidities were excluded, and no further analysis was conducted on the interpretation of comorbidities. The data in our study were collected from a single medical center, which might affect the objectivity of the text labels. Finally, we did not test the models against the interpretation of the radiologists; thus, we were unable to compare the similarities and differences between the interpretation of the models and specialists.

CONCLUSION

This study revealed that using deep learning to construct X-ray images and to distinguish between PE and PN, and using images with explicit signs of PE or PN and without interference for training, can produce an accuracy of over 80%. Moreover, an accuracy of 70% or higher was achieved even in the presence of interference. In addition, the recognition rate of normal images exceeded 90%; thus, this model can be potentially applied in clinical practice.

Currently, more than two-thirds of the world's population do not have access to professional interpretation of medical images, are unable to receive timely diagnosis reports, or cannot receive any diagnosis. During emergencies or the presence of large number of patients in medical centers (e.g., COVID-19 outbreak clusters), experienced radiologists are subject to human limitations, such as off duty hours, fatigue, and perceptual and cognitive biases; these limitations may lead to misjudgment. Although our model cannot completely replace clinicians. After testing, our model showed excellent performance on identifying pulmonary edema and also informative assistance on patients with pneumonia in elder patients after testing. It provides crucial image information in a timely manner to assist in clinical diagnosis.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Mackay Memorial Hospital Institutional Review Board (IRB). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

LL-R conducted experiments and wrote the manuscript. HM-Y built chest X-ray datasets. HS-T, TM-F, and CC-Y helped selecting patients and organized electric medical records and data assembling. CK-S provides administrative support. CH-W developed the methods, contributed to manuscript editing, and reified experiment instructions. All authors contributed to the article and approved the submitted version.

REFERENCES

1. Raoof S, Feigin D, Sung A, Raoof S, Irugulpati L, Rosenow EC 3rd. Interpretation of plain chest roentgenogram. *Chest.* (2012) 141:545–58. doi: 10.1378/chest.10-1302
2. Cherian T, Mulholland EK, Carlin JB, Ostensen H, Amin R, de Campo M, et al. Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies. *Bull World Health Organ.* (2005) 83:353–9.
3. Kesselman A, Soroosh G, Mollura DJ. RAD-AID conference on international radiology for developing countries: the evolving global radiology landscape. *J Am Coll Radiol.* (2016) (2015) 13:1139–44. doi: 10.1016/j.jacr.2016.03.028
4. Mollura DJ, Azene EM, Starikovsky A, Thelwell A, Iosifescu S, Kimble C, et al. White paper report of the RAD-AID conference on international radiology for developing countries: identifying challenges, opportunities, and strategies for imaging services in the developing world. *J Am Coll Radiol.* (2010) 7:495–500. doi: 10.1016/j.jacr.2010.01.018
5. Aydogdu M, Ozyilmaz E, Aksoy H, Gürsel G, Ekim N. Mortality prediction in community-acquired pneumonia requiring mechanical ventilation; values of pneumonia and intensive care unit severity scores. *Tuberk Toraks.* (2010) 58:25–34.
6. Singanayagam A, Singanayagam A, Elder DHJ, Chalmers JD. Is community-acquired pneumonia an independent risk factor for cardiovascular disease? *Eur Respir J.* (2012) 39:187–96. doi: 10.1183/09031936.00049111
7. Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *J Ambient Intell Humaniz Comput.* (2022). doi: 10.1007/s12652-021-03612-z
8. Chen H, Dou Q, Ni D, Cheng J-Z, Qin J, Li S, et al. *Automatic Fetal Ultrasound Standard Plane Detection Using Knowledge Transferred Recurrent Neural Networks BT - Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015.* In: Navab N, Hornegger J, Wells WM, Frangi A, editors. Cham: Springer International Publishing (2015) p. 507–14.
9. Yang S, Yin B, Cao W, Feng C, Fan G, He S. Diagnostic accuracy of deep learning in orthopaedic fractures: a systematic review and meta-analysis. *Clin Radiol.* (2020) 75:713.e17-713.e28. doi: 10.1016/j.crad.2020.05.021
10. Srinivasu PN, Ahmed S, Alhumam A, Kumar AB, Ijaz MF. An AW-HARIS based automated segmentation of human liver using CT images. *Computers, Materials and Continua.* (2021) 69:3303–19.
11. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *J Am Med Assoc.* (2016) 316:2402. doi: 10.1001/jama.2016.17216
12. Voets M, Møllersen K, Bongo LA. Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *PLoS One.* (2019) 14:e0217541. doi: 10.1371/journal.pone.0217541
13. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Corrigendum: dermatologist-level classification of skin cancer with deep neural networks. *Nature.* (2017) 546:686. doi: 10.1038/nature22985
14. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA.* (2017) 318:2199–210. doi: 10.1001/jama.2017.14585
15. Srinivasu PN, SivaSai JG, Ijaz MF, Bhoi AK, Kim W, Kang JJ. Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM. *Sensors.* (2021) 21:2852. doi: 10.3390/s21082852
16. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology.* (2017) 284:574–82. doi: 10.1148/radiol.2017162326
17. Zhou S, Zhang X, Zhang R. Identifying cardiomegaly in chestx-ray8 using transfer learning. *Stud Health Technol Inform.* (2019) 264:482–6. doi: 10.3233/SHTI190268
18. Rahaman MM, Li C, Yao Y, Kulwa F, Rahman MA, Wang Q, et al. Identification of COVID-19 samples from chest X-Ray images using deep learning: a comparison of transfer learning approaches. *J Xray Sci Technol.* (2020) 28:821–39. doi: 10.3233/XST-200715
19. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. *Int J Comput Vis.* (2015) 115:211–52. doi: 10.1007/s11263-015-0816-y
20. Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks. *Neural Inf Process Syst.* (2012) 25:1097–105. doi: 10.1145/3065386
21. Cicero M, Bilbily A, Colak E, Dowdell T, Gray B, Perampaladas K, et al. Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. *Invest Radiol.* (2017) 52:281–7. doi: 10.1097/RLI.0000000000000341
22. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Institute of Electrical and Electronics Engineers (IEEE) (2016). p. 770–8.
23. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *2016 IEEE Spoken Language Technology Workshop (SLT).* Institute of Electrical and Electronics Engineers (IEEE). (2016) 481–8. doi: 10.1109/SLT.2016.7846307
24. Lian J, Liu J, Zhang S, Gao K, Liu X, Zhang D, et al. A structure-aware relation network for thoracic diseases detection and segmentation. *IEEE Trans Med Imaging.* (2021) 40:2042–52. doi: 10.1109/TMI.2021.3070847
25. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* (2018) 15:1–17. doi: 10.1371/journal.pmed.1002686
26. Chotas HG, Ravin CE. Chest radiography: estimated lung volume and projected area obscured by the heart, mediastinum, and diaphragm. *Radiology.* (1994) 193:403–4. doi: 10.1148/radiology.193.2.7972752
27. Rigby D-M, Hacking L. Interpreting the chest radiograph. *Anaesth Intensive Care Med.* (2021) 22:354–8. doi: 10.1016/j.mpaic.2021.04.011
28. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. *PMLR.* (2015) 2015:448–56.

29. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. *In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR).* Institute of Electrical and Electronics Engineers (IEEE) (2015). p. 1–9.
30. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. *AAAI Conf Artif Intell* (2016).
31. Mandal M, Singh PK, Ijaz MF, Shafi J, Sarkar R. A tri-stage wrapper-filter feature selection framework for disease classification. *Sensors.* (2021) 21. doi: 10.3390/s21165571
32. Chakravarty A, Sarkar T, Ghosh N, Sethuraman R, Sheet D. Learning decision ensemble using a graph neural network for comorbidity aware chest radiograph screening. *Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Int Conf.* (2020) 2020:1234–7. doi: 10.1109/EMBC44109.2020.9176693
33. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks [Internet]. *arXiv.* (2014). doi: 10.48550/arXiv.1312.6199
34. Huang C, Li Y, Loy CC, Tang X. *Learning Deep Representation for Imbalanced Classification.* (2016). p. 5375–84.
35. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning [Internet]. *arXiv.* (2017). doi: 10.48550/arXiv.1711.05225
36. Berbaum K, Franken EA, Smith WL. The effect of comparison films upon resident interpretation of pediatric chest radiographs. *Invest Radiol.* (1985) 20:124–8. doi: 10.1097/00004424-198503000-00004

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.

frontiers | Frontiers in Medicine

ORIGINAL RESEARCH
published: 20 June 2022
doi: 10.3389/fmed.2022.928825

Identifying Distinct Risk Thresholds of Glycated Hemoglobin and Systolic Blood Pressure for Rapid Albuminuria Progression in Type 2 Diabetes From NHANES (1999–2018)

Jiahui Xu[1†], Yan Xue[2†], Qingguang Chen[1], Xu Han[1], Mengjie Cai[1], Jing Tian[1], Shenyi Jin[1] and Hao Lu[1]*

[1] Department of Endocrinology, Shuguang Hospital Affiliated to Shanghai University of Traditional Chinese Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai, China, [2] Laboratory of Cellular Immunity, Shuguang Hospital Affiliated to Shanghai University of Traditional Chinese Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai, China

OPEN ACCESS

Edited by:
Md. Mohaimenul Islam, Aesop Technology, Taiwan

Reviewed by:
Noriyuki Kitagawa, Kameoka Municipal Hospital, Japan
Xiangzhu Zhu, Vanderbilt University, United States

***Correspondence:**
Hao Lu
luhao403@163.com

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

Specialty section:
This article was submitted to Family Medicine and Primary Care, a section of the journal Frontiers in Medicine

Received: *26 April 2022*
Accepted: *12 May 2022*
Published: *20 June 2022*

Citation:
Xu J, Xue Y, Chen Q, Han X, Cai M, Tian J, Jin S and Lu H (2022) Identifying Distinct Risk Thresholds of Glycated Hemoglobin and Systolic Blood Pressure for Rapid Albuminuria Progression in Type 2 Diabetes From NHANES (1999–2018). Front. Med. 9:928825. doi: 10.3389/fmed.2022.928825

Background: It is widely recognized that glycated hemoglobin (HbA1c) and systolic blood pressure (SBP) are two key risk factors for albuminuria and renal function impairment in patients with type 2 diabetes mellitus (T2DM). Our study aimed to identify the specific numerical relationship of albumin/creatinine ratio (ACR) with HbA1c and SBP among a large population of adults with T2DM.

Method: A total of 8,626 patients with T2DM were included in the data analysis from the National Health and Nutrition Examination Surveys (NHANES) (1999-2018). The multiple linear regressions were used to examine the associations of ACR with HbA1c and SBP. Generalized additive models with smooth functions were performed to identify the non-linear relations between variables and interactions were also tested.

Results: Significantly threshold effects were observed between ACR and HbA1c or SBP after multivariable adjustment, with the risk threshold values HbA1c = 6.4% and SBP = 127 mmHg, respectively. Once above thresholds were exceeded, the lnACR increased dramatically with higher levels of HbA1c (β = 0.23, 95 CI%:0.14, 0.32, $P < 0.001$) and SBP (β = 0.03, 95 CI%:0.03, 0.04, $P < 0.001$). Subgroup analysis showed high protein diet was related to higher ACR. In addition, a higher risk of ACR progression was observed in central obesity participants with HbA1C ≥ 6.4% or hyperuricemia participants with SBP ≥ 127 mmHg among patients withT2DM.

Conclusion: We identified thresholds of HbA1c and SBP to stratify patients with T2DM through rapid albuminuria progression. These might provide a clinical reference value for preventing and controlling diabetes kidney disease.

Keywords: risk thresholds, glycated hemoglobin, systolic blood pressure, albuminuria, type 2 diabetes, NHANES

INTRODUCTION

Progression of albuminuria in diabetic patients is associated with impaired renal function and indicative of an increased risk of cardiovascular disease (CVD). Studies have demonstrated that in patients with type 2 diabetes mellitus (T2DM), microalbuminuria is considered an early marker for renal function decline, and elevated albuminuria was consistently correlated with the risk of end-stage kidney disease (1, 2). In addition, as an indicator of the systemic endothelial dysfunction response (3), increased albuminuria also predicts higher risks of myocardial infarction, heart failure, stroke, and cardiac death (4–6). Therefore, it is essential to assess albuminuria in diabetic patients. Since the albumin/creatinine ratio (ACR) is a reliable and sensitive index reflecting early kidney damage as well as relatively stable and convenient, ACR is commonly used to estimate the degree of urinary protein excretion clinically (7).

Although various risk factors could affect the development of albuminuria, abundant studies have confirmed that raised blood pressure and dysglycaemia are two critical risk factors for albuminuria (8–11). Cumulative evidence emphasizes that control of glycated hemoglobin (HbA1c) and systolic blood pressure (SBP) are significant in decreased ACR for both T2DM and Diabetic kidney disease (DKD) patients (12, 13). Previously, a study identified a 5.5% HbA1c level as the risk threshold for albuminuria prevalence in a large Chinese population over the age of 40 (14). Another study found a significantly increased risk of albuminuria in participants with HbA1c $\geq$ 7% compared with the normal urinary protein population. The above results remained stable in diabetic and non-diabetic populations (15). This might suggest a threshold effect between HbA1c and ACR levels, but a lack of large-scale population studies targeting patients with T2DM. In addition, the studies on the risk relationship between SBP and ACR have also been extensively reported. A meta-analysis included 31 cohorts in the world and demonstrated that each 20 mmHg increase in SBP was associated with a 1.5-fold higher prevalence of albuminuria (ACR $\geq$ 30 mg/g) in diabetes (11). It was also reported that only SBP $\leq$ 120 mmHg was associated with the lowest risk of new-onset microalbuminuria (16). However, almost all the above studies use a recommended cut-off point of 30 mg/g for ACR to explore the effects of HbA1c and SBP on the risk of albuminuria. Notably, A cohort study with an up to11-year follow-up period found that protein excretion levels, even with normal at baseline, are pronouncedly associated with increased mortality risk from CVD (17). A recent study also confirmed that a normal ACR range ($\leq$30 mg/g) was related to left ventricular hypertrophy in patients with T2DM (18). This suggested that the specific numerical changes of ACR and the risk thresholds might not be fully reflected when we simply treated ACR as a categorical variable with a 30 mg/g cut-off.

Thus, in this study, we treated ACR as a continuous variable and included a large-scale T2DM population to explore the specific association of ACR with SBP and HbA1c simultaneously.

RESEARCH DESIGN AND METHODS

Study Population

In this cross-sectional study, we merged all the National Health and Nutrition Examination Surveys (NHANES) data from 1999 to 2018. A total of 10,170 diabetes patients were identified according to the definition. We further identified 9,901 patients with T2DM after excluding pregnant woman (n = 47) and possible individuals with type 1 diabetes (n = 369). All the missing data for key variables, including ACR (n = 674), HbA1c (n = 251), and SBP (n = 369), were removed from the dataset. Eventually, 8,626 patients with T2DM were included in the final data analyses. The flow chart of the included study population is shown in **Figure 1**.

Definition of Diabetes

Diabetes was defined if each condition was satisfied in the following items according to the recent American Diabetes Association (ADA) recommendation (19): (1) Previous diagnosis of type 2 diabetes by doctors (2) Fasting blood glucose levels greater than or equal to 7.0 mmol/L (126 mg/dL) (3) Postprandial 2 h plasma glucose levels greater than or equal to 11.1 mmol/L (200 mg/dL) after a standard 75-g oral glucose tolerance test (4) HbA1c levels were 6.5% (48 mmol/mol) or higher (5) The use of insulin or hypoglycemic drugs. Possible type 1 diabetes patients were defined as those aged <20 years who were only treated with insulin (20).

Measurement of Main Variables

The albumin/creatinine ratio was calculated from random urine spot collections and reported as mg/g. Therein, the fluorescent immunoassay was employed to measure human urinary albumin and proved to be a reliable and accurate method. The Jaffé method was used to measure urine creatinine (period 1999–2007), and then the enzymatic method was used (period 2008–2018). HbA1c was tested by high-performance liquid chromatography after collecting venous whole blood specimens in EDTA. Above detection, operations were completed in the laboratory at the University of Minnesota and Columbia. More information on sample collection, transport, and processing was available in the NHANES manual. Blood pressure (BP) was measured by trained survey personnel when participants had rested for at least 5 min in a seated position. BP values included in the final analysis were the average of the three consecutive readings obtained with a standard mercury sphygmomanometer (interrupted or incomplete reading was replaced with fourth BP reading). Pulse pressure (PP) was calculated as systolic minus diastolic pressure.

Definition of Other Variables

Among the demographic parameters, marital status was divided into living with a partner and live without a partner; education level was divided into less than high school, high school, and more than high school. Cigarette smoking status was classified as current smokers (average cigarettes $\geq$ 1/day), past

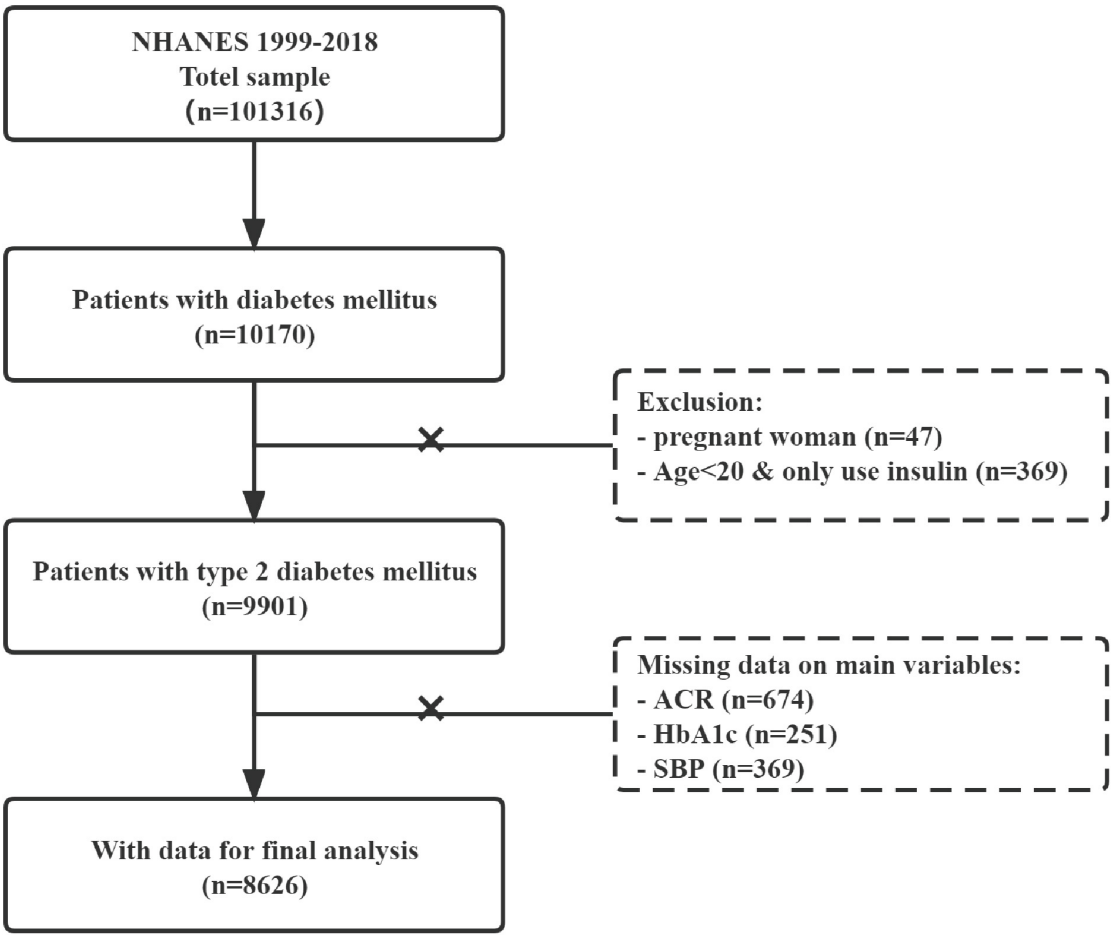

FIGURE 1 | Flow diagram of study population selection.

smokers (average cigarettes <1/day or ≥ 100-lifetime cigarettes but currently non-smoking), and never smokers (<100-lifetime cigarettes or never smoked). The consumption of alcohol was divided into two categories according to whether respondents had at least 12 drinks a year (21). Dietary intake, including dietary protein, sodium intake, and potassium intake, was assessed by two 24 h recalls (one in person and another by telephone 3–10 days later). Meanwhile, the sodium/potassium (Na/K) ratio was calculated for further analysis since the Na/K ratio was proved to have a stronger association with BP than either electrolyte examined alone (22). When obesity indicators were determined as categorical variables, body mass index (BMI, kg/m2) was grouped into normal weight (<25), overweight (≥25, <30), and obese (≥30). A waist circumference ≥102 cm for men and ≥88 cm for women indicates central obesity (23). Diabetes duration was analyzed as a categorical variable with <5 years, ≥5, <10 years, ≥10 years, and not recorded (missing data). The homeostasis model assessment of insulin resistance (HOMA-IR) was calculated with the formula [fasting glucose (mmol/L) × fasting insulin (μU/L)]/22.5. Estimate glomerular filtration rate (eGFR) was calculated based on the chronic kidney disease epidemiology collaboration (CKD-EPI) formula (24).

Statistical Processing and Analyses

To minimize bias brought by missing data, missing categorical covariates were coded as a separate category as appropriate, and missing continuous covariates were replaced by group means. In addition, allowing for the complex sampling design, all analyses were performed incorporating the sampling weights according to NHANES guidelines (25). First, new multi-year sample weights were calculated using ten survey cycles (using 4-year weights when combining the 1999–2000 and 2001–2002 survey cycles). Then the weights of the smallest subpopulation that includes all the variables were selected for final analysis. Finally, to estimate variance, Taylor series linearization was applied, and all estimates were weighted.

In the baseline data assessment, the study population was stratified into four groups according to ACR quartiles. Continuous variables are presented as means ± *SD*s, and categorical variables are reported as frequencies and percentages. ACR was transformed with the natural logarithm function

(LnACR) to stabilize variance prior to analysis. Comparison of continuous variables among groups was analyzed by one-way ANOVA or non-parametric test. The counting variables were analyzed by the chi-square test. Multiple linear regression models were performed to estimate the crude association of ACR with HbA1c and SBP after varying degrees of covariates adjustments. The fully adjusted model included covariates for age, sex, education level, marital status, smoking, alcohol consumption, diabetes duration, BMI, waist circumference, fasting plasma glucose (FPG), diastolic blood pressure (DBP), triglyceride (TG), uric acid (UA), eGFR, SBP/HbA1c, and dietary protein. Covariates listed above were screened based on their regression coefficients relative to ACR with a *P*-value of less than 0.1 (26). There was no multicollinearity effect among the covariates (variance inflation factor (VIF) = 1–4.7). It should be noted that PP was not included as a covariate because of strong collinearity among PP and SBP (VIF > 10). Also, SBP was more positively correlated to ACR than PP, which was consistent with previous studies (27, 28) and demonstrated a stronger relationship between SBP and risk of ACR. Generalized additive models with smooth functions captured the non-linear relationships of ACR with HbA1c and SBP. Then, the threshold levels of HbA1c and SBP were determined using a recursive approach. Likelihood ratio tests were used to assess the difference in fit between the one-line linear regression model with the two-piecewise linear regression model, and $P < 0.05$ was considered significant. Finally, interaction tests were performed between subgroups. Data were analyzed using statistical packages R (The R Foundation; version 3.4.3)[1] and EmpowerStats software (X&Y Solutions, Inc., Boston, MA, United States).[2]

RESULTS

Study Population Characteristics

The detailed clinical characteristics of the 8,626 patients with T2DM included in our study were listed in **Table 1**. When the study population was stratified into four groups according to ACR quartiles. Age, the percentage of participants living with a partner, proportion of participants with an educational level less than high school, the number of current smokers, the proportions of participants with a long diabetes duration (≥10 years), the proportions of participants taking antihypertensive medication, FPG, HbA1c, TG, SBP, PP, and UA levels all showed increased tendency between the four groups with elevated ACR level. BMI was significantly different across groups after being transformed into a categorical variable. No significant differences were observed in Na/K ratio, waist circumference, DBP, total cholesterol (TC), and alanine aminotransferase (ALT).

Association Between Albumin/Creatinine Ratio and HbA1c or Systolic Blood Pressure

To comprehensively explore the relationship of ACR with HbA1c and SBP, we conducted different linear regression models when the independent variables were both treated as continuous and categorical variables. Increased HbA1c and SBP levels (continuous variable) have consistently shown an association with increased lnACR level ($P < 0.001$) whether in the non-adjusted model, the multivariate-adjusted model I and II (**Table 2**). HbA1c and SBP were then transformed into categorical variables by fixed intervals. In the fully adjusted multivariable model II, compared with the reference group of HbA1c (HbA1c < 6), no significant elevated lnACR levels were observed in the second HbA1c group (β = 0.05, 95 CI%: −0.02, 0.12, $P = 0.156$), but the positive association became statistically significant from the third group (β = 0.18, 95 CI%:0.09, 0.27, $P < 0.001$) to highest HbA1c group (β = 0.81, 95 CI%:0.68, 0.94, $P < 0.001$) (**Table 2**). The Changes in SBP also displayed similar trends. Compared to the first group of SBP in multiple linear regression models, only the second group of SBP levels had no relationship with an increased level of lnACR (β = 0.02, 95 CI%: −0.09, 0.12, $P = 0.771$) (**Table 2**). The above results suggested that the positive linear relationships were not always consistent between ACR and HbA1c or SBP. Potential threshold effects might exist in the lower groups of HbA1c and SBP.

Non-linearity of Albumin/Creatinine Ratio With HbA1c and Systolic Blood Pressure

Generalized additive models with smooth functions further revealed the non-linear relationships between lnACR and HbA1c or SBP (**Figure 2**). Data were fitted with the segmented linear models, and two turning points were determined (HbA1c: 6.4%, SBP: 127 mmHg). The likelihood-ratio tests demonstrated that the two-piecewise linear regression models had a better fit ($P < 0.001$) (**Table 3**). However, the threshold effect of HbA1c became significant only after adjustment for confounders, while the threshold effect of SBP remained throughout whether or not the confounders were adjusted. After multivariate adjustment in model II, below the thresholds, no significant correlations were observed between lnACR and HbA1c or SBP. Above the thresholds, lnACR was increased significantly with the increment of HbA1c (β = 0.19, 95 CI%:0.16, 0.22, $P < 0.001$) and SBP (β = 0.03, 95 CI%:0.03, 0.04, $P < 0.001$) (**Table 3**). Notably, the corresponding ACR (mg/g) values for thresholds of HbA1c and SBP were 15.03 (14.44–15.8) and 12.55 (11.94–13.2), respectively, both values being in the normoalbuminuric range (ACR < 30 mg/g).

Combined Thresholds Analysis and Subgroups Analyses

We combined discovered thresholds and explored the comprehensive effect of HbA1c and SBP levels on changes in ACR. In parallel, subgroups analyses were performed separately based on different thresholds. When the study population was divided into four groups based on two thresholds, we discovered that the dose-dependent positive relationship between the groups and the risk of elevated lnACR levels was consistently present whether adjusted for covariates (**Table 4**). Compared with the population who had both HbA1c and SBP levels below the thresholds, the population simultaneous above the thresholds had the fastest increase in lnACR (β = 0.67, 95 CI%:0.58, 0.76,

[1]http://www.r-project.org

[2]www.empowerstats.net/cn/

TABLE 1 | The clinical characteristics of enrolled participants were stratified by albumin/creatinine ratio (ACR) quartiles.

Characteristic	ACR (mg/g)				*p*-value
	Q1(<6.58)*n* = 2,154	Q2(6.58 - 12.62)*n* = 2,158	Q3(12.62 - 40.46)*n* = 2,157	Q4(≥40.46)*n* = 2,157	
Age (years)	57.56 ± 13.59	60.76 ± 13.56	62.15 ± 14.09	64.08 ± 13.42	<0.001
Sex					<0.001
Male	1198 (55.62%)	1022 (47.36%)	1032 (47.84%)	1237 (57.35%)	
Female	956 (44.38%)	1136 (52.64%)	1125(52.16%)	920 (42.65%)	
Race					<0.001
Mexican American	379 (17.60%)	420 (19.46%)	439 (20.35%)	502 (23.27%)	
Other Hispanic	196 (9.10%)	207 (9.59%)	219 (10.15%)	195 (9.04%)	
Non-hispanic White	814 (37.79%)	836 (38.74%)	830 (38.48%)	749 (34.72%)	
Non-hispanic black	560 (26.00%)	466 (21.59%)	460 (21.33%)	530 (24.57%)	
Other race	205 (9.52%)	229 (10.61%)	209 (9.69%)	181 (8.39%)	
Marital status					<0.001
Living with partner	1399 (64.95%)	1296 (60.06%)	1264 (58.60%)	1203 (55.77%)	
Living without partner	740 (34.35%)	847 (39.25%)	882 (40.89%)	937 (43.44%)	
Not recorded	15 (0.70%)	15 (0.70%)	11 (0.51%)	17 (0.79%)	
Education level					<0.001
Less than high school	664 (30.83%)	754 (34.94%)	791 (36.67%)	946 (43.86%)	
High school	498 (23.12%)	516 (23.91%)	492 (22.81%)	460 (21.33%)	
More than high school	992 (46.05%)	888 (41.15%)	874 (40.52%)	751 (34.82%)	
Smoking					<0.001
Current	328 (15.23%)	335 (15.52)	314 (14.56)	346 (16.04)	
Past	700 (32.50)	720 (33.36%)	745 (34.54%)	821 (38.06%)	
Never	1126 (52.27%)	1103 (51.11%)	1098 (50.90%)	990 (45.90%)	
Alcohol consumption					<0.001
Yes	1306 (60.63%)	1178 (54.59%)	1153 (53.45%)	1199 (55.59%)	
No	707 (32.82%)	842 (39.02%)	856 (39.68%)	806 (37.37%)	
Not recorded	141 (6.55%)	138 (6.39%)	148 (6.86%)	152 (7.05%)	
Dietary protein (g/d)	79.28 ± 35.07	74.56 ± 32.40	73.88 ± 32.40	73.19 ± 33.52	<0.001
Sodium intake (mg/d)	3300.99 ± 1562.04	3130.04 ± 1446.84	3122.00 ± 1407.22	3035.04 ± 1459.83	<0.001
Potassium intake (mg/d)	2597.78 ± 1050.19	2516.54 ± 1053.73	2489.65 ± 1062.69	2382.29 ± 1006.80	<0.001
Na/K ratio	1.34 ± 0.50	1.31 ± 0.50	1.33 ± 0.50	1.34 ± 0.52	0.172
Diabetes duration (years)					<0.001
<5	364 (16.90%)	339 (15.71%)	328 (15.21%)	227 (10.52%)	
≥5, <10	217 (10.07%)	275 (12.74%)	253 (11.73%)	233 (10.80%)	
≥10	418 (19.41%)	461 (21.36%)	516 (23.92%)	803 (37.23%)	
Not recorded	1155 (53.62%)	1083 (50.19%)	1060 (49.14%)	894 (41.45%)	
BMI (kg/m^2)	32.03 ± 6.98	31.85 ± 7.23	31.68 ± 7.25	31.60 ± 7.04	0.141
BMI (kg/m^2)					0.022
<25	268 (12.44%)	322 (14.92%)	326 (15.11%)	343 (15.90%)	
≥25, <30	683 (31.71%)	644 (29.84%)	668 (30.97%)	616 (28.56%)	
≥ 30	1203 (55.85%)	1192 (55.24%)	1163 (53.92%)	1198 (55.54%)	
Waist circumference (cm)	107.55 ± 15.13	107.41 ± 15.26	107.58 ± 15.62	108.35 ± 15.22	0.121
Waist circumference (cm)					0.052
<102(male), < 88(female)	525 (24.37%)	459 (21.27%)	461 (21.37%)	483 (22.39%)	
≥102(male), ≥ 88(female)	1629 (75.63%)	1699 (78.73%)	1696 (78.63%)	1674 (77.61%)	
HOMA-IR					<0.001
Lower group	481 (22.33%)	453 (20.99%)	427 (19.80%)	387 (17.94%)	
Higher group	425 (19.73%)	472 (21.87%)	452 (20.96%)	400 (18.54%)	
Not recorded	1248 (57.94%)	1233 (57.14%)	1278 (59.25%)	1370 (63.51%)	
FPG (mmol/L)	7.62 ± 3.01	8.04 ± 3.27	8.79 ± 4.02	9.46 ± 4.59	<0.001
HbA1c (%)	6.69 ± 1.36	6.95 ± 1.54	7.30 ± 1.77	7.75 ± 2.07	<0.001
SBP (mmHg)	125.26 ± 15.49	129.13 ± 17.38	134.24 ± 20.34	141.95 ± 23.63	<0.001
DBP (mmHg)	69.15 ± 11.96	69.30 ± 12.99	69.45 ± 13.81	69.98 ± 15.25	0.225

(Continued)

TABLE 1 | (Continued)

Characteristic	ACR (mg/g)				*p*-value
	Q1(<6.58)*n* = 2,154	Q2(6.58 - 12.62)*n* = 2,158	Q3(12.62 - 40.46)*n* = 2,157	Q4(≥40.46)*n* = 2,157	
PP (mmHg)	56.11 ± 17.27	59.83 ± 18.62	64.79 ± 21.22	71.97 ± 24.71	<0.001
TC(mmol/L)	4.91 ± 1.15	4.93 ± 1.12	4.97 ± 1.24	4.99 ± 1.36	0.757
TG (mmol/L)	2.03 ± 1.64	2.05 ± 1.75	2.31 ± 2.08	2.39 ± 2.41	<0.001
HDL-C (mmol/L)	1.25 ± 0.36	1.28 ± 0.39	1.24 ± 0.37	1.24 ± 0.40	0.009
ALT (U/L)	27.51 ± 18.99	27.16 ± 20.55	27.22 ± 20.29	26.54 ± 38.09	0.652
AST (U/L)	26.44 ± 15.26	26.48 ± 22.14	26.51 ± 15.23	26.46 ± 21.53	0.002
Albumin (G/L)	41.48 ± 3.21	41.73 ± 3.18	41.68 ± 3.23	40.61 ± 3.80	<0.001
UA(umol/L)	343.55 ± 83.88	333.48 ± 87.50	335.04 ± 95.04	357.41 ± 101.19	<0.001
Scr (umol/L)	82.33 ± 23.14	78.18 ± 24.10	80.11 ± 30.13	105.23 ± 81.42	<0.001
eGFR (ml/min/1.73 m^2)	83.98 ± 23.92	86.71 ± 27.37	86.09 ± 29.60	75.84 ± 35.53	<0.001
Taking medication					
ACEI/ARB	913 (42.39%)	991 (45.92%)	1034 (47.94%)	1118 (51.83%)	<0.001
SGLT-2	8 (0.37%)	22 (1.02%)	15 (0.70%)	9 (0.42%)	0.026

ACR, albumin/creatinine ratio; Na/K ratio, sodium/potassium ratio; BMI, body mass index; HOMA-IR, homeostasis model assessment of insulin resistance; FPG, fasting plasma glucose; HbA1c, glycated hemoglobin; SBP, systolic blood pressure; DBP, diastolic blood pressure; PP, pulse pressure; TC, total cholesterol; TG, triglyceride; HDL-C, high-density lipoprotein cholesterol; ALT, alanine aminotransferase; AST, aspartate aminotransferase; UA, uric acid; Scr, serum creatinine; eGFR, estimated glomerular filtration rate; ACEI, angiotensin-converting enzyme inhibitor; ARB, angiotensin receptor blocker; SGLT-2, sodium-glucose cotransporter 2. Data are present as n (%) or the mean ± standard deviation.

TABLE 2 | The relationship between ACR and HbA1c or SBP using linear regression analysis.

	N	LnACR (mg/g)					
		Non-adjusted model		Multivariate-adjusted model I		Multivariate-adjusted model II	
		β (95CI)	*p*-value	β (95CI)	*p*-value	β (95CI)	*p*-value
HbA1C (%)(continuous variable)	8626	0.21 (0.19, 0.22)	<0.001	0.19 (0.17, 0.21)	<0.001	0.16 (0.13, 0.18)	<0.001
HbA1C (%)(categorical variable)							
<6.0	2256	reference		reference		reference	
6.0–7.0	2851	0.20 (0.13, 0.28)	<0.001	0.08 (0.01, 0.16)	0.030	0.05 (−0.02, 0.12)	0.160
7.0–8.0	1600	0.49 (0.40, 0.58)	<0.001	0.30 (0.21, 0.40)	<0.001	0.19 (0.10, 0.27)	<0.001
8.0–9.0	779	0.64 (0.52, 0.75)	<0.001	0.48 (0.37, 0.60)	<0.001	0.38 (0.26, 0.49)	<0.001
≥9.0	1140	1.08 (0.98, 1.19)	<0.001	1.02 (0.91, 1.12)	<0.001	0.81 (0.68, 0.94)	<0.001
SBP(mmHg)(continuous variable)	8626	0.03 (0.02, 0.03)	<0.001	0.02 (0.02, 0.03)	<0.001	0.02 (0.02, 0.02)	<0.001
SBP (mmHg)(categorical variable)							
<110	861	reference		reference		reference	
110–120	1487	−0.01 (−0.11, 0.10)	0.912	0.02 (−0.09, 0.12)	0.734	0.02 (−0.09, 0.12)	0.768
120–130	1919	0.21 (0.11, 0.32)	<0.001	0.20 (0.10, 0.30)	<0.001	0.17 (0.06, 0.27)	0.001
130–140	1708	0.37 (0.26, 0.47)	<0.001	0.33 (0.22, 0.44)	<0.001	0.31 (0.20, 0.41)	< 0.001
140–150	1091	0.72 (0.60, 0.84)	<0.001	0.68 (0.56, 0.80)	<0.001	0.65 (0.53, 0.77)	< 0.001
150–160	697	1.09 (0.95, 1.23)	<0.001	1.05 (0.92, 1.19)	<0.001	0.98 (0.84, 1.11)	< 0.001
≥160	863	1.76 (1.63, 1.90)	<0.001	1.70 (1.57, 1.84)	<0.001	1.62 (1.48, 1.75)	< 0.001

LnACR, ln-transformed albumin/creatinine ratio; HbA1c, glycated hemoglobin; SBP, systolic blood pressure. Multivariate-Adjusted Model I adjusted for: age, sex, marital status, education level, smoking, alcohol consumption, diabetes duration, body mass index (continuous), and waist circumference (continuous). Multivariate-Adjusted Model II adjusted for: age, sex, marital status, education level, smoking, alcohol consumption, diabetes duration, body mass index (continuous), waist circumference (continuous), fasting plasma glucose, glycated hemoglobin/systolic blood pressure, diastolic blood pressure, triglyceride, uric acid, estimated glomerular filtration rate and dietary protein.

$P < 0.001$). A rapid increase in lnACR level was more relevant to higher SBP levels above the threshold (≥127 mmHg) (**Table 4**). When subgroup analyses were carried out for patients with HbA1c ≥ 6.4%, significant interactions were observed both in the diabetes duration subgroup (interaction $P < 0.001$), waist circumference subgroup (interaction $P = 0.029$), dietary protein subgroup (interaction $P = 0.043$) and Na/K ratio subgroup (interaction $P = 0.02$) (**Figure 3**). In addition, there were also interaction effects between SBP with diabetes duration group (interaction $P < 0.001$), dietary protein subgroup (interaction $P < 0.001$), UA group (interaction $P < 0.001$), ACR group (interaction $P < 0.001$), and eGFR group (interaction $P < 0.001$)

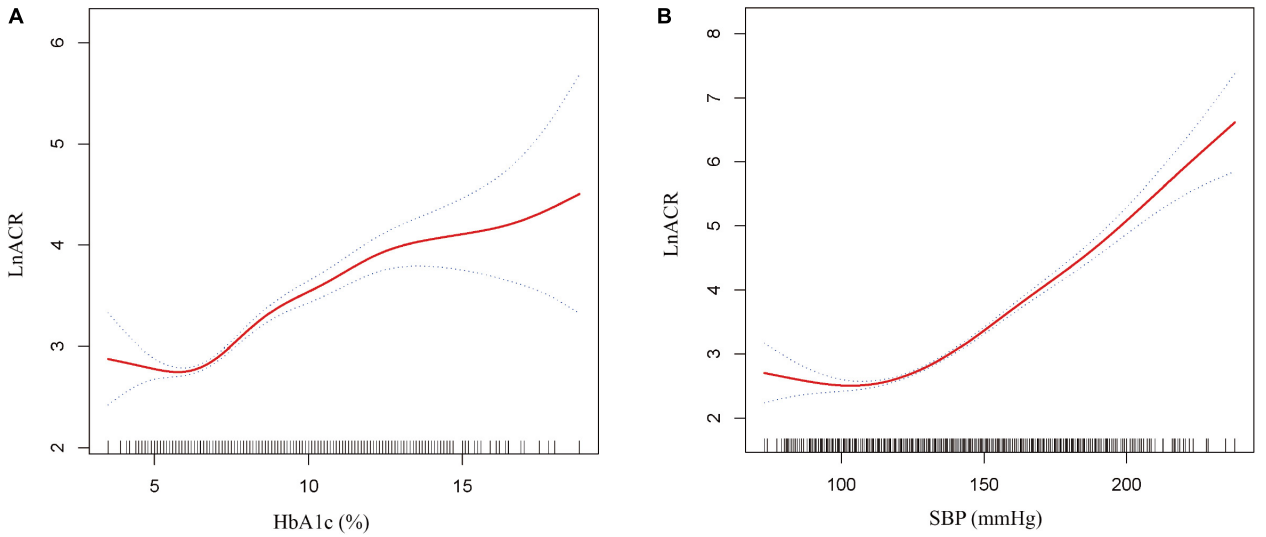

FIGURE 2 | Non-linear relationship between ACR and HbA1c or SBP. **(A)** lnACR with HbA1c **(B)** lnACR with SBP. The solid red line is the fitted curves and the dotted curves are the 95 CI of the fit. All analyses were adjusted for age, sex, marital status, education level, smoking, alcohol consumption, diabetes duration, body mass index (continuous), waist circumference (continuous), fasting plasma glucose, glycated hemoglobin/systolic blood pressure, diastolic blood pressure, triglyceride, uric acid, estimated glomerular filtration rate and dietary protein.

TABLE 3 | Threshold effect analysis of HbA1c or SBP on ACR using two-piecewise linear regression.

		LnACR (mg/g)					
	N	**Non-adjusted model**		**Multivariate-adjustedmodel I**		**Multivariate-adjustedmodel II**	
		β (95CI)	***p*-value**	**β (95CI)**	***p*-value**	**β (95CI)**	***p*-value**
HbA1C (%)							
<6.4	3184	0.25 (0.14, 0.37)	<0.001	0.07 (−0.05, 0.18)	0.254	0.07 (−0.04, 0.18)	0.258
≥6.4	5442	0.19 (0.17, 0.22)	<0.001	0.21 (0.19, 0.24)	<0.001	0.19 (0.16, 0.22)	<0.001
P for log-likelihood ratio test			0.092		0.002		<0.001
SBP (mmHg)							
<127	3760	0.01 (0.00, 0.01)	0.024	0.00 (0.00, 0.01)	0.031	0.00 (0.00, 0.01)	0.051
≥127	4866	0.03 (0.03, 0.04)	<0.001	0.03 (0.03, 0.04)	<0.001	0.03 (0.03, 0.04)	<0.001
P for log-likelihood ratio test			<0.001		<0.001		<0.001

LnACR, ln-transformed albumin/creatinine ratio; HbA1c, glycated hemoglobin; SBP, systolic blood pressure. Multivariate-Adjusted Model I adjusted for: age, sex, marital status, education level, smoking, alcohol consumption, diabetes duration, body mass index (continuous), and waist circumference (continuous). Multivariate-Adjusted Model II adjusted for: age, sex, marital status, education level, smoking, alcohol consumption, diabetes duration, body mass index (continuous), waist circumference (continuous), fasting plasma glucose, glycated hemoglobin/systolic blood pressure, diastolic blood pressure, triglyceride, uric acid, estimated glomerular filtration rate and dietary protein.

TABLE 4 | Analysis of the combined threshold effect of both HbA1and SBP on ACR.

HbA1c (%) & SBP (mmHg)		**LnACR (mg/g)**					
	N	**Non-adjusted model**	***p*-value**	**Multivariate-adjusted model I**	***p*-value**	**Multivariate-adjusted model II**	***p*-value**
<6.4, <127	1505	reference		reference		reference	
≥6.4, <127	2255	0.41 (0.32, 0.50)	<0.001	0.29 (0.20, 0.37)	<0.001	0.12 (0.03, 0.21)	0.006
<6.4, ≥127	1679	0.62 (0.53, 0.72)	<0.001	0.55 (0.45, 0.64)	<0.001	0.48 (0.39, 0.58)	<0.001
≥6.4, ≥127	3187	1.06 (0.98, 1.14)	<0.001	0.88 (0.80, 0.97)	<0.001	0.67 (0.58, 0.76)	<0.001

LnACR, ln-transformed albumin/creatinine ratio; HbA1c, glycated hemoglobin; SBP, systolic blood pressure. Multivariate-Adjusted Model I adjusted for: age, sex, marital status, education level, smoking, alcohol consumption, diabetes duration, body mass index (continuous), and waist circumference (continuous). Multivariate-Adjusted Model II adjusted for: age, sex, marital status, education level, smoking, alcohol consumption, diabetes duration, body mass index (continuous), waist circumference (continuous), fasting plasma glucose, diastolic blood pressure, triglyceride, uric acid, estimated glomerular filtration rate and dietary protein.

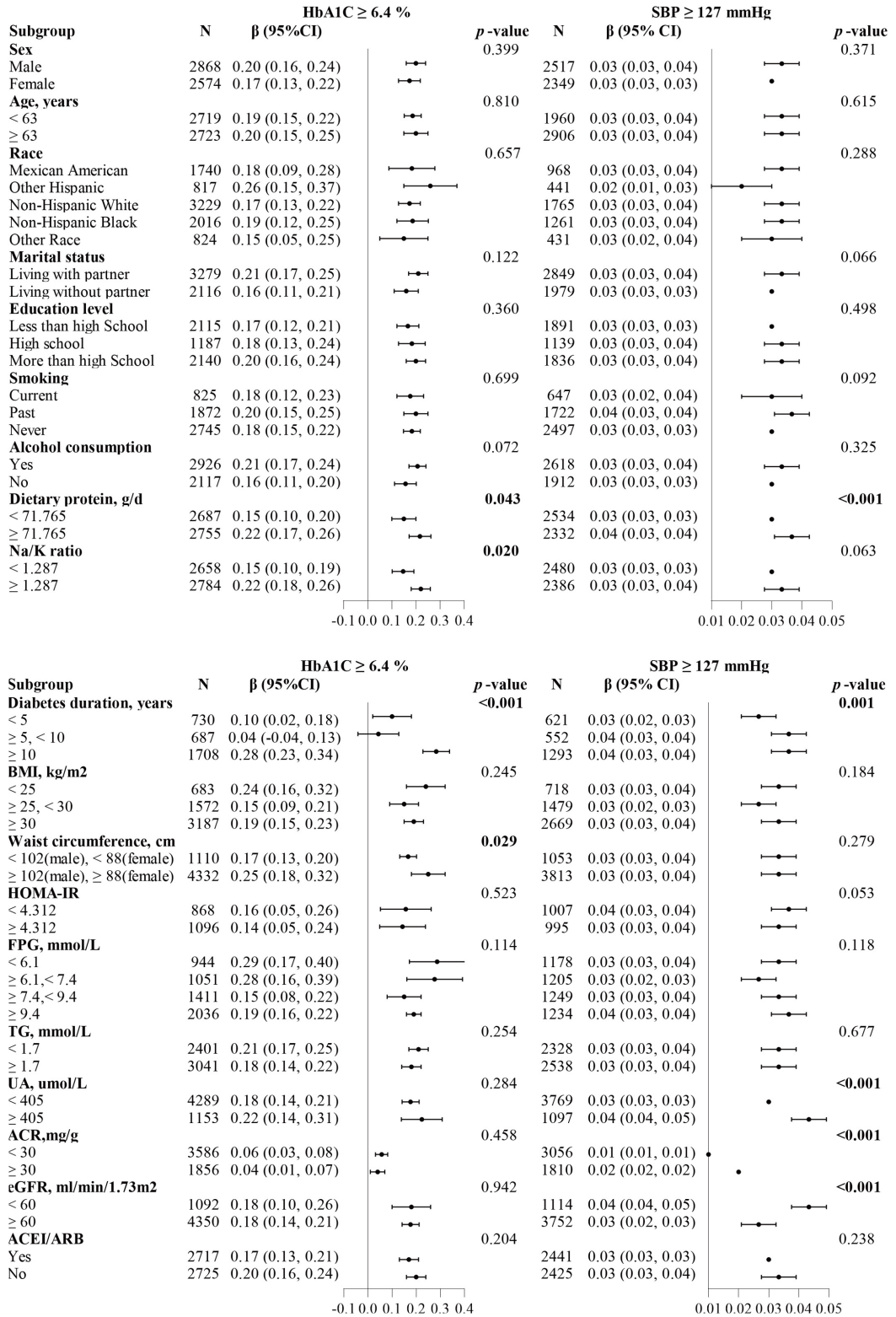

FIGURE 3 | Forest plots summarizing the subgroups analyses for ACR with HbA1c or SBP divided by thresholds (HbA1c ≥ 6.4%, SBP ≥ 127 mmHg). The dietary protein, Na/K ratio, and HOMA-IR subgroups were divided based on the median. Each subgroup analysis adjusted for age, sex, marital status, education level, smoking, alcohol consumption, diabetes duration, body mass index (continuous), waist circumference (continuous), fasting plasma glucose, glycated hemoglobin/systolic blood pressure, diastolic blood pressure, triglyceride, uric acid, estimated glomerular filtration rate and dietary protein, except the subgrouping variables.

among the patients with T2DM who had a SBP level above 127 mmHg (**Figure 3**).

DISCUSSION

Our study elaborated on the relationship curves between ACR and HbA1c or SBP in patients with T2DM and discovered the different risk thresholds of HbA1c and SBP (HbA1c = 6.4% and SBP = 127 mmHg) above which the risk of ACR increases significantly. Additionally, more pronounced risk relationships were detected in participants with longer-duration diabetes, central obesity, or hyperuricemia.

Previously, one study discovered the threshold effect between HbA1C and ACR among a Chinese population, but it has not been studied in diabetic people (14). We first confirmed a similar association in patients with T2DM, which suggested that there exists an obvious ACR rising period that we are easy to ignore before progression to microalbuminuria. The gap between the risk threshold of HbA1c obtained in our study (6.4%) and recommended HbA1c targets (7%) (29) might be related to the early control of the above period. Notably, to define target HbA1c control levels, not only the risk of ACR progression should be taken into account, but the incidence of renal endpoints, the ultimate risk of death, and the occurrence of adverse events. Appropriate glucose control (HbA1c < 7%) recommended by the guidelines was based on a famous landmark UKPDS study (30), while the ACCORD research highlighted that intensive glucose control (HbA1c < 6%) could not reduce microvascular outcome events (31). In addition, a large-scale study with up to 13 years of follow-up reported that strict control of glucose (HbA1c < 6.5%) in the first year after newly diagnosed type 2 diabetes was associated with lower risks of diabetic vascular complications and reduced mortality (32). The aforementioned studies implied that it might be reasonable to control the HbA1c level within 6–7%, and some newly diagnosed patients would benefit more with HbA1c values < 6.5%. The threshold value (HbA1c = 6.4%) obtained in our study was also within the above range. Furthermore, a large prospective cohort study of older German adults demonstrated that increasing HbA1c (≥6.4%) was closely associated with a more than a 3-fold increased risk of decreased renal function (33). This result was generally consistent with our findings, Further, it demonstrated that there might have both short-term and long-term renal function protection when the HbA1c level was controlled below 6.4% in patients with T2DM.

As another crucial risk factor for ACR, SBP exhibited a similar threshold effect to HbA1c. However, all extensive studies emphasized the approximate range of SBP control and did not reveal the specific threshold, nor did they evaluate the risky situations under continuous changes in SBP. The existing authoritative research (34–37) results showed that patients with T2DM had a relative positive benefit-risk balance with SBP control between 120 and 140 mmHg. A prospective study on T2DM veterans discovered a significant protective benefit from lowering SBP below 130 mmHg (38), which suggested that a tighter range (120–130 mmHg) for SBP control may be required. The 127 mmHg threshold of SBP obtained in our study is also within this range. Crucially, the risk threshold detected in our study could be instrumental in the future experiment design of SBP control levels to assess long-term effects.

Combined analyses of thresholds showed the lowest ACR levels when both HbA1c and SBP control levels were below the thresholds. This was in accordance with most other studies (39, 40). Of additional concern, compared to patients with T2DM with HbA1c ≥ 6.4% and SBP < 127 mmHg, a stronger association with elevated ACR was observed in subjects with HbA1c < 6.4% and SBP ≥ 127 mmHg. These findings implied that well-controlled SBP was likely to play a more significant role in reducing urine protein levels and should be elucidated by further studies. Finally, after complete adjustment for confounding factors, the results of the subgroup analysis partially explained the heterogeneity. Our results found that longer diabetes duration and higher protein intake had interacted with HbA1c (≥6.4%) and SBP (≥127 mmHg) in the risk of ACR progression. These, too, were in keeping with previous findings. Duration of diabetes was an unmodifiable risk factor of ACR in patients with T2DM (41) while a high protein diet can exacerbate hypertension and expedite glomerular damage (42). Additionally, our results showed central obesity and a higher Na/K intake ratio could impose an extra burden on the kidney in patients with T2DM who had HbA1c ≥ 6.4%. It has been reported that central obesity could aggravate insulin resistance (43), and lead to the progression of abnormal renal hemodynamics and podocyte injury (44). The higher Na/K intake ratio might cause endothelial insult and elevate urinary protein levels (45). When SBP ≥ 127 mmHg, a more rapid rise in ACR was observed in patients with T2DM with renal insufficiency or hyperuricemia. This may be closely related to compromised kidney regulation and marked glomerular hypertension caused by the combined effects of diabetes status, hypertension, and impaired kidney function (46). Hyperuricemia is recognized as one of the risk factors for the development and progression of diabetic kidney disease. It could activate the RAAS system, further increasing blood pressure levels to promote ACR progression in patients with T2DM (47). No significant differences were identified in the subgroup analyses of age, gender, education levels, marital status, smoking and alcohol consumption, and blood lipids, which indicated that our results remain stable across most subsamples.

There are two significant clinical implications in our study. First, we identified the risk thresholds of rapid ACR progression and provided valuable references for both early blockades of DKD occurrence and development. Different from the conventional studies dividing population-based on ACR ≥ 30 mg/g to explore the potential risk factors, we described consecutive changes in ACR and observed the risk thresholds of HbA1c and SBP at an earlier level of ACR. As for patients with T2DM with normal urine protein levels, tightly controlling HbA1c and SBP within the threshold levels emphasize no proteinuria and the maintenance of long-term stable normal urine protein

levels. For patients with T2DM along with proteinuria, the same control below the threshold levels may have positive significance in delaying the progression of DKD and even reversing to normal urinary protein levels (48). Second, our study further explored high-risk populations with rapid proteinuria progression, which provided partial references for individualized prevention and targeted intervention. For example, patients with T2DM diagnosed with unsatisfactory HbA1c level control should pay attention to weight management and moderately limit their protein and salt intake; patients with high SBP levels should not only reduce blood pressure reasonably but also need to check renal function regularly to prevent hyperuricemia.

Of course, our study has the following limitations. This study is cross-sectional and lacks longitudinal follow-up assessments, including primary endpoint and adverse events. More prospective studies based on our thresholds are needed in the future. In addition, it remains uncertain whether our results are generally applicable to other populations, such as Asian populations, since the enrolled participants are all from the United States.

CONCLUSION

In type 2 diabetic population, we identified distinct thresholds of HbA1c and SBP (HbA1c = 6.4% and SBP = 127 mmHg) beyond which an elevated albuminuria risk would become significant. Additionally, central obesity and higher Na/K intake ratio could further increase the albuminuria risk in patients with T2DM who had HbA1c $\geq$ 6.4% while hyperuricemia and higher protein intake have similar effects in patients with T2DM who had SBP $\geq$ 127 mmHg. Our findings might have important clinical implications for the early prevention and control of DKD.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention (CDC). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

HL: conception and design. JX and YX: drafting of the manuscript and data analysis. QC and XH: reviewed/edited the manuscript. MC and JT: selection of literature and interpretation. SJ: making figures and tables. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the National Natural Science Foundation of China, grant numbers 81503552 and 81874434, the Shanghai Municipal Key Clinical Specialty, grant number shslczdzk05401, and the Shanghai Key Laboratory of Traditional Chinese Clinical Medicine, grant number 14DZ2273200.

ACKNOWLEDGMENTS

We appreciate the useful discussions with HL.

REFERENCES

1. Coresh J, Heerspink HJL, Sang Y, Matsushita K, Arnlov J, Astor BC, et al. Change in albuminuria and subsequent risk of end-stage kidney disease: an individual participant-level consortium meta-analysis of observational studies. *Lancet Diabetes Endocrinol.* (2019) 7:115–27. doi: 10.1016/S2213-8587(18)30313-9
2. Neuen BL, Weldegiorgis M, Herrington WG, Ohkuma T, Smith M, Woodward M. Changes in GFR and albuminuria in routine clinical practice and the risk of kidney disease progression. *Am J Kidney Dis.* (2021) 78:350–360.e1. doi: 10.1053/j.ajkd.2021.02.335
3. Satchell SC, Tooke JE. What is the mechanism of microalbuminuria in diabetes: a role for the glomerular endothelium? *Diabetologia.* (2008) 51:714–25. doi: 10.1007/s00125-008-0961-8
4. Vaduganathan M, Pareek M, Kristensen AMD, Biering-Sorensen T, Byrne C, Almarzooq Z, et al. Prevention of heart failure events with intensive versus standard blood pressure lowering across the spectrum of kidney function and albuminuria: a sprint substudy. *Eur J Heart Fail.* (2021) 23:384–92. doi: 10.1002/ejhf.1971
5. Sacre JW, Magliano DJ, Shaw JE. Heart failure hospitalisation relative to major atherosclerotic events in type 2 diabetes with versus without chronic kidney disease: a meta-analysis of cardiovascular outcomes trials. *Diabetes Metab.* (2021) 47:101249. doi: 10.1016/j.diabet.2021.101249
6. Koo BK, Chung WY, Moon MK. Peripheral arterial endothelial dysfunction predicts future cardiovascular events in diabetic patients with albuminuria: a prospective cohort study. *Cardiovasc Diabetol.* (2020) 19:82. doi: 10.1186/s12933-020-01062-z
7. Sumida K, Nadkarni GN, Grams ME, Sang Y, Ballew SH, Coresh J, et al. Conversion of urine protein-creatinine ratio or urine dipstick protein to urine albumin-creatinine ratio for use in chronic kidney disease screening and prognosis : an individual participant-based meta-analysis. *Ann Intern Med.* (2020) 173:426–35. doi: 10.7326/M20-0529
8. Jiang W, Wang J, Shen X, Lu W, Wang Y, Li W, et al. Establishment and validation of a risk prediction model for early diabetic kidney disease based on a systematic review and meta-analysis of 20 cohorts. *Diabetes Care.* (2020) 43:925–33. doi: 10.2337/dc19-1897
9. Krolewski AS, Niewczas MA, Skupien J, Gohda T, Smiles A, Eckfeldt JH, et al. Early progressive renal decline precedes the onset of microalbuminuria and its progression to macroalbuminuria. *Diabetes Care.* (2014) 37:226–34. doi: 10.2337/dc13-0985
10. Elley CR, Robinson T, Moyes SA, Kenealy T, Collins J, Robinson E, et al. Derivation and validation of a renal risk score for people with

type 2 diabetes. *Diabetes Care.* (2013) 36:3113–20. doi: 10.2337/dc13-0190
11. Shin JI, Chang AR, Grams ME, Coresh J, Ballew SH, Surapaneni A, et al. Albuminuria testing in hypertension and diabetes: an individual-participant data meta-analysis in a global consortium. *Hypertension.* (2021) 78:1042–52. doi: 10.1161/HYPERTENSIONAHA.121.17323
12. Kinguchi S, Wakui H, Ito Y, Kondo Y, Azushima K, Osada U, et al. Improved home BP profile with dapagliflozin is associated with amelioration of albuminuria in japanese patients with diabetic nephropathy: the yokohama add-on inhibitory efficacy of dapagliflozin on albuminuria in japanese patients with type 2 diabetes study (Y-Aida study). *Cardiovasc Diabetol.* (2019) 18:110. doi: 10.1186/s12933-019-0912-3
13. Nordwall M, Abrahamsson M, Dhir M, Fredrikson M, Ludvigsson J, Arnqvist HJ. Impact of Hba1c, followed from onset of type 1 diabetes, on the development of severe retinopathy and nephropathy: the viss study (vascular diabetic complications in Southeast Sweden). *Diabetes Care.* (2015) 38:308–15. doi: 10.2337/dc14-1203
14. Lian H, Wu H, Ning J, Lin D, Huang C, Li F, et al. The risk threshold for hemoglobin A1c associated with albuminuria: a population-based study in China. *Front Endocrinol (Lausanne).* (2021) 12:673976. doi: 10.3389/fendo.2021.673976
15. Atkin SL, Butler AE, Hunt SC, Kilpatrick ES. The retinopathy-derived Hba1c threshold of 6.5% for type 2 diabetes also captures the risk of diabetic nephropathy in Nhanes. *Diabetes Obes Metab.* (2021) 23:2109–15. doi: 10.1111/dom.14449
16. Bohm M, Schumacher H, Teo KK, Lonn EM, Mahfoud F, Emrich I, et al. Renal outcomes and blood pressure patterns in diabetic and nondiabetic individuals at high cardiovascular risk. *J Hypertens.* (2021) 39:766–74. doi: 10.1097/HJH.0000000000002697
17. Sung KC, Ryu S, Lee JY, Lee SH, Cheong E, Hyun YY, et al. Urine Albumin/Creatinine Ratio Below 30 Mg/G Is a Predictor of Incident Hypertension and Cardiovascular Mortality. *J Am Heart Assoc.* (2016) 5:e003245. doi: 10.1161/JAHA.116.003245
18. Xie X, Peng Z, Li H, Li D, Tu Y, Bai Y, et al. Association of urine albumin/creatinine ratio below 30 mg/g and left ventricular hypertrophy in patients with type 2 diabetes. *Biomed Res Int.* (2020) 2020:5240153. doi: 10.1155/2020/5240153
19. American Diabetes Association Professional Practice Committee, Draznin B, Aroda VR, Bakris G, Benson G. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes-2022. *Diabetes Care.* (2022) 45(Suppl. 1):S17–38. doi: 10.2337/dc22-S002
20. Wang S, Wang Y, Wan X, Guo J, Zhang Y, Tian M, et al. Cobalamin intake and related biomarkers: examining associations with mortality risk among adults with type 2 diabetes in Nhanes. *Diabetes Care.* (2022) 45:276–84. doi: 10.2337/dc21-1674
21. Gong R, Luo G, Wang M, Ma L, Sun S, Wei X. Associations between Tg/Hdl ratio and insulin resistance in the us population: a cross-sectional study. *Endocr Connect.* (2021) 10:1502–12. doi: 10.1530/EC-21-0414
22. Va P, Dodd KW, Zhao L, Thompson-Paul AM, Mercado CI, Terry AL, et al. Evaluation of measurement error in 24-hour dietary recall for assessing sodium and potassium intake among US adults – national health and nutrition examination survey (Nhanes), 2014. *Am J Clin Nutr.* (2019) 109:1672–82. doi: 10.1093/ajcn/nqz044
23. Kushner RF, Ryan DH. Assessment and lifestyle management of patients with obesity: clinical recommendations from systematic reviews. *JAMA.* (2014) 312:943–52. doi: 10.1001/jama.2014.10432
24. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF III, Feldman HI, et al. A new equation to estimate glomerular filtration rate. *Ann Intern Med.* (2009) 150:604–12. doi: 10.7326/0003-4819-150-9-200905050-00006
25. Johnson CL, Paulose-Ram R, Ogden CL, Carroll MD, Kruszon-Moran D, Dohrmann SM, et al. National health and nutrition examination survey: analytic guidelines, 1999-2010. *Vital Health Stat.* (2013) 2:1–24.
26. Tang H, Liu N, Feng X, Yang Y, Fang Y, Zhuang S, et al. Circulating levels of Il-33 are elevated by obesity and positively correlated with metabolic disorders in Chinese adults. *J Transl Med.* (2021) 19:52. doi: 10.1186/s12967-021-02711-x
27. Sheen YJ, Lin JL, Li TC, Bau CT, Sheu WH. Systolic blood pressure as a predictor of incident albuminuria and rapid renal function decline in type 2 diabetic patients. *J Diabetes Complications.* (2014) 28:779–84. doi: 10.1016/j.jdiacomp.2014.08.002
28. Strandberg TE, Pitkala K. What is the most important component of blood pressure: systolic, diastolic or pulse pressure? *Curr Opin Nephrol Hypertens.* (2003) 12:293–7. doi: 10.1097/00041552-200305000-00011
29. American Diabetes Association. 6. Glycemic targets: standards of medical care in diabetes-2021. *Diabetes Care.* (2021) 44(Suppl. 1):S73–84. doi: 10.2337/dc21-S006
30. Turner R, Fox C, Matthews, Mcelroy H, Cull C, Holman R, et al. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (Ukpds 33). UK prospective diabetes study (Ukpds) group. *Lancet.* (1998) 352:837–53.
31. Ismail-Beigi F, Craven T, Banerji MA, Basile J, Calles J, Cohen RM, et al. Effect of intensive treatment of hyperglycaemia on microvascular outcomes in type 2 diabetes: an analysis of the accord randomised trial. *Lancet.* (2010) 376:419–30. doi: 10.1016/S0140-6736(10)60576-4
32. Laiteerapong N, Ham SA, Gao Y, Moffet HH, Liu JY, Huang ES, et al. The legacy effect in type 2 diabetes: impact of early glycemic control on future complications (the diabetes & aging study). *Diabetes Care.* (2019) 42:416–26. doi: 10.2337/dc17-1144
33. Schottker B, Brenner H, Koenig W, Muller H, Rothenbacher D. Prognostic association of Hba1c and fasting plasma glucose with reduced kidney function in subjects with and without diabetes mellitus. results from a population-based cohort study from Germany. *Prev Med.* (2013) 57:596–600.
34. Szyndler A. [Commentary to the articles: kaplan nm. vascular outcome in type 2 diabetes: an advance? lancet 2007; 370:804-5; Patel a; advance collaborative group, Macmahon S, Chalmers J, Neal B I Wsp. Effects of a fixed combination of perindopril and indapamide on macrovascular and microvascular outcomes in patients with type 2 diabetes mellitus (the advance trial): a randomised controlled trial. lancet 2007; 370:829-40]. *Kardiol Pol.* (2007) 65:1527–9; discussion 30.
35. Group AS, Cushman WC, Evans GW, Byington RP, Goff DC Jr., Grimm RH Jr., et al. Effects of intensive blood-pressure control in type 2 diabetes mellitus. *N Engl J Med.* (2010) 362:1575–85. doi: 10.1056/NEJMoa1001286
36. Group SR, Wright JT Jr., Williamson JD, Whelton PK, Snyder JK, Sink KM, et al. A randomized trial of intensive versus standard blood-pressure control. *N Engl J Med.* (2015) 373:2103–16. doi: 10.1056/NEJMoa1511939
37. Beddhu S, Greene T, Boucher R, Cushman WC, Wei G, Stoddard G, et al. Intensive systolic blood pressure control and incident chronic kidney disease in people with and without diabetes mellitus: secondary analyses of two randomised controlled trials. *Lancet Diabetes Endocrinol.* (2018) 6:555–63. doi: 10.1016/S2213-8587(18)30099-8
38. Anderson RJ, Bahn GD, Emanuele NV, Marks JB, Duckworth WC, Group VS. Blood pressure and pulse pressure effects on renal outcomes in the veterans affairs diabetes trial (Vadt). *Diabetes Care.* (2014) 37:2782–8. doi: 10.2337/dc14-0284
39. Zoungas S, de Galan BE, Ninomiya T, Grobbee D, Hamet P, Heller S, et al. Combined effects of routine blood pressure lowering and intensive glucose control on macrovascular and microvascular outcomes in patients with type 2 diabetes: new results from the advance trial. *Diabetes Care.* (2009) 32:2068–74. doi: 10.2337/dc09-0959
40. Patel A, Group AC, MacMahon S, Chalmers J, Neal B, Woodward M, et al. Effects of a fixed combination of perindopril and indapamide on macrovascular and microvascular outcomes in patients with type 2 diabetes mellitus (the advance trial): a randomised controlled trial. *Lancet.* (2007) 370:829–40. doi: 10.1016/S0140-6736(07)61303-8
41. Khitan Z, Nath T, Santhanam P. Machine learning approach to predicting albuminuria in persons with type 2 diabetes: an analysis of the look ahead cohort. *J Clin Hypertens (Greenwich).* (2021) 23:2137–45. doi: 10.1111/jch.14397
42. De Miguel C, Lund H, Mattson DL. High dietary protein exacerbates hypertension and renal damage in dahl SS rats by increasing infiltrating immune cells in the kidney. *Hypertension.* (2011) 57:269–74. doi: 10.1161/HYPERTENSIONAHA.110.154302

43. Lee CM, Huxley RR, Wildman RP, Woodward M. Indices of abdominal obesity are better discriminators of cardiovascular risk factors than BMI: a meta-analysis. *J Clin Epidemiol.* (2008) 61:646–53. doi: 10.1016/j.jclinepi.2007.08.012
44. De Cosmo S, Menzaghi C, Prudente S, Trischitta V. Role of insulin resistance in kidney dysfunction: insights into the mechanism and epidemiological evidence. *Nephrol Dial Transplant.* (2013) 28:29–36. doi: 10.1093/ndt/gfs290
45. Aaron KJ, Campbell RC, Judd SE, Sanders PW, Muntner P. Association of dietary sodium and potassium intakes with albuminuria in normal-weight, overweight, and obese participants in the reasons for geographic and racial differences in stroke (regards) study. *Am J Clin Nutr.* (2011) 94:1071–8. doi: 10.3945/ajcn.111.013094
46. Thomas MC, Brownlee M, Susztak K, Sharma K, Jandeleit-Dahm KA, Zoungas S, et al. Diabetic kidney disease. *Nat Rev Dis Prim.* (2015) 1:15018. doi: 10.1038/nrdp.2015.18
47. Mortada I. Hyperuricemia, type 2 diabetes mellitus, and hypertension: an emerging association. *Curr Hypertens Rep.* (2017) 19:69. doi: 10.1007/s11906-017-0770-x
48. Wong MG, Perkovic V, Chalmers J, Woodward M, Li Q, Cooper ME, et al. Long-term benefits of intensive glucose control for preventing end-stage kidney disease: ADVANCE-ON. *Diabetes Care.* (2016) 39:694–700.

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.

frontiers | Frontiers in Public Health

SYSTEMATIC REVIEW
published: 14 July 2022
doi: 10.3389/fpubh.2022.942595

Reliability of Evidence to Guide Decision-Making in the Use of Acupuncture for Postpartum Depression

Xiuwu Hu[1†‡], Qian Fan[2‡], Li Ma[3], Rui Jin[3†], Rui Gong[3], Xiaoying Zhao[3†], Fenfen Qiu[1] and Liang Zhou[1*†]*

[1] Nanchang Hongdu Hospital of Traditional Chinese Medicine, Nanchang, China, [2] Department of Acupuncture, Changshu Hospital Affiliated to Nanjing University of Chinese Medicine, Changshu, China, [3] Tianjin University of Traditional Chinese Medicine, Tianjin, China

Background: There is conflicting evidence on the effectiveness of acupuncture in the treatment of postpartum depression (PPD). This study aimed to assess previous systematic reviews/meta-analyses (SRs/MAs) on the effectiveness of acupuncture to treat PPD.

Method: SRs/MAs regarding the use of acupuncture for PPD were identified from the establishment of digital databases to November 2021. The Assessing the Methodological Quality of Systematic Reviews 2 (AMSTAR-2) was applied to evaluate the methodological quality of included SRs/MAs. The Grades of Recommendations, Assessment, Development and Evaluation (GRADE) was utilized to evaluate the evidence quality for outcomes of interest.

Results: Six studies that conducted quantitative syntheses were included. According to AMSTAR-2, the methodological quality of these SRs/MAs was critically low owing to limitations of items 2, 4, and 7. According to GRADE, no study included high-quality evidence and most studies included low-quality evidence.

Conclusions: Acupuncture m be beneficial for PPD, however, due to limitations of current evidence and inconsistent findings, further studies are needed to provide stronger evidence to draw definitive conclusions.

Keywords: evidence, decision-making, acupuncture, postpartum depression, overview

OPEN ACCESS

Edited by:
Md. Mohaimenul Islam,
AESOP TECHNOLOGY, Taiwan

Reviewed by:
Jinke Huang,
China Academy of Chinese Medical Sciences, China
Min Shen,
Zhejiang Chinese Medical University, China

***Correspondence:**
Fenfen Qiu
qiufenfen2022@126.com
Liang Zhou
zhouliang0131@126.com

***ORCID:**
Xiuwu Hu
orcid.org/0000-0002-2532-6977
Rui Jin
orcid.org/0000-0001-6124-9609
Xiaoying Zhao
orcid.org/0000-0002-0486-2834
Liang Zhou
orcid.org/0000-0003-0981-3887

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

Specialty section:
This article was submitted to Family Medicine and Primary Care, a section of the journal Frontiers in Public Health

Received: *12 May 2022*
Accepted: *09 June 2022*
Published: *14 July 2022*

Citation:
Hu X, Fan Q, Ma L, Jin R, Gong R, Zhao X, Qiu F and Zhou L (2022) Reliability of Evidence to Guide Decision-Making in the Use of Acupuncture for Postpartum Depression. Front. Public Health 10:942595. doi: 10.3389/fpubh.2022.942595

INTRODUCTION

Postpartum depression (PPD) is a mood disorder associated with childbirth, since its onset begins between the first day and 4 months after delivery (1). Typically, PPD occurs within 6 weeks postpartum and patients tend to recover in 3–6 months, while severe cases can persist for up to 2 years. The prevalence of PPD in first-time mothers is as high as 16% (2), and the recurrence rate of PPD in the second pregnancy reaches 30% of women (3, 4). PPD is characterized by a depressed mood, loss of interest, sleep disturbances, psychomotor agitation or retardation, feelings of worthlessness, and even suicidal thoughts and behaviors in severe cases (5). Given the high prevalence and deleterious impact of PPD, the development of effective treatments is needed.

Treatment of PPD includes pharmacotherapy, psychotherapy, or both, which is consistent with the treatment recommended in guidelines for major depression (6). However, these treatments vary in efficacy (7–9), are cost (10), while adverse events are common (11, 12). Therefore, more effective and safer treatments for PPD are still needed. In this regard, acupuncture is perceived as an effective and safe alternative (13). A number of systematic reviews (SRs)/meta-analyses (MAs) have evaluated the efficacy of acupuncture for PPD, however their findings are inconsistent and the evidence credibility is unclear. Therefore, we provide a critical evaluation of SRs/MAs on the use of acupuncture to treat PPD.

METHODS

This study followed the methodology of the Cochrane Handbook and high-quality studies (14–16).

Eligibility Criteria

The following eligibility criteria were used to screen studies: (a) SRs/MAs based on randomized controlled trials (RCTs) on the use of acupuncture to treat PPD; (b) participants diagnosed with PPD by a recognized guideline; (c) interventions included acupuncture therapy or acupuncture plus conventional medication (CM), while the control group was treated with CM, CM plus acupuncture, sham acupuncture, or other non-pharmacological therapy; (d) outcomes included the Hamilton Rating Scale for Depression (HAMD), Edinburgh Postnatal Depression Scale (EPDS), effective rate and estradiol levels. Repeated publications or studies lacking complete data were removed.

Search Strategy

Embase, PubMed, Web of Science, Cochrane Library, CNKI, CBM, Wanfang, and VIP were searched for studies published between database creation and November 2021. The following search terms were applied: postpartum depression, acupuncture, meta-analysis, and systematic review. **Table 1** presents the search strategy for the PubMed database.

TABLE 1 | Search strategy for the PubMed database.

Query	Search term
# 1	Postpartum depression [Mesh]
# 2	Postpartum depression[Title/Abstract] OR postnatal depression[Title/Abstract] OR post-partum depression[Title/Abstract] OR post-natal depression[Title/Abstract] OR post natal depression[Title/Abstract]
# 3	#1 OR #2
# 4	Acupuncture[Mesh]
# 5	Acupuncture[Title/Abstract] OR pharmacoacupuncture[Title/Abstract] OR acupotomy[Title/Abstract] OR acupotomies[Title/Abstract] OR pharmacopuncture[Title/Abstract] OR needle[Title/Abstract] OR needling[Title/Abstract] OR dry-needling[Title/Abstract] OR body-acupuncture[Title/Abstract] OR electroacupuncture[Title/Abstract] OR electro-acupuncture[Title/Abstract] OR auricular acupuncture[Title/Abstract]
# 6	#4 OR #5
# 7	Meta-analysis as Topic[Mesh]
# 8	Systematic review[Title/Abstract] OR meta-Analysis[Title/Abstract] OR meta-analysis [Title/Abstract] OR meta-analyses[Title/Abstract] OR meta-analysis [Title/Abstract]
# 9	#7 OR #8
# 10	#3 AND #6 AND #9

Data Collection and Extraction

Two independent evaluators screened abstracts and titles, and then assessed potentially eligible full texts for final inclusion. Disagreements were resolved through discussion with a third independent reviewer. The following data were extracted from included studies: first author, year of publication, country, sample size, interventions, outcomes, quality assessment methods, and summary estimates of effect.

Quality Assessment

Two independent evaluators assessed the methodological quality of SR/MA using the Assessment of Methodological Quality of Systematic Evaluation 2 (AMSTAR-2) (17). AMSTAR-2 consists of 16 items, each with three possible answers, i.e., "yes," "partially yes," or "no." When up to one non-critical item does not meet the requirements, the methodological quality is considered "high"; when more than one non-critical item does not meet the requirements, the methodological quality is considered "medium"; when one critical item does not meet the requirements, the methodological quality is considered "low" and when more than one critical item do not meet the requirements, the methodological quality is deemed "very low" (17).

Two independent evaluators used the Grade of Recommendation, Assessment, Development and Evaluation (GRADE) (18) to assess the quality of evidence for each outcome indicator. GRADE ranks the evidence according to risk of bias, indirectness, imprecision, inconsistency, and publication bias. Each outcome measure is rated on four levels, i.e., "high," "moderate," "low," or "very low" (18).

Data Synthesis and Presentation

A narrative synthesis was used in this overview. The characteristics and results of each SR/MA as well as results from AMSTAR 2 were summarized by tabulation. The GRADE evidence profile and summary of findings table were generated using the GRADE pro GDT online software.

RESULTS

Study Selection

The literature search identified 114 articles, of which 40 duplicates were removed. Titles and abstracts of 74 articles were screened, and 63 articles were subsequently excluded. The full

Abbreviations: PPD, postpartum depression; SR, Systematic review; MA, Meta-analysis; AMSTAR-2, Assessing the Methodological Quality of Systematic Reviews 2; GRADE, Grading of Recommendations, Assessment, Development, and Evaluation; RCTs, Randomized clinical trials; CM, conventional medication; HAMD, Hamilton Depression Scale; EPDS, Edinburgh Postnatal Depression Scale.

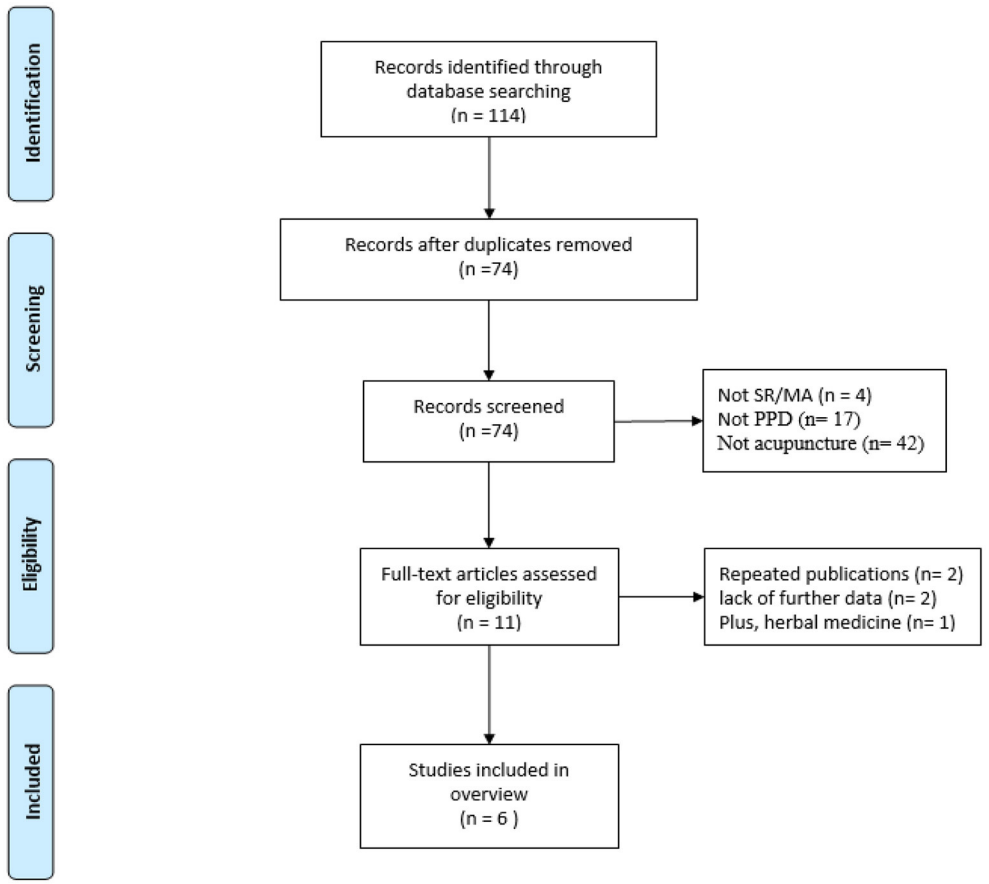

FIGURE 1 | Flow diagram of the literature selection.

text of the remaining 11 articles was read and five articles were excluded. Therefore, six papers were included in our analyses (19–25). The selection process is shown in **Figure 1**.

Study Characteristics

All included studies were conducted in China and published within the last 5 years. All studies were MAs with significant differences in sample size (5–14 studies, 27–934 participants). The interventions in the experimental group were acupuncture, or a combination of acupuncture and CM, while the control interventions were CM and/or sham acupuncture. Details on study characteristics are reported in **Table 2**.

Methodological Appraisal

All studies were rated as very low quality according to the AMSTAR-2 criteria. The key factors contributing to lower methodological quality were item 2 (only one review registered a protocol), item 4 (half of the studies did not provide a search strategy), and item 7 (all reviews did not provide a list of excluded studies). Detailed assessment results of AMSTAR-2 are shown in **Table 3**.

GRADE Evidence Quality Classification

A total of 19 outcome indicators were assessed. No indicator was deemed high, while two were moderate, 12 were low and five were very low quality of evidence. Risk of bias was the most common reason for downgrading the evidence, followed by inconsistency, imprecision, publication bias, and indirectness. Details are shown in **Table 4**.

Description of Efficacy

All studies used the HAMD to assess the severity of depression, and one review (20) concluded that acupuncture treatment improved depressive symptoms more significantly than CM, however, five reviews (19, 21–24) showed no significant difference between the two groups. Four reviews (19, 21–23) reported the EPDS of acupuncture vs. CM, in which three reviews showed that acupuncture was more effective than the control group (21–23) and one review showed no significant difference (19). The effective rate was reported in all reviews. Three of which revealed that acupuncture was more effective than the control group (19, 21, 23) while the other three reviews found no difference (20, 22, 24). Estradiol levels were reported in five reviews (19–23), in which three reviews found a significant effect

TABLE 2 | Characteristics of the included studies.

References	Country	Sample size	Treatment intervention	Control intervention	Quality assessment	Conclusion
Tong et al. (19)	China	12 (877)	AT; AT+CM	ST; CM	Cochrane criteria	Acupuncture has shown benefit in improving some symptoms of PPD, although the evidence is still inconclusive. High-quality studies are needed to confirm the effectiveness of acupuncture for PPD.
Li et al. (20)	China	8 (517)	AT	ST; CM	Cochrane criteria	Acupuncture treatment significantly improved HAMD scores, but had no significant effect on EPDS, clinical response, or serum estradiol levels.
Li et al. (21)	China	9 (653)	AT; AT+CM	ST; CM	Cochrane criteria	Acupuncture appears to be beneficial for PPD, however, the evidence is inconclusive. To confirm the effectiveness of acupuncture in PPD, further high-quality RCTs are needed.
Cao et al. (22)	China	13 (899)	AT	CM	Cochrane criteria	This study found no statistical difference between acupuncture and control groups in reducing HAMD scores and improving clinical effectiveness. Further studies are needed to validate these findings.
Wang et al. (23)	China	14 (934)	AT; AT+CM	ST; CM	Cochrane criteria	Acupuncture is effective in the treatment of PPD, but more high-quality and large sample size RCTs are needed to provide high-quality evidence.
Pang and Shi (24)	China	5 (279)	AT; AT+CM	ST; CM	Jadad	Acupuncture is as effective as CM and more effective than placebo to treat PPD. Acupuncture is safe and effective, although patients might experience fainting and pain during the procedure.

AT, acupuncture therapy; ST, sham acupuncture; CM, conventional medication.

TABLE 3 | AMSTAR-2 assessment results.

References	AMSTAR-2																Overall quality
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	
Tong et al. (19)	Y	PY	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	CL
Li et al. (20)	Y	PY	Y	PY	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	CL
Li et al. (21)	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	CL
Cao et al. (22)	Y	PY	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	CL
Wang et al. (23)	Y	PY	Y	PY	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	CL
Pang and Shi (24)	Y	PY	Y	PY	Y	Y	N	Y	Y	N	Y	Y	Y	Y	Y	N	CL

Y, Yes; PY, partial Yes; N, No; CL, Critically low; L, Low.

for acupuncture when compared to the control group (19, 21, 23) and one review found no difference (20, 22).

DISCUSSION

Acupuncture is routinely used in clinical therapy for PPD in China as a way to improve therapeutic effectiveness. Numerous SRs/MAs have evaluated the effectiveness of acupuncture for PPD, however, inconsistent results have been reported. In this context, a critical evaluation of different SRs/MAs and a summary of the scientific nature of the evidence is necessary (25). Furthermore, an overview can highlight deficiencies that need to be improved to guide future high-quality RCTs or SRs/MAs (26).

A total of six SR/MAs were included in this study, all of which were published in the past 5 years, suggesting that more researchers are beginning to study acupuncture as an alternative therapy for PPD. Nineteen outcome measures on the effectiveness of acupuncture to treat PPD were evaluated, and although most indicators suggested positive results, these

TABLE 4 | Certainty of evidences quality.

References	Outcomes	Simple	Limitations	Inconsistency	Indirectness	Imprecision	Publication bias	Quality
Tong et al. (19)	HAMD	MD −1.27 (−2.55, 0.01)	-1①	-1②	0	0	0	Low
	EPD	SMD −0.49 (−1.01, 0.02)	-1①	-1②	0	-1③	-1④	Very low
	Estradiol level	MD 63.99 (13.47, 114.51)	-1①	-1②	0	-1③	-1④	Very low
	Effect rate	RR 1.20 (1.09, 1.33)	-1①	0	0	0	0	Moderate
Li et al. (20)	HAMD	SMD −1.08 (−2.11, −0.05)	-1①	-1②	0	0	0	Low
	Estradiol levels	SMD 1.96 (−0.01, 3.93)	-1①	0	0	-1③	-1④	Very low
	Effect rate	RR 1.00 (0.89, 1.12)	-1①	-1②	0	0	0	Low
Li et al. (21)	HAMD	MD −1.38 (−3.40, 0.64)	-1①	-1②	0	0	0	Low
	EPDS	MD 1.08 (1.09, 3.26)	-1①	-1②	0	-1③	-1④	Very low
	Effective rate	RR 1.15 (1.06, 1.24)	-1①	0	0	-1③	0	Low
	Estradiol levels	MD 36.92 (23.14, 50.71)	-1①	-1②	0	0	0	Low
Cao et al. (22)	HAMD	MD 0.45 (−0.52,1.41)	-1①	-1②	0	0	0	Low
	EPDS	MD 0.55 (0.18, 0.92)	-1①	0	0	-1③	-1④	Very low
	Effective rate	RR 0.93 (0.70, 1.23)	-1①	-1②	0	0	0	Low
	Estradiol levels	MD 0.20 (−0.19, 0.58)	-1①	0	0	0	0	Moderate
Wang et al. (23)	HAMD	MD −1.27 (−2.55,0.01)	-1①	-1②	0	0	0	Low
	EPDS	MD −0.47 (−0.92, −0.03)	-1①	0	0	-1③	0	Low
	Estradiol levels	WMD 63.99 (13.39, 114.60)	-1①	-1②	0	0	0	Low
	Effective rate	OR 3.15 (2.19, 4.55)	-1①	0	0	-1③	0	Low
Pang and Shi (24)	HAMD	MD −1.03 (−2.58,0.52)	-1①	-1②	0	-1③	-1④	Very low
	Effective rate	RR 0.98 (0.84, 1.14)	-1①	0	0	-1③	-1④	Very low

RR, Risk Ratio; OR, odds ratio; SMD, SMD, standardized mean difference; WMD, Weighted Mean Difference; AT, acupuncture therapy; ST, sham acupuncture; CM, conventional medication; HAMD, Hamilton Rating Scale for Depression; EPDS, Edinburgh Postnatal Depression Scale. ①, The design of the experiment with a large bias in random, distributive hiding or blind; ②, The confidence interval overlaps less, the heterogeneity test P is very small, and the I2 is larger; ③, Confidence interval is not narrow enough; ④, Funnel graph asymmetry; ⑤, Fewer studies are included and there may be greater publication bias.

were inconsistent. Furthermore, although most of the included studies suggested that acupuncture was effective as a treatment for PPD, most authors did draw firm conclusions due to the low methodological quality of evidence or the small size of included trials. Indeed, all reviews were considered to be of very low quality according to AMSTAR-2 criteria. Therefore, our analysis concluded that acupuncture might be an effective treatment for PPD, but such conclusion must be treated with caution due to limitations of the current evidence.

Over recent years, AMSTAR-2 has become the most widely used tool to evaluate the methodological quality of SRs/MAs. All included studies had more than one critical flaw, so that there is very low confidence in their results. The key factors contributing to this setting were item 2 (only one review registered a protocol), item 4 (half of the studies did not provide a search strategy), and item 7 (all reviews did not provide a list of excluded studies). It was found that study protocols contribute to increased transparency of the methodology used and improve the overall methodological quality of SRs/MAs (27). The absence of a specific search strategy can result in an unreproducible search process, which leads to significant bias in included and excluded studies, undermining the scientific validity of findings. Likewise, by not presenting a list of excluded studies, authors can concur to incorrect exclusion of key literature, undermining the rigor of the report. Therefore, future SRs/MAs should address these identified deficiencies to develop high-quality studies and thus provide high-quality evidence.

In this study, authors of the included SRs/MAs did not draw definitive conclusions. Indeed, after rating the evidence using the GRADE system, we found that the certainty of evidence was unsatisfactory, indicating that findings of the included SRs/MAs are uncertain. Although all SRs/MAs evaluated only RCTs, the certainty of evidence was limited owing to the risk of bias (lack of blinding and allocation concealment), inconsistency, imprecision, or publication bias. The results of the methodological quality evaluation of RCTs showed that there is room for addressing random, distributed hidden or blind biases. Nevertheless, we must acknowledge that there are specificities of acupuncture therapy (inability to blind physicians and patients) that make the implementation of RCTs challenging. Improved standardization and precision of acupuncture techniques and procedures are urgently needed, as only a rigorously designed and implemented RCT can reduce the risk of bias and therefore assess the effectiveness of interventions (28).

To our knowledge, this is the first overview of SRs/MAs summarizing the current evidence on the use of acupuncture to treat PPD. The methodological and evidence qualities of the included SRs/MAs may help to inform evidence-based decision-making and guide future high-quality studies.

However, our study presents some limitations. First, the quality analysis demonstrated numerous methodological flaws in the performance of SRs/MAs, and the evidence quality was not satisfactory, making it impossible to draw firm conclusions about the use acupuncture for PPD. Second, the rapid growth in the number of SRs/MAs highlights challenges faced by healthcare decision-makers and researchers in keeping up with the evidence. This overview found that there were typically a large number of low-quality SRs/MAs. To help evidence-based practice, there is an urgent need for high-quality SRs/MAs that do not overlap and are up to date. Furthermore, widely used AMSTAR-2 tool and GRADE system are subjective evaluation tools, therefore the accuracy of assessments can vary. To mitigate this limitation, quality assessments were performed by two independent authors.

CONCLUSION

Acupuncture might be beneficial for PPD. However, due to limitations of the current evidence and inconsistent findings, further studies are needed to provide strong evidence to draw definitive conclusions.

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

XH and QF conceived the study and drafted the manuscript. LM, RJ, RG, and XZ help with the implementation of research. FQ and LZ provided guidance on the overview methodology. LM revised the manuscript. All authors read, critically reviewed, and approved the final manuscript as submitted.

FUNDING

This work was funded by Jiangxi Provincial Department of Science and Technology Key Research and Development Program General Project (No. 20192BBGL70037) and Changshu Science and Technology Development Plan Project (No. CS202136).

REFERENCES

1. Lee D, Yip A, Chiu H, Leung T, Chung T. A psychiatric epidemiological study of postpartum Chinese women. *Am J Psychiatry.* (2001) 158:220–6. doi: 10.1176/appi.ajp.158.2.220
2. Viguera AC, Tondo L, Koukopoulos AE, Reginaldi D, Lepri B, Baldessarini RJ. Episodes of mood disorders in 2,252 pregnancies and postpartum periods. *Am J Psychiatry.* (2011) 168:1179–85. doi: 10.1176/appi.ajp.2011.11010148
3. Payne JL, Maguire J. Pathophysiological mechanisms implicated in postpartum depression. *Front Neuroendocrinol.* (2019) 52:165–80. doi: 10.1016/j.yfrne.2018.12.001
4. Shorey S, Chee CYI, Ng ED, Chan YH, Tam WWS, Chong YS. Prevalence and incidence of postpartum depression among healthy mothers: a systematic review and meta-analysis. *J Psychiatr Res.* (2018) 104:235–48. doi: 10.1016/j.jpsychires.2018.08.001
5. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-V). 5th Edn.* Arlington, VA: American Psychiatric Association (2013).
6. Stein A, Pearson RM, Goodman SH, Rapa E, Rahman A, McCallum M, et al. Effects of perinatal mental disorders on the fetus and child. *Lancet.* (2014) 384:1800–19. doi: 10.1016/S0140-6736(14)61277-0
7. Cipriani A, Furukawa TA, Salanti G, Chaimani A, Atkinson LZ, Ogawa Y, et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Lancet.* (2018) 391:1357–66. doi: 10.1016/S0140-6736(17)32802-7
8. Cipriani A, Furukawa TA, Salanti G, Geddes JR, Higgins JP, Churchill R, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet.* (2009) 373:746–58. doi: 10.1016/S0140-6736(09)60046-5
9. Amick HR, Gartlehner G, Gaynes BN, Forneris C, Asher GN, Morgan LC, et al. Comparative benefits and harms of second generation antidepressants and cognitive behavioral therapies in initial treatment of major depressive disorder: systematic review and meta-analysis. *BMJ.* (2015) 351:h6019. doi: 10.1136/bmj.h6019
10. Karyotaki E, Tordrup D, Buntrock C, Bertollini R, Cuijpers P. Economic evidence for the clinical management of major depressive disorder: a systematic review and quality appraisal of economic evaluations alongside randomised controlled trials. *Epidemiol. Psychiatr. Sci.* (2017) 26:501–16. doi: 10.1017/S2045796016000421
11. Gartlehner G, Hansen RA, Morgan LC, Thaler K, Lux L, Van Noord M, et al. Comparative benefits and harms of second-generation antidepressants for treating major depressive disorder: an updated meta-analysis. *Ann. Intern Med.* (2011) 155:772–85. doi: 10.7326/0003-4819-155-11-201112060-00009
12. Nierenberg AA, Ostacher MJ, Huffman JC, Ametrano RM, Fava M, Perlis RH. A brief review of antidepressant efficacy, effectiveness, indications, and usage for major depressive disorder. *J Occup Environ Med.* (2008) 50:428–36. doi: 10.1097/JOM.0b013e31816b5034
13. Deligiannidis KM, Freeman MP. Complementary and alternative medicine for the treatment of depressive disorders in women. *Psychiatr Clin North Am.* (2010) 33:441–63. doi: 10.1016/j.psc.2010.01.002
14. Huang J, Liu J, Liu Z, Ma J, Ma J, Lv M, et al. Reliability of the evidence to guide decision-making in acupuncture for functional dyspepsia. *Front Public Health.* (2022) 10:842096. doi: 10.3389/fpubh.2022.842096
15. Hu C, Qin X, Ye R, Jiang M, Lu Y, Lin C. The role of traditional chinese medicine nursing for stroke: an umbrella review. *Evid Based Complement Alternat Med.* (2021) 2021:9918687. doi: 10.1155/2021/9918687
16. Huang J, Qin X, Shen M, Huang Y. An overview of systematic reviews and meta-analyses on acupuncture for post-stroke aphasia. *Eur J Integr Med.* (2020) 37:101133. doi: 10.1016/j.eujim.2020.101133
17. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ.* (2017) 358:j4008. doi: 10.1136/bmj.j4008
18. Pollock A, Farmer SE, Brady MC, Langhorne P, Mead GE, Mehrholz J, et al. An algorithm was developed to assign grade levels of evidence to comparisons within systematic reviews. *J Clin Epidemiol.* (2016) 70:106–10. doi: 10.1016/j.jclinepi.2015.08.013
19. Tong P, Dong LP, Yang Y, Shi YH, Sun T, Bo P. Traditional Chinese acupuncture and postpartum depression: a systematic review and meta-analysis. *J Chin Med Assoc.* (2019) 82:719–26. doi: 10.1097/JCMA.0000000000000140
20. Li W, Yin P, Lao L, Xu S. Effectiveness of acupuncture used for the management of postpartum depression: a systematic review and meta-analysis. *Biomed Res Int.* (2019) 2019:6597503. doi: 10.1155/2019/6597503

21. Li S, Zhong W, Peng W, Jiang G. Effectiveness of acupuncture in postpartum depression: a systematic review and meta-analysis. *Acupunct Med.* (2018) 36:295–301. doi: 10.1136/acupmed-2017-011530
22. Cao Y, Cao W, Yuan J, Li M, Li X, Yang K, Wen C. Efficacy and safety of acupuncture for postpartum depression: a systematic review. *Chinese J Evidence-Based Med*. (2021) 21:922–8. doi: 10.7507/1672-2531.202103078
23. Wang JF, Tan LJ, Mei QX, Zheng QH, Yang SB, Mei ZG. Meta-analysis on acupuncture for postpartum depression. World J Acupunct Moxibustion. (2017) 27:28–34. doi: 10.1016/S1003-5257(17)30096-X
24. Pang Y, Shi J. Clinical effect of acupuncture on postpartun depression: a meta-analysis. *J Liaoning Univ Tradit Chin Med.* (2016) 18:8–10. doi: 10.13194/j.issn.1673-842x.2016.07.002
25. Huang J, Shen M, Qin X, Wu M, Liang S, Huang Y. Acupuncture for the treatment of Alzheimer's disease: an overview of systematic reviews. *Front Aging Neurosci.* (2020) 12:574023. doi: 10.3389/fnagi.2020.574023
26. Hu C, Qin X, Jiang M, Tan M, Liu S, Lu Y, et al. Effects of tai chi exercise on balance function in stroke patients: an overview of systematic review. *Neural Plast.* (2022) 2022:3895514. doi: 10.1155/2022/3895514
27. Ge L, Tian JH, Li YN, Pan JX, Li G, Wei D, et al. Association between prospective registration and overall reporting and methodological quality of systematic reviews: a meta-epidemiological study. *J Clin Epidemiol.* (2018) 93:45–55. doi: 10.1016/j.jclinepi.2017.10.012
28. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomized trials. *BMJ.* (2010) 10:28–55. doi: 10.1016/j.ijsu.2011.10.001

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.

frontiers | Frontiers in Public Health

SYSTEMATIC REVIEW
published: 18 July 2022
doi: 10.3389/fpubh.2022.938113

Diagnostic Accuracy of Deep Learning and Radiomics in Lung Cancer Staging: A Systematic Review and Meta-Analysis

Xiushan Zheng[1†], Bo He[1†], Yunhai Hu[1†], Min Ren[1†], Zhiyuan Chen[1], Zhiguang Zhang[1], Jun Ma[1], Lanwei Ouyang[1], Hongmei Chu[1], Huan Gao[1], Wenjing He[2], Tianhu Liu[3] and Gang Li[3*]*

[1] Department of Thoracic Surgery, The 3rd Affiliated Hospital of Chengdu Medical College, Pidu District People's Hospital, Chengdu, China, [2] School of Electronic Engineering, Chengdu University of Technology, Chengdu, China, [3] Department of Cardiology, The 3rd Affiliated Hospital of Chengdu Medical College, Pidu District People's Hospital, Chengdu, China

OPEN ACCESS

Edited by:
Md. Mohaimenul Islam,
Aesop Technology, Taiwan

Reviewed by:
Hsuan-Chia Yang,
Taipei Medical University, Taiwan
Woon-Man Kung,
Chinese Culture University, Taiwan

***Correspondence:**
Tianhu Liu
lthzgl@163.com
Gang Li
ricon2001@126.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Family Medicine and Primary Care, a section of the journal Frontiers in Public Health

Received: 07 May 2022
Accepted: 15 June 2022
Published: 18 July 2022

Citation:
Zheng X, He B, Hu Y, Ren M, Chen Z, Zhang Z, Ma J, Ouyang L, Chu H, Gao H, He W, Liu T and Li G (2022) Diagnostic Accuracy of Deep Learning and Radiomics in Lung Cancer Staging: A Systematic Review and Meta-Analysis. Front. Public Health 10:938113. doi: 10.3389/fpubh.2022.938113

Background: Artificial intelligence has far surpassed previous related technologies in image recognition and is increasingly used in medical image analysis. We aimed to explore the diagnostic accuracy of the models based on deep learning or radiomics for lung cancer staging.

Methods: Studies were systematically reviewed using literature searches from PubMed, EMBASE, Web of Science, and Wanfang Database, according to PRISMA guidelines. Studies about the diagnostic accuracy of radiomics and deep learning, including the identifications of lung cancer, tumor types, malignant lung nodules and lymph node metastase, were included. After identifying the articles, the methodological quality was assessed using the QUADAS-2 checklist. We extracted the characteristic of each study; the sensitivity, specificity, and AUROC for lung cancer diagnosis were summarized for subgroup analysis.

Results: The systematic review identified 19 eligible studies, of which 14 used radiomics models and 5 used deep learning models. The pooled AUROC of 7 studies to determine whether patients had lung cancer was 0.83 (95% CI 0.78–0.88). The pooled AUROC of 9 studies to determine whether patients had NSCLC was 0.78 (95% CI 0.73–0.83). The pooled AUROC of the 6 studies that determined patients had malignant lung nodules was 0.79 (95% CI 0.77–0.82). The pooled AUROC of the other 6 studies that determined whether patients had lymph node metastases was 0.74 (95% CI 0.66–0.82).

Conclusion: The models based on deep learning or radiomics have the potential to improve diagnostic accuracy for lung cancer staging.

Systematic Review Registration: https://inplasy.com/inplasy-2022-3-0167/, identifier: INPLASY202230167.

Keywords: lung cancer, deep learning, radiomics, diagnostic accuracy, lymph node metastasis, meta-analysis

INTRODUCTION

Lung cancer is one of the most common malignancies globally and the leading cause of cancer-related death in the world. Its morbidity and cancer-related mortality rank first among malignant tumors. There are ∼2.2 million new cases and about 1.5 million deaths worldwide (1).

Radiomics and deep learning, as an innovative means to characterize lung lesions, can be applied to generate descriptive data, build predictive model, and correlate quantitative image features with phenotypes or gene-protein signatures, thus aiding in cancer detection, diagnosis, staging, treatment response prediction, and prognosis assessment and playing an increasingly important role in clinical decision-making, especially the management of malignant tumors (2).

Lung cancer staging is usually done by radiologists evaluating CT images of patients with lung cancer. The accuracy of diagnosis is affected by various factors, such as device performance, standardized imaging protocols, the experience of the reporting radiologist, and patient-specific factors. While radiomics involves using advanced computational algorithms to extract large numbers of researcher-defined features from images for defining related lung lesions, studies suggesting that deep learning algorithms can identify a more nuanced approach that eschews traditional radiology and statistical methods for cancer staging were extensively reported (3–6). Deep learning, as a new research direction in the field of machine learning (ML), is applied to learn the inherent laws and representation levels of sample data for feature recognition and model building (7). In the last decade, radiomics models and deep learning have made meaningful contributions to medical imaging diagnosis and related individual medicine (8).

This study aimed to perform a systematic review and meta-analysis of published data on lung cancer diagnosis and the diagnostic accuracy of deep learning algorithms and radiomics models for lung cancer staging.

METHODS

Search Strategy

This study followed the Preferred Reporting Item of the Guidelines for Systematic Reviews and Meta-Analysis (PRISMA), and selection criteria, data extraction, and data analysis were determined before study initiation. Any eligible studies in the PubMed, EMBASE, Web of Science, and Wanfang Database will be searched by Cancer, Radiomics, Deep Learning, Lung Cancer, and more. The search method is shown in **Table 1**. Search terms such as "radiomics," "deep learning," "lymph node metastasis," "non-small cell lung cancer," "malignant lung nodules," and "diagnostic accuracy." Use the Boolean operator AND to combine the results of different queries. We also manually searched the reference lists of included studies to identify any relevant articles. Both English and Chinese articles are considered eligible.

Abbreviations: CT, Computer tomography; MRI, Magnetic resonance imaging; AI, Artificial intelligence; ML, machine learning; LNM, lymph node metastasis; QUADAS-2, Quality assessment of diagnostic accuracy studies tool 2; AUROC, Area under the receiver operating characteristic curve; NSCLC, non-small cell lung cancer.

Study Selection

We selected publications for review if they met several of the following inclusion criteria: (1) patients with pathologically diagnosed lung cancer were included in the study; (2) radiomics or deep learning algorithms applied to lung cancer staging were evaluated. Exclusion criteria included: (1) informal publication types (e.g., reviews, letters to the editor, editorials, conference abstracts); (2) only focus on research on image segmentation or image feature extraction methods; (3) animal studies. After the removal of duplicates, titles and abstracts were identified by two independent reviewers using the Covidence systematic review software. Any disagreements will be resolved by consensus by arbitration by a third author.

Data Extraction

We reviewed data from selected primary studies using standardized forms, and two reviewers independently extracted data from each eligible study. Data extraction for each study included first author, country, year of publication, type of AI model, number of patients, patient characteristics (mean/median age, gender), type of malignancy, benign and malignant pulmonary nodules, lymph node metastasis. In addition, we extracted the area under the receiver operating characteristic curve (AUROC), along with sensitivity, specificity, accuracy, etc., for data processing and forest map production. The primary endpoint of this systematic review was AUROC.

Quality Assessment

Two independent reviewers will initially assess the risk of bias. A third reviewer will then review each study using the Quality Assessment of Studies for Diagnostic Accuracy (QUADAS-2) guidelines. The QUADAS-2 tool can assign a risk of bias rating of "low," "high," or "uncertain" based on the answer to "yes," "no," or "uncertain" to the relevant flag questions included in each section. For example, if the answer to all the landmark questions in a range is "yes," then it can be rated as low risk of bias; if all the informational questions are answered "no," then the risk of bias is rated as "high" (9). We summarized the risk of bias in individual studies in a narrative summary during the systematic review phase.

Statistical Analysis

The accuracy measures for this diagnostic meta-analysis included pooled sensitivity, pooled specificity, and their 95% confidence intervals (95% CI). Missing data is calculated

TABLE 1 | Search strategy.

Sources	Search in	MeSH terms	Limits	Search results
Web of science	Search manager	("deep learning" OR "convolutional neural network" OR "machine learning" OR "radiomics" OR "radiomic") AND ("CT" OR "MRI") AND ("Lymph node" OR "lymph node metastasis" OR "Benign and malignant pulmonary nodules")AND ("lung cancer" OR "non-small cell lung cancer" OR "NSCLC")	None	11
PubMed, (MEDLINE)	N/A	("deep learning" OR "convolutional neural network" OR "machine learning" OR "radiomics" OR "radiomic") AND ("CT" OR "MRI") AND ("Lymph node" OR "lymph node metastasis" OR "benign and malignant pulmonary nodules") AND ("lung cancer" OR "non-small cell lung cancer" OR "NSCLC")	None	30
EMBASE	Quick search	('deep learning'/exp OR "deep learning" OR "machine learning"/exp OR "machine learning" OR "radiomics"/exp OR "radiomics" OR "radiomic") AND ("ct"/exp OR "ct" OR "mri"/exp OR "mri") AND ("lymph node"/exp OR "lymph node" OR "lymph node metastasis"/exp OR "lymph node metastasis" OR "benign and malignant pulmonary nodules") AND ("lung cancer"/exp OR "non-small cell lung cancer" OR "NSCLC")	None	56
Wanfang database	N/A	("deep learning" OR "machine learning" OR "radiomics" OR "radiomic") AND ("CT" OR "MRI") AND ("Lymph node" OR "lymph node metastasis") AND ("lung cancer" OR "NSCLC")	None	5

TABLE 2 | Formulas.

Measure	Formula
Sensitivity	$\frac{TP}{P} = \frac{TP}{TP + FN}$
Specificity	$\frac{TN}{N} = \frac{TN}{TN + FP}$
Accuracy	$\frac{TP + TN}{P+N} = \frac{TP+TN}{TP + TN + FP + FN}$
PPV	$\frac{TP}{TP + FP}$
NPV	$\frac{TN}{TN + FN}$
SE	$\frac{(Upper\ Limit - Lower\ Limit)}{3.92}$
95% Confidence interval	$best\ estimate +/- (1.96) * (SE)$

P, condition positive; N, condition negative; FN, false negative; FP, false positive; TN, true negative and TP, true positive; PPV, positive predictive value; NPV, negative predictive value; Upper limit, upper limit of confidence interval; Lower limit, lower limit of confidence interval; SE, standard error.

using the formula in **Table 2**. At the same time, AUROC was calculated; an AUROC value close to 1.0 indicates that the test can discriminate almost perfectly, while an AUROC value close to 0.5 means poor discrimination (10, 11). The discordance index (I^2) was used (12). Heterogeneity was assessed as low, medium, and high, with upper limits for I^2 of 25, 50, and 75%, respectively. A forest plot was drawn to show the AUROC estimates relative to the summary pooled estimates for each study. In addition, we will draw a funnel plot to assess publication bias more intuitively. All statistical analyses were performed using STATA V16.0 software.

RESULTS

Study Selection

Our search identified 74 studies, with 56 screened after removing duplicates. Of these, 27 did not meet the inclusion criteria based on title and abstract. The remaining 29 full manuscripts were individually assessed, and, finally, 22 studies were eligible and included in our systematic review. Of these, 19 papers were available for meta-analysis, and five articles were excluded because of their insufficient data information. We outline the study selection process for review using the PRISMA flowchart (**Figure 1**).

Study Characteristics

Of the 19 included studies, 14 had sufficient data for a meta-analysis of AUROC (**Figure 2**). Regarding study design, 17 studies were retrospective, and two were prospective. Sixteen studies were single-center, and the other three were multicenter. Most of the patients are male, and the median age of 63 years (24–93 years) [**Table 3** (13–31)]. The malignancy type in twelve studies was NSCLC, and the malignancy type in the remaining studies was lung cancer. Seven studies used the diagnostic output per patient, and eight studies used the lymph node diagnostic output per node for metastases. While seven studies used post-operative pathology reports as reference standards, 11 used radiology reports.

Quality Assessment

According to the QUADAS-2 tool, the summary of this study's assessment is shown in **Figure 3**. The risk of bias in patient selection was low in 12 (74%) studies and high in 5 (26%) studies. The risk of bias for the index test was high in 2 studies (10%) and low in 17 studies (90%). The risk of bias for the reference standard test was low in 16 studies (85%), high in 2 studies (10%), and unclear in 1 study (5%). Process and timing made the risk of bias unclear for all 19 studies. **Table 4** shown individual evaluation of the risk of bias and applicability. Overall suitability issues are low. To assess the publication bias of the studies, a funnel plot was constructed (**Figure 4**). The shape of the funnel plot revealed asymmetry in the included studies, showing study heterogeneity.

Diagnostic Accuracy

Of the 19 studies eligible for quantitative analysis, 14 used radiomics and 5 used deep learning. For each outcome, on a per-patient basis, pooled estimates including specificity,

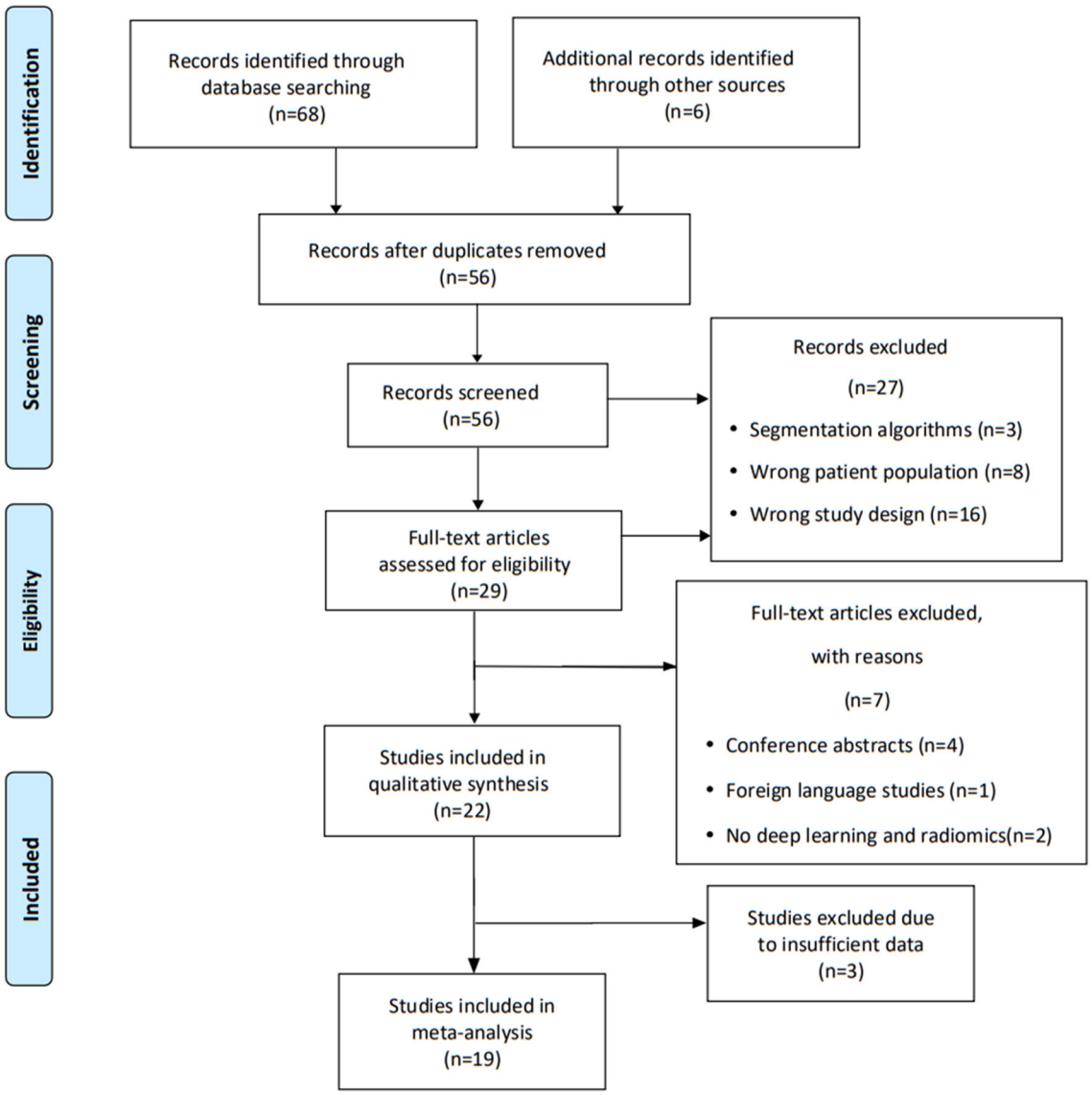

FIGURE 1 | PRISMA flow chart outlining the selection of studies for review.

sensitivity, and AUROC were generated with 95% confidence intervals. The categorized data extraction for each study report is shown in **Table 5**. The type of lung cancer, malignant lung nodules, lymph node metastases, and deep learning or radiomics models discussed in each study were considered.

The data from radiomics models showed high heterogeneity, except for AUROC and the sensitivity of each node. After removing the literature with insufficient data, the pooled AUROC of the 7 studies determining whether a patient had lung cancer was 0.83 (95% CI 0.78–0.88; **Figure 2A**), and the pooled sensitivity and specificity were 0.838 and 0.653, respectively, indicating high heterogeneity (I^2 = 65.3%, p = 0.008). For the 9 NSCLC studies that currently represent ∼85% of lung cancer, the pooled AUROC of radiomics was 0.78 (95% CI 0.73–0.83; **Figure 2B**), and the pooled sensitivity and specificity were 0.782 and 0.715, respectively, with higher heterogeneity (I^2 = 66.1%, p = 0.003). Among the six studies predicting benign or malignant pulmonary nodules, the pooled AUROC of radiomics was 0.79 (95% CI 0.77–0.82; **Figure 2C**), and the pooled sensitivity and specificity were 0.787 and 0.774, respectively, with heterogeneity relatively low (I^2 = 9.7%, p = 0.354). Among the 6 studies that predicted the accuracy of LNM in lung cancer patients, the pooled AUROC

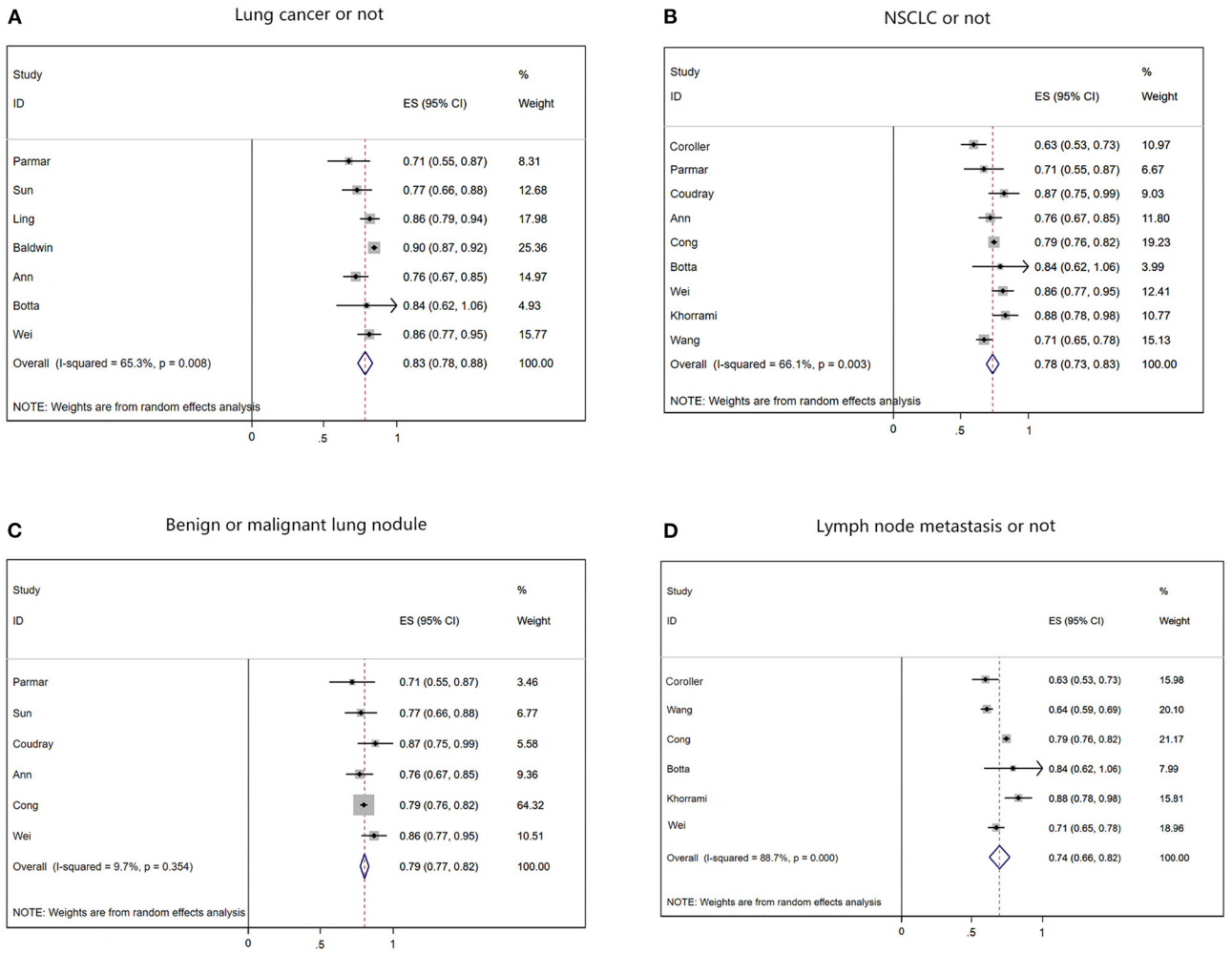

FIGURE 2 | Summary of forest plots for different classifications. **(A)** The forest plot of determine if a patient has lung cancer. **(B)** The forest plot of determining whether the cancer type is NSCLC. **(C)** The forest plot of predicting benign and malignant pulmonary nodules. **(D)** The forest plot of predicting lymph node metastasis in lung cancer.

of radiomics was 0.74 (95% CI 0.66–0.82; **Figure 2D**), and the pooled sensitivity and specificity were 0.661 and 0.598, respectively, with heterogeneity relatively high ($I^2 = 88.7\%$, $p = 0.000$).

DISCUSSION

During the diagnosis and treatment of lung cancer, many imaging data, such as CT, MRI, and PET, are generated. Doctors usually subjectively evaluate these data based on experience and make treatment plans (32). However, the features that doctors can observe from the image data with the naked eye are limited, and the potential of the image data is often not fully realized. Over the years, many researchers have tried to use complex mathematical and statistical algorithms to extract quantitative information that is hard to observe, even predicting cancer progression (33–35).

With the development of artificial intelligence technology, radiomics has emerged as the times require, using machine learning algorithms to mine high-throughput features from medical images and conduct modeling analysis. Increasing evidence shows that radiomics can be used for quantitative characterization of tumors for tasks such as disease diagnosis, treatment planning, and prognosis, which constitutes an important research direction for artificial intelligence technology in medical applications (36, 37). Radiomics is an emerging and rapidly developing field that integrates knowledge from radiology, oncology, and computer science and is an interdisciplinary subject that emphasizes the integration of medicine and engineering (38). With the rise of deep learning technology in recent

TABLE 3 | Selected characteristics of included studies.

References	Country	Year	Study design	Patients (% female patients)	Sample size for diagnostic accuracy	Mean or median age (SD; range), years	Imaging modality	Type of malignancy	AI model (Per-patient/per-node diagnostic output)	Reference standard	Classification criteria
Coroller et al. (13)	USA	2016	Retrospective single-center	85 (65%)	–	60.3	CT	NSCLC	Radiomics (per-patient)	Radiology	B D
Parmar et al. (14)	USA	2018	Retrospective single-center	1,194	–	68.3 (32–93)	CT	NSCLC	Deep learning (per-patient)	Pathology	A B C
Sun et al. (15)	China	2019	Retrospective single-center	385 (68%)	201	53.1 (±12.2)	CT	Lung Cancer	Radiomics (per-patient)	Radiology	A C
Ling et al. (16)	China	2019	Retrospective multi-center	229 (31.5%)	74	64 (59–81)	CT	Lung Cancer	Radiomics (per-patient)	Radiology	A
Coudray et al. (17)	USA	2018	Retrospective single-center	1,176	459	61 (51.3–72.8)	CT	NSCLC	Deep learning (per-patient)	Radiology	B C
Xu et al. (18)	China	2019	Retrospective single-center	179 (52.8%)	–	63 (32–93)	CT	NSCLC	Deep learning (per-patient)	Pathology	B D
Baldwin et al. (19)	UK	2020	Retrospective single-center	1,337	328	–	CT	Lung Cancer	Deep learning (per-patient)	–	A
Schroers et al. (20)	Germany	2019	Retrospective single-center	82 (38%)	50	61.5 (±5.0)	MRI	Lung Cancer	Radiomics (per-patient)	Pathology	A C
Wang et al. (21)	China	2019	Retrospective single-center	249 (39.8%)	–	61.4 (±8.96)	CT	Lung Cancer	Deep learning (per-patient)	Radiology	D
Leleu et al. (22)	France	2020	Retrospective single-center	215 (39%)	72	58.6 (±10.3)	CT	Lung Cancer	Radiomics (per-patient)	Pathology	A
Ann et al. (23)	USA	2019	Prospective multi-center	262	48	–	CT	NSCLC	Radiomics (per-patient)	Pathology	A B C
Cong et al. (24)	China	2020	Retrospective single-center	411 (50.4%)	141	59.62 (24–84)	CT	NSCLC	Radiomics (per-patient)	Radiology	B C D
Botta et al. (25)	Italy	2020	Retrospective single-center	270 (38%)	–	67.4 (61.0–72.6)	CT	NSCLC	Radiomics (per-patient)	Radiology	A B D
Wei et al. (26)	USA	2020	Retrospective multi-center	146 (39.7%)	–	65.72 (± 12.88)	PET/CT	NSCLC	Radiomics (per-node)	Radiology	A B C
Khorrami et al. (27)	USA	2019	Retrospective single-center	112	–	–	CT	NSCLC	Radiomics (per-patient)	Pathology	B D
Kirienko et al. (28)	Italy	2021	Retrospective single-center	149 (37.6%)	73	70 (41–84)	PET/CT	Lung Cancer	Radiomics (per-node)	Radiology	B C
Rossi et al. (29)	Italy	2020	Retrospective single-center	109	–	–	CT	NSCLC	Radiomics (per-patient)	Radiology	A B
Chai et al. (30)	China	2021	Retrospective single-center	198 (54%)	402	58.1 (± 8.5)	CT	NSCLC	Radiomics (per-node)	Pathology	A B D
Wang et al. (31)	China	2019	Retrospective single-center	717	386	—	CT	NSCLC	Radiomics (per-node)	Radiology	B D

A, Determine whether the patient has lung cancer; B, Determine whether the patient has non-small cell lung cancer; C, Determine whether the patient has malignant lung nodule; D, Determine whether the patient has lymph node metastasis.

years, the need for high precision and high stability in lung cancer staging has become more and more urgent (39).

To our knowledge, this is the first meta-analysis to summarize the diagnostic accuracy of deep learning and radiomics involving in lung cancer staging. We provided summarized data in this field and compared the identification effectiveness of lung cancer, tumor types, malignant lung nodules and lymph node metastase. In this article, the included studies mainly used radiomics ($n = 14$) rather than deep learning methods ($n = 5$). Of the five deep learning models, two were developed using transfer learning and three were developed using convolutional neural networks (CNN). Part of the reason there are relatively few deep learning models is that deep learning techniques are relatively new and prone to bias. The difference in the number of studies of the two AI models will lead to a significant deviation in the data ratio, affecting the ability comparison of the two models. Furthermore, most studies

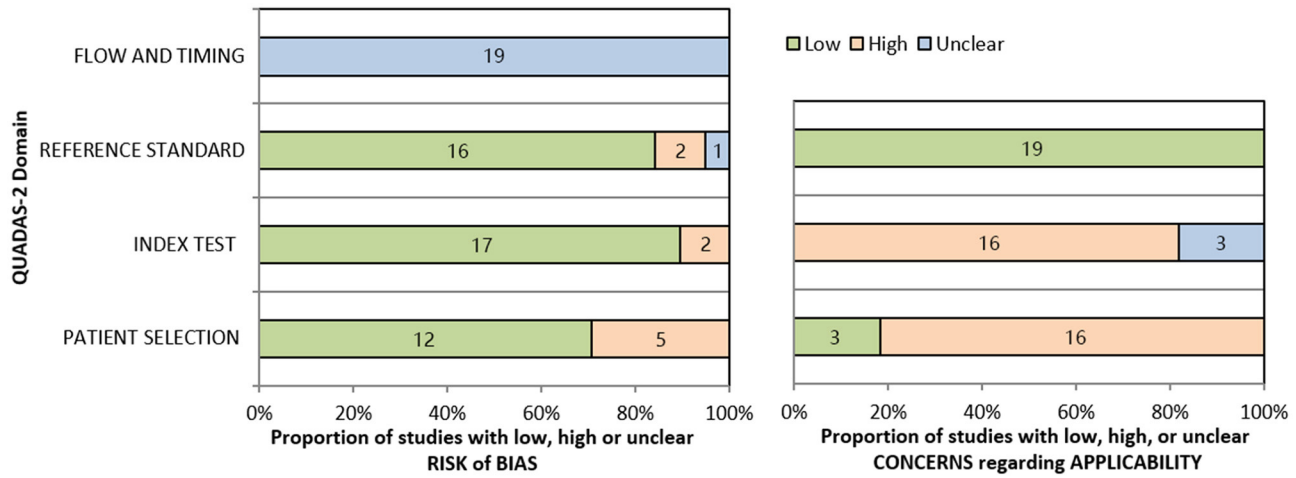

FIGURE 3 | Summary of QUADAS-2 assessments of included studies.

TABLE 4 | Quality assessment.

Source	Risk of bias								Applicability concerns		
	Patient selection			Index test			Reference standard	Flow and timing	Patient selection	Index test	Reference standard
	Was the statistical management adequate?	Were the inclusion/ exclusion criteria specified?	Was the type of study (retrospective or prospective) specified?	Were the imaging acquisition protocol and the segmentation method(s) detailed?	Was the image processing approach detailed?	Was the validation independent (i.e., no internal)?	Was the reference standard adequate?	Was there an appropriate interval between index test and reference standard?			
Chetan et al. (1)	Yes	Yes	Yes	Yes	Yes	No	Yes	Unclear	Yes	Yes	Unclear
Parmar et al. (2)	Yes	Yes	Yes	Yes	Yes	No	Yes	Unclear	Yes	Yes	Yes
Sun et al. (3)	Yes	Yes	Yes	Yes	Yes	No	Unclear	Unclear	Yes	Yes	Yes
Ling et al. (4)	Yes	Yes	Yes	Yes	Yes	No	Yes	Unclear	Yes	Yes	Yes
Coudray et al. (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	Unclear
Xu et al. (6)	Yes	No	Yes	Yes	Yes	No	Unclear	Unclear	Yes	Yes	Yes
Baldwin et al. (7)	Yes	Yes	Yes	Yes	Yes	No	Yes	Unclear	Yes	Yes	Yes
Schroers et al. (8)	Yes	Yes	Yes	Yes	Yes	No	Yes	Unclear	Yes	Yes	Yes
Wang et al. (9)	Yes	No	Yes	Yes	No	No	Unclear	Unclear	Yes	Yes	Unclear
Leleu et al. (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	Yes
Ann et al. (11)	Yes	Yes	Yes	Yes	Yes	Yes	Unclear	Unclear	Yes	Yes	Unclear
Cong et al. (12)	Yes	Yes	Yes	Yes	Yes	Yes	No	Unclear	Yes	Yes	Yes
Botta et al. (13)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	Unclear
Botta et al. (13)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	Unclear
Wei et al. (14)	Yes	Yes	Yes	Yes	Yes	No	Yes	Unclear	Yes	Yes	Yes
Khorrami et al. (15)	Yes	Yes	Yes	Yes	Yes	No	Unclear	Unclear	Yes	Yes	Yes
Kirienko et al. (16)	Yes	Yes	Yes	Yes	Yes	No	Unclear	Unclear	Yes	Yes	Unclear
Rossi et al. (17)	Yes	Yes	Yes	Yes	Yes	No	Yes	Unclear	Yes	Yes	Unclear
Chai et al. (18)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	Yes
Wang et al. (19)	Yes	Yes	Yes	Yes	Yes	No	Unclear	Unclear	Yes	Yes	Unclear

are retrospective in design, there are few prospective deep learning studies in lung cancer medical imaging staging, and most studies lack data and code availability. At the same time, most studies are single-center and use internal validation or resampling methods (cross-validation). However, internal validation tends to overestimate AUROC due to the lack of

generality of the models, limiting the integration of AI models into clinical settings (40). Therefore, predictive models validated externally by using images from different hospitals are needed to create reliable estimates of the performance levels of other sites (41).

This systematic review performed a statistical assessment of pooled data collected from 19 studies. However, our findings must take into account some limitations. First, while comprehensive, our search may have missed some studies that could have been included. Second, we calculated estimates of diagnostic performance using limited data as several studies reported incomplete data. Third, there may be geographic bias because the included studies were from geographically different quantitative distributions. Finally, the type of scanner used for diagnosis, the imaging protocol, and the criteria for lung cancer staging may affect the accuracy of the results. In the future, the clinical benefit of diagnostic lung cancer staging models must be rigorously evaluated against current diagnostic criteria, as not all models are applicable in clinical practice (42, 43). Under the current hot spot of artificial intelligence development, more and more deep learning studies have shown that deep learning big data extracted from patients' medical images can have good clinical application value in tumor staging of patients. Therefore, we can combine deep learning features to establish a radiomics combined with deep learning diagnostic model, so that the accuracy of lung cancer staging diagnosis of patients can be improved.

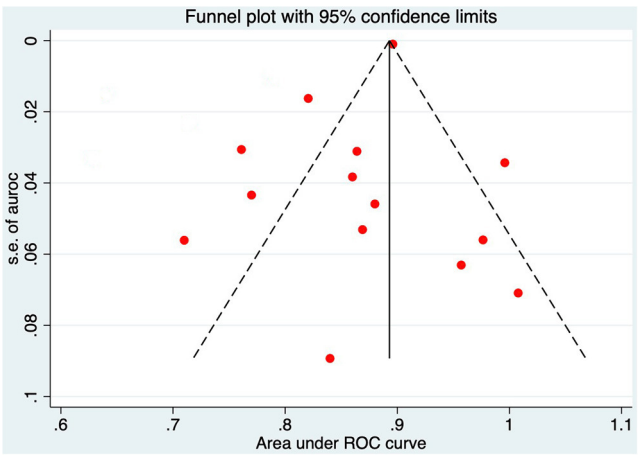

FIGURE 4 | Funnel plot of the area under the receiver operating characteristic in 14 studies.

CONCLUSION

The models based on deep learning or radiomics have the potential to improve diagnostic accuracy in the pathological staging of lung cancer with the purpose of providing individualized preoperative non-invasive auxiliary prediction means for clinicians and realizing valuable prediction for patients to obtain better treatment strategy. Future studies are welcomed to use standardized radiomics features, more robust tools of feature selection and model development to further improve the diagnostic accuracy of artificial intelligence in lung cancer staging.

TABLE 5 | Summary of AUROC for each study.

References	Sensitivity, %	Specificity, %	Accuracy, %	AUROC	95%CI	Standard error
Coroller et al. (13)	–	–	–	0.630	0.583–0.713	0.0331
Parmar et al. (14)	82.4	73.1	83.5	0.710	0.60–0.82	0.0561
Sun et al. (15)	–	–	–	0.770	0.69–0.86	0.0434
Ling et al. (16)	–	–	–	0.864	0.782–0.904	0.0311
Coudray et al. (17)	89.0	93.0	83.3	0.869	0.753–0.961	0.0531
Xu et al. (18)	–	–	63.5	0.670	–	–
Baldwin et al. (19)	99.57	28.03	40.01	0.896	0.876–0.915	0.0010
Schroers et al. (20)	86.95	93.25	88.89	–	–	–
Wang et al. (21)	64.04	58.97	61.47	0.640	0.61–0.67	0.0153
Leleu et al. (22)	–	–	72.6	–	–	–
Ann et al. (23)	79.9	75.2	65.8	0.761	0.59–0.71	0.0306
Cong et al. (24)	72.97	63.33	55.22	0.790	0.77–0.81	0.0102
Botta et al. (25)	–	–	–	0.840	0.63–0.98	0.0893
Wei et al. (26)	54.16	55.56	63.64	0.860	0.79–0.94	0.0383
Khorrami et al. (27)	61.34	57.16	63.81	0.880	0.79–0.97	0.0459
Kirienko et al. (28)	85.7	88.2	93.3	–	–	–
Rossi et al. (29)	100.0	66.7	85.7	0.850	–	–
Chai et al. (30)	–	–	95.3	–	–	–
Wang et al. (31)	–	–	72.4	0.712	0.678–0.770	0.0235

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

XZ and WH conceptualized the study. BH, YH, and MR collected the data. XZ, BH, and WH drafted the initial manuscript. ZC, ZZ, JM, LO, HC, and HG reviewed the included articles. YH and WH conducted the analyses. XZ, TL, and GL reviewed and revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Chengdu Science and Technology Program, Grant no. 2021007 and Sichuan Science and technology plan, Grant no. 2018JY0356.

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. (2018) 68:394–424. doi: 10.3322/caac.21492
2. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer*. (2018) 18:500–10. doi: 10.1038/s41568-018-0016-5
3. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, 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
4. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. *Magn Reson Imaging*. (2012) 30:1234–48. doi: 10.1016/j.mri.2012.06.010
5. Huang Y, Liu Z, He L, Chen X, Pan D, Ma Z, et al. Radiomics signature: a potential biomarker for the prediction of disease-free survival in early-stage (I or II) non-small cell lung cancer. *Radiology*. (2016) 281:947–57. doi: 10.1148/radiol.2016152234
6. Huang YQ, Liang CH, He L, Tian J, Liang CS, Chen X, et al. Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer. *J Clin Oncol*. (2016) 34:2157–64. doi: 10.1200/JCO.2015.65.9128
7. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. *Z Med Phys*. (2019) 29:102–27. doi: 10.1016/j.zemedi.2018.11.002
8. Benjamens S, Dhunnoo P, Mesko B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med*. (2020) 3:118. doi: 10.1038/s41746-020-00324-0
9. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med*. (2012) 9:1–12. doi: 10.1371/journal.pmed.1001221
10. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst*. (1959) 22:719–48.
11. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. (1986) 7:177–88. doi: 10.1016/0197-2456(86)90046-2
12. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. (2003) 327:557–60. doi: 10.1136/bmj.327.7414.557
13. Coroller TP, Agrawal V, Narayan V, Hou Y, Grossmann P, Lee SW, et al. Radiomic phenotype features predict pathological response in non-small cell lung cancer. *Radiother Oncol*. (2016) 119:480–6 doi: 10.1016/j.radonc.2016.04.004
14. Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, et al. Deep learning for lung cancer prognostication: a retrospective multi-cohort radiomics study. *PLoS Med*. (2018) 15:e1002711. doi: 10.1371/journal.pmed.1002711
15. Sun Y, Li C, Jin L, Gao P, Zhao W, Ma W, et al. Radiomics for lung adenocarcinoma manifesting as pure ground-glass nodules: invasive prediction. *Eur Radiol*. (2020) 30:3650–9. doi: 10.1007/s00330-020-06776-y
16. E L, Lu L, Li L, Yang H, Schwartz LH, Zhao B. Radiomics for classifying histological subtypes of lung cancer based on multiphasic contrast-enhanced computed tomography. *J Comput Assist Tomogr*. (2019) 43:300–6. doi: 10.1097/RCT.0000000000000836
17. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med*. (2018) 24:1559–67. doi: 10.1038/s41591-018-0177-5
18. Xu Y, Hosny A, Zeleznik R, Parmar C, Coroller T, Franco I, et al. Deep learning predicts lung cancer treatment response from serial medical imaging. *Clin Cancer Res*. (2019) 25:3266–75. doi: 10.1158/1078-0432.CCR-18-2495
19. Baldwin DR, Gustafson J, Pickup L, Arteta C, Novotny P, Declerck J, et al. External validation of a convolutional neural network artificial intelligence tool to predict malignancy in pulmonary nodules. *Thorax*. (2020) 75:306–12. doi: 10.1136/thoraxjnl-2019-214104
20. Meier-Schroers M, Homsi R, Schild HH, Thomas D. Lung cancer screening with MRI: characterization of nodules with different non-enhanced MRI sequences. *Acta Radiol*. (2019) 60:168–76. doi: 10.1177/0284185118778870
21. Wang S, Shi J, Ye Z, Dong D, Yu D, Zhou M, et al. Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. *Eur Respir J*. (2019) 53:1800986. doi: 10.1183/13993003.00986-2018
22. Leleu O, Basille D, Auquier M, Clarot C, Hoguet E, Pétigny V, et al. Lung cancer screening by low-dose CT scan: baseline results of a french prospective study. *Clin Lung Cancer*. (2020) 21:145–52. doi: 10.1016/j.cllc.2019.10.014
23. Trebeschi S, Drago SG, Birkbak NJ, Kurilova I, Călin AM, Delli Pizzi A, et al. Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. *Ann Oncol*. (2019) 30:998–1004. doi: 10.1093/annonc/mdz108
24. Cong M, Yao H, Liu H, Huang L, Shi G. Development and evaluation of a venous computed tomography radiomics model to predict lymph node metastasis from non-small cell lung cancer. *Medicine*. (2020) 99:e20074. doi: 10.1097/MD.0000000000020074
25. Botta F, Raimondi S, Rinaldi L, Bellerba F, Corso F, Bagnardi V, et al. Association of a CT-based clinical and radiomics score of non-small cell lung cancer (NSCLC) with lymph node status and overall survival. *Cancers*. (2020) 12:1432. doi: 10.3390/cancers12061432
26. Mu W, Tunali I, Gray JE, Qi J, Schabath MB, Gillies RJ. Radiomics of 18F-FDG PET/CT images predicts clinical benefit of advanced NSCLC patients to checkpoint blockade immunotherapy. *Eur J Nucl Med Mol Imaging*. (2020) 47:1168–82. doi: 10.1007/s00259-019-04625-9
27. Khorrami M, Prasanna P, Gupta A, Patil P, Velu PD, Thawani R, et al. Changes in CT radiomic features associated with lymphocyte distribution predict overall survival and response to immunotherapy in non-small cell lung cancer. *Cancer Immunol Res*. (2020) 8:108–19. doi: 10.1158/2326-6066.CIR-19-0476
28. Kirienko M, Sollini M, Corbetta M, Voulaz E, Gozzi N, Interlenghi M, et al. Radiomics and gene expression profile to characterise the disease and predict outcome in patients with lung cancer. *Eur J Nucl Med Mol Imaging*. (2021) 48:3643–55. doi: 10.1007/s00259-021-05371-7
29. Rossi G, Barabino E, Fedeli A, Ficarra G, Coco S, Russo A, et al. Radiomic detection of EGFR mutations in NSCLC. *Cancer Res*. (2021) 81:724–31. doi: 10.1158/0008-5472.CAN-20-0999
30. Yating C, Shu L, Xingyuan J, Fan W, Ye L. Combined model of radiomics features and clinical labels of peritumoral tissue to predict lymph node

metastasis in T1 non-small cell lung cancer. *Chin J Clin Med Imaging*. (2021) 32:470–5. doi: 10.12117/jccmi.2021.07.004
31. Chao W, Xia L, Di D, Liya Z, Zaiyi L, Changhong L, et al. Prediction of lymph node metastasis in non-small cell lung cancer based on radiomics. *Chin J Autom*. (2019) 45:1087–93. doi: 10.16383/j.aas.c160794
32. Wen PY, Macdonald DR, Reardon DA, Cloughesy TF, Sorensen AG, Galanis E, et al. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *J Clin Oncol*. (2010) 28:1963–72. doi: 10.1200/JCO.2009.26.3541
33. Chaunzwa TL, Christiani DC, Lanuti M, Shafer A, Aerts H. Using deep-learningradiomics to predict lung cancer histology. *J Clin Oncol*. (2018) 36:8545. doi: 10.1200/JCO.2018.36.15_suppl.8545
34. Chalkidou A, O'Doherty MJ, Marsden PK. False discovery rates in PET and CT studies with texture features: a systematic review. *PLoS ONE*. (2015) 10:e0124165. doi: 10.1371/journal.pone.0124165
35. Forghani R, Savadjiev P, Chatterjee A, Muthukrishnan N, Reinhold C, Forghani B. Radiomics and artificialintelligence for biomarker and prediction model development in oncology. *Comput Struct Biotechnol J*. (2019) 17:995–1008. doi: 10.1016/j.csbj.2019.07.001
36. Zacharaki EI, Wang S, Chawla S, Soo Yoo D, Wolf R, Melhem ER, et al. Classification of brain tumor type andgrade using MRI texture and shape in a machine learning scheme. *Magn Reson Med*. (2009) 62:1609–18. doi: 10.1002/mrm.22147
37. Qian Z, Li Y, Wang Y, Li L, Li R, Wang K, et al. Differentiation of glioblastoma from solitarybrain metastases using radiomic machine-learning classifiers. *Cancer Lett*. (2019) 451:128–35. doi: 10.1016/j.canlet.2019.02.054
38. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge betweenmedical imaging and personalized medicine. *Nat Rev Clinicaloncol*. (2017) 14:749. doi: 10.1038/nrclinonc.2017.141
39. Cai H, Cui C, Tian H, Zhang M, Li L. A novel approach to segment and classify regional lymph nodes on computed tomography images. *Comput Math Methods Med*. (2012) 2012:1–9. doi: 10.1155/2012/145926
40. Perone CS, Cohen-Adad J. Promises and limitations of deep learning for medical image segmentation. *J Med Artif Intel*. (2019) 2:1. doi: 10.21037/jmai.2019.01.01
41. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. *Korean J Radiol*. (2019) 20:405–10. doi: 10.3348/kjr.2019.0025
42. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. (2013) 10:e1001381. doi: 10.1371/journal.pmed.1001381
43. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against healthcare professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digital Health*. (2019) 1:e271–97. doi: 10.1016/S2589-7500(19)30123-2

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.

frontiers | Frontiers in Medicine

REVIEW
published: 28 July 2022
doi: 10.3389/fmed.2022.935080

The Application of Artificial Intelligence in the Diagnosis and Drug Resistance Prediction of Pulmonary Tuberculosis

Shufan Liang[1,2†], Jiechao Ma[3†], Gang Wang[2†], Jun Shao[1], Jingwei Li[1], Hui Deng[1,2], Chengdi Wang[1*] and Weimin Li[1*]*

[1] Department of Respiratory and Critical Care Medicine, Med-X Center for Manufacturing, Frontiers Science Center for Disease-Related Molecular Network, West China School of Medicine, West China Hospital, Sichuan University, Chengdu, China, [2] Precision Medicine Key Laboratory of Sichuan Province, Precision Medicine Research Center, West China Hospital, Sichuan University, Chengdu, China, [3] AI Lab, Deepwise Healthcare, Beijing, China

OPEN ACCESS

Edited by:
Mohaimenul Islam,
Aesop Technology, Taiwan

Reviewed by:
Michela Gabelloni,
University of Pisa, Italy
Hosna Salmani,
Iran University of Medical Sciences, Iran

***Correspondence:**
Hui Deng
huideng0923@hotmail.com
Chengdi Wang
chengdi_wang@scu.edu.cn
Weimin Li
weimi003@scu.edu.cn

[†] 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: 03 May 2022
Accepted: 13 June 2022
Published: 28 July 2022

Citation:
Liang S, Ma J, Wang G, Shao J, Li J, Deng H, Wang C and Li W (2022) The Application of Artificial Intelligence in the Diagnosis and Drug Resistance Prediction of Pulmonary Tuberculosis. Front. Med. 9:935080. doi: 10.3389/fmed.2022.935080

With the increasing incidence and mortality of pulmonary tuberculosis, in addition to tough and controversial disease management, time-wasting and resource-limited conventional approaches to the diagnosis and differential diagnosis of tuberculosis are still awkward issues, especially in countries with high tuberculosis burden and backwardness. In the meantime, the climbing proportion of drug-resistant tuberculosis poses a significant hazard to public health. Thus, auxiliary diagnostic tools with higher efficiency and accuracy are urgently required. Artificial intelligence (AI), which is not new but has recently grown in popularity, provides researchers with opportunities and technical underpinnings to develop novel, precise, rapid, and automated implements for pulmonary tuberculosis care, including but not limited to tuberculosis detection. In this review, we aimed to introduce representative AI methods, focusing on deep learning and radiomics, followed by definite descriptions of the state-of-the-art AI models developed using medical images and genetic data to detect pulmonary tuberculosis, distinguish the infection from other pulmonary diseases, and identify drug resistance of tuberculosis, with the purpose of assisting physicians in deciding the appropriate therapeutic schedule in the early stage of the disease. We also enumerated the challenges in maximizing the impact of AI in this field such as generalization and clinical utility of the deep learning models.

Keywords: pulmonary tuberculosis, artificial intelligence, deep learning, radiomics, machine learning

INTRODUCTION

Among the infectious diseases, tuberculosis (TB) is one of the major causes of mortality worldwide, leading to approximately 1.4 million deaths and 10 million new cases annually, according to the World Health Organization (WHO) Global Tuberculosis Report 2021 (1). In addition to the threat to public health posed by TB, the incidence of drug-resistant tuberculosis (DR-TB) continues to increase, resulting in difficulty in controlling the epidemic (2). Accurate detection methods based on bacteria, such as acid-fast bacilli or bacterial cultures, are time-consuming

and condition-limited. Gene testing to identify infection or drug resistance of the pathogen-*Mycobacterium tuberculosis* (*M. tuberculosis*) is inconvenient and restricted by the laboratory environment. Although medical images, such as chest radiographs [also called chest X-ray (CXR)] and computed tomography (CT), are comparatively inexpensive and more available, in certain developing countries or backward areas, there may be no advanced medical equipment or a lack of experienced radiologists to interpret the images, and the growing medical image data may add workload to the physicians. Therefore, automated, precise, efficient, and cost-effective assistance tools devoted to TB management demand prompt exploitation.

Over the past decades, with the vigorous development of computer technology, artificial intelligence (AI) has aroused a whopping level of attention in many fields, especially in image recognition. AI systems based on medical images or other meaningful clinical information have been utilized to screen, diagnose, assess severity, and predict prognosis in multiple diseases, such as brain tumor (3, 4), pneumonia (5), lung cancer (6), cardiovascular disease (7), and even tumor metastasis (8).

In addition, for better implementation of AI in the medical field, particular ethical concerns should also be considered. With the widespread development and utilization of AI, privacy and security during the management and transmission of data, as well as the informed consent of patients are emerging as critical ethical issues. Moreover, specific psychological and legal considerations have also been proposed. For instance, when an error by the automated system leads to a false diagnosis or improper therapeutic schedule resulting in harmful consequences, this may cause a dispute over who should be responsible for that mishap. In medical practice, owing to the opaqueness of the prediction generated by the algorithm, physicians may distrust the model (9). Furthermore, to verify the clinical relevance of the models, clinical trials are required, wherein more intractable issues, such as obtaining informed consent, are present; however, only a few clinical trials involving the use of AI systems have been performed. Collectively, to guide the appropriate adoption of AI systems, the establishment of effective ethical and legal frameworks is of great urgency (10, 11).

We searched the literature in PubMed, Embase, and Web of Science using a retrieval search strategy with the following keywords: "tuberculosis" and "artificial intelligence" or "deep learning" or "radiomics" or "machine learning," selecting quantified studies by the abstracts, and the flow diagram is demonstrated in **Supplementary Figure 1**. In this review, we mainly focused on approaches based on AI using CXR, CT, positron emission tomography (PET)/CT images, and genetic data associated with TB care. By describing the latest typical AI studies focusing on TB, we aimed to inform physicians and radiologists interested in AI for the precise diagnosis of TB to carry out optimal therapeutic regimens.

We started by briefly introducing AI, with deep learning and radiomics stressed; later, we provided a few definite examples of the application of AI in the medical field, especially in respiratory system. We then narrated the up-and-coming AI techniques in TB from three aspects according to the proposed use, namely, TB detection, discrimination between TB and other pulmonary diseases, and recognition of drug resistance of TB (**Figure 1**). Finally, we summarized the significance of previous studies, challenges, and prospects of developing more practical and accurate AI tools for TB in the future.

ARTIFICIAL INTELLIGENCE IN A NUTSHELL

AI is a technical science that studies and develops the theory, method, technology, and application of systems used to simulate and extend human intelligence. Deep learning, a hot topic in this field, which has been probed extensively, mostly leverages convolutional neural networks (CNNs) comprised of multiple layers, including input, convolutional, pooling, fully connected, and output layers, through which the specific predictions could derive from primary digitalized inputs, such as images, speech, gene sequences, and clinical text information (12, 13). What's more, other plentiful sorts of machine learning algorithms, such as logistic regression (LR), random forest (RF), support vector machine (SVM), and decision tree (DT), are valuable components of AI as well (14–18). Radiomics, designed to mine pathophysiological information from medical images, includes a common process involving data collection; identification of the region of interest (ROI); ROI segmentation; feature extraction, selection, and quantification; model establishment; and prediction making in the end (19, 20). The workflow of deep learning and radiomics is displayed in **Figure 2**.

The prosperity of AI applied to the medical field, especially in respiratory system, has attracted substantial attention with promising results, such as detection of pulmonary nodules (21) and prediction of treatment response or outcome of lung cancer (22, 23). Meanwhile, we have made excellent achievements, including diagnosis and discrimination of 2019 novel coronavirus pneumonia (24), predetermination of epidermal growth factor receptor (EGFR) gene mutation status, programmed death ligand-1 (PD-L1) expression level, and target therapy effect in patients with lung cancer (25–27).

As a noticeable disease in this system, AI applied to TB is presented as follows and summarized briefly in **Table 1**.

APPLICATION OF ARTIFICIAL INTELLIGENCE IN PULMONARY TUBERCULOSIS

Detection of Pulmonary Tuberculosis

Since the majority of patients with pulmonary tuberculosis (PTB) have abnormal chest CXR findings, such as cavities, centrilobular nodules, and consolidations (28), which are suggestive of the diagnosis of PTB, and CXR is comparatively widely available, WHO has recommended TB screening in high-risk populations by chest radiographs (29). Similarly, CT images demonstrate abnormalities when PTB occurs. These representative medical images are commonly utilized to train a deep learning model

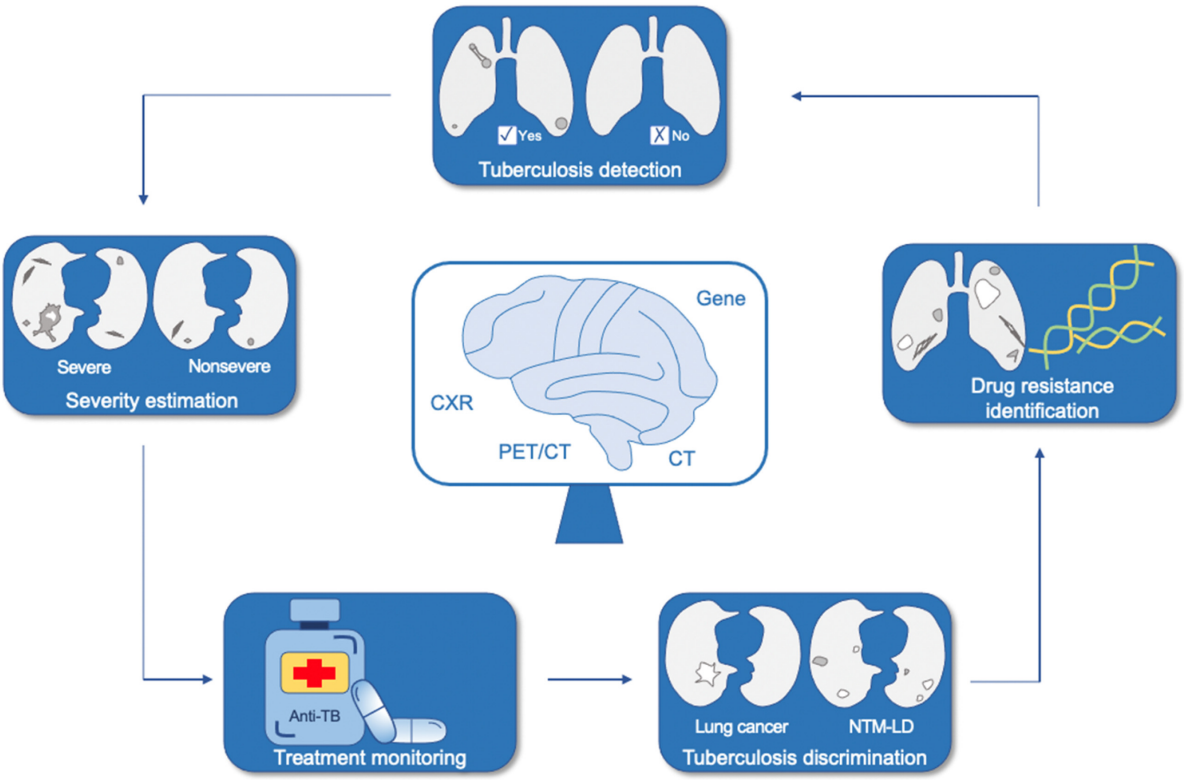

FIGURE 1 | Application of artificial intelligence in tuberculosis. NTM-LD, non-tuberculous mycobacterium lung disease; CXR, chest X-ray; CT, computed tomography; PET/CT, positron emission tomography/computed tomography.

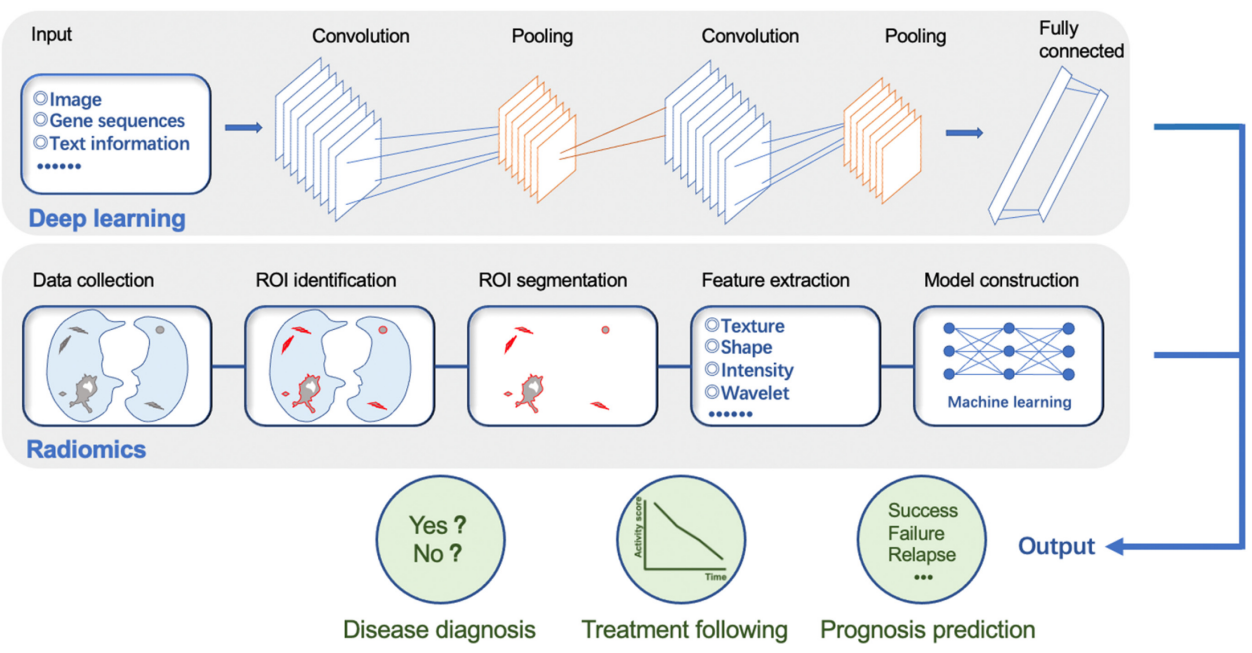

FIGURE 2 | The workflow of deep learning and radiomics. ROI, region of interest.

TABLE 1 | A brief summary of the included studies.

Section	Study proportion	Purpose	Reference standard	Primary materials	Algorithm	Evaluation indicators	References
Tuberculosis detection	48.5%	Diagnose pulmonary tuberculosis or disease evaluation	Pathogenic detection, radiology reports, clinical records, etc.	CXR and CT images	CNN and ML	AUC, sensitivity, specificity, accuracy, etc.	(31–41, 43–47)
Tuberculosis discrimination	18.2%	Discriminate between pulmonary tuberculosis and lung cancer or NTM-LD	Pathogenic detection, pathology, or follow-up confirmation	CT and PET/CT images	CNN and radiomics		(52–55, 59, 60)
Tuberculosis drug resistance prediction	33.3%	Recognize MDR-TB or drug resistance of *Mycobacterium tuberculosis* up to 14 anti-tuberculosis drugs	Drug susceptibility testing	CXR, CT images, and gene sequences	ANN, CNN, GNN, and ML		(63–65, 68–73)

CXR, chest X-ray; CT, computed tomography; CNN, convolutional neural network; ML, machine learning; AUC, area under the curve; NTM-LD, non-tuberculous mycobacterium lung disease; PET/CT, positron emission tomography/computed tomography; MDR-TB, multi-drug resistant tuberculosis; ANN, artificial neural network; GNN, graph neural network.

to detect PTB suffering. As early as 1999, an artificial neural network was exploited to predict active TB, taking advantage of radiographic findings, symptoms, and demographic variables, showing a favorable performance, which gave researchers powerful afflatus (30). Thereafter, abundant studies have been conducted to recognize the contagious disease using radiological images in slightly different forms (**Table 2**).

Detection of Pulmonary Tuberculosis Alone

Lakhani and Sundaram (31) adopted two deep CNNs to detect PTB on CXR images. Finally, the area under the curve (AUC) achieved a significant level at 0.99 [95% confidence interval (CI) 0.96–1.00] on account of a method named "resemble," which indicated that the ultimate PTB probability score was obtained from the two CNNs, with a different weighting of their outputs and choosing the best match. In addition, this study revealed that networks pretrained by daily color images outperformed untrained ones [AUC 0.98 pretrained vs. 0.90 untrained of AlexNet and 0.98 pretrained vs. 0.88 untrained of GoogLeNet ($P < 0.001$) in the test dataset]. Similarly, Hwang et al. (32) developed an automatic detection algorithm to classify active PTB using chest radiographs from a massive dataset containing 60,989 images which eventually manifested high performance both in lesion localization [area under the alternative free-response receiver operating characteristic curves (AUAFROC) 0.973–1.000] and disease classification (AUC 0.977–1.000), while the observer performance test showed that the algorithm had better behavior than physicians with different degrees of experience (AUC 0.993 vs. 0.664–0.925 in localization and 0.993 vs. 0.746–0.971 in classification). Another study developed an algorithm based on ResNet to detect PTB, and the model reached an accuracy of 96.73% with a heatmap generation for precise lesion location as well (33). Using 20,135 chest radiographs from 19,681 asymptomatic individuals, an out-of-sample test was conducted (34) to validate the screening performance of the deep learning-based automated detection (DLAD) algorithm developed by Hwang et al. (32). Five images from four active PTB cases confirmed by the bacteriological test were properly classified as having abnormal discoveries with specificities of 0.997 and 0.959 at high specificity and high sensitivity thresholds, respectively. Moreover, DLAD showed a decent performance in identifying radiologically relevant abnormalities with an AUC of 0.967 (95% CI 0.93–0.996). Likewise, to verify the performance of deep learning models on the general population, a study assessed five CNNs in two forms, namely, I-CNN (images input only) and D-CNN [images and demographic variables (age, sex, height, and weight) input] to detect PTB by CXR images in 39,677 workers from Korea. Among the five models, VGG19 achieved the highest performance in both the training and test cohorts, regardless of the demographic information input (AUC 0.9075 of I-CNN and 0.9213 of D-CNN in the test set), and the AUCs of the other four systems were all over 0.88 with D-CNN in the test set. Moreover, no statistical significance was observed when only a single demographic variable was included ($P > 0.05$) (35). Taking advantage of segmentation and augmentation, EfficientNetB3, the CNN structure, demonstrated incredibly high performance in PTB detection with an AUC of 0.999 (36). Moreover, a simplified network was proposed to surmount the trouble of overfitting and difficult deployment in mobile settings owing to the large scale of parameters and hardware requirements of the models, achieving an AUC of 0.925 through 5-fold cross-validation in the diagnosis of PTB on CXRs (37). Uniquely, different from the studies mentioned earlier, Rajaraman et al. blazed new trails to recognize findings consistent with PTB by lateral CXRs through deep learning, with an AUC up to 0.9491 (38).

TB is important not only in the general population but also in patients with specific conditions. Due to the high mortality caused by TB in human immunodeficiency virus (HIV)-positive patients with the conspicuous incidence and improper treatment, in South Africa, Rajpurkar et al. utilized CXRs, as well as certain clinical covariates, including age, temperature, hemoglobin, and white blood cell counts of 677 HIV-positive patients from two

TABLE 2 | Summary of AI applications in TB detection.

No.	References	Method	Reference standard	Dataset	Study population	Training/Validation/test cohort	Model names	Algorithm	Results
1	Lakhani and Sundaram (31)	Retrospective multi-center on CXR images	Sputum, radiology reports, radiologists, and clinical records.	1,007 participants	United States, China, and Belarus	Training: 685 Validation: 172 Test: 150	NA	CNN	AUC 0.99, Sen 97.3%, Spe 94.7%, Acc 96.0% of the ensemble method
2	Hwang et al. (32)	Retrospective multi-center on CXR images	Culture or PCR	62,433 CXR images	Korea, China, United States, etc.	Training: 60,089 Tuning: 450 Internal validation: 450 External validation: 1,444	DLAD	CNN	AUC 0.977–1.000 for TB classification, AUAFROC 0.973–1.000 for lesion localization; Sen 0.943–1.000, Spe 0.911–1.000 at high sensitivity cutoff
3	Nijiati et al. (33)	Retrospective single-center on CXR images	Symptoms, laboratory and radiological examinations	9,628 CXR images	China	Training: 7,703 Test: 1,925	NA	CNN	AUC 0.9902–0.9944, Sen 93.2–95.5%, Spe 95.78–98.05%, Acc 94.96–96.73% in the test set
4	Lee et al. (34)	Retrospective single-center on CXR images	Smear microscopy, culture, PCR, and radiologists	19,686 participants	Korea	Test: 19,686	DLAD	CNN	AUC 0.999, Sen 1.000, Spe 0.959–0.997, Acc 0.96–0.997
5	Heo et al. (35)	Retrospective single-center on CXR images	Radiologists	39,677 participants	Korea	Training: 2,000 Test: 37,677	D-CNN and I-CNN	CNN	AUC 0.9213, Sen 0.815, Spe 0.962 of D-CNN
6	Nafisah and Muhammad (36)	Retrospective multi-center on CXR images	NA	1,098 CXR images	United States, China, and Belarus	5-fold cross validation	NA	CNN	AUC 0.999, Acc 98.7%, recall 98.3%, precision 98.3%, Spe 99.0%
7	Pasa et al. (37)	Retrospective multi-center on CXR images	NA	1,104 participants	United States, China, and Belarus	5-fold cross validation	NA	CNN	AUC 0.925, Acc 86.2%
8	Rajaraman et al. (38)	Retrospective multi-center on CXR images	Radiologists	76,031 CXR images	United States and Spain	Training: test 9:1	NA	CNN	AUC 0.9274–0.9491, recall 0.7736–0.8113, precision 0.9524–0.9773, Acc 0.8585–0.8962

(Continued)

TABLE 2 | (Continued)

No.	References	Method	Reference standard	Dataset	Study population	Training/Validation/test cohort	Model names	Algorithm	Results
9	Rajpurkar et al. (39)	Retrospective multi-center on CXR images	Culture or Xpert MTB/RIF	677 participants	South Africa	Training: 563 Test: 114	CheXaid	Deep learning	AUC 0.83, Sen 0.67, Spe 0.87, Acc 0.78
10	Lee et al. (40)	Retrospective multi-center on CXR images	Sputum microscopy, culture or PCR	6,964 participants	Korea	Training: validation 7:3 Test: 455	NA	CNN	AUC 0.82–0.84, Spe 26–48.5% at the cutoff of 95% Sen in the test set
11	Yan et al. (41)	Retrospective multi-center on CT images	Culture	1,248 CT images	China and United States	Training: validation 8:2 External test: 356	NA	CNN	Acc 95.35–98.25%, recall 94.87–100%, precision 94.87–98.70%
12	Khan et al. (43)	Prospective single-center on CXR images	Culture	2,198 participants	Pakistan	Test: 2,198	qXR and CAD4TB	CNN	AUC 0.92, Sen 0.93, Spe 0.75 for qXR; AUC 0.87, Sen 0.93, Spe 0.69 for CAD4TB
13	Qin et al. (44)	Retrospective multi-center on CXR images	Xpert MTB/RIF	1,196 participants	Nepal and Cameroon	Test: 1,196	qXR, CAD4TB, and Lunit INSIGHT CXR	CNN	AUC 0.92–0.94, Sen 0.87–0.91, Spe 0.84–0.89, Acc 0.85–0.89
14	Qin et al. (45)	Retrospective multi-center on CXR images	Xpert MTB/RIF	23,954 participants	Bangladesh	Test: 23,954	qXR, CAD4TB, InferRead DR, etc.	CNN	AUC 84.89–90.81%, Sen 90.0–90.3%, Spe 61.1–74.3% when fixed at 90% Sen
15	Codlin et al. (46)	Retrospective multi-center on CXR images	Xpert MTB/RIF	1,032 participants	Viet Nam	Test: 1,032	qXR, CAD4TB, Genki, etc.	CNN	AUC 0.50–0.82, Spe 6.3–48.7%, Acc 17.8–54.7% when fixed at 95.5% Sen
16	Melendez et al. (47)	Retrospective single-center on CXR images	Culture	392 patients	South Africa	10-fold cross validation	CAD4TB	Machine learning	AUC 0.72–0.84, Spe 24–49%, NPV 95–98% when fixed at 95% Sen

AI, artificial intelligence; TB, tuberculosis; CXR, chest X-ray; NA, not available; CNN, convolutional neural network; AUC, area under the curve; Sen, sensitivity; Spe, specificity; Acc, accuracy; PCR, polymerase chain reaction; AUAFROC, area under the alternative free-response receiver-operating characteristic curve; CT, computed tomography.

hospitals to establish a deep learning algorithm, named CheXaid, which improved the clinicians' diagnostic accuracy slightly (0.65 vs. 0.60, $P = 0.002$). Interestingly, the performance of the algorithm alone was superior to that of clinicians assisted by AI (accuracy 0.79 vs. 0.65, $P < 0.001$). Moreover, the training strategy of adding clinical variables with CXRs improved the performance of the algorithm (AUC of 0.83 and 0.71 in the combination model and model alone) in this study and suggested the importance of integrating inputs in various modalities to enhance the power of the models (39).

Detection of Pulmonary Tuberculosis With Treatment Monitoring and Severity Estimation

Apart from detecting PTB, deep learning is capable to follow post-treatment changes and estimate the severity of it. Utilizing CXRs, the output of the model developed by Lee and his team elevated by 0.30 when the degree of smear positivity increased ($P < 0.001$) and decreased gradually during treatment; meanwhile, the model achieved AUCs over 0.82 in the two test sets for PTB diagnosis (40).

Owing to higher resolution and more subtle presentation, CT images provide more nuanced information on the lung region and play an important role in PTB diagnosis as well (28). Thus, Yan et al. (41) developed a model to detect PTB and quantitatively evaluate the disease burden, of which the quantified TB scores were obviously higher in severe patients than in non-severe ones and was well correlated with the CT scores assessed by radiologists. Moreover, the model demonstrated an accuracy of 83.37% for classifying the six pulmonary lesion types, such as consolidation and calcified granulomas in the validation set, while an accuracy of 98.25% was achieved for distinguishing active PTB patients from inactive individuals in the test set.

The two studies are unique as there is a lack of research targeting treatment monitoring and disease burden estimation of TB by AI methods, which inspires us to launch more relevant studies to give rein to their adjuvant role in the clinic.

Validation of Computer-Aided Pulmonary Tuberculosis Detection Systems

In addition to the models obtained from the original studies, the computer-aided detection (CAD) systems, such as qXR, CAD4TB, and Lunit INSIGHT CXR (42), which generate a PTB classification when the output is more than a defined threshold score, have been established to facilitate PTB detection using CXR images based on deep learning. Several studies have exclusively assessed the diagnostic ability of the application of various categories and versions in diverse datasets.

To identify the practicalities of qXR version 2.0 (qXRv2) and CAD4TB version 6.0 (CAD4TBv6) in detecting PTB in low- and middle-income countries with a high disease burden, Khan et al. (43) conducted a prospective single-center study with 2,198 individuals at the Indus Hospital, located in Karachi, Pakistan. Finally, qXRv2 attained a sensitivity of 0.93 (95% CI 0.89–0.95) and a specificity of 0.75 (95% CI 0.73–0.77), while CAD4TBv6 showed a specificity of 0.69 (95% CI 0.67–0.71) when matched with the same sensitivity, both reaching the Target Product Profile recommendations defined by WHO (sensitivity $\geq$ 0.90 and specificity $\geq$ 0.70). What's more, the sensitivity decreased obviously in smear-negative patients compared to that in smear-positive patients (0.80 in the negative group vs. 0.96 in the positive cohort of qXRv2 and 0.82 in the negative population vs. 0.97 in positive individuals of CAD4TBv6). This study is worth emphasizing because it is a rare prospective investigation validating CAD approaches.

Qin et al. (44) estimated three commercially available CAD tools, qXRv2, CAD4TBv6, and Lunit INSIGHT CXR, to triage PTB in 1,196 participants from Nepal and Cameroon, with AUCs above 90% [0.94 (95% CI 0.93–0.96) for Lunit INSIGHT CXR, 0.94 (95% CI 0.92–0.97) for qXRv2, and 0.92 (95% CI 0.90–0.95) for CAD4TBv6]. When the purpose was to reduce the Xpert test by 50%, the sensitivities of the three models maintained at 97–99%, with no statistical significance among them. Subsequently, the group assessed five AI algorithms in newer versions, including CAD4TB version 7 (CAD4TBv7), qXR version 3 (qXRv3), Infereread DR version 2, Lunit INSIGHT CXR version 4.9.0, and JF CXR-1 version 2, on a massive dataset comprising CXRs from 23,954 individuals. The performance of all of them surpassed that of three radiologists as a concrete manifestation that AI showed higher specificity and positive predictive values (PPVs) when matched with the same sensitivity (45). Another study evaluated a maximum of 12 CAD solutions to identify PTB in comparison with an experienced radiologist and an intermediate reader. The final results showed that qXRv3, CAD4TBv7, and Lunit INSIGHT CXR version3.1.0.0 achieved the highest AUC of 0.82. Meanwhile, five of them surpassed the intermediate reader in specificity and accuracy when holding at the same sensitivity, while only qXRv3 maintained comparable specificity when sensitivity reached the standard of the experienced reader [95.5% (95% CI 90.4–98.3%)] (46). The three studies mentioned (43, 45, 46) coincidentally discovered that, in groups with previous TB, the performance of AI systems would decline to some extent. In addition, when integrating clinical information with the CAD scores of CXRs generated by CAD4TB, the AUC of the combination framework reached 0.84, improving the performance of CAD4TB alone (47).

Although the verification results seem remarkable as a whole, more prospective validation tests need to be carried out in a real medical environment, after which these mercantile AI systems may be competent enough to supply convenient, efficient, and accurate tools for physicians worldwide, facilitating clinical decision making in the near future.

Discrimination Between Pulmonary Tuberculosis and Other Lung Diseases

In addition to detection, effort has been made to differentiate PTB from other pulmonary diseases (**Table 3**).

Discrimination Between Tuberculosis and Lung Cancer

Lung cancer is one of the primary causes of cancer death and is the most common tumor worldwide (48). Moreover, pulmonary tuberculosis granuloma (TBG) may present as lung adenocarcinoma (LAC) with the demonstration of similar solitary pulmonary nodules (49–51), resulting in diagnostic

TABLE 3 | Summary of AI applications in discrimination between pulmonary tuberculosis and other lung diseases.

No.	References	Method	Reference standard	Dataset	Study population	Discrimination	Training/Validation/ test cohort	Model names	Algorithm	Results
1	Feng et al. (52)	Retrospective multi-center on CT images	Histological diagnosis	550 patients	China	PTB and lung cancer	Training:218 Internal validation:140 External validation: 192	NA	DLN	AUC 0.809, Sen 0.908, Spe 0.608, Acc 0.828 in the external validation set
2	Zhuo et al. (53)	Retrospective multi-center on CT images	Surgical pathology, specimen culture or assay	313 patients	China	PTB and lung cancer	Training: validation 7:3	NA	Radiomics nomogram	AUC 0.99, Sen 0.9841, Spe 0.9000, Acc 0.9570 in the validation set
3	Hu et al. (54)	Retrospective multi-center on PET/CT images	Pathological or follow-up confirmation	235 patients	China	PTB and lung cancer	Training: 163 Validation: 72	NA	Radiomics nomogram	AUC 0.889, Sen 85%, Spe 78.12%, Acc 79.53% in the validation set
4	Du et al. (55)	Retrospective single-center on PET/CT images	Pathology	174 patients	China	PTB and lung cancer	Training: 122 Validation: 52	NA	Radiomics nomogram	AUC 0.93, Sen 0.86, Spe 0.83, Acc 0.85 in the validation set
5	Wang et al. (59)	Retrospective multi-center on CT images	Sputum acid-fast bacilli stain or culture	1,185 patients	China	MTB-LD and NTM-LD	Training: validation: test 8:1:1 External test: 80	NA	CNN	AUC 0.78, Sen 0.75, Spe 0.63, Acc 0.69 in the external test set
6	Yan et al. (60)	Retrospective multi-center on CT images	Sputum culture or smear	182 patients	China	MTB-LD and NTM-LD	Training: validation 8:2 External validation: 40	NA	Radiomics	AUC 0.84—0.98, Sen 0.61–0.97, Spe 0.61–0.97 in the external validation set

AI, artificial intelligence; CT, computed tomography; PTB, pulmonary tuberculosis; NA, not available; DLN, deep learning nomogram; AUC, area under the curve; Sen, sensitivity; Spe, specificity; Acc, accuracy; PET/CT: positron emission tomography/computed tomography; MTB-LD, Mycobacterium tuberculosis lung disease; NTM-LD, non-tuberculous mycobacterium lung disease; CNN, convolutional neural network.

confusion and treatment mistakes. A deep learning-based nomogram (DLN) using CT images was developed and validated to distinguish TBG from LAC (52). The DLN was constituted to compare with a clinical model including age, sex, and subjective findings on CT images, and a deep learning signature (DLS) model, with scores derived from 14 deep learning features constructed in advance, and showed better diagnostic performance than the clinical and DLS models. Comprised by age, sex, lobulated shape, and DLS score, DLN achieved both higher AUC and sensitivity than the other 2 models in the internal validation cohort, meanwhile showing an AUC of 0.809 in the external validation set. A radiomics nomogram based on CT images was proposed by another group, showing an AUC of 0.99 in the validation set to differentiate the two fundamentally different diseases which demonstrated similarities between each other (53). Analogously, to distinguish between solitary LAC and PTB, Hu et al. constructed a radiomic model containing a set of nine fluorine-18-fluorodeoxyglucose PET/CT (18F-FDG PET/CT) radiomic features, such as Histogram_Skewness and SHAPE_Sphericity (54). While developing a clinical model, they also constructed a complex model, which was a combination of the radiomic and clinical models using multivariate LR. Finally, the radiomic and complex models outperformed the clinical model, as the AUC of the complex model reached 0.909, while the radiomic and clinical models achieved 0.889 and 0.644 in the validation set. Furthermore, a similar study utilized a radiomic nomogram integrating the radiomic score (RAD-score) derived from a weighted linear combination of features selected from 18F-FDG PET/CT images and three semantic features to differentiate the two semblable image phenotypes. The diagnostic performance of the radiomic nomogram slightly surpassed that of the radiomic and semantic models with an AUC of 0.93 in the validation cohort; the decision curve also illustrated the net benefit of the nomogram (55).

Discrimination Between Tuberculosis and Non-tuberculous Mycobacterium Lung Disease

Given that non-tuberculous mycobacterium lung disease (NTM-LD) demonstrates an increasing incidence and prevalence in recent years (56, 57), due to similar clinical symptoms and CT imaging characteristics with mycobacterium pulmonary tuberculosis lung disease (MTB-LD) (58), it is crucial to distinguish the different infections as quickly as possible in the early stage to permit appropriate treatment implementation. A deep learning framework was developed by Wang and his colleagues to differentiate between NTM-LD and MTB-LD on chest CT images with an AUC of 0.86 and 0.78 in the internal test set and in the external test cohort, respectively (59). Moreover, the model surpassed three radiologists in almost every metric with higher diagnostic efficiency (1,000 times faster) and output class activation maps identifying abnormal lung areas without manual annotation. To achieve a similar purpose, another study leveraged radiomics by taking advantage of the features of cavities in CT images using six machine learning models (SVM, RF, LR, etc.) (60); 458 ROIs were depicted by two radiologists, with 29 optimal quantified image features, such as gradient and wavelet, selected subsequently. AUCs of the six models were up to over 0.98 in the training and validation sets.

These studies pioneered the application of AI for the discrimination of PTB from lung cancer and NTM-LD, with promising results encouraging investigators to develop more AI models using a variety of original training materials to differentiate PTB from more diseases.

Identification of Tuberculosis Drug Resistance

In the context of increasing incidence and intractable management of TB resistance, multiple examination approaches, including drug susceptibility testing (DST), Xpert MTB/RIF, line-probe assays, and whole-genome sequencing (WGS), have been explored to identify DR-TB (2). However, cost and time issues are still remaining. Hence, inexpensive, rapid, and accurate tools for automated detection of the antimicrobial resistance are of great concern (**Table 4**).

Drug-Resistant Tuberculosis Identification Based on Medical Images

Imaging manifestations of these two main categories of TB, sensitive or resistant to anti-tuberculosis therapy (ATT), differ depending on the phenotypes, as DR-TB could demonstrate larger lesions and thick-walled cavities on CXR images (61, 62). Jaeger et al. (63) trained an artificial neural network through cross-validation to identify patients with multi-drug resistant tuberculosis (MDR-TB) using CXRs, which achieved an AUC of only up to 66%. This unsatisfactory result may be explained by the small dataset containing only 135 cases. However, it is inspiring that the team used a larger dataset of 5,642 CXRs and various CNNs for the same purpose, and finally, a preferable outcome was obtained. With static or dynamic data augmentation, the AUC of InceptionV3 increased to 0.85. For custom CNNs, six-layer CNN expressed the best performance with an AUC of 0.74 (64). After the ImageCLEF2017 competition, a study utilized a small dataset from the match, which comprised CT images from 230 drug-sensitive and MDR-TB patients to implement a combination of a patch-based deep CNN and SVM, with an accuracy of 91.11% in predicting MDR-TB at the patient level and 79.8% at the patch level (65).

To date, the exploitation of using medical images to identify DR-TB has not been investigated thoroughly; hence, these studies are noteworthy because they could give us some instructions for future research orientation.

Drug-Resistant Tuberculosis Identification Based on Genetic Data

Besides medical images, genetic information could also serve as a diagnostic tool for TB. As introduced above, various molecular approaches are capable of detecting drug resistance, of which the theoretical proof is that the resistance occurrence in TB is caused by chromosomal mutations, passing along through vertical descent, in present genes. Meanwhile, rapid molecular tests using genomic information are more efficient than culture-based assays so they are adopted widely, and related gene data are available for scientific research (66). Therefore, numerous AI

TABLE 4 | Summary of AI applications in TB drug resistance identification.

No.	References	Method	Reference standard	Dataset	Study sample	Resistance identification	Training/Validation/ test cohort	Model names	Algorithm	Results
1	Jaeger et al. (63)	Retrospective multi-center on CXR images	NA	135 patients	Belarus	MDR-TB	5-fold cross validation	NA	ANN, CNN and ML	AUC 50–66%, Acc 0.62–0.66
2	Karki et al. (64)	Retrospective multi-center on CXR images	DST	5,642 CXR images	United States, China, etc.	DR-TB	10-fold cross validation	NA	CNN	AUC 0.85
3	Gao and Qian (65)	Retrospective multi-center on CT images	NA	230 patients	NA	MDR-TB	Training: 150 Validation: 35 Test: 45	NA	CNN and ML	Acc 64.71–91.11%
4	Yang et al. (68)	Retrospective multi-center on gene sequences	DST	8,388 isolates	European, Asia, and Africa	4 drugs and MDR-TB	Training: test 7:3	DeepAMR	ML	AUC 94.4–98.7%, Sen 87.3–96.3%, Spe 90.9–96.7%
5	Yang et al. (69)	Retrospective multi-center on gene sequences	DST	13,402 isolates	NA	4 drugs	Training: validation: test 4:2:2 or stratified cross validation	HGAT-AMR	GNN	AUC 72.83–99.10%, Sen 50.65–96.60%, Spe 79.50–98.87%
6	Yang et al. (70)	Retrospective multi-center on gene sequences	DST	1,839 isolates	United Kingdom	8 drugs and MDR-TB	Cross-validation	NA	ML	AUC 91–100%, Sen 84–97%, Spe 90–98%
7	Deelder et al. (71)	Retrospective multi-center on gene sequences	DST	16,688 isolates	NA	14 drugs and MDR-TB	5-fold cross validation	NA	ML	Acc 73.4–97.5%, Sen 0–92.8%, Spe 75.6–100%
8	Chen et al. (72)	Retrospective multi-center on gene sequences	DST	4,393 isolates	ReSeqTB Knowledgebase	10 drugs	10-fold cross validation Independent validation: 792	NA	WDNN and ML	AUC 0.937, Sen 87.9%, Spe 92.7% for the first-line drugs
9	Gröschel et al. (73)	Retrospective multi-center on gene sequences	DST	20,408 isolates	NCBI Nucleotide Database	10 drugs	Training: validation 3:1	GenTB	WDNN and ML	AUC 0.73–0.96, Sen 57–93%, Spe 78–100%
10	Kuang et al. (75)	Retrospective multi-center on gene sequences	DST	10,575 isolates	China, Cameroon, Uganda, etc.	8 drugs	10-fold cross validation	NA	CNN and ML	Acc 89.2–99.2%, Sen 93.4–100%, Spe 48.0–91.7%, F1 score 93.3–99.6%
11	Jiang et al. (76)	Retrospective multi-center on gene sequences	DST	12,378 isolates	NCBI-SRA Database	4 drugs	Training: validation: test 8:1:1 and 10-fold cross validation	HANN	Attentive neural network	AUC 93.66–99.05%, Sen 67.12–96.31%, Spe 92.52–98.84%

AI, artificial intelligence; TB, tuberculosis; CXR, chest X-ray; NA, not available; MDR-TB, multi-drug resistant tuberculosis; ANN, artificial neural network; CNN, convolutional neural network; ML, machine learning; AUC, area under the curve; Acc, accuracy; DST, drug susceptibility testing; DR-TB, drug-resistant tuberculosis; CT, computed tomography; Sen, sensitivity; Spe, specificity; GNN, graph neural network; WDNN, wide and deep neural network; SRA, sequence read archive.

studies based on gene sequences have been explored to identify drug resistance of *M. tuberculosis*, as follows.

As researched previously, deep learning using genomic data has been applied to reveal antibiotic resistance (67). Thus, with mutations for isolates input and phenotypes of drug resistance output, Yang et al. (68) developed "DeepAMR," a deep learning model with a deep denoising auto-encoder for multiple tasks to predict co-occurrent drug resistance of *M. tuberculosis*, comparing the model with conventional machine learning methods, including RF, SVM, and ensemble classification chains (ECC). The co-occurrence of rifampicin (RIF) and isoniazid (INH) resistance accounted for the majority of the dataset (n = 8,388). The results suggested that the model surpassed all other approaches in predicting resistance to the four first-line drugs, MDR-TB, and pan-susceptible tuberculosis (PANS-TB, isolates susceptible to any of the four first-line drugs), showing AUCs from 94.4 to 98.7% ($P < 0.05$). Later, utilizing a novel method using graphs translated from genetic data of *M. tuberculosis*, the team developed a graph neural network named "HGAT-AMR" to predict drug resistance in a sample consisting of 13,402 isolates tested for susceptibility to up to 11 drugs (69). HGAT–AMRi-E (HGAT–AMR trained on any available incomplete phenotype specimen for the multi-label learning task) and HGAT–AMRs (HGAT–AMR trained on individual subsets of different drugs for the single-label learning task) performed best in INH and RIF, respectively, with AUCs of 98.53 and 99.10%. Meanwhile, HGAT–AMRi-E demonstrated the highest sensitivity for INH, ethambutol (EMB), and pyrazinamide (PZA) at 94.91, 96.60, and 90.63%, respectively, and HGAT-AMR outperformed SVM and LR, unless in a condition of highly imbalanced data when an isolate had only been tested by INH and EMB, but not by other drugs. Favorable performance was yielded in machine learning models constructed by the group as well, with higher sensitivity compared to the previous rule-based method ($P < 0.01$) (70). Collecting 16,688 isolates of which the WGS and DST data are available to predict drug resistance, another study developed the gradient boosted tree, a machine learning method, reaching an accuracy of 95.5% in MDR-TB identification (71).

Similarly, to determine the drug resistance of *M. tuberculosis* strains by inputting gene sequences, Chen et al. compared the performance of three deep learning models (72). The wide and deep neural network (WDNN), constructed in the study, incorporating LR and deep multilayer perceptron, was presented in four forms, namely, kSD-WDNN (detecting preselected mutations), SD-WDNN (detecting single resistance), and 2 MD-WDNNs (detecting common mutations and for all mutations in multiple resistance), in which the most complex model MD-WDNN surpassed others in both first-line and second-line drugs, with average AUCs of 0.937 and 0.891 in the validation set. Subsequently, a correlative study developed a user-friendly online tool named GenTB based on genome sequencing to predict the antibiotic resistance (73), involving the WDNN and an RF algorithm constituted by Farhat et al. (74). After testing on 20,408 isolates, both GenTB-RF and GenTB-WDNN demonstrated satisfactory performance in first-line drugs with AUCs of more than 87% and with a slightly lower performance in second-line drugs. In particular, GenTB-RF reached the highest prediction for RIF [AUC 96% (95% CI 95–96%)]. Based on 1D CNN, using large and diverse *M. tuberculosis* isolates from six continents to verify the accuracy and steadiness of deep learning, another study developed a model which outperformed the advanced Mykrobe classifier which utilizes a De Bruijn graph to identify resistance profiles in antimicrobial-resistant prediction with higher F1 scores (75). Concurrently, it is worth mentioning that an innovative hierarchical attentive neural network has been constructed to predict the drug resistance of *M. tuberculosis* through genome-wide variants recently, discovering a potential gene related to drug resistance besides achieving supernal AUC and sensitivity in resistance recognition (76).

DISCUSSION

As described earlier, in terms of TB detection, discrimination, and drug resistance identification, AI showed a great potential, with performance approximate to or even better than that of physicians. Yet, there are still lots of challenges remaining, with the concurrence of prospects, as described below.

Challenges

First, DR-TB remains a critical issue worldwide, with an increasing incidence and tough management. Developing dependable AI systems using sufficient radiology-based data, which is more convenient than gene sequences to rapidly recognize patients with DR-TB to assist physicians in executing correct clinical decisions, is of great imperative.

Then, up till now, only a few studies have adopted deep learning or other AI approaches to predict TB relapse or treatment response to anti-tuberculosis drugs. An algorithm based on CNN was proposed to predict the persistence time needed to achieve culture negative in TB individuals with an unsatisfactory accuracy, regrettably (77). In addition, it has been revealed that the minimum inhibitory concentration grew higher with an increasing risk of relapse (78) and aggressive regimens may reduce the recurrence of MDR-TB after successful treatment (79). Thus, if a prediction of relapse can be made in advance, more precise and positive treatment could be carried out to reduce the hazard of returning.

Third, the generalization of these models in a broader population remains to be seen, since not all of those studies contain external tests, and research samples are not abundant or variable enough. However, studies, including external validation sets, demonstrated diminishing performance from training to external cohorts which gives us a hint of sustaining the reproducibility of the models to suit various individuals. Perhaps, multicenter studies in an enormous study population are capable of solving this problem, but the subsequent issues of data transmission efficiency and security in the process of data sharing deserve to be highlighted.

As for model modalities, since Lu et al. developed a fusion CNN integrated with images and basic clinical information to predict lung cancer (80) and the model CheXaid utilized CXRs with clinical variables to detect TB in HIV patients (39), it is being probed prevalently in the construction of a fusion

neural network, which is composed of several modules dealing with data at multiple scales. Thus, there is an incredible amount of untapped potential to develop AI models with the capacity to handle multimodal inputs. Furthermore, primary inputs, including images or data in other forms, are supposed to be standardized, while diversiform data obtained from different apparatuses may be at an uneven quality level.

Finally, to achieve the purpose of directing clinical practice, the practicality of these novel models should be tested in a real medical environment and seamlessly integrated into the routine workflow, especially in countries with high TB burden and a lack of advanced medical equipment and professional physicians. Owing to the prospective real-world clinical setting, the superior performance of retrospectively developed AI compared with that of human should be regarded with some care.

Prospect

Following tremendous progress in computational power and advanced techniques, AI is blooming increasingly in countless fields. In radiology, AI demonstrates remarkable performance in the detection, treatment monitoring, and prognosis prediction of multiple diseases, especially in oncology. With regard to TB, saving labor and time costs, AI is capable of improving detection efficiency and precision; therefore, medical institutions worldwide could benefit from these novel assistance tools. In the coming decades, after better integration with clinical workflow, AI will exert a brilliant influence on the entire duration of TB from screening, diagnosis, and treatment following to outcome prediction, meanwhile saving medical resources, avoiding inappropriate management, and improving the quality of life of patients.

CONCLUSION

AI-based approaches, including deep learning, radiomics, and other conventional machine learning methods applied to TB, provide a self-driven, convenient, and time-saving strategy to improve diagnostic efficiency and accuracy, outperforming radiologists. Nonetheless, the clinical utility of them remains to be verified, while pitfalls, such as reproducibility of the model and data standardization, need to be addressed as well. To summarize, in this review, we listed several studies focusing on AI-based assistance methods applied to TB detection, discrimination, and drug resistance identification using CXR, CT, PET/CT images, and genome data. Although most of these studies developed AI models with favorable performance, quite a few hurdles must be overcome along the way to maximize the potential of AI. Although TB is especially emphasized in this study, application of AI in other diseases is worth equivalent attention.

AUTHOR CONTRIBUTIONS

WL and CW contributed to supervision and conceptualization. SL, GW, JS, and JL designed the search strategy, performed the literature retrieval, and contributed to the original draft of the manuscript. CW, HD, JM, and SL reviewed and revised the final version of the manuscript. WL contributed to the funding acquisition. All authors read and approved the submitted version.

FUNDING

This study was supported by the National Natural Science Foundation of China (82100119, 91859203, and 81871890), the Science and Technology Project of Sichuan (2020YFG0473 and 2022ZDZX0018), the Chinese Postdoctoral Science Foundation (2021M692309), and the Postdoctoral Program of Sichuan University (2021SCU12018).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.935080/full#supplementary-material

REFERENCES

1. World Health O. *Global Tuberculosis Report 2021*. Geneva: World Health Organization (2021).
2. Lange C, Dheda K, Chesov D, Mandalakas AM, Udwadia Z, Horsburgh CR Jr. Management of drug-resistant tuberculosis. *Lancet*. (2019) 394:953–66. doi: 10.1016/s0140-6736(19)31882-3
3. Hollon TC, Pandian B, Adapa AR, Urias E, Save AV, Khalsa SSS, et al. Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nat Med.* (2020) 26:52–8. doi: 10.1038/s41591-019-0715-9
4. Martini ML, Oermann EK. Intraoperative brain tumour identification with deep learning. *Nat Rev Clin Oncol.* (2020) 17:200–1. doi: 10.1038/s41571-020-0343-9
5. 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. *Cell.* (2020) 182:1360. doi: 10.1016/j.cell.2020.08.029
6. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med.* (2019) 25:954–61. doi: 10.1038/s41591-019-0447-x
7. Gal R, van Velzen SGM, Hooning MJ, Emaus MJ, van der Leij F, Gregorowitsch ML, et al. Identification of risk of cardiovascular disease by automatic quantification of coronary artery calcifications on radiotherapy planning CT scans in patients with breast cancer. *JAMA Oncol.* (2021) 7:1024–32. doi: 10.1001/jamaoncol.2021.1144
8. Pan C, Schoppe O, Parra-Damas A, Cai R, Todorov MI, Gondi G, et al. Deep learning reveals cancer metastasis and therapeutic antibody targeting in the entire body. *Cell.* (2019) 179:1661–76.e19. doi: 10.1016/j.cell.2019.11.013
9. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med.* (2022) 28:31–8. doi: 10.1038/s41591-021-01614-0
10. Pesapane F, Volonté C, Codari M, Sardanelli F. Artificial intelligence as a medical device in radiology: ethical and. regulatory issues in Europe and the United States. *Insights Imaging.* (2018) 9:745–53. doi: 10.1007/s13244-018-0645-y

11. Coppola F, Faggioni L, Gabelloni M, De Vietro F, Mendola V, Cattabriga A, et al. Human, all too human? An all-around appraisal of the "artificial intelligence revolution" in medical imaging. *Front Psychol.* (2021) 12:710982. doi: 10.3389/fpsyg.2021.710982
12. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* (2015) 521:436–44. doi: 10.1038/nature14539
13. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw.* (2015) 61:85–117. doi: 10.1016/j.neunet.2014.09.003
14. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. *Radiographics.* (2017) 37:505–15. doi: 10.1148/rg.2017160130
15. Stoltzfus JC. Logistic regression: a brief primer. *Acad Emerg Med.* (2011) 18:1099–104. doi: 10.1111/j.1553-2712.2011.01185.x
16. Breiman L. Random forests. *Mach Learn.* (2001) 45:5–32.
17. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics Proteomics.* (2018) 15:41–51. doi: 10.21873/cgp.20063
18. Quinlan JR. Induction of decision trees. *Mach Learn.* (1986) 1:81–106. doi: 10.1007/BF00116251
19. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology.* (2016) 278:563–77. doi: 10.1148/radiol.2015151169
20. Zhou Y, Xu X, Song L, Wang C, Guo J, Zhang Y, et al. The application of artificial intelligence and radiomics in lung cancer. *Precis Clin Med.* (2020) 3:214–27. doi: 10.1093/pcmedi/pbaa028
21. Hua KL, Hsu CH, Hidayati SC, Cheng WH, Chen YJ. Computer-aided classification of lung nodules on computed tomography images *via* deep learning technique. *Onco Targets Ther.* (2015) 8:2015–22. doi: 10.2147/OTT.S80733
22. Bera K, Braman N, Gupta A, Velcheti V, Madabhushi A. Predicting cancer outcomes with radiomics and artificial intelligence in radiology. *Nat Rev Clin Oncol.* (2022) 19:132–46. doi: 10.1038/s41571-021-00560-7
23. Vaidya P, Bera K, Gupta A, Wang X, Corredor G, Fu P, et al. CT derived radiomic score for predicting the added benefit of adjuvant chemotherapy following surgery in stage I, II resectable non-small cell lung cancer: a retrospective multicohort study for outcome prediction. *Lancet Digit Health.* (2020) 2:e116–28. doi: 10.1016/s2589-7500(20)30002-9
24. Wang G, Liu X, Shen J, Wang C, Li Z, Ye L, et al. A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images. *Nat Biomed Eng.* (2021) 5:509–21. doi: 10.1038/s41551-021-00704-1
25. Wang S, Yu H, Gan Y, Wu Z, Li E, Li X, et al. Mining whole-lung information by artificial intelligence for predicting EGFR genotype and targeted therapy response in lung cancer: a multicohort study. *Lancet Digit Health.* (2022) 4:e309–19. doi: 10.1016/S2589-7500(22)00024-3
26. Wang C, Ma J, Shao J, Zhang S, Liu Z, Yu Y, et al. Predicting EGFR and PD-L1 status in NSCLC patients using multitask AI system based on CT images. *Front Immunol.* (2022) 13:813072. doi: 10.3389/fimmu.2022.813072
27. Wang C, Ma J, Shao J, Zhang S, Li J, Yan J, et al. Non-invasive measurement using deep learning algorithm based on multi-source features fusion to predict PD-L1 expression and survival in NSCLC. *Front Immunol.* (2022) 3:828560. doi: 10.3389/fimmu.2022.828560
28. Nachiappan AC, Rahbar K, Shi X, Guy ES, Mortani Barbosa EJ Jr, Shroff GS, et al. Pulmonary tuberculosis: role of radiology in diagnosis and management. *Radiographics.* (2017) 37:52–72. doi: 10.1148/rg.2017160032
29. World Health O. *Global Tuberculosis Report 2015*. Geneva: World Health Organization (2015).
30. El-Solh AA, Hsiao CB, Goodnough S, Serghani J, Grant BJ. Predicting active pulmonary tuberculosis using an artificial neural network. *Chest.* (1999) 116:968–73. doi: 10.1378/chest.116.4.968
31. Lakhani P, Sundaram B. Deep Learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology.* (2017) 284:574–82. doi: 10.1148/radiol.2017162326
32. Hwang EJ, Park S, Jin KN, Kim JI, Choi SY, Lee JH, et al. Development and validation of a deep learning-based automatic detection algorithm for active pulmonary tuberculosis on chest radiographs. *Clin Infect Dis.* (2019) 69:739–47. doi: 10.1093/cid/ciy967
33. Nijiati M, Ma J, Hu C, Tuersun A, Abulizi A, Kelimu A, et al. Artificial intelligence assisting the early detection of active pulmonary tuberculosis from chest X-rays: a population-based study. *Front Mol Biosci.* (2022) 9:874475. doi: 10.3389/fmolb.2022.874475
34. Lee JH, Park S, Hwang EJ, Goo JM, Lee WY, Lee S, et al. Deep learning-based automated detection algorithm for active pulmonary tuberculosis on chest radiographs: diagnostic performance in systematic screening of asymptomatic individuals. *Eur Radiol.* (2021) 31:1069–80. doi: 10.1007/s00330-020-07219-4
35. Heo SJ, Kim Y, Yun S, Lim SS, Kim J, Nam CM, et al. Deep learning algorithms with demographic information help to detect tuberculosis in chest radiographs in annual workers' health examination data. *Int J Res Public Health.* (2019) 16:250. doi: 10.3390/ijerph16020250
36. Nafisah SI, Muhammad G. Tuberculosis detection in chest radiograph using convolutional neural network architecture and explainable artificial intelligence. *Neural Comput Appl.* (2022) 1–21. doi: 10.1007/s00521-022-07258-6 [Epub ahead of print].
37. Pasa F, Golkov V, Pfeiffer F, Cremers D, Pfeiffer D. Efficient deep network architectures for fast chest X-ray tuberculosis screening and visualization. *Sci Rep.* (2019) 9:6268. doi: 10.1038/s41598-019-42557-4
38. Rajaraman S, Zamzmi G, Folio LR, Antani S. Detecting tuberculosis-consistent findings in lateral chest X-rays using an ensemble of CNNs and vision transformers. *Front Genet.* (2022) 13:864724. doi: 10.3389/fgene.2022.864724
39. Rajpurkar P, O'Connell C, Schechter A, Asnani N, Li J, Kiani A, et al. CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *NPJ Digit Med.* (2020) 3:115. doi: 10.1038/s41746-020-00322-2
40. Lee S, Yim JJ, Kwak N, Lee YL, Lee JK, Lee JY, et al. Deep learning to determine the activity of pulmonary tuberculosis on chest radiographs. *Radiology.* (2021) 301:435–42. doi: 10.1148/radiol.2021210063
41. Yan C, Wang L, Lin J, Xu J, Zhang T, Qi J, et al. A fully automatic artificial intelligence-based CT image analysis system for accurate detection, diagnosis, and quantitative severity evaluation of pulmonary tuberculosis. *Eur Radiol.* (2021) 32:2188–99. doi: 10.1007/s00330-021-08365-z
42. The Stop Tb Partnership. *FIND. Resource Center on Computer-Aided Detection Products for the Diagnosis of Tuberculosis*. (2020). Available online at: https://www.ai4hlth.org/ (accessed September 10, 2020).
43. Khan FA, Majidulla A, Tavaziva G, Nazish A, Abidi SK, Benedetti A, et al. Chest x-ray analysis with deep learning-based software as a triage test for pulmonary tuberculosis: a prospective study of diagnostic accuracy for culture-confirmed disease. *Lancet Digit Health.* (2020) 2:e573–81. doi: 10.1016/s2589-7500(20)30221-1
44. Qin ZZ, Sander MS, Rai B, Titahong CN, Sudrungrot S, Laah SN, et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: a multi-site evaluation of the diagnostic accuracy of three deep learning systems. *Sci Rep.* (2019) 9:15000. doi: 10.1038/s41598-019-51503-3
45. Qin ZZ, Ahmed S, Sarker MS, Paul K, Adel ASS, Naheyan T, et al. Tuberculosis detection from chest x-rays for triaging in a high tuberculosis-burden setting: an evaluation of five artificial intelligence algorithms. *Lancet Digit Health.* (2021) 3:e543–54. doi: 10.1016/s2589-7500(21)00116-3
46. Codlin AJ, Dao TP, Vo LNQ, Forse RJ, Van Truong V, Dang HM, et al. Independent evaluation of 12 artificial intelligence solutions for the detection of tuberculosis. *Sci Rep.* (2021) 11:23895. doi: 10.1038/s41598-021-03265-0
47. Melendez J, Sanchez CI, Philipsen RHHM, Maduskar P, Dawson R, Theron G, et al. An automated tuberculosis screening strategy combining X-ray-based computer-aided detection and clinical information. *Sci Rep.* (2016) 6:25265. doi: 10.1038/srep25265
48. Mao Y, Yang D, He J, Krasna MJ. Epidemiology of lung cancer. *Surgl Oncol Clin N Am.* (2016) 25:439–45. doi: 10.1016/j.soc.2016.02.001
49. Starnes SL, Reed MF, Meyer CA, Shipley RT, Jazieh AR, Pina EM, et al. Can lung cancer screening by computed tomography be effective in areas with endemic histoplasmosis? *J Thorac Cardiovasc Surg.* (2011) 141:688–93. doi: 10.1016/j.jtcvs.2010.08.045
50. MacMahon H, Naidich DP, Goo JM, Lee KS, Leung ANC, Mayo JR, et al. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner society 2017. *Radiology.* (2017) 284:228–43. doi: 10.1148/radiol.2017161659
51. Patel VK, Naik SK, Naidich DP, Travis WD, Weingarten JA, Lazzaro R, et al. A practical algorithmic approach to the diagnosis and management of solitary pulmonary nodules: part 2: pretest probability and algorithm. *Chest.* (2013) 143:840–6. doi: 10.1378/chest.12-1487

52. Feng B, Chen X, Chen Y, Lu S, Liu K, Li K, et al. Solitary solid pulmonary nodules: a CT-based deep learning nomogram helps differentiate tuberculosis granulomas from lung adenocarcinomas. *Eur Radiol.* (2020) 30:6497–507. doi: 10.1007/s00330-020-07024-z
53. Zhuo Y, Zhan Y, Zhang Z, Shan F, Shen J, Wang D, et al. Clinical and CT radiomics nomogram for preoperative differentiation of pulmonary adenocarcinoma from tuberculoma in solitary solid sodule. *Front Oncol.* (2021) 11:701598. doi: 10.3389/fonc.2021.701598
54. Hu Y, Zhao X, Zhang J, Han J, Dai M. Value of F-FDG PET/CT radiomic features to distinguish solitary lung adenocarcinoma from tuberculosis. *Eur J Nucl Med Mol Imaging.* (2021) 48:231–40. doi: 10.1007/s00259-020-04924-6
55. Du D, Gu J, Chen X, Lv W, Feng Q, Rahmim A, et al. Integration of PET/CT radiomics and semantic features for differentiation between active pulmonary tuberculosis and lung cancer. *Mol Imaging Biol.* (2021) 23:287–98. doi: 10.1007/s11307-020-01550-4
56. Lee H, Myung W, Koh WJ, Moon SM, Jhun BW. Epidemiology of nontuberculous mycobacterial infection, South Korea, 2007-2016. *Emerg Infect Dis.* (2019) 25:569–72. doi: 10.3201/eid2503.181597
57. Kendall BA, Winthrop KL. Update on the epidemiology of pulmonary nontuberculous mycobacterial infections. *Semin Respir Crit Care Med.* (2013) 34:87–94. doi: 10.1055/s-0033-1333567
58. Kwak N, Lee CH, Lee HJ, Kang YA, Lee JH, Han SK, et al. Non-tuberculous mycobacterial lung disease: diagnosis based on computed tomography of the chest. *Eur Radiol.* (2016) 26:4449–56. doi: 10.1007/s00330-016-4286-6
59. Wang L, Ding W, Mo Y, Shi D, Zhang S, Zhong L, et al. Distinguishing nontuberculous mycobacteria from *Mycobacterium tuberculosis* lung disease from CT images using a deep learning framework. *Eur J Nucl Med Mol Imaging.* (2021) 48:4293–306. doi: 10.1007/s00259-021-05432-x
60. Yan Q, Wang W, Zhao W, Zuo L, Wang D, Chai X, et al. Differentiating nontuberculous mycobacterium pulmonary disease from pulmonary tuberculosis through the analysis of the cavity features in CT images using radiomics. *BMC Pulm Med.* (2022) 22:4. doi: 10.1186/s12890-021-01766-2
61. Icksan AG, Napitupulu MRS, Nawas MA, Nurwidya F. Chest X-ray findings comparison between multi-drug-resistant tuberculosis and drug-sensitive tuberculosis. *J Nat Sci Biol Med.* (2018) 9:42–6. doi: 10.4103/jnsbm.JNSBM_79_17
62. Wang YXJ, Chung MJ, Skrahin A, Rosenthal A, Gabrielian A, Tartakovsky M. Radiological signs associated with pulmonary multi-drug resistant tuberculosis: an analysis of published evidences. *Quant Imaging Med Surg.* (2018) 8:161–73. doi: 10.21037/qims.2018.03.06
63. Jaeger S, Juarez-Espinosa OH, Candemir S, Poostchi M, Yang F, Kim L, et al. Detecting drug-resistant tuberculosis in chest radiographs. *Int J Comput Assist Radiol Surg.* (2018) 13:1915–25. doi: 10.1007/s11548-018-1857-9
64. Karki M, Kantipudi K, Yu H, Yang F, Kassim YM, Yaniv Z, et al. Identifying drug-resistant tuberculosis in chest radiographs: evaluation of CNN architectures and training strategies. In: *Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC).* Mexico (2021) 2964–7. doi: 10.1109/EMBC46164.2021.9630189
65. Gao XW, Qian Y. Prediction of multidrug-resistant TB from CT pulmonary images based on deep learning techniques. *Mol Pharm.* (2018) 15:4326–35. doi: 10.1021/acs.molpharmaceut.7b00875
66. Cohen KA, Manson AL, Desjardins CA, Abeel T, Earl AM. Deciphering drug resistance in *Mycobacterium tuberculosis* using whole-genome sequencing: progress, promise, and challenges. *Genome Med.* (2019) 11:45. doi: 10.1186/s13073-019-0660-8
67. Arango-Argoty G, Garner E, Pruden A, Heath LS, Vikesland P, Zhang L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome.* (2018) 6:23. doi: 10.1186/s40168-018-0401-z
68. Yang Y, Walker TM, Walker AS, Wilson DJ, Peto TEA, Crook DW, et al. DeepAMR for predicting co-occurrent resistance of *Mycobacterium tuberculosis*. *Bioinformatics.* (2019) 35:3240–9. doi: 10.1093/bioinformatics/btz067
69. Yang Y, Walker TM, Kouchaki S, Wang C, Peto TEA, Crook DW, et al. An end-to-end heterogeneous graph attention network for *Mycobacterium tuberculosis* drug-resistance prediction. *Brief Bioinform.* (2021) 22:bbab299. doi: 10.1093/bib/bbab299
70. Yang Y, Niehaus KE, Walker TM, Iqbal Z, Walker AS, Wilson DJ, et al. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics.* (2018) 34:1666–71. doi: 10.1093/bioinformatics/btx801
71. Deelder W, Christakoudi S, Phelan J, Benavente ED, Campino S, McNerney R, et al. Machine learning predicts accurately *Mycobacterium tuberculosis* drug resistance from whole genome sequencing data. *Front Genet.* (2019) 10:992. doi: 10.3389/fgene.2019.00922
72. Chen ML, Doddi A, Royer J, Freschi L, Schito M, Ezewudo M, et al. Beyond multidrug resistance: leveraging rare variants with machine and statistical learning models in *Mycobacterium tuberculosis* resistance prediction. *EBioMedicine.* (2019) 43:356–69. doi: 10.1016/j.ebiom.2019.04.016
73. Gröschel MI, Owens M, Freschi L, Vargas R Jr, Marin MG, Phelan J, et al. GenTB: a user-friendly genome-based predictor for tuberculosis resistance powered by machine learning. *Genome Med.* (2021) 13:138. doi: 10.1186/s13073-021-00953-4
74. Farhat MR, Sultana R, Iartchouk O, Bozeman S, Galagan J, Sisk P, et al. Genetic determinants of drug resistance in Mycobacterium tuberculosis and their diagnostic value. *Am J Respir Crit Care Med.* (2016) 194:621–30. doi: 10.1164/rccm.201510-2091OC
75. Kuang X, Wang F, Hernandez KM, Zhang Z, Grossman RL. Accurate and rapid prediction of tuberculosis drug resistance from genome sequence data using traditional machine learning algorithms and CNN. *Sci Rep.* (2022) 12:2427. doi: 10.1038/s41598-022-06449-4
76. Jiang Z, Lu Y, Liu Z, Wu W, Xu X, Dinnyés A, et al. Drug resistance prediction and resistance genes identification in Mycobacterium tuberculosis based on a hierarchical attentive neural network utilizing genome-wide variants. *Brief Bioinform.* (2022) 23:bbac041. doi: 10.1093/bib/bbac041
77. Higashiguchi M, Nishioka K, Kimura H, Matsumoto T. Prediction of the duration needed to achieve culture negativity in patients with active pulmonary tuberculosis using convolutional neural networks and chest radiography. *Respir Investig.* (2021) 59:421–7. doi: 10.1016/j.resinv.2021.01.004
78. Colangeli R, Jedrey H, Kim S, Connell R, Ma S, Chippada Venkata UD, et al. Bacterial factors that predict relapse after tuberculosis therapy. *N Engl J Med.* (2018) 379:823–33. doi: 10.1056/NEJMoa1715849
79. Ahmad Khan F, Gelmanova I, Franke M, Atwood S, Zemlyanaya N, Unakova I, et al. Aggressive regimens reduce risk of recurrence after successful treatment of MDR-TB. *Clin Infect Dis.* (2016) 63:214–20. doi: 10.1093/cid/ciw276
80. Lu MT, Raghu VK, Mayrhofer T, Aerts HJWL, Hoffmann U. Deep learning using chest radiographs to identify high-risk smokers for lung cancer screening computed tomography: development and validation of a prediction model. *Ann Intern Med.* (2020) 173:704–13. doi: 10.7326/m20-1868

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.

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 29 July 2022
DOI 10.3389/fmed.2022.950327

OPEN ACCESS

EDITED BY
Md. Mohaimenul Islam,
Aesop Technology, Taiwan

REVIEWED BY
Diego Pedro Pinto-Roa,
National University of Asunción, Paraguay
Michal Jasinski,
Wrocław University of Science and Technology, Poland

*CORRESPONDENCE
Changsop Yang
yangunja@kiom.re.kr
Jae-Dong Lee
ljdkhu@gmail.com

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

SPECIALTY SECTION
This article was submitted to Family Medicine and Primary Care, a section of the journal Frontiers in Medicine

RECEIVED 22 May 2022
ACCEPTED 11 July 2022
PUBLISHED 29 July 2022

CITATION
Lee H, Choi Y, Son B, Lim J, Lee S, Kang JW, Kim KH, Kim EJ, Yang C and Lee J-D (2022) Deep autoencoder-powered pattern identification of sleep disturbance using multi-site cross-sectional survey data. *Front. Med.* 9:950327. doi: 10.3389/fmed.2022.950327

Deep autoencoder-powered pattern identification of sleep disturbance using multi-site cross-sectional survey data

Hyeonhoon Lee[1], Yujin Choi[2], Byunwoo Son[3], Jinwoong Lim[1,4], Seunghoon Lee[5], Jung Won Kang[5], Kun Hyung Kim[6], Eun Jung Kim[7], Changsop Yang[2*†] and Jae-Dong Lee[5*†]

[1]Department of Clinical Korean Medicine, Graduate School, Kyung Hee University, Seoul, South Korea, [2]KM Science Research Division, Korea Institute of Oriental Medicine, Daejeon, South Korea, [3]Department of Korean Medicine, Combined Dispensary, 7th Corps, Republic of Korea Army, Icheon-si, South Korea, [4]Department of Acupuncture and Moxibustion, Wonkwang University Gwangju Korean Medicine Hospital, Gwangju, South Korea, [5]Department of Acupuncture and Moxibustion, College of Korean Medicine, Kyung Hee University, Seoul, South Korea, [6]School of Korean Medicine, Pusan National University, Yangsan, South Korea, [7]Department of Acupuncture and Moxibustion Medicine, Dongguk University Bundang Oriental Hospital, Seongnam-si, South Korea

Pattern identification (PI) is a diagnostic method used in Traditional East Asian medicine (TEAM) to select appropriate and personalized acupuncture points and herbal medicines for individual patients. Developing a reproducible PI model using clinical information is important as it would reflect the actual clinical setting and improve the effectiveness of TEAM treatment. In this paper, we suggest a novel deep learning-based PI model with feature extraction using a deep autoencoder and *k*-means clustering through a cross-sectional study of sleep disturbance patient data. The data were obtained from an anonymous electronic survey in the Republic of Korea Army (ROKA) members from August 16, 2021, to September 20, 2021. The survey instrument consisted of six sections: demographics, medical history, military duty, sleep-related assessments (Pittsburgh sleep quality index (PSQI), Berlin questionnaire, and sleeping environment), diet/nutrition-related assessments [dietary habit survey questionnaire and nutrition quotient (NQ)], and gastrointestinal-related assessments [gastrointestinal symptom rating scale (GSRS) and Bristol stool scale]. Principal component analysis (PCA) and a deep autoencoder were used to extract features, which were then clustered using the *k*-means clustering method. The Calinski-Harabasz index, silhouette coefficient, and within-cluster sum of squares were used for internal cluster validation and the final PSQI, Berlin questionnaire, GSRS, and NQ scores were used for external cluster validation. One-way analysis of variance followed by the Tukey test and chi-squared test were used for between-cluster comparisons. Among 4,869 survey responders, 2,579 patients with sleep disturbances were obtained after filtering using a PSQI score of >5. When comparing clustering performance using raw data and extracted features by PCA and the deep autoencoder, the best feature extraction method for clustering was the deep autoencoder (16 nodes for the first and third hidden layers, and two nodes for the second

hidden layer). Our model could cluster three different PI types because the optimal number of clusters was determined to be three *via* the elbow method. After external cluster validation, three PI types were differentiated by changes in sleep quality, dietary habits, and concomitant gastrointestinal symptoms. This model may be applied to the development of artificial intelligence-based clinical decision support systems through electronic medical records and clinical trial protocols for evaluating the effectiveness of TEAM treatment.

KEYWORDS

deep autoencoder, deep learning, pattern identification, clustering, sleep

Introduction

Pattern identification (PI), a diagnostic method in Traditional East Asian medicine (TEAM), is a meaningful step for TEAM doctors when making treatment decisions such as selection of an appropriate acupuncture point and herbal medicine. It uses clinical information based on traditional diagnostic criteria, which include observation, listening, questioning, and pulse detection (1). Particularly, the use of PI in selecting an optimal combination with a few acupuncture points has been an important research subject to reveal those used in actual clinical practice (2, 3). Most clinical trials on the effectiveness of acupuncture treatment used a fixed-point approach, which is different from the clinical practice that uses a more individualized approach (4). Although some study designs such as conventional randomized clinical trials (RCTs) with a personalized acupuncture protocol or a pragmatic clinical trial have been suggested to overcome the gap between acupuncture research and clinical practice, the results of an individualized approach vs. a fixed-point approach are still controversial (5–9). Nonetheless, several recent experimental studies have supported the significance of acupuncture point selection (10–13). Therefore some studies with data-mining methods were conducted using RCT data, medical records, virtual diagnosis data, and classical medical texts to systematically prove the relationship between symptoms, diseases, PI, and acupuncture point selections (3, 14–16).

Artificial intelligence (AI) techniques have also emerged in the research of TEAM. Previous studies used artificial neural network models to differentiate patterns for acupuncture point selections (17, 18), and clustering algorithms to discover the combination rules of herbal medicine (19). Also, the recent deep learning models such as bidirectional encoder representations from transformers generated some new herbal medicine prescriptions from a few medical records (20, 21). However, to the best of our knowledge, there are few AI studies that assist PI from large amounts of clinical information, though most clinical guidelines recommend a PI process by a TEAM doctor prior to providing acupuncture or herbal medicine treatment (22, 23).

With the appropriate PI, a wide variety of conditions can be addressed by TEAM treatment. Sleep disturbances were one of the major target conditions for TEAM treatment in several previous studies (19, 24–28). The Korean Medicine Clinical Practice Guidelines for insomnia disorder, which was officially developed by research funded by the government, suggest that TEAM doctors may consider six types of PI before TEAM treatment (29). Furthermore, a recent systematic review for insomnia showed that acupuncture treatment using PI significantly improved the total effectiveness rate compared to conventional medication (30). However, the effect of TEAM treatment using PI is not reproducible since PI types and processes may be inconsistent among TEAM doctors in clinical settings. Therefore, the development of a model that can consistently produce the same PI for certain patient details required.

In this paper, we suggested a novel data-driven PI method for TEAM treatment using emerging bioinformatics techniques in combination with feature extraction using a deep autoencoder, one of the self-supervised deep learning models, and clustering using *k*-means clustering, an unsupervised machine learning model. To develop a new model using various types of clinical information as input data and provide reproducible PI as an output for TEAM treatment decisions in patients with sleep disturbances, we used cross-sectional study data which examined the association between sleep and diet/digestion in Republic of Korea Army (ROKA) active duty service members.

Materials and methods

Study population

A multi-site cross-sectional study was conducted using an anonymous electronic survey. The study was posted in five units of the ROKA through printed recruitment posters and electronic

bulletin boards from August 16, 2021, to September 20, 2021. The participants were recruited during the same period. The original aim of this study was to examine the association between sleep and diet/digestion in ROKA active duty service members. The results will be published in another paper.

Among active duty service members in five units of the ROKA who met the inclusion criteria, the participants who provided informed consent were enrolled in the study. The inclusion criteria were (1) age 19 years or over; (2) active duty service members (private, private first class, corporal, and sergeants) who completed the basic military training course, and (3) those who voluntarily agreed to participate in the study. There were no exclusion criteria.

Sample size calculation for cross-sectional study

Assuming that the total number of all active duty service members in the ROKA is approximately 300,000, the sample size was calculated using the following equation. The margin of error was 3% and the confidence level was 95%, and the sample size result was 1,064 (the target number of completed surveys).

$$Sample\ size = \frac{\frac{z^2 \times p(1-p)}{e^2}}{1 + (\frac{z^2 \times p(1-p)}{e^2 N})}$$
$$(N = number\ in\ the\ populations;$$
$$e = margin\ of\ error;\ z = Z - score)$$

Previous studies using surveys showed that various factors such as the survey method, survey content, and participant compensation were associated with the response rate of the study subjects. In particular, the online survey method is known to have about a 10% lower response rate compared to other media, but the actual response rate was different in each study (31). In this study, referring to the response rate (3.4%) reported in a previous study that conducted a health-related survey in adult males, the response rate was set to 3%, and the target number of questionnaires was determined to be 35,467 (31).

Survey instrument of cross-sectional study

The survey instruments were refined to reveal the military environment by healthcare professionals (seven TEAM doctors including five military doctors). This involved the refinement of the questionnaire by changing the phrasing and modifying questions to clarify the premise of each item within the questionnaire. The final questionnaire was designed and distributed through the web-based application Survey Monkey.

The survey consisted of six sections: (1) demographics (birth, recruitment date, height, weight, military identification number, rank, military unit, education, smoking status, alcohol consumption habits, caffeine consumption, exercise, and physical grade); (2) medical history (present/past history of sleep disorders, present/past history of gastrointestinal disorders, present/past history of general diseases including hypertension, diabetes, hyperlipidemia, and cardiac disease, stress status, and drug history); (3) military duty (branch, position, night shift with or without tomorrow duty-off, and its effect on sleep and/or fatigue); (4) sleep-related assessments (Pittsburgh sleep quality index (PSQI), Berlin questionnaire, and sleeping environment); (5) diet/nutrition-related assessments [dietary habit survey questionnaire and nutrition (32) quotient (NQ)]; and (6) gastrointestinal-related assessments [gastrointestinal symptom rating scale (GSRS) and Bristol stool scale (BSS)].

The PSQI, a self-assessment questionnaire to evaluate sleep quality within the past month, contains 19 items consisting of seven component scores, including sleep quality, sleep latency, sleep duration, daytime dysfunction, sleep efficiency, sleep disturbances, and sleeping medication use (33). A final score of >5 out of 21 indicates significant sleep disturbance.

The Berlin questionnaire has 11 questions grouped into three categories (34). The first category comprises five questions concerning snoring, witnessed apnea, and the frequency of such events. The second category comprises four questions addressing daytime sleepiness, with a sub-question on drowsy driving. The third category comprises two questions concerning a history of high blood pressure (> 140/90 mmHg) and a body mass index (BMI) of >30 kg/m^2. Categories 1 and 2 were considered positive if there were two positive responses in each category, while category 3 was considered positive with a self-report of high blood pressure and/or a BMI of > 30 kg/m^2. The study patients were scored as being at high risk of having obstructive sleep apnea (OSA) if the scores were positive for two or more of the three categories.

The dietary habit survey questionnaire consists of 25 items to evaluate the dietary habits of Korean adults (35). It includes the number of meals per day, mealtime regularity, the amount consumed, time taken for a meal, the frequency of missed meals, the frequency of having breakfast, the reason for missing breakfast, the frequency of dinners with family, the frequency of overeating, meal at which overeating occurred (breakfast, lunch, dinner or not), the frequency of eating out, the frequency of eating snacks, the time of eating snacks, the types of snacks, the time of late-night meals, whether certain foods were not eaten, the reasons for not eating certain foods, and the frequency of food intake (grains, meat, fish, eggs and legumes, fruits, vegetables, milk and dairy products, fatty foods, instant foods, and fast foods).

The NQ comprehensively evaluates the nutritional status and meal quality of individuals or groups of Korean adults through a checklist consisting of 21 items (36). It provides

the global NQ score (NQ global), and scores for four factors: nutritional balance (NQ balance), food diversity (NQ diversity), moderation in the amount of food eaten (NQ moderation), and dietary behavior (NQ behavior). It is considered "good" if the score is 58 or higher, and "monitoring is necessary" if it is below 58.

The GSRS evaluates gastrointestinal symptoms *via* an inquiry table consisting of 15 items for the evaluation of general gastrointestinal symptoms (37). Each GSRS item is rated on a 7-point Likert scale ranging from "no discomfort" to "very severe discomfort."

The BSS examines the stool status in the past 24 h (32). The score is based on a one to seven scale where one corresponds to a hard stool and seven corresponds to watery diarrhea.

Data preprocessing

Data preprocessing to improve data quality and impute missing values was performed in three steps. In the first step, from all survey responders, the participants who provided multiple responses were eliminated to ensure survey reliability. Second, the participants who did not meet the inclusion criteria were removed. The participants who completed the survey remained. Last, a few samples with outliers, which might be caused by miswriting in open question items such as height, weight, smoking amount, and smoking duration were also eliminated after exploratory data analysis (EDA).

Each PSQI, Berlin questionnaire, NQ, and GSRS score was calculated and the remaining questionnaire responses were used for input data. In clinical practice, TEAM doctors' questions to patients are closer to each item of the questionnaire, and conversely, calculating each questionnaire's scores one by one is closer to the purpose of the clinical study. The calculated scores were used for external cluster evaluation.

Since this study was conducted to examine patients with sleep disturbances, the participants with PSQI scores of over five were collected as a total data set. Then, the data set was randomly split into a training set (80%) and test set (20%) for evaluating the machine learning models.

Feature extraction

The autoencoder is a simple unsupervised learning model. It learns hidden features through encoding and decoding unlabeled data. Consider a d-dimensional data set $X = \{x_1, x_2, \ldots, x_d\}$, where d is the number of variables presented at the input layer. The autoencoder attempts to reconstruct X at the output layer, which is the same as the identity function $f(x) = x$ (38). Then, the hidden layer is forced to learn a compressed representation of the data X from the input layer, which is reconstructed at the output layer as $\hat{X}$. The optimized model can be evaluated by the root mean squared error (RMSE) between X and $\hat{X}$.

In this study, we built a symmetric deep autoencoder model composed of d-dimensional input and output layers, and three hidden layers: J nodes for the second hidden layer (bottleneck), and $8 \times J$ nodes for the first and third hidden layers. Also, a grid search using $1 \leq J \leq 10$ was conducted to find the optimal number of nodes in the hidden layers. When compiling the model, RMSE and Adam were applied as the loss function and training optimizer, respectively. For the training process with 10-fold cross-validation, the batch size and the number of epochs were set to 64 and 100, respectively. Finally, representative nodes in the second hidden layer were used to extract features for the clustering process. We also conducted principal component analysis (PCA), one of the conventional feature extraction methods, before k-means clustering.

K-means clustering

K-means clustering is an unsupervised machine learning algorithm (39). This algorithm is less computationally intensive for processing our large study data than hierarchical clustering. Also, the number of clusters (k) can be predefined by this algorithm to reveal our prior medical knowledge since the number of PI types is generally $\leq$10 in TEAM. Consider a d-dimensional data set $X = \{x_1, x_2, \ldots, x_n\}$, where d is the number of variables, this algorithm aims to partition the n observations into k ($\leq n$) sets $S = \{S_1, S_2, \ldots, S_k\}$ to minimize the within-cluster sum of squares (WCSS). Formally, the objective is to find:

$$\arg\min_{s} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$
$$(\mu_i = mean\ of\ points\ in\ S_i)$$

In this study, k-means clustering was performed on the data set using raw data and PCA-extracted and deep autoencoder-extracted features. The performance of the clusters was compared between each input type. We set the candidate number of clusters from $k = 1$ to $k = 10$, and 300 iterations for each k using the expectation-maximization style algorithm.

Cluster evaluation

Cluster evaluation was conducted in two parts, internal cluster evaluation and external cluster evaluation. The Calinski-Harabasz index and silhouette coefficient were initially assessed for internal cluster evaluation (40). The optimal number of clusters was determined by the elbow method after plotting the WCSS with k values. All the above processes were conducted using the training set only. After determining the whole PI model including the feature extraction and clustering methods,

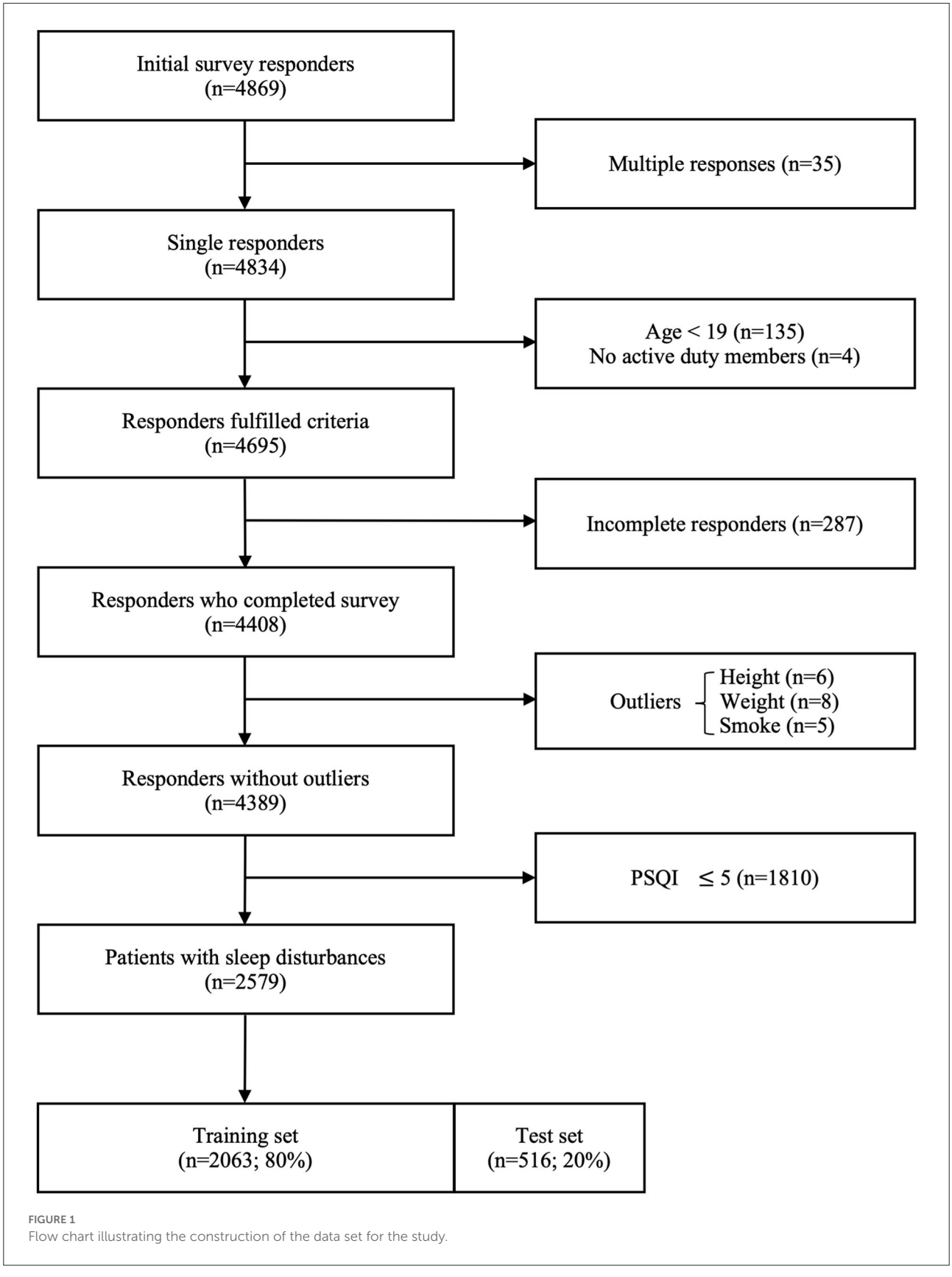

FIGURE 1
Flow chart illustrating the construction of the data set for the study.

the test set was inferred by the trained PI model. The PSQI, Berlin questionnaire, GSRS, and NQ scores, which were not used in feature extraction, were compared by external cluster evaluation.

Statistical analysis

Summaries of the continuous variables are presented as means and standard deviations, and the categorical variables are presented as frequencies and percentages. For continuous variables, one-way analysis of variance (ANOVA) was used for comparing means among three clusters, followed by the Tukey-Kramer test for *post-hoc* multiple comparisons between two clusters with unequal sample sizes. For categorical variables, the chi-squared test was also performed. Statistical significance was set at $p < 0.05$.

Tools

Python 3.8.0 (Python Software Foundation, Wilmington, DE, USA) was used for data preprocessing, model development and validation, visualization, and statistical analysis. The Python libraries Pandas and Numpy were adopted for data preprocessing; Scikit-learn was used for data preprocessing, PCA, and *k*-means clustering; Keras with Tensorflow backend for building and evaluating the deep autoencoder model; Statsmodels for statistical analysis of comparisons between clusters, and Seaborn with Matplotlib for data visualization. Google Colab, a cloud service for machine learning research, was used in this study. It provides various libraries and frameworks for deep learning and a robust graphics processing unit.

Results

Data set construction

Of a total of 4,869 survey responders, 35 multiple responders, and 139 responders who did not meet the inclusion criteria were excluded. A total of 4,408 responders completed the survey. After removing a few outliers for height (below 110 cm or above 200 cm), weight (below 40 kg or above 160 kg), smoking amount (above five packs per day), and smoking duration (above 20 years) through EDA, 4,389 responses remained. The data set of 2,579 patients with sleep disturbances was obtained after filtering by PSQI scores of >5, which were randomly split into a training

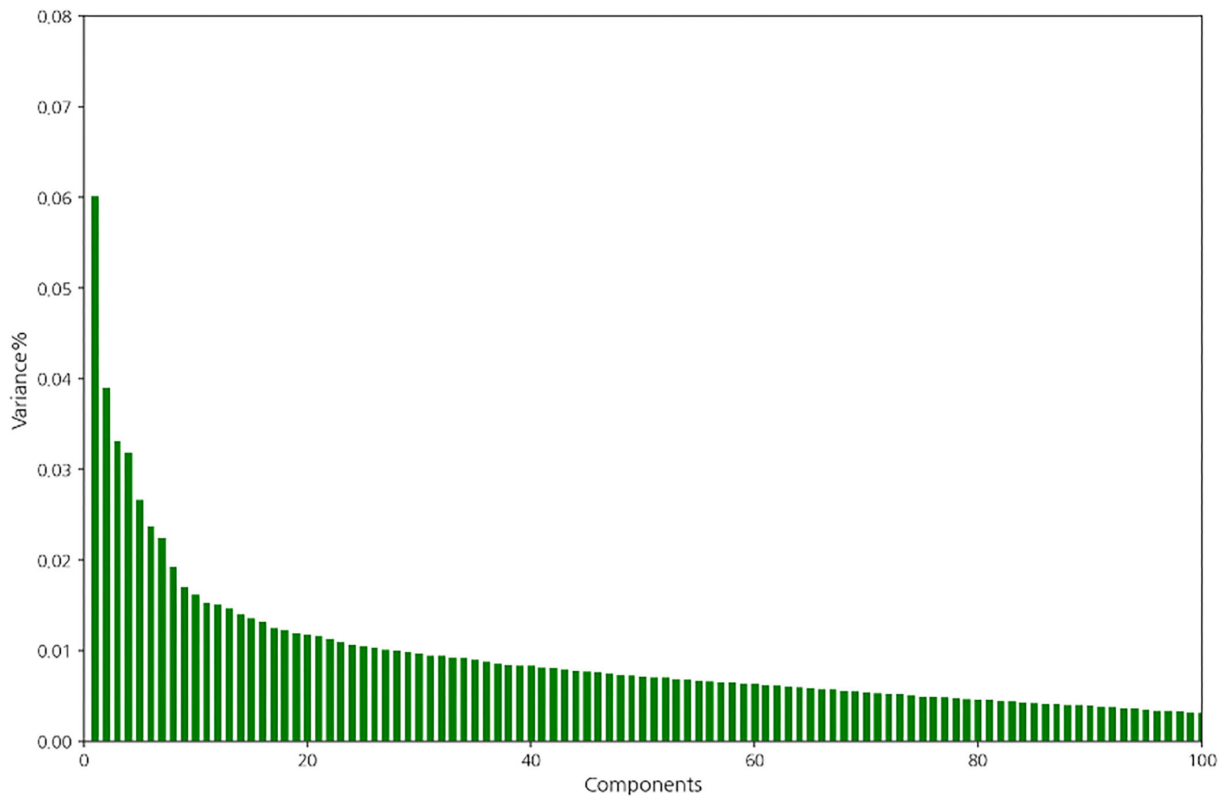

FIGURE 2
Variance of the components in the training set.

set (n = 2,063; 80%) and a test set (n = 516; 20%). The flow chart of the data set construction process is shown in Figure 1.

Feature extraction using PCA

For comparison with the main feature extraction method, the deep autoencoder, PCA was first conducted using the training set. It showed that variance dropped off when the number of components was four, and the first four components explained the majority of the variance in the training set (Figure 2). Therefore, feature extraction using PCA was conducted with four components.

Feature extraction using deep autoencoder

Ten-fold cross-validation was conducted while training the deep autoencoder. The mean RMSE of the training set and validation set after 100 epochs (Table 1), and the change in RMSE of the validation set during training (Figure 3) are presented in each deep autoencoder architecture (the number of nodes in the second hidden layer).

Internal cluster validation

The Calinski-Harabasz index and silhouette coefficient after k-means clustering ($2 \leq k \leq 10$) are presented in Figure 4; Supplementary Table 1. The performance of clustering after feature extraction with the deep autoencoder was much better than that with raw data or PCA. Comparing the results of clustering after all feature extraction methods including PCA and the deep autoencoder in this study, the deep autoencoder (J = 2)—which presented the highest values of both the Calinski-Harabasz index and the silhouette coefficient in the small numbers of clusters ($k \leq 4$)–might be the best feature extraction method for k-means clustering. The final deep autoencoder model architecture is shown in Figure 5. Also, considering both the Calinski-Harabasz index and the silhouette coefficient, k = 2 or 3 might be candidate clustering numbers. Finally, the optimal number (k = 3) of clusters was determined by the elbow method, a heuristic approach for determining the appropriate point for the local optimum (41, 42), as shown in Figure 6.

External cluster validation

The patient characteristics in each cluster of the training set and test set are presented in Tables 2, 3 respectively. Among the clusters, the PSQI ($p < 0.001$), GSRS ($p < 0.001$), NQ balance ($p = 0.008$), NQ moderation ($p < 0.001$), NQ behavior ($p < 0.001$), and Berlin scores ($p < 0.001$) were significantly different in the training set, and PSQI ($p < 0.001$), GSRS ($p < 0.001$), NQ global ($p < 0.001$), NQ moderation ($p < 0.001$), and Berlin scores ($p < 0.001$) were significantly different in the test set (Table 4).

TABLE 1 The mean RMSE for each model.

The number of nodes in the second hidden layer (J)	RMSE	
	Training set	Validation set
1	0.820 ± 0.004	0.821 ± 0.013
2	0.796 ± 0.002	0.802 ± 0.014
3	0.776 ± 0.002	0.787 ± 0.014
4	0.760 ± 0.002	0.774 ± 0.012
5	0.745 ± 0.003	0.763 ± 0.013
6	0.732 ± 0.003	0.752 ± 0.013
7	0.719 ± 0.003	0.744 ± 0.012
8	0.708 ± 0.004	0.737 ± 0.012
9	0.696 ± 0.004	0.732 ± 0.011
10	0.686 ± 0.004	0.725 ± 0.012

Values are presented as the mean ± standard deviation.
RMSE, root mean squared error.

Through *post-hoc* analysis (Table 4), the mean PSQI score of cluster A was significantly lower than that of cluster B (−3.24, 95% confidence interval (CI) −3.84, −2.64], $p < 0.001$) and cluster C (−1.62, 95% CI [−1.91, −1.34], $p < 0.001$) in the training set. The mean PSQI score of cluster B was significantly higher than that of cluster C (1.62, 95% CI [0.99, 2.24], $p < 0.001$) in the training set. The mean PSQI score of cluster A was significantly lower than that of cluster B (−2.78, 95% CI [−4.02, −1.53], $p < 0.001$) and cluster C (−1.68, 95% CI [−2.24, −1.12], $p < 0.001$) in the test set. The mean PSQI score of cluster B was also higher but not significantly different than that of cluster C (1.09, 95% CI [−0.20, 2.39], $p = 0.117$) in the test set.

The mean GSRS score of cluster A was significantly lower than that of cluster B (−15.76, 95% CI [−16.55, −14.98], $p < 0.001$) and cluster C (−4.21, 95% CI [−4.59, −3.84], $p < 0.001$) in the training set. The mean GSRS score of cluster B was significantly higher than that of cluster C (11.55, 95% CI [10.73, 12.37], $p < 0.001$) in the training set. The mean GSRS score of cluster A was significantly lower than that of cluster B (−17.69, 95% CI [−19.28, −16.10], $p < 0.001$) and cluster C (−4.44, 95% CI [−5.16, −3.72], $p < 0.001$) in the test set. The mean GSRS score of cluster B was significantly higher than that of cluster C (13.25, 95% CI [11.60, 14.91], $p < 0.001$) in the test set.

The mean NQ global score of cluster A was lower but not statistically different from that of cluster B (−1.27, 95% CI [−3.59, 1.05], $p = 0.404$) and cluster C (−0.60, 95% CI [−1.71, 0.50], $p = 0.406$) in the training set. The mean NQ global score

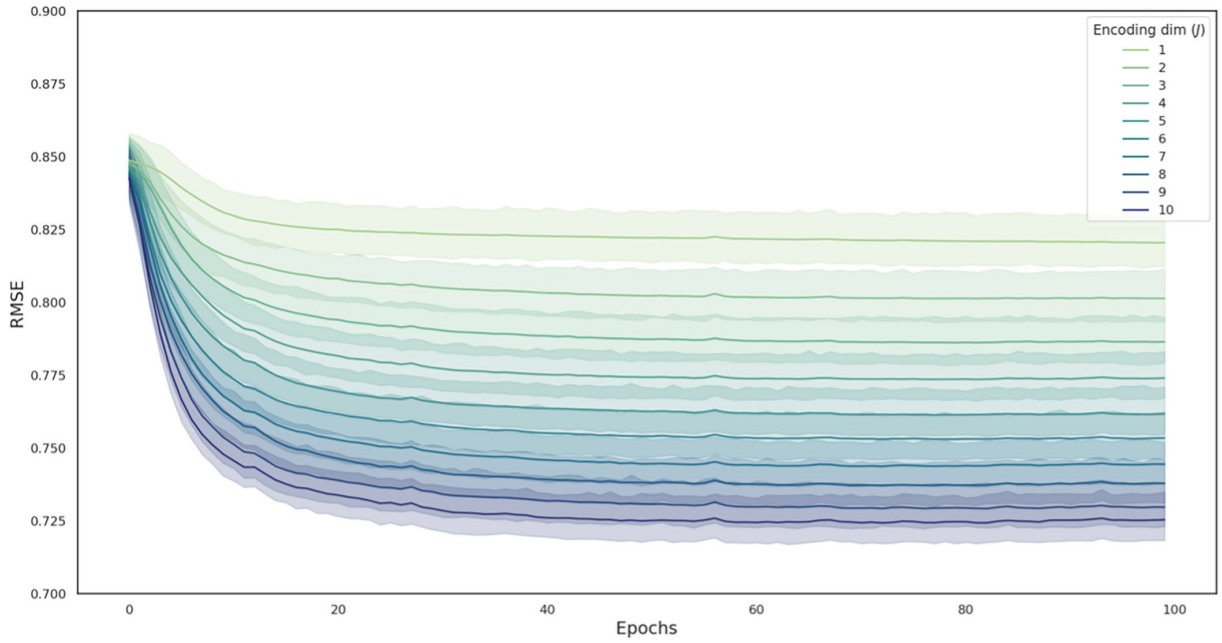

FIGURE 3
The change in RMSE during the model training process in each deep autoencoder architecture. Curves are averaged over 10 folds, with the shaded area representing the 95% confidence interval across folds. RMSE, root mean squared error.

of cluster B was higher but not significantly different than that of cluster C (0.67, 95% CI [−1.76, 3.09], $p = 0.774$) in the training set. The mean NQ global score of cluster A was lower but not statistically different than that of cluster B (−4.41, 95% CI [−9.31, 0.50], $p = 0.089$) and significantly lower than that of cluster C (−3.52, 95% CI [−5.74, −1.30], $p < 0.001$) in the test set. The mean NQ global score of cluster B was higher but not significantly different than that of cluster C (0.88, 95% CI [−4.23, 6.00], $p = 0.900$) in the test set.

The mean NQ balance score of cluster A was higher but not statistically different than that of cluster B (2.18, 95% CI [−1.51, 5.88], $p = 0.349$) and significantly higher than that of cluster C (2.20, 95% CI [0.44, 3.96], $p = 0.010$) in the training set. The mean NQ balance score of cluster B was higher but not significantly different than that of cluster C (0.02, 95% CI [−3.85, 3.88], $p = 0.900$) in the training set. The mean NQ balance score of cluster A was lower but not statistically different than that of cluster B (−3.66, 95% CI [−11.39, 4.08], $p = 0.508$) and cluster C (−1.42, 95% CI [−4.92, 2.08], $p = 0.599$) in the test set. The mean NQ balance score of cluster B was higher but not significantly different than that of cluster C (2.24, 95% CI [−5.82, 10.30], $p = 0.770$) in the test set.

The mean NQ diversity score of cluster A was higher but not statistically different than that of cluster B (0.73, 95% CI [−0.33, 1.79], $p = 0.240$) and lower but not statistically different than that of cluster C (−0.16, 95% CI [−0.66, 3.48], $p = 0.728$) in the training set. The mean NQ diversity score of cluster B was lower but not significantly different than that of cluster C (−0.89, 95% CI [−1.99, 0.22], $p = 0.147$) in the training set. The mean NQ diversity score of cluster A was higher but not statistically different than that of cluster B (0.72, 95% CI [−1.43, 2.86], $p = 0.694$) and lower but not statistically different than that of cluster C (−0.54, 95% CI [−1.51, 0.43], $p = 0.386$) in the test set. The mean NQ diversity score of cluster B was lower but not significantly different than that of cluster C (−1.26, 95% CI [−3.49, 0.97], $p = 0.382$) in the test set.

The mean NQ moderation score of cluster A was significantly lower than that of cluster B (−4.25, 95% CI [−5.35, −3.15], $p < 0.001$) and cluster C (−2.48, 95% CI [−3.00, −1.95], $p < 0.001$) in the training set. The mean NQ moderation score of cluster B was significantly higher than that of cluster C (1.77, 95% CI [0.62, 2.92], $p < 0.001$) in the training set. The mean NQ moderation score of cluster A was significantly lower than that of cluster B (−4.89, 95% CI [−7.13, −2.64], $p < 0.001$) and cluster C (−3.28, 95% CI [−4.29, −2.26], $p < 0.001$) in the test set. The mean NQ moderation score of cluster B was also higher but not significantly different than that of cluster C (1.61, 95% CI [−0.73, 3.95], $p = 0.238$) in the test set.

The mean NQ behavior score of cluster A was significantly higher than that of cluster B (1.72, 95% CI [0.85, 2.58], $p < 0.001$) and cluster C (1.49, 95% CI [1.08, 1.90], $p < 0.001$) in the training set. The mean NQ behavior score of cluster B was lower but not significantly different than that of cluster C (−0.23, 95% CI [−1.13, 0.68], $p = 0.806$) in the training set. The mean

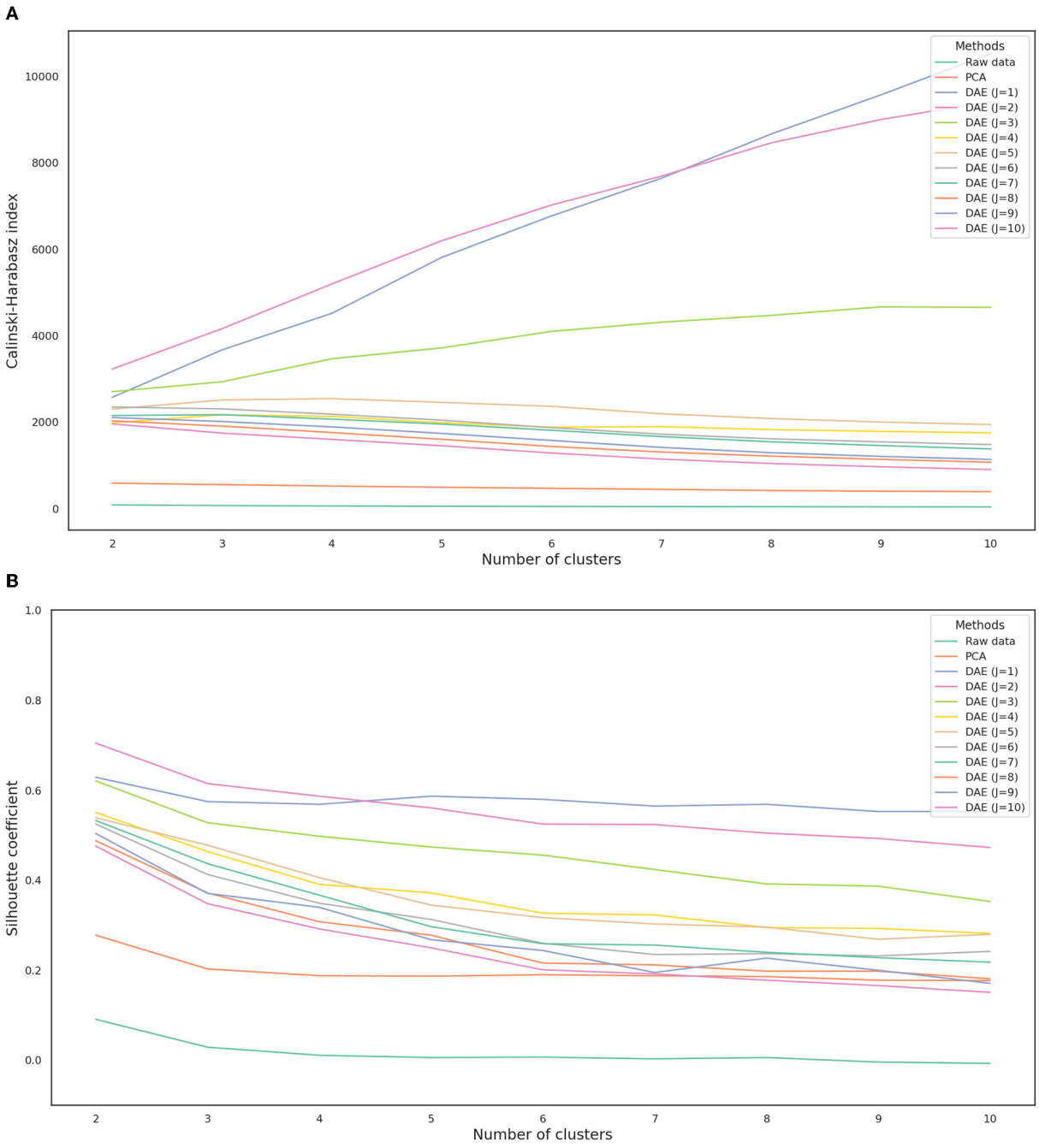

FIGURE 4
Calinski-Harabasz index **(A)** and silhouette coefficient **(B)** depending on each feature extraction method. *J* is the number of nodes in the second hidden layer. DAE, deep autoencoder; PCA, principal component analysis.

NQ behavior score of cluster A was higher but not significantly different than that of cluster B (0.70, 95% CI [−1.15, 2.55], $p = 0.639$) and cluster C (0.65, 95% CI [−0.18, 1.49], $p = 0.159$) in the training set. The mean NQ behavior score of cluster B was lower but not significantly different than that of cluster C (−0.04, 95% CI [−1.97, 1.89], $p = 0.900$) in the test set.

The Berlin score showed that cluster A had a significantly lower risk of OSA than that of cluster B (odds ratio (OR) = 0.24, 95% CI [0.16, 0.38], $X^2 = 42.61$, $p < 0.001$) and cluster C (OR = 0.30, 95% CI [0.23, 0.38], $X^2 = 103.92$, $p < 0.001$) in the training set. The Berlin score showed that cluster B had a higher risk than that of cluster C (OR = 1.22, 95% CI [0.78, 1.91], $X^2 = 0.57$, $p =$

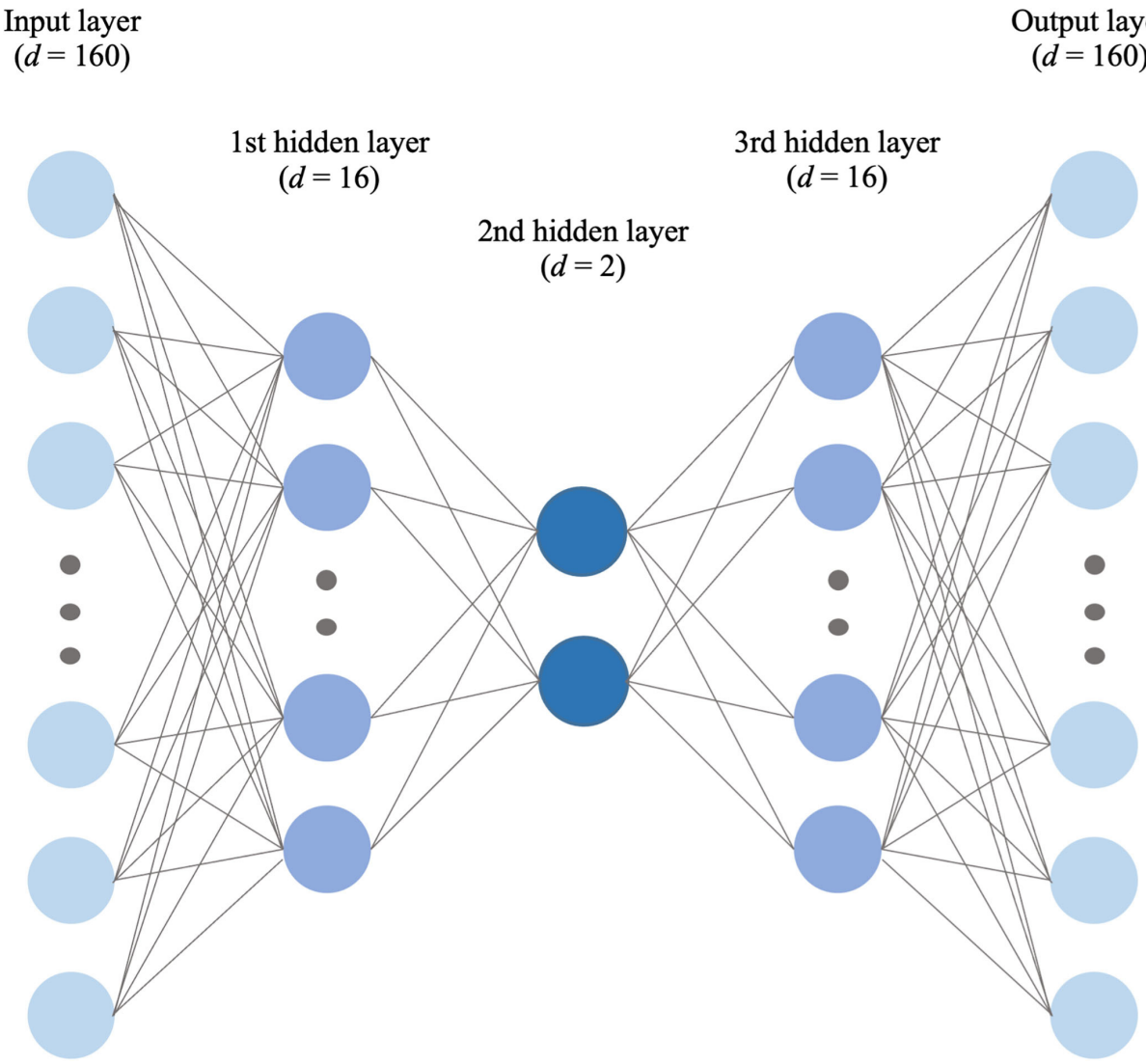

FIGURE 5
The final selected deep autoencoder model architecture for feature extraction.

0.452) in the training set. The Berlin score showed that cluster A had a significantly lower risk of OSA than that of cluster B (OR = 0.27, 95% CI [0.11, 0.68], $X^2 = 7.00$, $p = 0.008$) and cluster C (OR = 0.37, 95% CI [0.23, 0.59], $X^2 = 16.95$, $p < 0.001$) in the test set. The Berlin score showed that cluster B had a higher risk than that of cluster C (OR = 1.36, 95% CI [0.53, 3.48], $X^2 = 0.16$, $p = 0.693$) in the test set.

Three–dimensional clustering visualizations were presented with the major components that were statistically different by multi-comparison and *post-hoc* analysis in both the training and test sets, and statistically different by multi-comparison only in both the training and test sets; NQ moderation between cluster B and C was not statistically different by *post-hoc* analysis in the test set (Figure 7).

Discussion

This study demonstrated that the deep autoencoder method was a better feature extraction method for the clustering of sleep disturbances than PCA. This result is comparable to that of other studies in that the autoencoder effectively reduces the high-dimensionality of the various types of data since it can learn non-linear feature representations (43–45). Specifically, based on internal cluster validation and the elbow method, the best architecture of the deep autoencoder for extracting features for clustering our study samples with sleep disturbances was 16 nodes for the first and third hidden layers, and two nodes for the second hidden layer, while the optimal number of clusters was considered to be three. After

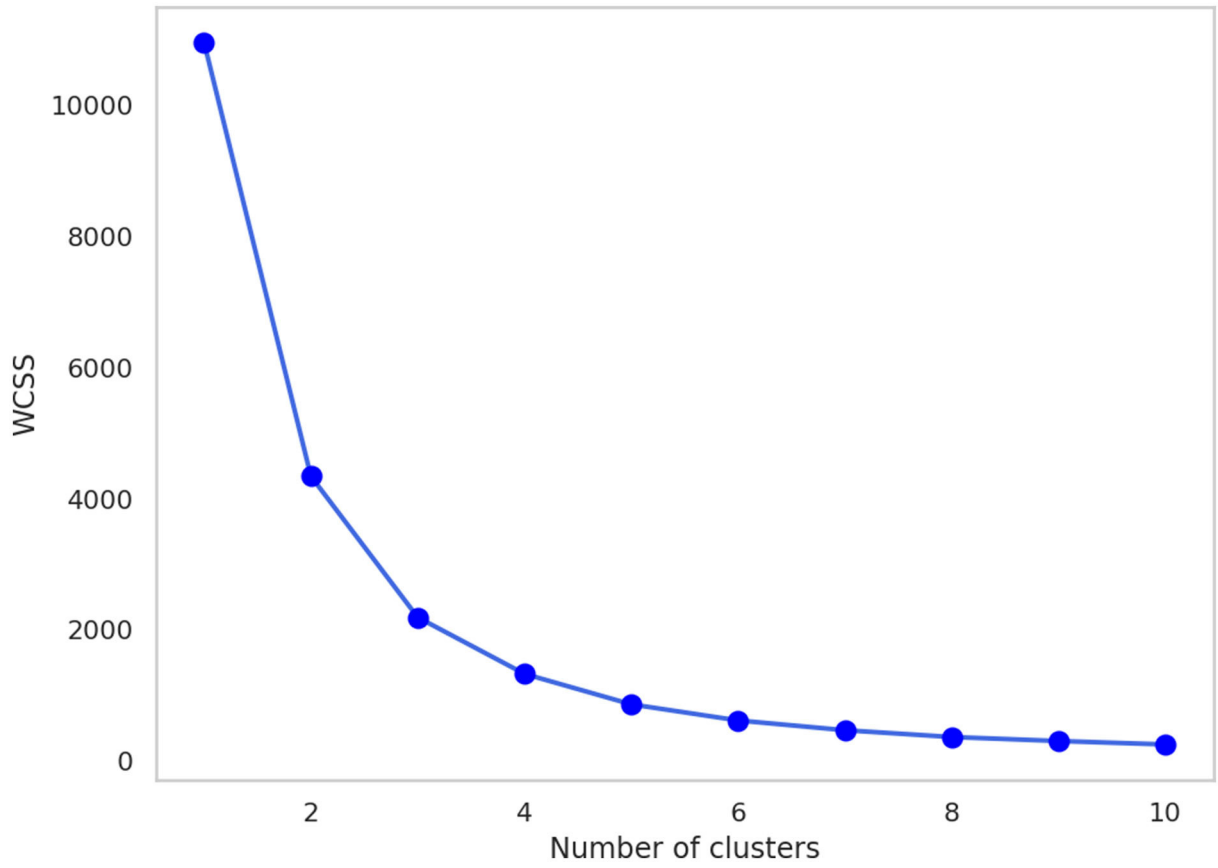

FIGURE 6
The change in WCSS with the number of clusters after feature extraction by the deep autoencoder ($J = 2$). WCSS, within-cluster sum of squares.

external cluster validation, three PI types were differentiated by changes in sleep quality, dietary habits, and concomitant gastrointestinal symptoms.

PI has been used in TEAM for the personalized care of various conditions including sleep disorders. As the accurate diagnosis and precise evaluation of individual patients are the key for personalized care in conventional medicine, PI, as well as diagnosis according to the International Classification of Diseases, Tenth Revision (ICD-10), is an important principle in personalized TEAM treatments such as acupuncture point selections and combinations of herbal medicines. Although several previous studies have tried to standardize PI and suggested new methods for PI in different types of data, it is considered a "black box" in which the external validity or usability in clinical TEAM practice cannot be ensured (14, 15, 46–48). Therefore, in another aspect of PI standardization, we proposed a new paradigm, the clinical data-driven PI model, applying advanced machine learning techniques. The PI model is flexible in the data characteristics that can be used and is reproducible for certain data to enhance the effectiveness of TEAM treatment in clinical practice (Figure 8).

There were three main aspects to this study, data type, feature extraction, and clustering. First, whole raw data from a cross-sectional study were used. The cross-sectional study data were generally composed of fundamental clinical information such as age, sex, and medical history, and symptoms, and/or a disease-related questionnaire. Particularly, our used cross-sectional study included several questionnaires with different domains including sleep, diet, nutrition, and gastrointestinal status. Since TEAM doctors usually ask not only about sleep conditions but also about other conditions to select the appropriate acupuncture points and/or herbal medicines for treating insomnia patients (28, 30), this type of data was suitable for reflecting clinical settings. Furthermore, this data type may assure external cluster validation. Clustering validation, which measures the goodness of clustering results, can be categorized into two methods: internal cluster validation and external cluster validation (49). The internal cluster validation is conducted without the need to obtain any additional information, such as evaluating the average between- and within-cluster sums of squares (Calinski-Harabasz index), or the difference of the between- and within-cluster distances

TABLE 2 Patient characteristics in each training set cluster.

Characteristics	Cluster A ($n = 1,396$)	Cluster B ($n = 98$)	Cluster C ($n = 569$)
Age, years (mean ± SD)	20.8 ± 1.2	21.2 ± 1.6	21.1 ± 1.4
Height, cm (mean ± SD)	174.6 ± 5.6	175.2 ± 6.0	174.8 ± 5.6
Weight, kg (mean ± SD)	72.2 ± 9.2	75.1 ± 11.1	73.8 ± 11.1
BMI, kg/m^2 (mean ± SD)	23.6 ± 2.6	24.4 ± 3.2	24.1 ± 3.2
Smoking status			
Never, *n* (%)	739 (52.9)	53 (54.1)	244 (42.9)
Past, *n* (%)	84 (6.0)	4 (4.1)	44 (7.7)
Active, *n* (%)	573 (41.0)	41 (41.8)	281 (49.4)
Pack-years (mean± SD)	0.97 ± 1.95	1.65 ± 3.07	1.57 ± 2.48
Alcohol, *n* (%)			
<1 time/month	322 (23.1)	28 (28.6)	113 (19.9)
<1 time/week	396 (28.4)	26 (26.5)	130 (22.8)
1–2 times/week	493 (35.3)	25 (25.5)	199 (35.0)
3–7 times/week	185 (13.3)	19 (19.4)	127 (22.3)
Caffeine, *n* (%)			
Coffee			
<1 cup/week	511 (36.6)	34 (34.7)	193 (33.9)
1–2 cups/week	299 (21.4)	16 (16.3)	114 (20.0)
3–6 cups/week	252 (18.1)	15 (15.3)	90 (15.8)
1 cup/day	184 (13.2)	13 (13.3)	77 (13.5)
2 cups/day	102 (7.3)	11 (11.2)	59 (10.4)
≥3 cups/day	48 (3.4)	9 (9.2)	36 (6.3)
Energy drink			
<1 cup/week	907 (65.0)	46 (46.9)	335 (58.9)
1–2 cups/week	266 (19.1)	17 (17.3)	88 (15.5)
3–6 cups/week	113 (8.1)	19 (19.4)	69 (12.1)
1 cup/day	72 (5.2)	8 (8.2)	43 (7.6)
2 cups/day	22 (1.6)	3 (3.1)	23 (4.0)
≥3 cups/day	16 (1.1)	5 (5.1)	11 (1.9)
Bacchus ®			
<1 cup/week	1,245 (90.9)	77 (78.6)	484 (85.1)
1–2 cups/week	121 (8.8)	13 (13.3)	60 (10.5)
3–6 cups/week	16 (1.2)	3 (3.1)	12 (2.1)
1 cup//day	10 (0.7)	2 (2.0)	7 (1.2)
2 cups/day	3 (0.2)	2 (2.0)	3 (0.5)
≥ 3 cups/day	1 (0.1)	1 (1.0)	3 (0.5)
Rank, *n* (%)			
Private	80 (5.7)	2 (2.0)	15 (1.2)
Private first class	606 (43.4)	38 (38.8)	211 (8.4)
Corporal	560 (40.1)	45 (45.9)	251 (11.6)
Sergeant	150 (12.9)	13 (13.3)	92 (3.7)
Education, *n* (%)			
Elementary school	1 (0.1)	0 (0.0)	0 (0.0)
Middle school	1 (0.1)	0 (0.0)	3 (0.5)
High school	1,326 (95.0)	89 (90.8)	518 (91.0)
University or college	68 (4.9)	9 (9.2)	48 (8.4)

(Continued)

TABLE 2 Continued

Characteristics	Cluster A ($n = 1,396$)	Cluster B ($n = 98$)	Cluster C ($n = 569$)
Exercise, *n* (%)			
<1 day/week	217 (15.5)	27 (27.6)	121 (21.3)
1–2 days/week	331 (23.7)	20 (20.4)	140 (24.6)
3–4 days/week	399 (28.6)	27 (27.6)	162 (28.5)
≥5 days/week	449 (32.2)	24 (24.5)	146 (25.7)
Physical grade i*n* the military, *n* (%)			
First	413(29.6)	17 (17.3)	142 (25.0)
Second	596 (42.7)	43 (43.9)	224 (39.4)
Third	379 (27.1)	38 (38.8)	198 (34.8)
Fourth	4 (0.3)	0 (0.0)	3 (0.5)
Fifth or above	4 (0.3)	0 (0.0)	2 (0.4)
Sleep disorders, *n* (%)			
Present history			
None	1,306 (93.6)	65 (66.3)	470 (82.6)
Insomnia	37 (2.7)	16 (16.3)	50 (8.8)
Narcolepsy	30 (2.1)	17 (17.3)	27 (4.7)
Obstructive sleep apnea	5 (0.4)	6 (6.1)	12 (2.1)
Restless leg syndrome	7 (0.5)	2 (2.0)	14 (2.5)
Periodic limb movement	7 (0.5)	4 (4.1)	8 (1.4)
Past history			
None	1,353 (96.9)	80 (81.6)	526 (92.4)
Insomnia	32 (2.3)	11 (11.2)	31 (5.4)
Narcolepsy	3 (0.2)	8 (8.2)	3 (0.5)
Obstructive sleep apnea	4 (0.3)	2 (2.0)	7 (1.2)
Restless leg syndrome	2 (0.1)	1 (1.0)	4 (0.7)
Periodic limb movement	1 (0.1)	1 (1.0)	1 (0.2)
Gastrointestinal disorders, *n* (%)			
Present history			
None	1,337 (95.8)	67 (68.4)	487 (85.6)
Gastroesophageal reflux	23 (1.6)	19 (19.4)	41 (7.2)
Gastric ulcer	0 (0.0)	0 (0.0)	3 (0.5)
Duodenal ulcer	0 (0.0)	0 (0.0)	0 (0.0)
Irritable bowel syndrome	37 (2.7)	17 (17.3)	40 (7.0)
Past history			
None	1,263 (90.5)	63 (64.3)	440 (77.3)
Gastroesophageal reflux	74 (5.3)	26 (26.5)	74 (13.0)
Gastric ulcer	1 (0.1)	1 (1.0)	7 (1.2)
Duodenal ulcer	1 (0.1)	1 (1.0)	1 (0.2)
Irritable bowel syndrome	62 (4.4)	17 (17.3)	60 (10.5)
General diseases, *n* (%)			
None	1,322 (94.7)	85 (86.7)	527 (92.6)
Hypertension	65 (4.7)	12 (12.2)	34 (6.0)
Diabetes	5 (0.4)	2 (2.0)	2 (0.4)
Hyperlipidemia	4 (0.3)	1 (1.0)	5 (0.9)
Cardiac diseases	9 (0.6)	2 (2.0)	4 (0.7)
Medications, *n* (%)			

(Continued)

TABLE 2 Continued

Characteristics	Cluster A ($n = 1,396$)	Cluster B ($n = 98$)	Cluster C ($n = 569$)
Sleeping pills	12 (0.9)	9 (9.2)	15 (2.6)
Sleep health supplements	5 (0.4)	5 (5.1)	12 (2.1)
Oral steroids	6 (0.4)	4 (4.1)	6 (1.1)
Melatonin	2 (0.1)	0 (0.0)	2 (0.4)
Anticonvulsants	0 (0.0)	0 (0.0)	2 (0.4)
Antidepressants	19 (1.4)	9 (9.2)	14 (2.5)
Beta blockers	1 (0.1)	0 (0.0)	0 (0.0)
Bronchodilators	4 (0.3)	2 (2.0)	5 (0.9)
Stimulants	3 (0.2)	4 (4.1)	6 (1.1)
Antihistamines	31 (2.2)	3 (3.1)	12 (2.1)
Weight loss pills	4 (0.3)	3 (3.1)	6 (1.1)
Weight loss supplements	26 (1.9)	8 (8.2)	21 (3.7)
Digestive pills	31 (2.2)	13 (13.3)	33 (5.8)
Digestive supplements	54 (3.9)	12 (12.2)	52 (9.1)
Stress* (mean ± SD)	3.46 ± 1.57	3.41 ± 1.16	3.40 ± 1.44
Night shift with tomorrow duty-off			
Frequency, *n* (%)			
None	781 (55.9)	47 (48.0)	289 (50.8)
1 time/month	58 (4.2)	7 (7.1)	29 (5.1)
2 times/month	115 (8.2)	9 (9.2)	36 (6.3)
3 times/month	134 (9.6)	10 (10.2)	61 (10.7)
4 times/month	94 (6.7)	9 (9.2)	44 (7.7)
≥5 times/month	214 (15.3)	16 (16.3)	110 (19.3)
Sleep disturbance or fatigue* (mean ± SD)	4.07 ± 0.87	4.43 ± 0.85	4.25 ± 0.89
Night shift without tomorrow duty-off			
Frequency, *n* (%)			
None	710 (50.9)	44 (44.9)	210 (36.9)
1 time/month	45 (3.2)	5 (5.1)	18 (3.2)
2 times/month	58 (4.2)	9 (9.2)	21 (3.7)
3 times/month	52 (3.7)	5 (5.1)	22 (3.9)
4 times/month	62 (4.4)	3 (3.1)	29 (5.1)
≥5 times/month	469 (33.6)	32 (32.7)	169 (29.7)
Sleep disturbance or fatigue* (mean ± SD)	4.28 ± 0.85	4.50 ± 0.77	4.42 ± 0.85

Values are presented as the mean ± standard deviation (range) or number (%).
*Five-point Likert scale.
BMI, body mass index.

(silhouette coefficient). On the other hand, the external cluster validation is conducted with other external information, such as a true class of cluster or previous knowledge about data. In this study, instead of obtaining the true labels of each cluster, which require large amounts of cost and time for TEAM doctors, we used the final scores of PSQI, NQ, GSRS, and the Berlin score, which were not used in the input features of clustering, but which could be calculated using specific non-linear functions respectively to externally compare the clustering results.

Second, feature extraction was conducted by a deep autoencoder model. Two methods have been used before the clustering process, feature selection (selecting a small subset of actual features from the data) and feature extraction (constructing a small set of artificial features from the data). Most clinical studies conducted feature selection through statistical methods such as the *t*-test or chi-squared test between two groups or it was determined by clinical experience or medical knowledge. However, in a large series of data, so-called high-dimensional data, it was difficult to find the best feature

TABLE 3 Patient characteristics in each test set cluster.

Characteristics	Cluster A ($n = 352$)	Cluster B ($n = 22$)	Cluster C ($n = 142$)
Age, years (mean ± SD)	21.0 ± 1.4	21.9 ± 2.2	21.1 ± 1.5
Height, cm (mean ± SD)	174.2 ± 5.5	175.3 ± 4.7	174.9 ± 5.1
Weight, kg (mean ± SD)	72.5 ± 9.7	74.9 ± 7.8	74.6 ± 10.9
BMI, kg/m^2 (mean ± SD)	23.8 ± 2.6	24.4 ± 2.5	24.4 ± 3.1
Smoking status			
Never, *n* (%)	186 (52.8)	6 (27.3)	60 (42.3)
Past, *n* (%)	19 (5.4)	3 (13.6)	11 (7.7)
Active, *n* (%)	147 (41.8)	13 (59.1)	71 (50.)
Pack-years (mean± SD)	0.94 ± 1.80	2.07 ± 2.99	1.78 ± 2.68
Alcohol, *n* (%)			
<1 time/month	77 (21.9)	6 (27.3)	25 (17.6)
<1 time/week	107 (30.4)	4 (18.2)	35 (24.6)
1–2 times/week	122 (34.7)	8 (36.4)	41 (28.9)
3–7 times/week	46 (13.1)	4 (18.2)	41 (28.9)
Caffeine, *n* (%)			
Coffee			
<1 cup/week	131 (37.2)	8 (36.4)	35 (24.6)
1–2 cups/week	89 (25.3)	2 (9.1)	31 (21.8)
3–6 cups/week	56 (15.9)	4 (18.2)	24 (16.9)
1 cup/day	44 (12.5)	6 (27.3)	24 (16.9)
2 cups/day	20 (5.7)	1 (4.5)	21 (14.8)
≥3 cups/day	12 (3.4)	1 (4.5)	7 (4.9)
Energy drink			
<1 cup/week	234 (66.5)	15 (68.2)	75 (52.8)
1–2 cups/week	65 (18.5)	3 (13.6)	37 (26.1)
3–6 cups/week	29 (8.2)	3 (13.6)	10 (7.0)
1 cup/day	13 (3.7)	1 (4.5)	11 (7.7)
2 cups/day	7 (2.0)	0 (0.0)	4 (2.8)
≥3 cups/day	4 (1.1)	0 (0.0)	5 (3.5)
Bacchus ®			
<1 cup/week	321 (91.2)	18 (81.8)	118 (83.1)
1–2 cups/week	27 (7.7)	3 (13.6)	12 (8.5)
3–6 cups/week	3 (0.9)	1 (4.5)	7 (4.9)
1 cup/day	0 (0.0)	0 (0.0)	3 (2.1)
2 cups/day	1 (0.3)	0 (0.0)	1 (0.7)
≥3 cups/day	0 (0.0)	0 (0.0)	1 (0.7)
Rank, *n* (%)			
Private	25 (7.1)	2 (9.1)	7 (4.9)
Private first class	142 (40.3)	3 (13.6)	48 (33.8)
Corporal	136 (38.6)	10 (45.5)	66 (46.5)
Sergeant	49 (13.9)	7 (31.8)	21 (14.8)
Education, *n* (%)			
Elementary school	0 (0.0)	0 (0.0)	0 (0.0)
Middle school	0 (0.0)	1 (4.5)	1 (0.7)
High school	329 (93.5)	19 (86.4)	134 (94.4)
University or college	23 (6.5)	2 (9.1)	7 (4.9)

(Continued)

TABLE 3 Continued

Characteristics	Cluster A ($n = 352$)	Cluster B ($n = 22$)	Cluster C ($n = 142$)
Exercise, *n* (%)			
<1 day/week	50 (14.2)	6 (27.3)	23 (16.2)
1–2 days/week	91 (25.9)	5 (22.7)	35 (24.6)
3–4 days/week	88 (25.0)	5 (22.7)	36 (25.4)
≥5 days/week	123 (34.9)	6 (27.3)	48 (33.8)
Physical grade in the military, *n* (%)			
First	103 (29.3)	8 (36.4)	31 (21.8)
Second	141 (40.1)	6 (27.3)	63 (44.4)
Third	107 (30.4)	7 (31.8)	47 (33.1)
Fourth	1 (0.3)	1 (4.5)	1 (0.7)
Fifth or above	0 (0.0)	0 (0.0)	0 (0.0)
Sleep disorders, *n* (%)			
Present history			
None	334 (90.9)	18 (81.8)	114 (80.3)
Insomnia	10 (2.8)	3 (13.6)	14 (9.9)
Narcolepsy	5 (1.4)	2 (9.1)	7 (4.9)
Obstructive sleep apnea	1 (0.3)	0 (0.0)	2 (1.4)
Restless leg syndrome	1 (0.3)	1 (4.5)	6 (4.2)
Periodic limb movement	2 (0.6)	1 (4.5)	5 (3.5)
Past history			
None	344 (97.7)	20 (90.9)	131 (92.3)
Insomnia	5 (1.4)	2 (9.1)	6 (4.2)
Narcolepsy	2 (0.6)	0 (0.0)	3 (2.1)
Obstructive sleep apnea	0 (0.0)	0 (0.0)	1 (0.7)
Restless leg syndrome	1 (0.3)	0 (0.0)	1 (0.7)
Periodic limb movement	0 (0.3)	0 (0.0)	1 (0.7)
Gastrointestinal disorders, *n* (%)			
Present history			
None	329 (93.5)	20 (90.9)	130 (91.5)
Gastroesophageal reflux	13 (3.7)	2 (9.1)	3 (2.1)
Gastric ulcer	0 (0.0)	0 (0.0)	0 (0.0)
Duodenal ulcer	0 (0.0)	0 (0.0)	0 (0.0)
Irritable bowel syndrome	12 (3.4)	0 (0.0)	10 (7.0)
Past history			
None	300 (85.2)	15 (68.2)	119 (83.8)
Gastroesophageal reflux	22 (6.3)	5 (22.7)	11 (7.7)
Gastric ulcer	5 (1.4)	1 (4.5)	0 (0.0)
Duodenal ulcer	1 (0.3)	0 (0.0)	0 (0.0)
Irritable bowel syndrome	27 (7.7)	3 (13.6)	14 (9.9)
General diseases, *n* (%)			
None	331 (94.0)	12 (54.5)	129 (90.8)
Hypertension	18 (5.1)	5 (22.7)	10 (7.0)
Diabetes	1 (0.3)	0 (0.0)	2 (1.4)
Hyperlipidemia	3 (0.9)	0 (0.0)	2 (1.4)
Cardiac diseases	1 (0.3)	2 (9.1)	2 (1.4)
Medications, *n* (%)			

(Continued)

TABLE 3 Continued

Characteristics	Cluster A ($n = 352$)	Cluster B ($n = 22$)	Cluster C ($n = 142$)
Sleeping pills	0 (0.0)	1 (4.5)	3 (2.1)
Sleep health supplements	1 (0.3)	0 (0.0)	1 (0.7)
Oral steroids	2 (0.6)	1 (4.5)	1 (0.7)
Melatonin	0 (0.0)	1 (4.5)	0 (0.0)
Anticonvulsants	0 (0.0)	0 (0.0)	0 (0.0)
Antidepressants	2 (0.6)	2 (9.1)	3 (2.1)
Beta blockers	0 (0.0)	0 (0.0)	0 (0.0)
Bronchodilators	1 (0.3)	0 (0.0)	1 (0.7)
Stimulants	1 (0.3)	0 (0.0)	0 (0.0)
Antihistamines	10 (2.8)	1 (4.5)	4 (2.8)
Weight loss pills	0 (0.0)	0 (0.0)	0 (0.0)
Weight loss supplements	3 (0.9)	1 (4.5)	10 (7.0)
Digestive pills	9 (2.6)	2 (9.1)	6 (4.2)
Digestive supplements	12 (3.4)	2 (9.1)	9 (6.3)
Stress* (mean ± SD)	3.57 ± 1.57	3.41 ± 1.30	3.42 ± 1.38
Night shift with tomorrow duty-off			
Frequency, *n* (%)			
None	192 (54.5)	10 (45.5)	72 (50.7)
1 time/month	19 (5.4)	3 (13.6)	7 (4.9)
2 times/month	36 (10.2)	2 (9.1)	7 (4.9)
3 times/month	32 (9.1)	3 (13.6)	13 (9.2)
4 times/month	21 (6.0)	2 (9.1)	7 (4.9)
≥5 times/month	52 (14.8)	2 (9.1)	36 (25.4)
Sleep disturbance or fatigue* (mean ± SD)	3.96 ± 0.89	4.13 ± 0.74	4.43 ± 0.80
Night shift without tomorrow duty-off			
Frequency, *n* (%)			
None	175 (49.7)	10 (45.5)	84 (59.2)
1 time/month	19 (5.4)	2 (9.1)	6 (4.2)
2 times/month	13 (3.7)	2 (9.1)	5 (3.5)
3 times/month	15 (4.3)	1 (4.5)	4 (2.8)
4 times/month	16 (4.5)	0 (0.0)	13 (9.2)
≥5 times/month	114 (32.4)	7 (31.8)	30 (21.1)
Sleep disturbance or fatigue* (mean ± SD)	4.24 ± 0.88	4.67 ± 0.62	4.42 ± 0.76

Values are presented as the mean ± standard deviation (range) or number (%).
*Five-point Likert scale.
BMI, body mass index.

selection strategy for efficiently reducing the dimension of the data (50). Therefore, some algorithms such as PCA and the autoencoder have been suggested for feature extraction (51), very similar to a TEAM doctor's PI process made by observing patients with not just a few pieces of clinical information but comprehensively, using a lot of clinical information. This characteristic of TEAM doctors' decision-making might also be related to the reason why deep autoencoder model extraction was much more efficient than that of other methods in our study. As decision-making in TEAM is complex and the interactions between clinical information and PI are non-linear, autoencoder architecture learning non-linear mapping allows for the transformation of high-dimensional data into more clustering-friendly representations, whereas PCA is fundamentally limited to linear embedding, and it is possible to lose essential features (38). Another strength of using the deep autoencoder for feature extraction is that it can extract features from non-quantizable questionnaire responses (e.g., dietary habit survey questionnaire), which does not use a formula to generate a single score, and efficiently prevents the curse

TABLE 4 Results of external cluster validation in the training set and test set.

	Cluster A	Cluster B	Cluster C			Cluster A vs. B		Cluster A vs. C		Cluster B vs. C	
	mean ± SD or ratio	mean ± SD or ratio	mean ± SD or ratio	F/X^2	*p*-value	Difference or OR [95% CI]	*p*-value	Difference or OR [95% CI]	*p*-value	Difference or OR [95% CI]	*p*-value
Training set	*n* = 1,396	*n* = 98	*n* = 569								
PSQI	8.33 ± 2.20	11.57 ± 3.31	9.96 ± 2.81	149.27	<0.001	−3.24 [−3.84, −2.64]	<0.001	−1.62 [−1.91, −1.34]	<0.001	1.62 [0.99, 2.24]	<0.001
GSRS	2.18 ± 2.15	17.94 ± 7.38	6.39 ± 4.09	1,305.82	<0.001	−15.76 [−16.55, −14.98]	<0.001	−4.21 [−4.59, −3.84]	<0.001	11.55 [10.73, 12.37]	<0.001
NQ global	41.51 ± 8.78	42.78 ± 11.24	42.11 ± 10.65	1.45	0.234	−1.27 [−3.59, 1.05]	0.404	−0.60 [−1.71, 0.50]	0.406	0.67 [−1.76, 3.09]	0.774
NQ balance	35.44 ± 14.33	33.25 ± 15.37	33.24 ± 16.73	4.79	0.008	2.18 [−1.51, 5.88]	0.349	2.20 [0.44, 3.96]	0.010	0.02 [−3.85, 3.88]	0.900
NQ diversity	12.67 ± 4.21	11.95 ± 4.54	12.83 ± 4.55	1.76	0.172	0.73 [−0.33, 1.79]	0.240	−0.16 [−0.66, 3.48]	0.728	−0.89 [−1.99, 0.22]	0.147
NQ moderation	8.65 ± 4.18	12.9 ± 5.85	11.12 ± 4.91	90.63	<0.001	−4.25 [−5.35, −3.15]	<0.001	−2.48 [−3.00, −1.95]	<0.001	1.77 [0.62, 2.92]	<0.001
NQ behavior	11.38 ± 3.46	9.67 ± 3.74	9.90 ± 3.66	42.07	<0.001	1.72 [0.85, 2.58]	<0.001	1.49 [1.08, 1.90]	<0.001	−0.23 [−1.13, 0.68]	0.806
Berlin score (low/high)	1,230/166	63/35	391/178	122.00	<0.001	0.24 [0.16, 0.38]	<0.001	0.30 [0.23, 0.38]	<0.001	1.22 [0.78, 1.91]	0.452
Test set	n=352	n=22	n=142								
PSQI	8.36 ± 2.12	11.14 ± 3.45	10.04 ± 2.84	34.32	<0.001	−2.78 [−4.02, −1.53]	<0.001	−1.68 [−2.24, −1.12]	<0.001	1.09 [−0.20, 2.39]	0.117
GSRS	2.08 ± 2.05	19.77 ± 8.15	6.52 ± 3.73	406.57	<0.001	−17.69 [−19.28, −16.10]	<0.001	−4.44 [−5.16, −3.72]	<0.001	13.25 [11.60, 14.91]	<0.001
NQ global	40.30 ± 8.28	44.71 ± 17.48	43.83 ± 10.59	8.31	<0.001	−4.41 [−9.31, 0.50]	0.089	−3.52 [−5.74, −1.30]	<0.001	0.88 [−4.23, 6.00]	0.900
NQ balance	33.89 ± 13.10	37.55 ± 24.27	35.30 ± 17.34	0.95	0.387	−3.66 [−11.39, 4.08]	0.508	−1.42 [−4.92, 2.08]	0.599	2.24 [−5.82, 10.30]	0.770
NQ diversity	12.33 ± 4.05	11.62 ± 5.69	12.88 ± 4.12	1.34	0.263	0.72 [−1.43, 2.86]	0.694	−0.54 [−1.51, 0.43]	0.386	−1.26 [−3.49, 0.97]	0.382
NQ moderation	8.43 ± 4.02	12.32 ± 5.49	11.71 ± 4.88	37.53	<0.001	−4.89 [−7.13, −2.64]	<0.001	−3.28 [−4.29, −2.26]	<0.001	1.61 [−0.73, 3.95]	0.238

(Continued)

10.3389/fmed.2022.950327

TABLE 4 Continued

	Cluster A	Cluster B	Cluster C			Cluster A vs. B		Cluster A vs. C		Cluster B vs. C	
	mean ± SD or ratio	mean ± SD or ratio	mean ± SD or ratio	F/X^2	*p*-value	Difference or OR [95% CI]	*p*-value	Difference or OR [95% CI]	*p*-value	Difference or OR [95% CI]	*p*-value
NQ behavior	11.12 ± 3.41	10.42 ± 4.12	10.47 ± 3.91	1.90	0.151	0.70 [−1.15, 2.55]	0.639	0.65 [−0.18, 1.49]	0.159	−0.04 [−1.97, 1.89]	0.900
Berlin score (low/high)	305/47	14/8	100/42	22.09	<0.001	0.27 [0.11, 0.68]	0.008	0.37 [0.23, 0.59]	<0.001	1.36 [0.53, 3.48]	0.693

GSRS, gastrointestinal symptom rating scale; NQ, nutrition quotient; OR, odds ratio; PSQI, Pittsburgh sleep quality index; SD, standard deviation.

of dimensionality without suffering from high computational complexity in large-scale data (38).

Third, *k*-means clustering, an unsupervised machine learning algorithm serving as a powerful computational method to analyze high-dimensional data in the form of sequences or expressions, was used in this study (38). It does not need data labeling, which is costly and time-consuming in biomedical research using supervised learning. In addition, even if data labeling is performed by several TEAM doctors, the labeling results are highly likely to be inconsistent because the types of PI are inconsistent among TEAM doctors and each TEAM society and different depending upon the disease. Therefore, a data-driven approach to PI for TEAM research, which is flexible for changes in data and reproducible for certain data, might be more reasonable than a standardization approach using a few TEAM research experts.

Each cluster of sleep disturbance patients could be differentiated, as shown in Figure 7. The patients in cluster A had relatively mild sleep disturbances, severe immoderation in the amount of food consumed, and good gastrointestinal status compared to the other clusters. The patients in cluster B had relatively severe sleep disturbances, mild immoderation in the amount of food consumed, and severe gastrointestinal problems compared to the other clusters. The patients in cluster C had relatively moderate sleep disturbances, moderate immoderation in the amount of food consumed, and mild-to-moderate gastrointestinal problems compared to the other clusters. Although the statistical analysis of the Berlin score indicated that cluster A had a much lower risk than the other two clusters, it could not be observed in the 3-dimensional visualizations.

The clustering results can be interpreted in two aspects, the changes in sleep quality and the concomitant symptoms. As sleep quality deteriorates, the appetite associated with food moderation decreases, and the condition of the gastrointestinal system worsens. Based on a recent systematic review and meta-analysis of acupuncture using PI and TEAM clinical guidelines for insomnia patients, cluster A may be matched to the "stomach disharmony pattern" type using ST36, CV12, and ST25; cluster C may be matched to the "pattern of lingering phlegm" type using ST40 and CV12; and cluster B may be matched to the "pattern of dual deficiency of heart and spleen" type using CV12, ST36, and ST40 (29, 30). This clustering model can automatically and consistently provide the same PI for a certain patient, which ensures reliability for both TEAM doctors and patients. However, it should be noted that this clustering model is flexible to the number of patient data, changes in patient features, or changes in the target disease, so-called "transfer learning" and "fine-tuning" in machine learning techniques (52), which might provide a different output for the number or types of patterns identified. Therefore, the novel PI model in the present study can be advanced, modified, or expanded for other studies.

The applications of this study include AI-based clinical decision support systems (CDSSs) through electronic medical

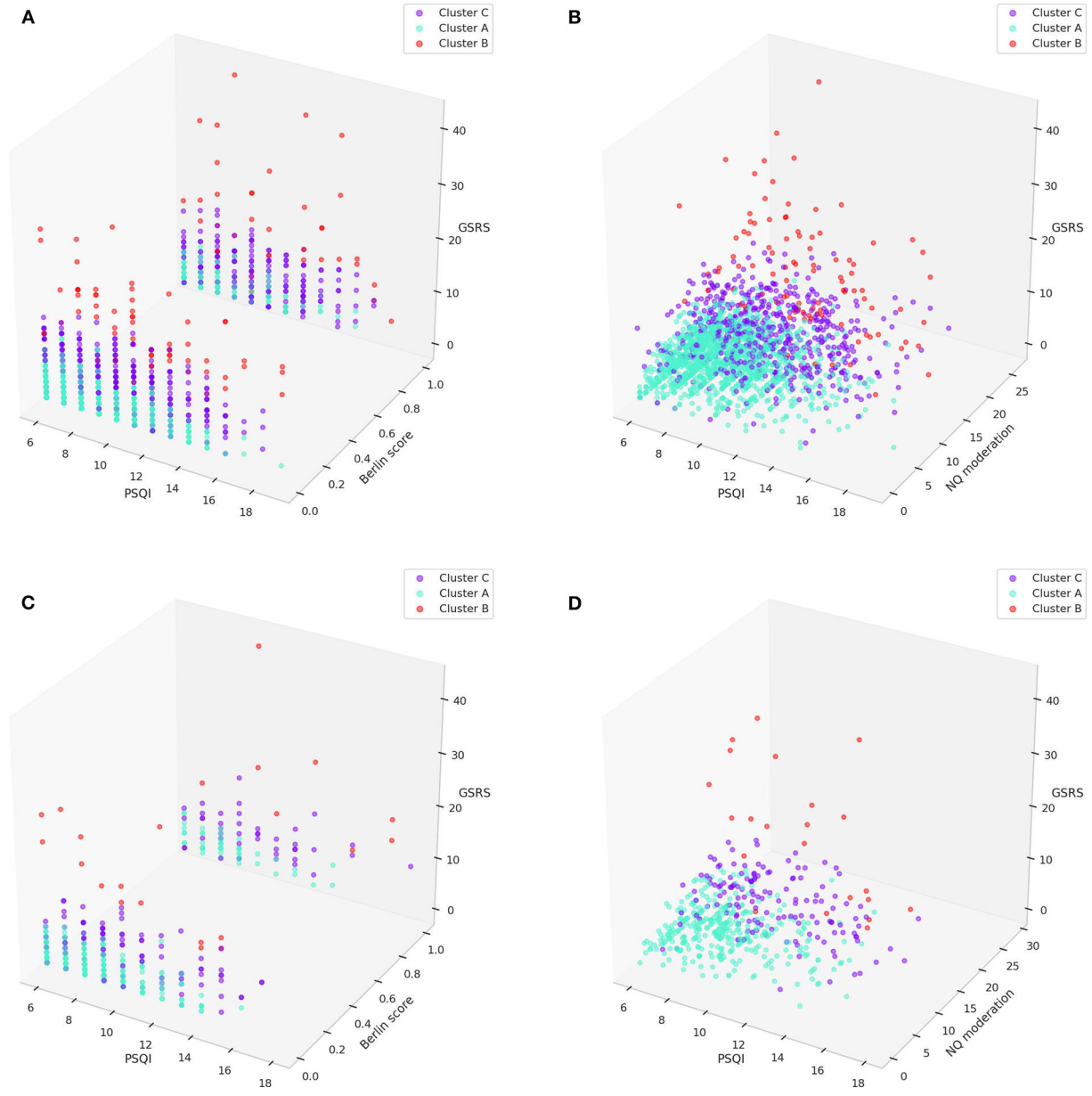

FIGURE 7
3D visualization of clusters with major components (PSQI, Berlin, GSRS scores, and NQ moderation) in the training set **(A,B)** and test set **(C,D)**. GSRS, gastrointestinal symptom rating scale; NQ, nutrition quotient; PSQI, Pittsburgh sleep quality index.

records (EMRs) and clinical trial protocols for evaluating the effectiveness of TEAM treatment. If a TEAM doctor in clinical practice obtains clinical data from insomnia patients and documents them in the EMR, the PI model in AI-based CDSSs suggests the candidate PI with the associated probability and recommends a fundamental combination of acupuncture points and herbal medicines. In addition, most pragmatic trial protocols with individualized TEAM treatment depend completely upon (one person or more) the TEAM doctor's PI for each patient. The reliability and validity of PI itself, which might affect the effect size of individualized TEAM treatments, are limited. However, the PI model in this study could suggest a consistent PI technique for patients with similar features, although the model's effect on the results of individualized TEAM treatment should be validated in a prospective clinical trial.

Some limitations of this study follow. First, this cross-sectional study data might not be fully sufficient to mimic the

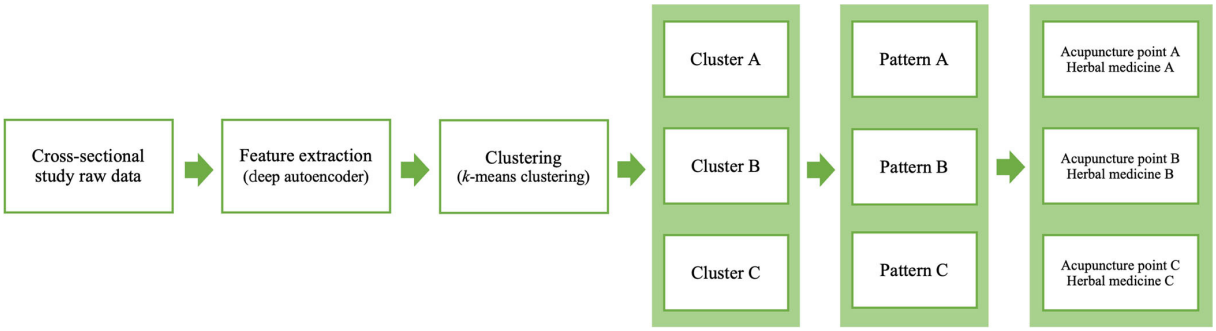

FIGURE 8
Flow diagram of the proposed PI model for acupuncture point and herbal medicine selection.

interaction between doctors and patients in clinical practices. Some data obtained from free medical notes or an AI speaker in clinical settings might be helpful to overcome this limitation. Second, since this data was obtained from a single sample of sleep disturbances in the ROKA, another study sample is required for external validation of our proposed model. Third, this study sequentially used a feature extraction model and a clustering model separately. Emerging machine learning research such as a deep clustering network, which optimizes the feature extraction model and the clustering model simultaneously, might perform better than the techniques used in our study. This will be considered in future studies. Fourth, the PI data used for each patient made by TEAM doctors were limited in this study. However, the correlation between our model's output and actual PI by TEAM doctors in this study should be observed to externally and more robustly validate our clustering results. Fifth, although all features of data were included to reflect a clinical setting wherein TEAM doctors might consider all information of patients as much as possible to find the appropriate PI, the feature selection algorithms, such as univariate statistical test, Lasso regularization, or Boruta algorithm can be applied in future studies to improve upon our results. Finally, the specific combinations of acupuncture points and herbal medicines after PI process were not represented in this study. Although this study revealed the basic concepts of the novel data-driven PI model, more research such as a systematic review of published clinical articles, including case series, or a survey of TEAM doctors is required to recommend the appropriate acupuncture points and/or herbal medicines after the determination of PI.

Data availability statement

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

Code availability

Code for this paper is provided at https://github.com/HyeonhoonLee/DeepPI.

Ethics statement

The studies involving human participants were reviewed and approved by Institutional Review Board of the Armed Forces Medical Command (AFMC-202107-HR-054-02). The patients/participants provided their written informed consent to participate in this study.

Author contributions

HL and YC designed the research study. CY and J-DL supervised the study. HL, YC, BS, and SL collected and analyzed the survey data. HL drafted the manuscript. SL, JK, KK, and EK provided critical comments for improvement of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by grants from Korea Institute of Oriental Medicine (KSN2022210).

Acknowledgments

This paper was modified and developed from the Ph.D. thesis of HL.

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.2022.950327/full#supplementary-material

References

1. Zhao C, Li GZ, Wang C, Niu J. Advances in patient classification for traditional chinese medicine: a machine learning perspective. *Evid Based Complement Alternat Med.* (2015) 2015:376716. doi: 10.1155/2015/376716

2. Choi EM, Jiang F, Longhurst JC. Point specificity in acupuncture. *Chin Med.* (2012) 7:4. doi: 10.1186/1749-8546-7-4

3. Lee YS Ryu Y, Yoon DE, Kim CH, Hong G, Hwang YC, et al. Commonality and specificity of acupuncture point selections. *Evid Based Complement Alternat Med.* (2020) 2020:2948292. doi: 10.1155/2020/2948292

4. Liu W, Cohen L. Overcoming barriers for clinical research of acupuncture. *Med Acupunct.* (2020) 32:348–51. doi: 10.1089/acu.2020.1480

5. Pach D, Yang-Strobel X, Ludtke R, Roll S, Icke K, Brinkhaus B, et al. Standardized vs. individualized acupuncture for chronic low back pain: a randomized controlled trial. *Evid Based Complement Alternat Med.* (2013) 2013:125937. doi: 10.1155/2013/125937

6. Ko SJ, Kuo B, Kim SK, Lee H, Kim J, Han G, et al. Individualized acupuncture for symptom relief in functional dyspepsia: a randomized controlled trial. *J Altern Complement Med.* (2016) 22:997–1006. doi: 10.1089/acm.2016.0208

7. Nielsen A, Ocker L, Majd I, Draisin JA, Taromina K, Maggenti MT, et al. Acupuncture intervention protocol: consensus process for a pragmatic randomized controlled trial of acupuncture for management of chronic low back pain in older adults: an nih heal initiative funded project. *Glob Adv Health Med.* (2021) 10:21649561211007091. doi: 10.1177/21649561211007091

8. Brinkhaus B, Roll S, Jena S, Icke K, Adam D, Binting S, et al. Acupuncture in patients with allergic asthma: a randomized pragmatic trial. *J Altern Complement Med.* (2017) 23:268–77. doi: 10.1089/acm.2016.0357

9. Chung VC, Wong CH, Wu IX, Ching JY, Cheung WK, Yip BH, et al. Electroacupuncture plus on-demand gastrocaine for refractory functional dyspepsia: pragmatic randomized trial. *J Gastroenterol Hepatol.* (2019) 34:2077–85. doi: 10.1111/jgh.14737

10. Liu S, Wang Z, Su Y, Qi L, Yang W, Fu M, et al. A Neuroanatomical basis for electroacupuncture to drive the vagal-adrenal axis. *Nature.* (2021) 598:641–5. doi: 10.1038/s41586-021-04001-4

11. Lee DY, Jiu YR, Hsieh CL. Electroacupuncture at Zusanli and at Neiguan characterized point specificity in the brain by metabolomic analysis. *Sci Rep.* (2020) 10:10717. doi: 10.1038/s41598-020-67766-0

12. Liu S, Wang ZF, Su YS, Ray RS, Jing XH, Wang YQ, et al. Somatotopic organization and intensity dependence in driving distinct Npy-expressing sympathetic pathways by electroacupuncture. *Neuron.* (2020) 108:436-50 e7. doi: 10.1016/j.neuron.2020.07.015

13. Ma Q. Somato-Autonomic reflexes of acupuncture. *Med Acupunct.* (2020) 32:362–6. doi: 10.1089/acu.2020.1488

14. Kim CH, Yoon DE, Lee YS, Jung WM, Kim JH, Chae Y. Revealing associations between diagnosis patterns and acupoint prescriptions using medical data extracted from case reports. *J Clin Med.* (2019) 8:1663. doi: 10.3390/jcm8101663

15. Hwang YC, Lee IS Ryu Y, Lee YS, Chae Y. Identification of acupoint indication from reverse inference: data mining of randomized controlled clinical trials. *J Clin Med.* (2020) 9:3027. doi: 10.3390/jcm9093027

16. Hwang YC, Lee IS, Ryu Y, Lee MS, Chae Y. Exploring traditional acupuncture point selection patterns for pain control: data mining of randomised controlled clinical trials. *Acupunct Med.* (2020):964528420926173. doi: 10.1177/0964528420926173

17. Jung WM, Park IS, Lee YS, Kim CE, Lee H, Hahm DH, et al. Characterization of hidden rules linking symptoms and selection of acupoint using an artificial neural network model. *Front Med.* (2019) 13:112–20. doi: 10.1007/s11684-017-0582-z

18. Huang WT, Hung HH, Kao YW, Ou SC, Lin YC, Cheng WZ, et al. Application of neural network and cluster analyses to differentiate tcm patterns in patients with breast cancer. *Front Pharmacol.* (2020) 11:670. doi: 10.3389/fphar.2020.00670

19. Lee SH, Lim SM. Acupuncture for insomnia after stroke: a systematic review and meta-analysis. *BMC Complement Altern Med.* (2016) 16:228. doi: 10.1186/s12906-016-1220-z

20. Han N, Qiao S, Yuan G, Huang P, Liu D, Yue K, et al. Novel Chinese herbal medicine clustering algorithm *via* artificial bee colony optimization. *Artif Intell Med.* (2019) 101:101760. doi: 10.1016/j.artmed.2019.101760

21. Liu Z, Luo C, Fu D, Gui J, Zheng Z, Qi L, et al. A novel transfer learning model for traditional herbal medicine prescription generation from unstructured resources and knowledge. *Artif Intell Med.* (2022) 124:102232. doi: 10.1016/j.artmed.2021.102232

22. Ang L, Lee HW, Choi JY, Zhang J, Soo Lee M. Herbal medicine and pattern identification for treating COVID-19: a rapid review of guidelines. *Integr Med Res.* (2020) 9:100407. doi: 10.1016/j.imr.2020.100407

23. Xu Q, Guo Q, Wang CX, Zhang S, Wen CB, Sun T, et al. Network differentiation: a computational method of pathogenesis diagnosis in traditional chinese medicine based on systems science. *Artif Intell Med.* (2021) 118:102134. doi: 10.1016/j.artmed.2021.102134

24. Yin X, Gou M, Xu J, Dong B, Yin P, Masquelin F, et al. Efficacy and safety of acupuncture treatment on primary insomnia: a randomized controlled trial. *Sleep Med.* (2017) 37:193–200. doi: 10.1016/j.sleep.2017.02.012

25. Pei W, Peng R, Gu Y, Zhou X, Ruan J. Research trends of acupuncture therapy on insomnia in two decades (from 1999 to 2018):a bibliometric analysis. *BMC Complement Altern Med.* (2019) 19:225. doi: 10.1186/s12906-019-2606-5

26. Abanes JJ, Ridner SH, Dietrich MS, Hiers C, Rhoten B. Acupuncture for Sleep Disturbances in Post-Deployment Military Service Members: A Randomized Controlled Trial. *Clin Nurs Res.* (2022) 31:239–50. doi: 10.1177/10547738211030602

27. Choi Y, Kim Y, Kwon O, Chung SY, Cho SH. Effect of herbal medicine (Huanglian-Jie-Du Granule) for somatic symptoms and insomnia in patients with Hwa-Byung: a randomized controlled trial. *Integr Med Res.* (2021) 10:100453. doi: 10.1016/j.imr.2020.100453

28. Leach MJ, Page AT. Herbal medicine for insomnia: a systematic review and meta-analysis. *Sleep Med Rev.* (2015) 24:1–12. doi: 10.1016/j.smrv.2014.12.003

29. Lim JH, Jeong JH, Kim SH, Kim KO, Lee SY, Lee SH, et al. The pilot survey of the perception on the practice pattern, diagnosis, and treatment on Korean medicine insomnia: focusing on the difference between korean medical neuropsychiatry specialists and Korean medical general practitioners. *Evid Based Complement Alternat Med.* (2018) 2018:9152705. doi: 10.1155/2018/9152705

30. Kim SH, Jeong JH, Lim JH, Kim BK. Acupuncture using pattern-identification for the treatment of insomnia disorder: a systematic review and meta-analysis of randomized controlled trials. *Integr Med Res.* (2019) 8:216–26. doi: 10.1016/j.imr.2019.08.002

31. Fan W, Yan Z. Factors affecting response rates of the web survey: a systematic review. *Comput Human Behav.* (2010) 26:132–9. doi: 10.1016/j.chb.2009.10.015

32. Lewis SJ, Heaton KW. Stool form scale as a useful guide to intestinal transit time. *Scand J Gastroenterol.* (1997) 32:920–4. doi: 10.3109/00365529709011203

33. Backhaus J, Junghanns K, Broocks A, Riemann D, Hohagen F. Test-retest reliability and validity of the Pittsburgh sleep quality index in primary insomnia. *J Psychosom Res.* (2002) 53:737–40. doi: 10.1016/S0022-3999(02)00330-6

34. Netzer NC, Stoohs RA, Netzer CM, Clark K, Strohl KP. Using the Berlin questionnaire to identify patients at risk for the sleep apnea syndrome. *Ann Intern Med.* (1999) 131:485–91. doi: 10.7326/0003-4819-131-7-199910050-00002

35. Jo JS, Kim KN. Development of a questionnaire for dietary habit survey of Korean adults. *Korean J Community Nutr.* (2014) 19:258–73. doi: 10.5720/kjcn.2014.19.3.258

36. Lee J-S, Kim H-Y, Hwang J-Y, Kwon S, Chung HR, Kwak T-K, et al. Development of nutrition quotient for korean adults: item selection and validation of factor structure. *J Nutr Health.* (2018) 51:340–56. doi: 10.4163/jnh.2018.51.4.340

37. Kwon S, Jung H-K, Hong JH, Park HS. Diagnostic validity of the Korean Gastrointestinal Symptom Rating Scale (KGSRS) in the assessment of gastro-esophageal reflux disease. *Ewha Med J.* (2008) 31:73–80. doi: 10.12771/emj.2008.31.2.73

38. Karim MR, Beyan O, Zappa A, Costa IG, Rebholz-Schuhmann D, Cochez M, et al. Deep learning-based clustering approaches for bioinformatics. *Brief Bioinform.* (2021) 22:393–415. doi: 10.1093/bib/bbz170

39. Ahmed M, Seraj R, Islam SMS. The K-means algorithm: a comprehensive survey and performance evaluation. *Electronics.* (2020) 9:1295. doi: 10.3390/electronics9081295

40. Wang X, Xu Y. *An Improved Index For Clustering Validation Based on Silhouette Index and Calinski-Harabasz Index.* IOP Conference Series: Materials Science and Engineering. (2019) 569:052024. doi: 10.1088/1757-899X/569/5/052024

41. Sammouda R, El-Zaart A. An optimized approach for prostate image segmentation using K-means clustering algorithm with elbow method. *Comput Intell Neurosci.* (2021) 2021:4553832. doi: 10.1155/2021/4553832

42. Lee M, Lee S, Park J, Seo S. Clustering and characterization of the lactation curves of dairy cows using K-medoids clustering algorithm. *Animals.*(2020) 10:1348. doi: 10.3390/ani10081348

43. Portnova-Fahreeva AA, Rizzoglio F, Nisky I, Casadio M, Mussa-Ivaldi FA, Rombokas E. Linear and non-linear dimensionality-reduction techniques on full hand kinematics. *Front Bioeng Biotechnol.* (2020) 8:429. doi: 10.3389/fbioe.2020.00429

44. Tran B, Tran D, Nguyen H, Ro S, Nguyen T. Sccan: single-cell clustering using autoencoder and network fusion. *Sci Rep.* (2022) 12:10267. doi: 10.1038/s41598-022-14218-6

45. Nasser M, Salim N, Saeed F, Basurra S, Rabiu I, Hamza H, et al. Feature reduction for molecular similarity searching based on autoencoder deep learning. *Biomolecules.* (2022) 12:508. doi: 10.3390/biom12040508

46. Lee JA, Park TY, Lee J, Moon TW, Choi J, Kang BK, et al. Developing indicators of pattern identification in patients with stroke using traditional Korean medicine. *BMC Res Notes.* (2012) 5:136. doi: 10.1186/1756-0500-5-136

47. Jang E, Lee EJ, Yun Y, Park YC, Jung IC. Suggestion of Standard Process in Developing Questionnaire of Pattern Identification. *J Physiol Pathol Korean Med.* (2016) 30:190–200. doi: 10.15188/kjopp.2016.06.30.3.190

48. Wang J, Guo Y, Li GL. Current status of standardization of traditional Chinese medicine in China. *Evid Based Complement Alternat Med.* (2016) 2016:9123103. doi: 10.1155/2016/9123103

49. Liu Y, Li Z, Xiong H, Gao X, Wu J, Wu S. Understanding and enhancement of internal clustering validation measures. *IEEE Trans Cybern.* (2013) 43:982–94. doi: 10.1109/TSMCB.2012.2220543

50. Tadist K, Najah S, Nikolov NS, Mrabti F, Zahi A. Feature selection methods and genomic big data: a systematic review. *J Big Data.* (2019) 6:79. doi: 10.1186/s40537-019-0241-0

51. Bakrania MR, Rae IJ, Walsh AP, Verscharen D, Smith AW. Using dimensionality reduction and clustering techniques to classify space plasma regimes. *Front Astron Space Sci.* (2020) 7:80. doi: 10.3389/fspas.2020.593516

52. Weiss K, Khoshgoftaar TM, Wang D, A. Survey of Transfer Learning. *J Big Data.* (2016) 3:9. doi: 10.1186/s40537-016-0043-6

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 30 August 2022
DOI 10.3389/fmed.2022.955765

OPEN ACCESS

EDITED BY
Ming-Chin Lin,
Taipei Medical University, Taiwan

REVIEWED BY
Mehrad Aria,
Azarbaijan Shahid Madani University, Iran
Francis Jesmar Perez Montalbo,
Batangas State University, Philippines

*CORRESPONDENCE
Saleh Albahli
salbahli@qu.edu.sa

SPECIALTY SECTION
This article was submitted to Family Medicine and Primary Care, a section of the journal Frontiers in Medicine

RECEIVED 29 May 2022
ACCEPTED 21 July 2022
PUBLISHED 30 August 2022

CITATION
Albahli S and Nazir T (2022) AI-CenterNet CXR: An artificial intelligence (AI) enabled system for localization and classification of chest X-ray disease. *Front. Med.* 9:955765. doi: 10.3389/fmed.2022.955765

AI-CenterNet CXR: An artificial intelligence (AI) enabled system for localization and classification of chest X-ray disease

Saleh Albahli[1]* and Tahira Nazir[2]

[1]Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia, [2]Faculty of Computing, Riphah International University, Islamabad, Pakistan

Machine learning techniques have lately attracted a lot of attention for their potential to execute expert-level clinical tasks, notably in the area of medical image analysis. Chest radiography is one of the most often utilized diagnostic imaging modalities in medical practice, and it necessitates timely coverage regarding the presence of probable abnormalities and disease diagnoses in the images. Computer-aided solutions for the identification of chest illness using chest radiography are being developed in medical imaging research. However, accurate localization and categorization of specific disorders in chest X-ray images is still a challenging problem due to the complex nature of radiographs, presence of different distortions, high inter-class similarities, and intra-class variations in abnormalities. In this work, we have presented an Artificial Intelligence (AI)-enabled fully automated approach using an end-to-end deep learning technique to improve the accuracy of thoracic illness diagnosis. We proposed AI-CenterNet CXR, a customized CenterNet model with an improved feature extraction network for the recognition of multi-label chest diseases. The enhanced backbone computes deep key points that improve the abnormality localization accuracy and, thus, overall disease classification performance. Moreover, the proposed architecture is lightweight and computationally efficient in comparison to the original CenterNet model. We have performed extensive experimentation to validate the effectiveness of the proposed technique using the National Institutes of Health (NIH) Chest X-ray dataset. Our method achieved an overall Area Under the Curve (AUC) of 0.888 and an average IOU of 0.801 to detect and classify the eight types of chest abnormalities. Both the qualitative and quantitative findings reveal that the suggested approach outperforms the existing methods, indicating the efficacy of our approach.

KEYWORDS

DenseNet, localization, CenterNet, chest X-ray images, deep learning

Introduction

The easier availability of multimedia content such as digital images and videos has enhanced the growth of tasks performed in the field of computer vision (CV). The well-known applications of CV involve object detection (1), object tracking (2), medical image analysis (3–5), text analysis (6, 7), and video processing (8). The usage of CV

approaches in the area of medical image analysis is assisting the practitioners to perform their jobs quickly and accurately. One of such applications is chest X-ray (CXR) analysis. The CXR is the highest employed modality in the world to identify several thoracic abnormalities such as pneumonia, COVID-19, atelectasis, and lung nodule. The easier and more economic behavior of CXR leads to significant medical inspections every day (9). However, the manual examination of CXR is highly reliant on the availability of domain specialists. Moreover, the manual CXR study is a taunting and time-taking activity accompanying high chances of wrong predictions. Whereas, the automated CXR recognition system can fasten this process and increase the accuracy of the system as well.

Chest abnormalities are the major reasons of deaths and disability around the globe with about 65 million people suffering from one disease or the other and 3 million demises per year. Hence, timely identification of such diseases can save the lives of patients and protect them from painful treatment procedures (10). Therefore, to tackle the problems of manual CXR inspection, the researchers have focused their attention to present reliable automated solutions. Initially, the handcrafted feature computation approaches were used for the classification of several CXR abnormalities. Such methods are simple in nature and can work well-with a small amount of data (11, 12). However, the handcrafted key points calculation methods need extensive domain information and take huge time to produce accurate results. Furthermore, there remains always a trade-off between time complexity and classification results for such techniques. The employment of huge key points enhances the recognition power of these methods but at the cost of the increased computational burden (12). The usage of small key points causes increase in the efficiency of the hand-coded approaches but results in missing acquiring the significant aspect of image modality which in turn decreases the classification results. Due to such reasons, these methods are not found to be proficient for the CXR analysis (13).

Now, the success of Artificial Intelligence (AI)-based techniques in the automatic diagnosis of medical diseases is astonishing. AI, when applied to the medical field, helps with managing, diagnosing, and treating patients. This reduces the stress of physicians and also serves as a helping hand to them. It also helps on the administrative side by automating and managing a large portion of the administrative burden (14). Recently, the advancement of deep learning (DL) frameworks is attracting the attention of the research community to use them for digital image processing including the CXR examination (15, 16). Numerous well-explored DL models such as CNN (17) and Recurrent neural networks (RNNs) (18) are used for segmentation and classification problems. This makes deep learning a very powerful tool in healthcare, as most of the work being done is categorized as either a classification or a segmentation task. The empowerment of DL approaches has made them highly suitable for medical image analysis as these frameworks are capable of computing a more discriminative set of key point vectors without the need for area specialists. The CNN models are inspired by the working of human brains to visualize and recall several objects. The well-known CNN models i.e., VGG (19), ResNet (20), DenseNet (21), and EfficientNet (22) are highly used for several image classification tasks. Such methods can exhibit reliable performance with minimum processing time (23–25). The main idea of using the DL-based techniques for the medical image examination is that these approaches are capable of computing the fundamental information of the input samples and can deal with complex image distortions such as intensity and color variations, noise, blurring, and size changes.

Although existing techniques have acquired inspiring CXR classification results; however, there is space for enhancement both in terms of computational complexity and classification accuracy. Hence, a more comprehensive investigation of the existing traditional machine learning (ML) and DL frameworks is required that can increase the CXR-related disease classification performance. The major problem of ML methods for the CXR abnormality classification is their low effectiveness with increased computational time (26). The power of DL approaches to resolve complicated real-world issues is remarkable in comparison to human brain intellect. While the DL approach resolves problems of ML techniques, however, increased the model complexity as well. Hence, there is a need for a more robust approach to the CXR-related disease classification.

The timely and accurate classification of several CXR diseases is a complex job due to the extensive similarities found among different chest abnormalities. Besides, the incidence of noise, blurring, light variation, and intensity changes in the input samples further complicates the classification procedure. To tackle the problems of existing methods, we have presented a novel framework namely AI-CenterNet CXR to detect and classify eight types of chest abnormalities. More clearly, we have presented the DenseNet-41-based CenterNet approach, where the key points from the input samples are computed by using the DenseNet-41 model. The computed features are later localized and classified by the one-stage object detector of the CenterNet model. The experimental results show that our technique is capable of discriminating various types of chest diseases effectively under the presence of different image distortions. The key contributions of our work are:

- We proposed a novel AI-enabled framework namely AI-CenterNet CXR with DenseNet-41 as a feature extractor to enhance the identification and classification results of eight types of chest abnormalities.
- The presented method is capable of accurately locating and classifying the diseased portion from the X-ray samples because of the effectiveness of the CenterNet technique.

- We have improved the classification performance because of the ability of the AI-CenterNet CXR model to better deal with the model's over-tuned training data.
- We have presented a computationally robust model to classify several CXR abnormalities due to the one-stage object detector framework of CenterNet.
- Huge evaluation is presented, and extensive experimentation is performed against the latest approaches for the CXR disease classification on a complex dataset namely NIH Chest X-ray to show the accurateness of our approach.

Related work

A lot of research work is proposed in the area of CXR disease detection. This section provided a brief review of previous research done for the detection of multi-class chest diseases from medical images. Ayan and Ünver (27) proposed a DL-based method using Xception and Vgg16 CNN models for the diagnosis of pneumonia. Initially, different data augmentation techniques, such as rotation, zooming, and flipping, were applied to the input images to increase diversity and avoid overfitting. Then, the DL models were fine-tuned using transfer learning to extract discriminative key points. The results showed that the Xception network achieved better classification accuracy as compared to Vgg16; however, the performance can be further improved. Bhandary et al. (28) suggested a DL-based framework for the identification of pneumonia and lung cancer that included two different models. The first network was based on a modified AlexNet (MAN) model to identify pneumonia class. The second network was built using an ensemble strategy that combined handcrafted features collected by the Haralick and Hu approach (29) with deep features from the MAN model. For classification, the Support Vector Machine (SVM) classifier was employed and its performance was compared with the softmax classifier. This technique attained a classification accuracy of 97.27% using CT images from LIDC-IDRI benchmark dataset. In (30), the authors evaluated the performance of different pre-trained CNN models such as GoogLeNet, InceptionNet, and ResNet using different image sizes and transfer learning. Moreover, the network visualization was used to analyze the features learned by these models. The results showed that shallow networks, such as GoogleNet, outperform deeper network architectures for discriminating between healthy and abnormal chest X-rays. Rajpurkar et al. (31) presented a DL-based model namely CheXNet to identify different illnesses in chest. The model was comprised of 121 layers utilizing dense connectivity and batch normalization. The authors retrained the ChexNet model, which had previously been trained on ImageNet data, using the CXR dataset. This approach achieved an F1 score of 43.5% and Area Under the Receiver Operating Characteristic curve (AUROC) of 0.801. In (32), the author proposed a DL model for COVID-19 illness categorization across a wide range of other chest diseases (multi-class classification) from chest x-rays. They employed a Generative Adversarial Networks (GAN)-based approach to generate synthetic images to solve the issues of class imbalance data. The author analyzed the performance using various scenarios such as data augmentation, transfer learning, and imbalanced class data. The results showed that the ResNet-based model yields higher accuracy of 87% with balanced data. Ho and Gwak (33) designed a two-stage approach for the precise identification of 14 different diseases from chest x-ray images. Initially, the abnormal region was localized using activation weights obtained from the last convolutional layer of fine-tuned DenseNet-121 network. Then, classification was performed by using a combination of handcrafted feature extractors i.e., SIFT, HOG, LBP, GIST, and deep features. Several supervised learning classifiers such as SVM, KNN, AdaBoost, and others were used to classify hybrid features. The experimental findings showed that the Extreme Learning Machines (ELM) classifier performs well in comparison to other classifiers, with an accuracy of 0.8462. In (34), the authors developed a CNN-based network comprising three convolutional layers for the identification of 12 different diseases using the CXR samples. They investigated the performance against competitive NN and backpropagation NN with unsupervised learning. The results demonstrated that the proposed CNN attains high recognition rates and better generalization power due to robust feature learning. However, computation time and convergence iterations were slightly higher.

In (35), the authors designed a multi-scale attention network for enhanced multi-class chest disease identification accuracy. The proposed network employed DenseNet169 as a backbone with a multi-scale attention block that fused local characteristics gathered at different scales with global features. A novel loss function using perceptual and multi-label balance was also introduced to solve issues of data imbalance. This approach achieves an AUROC of 0.850 on CheXpert and 0.815 on the CXR dataset. Ma et al. (36) suggested a cross-attention-based, end-to-end architecture to address class unbalanced multi-label x-ray chest illness classification. The model comprised a feature extraction network based on densenet121 and densenet169 as its backbone and a loss function based on an attention loss and multi-label balance loss for better key point representation through mutual attention. This model showed an improved AUROC of 0.817 on the Chest X-ray14 dataset. Wang and Xia (37) presented the ChestNet model to improve the accuracy of multi-class thoracic illness diagnosis using chest radiography. The model was comprised of two sub-networks: classification and attention network. The classification network was based on a pre-trained ResNet-152 model that was used to extract unified key points. The attention network was used to investigate the relationship between class labels and abnormal regions by using the extracted key points. The suggested model outperformed the existing models in classification using the CXR dataset. Ouyang et al. (38) presented an approach to simultaneously

perform both abnormality localization and multi-label chest disease classification. The model was based on the hierarchical visual attention mechanism comprising three levels and was trained using a weakly supervised learning algorithm due to the limited number of available box annotations for the abnormal region. This approach exhibited a mean AUC score of 0.819 over the CXR dataset.

Pan et al. (39) used pre-trained DenseNet and MobileNetV2 models for categorizing chest radiographs as healthy or diseased. They evaluated these models for 14 different chest pathologies. To analyze the generalization ability, the authors utilized two different datasets. The results showed that MobileNetV2 outperformed the DenseNet model in the majority of scenarios. Albahli and Yar (40) presented a multilevel classification approach using DL to diagnose COVID-19 and other chest disorders using CXR images. Initially, the first model was used to classify the input into three classes: normal, COVID-19 affected, and other. The second model was then used to perform classification into 14 chest and associated disorders. The suggested approach was evaluated using different pre-trained DL models such as ResNet50, NasNetLarge, Xception, InceptionV3, and InceptionResNetV2. The results exhibit that ResNet50 performed best with an average accuracy of 71.905% for COVID-19 identification and 66.634% for other diseases. Alqudah et al. (41) introduced an approach for the diagnosis of bacterial and viral pneumonia from healthy chest radiographs. Initially, a modified CNN model pre-trained on other medical images was fine-tuned to learn pneumonia disease-specific features. Then, classification was performed using different classifiers such as softmax classifier, SVM, and KNN. The results exhibit that SVM outperformed the other classifiers; however, the performance was evaluated on the limited dataset. Kim et al. (42) presented an end-to-end learning approach to perform multi-label lung disease classification. Initially, the input images were preprocessed by applying crop and resize operations to remove meaningless information from images. Then, the pre-trained EfficientNetv2 model was fine-tuned using input images for the extraction of discriminative key point vector and then classified into respective classes. This method depicts improved results for three-class classification; however, the model suffers from overfitting and performance degrades on increasing the number of classes. Baltruschat et al. (43) examined the execution of various ResNet-based models for the task of multi-label chest x-ray images. The authors extended the architecture and incorporated non-image features such as the patient's age, gender, and image acquisition category in the network for improved classification. The results show that ResNet-38 with integrated meta-information performed best with an AUC of 0.727 as compared to others. Ibrahim et al. (44) presented a DL-based multi-class identification method using both CXR and CT images. The authors compared four different custom architectures based on VGG19, ResNet152V2, and Gated Recurrent Unit (GRU). The results exhibit that custom VGG-19 outperformed the other models (i.e., ResNet152V2, ResNet152V2 followed by GRU and Bi-GRU) by attaining an accuracy of 98.05% on both X-ray and CT images; however, the approach suffers from data overfitting issues. Ge et al. (45) presented a multi-label CXR disease diagnosis approach using illness and health label dependencies. The model was comprised of two distinct sub-CNNs that were trained using pairs of different loss functions, i.e., binary cross-entropy, multi-label softmax loss, and correlation loss. The authors further introduced bilinear pooling to compute meaningful features for fine-grained categorization. This method (45) exhibits an AUC of 0.8398 using ResNet as base model; however, it suffers from high computational complexity.

The studies described above have shown remarkable outcomes; however, they are limited to the identification of a few chest-related diseases and lack generalizability for the classification of multiple chest illnesses. A review of approaches for recognizing chest diseases from the literature is given in Table 1. It can be seen that there is still potential for improvement in performing multi-label chest disease classification in terms of accuracy, computation complexity, and generalization ability.

Proposed methodology

Chest X-ray disease detection is based on two essential modules: the first is the Localization of chest disease pathologies, and the other is a classification of chest disease into eight categories. The complete functionality of our novel method is described in Figure 1.

For the classification of Chest X-ray disease, we have presented the novel method named CenterNet with Densenet-41. For training of our model, we have the publicly available dataset having eight classes and also their bounding boxes values of disease pathologies. So, we can perform localization of chest X-ray disease lesions directly from images due to the availability of bounding box ground truths. The proposed CenterNet method recognizes the region of interest (ROI) in feature extraction using DenseNet-41, afterward the localized areas are classified into eight classes of chest diseases. Moreover, we have evaluated all samples as per parameters in the field of CV.

CenterNet

Feature extraction is an essential step for recognizing the regions in images and also for classification. So, efficient features are required to correctly locate the disease areas from CXR images and recognize their categories into eight classes. However, this task is challenging due to the overfitting problem which occurs because of the large feature vector. Another challenge is the skip of essential areas (such as texture, shape,

TABLE 1 A comparison of the multi-class chest disease diagnosis.

Reference	Methodology	Findings	Gaps identified
Ayan and Ünver (27)	VGG16 Xception network	Accuracy = 0.87% (VGG16) Accuracy = 0.82% (Xception network)	The accuracy can be improved by combining features from both networks
Bhandary et al. (28)	Modified AlexNet (MAN) and Haralick and Hu approach	Accuracy = 97.27%	The generalization performance of the model can be enhanced
Tataru et al. (30)	GoogLeNet, InceptionNet, and ResNet	Accuracy = 80%, F1 score of 0.66	The performance can be improved by the inclusion of a segmentation approach to allow the network to learn more disease-specific attributes
Rajpurkar et al. (31)	Novel CNN (121-layer)	F1 score = 43.5% and AUROC = 0.801	Performance requires further improvement
Albahli (32)	Novel CNN	Accuracy = 87%	Performance needs improvement
Ho and Gwak (33)	A hybrid model with a DenseNet-121 network and hand-crafted feature extractor i.e., SIFT, HOG, LBP, GIST, and different ML classifiers such as SVM, KNN, AdaBoost, and others	Accuracy = 0.8462, F1-score = 0.9413, AUC = 0.8097	Requires improvement in the generalization ability of the model
Abiyev and Ma'aita (34)	Novel CNN	Accuracy = 92.4%	The model can be made deeper to enhance performance
Xu et al. (35)	Densenet169 with multi-scale attention network	AUROC = 0.850 AUROC = 0.815	The performance can be improved further
Ma et al. (36)	Densenet121 and densenet169 with cross attention	AUROC = 0.817 AUROC = 0.775	The model is computationally complex
Wang and Xia (37)	ResNet-152 with attention network	AUC = 0.781	The model is computationally complex and suffers from high inference time
Ouyang et al. (38)	ResNet with a hierarchical visual attention mechanism	AUC = 0.819 AUC = 0.9166	The model is dependent on the availability of box annotations
Pan et al. (39)	DenseNet and MobileNetV2	AUROC = 0.924 AUROC = 0.900	The generalizability of the model requires improvement
Albahli and Yar (40)	ResNet50, NasNetLarge, Xception, InceptionV3, and InceptionResNetV2	AUC = 96.9, Sensitivity = 93.4, Specificity = 93.72	The images were segmented before the classification
Alqudah et al. (41)	Novel CNN with softmax classifier, SVM, and KNN	Accuracy = 94%, Sensitivity = 93.33%, Specificity = 96.68%	Performed classification between Normal vs. Bacterial Pneumonia vs. Viral Pneumonia classes
Kim et al. (42)	EfficientNetv2	Accuracy = 82.15%, Sensitivity = 81.40%, Specificity = 91.65%	The evaluation was performed on 4 classes only Pneumonia, Pneumothorax, Tuberculosis, and Normal class
Baltruschat et al. (43)	ResNet38, ResNet50, ResNet101	AUC = 0.822	The performance can be improved further
Ibrahim et al. (44)	CustomVGG19, ResNet152V2, ResNet152V2-GRU, and ResNet152V2-BiGRU	Accuracy = 98.05%, Recall = 98.05%, Specificity = 99.5%, F1-score = 98.24%, AUC = 99.66%	The model is evaluated only using COVID-19, Pneumonia, Lung Cancer, and Normal classes
Ge et al. (45)	ResNet and DenseNet with novel multi-loss function	AUC = 0.8398 (ResNet) AUC = 0.8392 (DenseNet)	The model is evaluated using only four classes

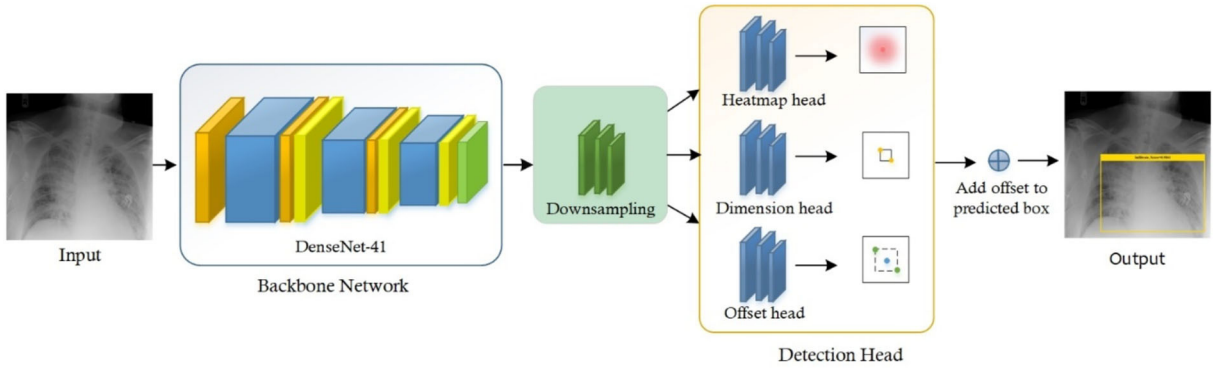

FIGURE 1
Flow diagram of proposed method namely AI CenterNet CXR.

and color changes) of the model due to the small set of the feature vector.

To accomplish the robust and efficient feature vector, it is essential to apply an automated key points extraction approach, avoiding the handcrafted feature methods. Because the handcrafted approaches of features extraction are not effective in correctly recognizing the disease lesions from the CXR images due to different variations, positions, and textures of lesions. To tackle all these problems, we have presented an efficient and novel method, which is the DL method and based on CenterNet. The presented approach named Efficient CenterNet has the ability to directly extract the features efficiently from CXR images. CenterNet has the convolution filter (CF) for key points calculation that extracts the structure of disease areas from images. The inspiration for using the one-stage method i.e., CenterNet (26) over the other object detectors e.g., RCNNs (28) and (15, 29) for chest disease identification is that these are complex structures and take more time due to the two-stage approach. Faster-RCNN uses Region Proposal Network (RPN) for localization of objects from images, then collective features intimate with each ROIs split detection heads and detect the class of object with bounding box. However, these approaches are economically not robust and are not applicable to real-world requirements of object localization. The DL approach CenterNet addresses the issues of the abovementioned methods by identifying features and also the location of ROIs in input parallelly. Moreover, the one-step technique is the ability of CenterNet that makes it more accurate and timelier efficient.

For recognizing and categorizing CXR diseases, it is challenging to locate the features of ROIs because of numerous factors i.e., finding the actual location of ROIs due to extreme color and light variations, and other is finding the category of each object. CenterNet can precisely classify and detect the disease areas of numerous categories through heatmaps, which switched the two-stage into a one-step object detector. The heatmap unit acts by utilizing the center features that accomplish greater recall values, which facilitate to decrease in the computation cost of feature extraction.

Customize centernet

The conventional CenterNet (30) used ResNet-101 for computing features to execute medical image analysis. However, this method i.e., ResNet employs skip connections to prevent non-linear conversions, which reason the immediate gradient flow from the previous to the next layers through the identity module. Figure 2 describes the Res-Net-101 technique that encompasses huge parameters and ultimately produces the vanishing gradient problem. To overcome the above issue, we proposed a DenseNet-41 for feature extraction that is densely accompanying the convolution approach. In the presented approach, DenseNet is utilized as a backbone network of CenterNet, which makes CenterNet more efficient due to a smaller number of parameters than ResNet-101. The introduced network consists of numerous Dense Blocks (D_B), which consecutively join up by additional convolutional and pooling layers among successive D_Bs. The DenseNet can exhibit the complex renovation that facilitates overwhelming the challenge of the inadequacy of the output position information for the upper-level key points, in some measure. Moreover, this method encourages feature reproduction, which makes them highly convenient for Chest X-ray disease localization and improves the training procedure. So, we introduced the DenseNet-41 (31) in CenterNet approach for feature extraction from Chest Xray Images.

DenseNet-41 feature extractor

DenseNet-41 encompasses four D_Bs along with the equal layers as employed in ResNet-101. The DenseNet-41 has less no of parameters than Resnet-101, which makes it computationally

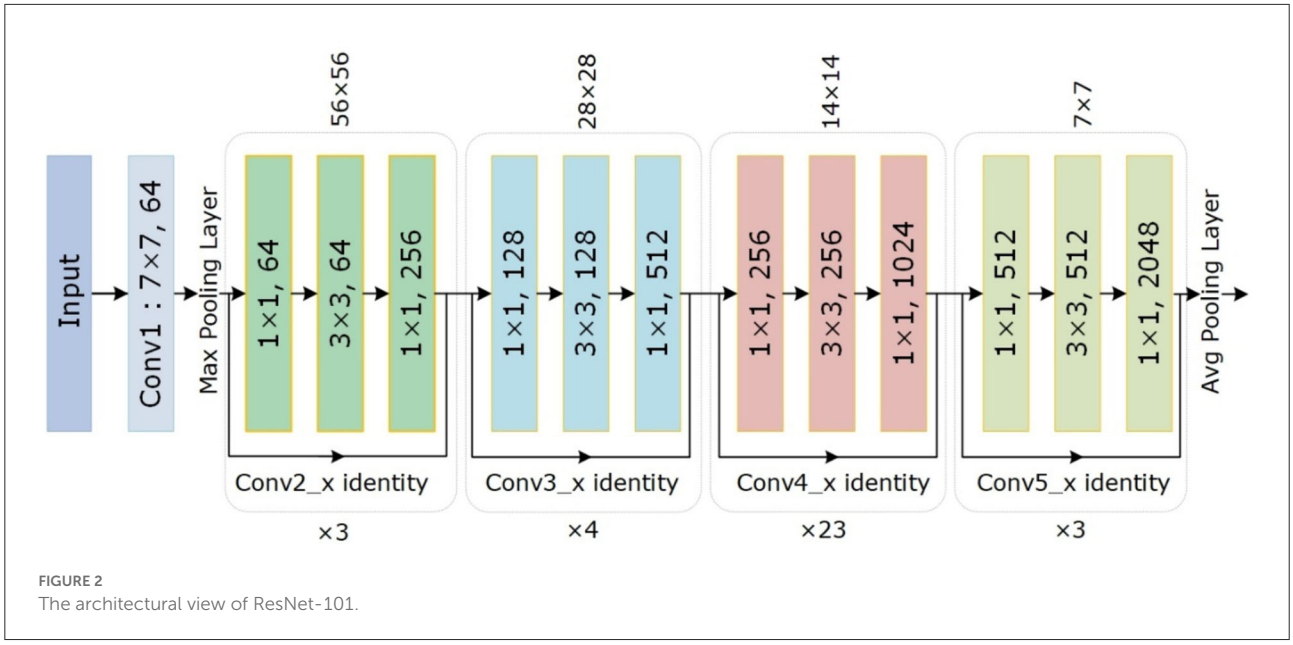

FIGURE 2
The architectural view of ResNet-101.

TABLE 2 Description of DenseNet-41.

Layer		DenseNet-41	
		Size	Stride
Con L1		7×7 *conv*	2
Pool L1		3×3 max *_pool*	2
Dense B1		$\begin{bmatrix} 1 \times 1 & conv \\ 3 \times 3 & conv \end{bmatrix} \times 3$	1
Transition L1	Con L2	1×1 *conv*	1
	Pool L2	2×2 *avg_pool*	2
Dense B2		$\begin{bmatrix} 1 \times 1 & conv \\ 3 \times 3 & conv \end{bmatrix} \times 6$	1
Transition L2	Con L3	1×1 *conv*	1
	Pool L3	2×2 *avg_pool*	2
Dense B3		$\begin{bmatrix} 1 \times 1 & conv \\ 3 \times 3 & conv \end{bmatrix} \times 6$	1
Transition L3	Con L4	1×1 *conv*	1
	Pool L4	2×2 *avg_pool*	2
Dense B4		$\begin{bmatrix} 1 \times 1 & conv \\ 3 \times 3 & conv \end{bmatrix} \times 3$	1
Classification Layer		7×7 *avg_pool*	
		FC layer	
		SoftMax	

efficient for feature computation of disease detection. Table 2 has the description of DenseNet-41, including D_Bs (as shown in Figure 3), convolutional, and transition layers (T_L).

The D_B is the vital component of DenseNet, l × l × m0 demonstrates the key points maps (KM) of the *L*-1 layer. N specifies the dimension of KM, whereas all channels are characterized by m0. P(.) is the non-linear conversion that contains different modules i.e., batch normalization (BaN), ReLU activation method, a 1 × 1 Conv layer (C_L), utilized to lessen all the channels, and a 3 × 3 C_L, used for features reorganization. Dense links are represented by long-dashed arrows, which are utilized to join the *L*-1 to the *L* layer and combined them through the result of the P(.). Lastly, l × l × (m0 + 2m) is the result of the L + 1 layer.

The numerous dense connections enhance KMs; so, the T_L is activated for reduction in feature size from the previous DB, which is briefly explained in (32, 33). The calculated key points are down-sampled with the four stride rate, after that these features are utilized for the estimation of various heads, illustrated in the proceeding subsections.

FMs increase because of vast dense links, so the T_L is represented to decrease the size of the feature map from the preceding D_B (32, 33). The feature set comes from the DenseNet-41 is put down using four stride rate and then transfer to calculate the several heads which are explained below:

Heatmap head

This head offers a key points approximation on the reduced deep key points from the DenseNet-41 to find the diseased portions with their category. The respective features are the center of bbox when localize the ROIs can be calculated as follows:

$$\hat{O}_{i,j,c} = \exp\left(-{}^{(i-\hat{p}_i)^2+(j-\hat{p}_j)^2}\big/_{2\sigma_p^2}\right) \quad (1)$$

where i and j are the original feature values, $\hat{p}_i$ and $\hat{p}_j$ are the positions of estimated down-sampled features, σ_p displays

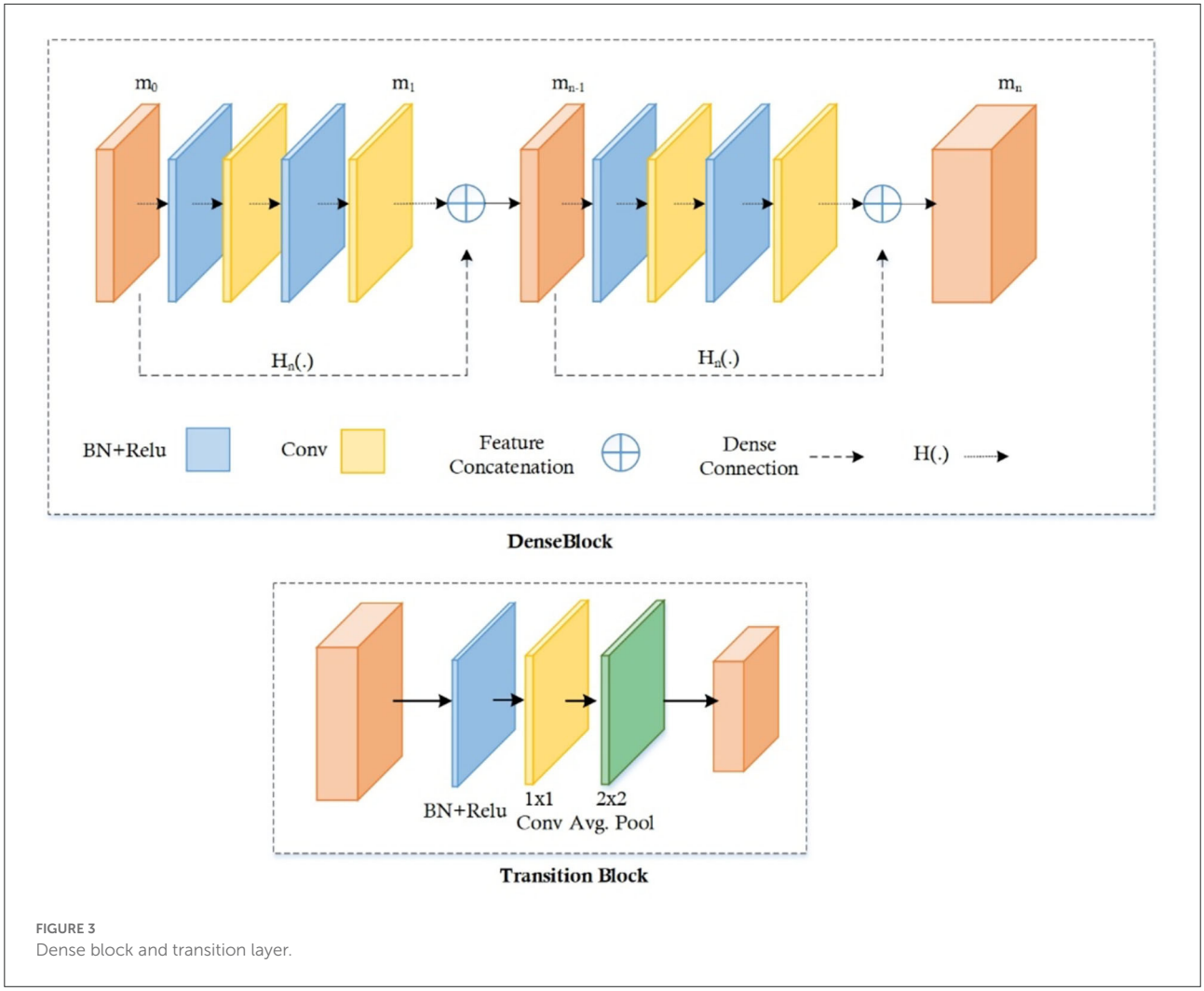

FIGURE 3
Dense block and transition layer.

the region size-adaptive standard deviation, c is the total of categories, and $o_{x,y,c}$ shows the center for a candidate features, in case it is marked as 1 means affected; or else, considered as healthy.

Dimension head

This type of head is utilized for the prediction of values of bbox, which is responsible for computing the dimensions of the box. The width and height of the bbox can be computed by the L1 norm i.e., $(x_2\text{-}x_1, y_2\text{-}y_1)$, for the k object with values (x_1, x_2, y_1, y_2).

Offset head

After applying down-sampling on input images, the discretization error appears that needs to be minimized. So, the offset head is calculated for this purpose and then the center points are again represented in the high-resolution input image.

Multitask loss

Multitask loss is the technique to improve the performance of DL-based approaches like CenterNet, our proposed technique used this type of loss for performance enhancement with accurate localization and classification of disease regions. So, the multitask loss is represented with L on every head, which can be estimated as follows:

$$L_{centernet} = L_{map} + \lambda_{\dim} L_{\dim} + \lambda_{off} L_{off} \tag{2}$$

The total loss calculated by our method is $L_{CenterNet}$, in which heatmaps, offset, and dimension head losses are described by L_{map}, $L_{\dim}$, and L_{off}, respectively. And $\lambda_{\dim}$ and λ_{off} are equal to constant values of 0.1 and 1 simultaneously.

The L_{map} is calculated through the following equation:

$$L_{map} = \frac{-1}{n} \sum_{i,j,c} \begin{cases} (1 - \hat{O}_{i,j,c})^{\alpha} \log(\hat{O}_{i,j,c}) \; if \; \hat{O}_{i,j,c} = 1 \\ (1 - O_{i,j,c})^{\beta} (\hat{O}_{i,j,c})^{\alpha} \log(1 - \hat{O}_{i,j,c}) \; otherwise \end{cases} \tag{3}$$

The total key points are shown by n. $O_{i,j,c}$ is the center of the original feature center, whereas $\mathbf{o}_{i,j,c}$ is the estimated value of the center. Hypermeters of loss in our case is described by α and β having the values of 2 and 4 for the whole test.

The L_{dim} can be estimated by using the Equation 4,

$$L_{\text{dim}} = \frac{1}{n}\sum_{k=1}^{n}\left|\hat{b}_k - b_k\right| \tag{4}$$

where b_k is the actual and $\hat{b_k}$ is the predicted bbox coordinates, total samples are shown by n.

Ultimately, the L_{off} is determined by the Equation 5:

$$Loff = \frac{1}{n}\sum_{p}\left|\hat{F}_{\hat{p}} - (^{p}/_{R} - \hat{p})\right| \tag{5}$$

The predicted offset rate is denoted by $\hat{F}$, while R is the resultant stride. The real key point is p, while $\hat{p}$ is the down-sampled value.

Creation of bounding box

Lastly, the estimated values with each category are processed separately which are gained through heatmaps. In this work, we have utilized the 8 nearest neighbors, and then the highest 100 values are considered.

Let $\hat{Q}$ is producing N-related center points of class c using Equation 6:

$$Q = \left\{(\hat{x}_j, \hat{y}_j)\right\}_{j=1}^{N} \tag{6}$$

where the location of every estimated point is symbolized as $(\hat{x}_j, \hat{y}_j)$. We have utilized all values of key points denoted by $\hat{Q}$, and bbox and coordinates can be found through Equation 7:

$$\begin{aligned}&(\hat{x_j} + \partial\hat{x_j} - \hat{w}_j/2,\ \hat{y_j} + \partial\hat{y_j} - \hat{h_j}/2,\\ &\hat{x_j} + \partial\hat{x_j} + \hat{w}_j/2,\ \hat{y_j} + \partial\hat{y_j} + \hat{h_j}/2)\end{aligned} \tag{7}$$

In Equation 7, $(\partial\hat{x_j}, \partial\hat{y_j})$ = offset prediction, while $(\hat{w_j},\ \hat{h_j})$ = size prediction.

The final bbox is created immediately from the valuation of the features with no usage of IoU-based non-maxima suppression.

Detection process

CenterNet is an efficient technique as compared to other methods, which are explained in previous sections. So, in this method, input X-ray image along their bbox is given to the trained framework, whereas the CenterNet estimates its center values of disease regions. The complete flow of the introduced solution is described in Algorithm 1.

Input	1. CXR images from the NIH dataset
	2. Eight categories of diseases are nominated i.e., Atelectasis (AT), Cardiomegaly (CD), Effusion (EF), Infiltration (IF), Mass (M), Nodule (ND), Pneumonia (PN), and Pneumothorax (PX)
	3. Bonding boxes containing the region of interest
Output	4. Localized region identifying the diseased area
	5. Output label with a classification score
	6. Trained model
Environment	7. Python with TensorFlow and others requires libraries
	8. GPU-based machine
Configuration	9. Importing samples
	10. Distribution of dataset into train, validation, and test sets
Data Configuration	11. Batch size = 16
Directories Configurations	12. Generate 2-folders of the samples with their output labels and bounding box values employed for the model training and validation, respectively
Training and Testing	13. Create the CenterNet model with the Dense-41 base network and fine-tuned it on the NIH images to perform the model training.
	14. Samples from the test set are used to evaluate the trained model performance for the CXR disease classification.
Validation	15. Compilation of model with 25 epochs along with the 0.001 learning rate
	16. The multi-loss function is used to measure three types of losses i.e., heatmaps, offset and dimension head losses for model performance optimization
Evaluation	17. Measure model performance by using standard metrics:
	• mAP • IOU • Confusion matrix • Precision • Recall • Accuracy • F1-Score • Error rate • AUC • Test time

Algorithm 1. Flow of the introduced method.

Experiment and results

In this portion of the paper, we have provided detailed information about the dataset being used for the model verification. Further, we have elaborated on the evaluation measures that are used to compute the quantitative results of our approach. Besides, extensive experiments have been performed to test the proposed approach in numerous ways to show its robustness for CXR disease detection and classification. We have performed the experiments in Python language by using an Nvidia GTX1070 GPU-based system. In the presented technique for CXR recognition, the CenterNet model is employed with pre-trained weights obtained from the MS-COCO dataset, and transfer learning is carried out on the NIH X-ray dataset to modify it for the chest disease classification.

Dataset

For model training and testing, we have used a standard dataset of CXR namely the NIH Chest X-ray dataset (46). The employed database comprises a total of 112,120 samples from 30,805 subjects. The details about the entire NIH CXR dataset are shown in Figure 4. The outer layer in the figure shows the number of images in the respective class, and the second outer layer represents all the 14 classes. The complete dataset has 14 classes, however, the dataset contains the annotations for eight types of chest diseases such as AT, CD, EF, IF, M, ND, PN, and PX, respectively. There are a total of 984 annotated samples available for model training, which are marked by a panel of radiologists. As the proposed work is concerned with the employment of an object detection-based model for the CXR classification, therefore, we have considered the abovementioned eight diseases for our approach. A few samples from the NIH CXR dataset are presented in Figure 5. The used dataset is quite complex in nature due to the presence of intense light variations, noise, blurring, color changes, and class imbalance problems.

Performance metrics

To assess the CXR detection and classification performance of the proposed Custom-CenterNet model, we have utilized several standard metrics used in the area of object detection and classification domain. We have used the mean average precision (mAP), Intersection over Union (IOU), precision, accuracy, and recall, metrics for performance analysis. The mathematical description of the accuracy measure is given in Equation 8:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

Equation 9 depicts the mathematical formulation of AP, and equation 10 is the mAP measure, where AP shows the average precision for all classes and t is the test sample. T is representing all test samples:

$$AP = \int_0^1 p(r)dr \tag{9}$$

Here, p(r) is the accuracy of the target area or detection:

$$mAP := \sum_{i=1}^{T} AP(t_i)/T \tag{10}$$

Figures 6–8 explain the visual demonstration form of IOU, precision, and recall, respectively.

Localization results

An effective CXR disease classification should be capable of correctly recognizing and classifying all categories of diseases. For this reason, we have performed an analysis to check the CXR abnormalities detection and classification performance of our approach. The test images from the NIH CXR dataset are applied to confirm the localization and categorization power of the custom CenterNet approach, and visual samples are reported in Figure 9. We have reported some test results in Figure 9 for all eight classes, which include class labels and confidence scores. The first row is showing the localization results of the Atelectasis class, the second row is for the Cardiomegaly class. Similarly, the remaining six rows in Figure 9 show the detection results for Effusion, Infiltration, Mass, Nodule, Pneumonia, and Pneumothorax, respectively. From localization results, we analyze that this dataset has both smaller and larger disease regions such as Effusion and Nodule diseases have smaller affected areas, while others have larger affected regions. So, our model can detect both the smaller and larger regions precisely with better results. The samples shown in Figure 9 having different intensity variations are depicting that our model can accurately identify the diseased portion and can differentiate several chest diseases efficiently. Moreover, the model is capable of reliably locating the diseased portion for the distorted samples, which are depicting the robustness of our method. For example, in Figure 9, the second case of the last row has a smaller region and is also similar to the background area, but our method detected it accurately. To numerically discuss the localization ability of the DenseNet41-based CenterNet approach, we have computed the mAP score which is the standard evaluation metric and we have acquired the mAP score of 0.91. From both the visual and quantitative results analysis, we can say that the proposed custom CenterNet

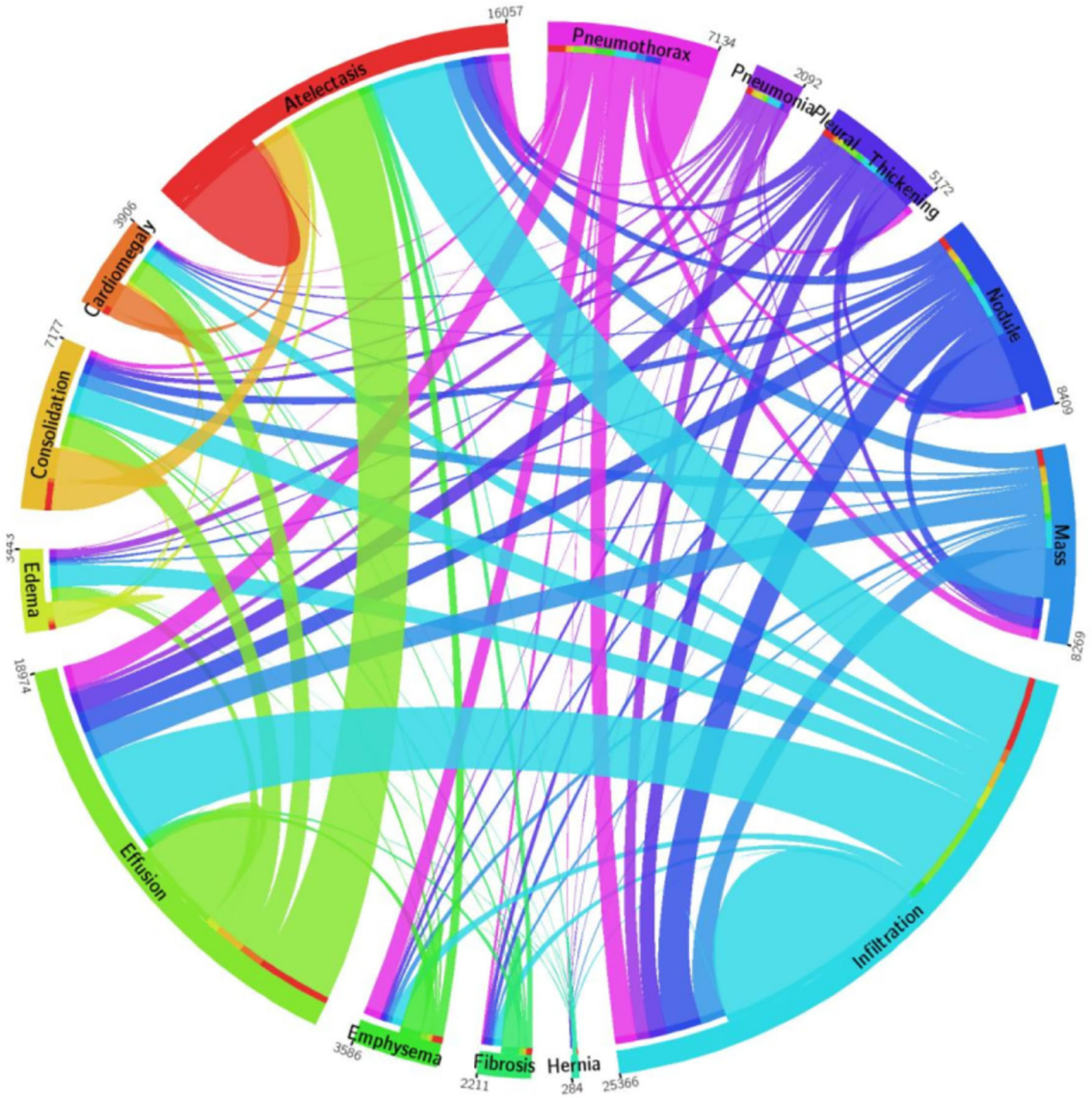

FIGURE 4
A pictorial view of sample information from the NIH Chest X-ray dataset (47).

model can be reliably applied for the CXR disease identification and classification.

Class-wise results

Here, we have elaborated the class-wise results of our approach to elaborate the recognition power of our approach in categorizing eight types of chest diseases from the X-ray image modality. For this reason, we employed the DenseNet41-based CenterNet framework on all the suspected samples from the NIH CXR database and computed the performance in the forms of precision, recall, accuracy, and F1 measure.

Firstly, we have reported the category-wise obtained precision values for our approach as this metric permits us to check how much a model is competent in discriminating the diseased images from the normal samples. The acquired results are shown in Figure 10 from where it is quite visible that our approach has correctly detected the affected samples. More clearly, we have obtained the average precision value of 89%, which is showing the efficacy of the presented technique.

Moreover, we have computed the recall evaluation metric as it allows us to analyze how much a framework is capable of differentiating the different diseases from each other. The obtained AP and recall values are shown in Figure 11, which is clearly showing that our proposed model is empowered to

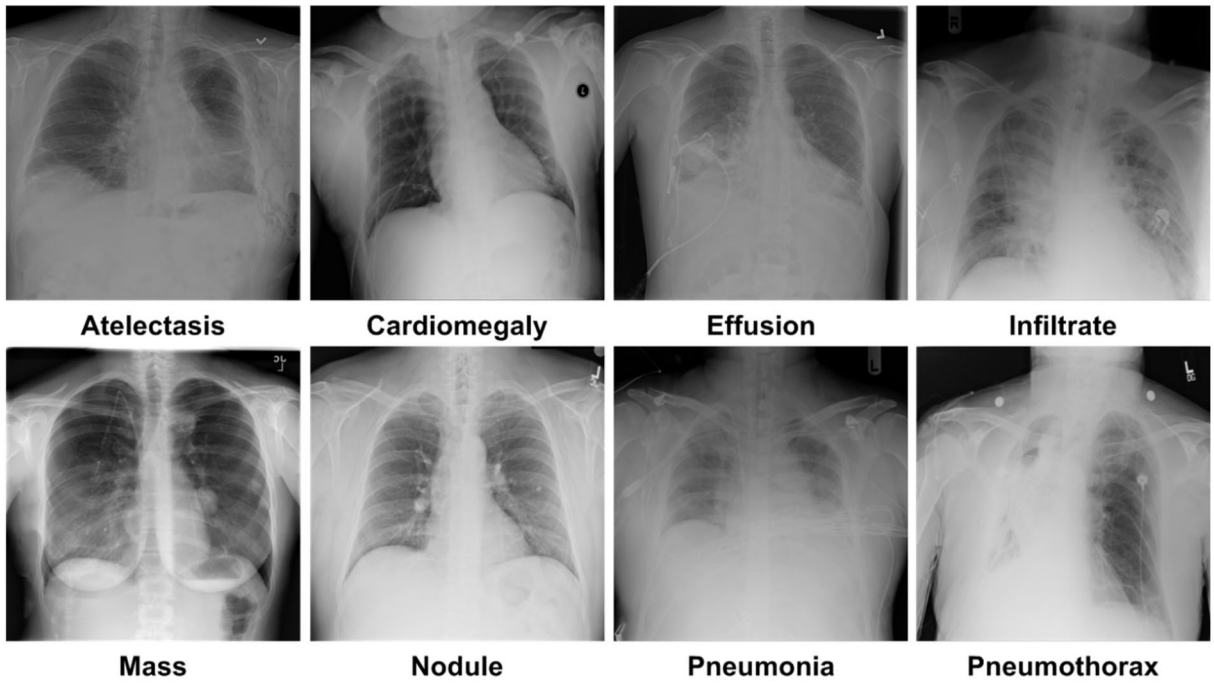

FIGURE 5
Samples images of NIH CXR dataset.

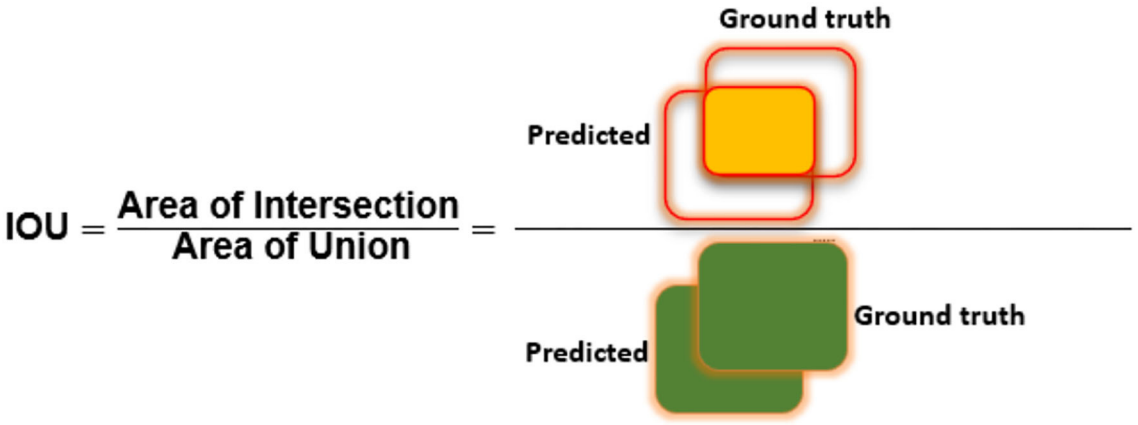

FIGURE 6
Visual depiction of IOU metric.

correctly recognize all eight types of CXR abnormalities and shows an average recall value of 91%.

Furthermore, we have computed the F1-Score as the more the value of the F1-Score the better the model performance. The calculated F1-Score along with the error rate for all eight classes of CXR abnormalities are shown in Figure 12. The custom CenterNet approach shows the maximum F1 score of 94.30% along with the minimum error rate of 5.70% for the Pneumothorax class while reporting the lowest F1-Score of 87.88% along with a maximum error rate of 12.12% for the Nodule abnormality. More clearly, we have attained the average F1-Score and error rate of 89.99 and 10.01%, respectively.

Furthermore, we have reported the confusion matrix to further demonstrate the CXR abnormality categorization power of the proposed approach as the confusion matrix is capable of showing the classification performance of a model in a viable manner by showing the actual and predicted values. More

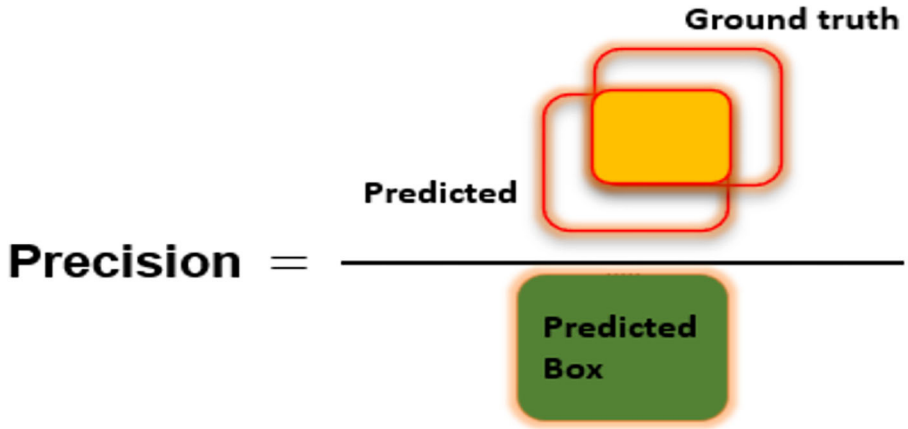

FIGURE 7
Visual demonstration of Precision metric.

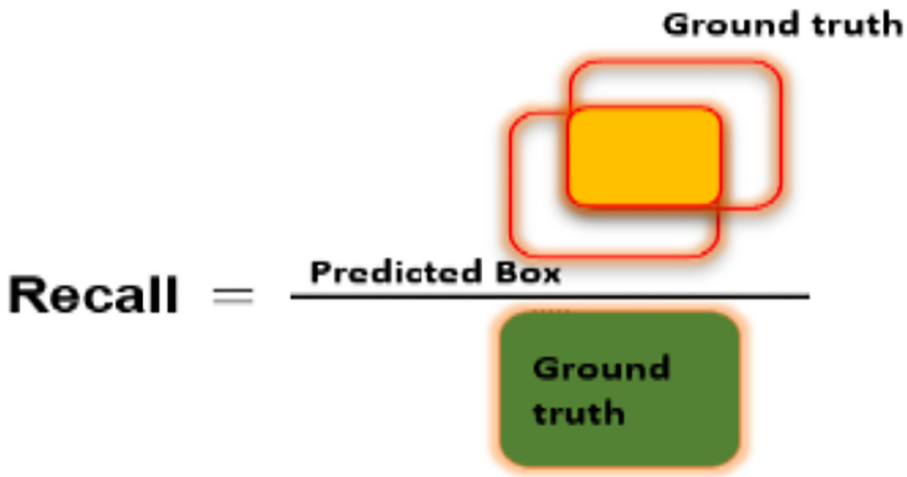

FIGURE 8
Pictorial representation of Recall measure.

descriptively, we have acquired the True Positive Rate (TPR) of 89.55, 87.65, 87.69, 93.33, 90.87, 89.98, 93.78, and 94.82%. It is quite evident from Figure 13 that the presented method can efficiently discriminate the affected regions of several classes of CXR diseases.

Finally, we have calculated the accuracy values for all eight classes of the CXR diseases, and values are shown in Figure 14 from where it is quite evident that the proposed approach shows robust classification results for all classes. More clearly, we have acquired an average accuracy value of 92.21%. Based on the conducted analysis, we can say that our approach shows better classification performance in terms of all performance measures due to its efficient feature computation power.

Evaluation of proposed model

In this section, we have provided a comparison of the classification results of our approach against other DL-based methods. For this reason, we have selected the AlexNet (48), GoogleNet (49), VGG16 (50), and ResNet50 (51) models by considering their results for the CXR disease classification as mentioned in (52).

Initially, we performed the class-wise performance analysis of our approach with the nominated DL approaches, and the results are elaborated in Table 3. It can be seen from the table that the DenseNet41-based CenterNet model has outperformed the other approaches for all categories of diseases. More clearly, for the AT and CD diseases, the selected DL methods show

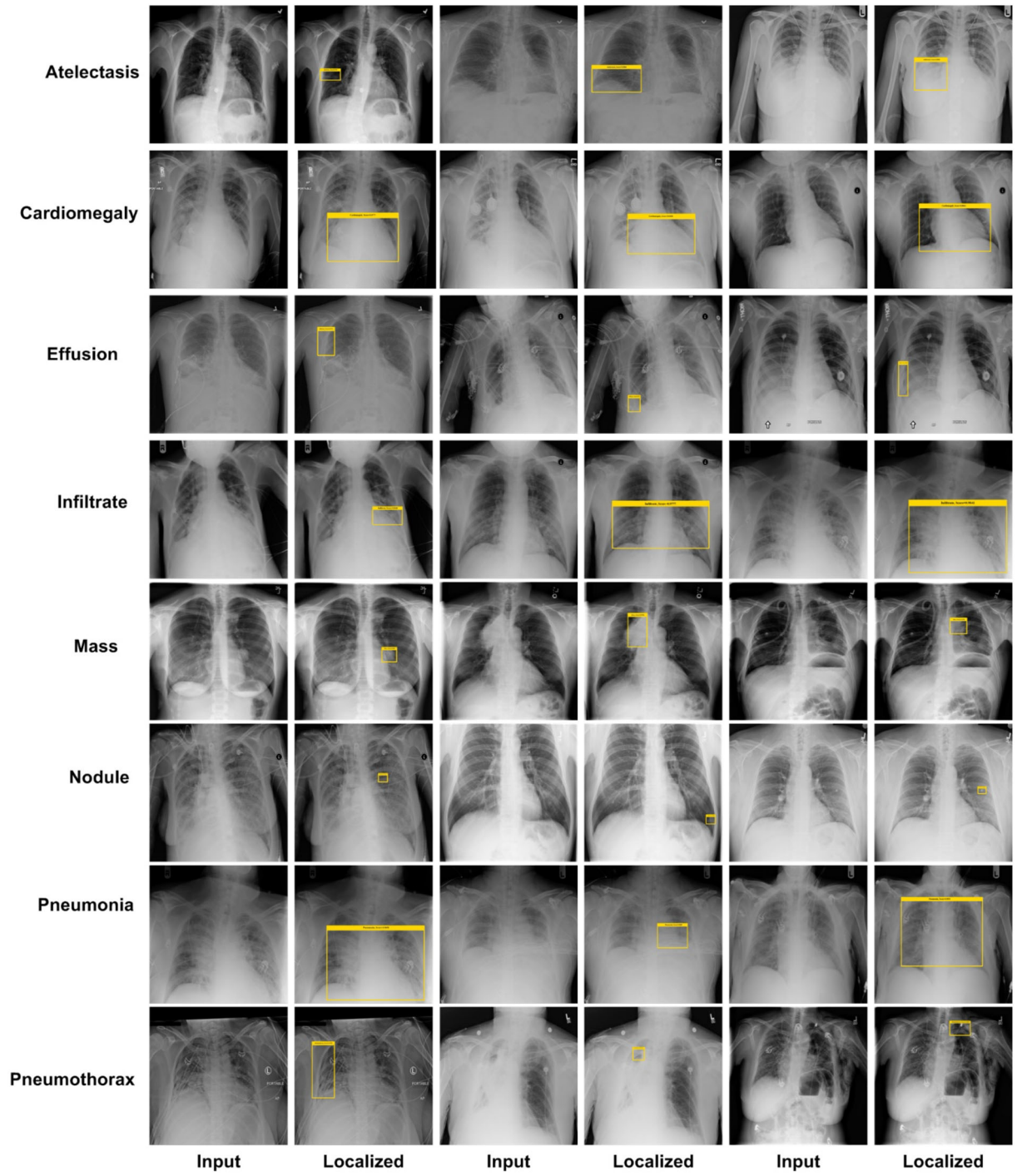

FIGURE 9
Localization results of the proposed method.

the average values of 0.65 and 0.73 which are 0.88 and 0.95 for our case. So, for the AT and CD diseases classification, we have shown performance gains of 22.75 and 22.25%, respectively. While for the EF, IF, and M chest diseases, we have given the average values of 0.78, 0.91, and 0.85, while the comparative methods show the average values of 0.68, 0.60, and 0.54, respectively, so we have shown the performance gains of 24.5, 17.75, and 36.75% for the mentioned diseases, respectively. Similarly, for the ND, PN, and PX chest diseases, the peer approaches report the average values of 0.645, 0.57, and 0.735,

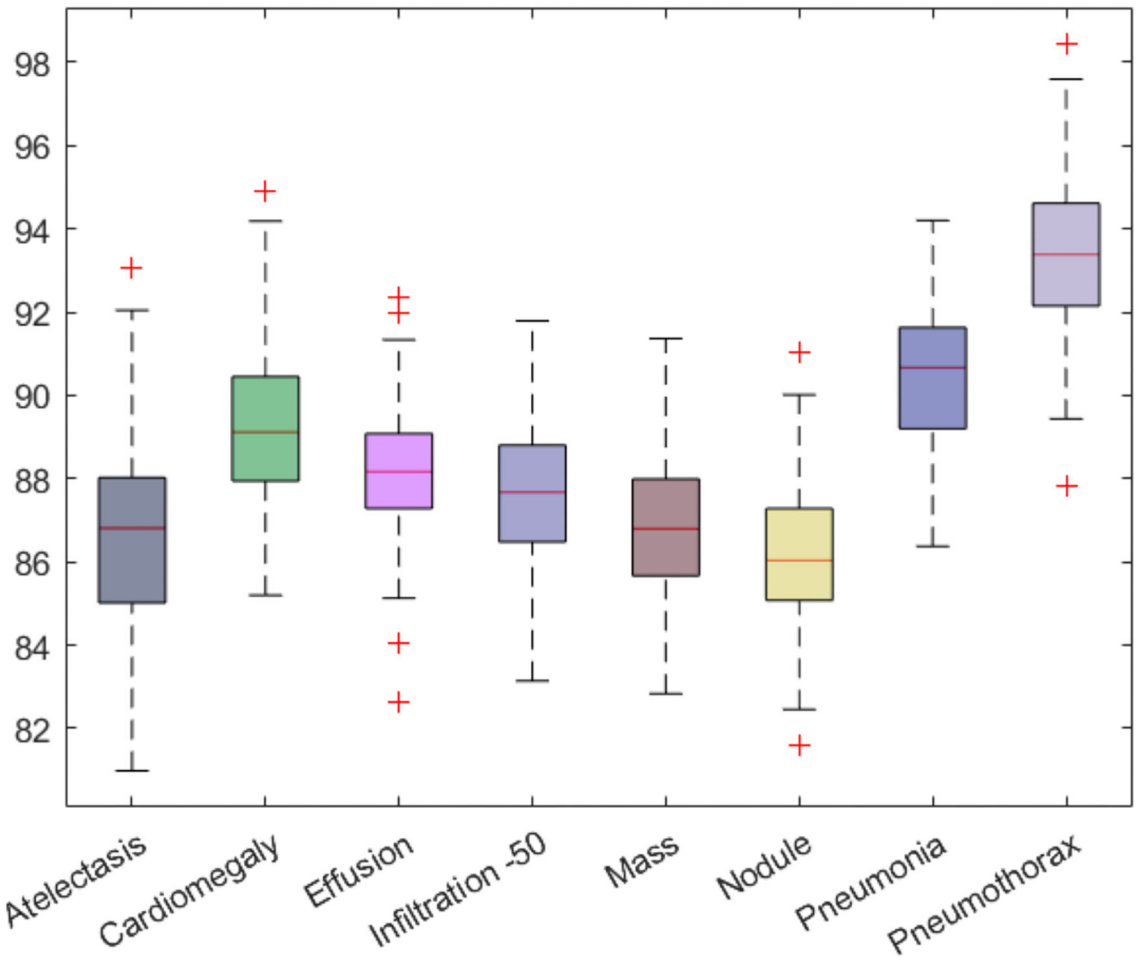

FIGURE 10
Class-wise precision values for the custom CenterNet model.

which are 0.85, 0.84, and 0.96 for the proposed approach. Hence, we have presented the 20.5, 27, and 22.5% of performance gains for the ND, PN, and PX chest disease classification, respectively. Entirely, for all diseases, the competent methods attain the average AUC value of 0.645, while our work acquires 0.887, hence we have provided an overall performance gain of 24.20%.

In the second phase, we assessed the custom CenterNet approach with the nominated DL approaches by comparing the results on the entire dataset using several standard metrics, namely, precision, recall, accuracy, and F1-measure. The comparative analysis is shown in Table 4 from where it is quite clear that the proposed framework is more efficient for CXR abnormality categorization. We have obtained the highest performance values for all the evaluation measures with the values of 89, 91, 92.21, and 89.99% for the precision, recall, accuracy, and F1-Score, respectively. The second largest results are shown by the EfficientNet with the values 87.74, 88.95, 88.01, and 87.61% for the precision, recall, accuracy, and F1-Score respectively. DenseNet-121 attained better results, however, this model is computationally complex as compared to our proposed DenseNet-41. Furthermore, the ResNet50 model the values of 77, 75, 77.63, and 75.99% for the precision, recall, accuracy, and F1-Score, respectively. Moreover, the AlexNet model shows the lowest classification results with values of 65, 66.14, 67.45, and 65.57% for the precision, recall, accuracy, and F1-Score, respectively. From the conducted analysis, we can say that the proposed DenseNet41-based CenterNet model is quite efficient to recognize each category of chest diseases and show robust performance on the entire dataset as compared to the other DL-based approaches. The main cause for the enhanced classification results of our model is because of the usage of the DenseNet41 as its base network, as this model employs the shallow network architecture which permits it to select a more reliable set of images key points. While comparatively, the selected DL-based approaches are quite complex in structure and unable to perform well for the samples with intense light, and color variations causes decrease in their performance for the CXR abnormalities recognition. So, we can say that our model

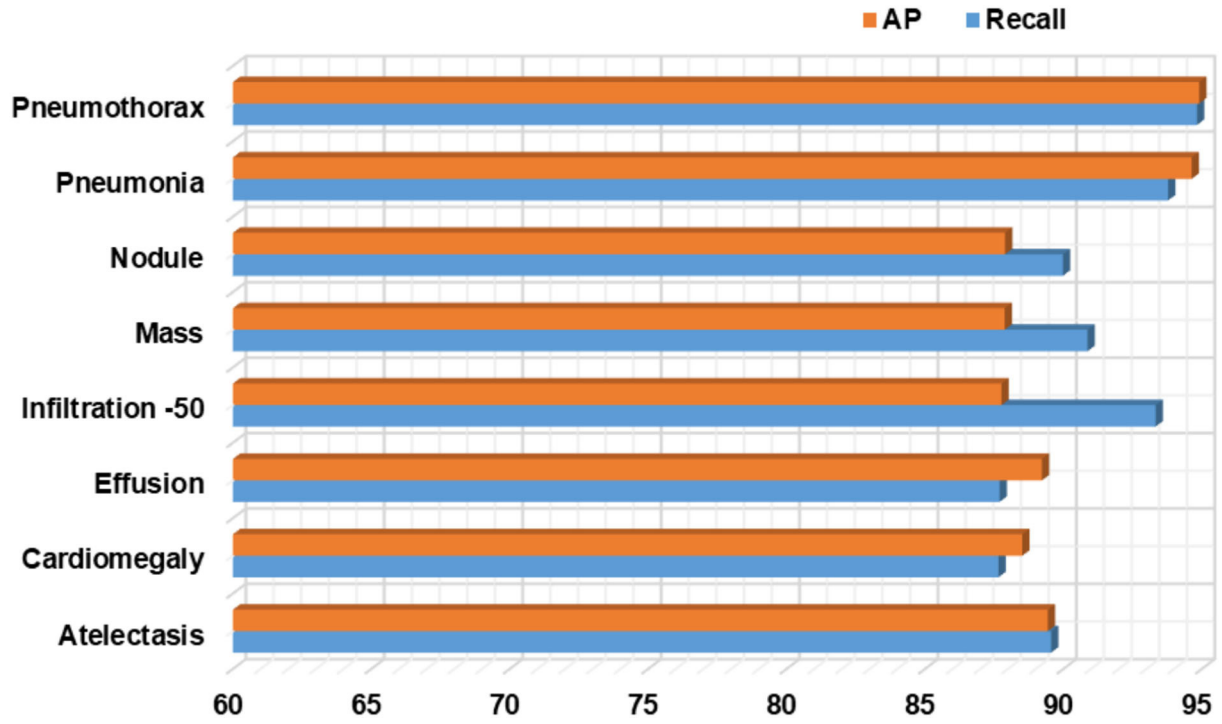

FIGURE 11
Class-wise AP and recall values for the proposed Custom CenterNet approach.

presents an efficient and effective solution for classifying chest disease from the X-ray image modality.

Comparison with other object detection models

Here, we experimented to analyze the results of our approach by comparing it against several other DL-based object recognition approaches for the CXR abnormality categorization. For this reason, we have taken both the one and two-stage techniques. The major distinction between the one and two-stage object detection models is that in the case of two-stage approaches, initially numerous region proposals are created to identify the location of the diseased portion, and then the associated class is determined. While for the one-stage object detection methods, the position and class of RoI are determined in a single step. In the case of two-stage approaches, we have chosen the Fast-RCNN (53), Faster-RCNN (4, 54), and Mask-RCNN (55) models, while for the other, we have taken the RetinaNet (56) and conventional CenterNet (21) models.

For performance comparison, we have used the mAP performance measure as it is the highly designated metric used in the area of object recognition. Additionally, the test time of all competitor methods is also considered to discuss the computational efficiency as well. The obtained comparison is shown in Table 5 from which it is quite evident that our approach is proficient for CXR disease classification both in terms of performance results and test time with the values of 0.91 and 0.21 s, respectively. The Fast-RCNN model employs the hardcoded-based approaches for its key points computation that are unable to tackle the image distortions reliably. The Faster-RCNN and Mask-RCNN approaches have tackled the issues of the Fast-RCNN model; however, these are computationally inefficient due to their two-stage networks. Whereas, the RetinaNet approach is unable to learn the discriminative anchors for the acentric key points of suspected samples. We also compared our model with the YOLO object detector, it achieved a 0.76 mAP value and the test time is 0.22 s. This model is faster, however, attained a low localization rate because it strives to detect small regions of disease from the images.

The conventional CenterNet model shows better performance; however, still unable to generalize to real-world scenarios due to its high computational cost. The proposed approach that is the DenseNet41-based model has better addressed the limitations of existing approaches by identifying the diseased portion in a more viable manner. The major cause for the better performance of our model is due to the employment of the DenseNet41 model as a feature extractor, which empowers it to better designate the image features which in turn enhances its recognition power and reduces its time complexity as well.

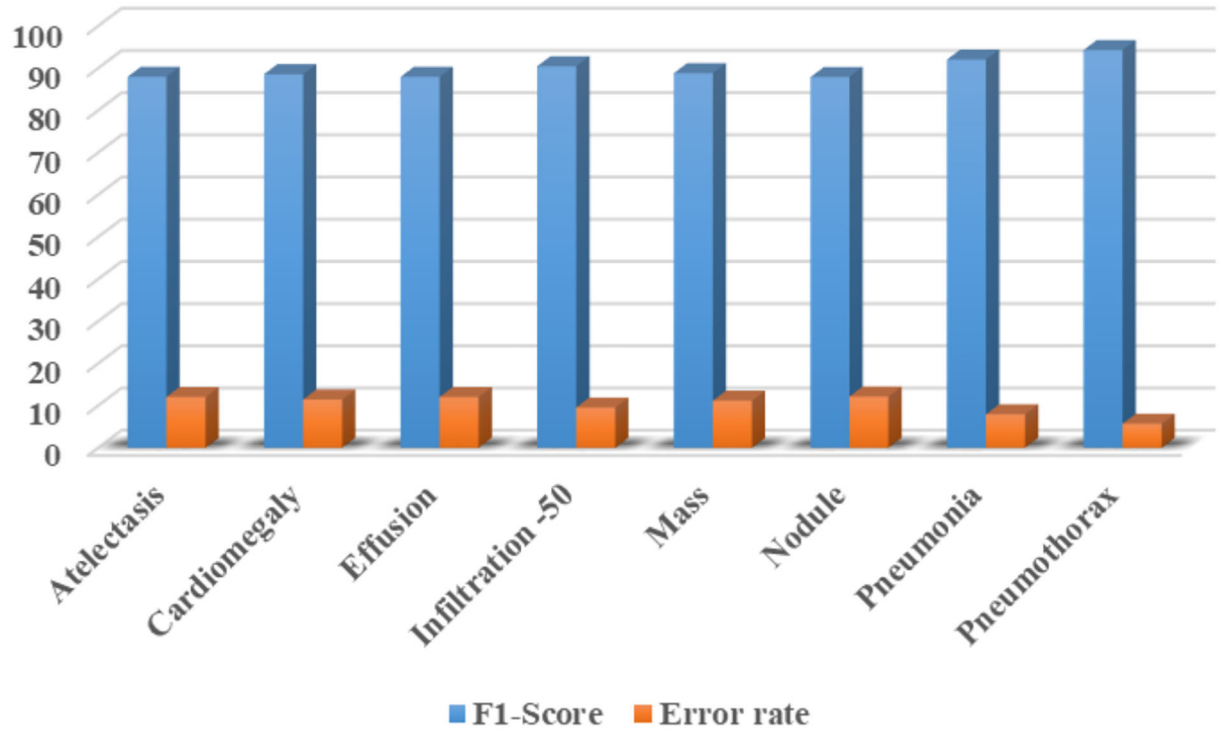

FIGURE 12
Class-wise F1-Score along with the error rate for CXR diseases classification using custom CenterNet model.

True Class \ Predicted Class	Atelectasis	Cardiomegaly	Effusion	Infiltration-50	Mass	Nodule	Pneumonia	Pneumothorax
Atelectasis	89.55%	1.20%	1%	1.30%	2.20%	2.10%	1.40%	1.25%
Cardiomegaly	1.20%	87.65%	1.37%	1.50%	2.72%	2.60%	1.90%	1.06%
Effusion	2.30%	1.60%	87.69%	1%	1.85%	2.70%	1.52%	1.34%
Infiltration-50	2.40%	2.70%	0.50%	93.33%	0.50%	0.07%	0.25%	0.25%
Mass	1.70%	2.10%	1.81%	1%	90.870%	2.33%	0.10%	0.09%
Nodule	1.10%	1.75%	2.26%	1.42%	1.50%	89.98%	1.03%	0.96%
Pneumonia	1%	1.65%	2.57%	0.25%	0.32%	0.20%	93.78%	0.23%
Pneumothorax	0.75%	1.35%	2.80%	0.20%	0.04%	0.02%	0.02%	94.82%

FIGURE 13
Confusion matrix obtained for CXR disease classification with the custom CenterNet.

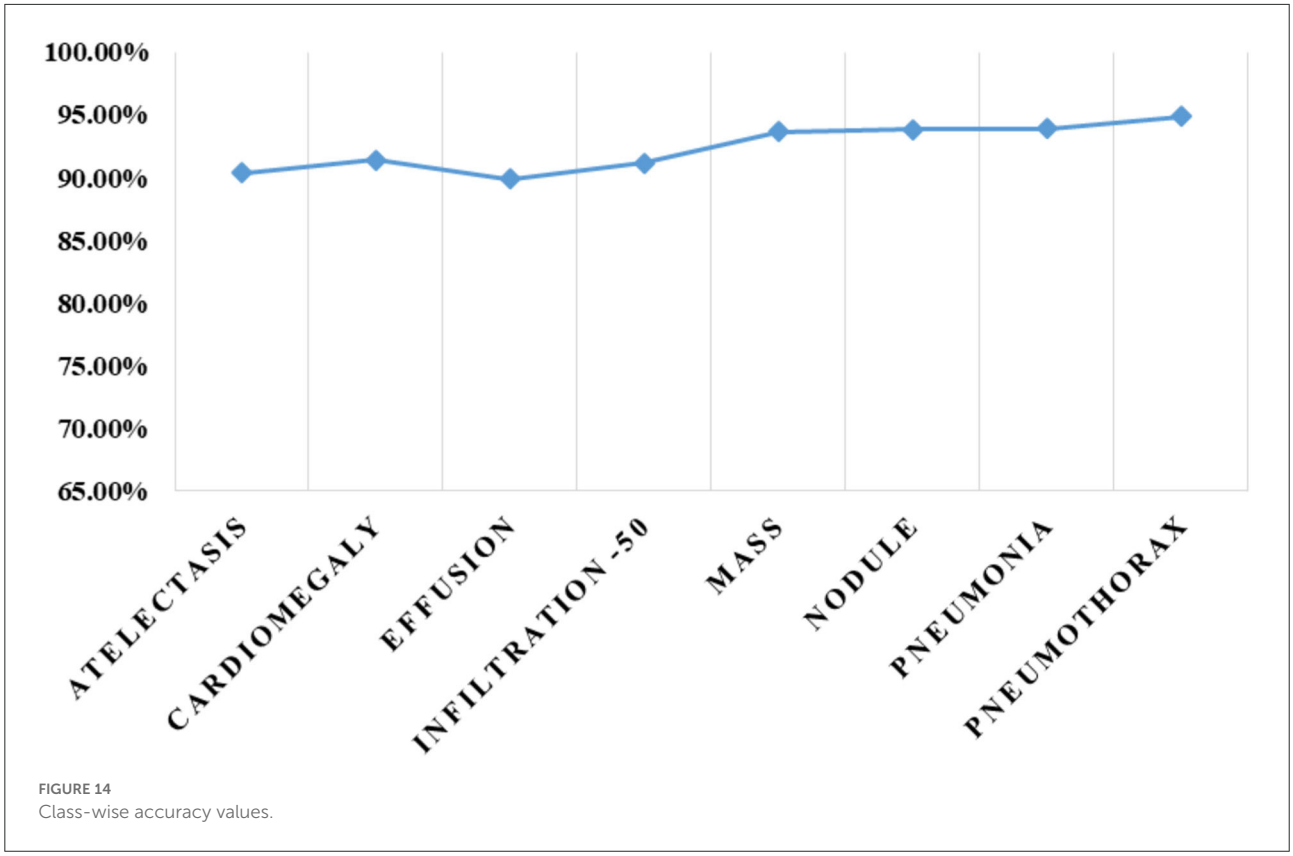

FIGURE 14
Class-wise accuracy values.

TABLE 3 Comparison with base models in terms of the AUC metric.

Model	Atelectasis	Cardiomegaly	Effusion	Infiltrate	Mass	Nodule	Pneumonia	Pneumothorax
AlexNet	0.64	0.69	0.66	0.60	0.56	0.65	0.55	0.74
GoogLeNet	0.630	0.70	0.69	0.61	0.54	0.56	0.59	0.78
VGG16	0.63	0.71	0.65	0.59	0.51	0.65	0.51	0.63
ResNet50	0.71	0.81	0.74	0.61	**0.56**	0.72	0.63	0.79
Proposed	**0.88**	**0.95**	**0.93**	**0.78**	**0.91**	**0.85**	**0.84**	**0.96**

The bold means highest AUC metric.

TABLE 4 Comparative comparison with base models.

Model	Precision	Recall	Accuracy	F1-Score
AlexNet	65.00%	66.14%	67.45%	65.57%
GoogLeNet	69.53%	71.88%	70.35%	70.69%
VGG16	72.00%	74.32%	75.41%	73.14%
ResNet-50	77.00%	75.00%	77.63%	75.99%
Inception V4	79.32%	75.65%	79.32%	79.22%
DenseNet-121	83.01%	81.84%	83.21%	82.87%
EfficientNet	87.74%	88.95%	88.01%	87.61%
Proposed	89.00%	91.00%	92.21%	89.99%

Comparative analysis against ML classifiers

We have further explained the robustness of our approach for the CXR disease recognition by evaluating its results against the Conventional ML-based classifiers. For this reason, we have nominated two renowned ML classifiers named the SVM and KNN, and obtained values are shown in Table 6. The values in Table are clearly showing that the presented approach obtains the highest AUC with the value of 0.887. The second highest result is attained by the SVM classifier with the value of 0.745, while the KNN classifier shows the lowest

TABLE 5 Comparison with object detection models.

Model	Base	mAP	Test time (sec/img)
Fast-RCNN	VGG-16	0.65	0.28
Faster-RCNN	VGG-16	0.77	0.25
Mask-RCNN	ResNet-101	0.79	0.23
RetinaNet	ResNet-101	0.63	0.27
YOLO	ResNet-50	0.76	0.22
CenterNet	ResNet-101	0.82	0.25
Proposed CenterNet	DenseNet-41	0.91	0.21

TABLE 6 Comparison with ML-based classifiers.

Classifier	AUC
SVM (57)	0.745
KNN (57)	0.721
Proposed	**0.887**

The bold means highest AUC metric.

value of 0.721, respectively. More descriptively, the comparative classifiers show the average value of 0.733, which is 0.887 for the proposed work. So, we have given a performance of 15.40%. The comparative analysis is clearly depicting that the presented custom CenterNet is more proficient in classifying the several diseases of the chest from the X-ray image modality because of its high recognition ability.

Comparative analysis with state-of-the-art methods

In this part, a comparative analysis is executed in comparison to several latest approaches introduced for the CXR disease classification employing the same dataset. For a fair comparison, the highest average results reported in (52, 58–62) are taken and evaluated against our obtained average results.

Initially, we have compared the proposed approach in terms of the AUC metric and the obtained comparison is reported in Table 7. Wang et al. (58) proposed a DL-based approach for the CXR disease classification, where the CNN-RNN framework was introduced to compute the deep features from the input samples and perform the classification task. The work (58) acquired an average AUC value of 0.753. Another DL-based approach was presented in (59) employing the concept of boosted cascaded convents and attained the average AUC value of 0.778. Liu et al. (60) introduced an approach namely the Contrast-Induced Attention Network (CIA-Net) that used the concept of constructive learning to perform the CXR abnormalities recognition and show the average AUC value of 0.801. Seyyed-Kalantari et al. (61) presented a CNN-based approach to categorize several diseases of the chest via employing the X-ray modality and obtained the average AUC value of 0.821. Han et al. (62) presented a residual-based approach for recognizing several CXR diseases and acquired an average AUC value of 0.838. While in comparison, the presented approach acquired the highest value of the AUC measure with the value of 0.837. More descriptively, for the AT disease, the competent approaches show an average value of 0.786 and 0.880 in our work; hence, we presented a performance gain of 9.40%. For the CD, EF, and IN classes, the competitor methods show the average values of 0.894, 0.856, and 0.698, respectively, which are 0.99, 0.93, and 0.95 for our technique. Therefore, for the CD, EF, and IN classes, the custom CenterNet approach shows the average performance gains of 9.6, 7.4, and 15.2%, respectively. Similarly, for the M, ND, PN, and PX classes, the presented framework provides the average performance gains of 10.2, 10.8, 9.6, and 0.4%, respectively. While collectively, the approaches in (58–62) show the average AUC value of 0.789, while our method shows the average AUC value of 0.888 and presented the performance gain of 8.98%, which is showing the robustness of our approach for the CXR abnormalities classification.

Secondly, the performance comparison of our work in terms of IOU is discussed against the latest methods reported in (52), and obtained comparison is presented in Table 8. Wang et al. (52) introduced a deep CNN model for identifying and classifying the CXR diseases and attained the average IOU value of 0.569. Similarly, a CNN-based approach was introduced in (62) and acquired an average IOU value of 0.746. Li et al. (63) proposed a Residual-based approach for classifying the CXR abnormalities and attained an average IOU value of 0.728. In comparison, our proposed custom CenterNet model exhibits the average IOU value of 0.801 which is the greatest among all peer methods. More clearly, the peer techniques show the average IOU value of 0.681 which is 0.801 for the proposed solution. Hence, for the IOU measure, the custom CenterNet model gives the average performance gain of 12%.

From the conducted analysis, it is quite clear that the proposed approach for the CXR disease classification is more competent in terms of both IOU and AUC evaluation measures as compared to the latest approaches. The major reason for the robust recognition power of the proposed solution is due to the more discriminative feature computation ability of our model, which assists it to recognize all categories of disease in an efficient manner. While in comparison, the approaches in (52, 58–62) are quite complex in structure which results in the model over-fitting issue. Moreover, the approaches are unable to deal with several distortions of suspected samples such as color and light variations which make them inefficient to capture the image information accurately. While in comparison, our technique is more effective to tackle the transformation changes in the suspected samples. Hence, we can say that the presented custom CenterNet is more competent for CXR disease recognition and categorization.

TABLE 7 Comparison of latest approaches in terms of the AUC metric.

Approach	Atelectasis	Cardiomegaly	Effusion	Infiltrate	Mass	Nodule	Pneumonia	Pneumothorax
Wang et al. (58)	0.73	0.84	0.79	0.67	0.73	0.69	0.72	0.85
Kumar et al. (59)	0.76	0.91	0.86	0.69	0.75	0.67	0.72	0.86
Liu et al. (60)	0.79	0.87	0.88	0.69	0.81	0.73	0.75	0.89
Seyyed-Kalantari et al. (61)	0.81	0.92	0.87	0.72	0.83	0.78	0.76	0.88
Han et al. (62)	0.84	0.93	0.88	0.72	0.87	0.79	0.77	**0.90**
Proposed	**0.88**	**0.99**	**0.93**	**0.95**	**0.90**	**0.84**	**0.84**	0.88

The bold means highest AUC metric.

TABLE 8 Comparison of latest techniques in terms of the IOU metric.

Approach	Atelectasis	Cardiomegaly	Effusion	Infiltrate	Mass	Nodule	Pneumonia	Pneumothorax
Wang et al. (52)	0.69	0.94	0.66	0.71	0.40	0.14	0.63	0.38
Han et al. (62)	0.72	0.96	0.88	0.93	0.74	0.45	0.65	0.64
Li et al. (63)	0.71	0.98	0.87	0.92	0.71	0.40	0.60	0.63
Proposed	**0.76**	**0.99**	**0.93**	**0.95**	**0.74**	**0.56**	**0.74**	**0.74**

The bold means highest IOU metric.

Conclusion

In our work, we presented AI-CenterNet CXR, an end-to-end DL-based framework for the automated recognition and categorization of thoracic illness from chest radiographs. Our method is based on a CenterNet model that uses the DenseNet network for the computation of effective image attributes. More specifically, we integrated the DenseNet-41 network to extract a discriminative set of key points from the chest x-rays for the accurate identification of abnormalities. Moreover, due to the one-stage object detector framework CenterNet model, the suggested architecture is computationally robust to classify several CXR abnormalities. We conducted extensive experiments using the NIH CXR dataset to show the effectiveness of the proposed approach. Our technique attained an overall AUC of 0.888, an average precision value of 89%, a recall value of 91%, and an IOU of 0.801 to identify and classify eight categories of chest illness. According to the results, the proposed technique outperforms existing approaches in terms of both time and computational complexity. Moreover, the approach can correctly identify the aberrant regions and categorize the various types of chest illness in the presence of distortions, significant inter-class similarities, and intra-class variances. In the future, we will incorporate fourteen classes and perform experiments on other latest DL-based models.

Data availability statement

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

Author contributions

SA: conceptualization, methodology, software, and writing-original draft preparation. TN: data curation, writing-original draft preparation, validation, supervision, and writing-reviewing and editing. All authors contributed to the article and approved the submitted version.

Acknowledgments

The researchers would like to thank the Deanship of Scientific Research, Qassim University for funding the publication of this project.

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. Nawaz M, Nazir T, Javed A, Tariq U, Yong HS, Khan MA, et al. An efficient deep learning approach to automatic glaucoma detection using optic disc and optic cup localization. *Sensors*. (2022) 22:434. doi: 10.3390/s22020434

2. Nawaz M, Mehmood Z, Nazir T, Masood M, Tariq U, Munshi AM, et al. Image authenticity detection using DWT and circular block-based LTrP features. *CMC-Comp Mat Cont*. (2021) 69:1927–44. doi: 10.32604/cmc.2021.018052

3. Nawaz M, Nazir T, Masood M, Mehmood A, Mahum R, Khan MA, et al. Analysis of brain MRI images using improved cornernet approach. *Diagnostics*. (2021) 11:1856. doi: 10.3390/diagnostics11101856

4. Nawaz M, Masood M, Javed A, Iqbal J, Nazir T, Mehmood A, et al. Melanoma localization and classification through faster region-based convolutional neural network and SVM. *Multim Tools Appl*. (2021) 809:28953–74. doi: 10.1007/s11042-021-11120-7

5. Nawaz M, Nazir T, Masood M. Glaucoma detection using tetragonal local octa patterns and SVM from retinal images. *Int Ar J Inform Technol*. (2021) 18:686–93. doi: 10.34028/iajit/18/5/8

6. Albahli S, Nawaz M, Javed A, Irtaza A. An improved faster-RCNN model for handwritten character recognition. *Ar J Sci Engin*. (2021) 46:8509–23. doi: 10.1007/s13369-021-05471-4

7. Albahli S, Nazir T, Mehmood A, Irtaza A, Alkhalifah A, Albattah W. AEI-DNET: a novel densenet model with an autoencoder for the stock market predictions using stock technical indicators. *Electronics*. (2022) 11:611. doi: 10.3390/electronics11040611

8. Rafique R, Nawaz M, Kibriya H, Masood M. DeepFake detection using error level analysis and deep learning. *2021 4th Int Conf on Comp Inform Sci (ICCIS)*. (2021) 4:1–4. doi: 10.1109/ICCIS54243.2021.9676375

9. de Moura J, Novo J, Ortega M. Fully automatic deep convolutional approaches for the analysis of COVID-19 using chest X-ray images. *App Soft Comp*. (2022) 115:108190. doi: 10.1016/j.asoc.2021.108190

10. Gayathri J, Abraham B, Sujarani M, Nair MS. A computer-aided diagnosis system for the classification of COVID-19 and non-COVID-19 pneumonia on chest X-ray images by integrating CNN with sparse autoencoder and feed forward neural network. *Comp Biol Med*. (2022) 141:105134. doi: 10.1016/j.compbiomed.2021.105134

11. Bures M, Klima M, Rechtberger V, Ahmed BS, Hindy H, Bellekens X. Review of specific features and challenges in the current internet of things systems impacting their security and reliability. *World Conf Inform Sys Technol*. (2021):546–56: Springer. doi: 10.1007/978-3-030-72660-7_52

12. Bures M, Macik M, Ahmed BS, Rechtberger V, Slavik P. Testing the usability and accessibility of smart tv applications using an automated model-based approach. *IEEE Transact Consum Elect*. (2020) 66:134–43. doi: 10.1109/TCE.2020.2986049

13. Tan T, Das B, Soni R, Fejes M, Yang H, Ranjan S, et al. Multi-modal trained artificial intelligence solution to triage chest X-ray for COVID-19 using pristine ground-truth, vs. radiologists. *Neurocomputing*. (2022) 485:36–46. doi: 10.1016/j.neucom.2022.02.040

14. Ayalew M, Salau AO, Abeje BT, Enyew B. Detection and classification of COVID-19 disease from X-ray images using convolutional neural networks and histogram of oriented gradients. *Biomed Signal Process Control*. (2022) 74:103530. doi: 10.1016/j.bspc.2022.103530

15. Dewi RC, Chen X, Jiang, H, Yu. Deep convolutional neural network for enhancing traffic sign recognition developed on Yolo V4. *Multimedia Tools Appls*. (2022) 3:1–25. doi: 10.1007/s11042-022-12962-5

16. Aria M, Nourani E, Golzari Oskouei A. ADA-COVID: adversarial deep domain adaptation-based diagnosis of COVID-19 from lung CT scans using triplet embeddings. *Comp Intell Neurosci*. (2022) 2022:640. doi: 10.1155/2022/2564022

17. Fukushima KJ. Biological cybernetics neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Comp Coop Neural Nets*. (1980) 36:193–202. doi: 10.1007/BF003 44251

18. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. *IEEE Transact Signal Process*. (1997) 45:2673–81. doi: 10.1109/78.650093

19. M. Nawaz et al. Skin cancer detection from dermoscopic images using deep learning and fuzzy k-means clustering. *Micros Res Tech*. (2022) 85:339–51. doi: 10.1002/jemt.23908

20. Wu Z, Shen C, Van Den Hengel A. Wider or deeper: revisiting the resnet model for visual recognition. *Patt Recog*. (2019) 90:119–33. doi: 10.1016/j.patcog.2019.01.006

21. Albattah W, Nawaz M, Javed A, Masood, Albahli S. A novel deep learning method for detection and classification of plant diseases. *Comp Intell Sys*. (2022) 8:507–24. doi: 10.1007/s40747-021-00536-1

22. Marques G, Agarwal D, de la Torre Díez I. Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network. *App Soft comp*. (2020) 96:106691. doi: 10.1016/j.asoc.2020.106691

23. Gupta S, Gupta, Katarya R. InstaCovNet-19: a deep learning classification model for the detection of COVID-19 patients using chest X-ray. *App Soft Comp*. (2021) 99:106859. doi: 10.1016/j.asoc.2020.106859

24. Demir F. DeepCoroNet: a deep LSTM approach for automated detection of COVID-19 cases from chest X-ray images. *App Soft Comp*. (2021) 103:107160. doi: 10.1016/j.asoc.2021.107160

25. Pathan S, Siddalingaswamy P, Ali T. Automated detection of Covid-19 from Chest X-ray scans using an optimized CNN architecture. *App Soft Comput*. (2021) 104:107238. doi: 10.1016/j.asoc.2021.107238

26. Nazir T, Nawaz M, Javed A, Malik KM, Saudagar AK, Khan MB, et al. COVID-DAI: a novel framework for COVID-19 detection and infection growth estimation using computed tomography images. *Micros Res Tech*. (2022) 85:2313–30. doi: 10.1002/jemt.24088

27. Ayan E, Ünver HM. Diagnosis of pneumonia from chest X-ray images using deep learning. *2019 Scien Meeting Electrical-Electronics Biomed Engin Comp Sci (EBBT)*. (2019) 8:1–5: doi: 10.1109/EBBT.2019.8741582

28. Bhandary A, Prabhu GA, Rajinikanth V, Thanaraj KP, Satapathy SC, Robbins DE, et al. Deep-learning framework to detect lung abnormality–A study with chest X-Ray and lung CT scan images. *Pattern Recog Lett*. (2020) 129:271–8. doi: 10.1016/j.patrec.2019.11.013

29. Huang Z, Leng J. Analysis of Hu's moment invariants on image scaling and rotation. In: *2010 2nd International Conference on Computer Engineering and Technology. Vol. 7*. Chengdu: IEEE (2010). p. V7-476–480.

30. Tataru C, Yi D, Shenoyas A, Ma A. Deep learning for abnormality detection in chest X-Ray images. In: *IEEE Conference on Deep Learning*. Adelaide: IEEE (2017).

31. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv [Preprint]*. (2017) 25:05225. doi: 10.48550/arXiv.1711.05225

32. Albahli S. A deep neural network to distinguish covid-19 from other chest diseases using x-ray images. *Curr Med Imag*. (2021) 17:109–19. doi: 10.2174/1573405616666200604163954

33. Ho TK, Gwak J. Multiple feature integration for classification of thoracic disease in chest radiography. *App Sci*. (2019) 9:4130. doi: 10.3390/app9194130

34. Abiyev RH, Ma'aita MK. Deep convolutional neural networks for chest diseases detection. *J Healthcare Engin*. (2018) 2018:8538. doi: 10.1155/2018/4168538

35. Xu J, Li H, Li X. MS-ANet: deep learning for automated multi-label thoracic disease detection and classification. *PeerJ Comp Sci*. (2021) 7:e541. doi: 10.7717/peerj-cs.541

36. Ma C, Wang H, Hoi SC. Multi-label thoracic disease image classification with cross-attention networks. *Int Conf Med Image Comp Com-Assist Intervent*. (2019):730–8. doi: 10.1007/978-3-030-32226-7_81

37. Wang H, Xia Y. Chestnet: a deep neural network for classification of thoracic diseases on chest radiography. *arXiv [Preprint]*. (2018) 24:03058. doi: 10.1109/JBHI.2019.2928369

38. Ouyang X, Karanam S, Wu Z, Chen T, Huo J, Zhou XS, et al. Learning hierarchical attention for weakly-supervised chest X-ray abnormality localization and diagnosis. *IEEE Transacts medical Imag*. (2020) 40:2698–710. doi: 10.1109/TMI.2020.3042773

39. Pan I, Agarwal S, Merck D. Generalizable inter-institutional classification of abnormal chest radiographs using efficient convolutional neural networks. *J Digit Imag*. (2019) 32:888–96. doi: 10.1007/s10278-019-00180-9

40. Albahli S, Yar GN. Fast and accurate detection of covid-19 along with 14 other chest pathologies using a multi-level classification: algorithm development and validation study. *J Med Int Res*. (2021) 23:e23693. doi: 10.2196/23693

41. Alqudah M, Qazan S, Masad IS. Artificial intelligence framework for efficient detection and classification of pneumonia using chest radiography images. *J Med Biol Engin*. (2021) 41:99–609. doi: 10.21203/rs.3.rs-66836/v2

42. Kim S, Rim B, Choi S, Lee A, Min S, Hong M. Deep learning in multi-class lung diseases' classification on chest X-ray images. *Diagnostics*. (2022) 12:915. doi: 10.3390/diagnostics12040915

43. Baltruschat M, Nickisch H, Grass M, Knopp T, Saalbach A. Comparison of deep learning approaches for multi-label chest X-ray classification. *Sci Rep*. (2019) 9:1–10. doi: 10.1038/s41598-019-42294-8

44. Ibrahim DM, Elshennawy EM, Sarhan AM. Deep-chest: multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases. *Comp Biol Med*. (2021) 132:104348. doi: 10.1016/j.compbiomed.2021.104348

45. Ge D, Mahapatra X, Chang Z, Chen L, Chi, Lu H. Improving multi-label chest X-ray disease diagnosis by exploiting disease and health labels dependencies. *Multi Tools Appl*. (2020) 79:14889–902. doi: 10.1007/s11042-019-08260-2

46. Peng Y, Wang X, Lu, L Bagheri M, Summers R, Lu Z. NegBio: A high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits Transl Sci Proc*. (2018) 2018:188. doi: 10.7717/peerj-cs.913

47. Summers R. *NIH Chest X-ray Dataset of 14 Common Thorax Disease Categories* (2019).

48. Iandola FN, Han S, Moskewicz MW, Ashraf K, W. J. Dally, and K. Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv [Preprint]*. (2016). doi: 10.48550/arXiv.1602.07360

49. Ballester P, Araujo RM. On the performance of GoogLeNet and AlexNet applied to sketches. *Thirtieth AAAI Conf Artif Intell*. (2016) 3:10171. doi: 10.1609/aaai.v30i1.10171

50. Qassim H, Verma A, Feinzimer D. Compressed residual-VGG16 CNN model for big data places image recognition. *2018 IEEE 8th Annual Comput Commun Workshop Conf (CCWC)*. (2018) 5:169–75. doi: 10.1109/CCWC.2018.8301729

51. Theckedath D, Sedamkar R. Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks. *SN Comp Sci*. (2020) 1:1–7. doi: 10.1007/s42979-020-0114-9

52. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *Proceed IEEE Conf Comp Vision Pattern Recog*. (2017) 5:2097–106. doi: 10.1109/CVPR.2017.369

53. Girshick R. Fast r-cnn. *Proceed IEEE Int Conf Comp Vision*. (2015) 2:1440–8. doi: 10.1109/ICCV.2015.169

54. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transact Patt Anal Mach Intel*. (2016) 39:1137–49. doi: 10.1109/TPAMI.2016.2577031

55. Masood M, Nazir T, Nawaz M, Javed A, Iqbal M, Mehmood A. Brain tumor localization and segmentation using mask RCNN. *Front Comput Sci*. (2021) 15:1–3. doi: 10.1007/s11704-020-0105-y

56. Wang Y, Wang C, Zhang H, Dong Y, Wei S. Automatic ship detection based on RetinaNet using multi-resolution Gaofen-3 imagery. *Rem Sens*. (2019) 11:531. doi: 10.3390/rs11050531

57. Mohameth F, Bingcai C, Sada KA. Plant disease detection with deep learning and feature extraction using plant village. *J Comp Commun*. (2020) 8:10–22. doi: 10.4236/jcc.2020.86002

58. Wang X, Peng Y, Lu L, Lu Z, Summers RM. Tienet: text-image embedding network for common thorax disease classification and reporting in chest x-rays. *Proceed IEEE Conf Comp Vision Pattern Recog*. (2018) 43:9049–58. doi: 10.1109/CVPR.2018.00943

59. Kumar P, Grewal M, Srivastava MM. Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs. *Int Conf Image Ana Recog*. (2018) 56:546–52. doi: 10.1007/978-3-319-93000-8_62

60. Liu G, Zhao Y, Fei M, Zhang Y, Wang, Yu Y. Align, attend, and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision. *Proceed IEEE/CVF Int Conf Comp Vision*. (2019) 32:10632–41. doi: 10.1109/ICCV.2019.01073

61. Seyyed-Kalantari G, Liu M, McDermott IY, Chen M, Ghassemi. CheXclusion: fairness gaps in deep chest X-ray classifiers. in *Biocomputing 2021: Proceedings of the Pacific Symposium*. *World Sci*. (2020) 4:232–43. doi: 10.1142/9789811232701_0022

62. Han Y, Chen C, Tewfik A, Glicksberg B, Ding Y, Peng Y, et al. Knowledge-augmented contrastive learning for abnormality classification and localization in chest X-rays with radiomics using a feedback loop. *Proc IEEE/CVF Winter Conf Appl Comput Vis*. (2022). 7:2465–74. doi: 10.1109/WACV51458.2022.00185

63. Li Z, Wang C, Han M, Xue Y, Wei W, Li LJ, et al. Thoracic disease identification and localization with limited supervision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake, UT: IEEE (2018). p. 8290–9.

frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 04 October 2022
DOI 10.3389/fpubh.2022.984621

OPEN ACCESS

EDITED BY
Ming-Chin Lin,
Taipei Medical University, Taiwan

REVIEWED BY
Wenke Cheng,
Leipzig University, Germany
Yunchao Xie,
University of Missouri, United States

*CORRESPONDENCE
Dehua Yu
shgprc@yeah.net
Zhaoxin Wang
supercell002@sina.com
Jianwei Shi
shijianwei_amy@126.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Family Medicine and Primary Care, a section of the journal Frontiers in Public Health

RECEIVED 02 July 2022
ACCEPTED 12 September 2022
PUBLISHED 04 October 2022

CITATION
Chen N, Fan F, Geng J, Yang Y, Gao Y, Jin H, Chu Q, Yu D, Wang Z and Shi J (2022) Evaluating the risk of hypertension in residents in primary care in Shanghai, China with machine learning algorithms.
Front. Public Health 10:984621.
doi: 10.3389/fpubh.2022.984621

Evaluating the risk of hypertension in residents in primary care in Shanghai, China with machine learning algorithms

Ning Chen[1†], Feng Fan[2†], Jinsong Geng[3], Yan Yang[4], Ya Gao[1], Hua Jin[5,6,7,8], Qiao Chu[1], Dehua Yu[5,6,7,8*], Zhaoxin Wang[9,10,11*] and Jianwei Shi[5,6,10*]

[1]School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [2]School of Medicine, Tongji University, Shanghai, China, [3]School of Medicine, Nantong University, Nantong, China, [4]School of Economics and Management, Tongji University, Shanghai, China, [5]Department of General Practice, Yangpu Hospital, Tongji University School of Medicine, Shanghai, China, [6]Shanghai General Practice and Community Health Development Research Center, Shanghai, China, [7]Academic Department of General Practice, Tongji University School of Medicine, Shanghai, China, [8]Clinical Research Center for General Practice, Tongji University, Shanghai, China, [9]The First Affiliated Hospital of Hainan Medical University, Haikou, China, [10]Department of Social Medicine and Health Management, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [11]School of Management, Hainan Medical University, Haikou, China

Objective: The prevention of hypertension in primary care requires an effective and suitable hypertension risk assessment model. The aim of this study was to develop and compare the performances of three machine learning algorithms in predicting the risk of hypertension for residents in primary care in Shanghai, China.

Methods: A dataset of 40,261 subjects over the age of 35 years was extracted from Electronic Healthcare Records of 47 community health centers from 2017 to 2019 in the Pudong district of Shanghai. Embedded methods were applied for feature selection. Machine learning algorithms, XGBoost, random forest, and logistic regression analyses were adopted in the process of model construction. The performance of models was evaluated by calculating the area under the receiver operating characteristic curve, sensitivity, specificity, positive predictive value, negative predictive value, accuracy and F1-score.

Results: The XGBoost model outperformed the other two models and achieved an AUC of 0.765 in the testing set. Twenty features were selected to construct the model, including age, diabetes status, urinary protein level, BMI, elderly health self-assessment, creatinine level, systolic blood pressure measured on the upper right arm, waist circumference, smoking status, low-density lipoprotein cholesterol level, high-density lipoprotein cholesterol level, frequency of drinking, glucose level, urea nitrogen level, total cholesterol level, diastolic blood pressure measured on the upper right arm, exercise frequency, time spent engaged in exercise, high salt consumption, and triglyceride level.

Conclusions: XGBoost outperformed random forest and logistic regression in predicting the risk of hypertension in primary care. The integration of this risk assessment model into primary care facilities may improve the prevention and management of hypertension in residents.

KEYWORDS

hypertension, risk assessment model, risk of hypertension, machine learning algorithms, primary care

Introduction

Hypertension is becoming increasingly common in primary care. It is accompanied by the occurrence and development of a series of cardiovascular events, disability and even premature death if not detected early and managed well (1). An estimated 245 million adults are diagnosed with hypertension in China (2). An early warning after accurately evaluating the risk of hypertension in primary care patients can alert individuals in the healthy population or subhealthy population with unhealthy lifestyles to take measures to slow or stop the progression of hypertension. Similar practices have been implemented in foreign countries. For instance, management of risk factors for various chronic diseases has been implemented in primary care in Australia (3). Risk assessment models are a cost-effective measure for identifying high-risk individuals with chronic diseases (4, 5). Nevertheless, few existing models can be applied to the health management services provided in primary care. The most intractable problem is that most of these models are targeted at patients in a hospital setting (6); thus, the data input into the models are all extracted from the EHRs of hospitals, which may not be readily available in primary care settings and suitable for general practitioners to implement.

Machine learning (ML) is a nuclear branch of artificial intelligence that has been employed everywhere knowingly or unknowingly, not only in industry and the military but also in medicine and healthcare (7). As a modern data mining, extraction, and analysis technology, ML has the extraordinary ability to automatically train itself and improve its performance without human instruction or elaborate programming (8, 9). With the ability to identify a pattern or make a decision based on the knowledge input, ML algorithms have demonstrated their excellent performance in the area of risk evaluation of diseases. Higher accuracy separates ML algorithms from various other statistical methods. Highly precise risk prediction models for future hypertension were constructed using artificial intelligence techniques in Japan (10). Health check-up data from 18,258 Japanese individuals were utilized to develop a risk prediction model for new-onset hypertension by machine learning techniques. The XGBoost and ensemble models outperformed the logistic regression models [area under the receiver operating characteristic curve (AUC) = 0.859], with AUCs of 0.877 and 0.881, respectively. A study based on several easy-to-collect risk factors to predict the risk of hypertension also revealed that the random forest (AUC = 0.92), CatBoost (AUC = 0.87), and MLP neural network (AUC = 0.78) models performed better than the logistic regression analysis (AUC = 0.77) (11). Although ML is applicable in an extensive range of contexts, the ML algorithm technique alone is insufficient to solve real-world problems (12). Thus, health and medical data in a primary care setting were utilized to facilitate the practical implementation of the risk assessment model for residents in primary care.

The objective of this study is to develop and compare the performances of three ML algorithms on predicting the risk of hypertension for residents over the age of 35 years in primary care in Shanghai, China.

Materials and methods

Data source

The dataset was extracted from the electronic healthcare records of 47 community health centers in the Pudong district of Shanghai. Health records, health examinations and other health-related data of community residents over 35 years old from 2017 to 2019 were collected as the original set of data. A total of 40,261 subjects were enrolled in the study. The dataset included 20 variables containing information regarding demographic characteristics, diagnosis, biochemical indicators and lifestyles. The characteristics of the participants in primary care are shown in Table 1.

Definition of hypertension

Hypertension was defined as (1) systolic blood pressure (SBP) $\geq$140 mmHg and/or diastolic blood pressure (DBP) $\geq$ 90 mmHg, which was measured three times on different days in the clinic without the use of antihypertensive drugs, according to Chinese guidelines for the prevention and treatment of hypertension (2018 revised edition) (13) and/or (2) a diagnosis of hypertension by a physician and/or (3) antihypertension treatment.

TABLE 1 Characteristics of the participants in primary care settings.

Feature	Hypertension ($n = 25,038$)	Normal ($n = 15,223$)	χ^2	P
Age*	72.00 (68.00–78.00)	70.00 (66.00–75.00)	683.51[a]	<0.01
Diabetes status			2077.18[b]	<0.01
No	16,512 (65.95)	13,177 (86.56)		
Yes	8,526 (34.05)	2,046 (13.44)		
Urinary protein level			32.33[b]	<0.01
Negative	8,261 (32.99)	8,392 (55.13)		
Positive	581 (2.32)	405 (2.66)		
BMI*	24.98 (23.01–27.30)	24.16 (22.10–26.30)	458.44[a]	<0.01
EHSA			563.15[b]	<0.01
1	6,973 (27.85)	5,973 (39.24)		
2	12,604 (50.34)	6,387 (41.96)		
3	358 (1.43)	219 (1.44)		
4	277 (1.11)	149 (0.98)		
5	163 (0.65)	46 (0.30)		
Cr level*	69.00 (58.00–84.00)	66.00 (56.00–77.70)	229.09[a]	<0.01
SBP*	140.00 (130.00–153.00)	139.00 (126.00–148.00)	326.93[a]	<0.01
WC*	87.00 (81.00–93.00)	85.00 (79.00–91.00)	157.52[a]	<0.01
Smoking status			200.85[b]	<0.01
1	19,171 (76.57)	10,238 (67.25)		
2	1,159 (4.63)	857 (5.63)		
3	2,028 (8.10)	1,700 (11.17)		
LDL-C level*	2.89 (2.20–3.41)	2.99 (2.46–3.63)	402.35[a]	<0.01
HDL-C level*	1.35 (1.11–1.54)	1.40 (1.20–1.66)	586.65[a]	<0.01
Frequency of drinking			97.64[b]	<0.01
1	18,096 (72.27)	9,837 (64.62)		
2	2,753 (11.00)	1,771 (11.63)		
3	199 (0.79)	151 (0.99)		
4	918 (3.67)	764 (5.02)		
Glucose level*	5.60 (5.13–6.90)	5.50 (5.00–6.33)	247.31[a]	<0.01
Urea nitrogen level*	5.63 (4.80–6.83)	5.63 (4.80–6.37)	306.45[a]	<0.01
TC level*	4.82 (4.01–5.52)	4.99 (4.35–5.72)	267.34[a]	<0.01
DPB*	78.00 (72.00–84.00)	78.00 (70.00–82.00)	235.77[a]	<0.01
Exercise frequency			17.48[b]	<0.01
1	14,751 (58.91)	8,460 (55.57)		
2	815 (3.26)	391 (2.57)		
3	1,495 (5.97)	926 (6.08)		
4	5,471 (21.85)	3,331 (21.88)		
High salt consumption			17.24[b]	<0.01
No	24,938 (99.60)	15,199 (99.80)		
Yes	100 (0.40)	24 (0.20)		
TG level*	1.39 (1.12–1.84)	1.39 (1.00–1.80)	13.22[a]	<0.01
Time spent engaged in exercise*	30.00 (30.00–30.00)	30.00 (30.00–30.00)	0.41[a]	0.52

*Refers to nonnormally distributed measurement data, reported as the median (25th percentile, 75th percentile). [a]refers to results of the rank sum test. [b]refers to the results of the chi-square test.

Inclusion and exclusion criteria

The sample data that fulfilled the following inclusion criteria were obtained for further analysis in this study: community residents over 35 years of age. The chapter "Health Management Service Specifications for Hypertension Patients" in "National Basic Public Health Service Specifications (the Third Edition)" specified that one of the services is to "Provide free blood

pressure measurement once a year for permanent residents aged 35 years old and over in area of responsibility" (14). Therefore, we chose community residents aged 35 years and older as our subjects. The exclusion criteria were: (1) individuals who were unable to provide informed consent, (2) those have any diagnosis of secondary or gestational hypertension, and (3) those who could not cooperate with the investigation because of a long-term outing or a lack of electronic healthcare records.

Data processing

Outliers were handled by interquartile range (IQR). The IQR is evaluated as IQR = Q3–Q1. Q3 is the upper quartile, and Q1 is the lower quartile. Outliers were defined as records that fell below Q1–(1.5*IQR) or above Q3+ (1.5*IQR).

Missing values, such as data with null rows and columns, which did not have a single value or number available, were deleted. Different methods, such as the mean values, median values, mode values, feature combinations and null values, were adopted for dealing with the individual missing values according to the characteristics of different variables. In total, 5.62% of missing values were found in the whole dataset.

Discretization was performed by splitting the range of the continuous variables into intervals to save time needed to build the risk assessment model and improve the assessment results (15).

Feature selection

Feature selection, which is one of the essential parts of building a good prediction model, was employed in this study to improve the prediction accuracy by choosing the most important variables. Moreover, it facilitates a reduction in the resources (time and space) needed to construct the model (16). The embedded method was applied in this study for feature selection. It integrates the feature selection process with the model training process. This method considers variable interactions and is less computationally demanding than the wrapper method (17).

Twenty features were selected to construct the model, from the 127 variables (see the Supplementary Files): age, diabetes status, urinary protein level, BMI, elderly health self-assessment (EHSA), creatinine (Cr) level, systolic blood pressure measured on the upper right arm (SBP), waist circumference (WC), smoking status, low-density lipoprotein cholesterol (LDL-C) level, high-density lipoprotein cholesterol (HDL-C) level, frequency of drinking, glucose level, urea nitrogen level, total cholesterol (TC) level, diastolic blood pressure of the upper right arm (DBP), exercise frequency, time spent engaged in exercise, high salt consumption, and triglyceride (TG) level.

Machine learning algorithms

Extreme Gradient Boosting (XGBoost) is a supervised ML algorithm (18). It is a scalable end-to-end tree boosting system (19). XGBoost can automatically perform parallel computations and is generally more than 10 times faster than GBM (20). Its input types include dense matrix, sparse matrix, data file and xgb.dmatrix. XGBoost accepts sparse input for both tree and linear booster and is optimized for sparse input. It supports customized objective and evaluation functions, and performs better on several different datasets.

Random forest is a supervised classification algorithm (21). It works by learning simple decision rules extracted from the data features and overcomes the limitation of overfitting of the decision trees (22).

Logistic regression is an algorithm that classifies values through the application of a logistic function to coefficients calculated using a linear regression equation (23). It requires that the dependent variable be a second-level score or a second-level evaluation.

Model evaluation and validation

A confusion matrix was employed to evaluate the performance of the models based on ML algorithms for the assessment of hypertension risk. The distinguishing abilities of the risk assessment model were evaluated with the receiver operator characteristic (ROC) curve and the AUC (24). The performance of the models was evaluated by calculating the sensitivity (true positive rate, TPR), specificity (true negative rate, TNR), positive predictive value (PPV), negative predictive value (NPV), accuracy (ACC), and F1-score (25, 26).

Determination of the cut-off point

The evaluations were kinds of probabilities; thus, a cut-off point was needed to classify the prediction probabilities. The probability of having hypertension was represented by "P" in the model. The cut-off point was utilized to classify the evaluated probabilities belonging to the positive results or negative results. We adopted a cut-off point of 0.5 in this study, which meant that participants were evaluated to be at high risk of hypertension when $P \geq 0.5$; otherwise, they were not.

Statistical analysis

Basic descriptive statistics were used to depict the characteristics of the subjects, including demographic characteristics and health-related factors. All normally distributed measurement data are depicted as the mean

± standard deviation (X ± SD), nonnormally distributed measurement data are reported as the median (25th percentile, 75th percentile), and the counting data are expressed as the frequency and proportion. Between groups, normally distributed measurement data were compared by *T*-test, nonnormally distributed measurement data were compared by rank sum test, and the counting data were analyzed by chi-square test. $P < 0.05$ were considered statistically significant. All statistical analyses were performed using IBM SPSS Statistics version 22.0 (IBM Corp., Armonk, NY, USA).

For the assessment models, ML algorithms, XGBoost, random forest and logistic regression were utilized for the evaluation of the risk of hypertension and the effects of the risk factors. Python 3.7.3 was used for the construction of the risk assessment models of hypertension.

Reporting guidelines

Results are presented in accordance with the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines. STROBE and RECORD guidelines for observational studies and studies using routinely collected health data were also considered. The study was conducted in accordance with relevant institutional guidelines.

Results

Characteristics of the study population

A total of 40,261 subjects were included, with a mean age of 72.429 ± 7.643 years, and the mean age of patients with hypertension was 73.216 ± 7.696 years. The sample prevalence of hypertension was almost 62.19%. The differences in age, diabetes status, urinary protein level, BMI, EHSA, Cr level, SBP, WC, smoking status, LDL-C level, HDL-C level, frequency of drinking, glucose level, urea nitrogen level, TC level, DBP, exercise frequency, high salt consumption, and TG level between participants with hypertension and normotensive participants were statistically significant ($P < 0.01$). There were no statistically significant differences ($P > 0.05$) in terms of time spent engaged in exercise. The characteristics of the study participants are summarized in Table 1.

Construction of the risk assessment models

The training set and validation set were utilized to determine the optimal parameters for XGBoost, random forest and logistic regression. The parameters of each model under optimal performance are exhibited in Table 2. For other unlisted parameters in the three ML algorithms, default values were set.

TABLE 2 Configuration of parameters in each ML algorithm.

ML algorithm	Parameter	Value range	Optimal value
XGBoost	learning_rate	[0, 0.3]	0.05
	n_estimators	[100, 500]	200
	gamma	[0, 20]	5
	subsample	[0, 0.9]	0.4
	colsample_bytree	[0.5, 0.9]	0.9
	min_child_weight	(1, 6)	5
	max_depth	(2, 8)	6
	objective	-	binary:logistic
Random forest	n _estimators	[1, 50]	40
	criterion	gini	gini
	max_depth	none	none
	min_samples_split	[5, 200]	200
	min_samples_leaf	[1, 50]	1
	max_features	auto	auto
Logistic regression	C	[0, 200]	100
	class_weight	none	none
	max_iter	[10, 100]	10
	solver	-	liblinear

Feature importance

The significant features of the XGBoost model, random forest model and logistic regression model are listed in Figures 1–3, respectively. The urea nitrogen level was the highest ranked feature for predicting hypertension in both the XGBoost model and the random forest model. BMI, SBP, TG level, Cr level, LDL-C level, and glucose level were ranked in the top 10 in all three models.

Model performance

We utilized various methods and evaluation metrics to assess the performances of the XGBoost, random forest, and logistic regression models in the training, validation, and testing sets. Overall, the XGBoost model outperformed the other two models in TPR (0.864), TNR (0.488), PPV (0.735), NPV (0.686), ACC (0.722), F1-score (0.795), and AUC (0.765) in the testing set (Table 3).

Figure 4 summarizes the ROC curve areas obtained from the XGBoost model, random forest model and logistic regression model in the testing set. The areas under the ROC curves were different among the three models. The AUCs for the test set

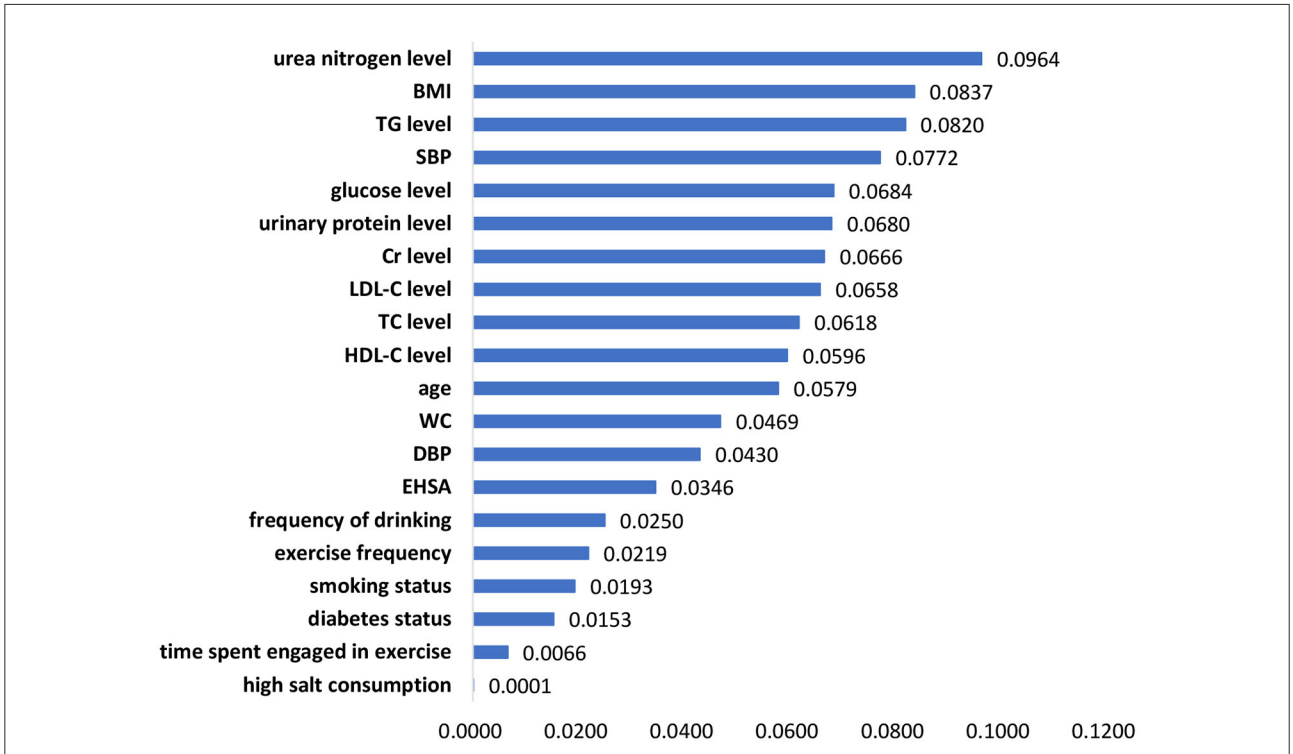

FIGURE 1
Feature importance in the XGBoost model.

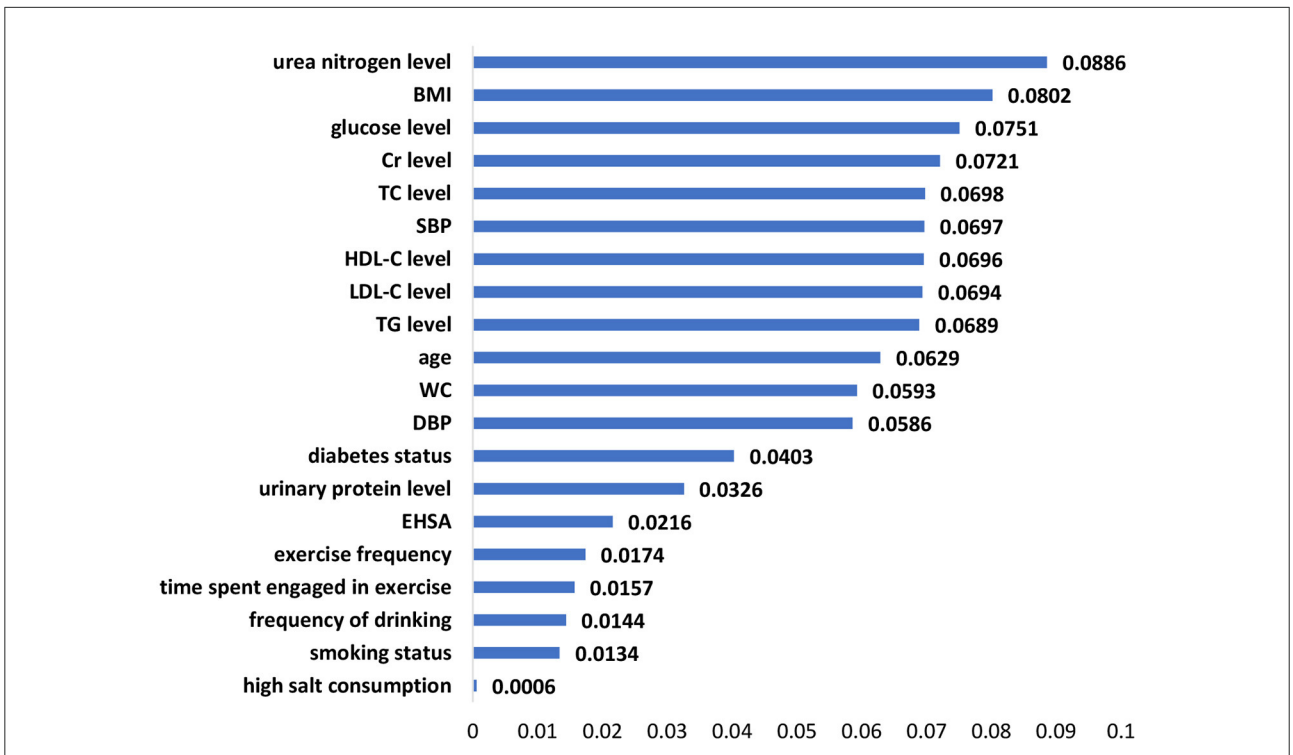

FIGURE 2
Feature importance in the random forest model.

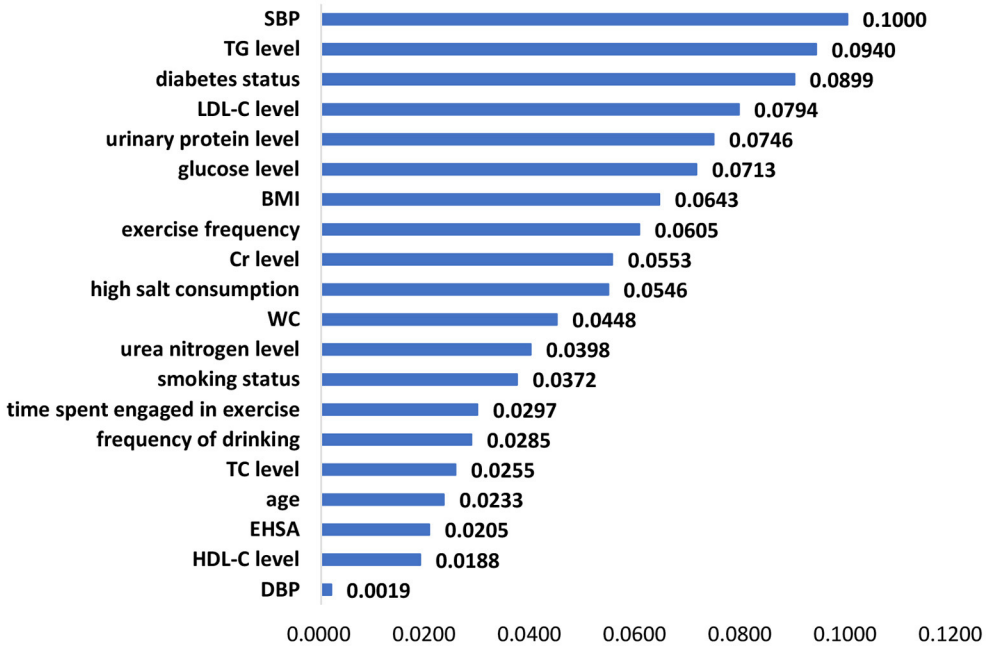

FIGURE 3
Feature importance in the logistic regression model.

TABLE 3 The fitting results for the XGBoost, random forest, and logistic regression models for the training, validation, and testing sets.

ML algorithm	Dataset	TPR	TNR	PPV	NPV	ACC	F1-Score	AUC
XGBoost	Training	0.886	0.530	0.756	0.739	0.752	0.816	0.818
	Validation	0.862	0.480	0.732	0.678	0.717	0.791	0.753
	Testing	0.864	0.488	0.735	0.686	0.722	0.795	0.765
Random forest	Training	0.896	0.434	0.723	0.718	0.722	0.800	0.782
	Validation	0.871	0.446	0.721	0.678	0.711	0.789	0.745
	Testing	0.816	0.548	0.748	0.644	0.714	0.780	0.756
Logistic regression	Training	0.827	0.411	0.698	0.591	0.670	0.757	0.705
	Validation	0.822	0.418	0.699	0.588	0.669	0.756	0.692
	Testing	0.829	0.430	0.705	0.604	0.678	0.762	0.707

were 0.765 for XGBoost, 0.756 for random forest, and 0.707 for logistic regression (Table 4). The AUC of the XGBoost model was higher than that of the random forest and logistic regression models. Our results demonstrated that the XGBoost model had better predictive performance than the random forest and logistic regression models.

Discussion

Among the 20 selected features in this study, BMI, SBP, TG level, Cr level, LDL-C level, and glucose level had a strong effect on hypertension prediction and were included among the top 10 in the ranking of the feature importance for all three models. Similar to the results of previous studies, features such as age (27–29), BMI (28, 30), diabetes status (28), Cr level (26), blood pressure (29), WC (31), smoking status (28), LDL-C level (26, 28), HDL-C level (26), drinking (28), glucose level (32), TC level (26, 27), exercise (33), salt intake (34), and TG level (27) were identified as predictors of hypertension in the risk assessment model of hypertension.

However, to the best of our knowledge, urinary protein level, urea nitrogen level, and EHSA entered the models as new components that have not been included in risk evaluation models of hypertension in previous studies.

A study collected data from three exams in the Strong Heart Study, explored the risk factors for hypertension by means of generalized linear models and demonstrated

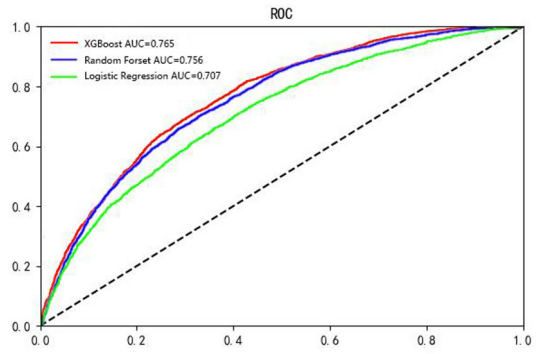

FIGURE 4
The ROC curves obtained from the XGBoost model, random forest model and logistic regression model. X axis: 1-specificity, Y axis: sensitivity. The reference line is shown as a dashed line (the black line).

TABLE 4 AUCs for the XGBoost, random forest, and logistic regression models for the training, validation, and testing sets.

ML algorithm	Dataset	AUC
XGBoost	Training	0.818
	Validation	0.753
	Testing	0.765
Random forest	Training	0.782
	Validation	0.745
	Testing	0.756
Logistic regression	Training	0.705
	Validation	0.692
	Testing	0.707

that systolic blood pressure was significantly and positively associated with albuminuria, age, and obesity and negatively associated with smoking. Moreover, participants with more severe albuminuria status or older age developed higher SBP, while DBP was not significantly affected by the albuminuria status (35). This study in American Indians revealed that having macro/microalbuminuria is a significant risk factor for hypertension, which can explain why urinary protein level was selected as one of the features in our model to some extent. Urinary protein level may also affect the development of hypertension in Chinese individuals or facilitate the risk assessment of hypertension in Chinese individuals. Furthermore, Kim et al. reported that subjects with high normal BP had an independently significant association with microalbuminuria by performing a multiple logistic regression analysis, with an odds ratio of 1.692 and a 95% confidence interval of 1.097 to 2.611 (36). These results from a Korean population indicated that compared to individuals with normal BP, those with high normal BP have more risk factors for hypertension and cardiovascular diseases, for instance, albuminuria. Since the incidence of urinary protein was significantly higher in the prehypertensive population than in the normal population, urinary protein level should receive attention in future predictive studies and intervention measures.

Although we rarely found urea nitrogen level to be included as a predictive factor in the risk prediction models, it was found to be a significant risk factors for hypertension. A case-control study conducted among university staff found that staff with high serum urea levels had a higher risk of hypertension than those with normal urea levels (OR = 1.452), which implies that the level of urea is also very important as one of the risk factors for hypertension (37). Not coincidentally, this phenomenon has been found among middle-aged and elderly people. SBP was positively correlated with the blood urea nitrogen concentration ($r = 0.16424$, $P = 0.0105$) and the blood uric acid concentration ($r = 0.16023$, $P = 0.0126$) among middle-aged and older-aged populations in Guangzhou, China, as well as DBP (blood urea nitrogen concentration: $r = 0.13506$, $P = 0.0358$; blood uric acid concentration: $r = 0.16562$, $P = 0.0099$) (38). The results of stepwise regression analysis also indicated that there was still a significant positive correlation between SBP, DBP and concentrations of blood urea nitrogen and blood uric acid. The role of urea nitrogen level, one of the features entered into our risk assessment model, in the occurrence and development of hypertension still needs to be further investigated.

EHSA was also one of the predictors entered into our model. Kaplan and Camacho have already reported that the association between level of perceived health and mortality persisted in multiple logistic analyses controlling for age, sex, physical health status, health practices, social network participation, income, education, health relative to peers of the same age, anomy, morale, depression, and happiness (39). The results reminded us that self-assessment of health might serve as a comprehensive reflection of unmeasurable factors and as an indication of some underlying diseases or an early stage of the diseases. Evidence has shown that psychosocial factors exert strong effects on health status measures (40). Zhang et al. revealed that the proportion of elderly individuals with poor or normal health self-assessments who were suffering from common chronic diseases was significantly increased (41). The health self-assessment epitomizes the health concept and self-perception of health status of elderly individuals to some extent, which might have an underlying predictive value on the prediction of the risk of hypertension and should thus be given more attention in future research, as well as the practice in primary care.

Unlike traditional risk assessment methods, our study employed ML algorithms for model construction. XGBoost exhibited the best performance compared to random forest and logistic regression. Logistic regression assumes that every variable should be independent, and the model possesses only a linear partition surface. However, the associations between

exposure factors and diseases are often affected by various confounding factors, which leads to the large deviation and low accuracy when fitting the model through logistic inference. In contrast, XGBoost and random forest are nonparametric algorithms (42) that do not assume that a functional relationship between the features and outcomes exists, as required by logistic regression models. A greedy algorithm is executed to determine the optimal splits in the data that reduce the entropy of the outcome to the utmost extent during every split. As a result, once a feature is selected, the significance of any highly related feature will decrease greatly due to the completion of the effective split done by the original feature previously. Consequently, the entropy of the outcome will no longer be reduced effectively by related features. Therefore, XGBoost and random forest are robust to related features. The reason why XGBoost outperforms the other methods may be that it introduces the regularized loss function (43) and combines gradient lifting algorithms and decision trees, which preserves the correlation between features during the modeling process (44).

In terms of performance, the XGBoost-based hypertension prediction model proposed by the Japanese group showed an AUC of 0.877 (10), while the hypertension risk assessment model proposed in this study exhibited an AUC of 0.765. The explanation for this discrepancy may be the difference in ethnic populations. According to previous studies, different ethnic populations have different characteristics of hypertension, which may affect the discrepancies in the AUCs for different models (45, 46). Meanwhile, the difference between age range of the subjects may also contribute to the discrepancy in the model performance. For instance, in a study regarding assessing the relationship between nerves and cancer using machine learning methods, the authors found that the performance of the model trained on the young dataset was much better than that trained on the elderly dataset and the whole age dataset, and the performance of the model trained on the whole age dataset was slightly better or similar to that trained on the elderly dataset (47). The findings from these studies suggested that we should further investigate the effect of the difference in subjects' age range on the performance of hypertension models in the future. Compared with other models used to predict hypertension (11), the results from the proposed XGBoost prediction model in the present study did not show a higher AUC. The variable selection may partially explain the discrepancy.

After the risk assessment of hypertension, subsequent interventions and management to prevent or postpone the occurrence and development of hypertension are crucially important in high-risk populations. Continuous monitoring and management are imperative for high-risk patients. On the one hand, realtimeness and continuity monitoring can detect any problem without delay. On the other hand, early signs of detected symptoms can alert both general practitioners (GPs) and individuals in a timely manner. For high-risk populations, corresponding individual intervention strategies targeting the main risk factors should be prescribed by GPs in primary care. For instance, lifestyle factors such as exercise, eating habits, and drinking habits can be improved under the guidance of GPs after risk assessment. Evidence has revealed that a high concentration of parks or playgrounds in residential areas may reduce the risk of hypertension, which is mainly attributable to the cultivation and formation of exercise habits and implies the importance of interventions in communities (48).

However, there were several limitations in our study. One of the limitations of the study was that it had a cross-sectional design, and the results could not indicate causality in this situation. A prospective cohort study is needed to further identify the cause-and-effect relationships. Second, the risk assessment model was designed considering only variables available in the setting of primary care, and variables regarding mental health and hereditary factors were not included. Third, we measured several variables, such as age, urinary protein level, BMI, and Cr level, on only a single occasion and did not take changes in these variables into consideration.

In conclusion, XGBoost outperformed random forest and logistic regression models in predicting the risk of hypertension in primary care settings. Early identification and the corresponding preventive strategies in primary care remain insufficient in China. Integration of such a risk assessment model into primary care may help general practitioners target populations at high-risk for hypertension, tailor the corresponding preventive measures and treatment strategies to those at high risk, improve the awareness of residents regarding health risks and their adherence toward targeted intervention, and eventually facilitate individuals' health and quality of life while decreasing healthcare costs.

Data availability statement

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

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committees of Tongji University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

NC was involved in designing the study, analyzing the results, and wrote the manuscript. FF performed data collection, proofread the manuscript, and modified the format. DY, ZW, and JS supervised the work and were involved in the study

design. JG and YY helped with data interpretation and graphing. YG, HJ, and QC revised the manuscript. All authors reviewed and approved the final version of the manuscript.

Funding

This study was supported by grants from Soft Science Project of the Shanghai Science and Technology Commission (22692107200), Shanghai Education Science Research Project (C2021039), the Natural Science Foundation of China (71774116 and 71603182), Shanghai Public Health Outstanding Young Personnel Training Program (GWV-10.2-XD07), National Key Research and Development Program of China (2018YFC2000700, SQ2022YFC3600172), and Shanghai Pujiang Program (2020PJC080). The funding agencies had no role in the design of this study nor any role during its execution, analyses, data interpretation, or decision to submit results.

Acknowledgments

We are grateful to all the participants involved in this study.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2022.984621/full#supplementary-material

References

1. Seo MJ, Ahn SG, Lee YJ, Kim JK. Development of the hypertension index model in general adult using the Korea national health and nutritional examination survey and the Korean genome and epidemiology study. *J Pers Med.* (2021) 11:968. doi: 10.3390/jpm11100968

2. Wang ZW, Chen Z, Zhang LF, Wang X, Hao G, Zhang ZG, et al. Status of hypertension in China: results from the China hypertension survey, 2012–2015. *Circulation.* (2018) 137:2344–56. doi: 10.1161/CIRCULATIONAHA.117.032380

3. The Royal Australian College of General Practitioners. *Guidelines for Preventive Activities in General Practice.* 9th edition. Available online at: https://www.racgp.org.au/getattachment/1ad1a26f-9c8b-4e3c-b45b-3237272b3a04/Guidelines-for-preventive-activities-in-general-practice.aspx (accessed December 9, 2021).

4. Chen X, Wu Z, Chen Y, Wang X, Zhu J, Wang N, et al. Risk score model of type 2 diabetes prediction for rural Chinese adults: the Rural Deqing Cohort Study. *J Endocrinol Invest.* (2017) 40:1115–23. doi: 10.1007/s40618-017-0680-4

5. Hart GR, Roffman DA, Decker R, Deng J. A multi-parameterized artificial neural network for lung cancer risk prediction. *PLoS ONE.* (2018) 13:e0205264. doi: 10.1371/journal.pone.0205264

6. Andriani P, Chamidah N. Modelling of hypertension risk factors using logistic regression to prevent hypertension in Indonesia. *J Phys Conf Ser.* (2019) 1306:012027. doi: 10.1088/1742-6596/1306/1/012027

7. Dash SS, Nayak SK, Mishra D. A review on machine learning algorithms. *Intelligent and Cloud Computing.* (2021) 2:495–507. doi: 10.1007/978-981-15-6202-0_51

8. Alpaydin E. *Introduction to Machine Learning.* Cambridge: MIT press (2014).

9. Marsland S. *Machine Learning: An Algorithmic Perspective.* Florida: CRC press (2015).

10. Kanegae H, Suzuki K, Fukatani K, Ito T, Harada N, Kario K. Highly precise risk prediction model for new-onset hypertension using artificial intelligence techniques. *J Clin Hypertens.* (2020) 22:445–50. doi: 10.1111/jch.13759

11. Zhao HH, Zhang XY, Xu Y, Gao LS, Ma ZC, Sun YN, et al. Predicting the risk of hypertension based on several easy-to-collect risk factors: a machine learning method. *Front Public Health.* (2021) 9:619429. doi: 10.3389/fpubh.2021.619429

12. Benton WC. Machine learning systems and intelligent applications. *IEEE Software.* (2020) 37:43–9. doi: 10.1109/MS.2020.2985224

13. Writing Group of 2018 Chinese Guidelines for the Management of Hypertension. 2018 Chinese guidelines for the management of hypertension. *Chin J Cardiovasc Med.* (2019) 24:24–56. Available online at: https://kns.cnki.net/KXReader/Detail?invoice=mmxQHXTLYBslWiQ8dVj4qX86LVtlqnzSOMxWui9DdJnArA1bL1jbq27wZdQFz4vFFY2YwM2H2r9McXmfq0V42SlfIfqvqpfNOfDEaxHTJqBQihD1thTZxR0mcafyuP%2Bp8hksOj%2FLDyIMXyOOm7bb9G6Xl9eNJ5Bt6%2Fh9DfJ9CKI%3D&DBCODE=CJFD&FileName=XIXG201901003&TABLEName=cjfdlast2019&nonce=E48DCA6E53304D7089F6BE689E4F1861&uid=&TIMESTAMP=1663581215503

14. National Health and Family Planning Commission of the People's Republic of China. National Basic Public Health Service Specifications (the Third Edition). Available online at: http://www.nhc.gov.cn/ewebeditor/uploadfile/2017/04/20170417104506514.pdf (accessed August 15, 2022).

15. Kurgan L, Cios KJ. Discretization algorithm that uses class-attribute interdependence maximization. In: *IC-AI'2001: Proceedings of the International Conference on Artificial Intelligence, VOLS I-III.* (2001). p. 980–6. Available online at: https://www.webofscience.com/wos/alldb/full-record/WOS:000173960400153

16. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res.* (2003) 3:1157–82. doi: 10.1162/153244303322753616

17. Shi X, Nikolic G, Epelde G, Arrue M, Van-Dierdonck JB, Bilbao R, et al. An ensemble-based feature selection framework to select risk factors of childhood obesity for policy decision making. *BMC Med Inform Decis Mak.* (2021) 21:222. doi: 10.1186/s12911-021-01580-0

18. Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics.* (2012) 28:2747–54. doi: 10.1093/bioinformatics/bts526

19. Chen TQ, Guestrin C. XGBoost: A scalable tree boosting system. In: *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).* San Francisco, CA: ACM (2016). p. 785–94.

20. Chen TQ, He T. *xgboost: eXtreme Gradient Boosting.* Available online at: https://cran.microsoft.com/snapshot/2017-12-11/web/packages/xgboost/vignettes/xgboost.pdf (accessed August 25, 2022).

21. Prasad AM, Iverson LR, Liaw A. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems.* (2006) 9:181–99. doi: 10.1007/s10021-005-0054-1

22. Sakr S, Elshawi R, Ahmed A, Qureshi WT, Brawner C, Keteyian S, et al. Using machine learning on cardiorespiratory fitness data for predicting hypertension: the Henry Ford ExercIse Testing (FIT) Project. *PLoS ONE.* (2018) 13:e0195344. doi: 10.1371/journal.pone.0195344

23. Buya S, Tongkumchum P, Owusu BE. Modelling of land-use change in Thailand using binary logistic regression and multinomial logistic regression. *Arab J Geosci.* (2020) 13:12. doi: 10.1007/s12517-020-05451-2

24. Moons KGM, Harrell FE, Steyerberg EW. Should scoring rules be based on odds ratios or regression coefficients? *J Clin Epidemiol.* (2002) 55:1054–5. doi: 10.1016/S0895-4356(02)00453-5

25. Cai QC, Yu ED, Xiao Y, Bai WY, Chen X, He LP, et al. Derivation and validation of a prediction rule for estimating advanced colorectal neoplasm risk in average-risk Chinese. *Am J Epidemiol.* (2012) 175:584–93. doi: 10.1093/aje/kwr337

26. Lavrac N. Selected techniques for data mining in medicine. *Artif Intell Med.* (1999) 16:3–23. doi: 10.1016/S0933-3657(98)00062-1

27. Ren ZG, Rao BC, Xie SQ Li A, Wang LJ, Cui GY, et al. A novel predicted model for hypertension based on a large cross-sectional study. *Sci Rep.* (2020) 10:10615. doi: 10.1038/s41598-020-64980-8

28. Kshirsagar AV, Chiu YL, Bomback AS, August PA, Viera AJ, Colindres RE, et al. A hypertension risk score for middle-aged and older adults. *J Clin Hypertens.* (2010) 12:800–8. doi: 10.1111/j.1751-7176.2010.00343.x

29. Kanegae H, Oikawa T, Suzuki K, Okawara Y, Kario K. Developing and validating a new precise risk-prediction model for new-onset hypertension: The Jichi Genki hypertension prediction model (JG model). *J Clin Hypertens.* (2018) 20:880–90. doi: 10.1111/jch.13270

30. Akdag R, Fenkci S, Degirmencioglu S, Rota S, Sermez Y, Camdeviren H, et al. Determination of risk factors for hypertension through the classification tree method. *Adv Ther.* (2006) 23:885–92. doi: 10.1007/BF02850210

31. Xu F, Zhu JC, Sun N, Wang L, Xie C, Tang QX, et al. Development and validation of prediction models for hypertension risks in rural Chinese populations. *J Glob Health.* (2019) 9:020601. doi: 10.7189/jogh.09.020601

32. Chien KL, Hsu HC, Su TC, Chang WT, Sung FC, Chen MF, et al. Prediction models for the risk of new-onset hypertension in ethnic Chinese in Taiwan. *J Hum Hypertens.* (2011) 25:294–303. doi: 10.1038/jhh.2010.63

33. Niiranen TJ, Havulinna AS, Langen VL, Salomaa V, Jula AM. Prediction of blood pressure and blood pressure change with a genetic risk score. *J Clin Hypertens.* (2016) 18:181–6. doi: 10.1111/jch.12702

34. Xu YZ, Liu JB, Wang JW, Fan QL, Luo YY, Zhan HF, et al. Establishment and verification of a nomogram prediction model of hypertension risk in Xinjiang Kazakhs. *Medicine.* (2021) 100:e27600. doi: 10.1097/MD.0000000000027600

35. Wang WY, Lee ET, Fabsitz RR, Devereux R, Best L, Welty TK, et al. A longitudinal study of hypertension risk factors and their relation to cardiovascular disease: the Strong Heart Study. *Hypertension.* (2006) 47:403–9. doi: 10.1161/01.HYP.0000200710.29498.80

36. Kim BJ, Lee HJ, Sung KC, Kim BS, Kang JH, Lee MH, et al. Comparison of microalbuminuria in 2 blood pressure categories of prehypertensive subjects. *Cir J.* (2007) 71:1283–7. doi: 10.1253/circj.71.1283

37. Guan XP, Xiang H, Xia H. Risk factors of essential hypertension among university staff: a case-conrol study. *Chin J Public Health.* (2011) 27:501–3. Available online at: https://kns.cnki.net/KXReader/Detail?invoice=P3sGsUoJPKzESgeaIO90OqypazL6%2FaV6aKlOyyEtHUjyNMEkFt0r7As7IDJ1%2Fb4Ul2gWu5h3GAHTUpsNhIhDJ3AXK6qQ0nvjThJG5jy69hiTpMTSgcb2WpDM2vmmF3%2BtJugjJFak%2Bw%2FroyHgcAdV3yCNnSvHcZaL%2FaIBDwlX22Y%3D&DBCODE=CJFD&FileName=ZGGW201104058&TABLEName=cjfd2011&nonce=C6563EFF8D7B436BA206C2A8A9DFCD4D&uid=&TIMESTAMP=1663583554744

38. Xiao M, Li H, Shi ML, Deng ML, Mai JZ, Liu XQ, et al. Relationship between blood pressure and blood uric acid, urea nitrogen in middle and older-aged population in Guangzhou. *South China Journal of Cardiovascular Diseases.* (2009) 15:457–60. doi: 10.3969/j.issn.1007-9688.2009.06.012

39. Kaplan GA, Camacho T. Perceived health and mortality: a nine-year follow-up of the human population laboratory cohort. *Am J Epidemiol.* (1983) 117:292–304. doi: 10.1093/oxfordjournals.aje.a113541

40. Ring D, Kadzielski J, Fabian L, Zurakowski D, Malhotra LR, Jupiter JB, et al. Self-reported upper extremity health status correlates with depression. *J Bone Joint Surg Am.* (2006) 88:1983–8. doi: 10.2106/00004623-200609000-00012

41. Zhang FM, Xu HJ. Research on the relationship between self-assessment of health and chronic diseases in elderly population. *Chin J Gerontol.* (2008) 28:2353–5. Available online at: https://kns.cnki.net/KXReader/Detail?invoice=N9Z0FR7NKGLgehblrtqdPeDcE5qOG2iMpGYxCX7YPD1NsFOumGhc%2F67eXgkGmlawtzKNNUAzTgsK3GZ5bmzbCzuuBrFmsX5p4GqFo4LhgB2DVhvH4GePaUe6Q3iohh0nBxD2Ai3Zgz1iYgnJeKoWgIgwO6SDOhQhQQGcGTCaggM%3D&DBCODE=CJFD&FileName=ZLXZ200823029&TABLEName=cjfd2008&nonce=4B936EF422F744949B8210E480541DE9&uid=&TIMESTAMP=1663584334018

42. Chen TQ, He T, Benesty M, Tang Y. Understand your dataset with XGBoost. Available online at: https://cran.r-project.org/web/packages/xgboost/vignettes/discoverYourData.html (accessed December 15, 2021).

43. Pan BY. Application of XGBoost algorithm in hourly PM2.5 concentration prediction. *3rd international conference on advances in energy resources and environment engineering book series: IOP conference series-earth and environmental science.* (2018) 113:012127. doi: 10.1088/1755-1315/113/1/012127

44. Thomas J, Hepp T, Mayr A, Bischl B. Probing for Sparse and Fast Variable Selection with Model-Based Boosting. *Comput Math Methods Med.* (2017) 2017:1421409. doi: 10.1155/2017/1421409

45. Brown MJ. Hypertension and ethnic group. *BMJ Bri Med J.* (2006) 332:8336B. doi: 10.1136/bmj.332.7545.833

46. Kramer H, Han C, Post W, Goff D, Diez-Roux A, Cooper R, et al. Racial/ethnic differences in hypertension and hypertension treatment and control in the multi-ethnic study of atherosclerosis (MESA). *Am J Hypert.* (2004) 17:963–970. doi: 10.1016/j.amjhyper.2004.06.001

47. Wang FL. *Function Research and Biomarker Identification of Nervous System in Cancer* [dissertation / master's thesis]. [Changchun (Jilin)]: Jilin University (2022).

48. Ye CY, Fu TY, Hao SY, Zhang Y, Wang O, Jin B, et al. Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning. *J Med Internet Res.* (2018) 20:e22. doi: 10.2196/jmir.9268

frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 31 October 2022
DOI 10.3389/fpubh.2022.1003002

OPEN ACCESS

EDITED BY
Ming-Chin Lin,
Taipei Medical University, Taiwan

REVIEWED BY
Rakeeb Ahmad Mir,
Central University of Kashmir, India
Sylvie Amu,
University College Cork, Ireland

*CORRESPONDENCE
Xudong Tang
txdly@sina.com

SPECIALTY SECTION
This article was submitted to Family Medicine and Primary Care, a section of the journal Frontiers in Public Health

RECEIVED 25 July 2022
ACCEPTED 17 October 2022
PUBLISHED 31 October 2022

CITATION
Huang J, Zhang J, Wang F, Zhang B and Tang X (2022) Revealing immune infiltrate characteristics and potential diagnostic value of immune-related genes in ulcerative colitis: An integrative genomic analysis. *Front. Public Health* 10:1003002. doi: 10.3389/fpubh.2022.1003002

Revealing immune infiltrate characteristics and potential diagnostic value of immune-related genes in ulcerative colitis: An integrative genomic analysis

Jinke Huang[1], Jiaqi Zhang[1,2], Fengyun Wang[1,2], Beihua Zhang[1,2] and Xudong Tang[1,2]*

[1]Department of Gastroenterology, Xiyuan Hospital of China Academy of Chinese Medical Sciences, Beijing, China, [2]Institute of Digestive Diseases, Xiyuan Hospital of China Academy of Chinese Medical Sciences, Beijing, China

Objectives: Ulcerative colitis (UC) is an autoimmune disease of the colon. The aim of this study was to explore the characteristics of immune infiltrates in UC patients and identify immune-related diagnostic biomarkers for UC.

Methods: Three gene expression profiles were acquired from the GEO database, followed by identification of differentially expressed genes (DEGs) by Linear Modeling of Microarray Data. Enrichment analysis of Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Disease Ontology (DO) were performed to analyze the biological functions of DEGs. Subsequently, the single sample gene set enrichment analysis (ssGSEA) was performed to identify immune infiltration characteristics of UC. Correlations between diagnostic genes and immune infiltration were explored to identify markers with the greatest diagnostic potential, and a UC diagnostic model was subsequently constructed. Finally, the prediction performance of the model was quantified by nomogram, non-correlated nomogram, and ROC curve.

Results: A total of 3111 DEGs (1,608 up-regulated and 1,503 down-regulated genes) were identified. DEGs were significantly involved in the immune system and UC-related pathways. Immune infiltration profiles of colonic tissue were significantly different between healthy individuals and UC patients. High proportions of resting of aDCs, B cells, $CD8^+$ T cells, DCs, iDCs, Macrophages, Neutrophils, pDCs, T helper cells, Tfh, Th1 cells, Th2 cells, TIL and Treg were found in UC samples. A 5-gene based diagnostic prediction model was constructed and the results of nomogram, non-correlated nomogram and ROC curve suggested the powerful diagnostic value of the model.

Conclusions: This study identified the immune infiltrate characteristics and 5 immune-related genes for UC. The model based on the immune-related genes facilitates the early diagnosis of UC and provides a basis for the evaluation of the prognosis of UC.

KEYWORDS

ulcerative colitis, immune infiltration, genes, diagnostic value, genomic analysis

Introduction

Ulcerative colitis (UC) is a complex disease characterized by chronic inflammation of the colon (1). Worldwide, UC is estimated to affect 9–100,000 people annually, and the incidence is increasing year by year (2). The growing number of UC patients places a heavy economic burden on society, with direct and indirect costs ranging from $8.1–14.9 billion per year in the United States and Ă12.5–29.1 billion in Europe (3). The treatment goal in UC is the induction and maintenance of remission. Although the therapeutic armamentarium is expanding, the treatment of UC is highly challenging because of its incompletely understood pathogenesis (4). Therefore, an in-depth understanding of disease pathogenesis and identification of biomarkers of disease progression at the molecular level may provide new ideas for the early diagnosis of UC.

The etiology and pathogenesis of UC are not fully understood, and it is mainly thought to be caused by an enhanced immune response to the gut microbiota in genetically susceptible individuals (5). Many studies have investigated the function of various immune cells, but it has been challenging to predict the role of all immune subsets in UC in an integrated manner. Initial activation of innate immunity causes a non-specific response, and then, sustained stimulation of inflammation will activate adaptive immunity, which may lead to persistent chronic inflammation (6). Accumulating evidence suggests that both innate and adaptive immune abnormalities are responsible for the abnormal inflammatory response in the gut (7). Inflammation associated with inflammatory bowel disease (IBD) has been reported to be closely associated with aberrant immune response elicited by CD4 T cells and dendritic cells (8–10). IRF5 contributes to the regulation of T cell signaling and modulates cytokine secretion to promote inflammation in UC (11). Neutrophil HGF-MET signaling can also contribute to the progression of UC (12). Furthermore, infiltrating immune cells are present in the intestinal mucosa of individuals with UC (13), and increased immune cell infiltration may correlate with the severity and recurrence of UC (14, 15). All these findings suggest a key role of immune cells in the pathogenesis of UC, and molecules associated with these cells may serve as new biomarkers for UC.

Gene chip is a genetic detection technology that can detect all expression information of all genes from a sample and reveal numerous genes activated in different tissues and their physiological and pathological states (16, 17). Currently, microarray technology integrated with bioinformatics analysis

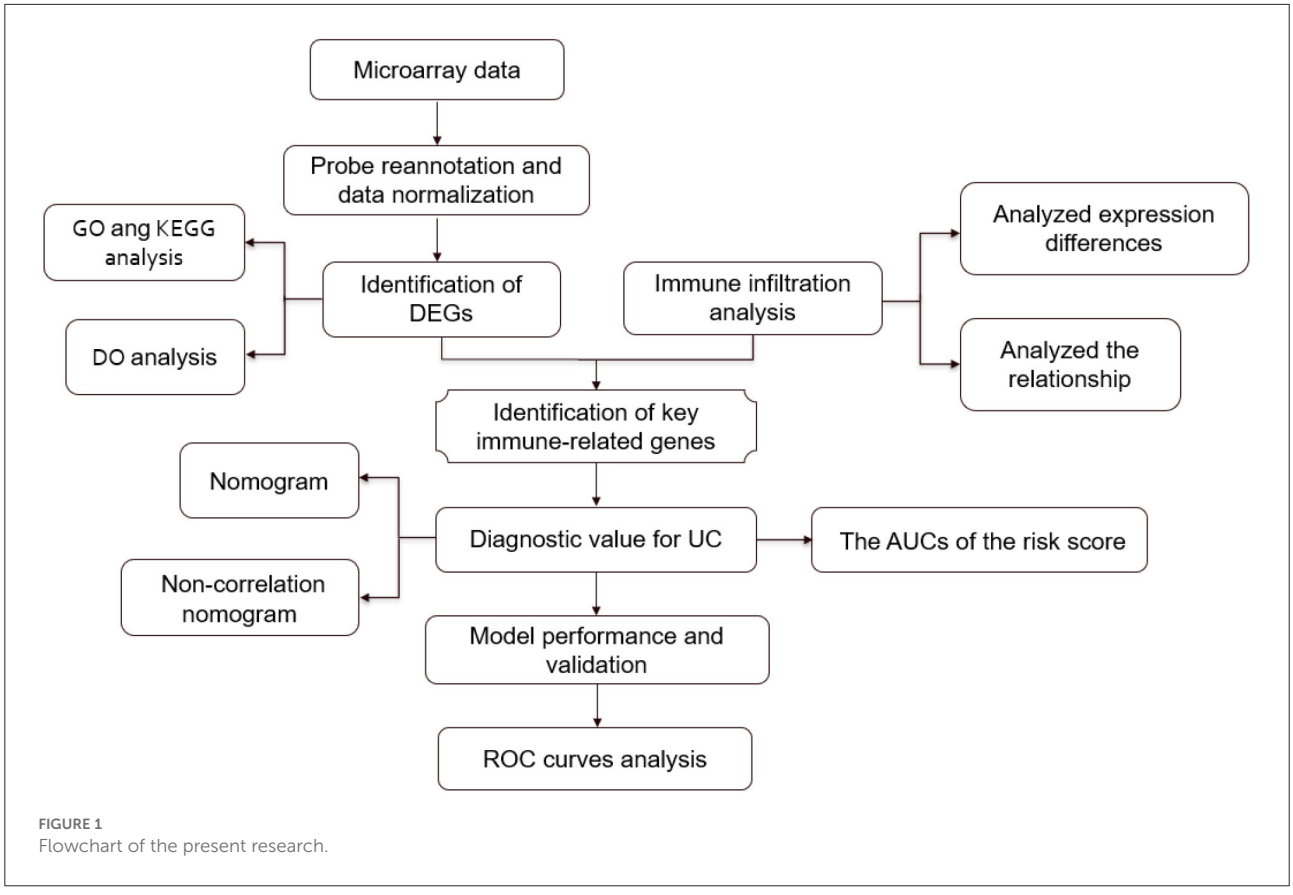

FIGURE 1
Flowchart of the present research.

TABLE 1 Details of gene expression profiles.

Dataset	Platform	Tissue	Normal	UC	Reference (PMID)	Hyperlinks
GSE87473	GPL13158	Colon	106	21	29401083	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87473
GSE75214	GPL6244	Colon	97	11	28885228	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75214
GSE92415	GPL13158	Colon	87	21	23735746	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92415

TABLE 2 R packages for bioinformatics analysis.

R package	Version	Description
Limma	3.50.3	Probe reannotation and data normalization, identification of DEGs
Pheatmap	1.0.12	Plotting heat map
Ggplot2	3.3.6	Plotting volcano plot, visualization of enrichment analysis results
Clusterprofiler	4.2.2	GO and KEGG enrichment analyses
DOSE	3.20.1	Disease ontology enrichment analysis
Corrplot	0.92	Correlation matrix visualization
Ggpubr	0.4.0	Plotting boxplot
Psych	2.2.5	Correlation analysis of DEGs and immune infiltration
RMS	6.3-0	Construction of diagnostic model
ROCR	1.0-11	ROC analysis

has been widely used to explore pathological features and identify potential novel biomarkers for various diseases (18, 19). Based on large-scale microarray gene expression data, this study applied integrated bioinformatics analysis to explore the molecular mechanisms of UC. Moreover, we focused on identifying core genes associated with immune infiltrating cells and used these core biomarkers to construct a risk prediction model for UC with the aim of providing new ideas for early diagnosis of UC. The flow chart of the present research is shown in Figure 1.

Materials and methods

Microarray data acquisition

Gene expression profiles were acquired from GEO database (www.ncbi.nlm.nih.gov/geo/) (20) with the following criteria: (a) patients were diagnosed as UC; (b) data on colonic tissue from healthy controls and UC patients from the same GEO platform; (c) datasets inclusion with at least 10 UC and healthy tissue samples; (d) GEO platforms containing >5,000 genes. Finally, three gene expression profiles (GSE87473, GSE92415, and GSE 75214) were included. Table 1 shows the details of the gene expression profiles.

Identification of differentially expressed genes

After downloading the microarray expression matrices, probe names were converted to gene symbols using R software (version 4.1.2). DEGs between UC colon tissue and healthy colon tissue were screened by the "limma" package in R software, and the threshold for DEG was set to $|\log_2$ Fold change (FC)$| > 0.5$ and p-value < 0.05.

Functional analysis of DEGs

Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Disease Ontology (DO) enrichment analyses were performed on the screened DEGs from GSE87473 by the "clusterProfiler," and "DOSE" packages in R software. The threshold for enrichment analysis was set to p-value < 0.05. The "ggplot2" package was applied to visualize the enrichment results.

Immune infiltration analysis

Single sample gene set enrichment analysis (ssGSEA) method was applied to quantify the enrichment scores of immune cells (activated DCs (aDCs), B cells, CD8+ T cells, DCs, interdigitating DCs (iDCs), macrophages, mast cells, neutrophils, natural killer (NK) cells, plasmacytoid DCs (pDCs), T helper cells, T follicular helper (Tfh) cells, T helper1 (Th1) cells, T helper 2 (Th2) cells, tumor infiltrating lymphocytes (TIL), regulatory T (Treg) cells), and immune functions (antigen presenting cell (APC) co-inhibition, APC co-stimulation, chemokine receptors (CCR), check-point, cytolytic activity, human leukocyte antigen (HLA), inflammation-promoting, major histocompatibility complex (MHC) class I, Parainflammation, T cell co-inhibition, T cell co-stimulation, type-I interferon (IFN) response, type II IFN response) between control and UC group. A p-value < 0.05 was used to filter the samples. Heat map of the 29 types of immune cells and immune function in samples was produced by the "pheatmap" package. Levels of immune cells and immune function between UC and control samples were visualized by the "ggpubr" package. Correlative heat map was performed using

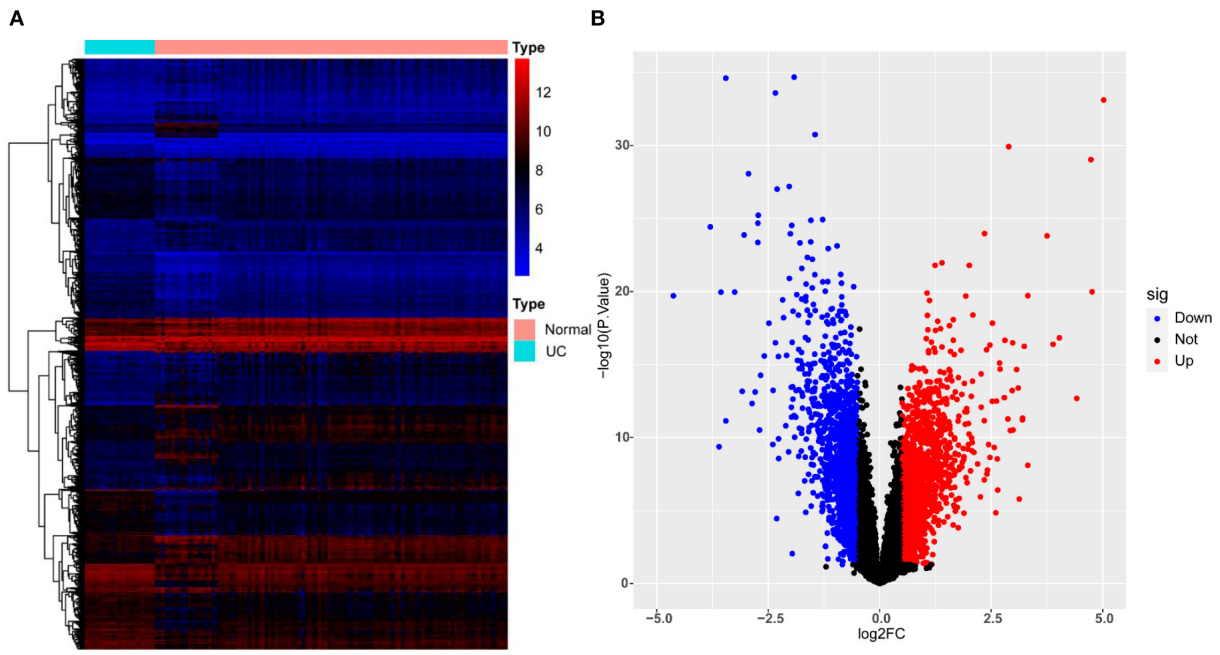

FIGURE 2
Heatmap and volcano plot for the DEGs identified from GSE87473. **(A)** Heat map for DEGs. Each row of the heat map represents one DEG, and each column represents one sample, either healthy or UC; the red and blue colors represent upregulated and downregulated DEGs, respectively. **(B)** Volcano plot for DEGs. Red plot points represent upregulated DEGs, and blue plot points show downregulated DEGs.

the "corrplot" package to reveal the correlation of immune cells and immune function.

The construction of diagnostic model

The top genes with the most significant differences in expression between healthy and UC samples were considered as diagnostic genes for UC, and they were evaluated by the "psych" package for their relevance to immune cell and immune function subtypes. After the feature selection, the diagnostic genes most strongly associated with immune infiltration were used to construct diagnostic models with "rms" package. The prediction performance of the model was quantified by nomogram, non-correlated nomogram, and receiver operating characteristic (ROC) curve which was performed with "ROCR" package.

Model performance and validation

The expression data of GSE75214 and GSE92415 were used to verify the robustness of diagnostic model. The area under the curve (AUC) from a ROC curve analysis was calculated to test the diagnostic performance of the model: ROC-AUC $\geq$ 0.9 indicates outstanding discrimination; 0.8 $\leq$ ROC-AUC $<$ 0.9 indicates excellent discrimination; 0.7 $\leq$ ROC-AUC $<$ 0.8 indicates acceptable discrimination; and ROC = 0.5 indicates no discrimination (21).

Statistical analysis

Categorical variables were presented as percentages, while continuous variables were presented as the mean ± standard deviation. All data analyses in this study were performed using R software (version 4.1.2), and the main packages that were used for t bioinformatics analysis are provided in Table 2. A p-value < 0.05 was considered significant for screening DEGs, enrichment analysis, correlation analysis, and immune infiltration analysis.

Results

Differential gene screening

3111 DEGs (1,608 up-regulated and 1,503 down-regulated genes) were identified from GSE87473. The top 10 up-regulated DEGs involved were: DUOX2, MMP3, SLC6A14, DEFB4A, TNIP3, S100A8, CXCL1, DUOXA2, REG1A, and MMP10, CALU while the top 10 down-regulated DEGs were: AQP8, SLC51A, CLDN8, HMGCS2, DPP10-AS1, PCK1, ABCG2, SLC26A2, GBA3, and MEP1B. Figure 2 presents the details of the heatmap and volcano plot of DEGs.

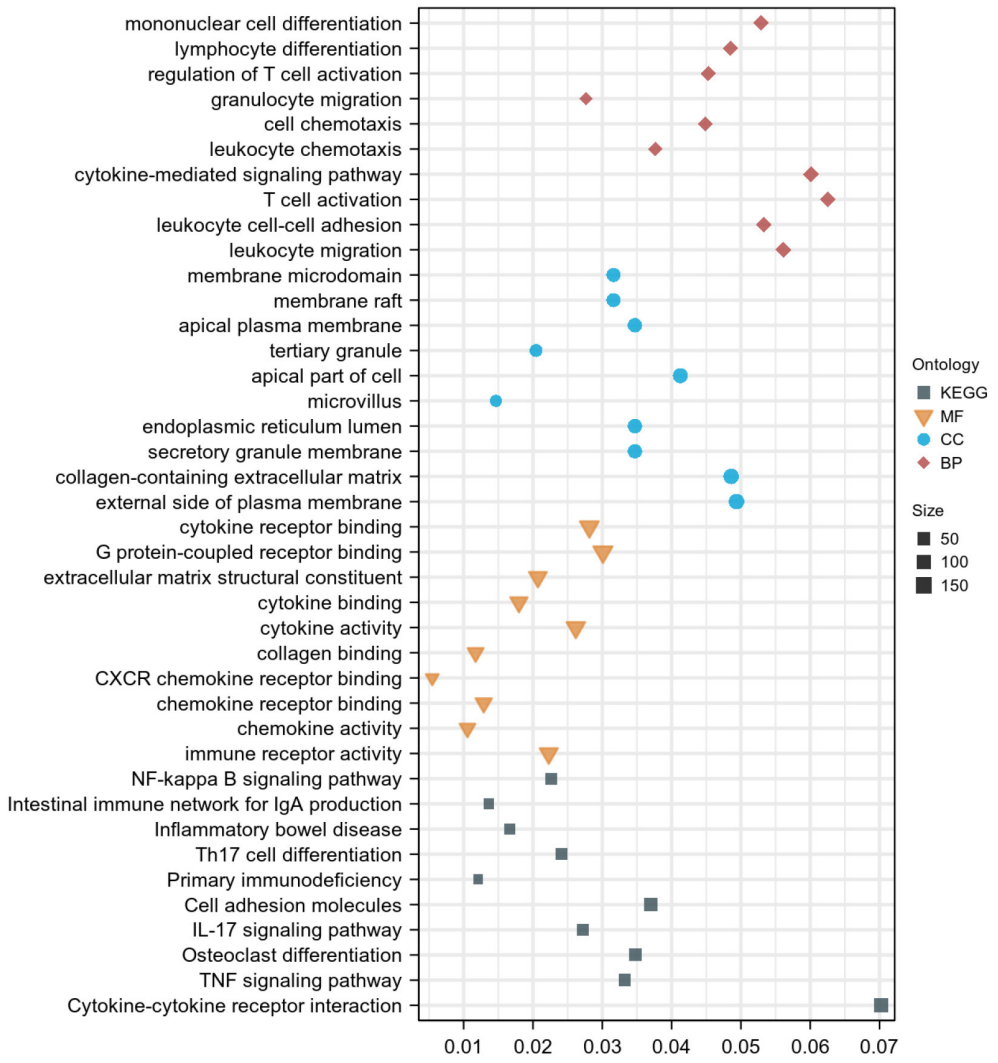

FIGURE 3
GO and KEGG enrichment analysis of DEGs. Dot graph of GO and KEGG enrichment analysis, the size of the dot represents the number of enriched genes. The figure shows the terms with $p < 0.05$.

Functional enrichment analysis

For GO analysis, DEGs were significantly enriched in the following process: leukocyte migration, leukocyte cell-cell adhesion, T cell activation, cytokine-mediated signaling pathway, leukocyte chemotaxis, cell chemotaxis, granulocyte migration, regulation of T cell activation, lymphocyte differentiation, mononuclear cell differentiation, immune receptor activity, chemokine activity, chemokine receptor binding, CXCR chemokine receptor binding, collagen binding, cytokine activity, cytokine binding, extracellular matrix structural constituent, and G protein–coupled receptor binding (Figure 3). For KEGG analysis, genes were significantly enriched in immune-related pathways such as TNF signaling pathway, Osteoclast differentiation, IL-17 signaling pathway, Th17 cell differentiation, and NF-kappa B signaling pathway (Figure 3). For DO analysis, DEGs were significantly enriched in infectious diseases, inflammatory diseases, and cancer (Figure 4).

Immune infiltration analysis

The normalized enrichment score of immune infiltrates is presented in the heat map (Figure 5). The results of differential analysis of immune cell revealed that UC patients had a higher level of aDCs, B cells, $CD8^+$ T cells, DCs, iDCs, Macrophages, Neutrophils, pDCs, T helper cells, Tfh, Th1 cells, Th2 cells, TIL and Treg than healthy subjects (Table 3; Figure 6A).

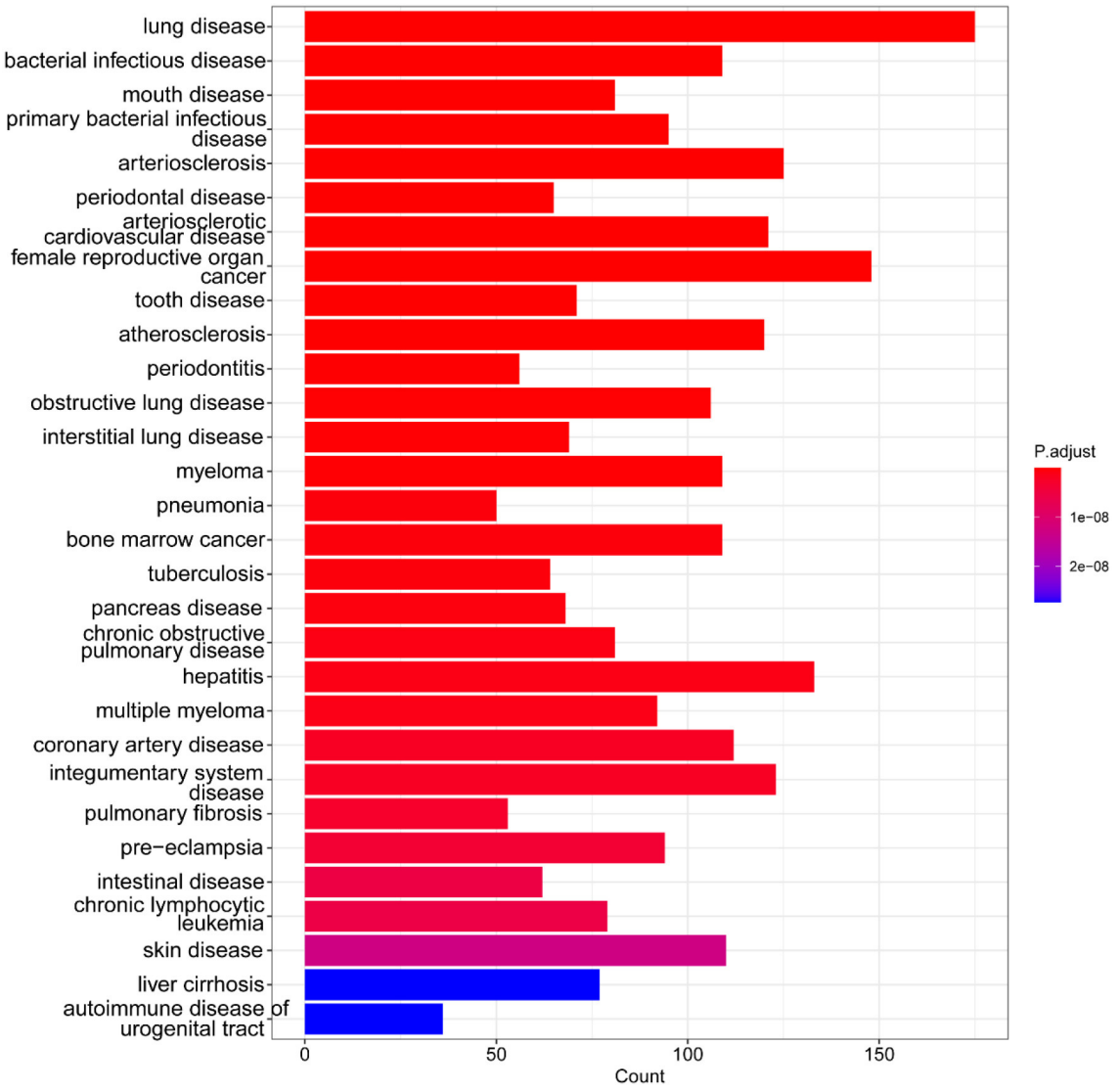

FIGURE 4
DO enrichment analysis of DEGs. Bar graph of DO enrichment analyses, the lengths of the bars represent the number of enriched genes, color represents the significance, increasing gradually from blue to red. The figure shows the terms with $p < 0.05$.

The results of differential immune function analysis revealed that significant APC co-inhibition, APC co-stimulation, CCR, Check-point, HLA, Inflammation-promoting, MHC class I, Parainflammation, T cell co-inhibition, T cell co-stimulation, Type I IFN Reponse, and Type II IFN Reponse were observed in UC patients (Table 3; Figure 6B). Details of these biomarkers in patients with UC are presented in Table 3.

The correlation analysis revealed that activated B cells were not related to Mast cells; $CD8^+$ T cells were not related to iDCs; iDCs were not related to NK cells; Macrophages were not related to Mast cells or NK cells; Mast cells were not related to Neutrophils, pDCs, Th2 cells or Treg; and NK cells were not related to Treg (Figure 7A). However, strong correlations were observed for all other types of immune cells and immune function (Figure 7).

The construction of diagnostic model

Results of Pearson correlation analysis revealed that all upregulated diagnostic genes were significantly positively correlated with almost all immune cell subtypes and immune function subtypes (except $CD8^+$ T cells and NK cells). Similarly, almost all down-regulated diagnostic genes were negatively correlated with almost all immune cell subtypes and immune function subtypes (except $CD8^+$ T cells and NK cells) (Figure 8).

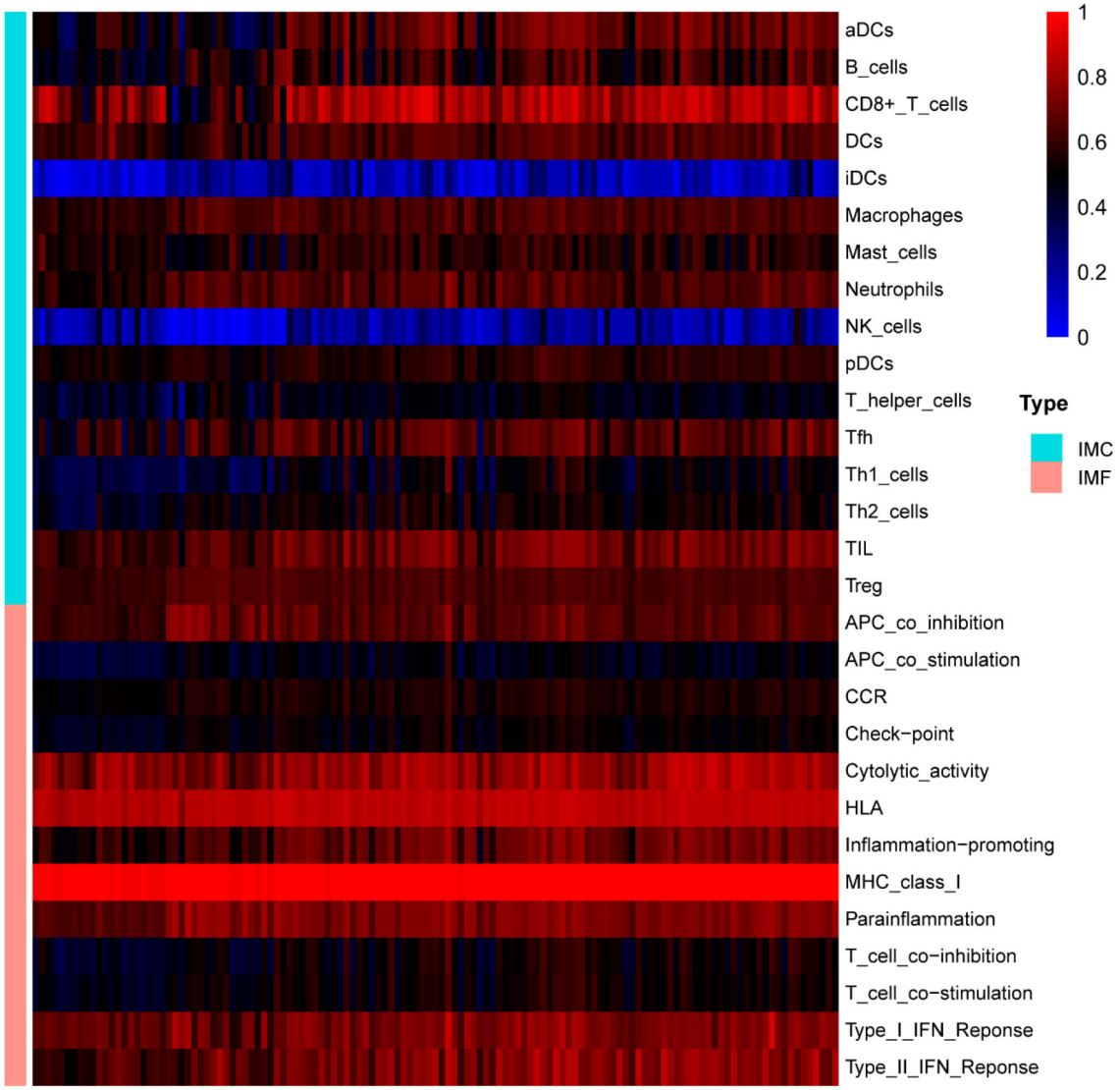

FIGURE 5
The overall landscape of immune infiltration in UC. Each row represents one sample, either healthy or UC, each column represents a type of immune cell or immune function. IMC, immune cell; IMF, immune function.

Inspired by the above results, five diagnostic genes most associated with immune infiltration (HMGCS2, CLDN8, AQP8, DEFB4A, MMP3) were used to construct a diagnostic model for UC. Details of these biomarkers in patients with UC are presented in Table 4.

The nomogram showed the diagnostic efficacy of the model constructed with these predicted diagnostic genes for UC (Figure 9A). Based on the calibration curve predicted by the uncorrelated nomogram, the performance of the column line plot was close to the ideal model, suggesting that the predictive value of the model is credible (Figure 9B). Similarly, ROC-AUC of the risk score was 0.897, which indicates excellent discrimination of the model (Figure 9C).

Model performance and validation

To go step further validation, ROC curves were applied to assesses the prediction accuracy of the model. The ROC-AUC of the risk score was 0.871 in GSE75214 and 0.908 in GSE92415, respectively, indicating excellent discrimination of the model (Figure 10).

Discussion

The development of UC involves genetic susceptibility, environmental factors and disturbances in the gut microbiota

TABLE 3 Results of immune infiltration analysis.

Immune infiltrates	Control		UC		*P*-value
	Median, interquartile range	Mean ± standard deviation	Median, interquartile range	Mean ± standard deviation	
aDCs	0.501 [0.468, 0.570]	0.495 ± 0.103	0.683 [0.584, 0.731]	0.641 ± 0.120	<0.05
B cells	0.46 [0.410, 0489]	0.459 ± 0.058	0.582 [0.508, 0.662]	0.584 ± 0.097	<0.05
$CD8^+$ T cells	0.723 [0.613, 0.822]	0.716 ± 0.136	0.825 [0.723, 0.873]	0.775 ± 0.147	<0.05
DCs	0.603 [0.578, 0.628]	0.608 ± 0.053	0.649 [0.607, 0.677]	0.636 ± 0.063	<0.05
iDCs	0.072 [0.052, 0.104]	0.085 ± 0.050	0.157 [0.109, 0.240]	0.180 ± 0.091	<0.05
Macrophages	0.566 [0.549, 0.592]	0.571 ± 0.034	0.644 [0.622, 0.664]	0.643 ± 0.037	<0.05
Mast cells	0.556 [0.527, 0.583]	0.558 ± 0.049	0.567 [0.527, 0.609]	0.562 ± 0.067	>0.05
Neutrophils	0.543 [0.514, 0.589]	0.553 ± 0.043	0.655 [0.613, 0.678]	0.646 ± 0.050	<0.05
NK cells	0.17 [0.138, 0.270]	0.206 ± 0.103	0.182 [0.114, 0.241]	0.185 ± 0.101	>0.05
pDCs	0.534 [0.507, 0.546]	0.531 ± 0.027	0.580 [0.557, 0.597]	0.575 ± 0.039	<0.05
T helper cells	0.431 [0.387, 0.462]	0.424 ± 0.048	0.471 [0.440, 0.499]	0.467 ± 0.058	<0.05
Tfh	0.526 [0.447, 0.606]	0.523 ± 0.104	0.661 [0.590, 0.698]	0.641 ± 0.084	<0.05
Th1 cells	0.378 [0.361, 0.390]	0.381 ± 0.036	0.482 [0.436, 0.536]	0.479 ± 0.083	<0.05
Th2 cells	0.429 [0.376, 0.460]	0.426 ± 0.057	0.535 [0.503, 0.564]	0.528 ± 0.053	<0.05
TIL	0.601 [0.562, 0.620]	0.600 ± 0.045	0.707 [0.658, 0.756]	0.701 ± 0.068	<0.05
Treg	0.600 [0.589, 0.611]	0.600 ± 0.015	0.642 [0.628, 0.652]	0.641 ± 0.021	<0.05
APC co-inhibition	0.606 [0.586, 0.617]	0.601 ± 0.023	0.667 [0.637, 0.701]	0.671 ± 0.051	<0.05
APC co-stimulation	0.396 [0.381, 0.416]	0.396 ± 0.021	0.472 [0.435, 0.503]	0.470 ± 0.045	<0.05
CCR	0.490[0.482, 0.504]	0.492 ± 0.014	0.553 [0.539, 0.576]	0.554 ± 0.029	<0.05
Check-point	0.437 [0.430, 0.452]	0.444 ± 0.022	0.509 [0.485, 0.538]	0.509 ± 0.038	<0.05
Cytolytic activity	0.768 [0.706, 0.814]	0.755 ± 0.076	0.797 [0.746, 0.837]	0.787 ± 0.069	>0.05
HLA	0.840 [0.827, 0.854]	0.839 ± 0.020	0.864 [0.854, 0.874]	0.863 ± 0.020	<0.05
Inflammation-promoting	0.546 [0.507, 0.594]	0.552 ± 0.062	0.703 [0.659, 0.733]	0.687 ± 0.075	<0.05
MHC class I	0.982 [0.977, 0.986]	0.981 ± 0.006	0.993 [0.991, 0.995]	0.992 ± 0.006	<0.05
Parainflammation	0.645 [0.635, 0.666]	0.651 ± 0.022	0.753 [0.724, 0.777]	0.748 ± 0.041	<0.05
T cell co-inhibition	0.423 [0.416, 0.446]	0.434 ± 0.039	0.524 [0.477, 0.563]	0.518 ± 0.066	<0.05
T cell co-stimulation	0.441 [0.423,0.468]	0.445 ± 0.031	0.532 [0.489, 0.560]	0.526 ± 0.052	<0.05
Type I IFN response	0.689 [0.671, 0.707]	0.682 ± 0.038	0.730 [0.703, 0.766]	0.729 ± 0.064	<0.05
Type II IFN response	0.606 [0.561, 0.644]	0.608 ± 0.063	0.740 [0.686, 0.792]	0.734 ± 0.080	<0.05

and is characterized by an abnormal mucosal immune response and a persistent inflammatory infiltrate. Pathological mechanisms that have not been fully elucidated have limited the development of early diagnosis and treatment of UC. Currently, integrated microarray-based bioinformatics analysis has been widely used to identify potential novel biomarkers for various diseases, which is important for specifying molecular markers and diagnosing UC early in the disease process. In this study, DEGs were identified to be mainly enriched in immune response-related pathways. Immune infiltration analysis suggested significant differences in immune cell and immune function types between UC patients and controls. Correlation analysis of DEGs and immune microenvironment was performed, and five immune-related genes (HMGCS2, CLDN8, AQP8, DEFB4A, MMP3) were identified and were subsequently used to construct a diagnostic model of UC. Encouragingly, this model showed good diagnostic performance for UC, and data from the other two datasets further validated the accuracy of this model for UC diagnosis.

With data from GEO, 3111 DEGs (1,608 up-regulated genes and 1,503 down-regulated genes) were obtained in UC by comprehensive analysis of the microarray dataset (GSE87473). Further GO analysis indicated that these DEGs were significantly enriched in the leukocyte migration, leukocyte cell-cell adhesion, T cell activation, cytokine-mediated signaling pathway, leukocyte chemotaxis, cell chemotaxis, granulocyte migration, regulation of T cell activation, lymphocyte differentiation, mononuclear cell differentiation, immune receptor activity, chemokine activity, chemokine receptor binding, CXCR chemokine receptor binding, collagen binding,

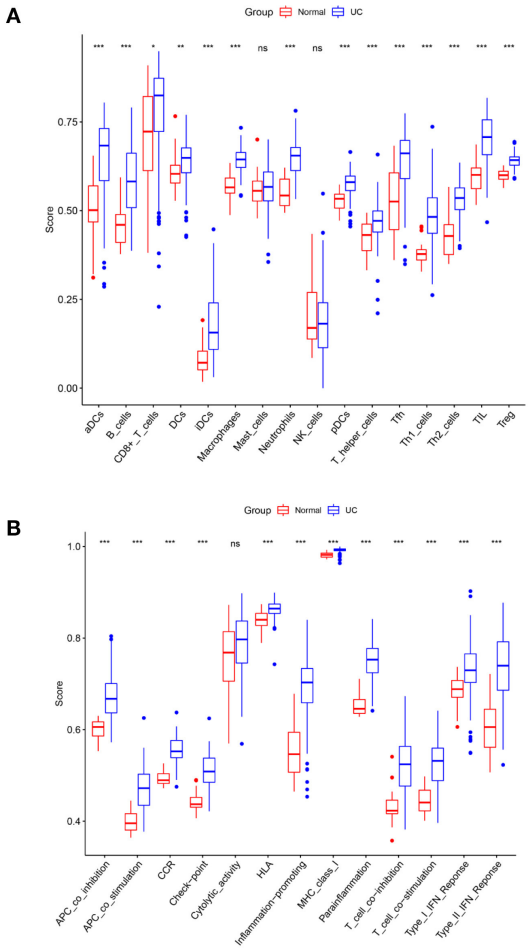

FIGURE 6
Comparison of immune cell subtypes **(A)** and immune function subtypes **(B)** between UC patients and controls. Red and blue colors represent healthy and UC samples, respectively. Differences between groups are indicated by "*". $*p < 0.05$; $**p < 0.01$; $***p < 0.001$.

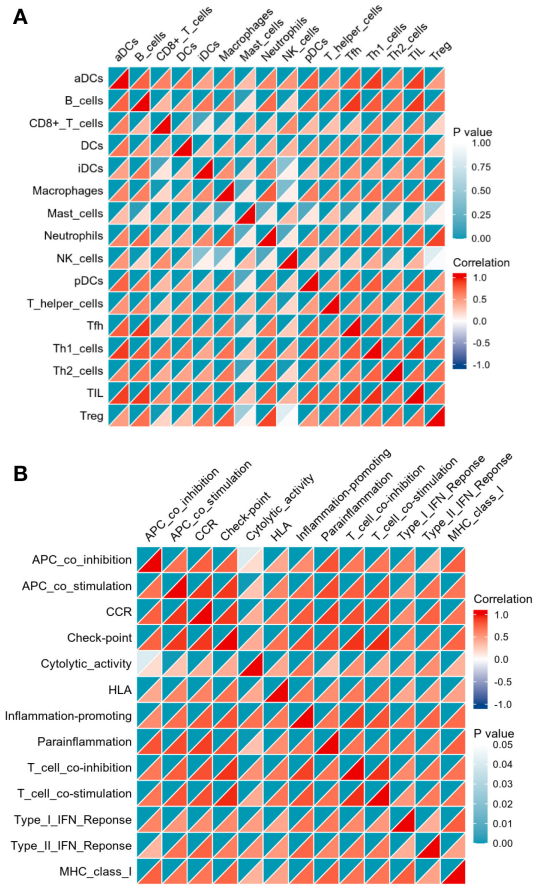

FIGURE 7
Correlation matrix of immune cell subtypes **(A)** and immune function subtypes **(B)**. Both horizontal and vertical axes demonstrate immune cell subtypes and immune function subtypes (blue for negative correlation, white for weak or no correlation), red for positive correlation). Color represents the significance, increasing gradually from white to indigo.

cytokine activity, cytokine binding, extracellular matrix structural constituent, and G protein–coupled receptor binding. For KEGG analysis, genes were significantly enriched in immune-related pathways such as TNF signaling pathway, Osteoclast differentiation, IL-17 signaling pathway, Th17 cell differentiation, and NF-kappa B signaling pathway. For DO analysis, DEGs were significantly enriched in lung disease, intestinal disease, mouth disease, primary bacterial infectious disease, arteriosclerosis, IBD, arteriosclerotic cardiovascular disease, female reproductive organ cancer, and tooth disease. In summary, results of the bioinformatics analyses suggested that these DEGs were closely related to immune cell infiltration in UC. These findings further validate the key role of immune abnormalities in the pathological progression of UC (22).

Inspired by the results of functional analysis of DEGs, immune infiltration analysis was further performed. The results revealed that there were significant differences in aDCs, B cells, $CD8^+$ Tcells, DCs, iDCs, Macrophages, Neutrophils, pDCs, T helper cells, Tfh, Th1 cells, Th2 cells, TIL and Treg between colon tissue in UC patients and the healthy group. In fact, most of the above immune cells have been reported to be in an abnormal state in UC by previous studies (23, 24). Therefore, these identified immune cells may be involved in the development and progression of UC. Top genes with the most significant expression differences between healthy and UC samples were considered as potential diagnostic genes for UC. After feature selection, HMGCS2, CLDN8, AQP8, DEFB4A, and MMP3 were identified as having strong association with immune infiltration, and thus they may be key genes that

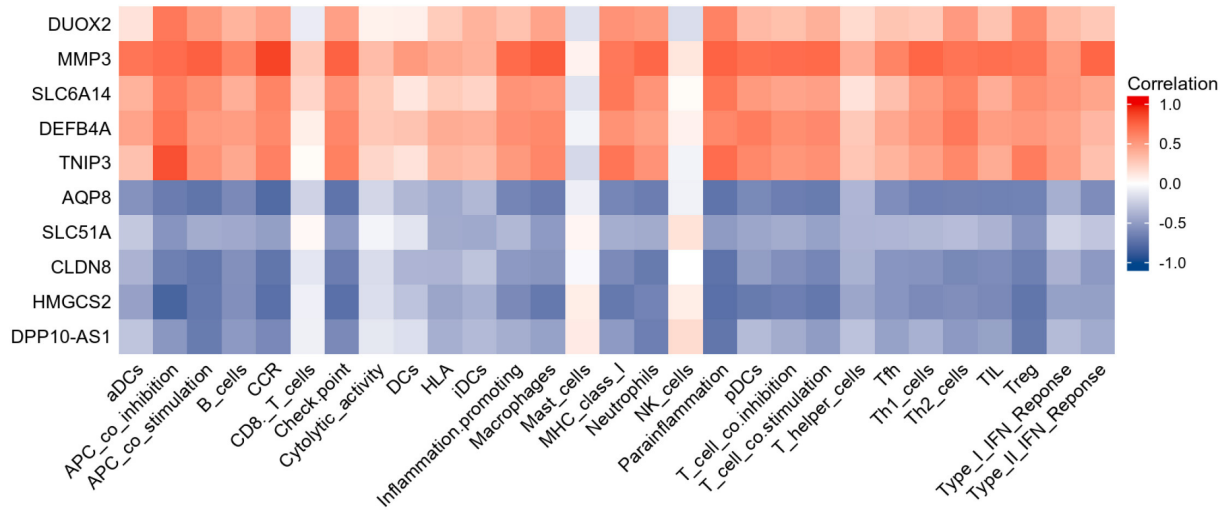

FIGURE 8
Heatmap showing the relationship between candidate genes and immune infiltrate. Red is positive correlation and blue is negative correlation.

TABLE 4 Details of genes in diagnostic model.

Genes	Sig	Log2FC	Average expression
AQP8	Down	−4.626352947	8.071990976
CLDN8	Down	−3.601376685	6.380647165
HMGCS2	Down	−3.559635503	7.584732929
DEFB4A	Up	4.418274083	8.68987552
MMP3	Up	4.763273249	9.167130008

identify the immune features of UC and are involved in UC immune regulation.

Inspired by the results of functional analysis of DEGs and immune infiltration analysis, HMGCS2, CLDN8, AQP8, DEFB4A, and MMP3 were applied to construct a diagnostic model for UC. The nomogram showed the well diagnostic efficacy of the model constructed with these predicted diagnostic genes for UC. The calibration curve for the uncorrelated nomogram prediction showed that the performance of the column line plot was close to the ideal model, suggesting that the predictive value of the model is credible. Similarly, ROC-AUC of the risk score was 0.897, suggesting a high diagnostic efficiency of the diagnostic marker gene model. To go step further validation, data from GSE75214 and GSE92415 were applied to assesses the prediction accuracy of the model. The results revealed that ROC-AUC of the risk score were 0.871 in GSE75214 and 0.908 in GSE92415, respectively, indicating excellent discrimination of the diagnostic model.

A total of five immune-related genes were included in the diagnostic model. The protein encoded by HMGCS2 belongs to the HMG-CoA synthase family of mitochondrial enzymes that catalyze the first reaction of ketogenesis, which is a crucial alternative metabolic pathway and is involved in the regulation of the body's immune function (25). Restoration of ketogenesis enhances immune cell effects (26) and attenuates the activation of pro-inflammatory macrophages (27). The protein levels of HMGCS2 in the intestinal epithelium of UC patients were reported to be sharply decreased compared to healthy samples (28). Increased ketogenesis may help to counteract intestinal inflammation, and conversely, its suppression may exacerbate intestinal pathology (28). The CLDN8 gene encodes a member protein of the claudin family. Claudins are integral membrane proteins and components of tight junction chains that play a key role in maintaining the integrity of the intestinal mucosal barrier. CLDN8 was reported to be significantly downregulated in the biological colon of IBD patients, and similar results were observed in colitis mice (29). AQP8 encodes an epithelial water transport protein specifically expressed in colonic absorptive cells, and it was found to be significantly downregulated in UC patients compared to healthy controls (30, 31). In addition, AQP8 was observed to promote H_2O_2 diffusion in experimental mouse models, which suggested its balance and regulatory effects on antioxidant pathways (32). DEFB4A encodes defensin, beta 4, an antibiotic peptide locally regulated by inflammation, which has been shown to be involved in the pathological process of IBD. Results of high-throughput sequencing suggested that the composition of the microbiota differs significantly between UC and non-IBD. Alterations in the microbiota can affect antimicrobial peptide expression, which in turn is involved in the progression of IBD (33). Various studies have highlighted the involvement of specific matrix metalloprotease in IBD: MMP3 transcript or protein levels are upregulated in the mucosa of inflammatory IBD or in the

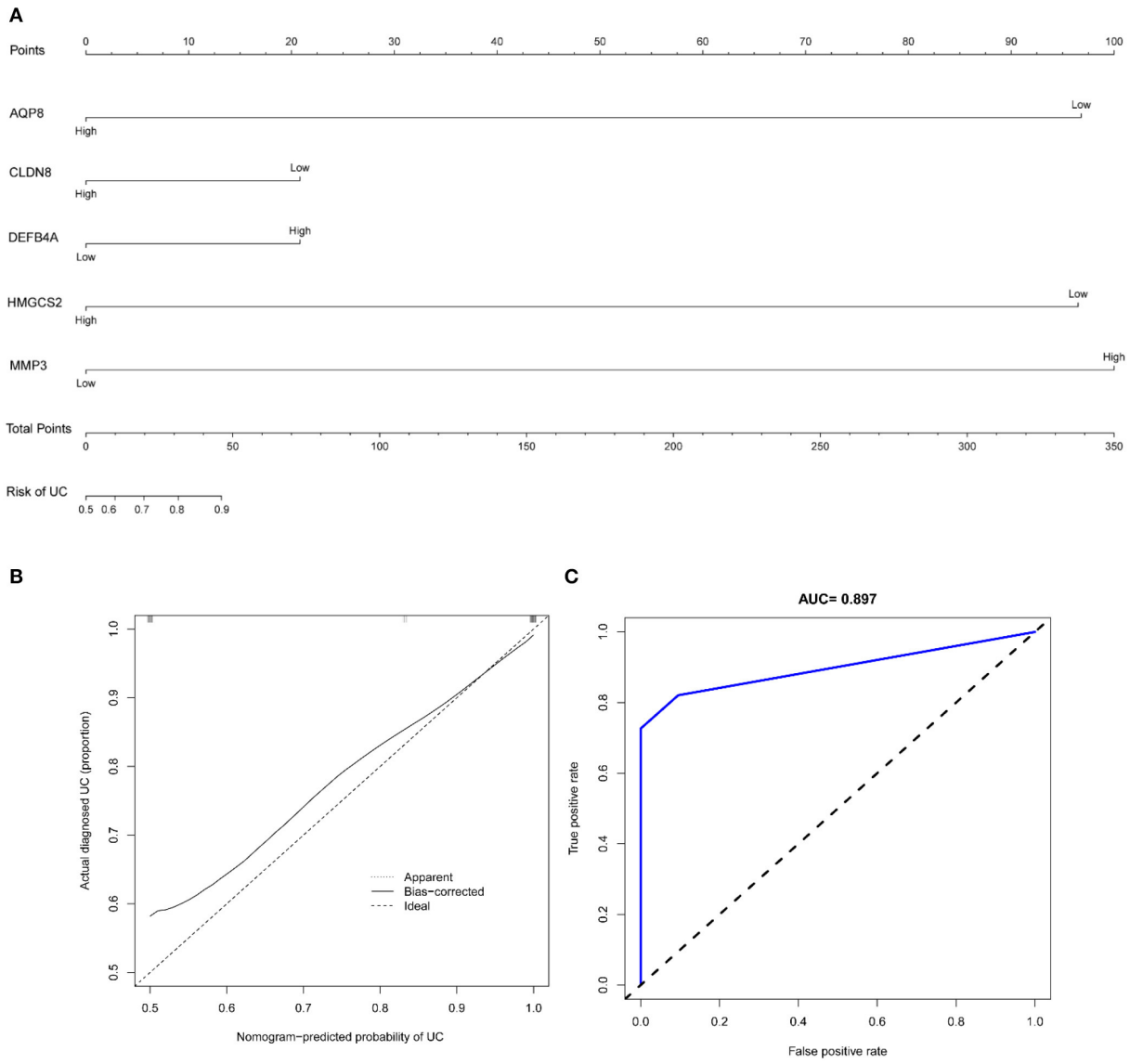

FIGURE 9
(A) Nomogram of diagnostic marker genes; **(B)** calibration curve of non-correlation nomogram prediction in the cohort; **(C)** ROC curve for the diagnostic efficacy of diagnostic model.

serum of IBD patients, and MMP protein hydrolytic activity is increased in epithelial-derived cells of inflammatory IBD (34–36). It was found that MMP3 expression was significantly upregulated in inflammatory colonic segments of IBD patients compared to non-inflammatory regions (37). In addition, MMP3 serum assay possesses a suggestive role for early response to infliximab treatment of UC (38). In summary, these key genes are all involved in the development of UC, and their inclusion in our diagnostic model of UC is reasonable.

Although this study applied a relatively large sample size to characterize the immune microenvironment and construct a diagnostic model for UC by integrating the GEO dataset, limitations should be acknowledged. First, this study explored the infiltration of immune cells by ssGSEA and found that immune cells play an important role in the pathological progression of UC. Therefore, it is crucial to validate our findings by flow cytometry. Second, although the present model may serve as a valid predictive tool for UC diagnosis, the true predictive value of the model should be prospectively validated in future independent and multicentered-studies with larger sample sizes.

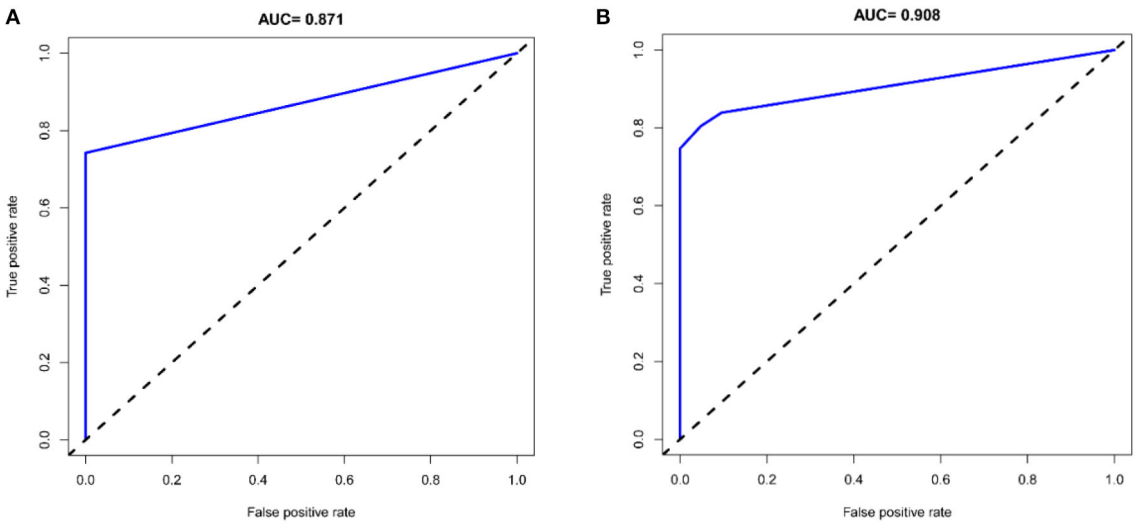

FIGURE 10
ROC curves for the model in the validation dataset of GSE75214 **(A)** and GSE92415 **(B)**.

Conclusion

In conclusion, this study identified the immune infiltrate characteristics and five immune-related genes for UC. The model based on the immune -related genes facilitates the early diagnosis of UC and provides a basis for the evaluation of the prognosis of UC.

Data availability statement

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

Author contributions

JH drafted the manuscript. JZ, FW, and BZ helped with implementation of this work. XT contributed to the methodology, review, and editing of the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (No. 81830118), China Academy of Chinese Medical Sciences Innovation Fund (No. CI 2021A01012), China Academy of Chinese Medical Sciences Excellent Young Talent Cultivation Fund (No. ZZ 15-YQ-002), and Administration of Traditional Chinese Medicine Digestive Refractory Disease Inheritance and Innovation Team Project (No. ZYYCXTD-C-C202010).

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/fpubh.2022.1003002/full#supplementary-material

References

1. Huang J, Zhang J, Ma J, Ma J, Liu J, Wang FJ, et al. Inhibiting ferroptosis: a novel approach for ulcerative colitis therapeutics. *Oxid Med Cell Longev.* (2022) 2022:9678625. doi: 10.1155/2022/9678625

2. Talley NJ, Abreu MT, Achkar JP, Bernstein CN, Dubinsky MC, Hanauer SB, et al. An evidence-based systematic review on medical therapies for inflammatory bowel disease. *Am J Gastroenterol.* (2011) 106:S2–25. doi: 10.1038/ajg.2011.58

3. Cohen RD Yu AP, Wu EQ, Xie J, Mulani PM, Chao J. Systematic review: the costs of ulcerative colitis in Western countries. *Aliment Pharmacol Ther.* (2010) 31:693–707. doi: 10.1111/j.1365-2036.2010.04234.x

4. Agrawal M, Spencer EA, Colombel JF, Ungaro RC. Approach to the management of recently diagnosed inflammatory bowel disease patients: a user's guide for adult and pediatric gastroenterologists. *Gastroenterology.* (2021) 161:47–65. doi: 10.1053/j.gastro.2021.04.063

5. Du L, Ha C. Epidemiology and pathogenesis of ulcerative colitis. *Gastroenterol Clin North Am.* (2020) 49:643–54. doi: 10.1016/j.gtc.2020.07.005

6. Bai X, Liu W, Chen H, Zuo T, Wu X. Immune cell landscaping reveals distinct immune signatures of inflammatory bowel disease. *Front Immunol.* (2022) 13:861790. doi: 10.3389/fimmu.2022.861790

7. Kałuzna A, Olczyk P, Komosińska-Vassev K. The role of innate and adaptive immune cells in the pathogenesis and development of the inflammatory response in ulcerative colitis. *J Clin Med.* (2022) 11:400. doi: 10.3390/jcm11020400

8. Wang Y, Zhang H, He H, Ai K, Yu W, Xiao X, et al. LRCH1 suppresses migration of CD4+ T cells and refers to disease activity in ulcerative colitis. *Int J Med Sci.* (2020) 17:599–608. doi: 10.7150/ijms.39106

9. Yang W, Liu H, Xu L, Yu T, Zhao X, Yao S, et al. GPR120 inhibits colitis through regulation of CD4+ T cell interleukin 10 production. *Gastroenterology.* (2022) 162:150–65. doi: 10.1053/j.gastro.2021.09.018

10. Yang ZJ, Wang BY, Wang TT, Wang FF, Guo YX, Hua RX, et al. Functions of Dendritic Cells and Its Association with Intestinal Diseases. *Cells.* (2021) 10:583. doi: 10.3390/cells10030583

11. Yan J, Pandey SP, Barnes BJ, Turner JR, Abraham C, T. Cell-intrinsic IRF5 regulates T cell signaling, migration, and differentiation and promotes intestinal inflammation. *Cell Rep.* (2020) 31:107820. doi: 10.1016/j.celrep.2020.107820

12. Stakenborg M, Verstockt B, Meroni E, Goverse G, De Simone V, Verstockt S, et al. Neutrophilic HGF-MET signalling exacerbates intestinal inflammation. *J Crohns Colitis.* (2020) 14:1748–58. doi: 10.1093/ecco-jcc/jjaa121

13. Hone Lopez S, Kats-Ugurlu G, Renken RJ, Buikema HJ, de Groot MR, Visschedijk MC, et al. Immune checkpoint inhibitor treatment induces colitis with heavy infiltration of CD8 + T cells and an infiltration pattern that resembles ulcerative colitis. *Virchows Arch.* (2021) 479:1119–29. doi: 10.1007/s00428-021-03170-x

14. Hegazy AN, West NR, Stubbington MJT, Wendt E, Suijker KIM, Datsi A, et al. Circulating and tissue-resident CD4+ T cells with reactivity to intestinal microbiota are abundant in healthy individuals and function is altered during inflammation. *Gastroenterology.* (2017) 153:1320–37. doi: 10.1053/j.gastro.2017.07.047

15. Hart AL, Al-Hassi HO, Rigby RJ, Bell SJ, Emmanuel AV, Knight SC, et al. Characteristics of intestinal dendritic cells in inflammatory bowel diseases. *Gastroenterology.* (2005) 129:50–65. doi: 10.1053/j.gastro.2005.05.013

16. Lu Z, Su H. Employing gene chip technology for monitoring and assessing soil heavy metal pollution. *Environ Monit Assess.* (2021) 194:2. doi: 10.1007/s10661-021-09650-6

17. Yu Z, Ma X, Zhang W, Chang X, An L, Niu M, et al. Microarray data mining and preliminary bioinformatics analysis of hepatitis D virus-associated hepatocellular carcinoma. *Biomed Res Int.* (2021) 2021:1093702. doi: 10.1155/2021/1093702

18. Zhao E, Zhou C, Chen S, A. signature of 14 immune-related gene pairs predicts overall survival in gastric cancer. *Clin Transl Oncol.* (2021) 23:265–74. doi: 10.1007/s12094-020-02414-7

19. Zhao E, Xie H, Zhang Y. Identification of differentially expressed genes associated with idiopathic pulmonary arterial hypertension by integrated bioinformatics approaches. *J Comput Biol.* (2021) 28:79–88. doi: 10.1089/cmb.2019.0433

20. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, et al. NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.* (2007) 35:D760–5. doi: 10.1093/nar/gkl887

21. Omlor W, Wahl AS, Sipilä P, Lütcke H, Laurenczy B, Chen IW, et al. Context-dependent limb movement encoding in neuronal populations of motor cortex. *Nat Commun.* (2019) 10:4812. doi: 10.1038/s41467-019-12670-z

22. van Unen V, Ouboter LF, Li N, Schreurs M, Abdelaal T, Kooy-Winkelaar Y, et al. Identification of a disease-associated network of intestinal immune cells in treatment-naive inflammatory bowel disease. *Front Immunol.* (2022) 13:893803. doi: 10.3389/fimmu.2022.893803

23. Luo Y, Liu S, Li H, Hou J, Lin W, Xu Z, et al. Mass cytometry and single-cell transcriptome analyses reveal the immune cell characteristics of ulcerative colitis. *Front Mol Biosci.* (2022) 9:859645. doi: 10.3389/fmolb.2022.859645

24. Mitsialis V, Wall S, Liu P, Ordovas-Montanes J, Parmet T, Vukovic M, et al. Single-cell analyses of colon and blood reveal distinct immune cell signatures of ulcerative colitis and crohn's disease. *Gastroenterology.* (2020) 159:591–608. doi: 10.1053/j.gastro.2020.04.074

25. Goldberg EL, Shchukina I, Asher JL, Sidorov S, Artyomov MN, Dixit VD. Ketogenesis activates metabolically protective γδ T cells in visceral adipose tissue. *Nat Metab.* (2020) 2:50–61. doi: 10.1038/s42255-019-0160-6

26. Wei R, Zhou Y, Li C, Rychahou P, Zhang S, Titlow WB, et al. Ketogenesis attenuates KLF5-dependent production of CXCL12 to overcome the immunosuppressive tumor microenvironment in colorectal cancer. *Cancer Res.* (2022) 82:1575–88. doi: 10.1158/0008-5472.CAN-21-2778

27. Zhang L, Shi J, Du D, Niu N, Liu S, Yang X, et al. Ketogenesis acts as an endogenous protective programme to restrain inflammatory macrophage activation during acute pancreatitis. *EBioMedicine.* (2022) 78:103959. doi: 10.1016/j.ebiom.2022.103959

28. Kim JT, Napier DL, Kim J, Li C, Lee EY, Weiss HL, et al. Ketogenesis alleviates TNFα-induced apoptosis and inflammatory responses in intestinal cells. *Free Radic Biol Med.* (2021) 172:90–100. doi: 10.1016/j.freeradbiomed.2021.05.032

29. Wang H, Chao K, Ng SC, Bai AH, Yu Q, Yu J, et al. Pro-inflammatory miR-223 mediates the cross-talk between the IL23 pathway and the intestinal barrier in inflammatory bowel disease. *Genome Biol.* (2016) 17:58. doi: 10.1186/s13059-016-0901-8

30. Dotti I, Mora-Buch R, Ferrer-Picón E, Planell N, Jung P, Masamunt MC, et al. Alterations in the epithelial stem cell compartment could contribute to permanent changes in the mucosa of patients with ulcerative colitis. *Gut.* (2017) 66:2069–79. doi: 10.1136/gutjnl-2016-312609

31. Min M, Peng LH, Sun G, Guo MZ, Qiu ZW, Yang YS. Aquaporin 8 expression is reduced and regulated by microRNAs in patients with ulcerative colitis. *Chin Med J.* (2013) 126:1532–7. doi: 10.3760/cma.j.issn.0366-6999.20122989

32. Te Velde AA, Pronk I, de Kort F, Stokkers PC. Glutathione peroxidase 2 and aquaporin 8 as new markers for colonic inflammation in experimental colitis and inflammatory bowel diseases: an important role for H2O2? *Eur J Gastroenterol Hepatol.* (2008) 20:555–60. doi: 10.1097/MEG.0b013e3282f45751

33. Jalanka J, Cheng J, Hiippala K, Ritari J, Salojärvi J, Ruuska T, et al. Colonic mucosal microbiota and association of bacterial taxa with the expression of host antimicrobial peptides in pediatric ulcerative colitis. *Int J Mol Sci.* (2020) 21:6044. doi: 10.3390/ijms21176044

34. Meijer MJ, Mieremet-Ooms MA, van der Zon AM, van Duijn W, van Hogezand RA, Sier CF, et al. Increased mucosal matrix metalloproteinase-1,−2,−3 and−9 activity in patients with inflammatory bowel disease and the relation with Crohn's disease phenotype. *Dig Liver Dis.* (2007) 39:733–9. doi: 10.1016/j.dld.2007.05.010

35. Hu J, Van den Steen PE, Sang QX, Opdenakker G. Matrix metalloproteinase inhibitors as therapy for inflammatory and vascular diseases. *Nat Rev Drug Discov.* (2007) 6:480–98. doi: 10.1038/nrd2308

36. Lakatos G, Hritz I, Varga MZ, Juhász M, Miheller P, Cierny G, et al. The impact of matrix metalloproteinases and their tissue inhibitors in inflammatory bowel diseases. *Dig Dis.* (2012) 30:289–95. doi: 10.1159/000336995

37. Medina C, Santana A, Paz-Cabrera MC, Parra-Blanco A, Nicolás D, Gimeno-Garcia AZ, et al. Increased activity and expression of gelatinases in ischemic colitis. *Dig Dis Sci.* (2006) 51:2393–9. doi: 10.1007/s10620-006-9255-5

38. Barberio B, D'Incà R, Facchin S, Dalla Gasperina M, Fohom Tagne CA, Cardin R, et al. Matrix metalloproteinase 3 predicts therapeutic response in inflammatory bowel disease patients treated with infliximab. *Inflamm Bowel Dis.* (2020) 26:756–63. doi: 10.1093/ibd/izz195

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