Cong *et al.*
European Journal of Medical Research (2025) 30:288
https://doi.org/10.1186/s40001-025-02501-x

European Journal of Medical Research

RESEARCH

Open Access

Diagnosis accuracy of machine learning for idiopathic pulmonary fibrosis: a systematic review and meta-analysis

Li Cong[1†], Ying Chen[1†], Xiaolei He[2], MaiLiKai KuErBan[2], Chao Wu[3] and Liping Chen[1*]

Abstract

Background The diagnosis of idiopathic pulmonary fibrosis (IPF) is complex, which requires lung biopsy, if necessary, and multidisciplinary discussions with specialists. Clinical diagnosis of the two ailments is particularly challenging due to the impact of interobserver variability. Several studies have endeavored to utilize image-based machine learning to diagnose IPF and its subtype of usual interstitial pneumonia (UIP). However, the diagnostic accuracy of this approach lacks evidence-based support.

Objective We conducted a systematic review and meta-analysis to explore the diagnostic efficiency of image-based machine learning (ML) for IPF.

Data sources and methods We comprehensively searched PubMed, Cochrane, Embase, and Web of Science databases up to August 24, 2024. During the meta-analysis, we carried out subgroup analyses by imaging source (computed radiography/computed tomography) and modeling type (deep learning/other) to evaluate its diagnostic performance for IPF.

Results The meta-analysis findings indicated that in the diagnosis of IPF, the C-index, sensitivity, and specificity of ML were 0.93 (95% CI 0.89–0.97), 0.79 (95% CI 0.73–0.83), and 0.84 (95% CI 0.79–0.88), respectively. The sensitivity of radiologists/clinicians in diagnosing IPF was 0.69 (95% CI 0.56–0.79), with a specificity of 0.93 (95% CI 0.74–0.98). For UIP diagnosis, the C-index of ML was 0.91 (95% CI 0.87–0.94), with a sensitivity of 0.92 (95% CI 0.80–0.97) and a specificity of 0.92 (95%CI 0.82–0.97). In contrast, the sensitivity of radiologists/clinicians in diagnosing UIP was 0.69 (95% CI 0.50–0.84), with a specificity of 0.90 (95% CI 0.82–0.94).

Conclusions Image-based machine learning techniques demonstrate robust data processing and recognition capabilities, providing strong support for accurate diagnosis of idiopathic pulmonary fibrosis and usual interstitial pneumonia. Future multicenter large-scale studies are warranted to develop more intelligent evaluation tools to further enhance clinical diagnostic efficiency.

Trial registration This study protocol was registered with PROSPERO (CRD42022383162).

Keywords Idiopathic pulmonary fibrosis, Deep learning, Machine learning

†Li Cong and Ying Chen have contributed equally to this work.

*Correspondence:
Liping Chen
mezatecong@163.com
Full list of author information is available at the end of the article

Introduction

Idiopathic pulmonary fibrosis (IPF) is a chronic interstitial lung disease of unknown origin, marked by progressive lung constriction. A study published in 2021 indicated that the incidence rate ranged from 0.9 to 13.0 per 100,000 people, and the prevalence rate varied from 3.3 to 45.1 per 100,000 people [1]. IPF, as the most prevalent idiopathic interstitial pneumonia (IIP), is linked to the poorest prognosis, and the estimated median survival period post-diagnosis is 3–5 years [2–5]. Early diagnosis is imperative for tailoring treatment plans, such as choosing between anti-fibrotic treatment and the treatment of pulmonary fibrosis due to other causes [6]. The diagnosis of IPF necessitates multidisciplinary discussions and collaborations among clinicians, radiologists, and pathologists [7]. However, the prolonged time from referral to multidisciplinary diagnosis is notable due to interobserver variabilities [8, 9]. Therefore, a diagnostic method that mitigates observer variability, and can accurately and swiftly differentiate between IPF and non-IPF interstitial lung diseases, is necessary in clinical practice.

The *Diagnosis and Treatment Guidelines for Idiopathic Pulmonary Fibrosis* for the first time included radiological UIP patterns in the definition of IPF, emphasizing the importance and diagnostic role of identifying high-resolution computed tomography (HRCT) UIP patterns. With the increasing importance of medical imaging in precision medicine for disease diagnosis, prognosis, and treatment planning [10], computed tomography (CT) emerges as a valuable tool providing visual data to enhance decision-making [11]. However, qualitative CT assessment remains challenging, leading to common discrepancies even among experienced experts [12, 13]. Hence, there is an urgent demand for an automated clinical tool that can assist clinicians in making precise and timely diagnoses.

Computer-aided diagnosis empowers doctors to leverage information technology (IT) to interpret and utilize various imaging techniques. The primary objective is to shorten diagnosis time and enhance accuracy, with IT serving as a supportive or even independent diagnostic option [14]. Computer-aided diagnostic algorithms fall within the realm of artificial intelligence (AI), mimicking human thinking. With the increasing wealth of imaging data and the availability of computing resources, AI is gaining popularity [15]. Quantitative imaging techniques in medical imaging are increasingly used in an exponential way [16]. In this context, some researchers are striving to develop tools that can assist in the diagnosis of IPF and UIP. Nevertheless, there is presently insufficient evidence-based backing for the detailed diagnostic value. To address this gap, we conducted a systematic review and meta-analysis to explore the diagnostic efficacy of image-based machine learning (ML) for IPF and UIP.

Methods

Study registration

Our study adhered to the preferred reporting guidelines (PRISMA 2020) for systematic review and was prospectively registered on Prospero (CRD42022383162).

Eligibility criteria

Inclusion criteria

(1) The subjects were patients with suspected interstitial lung disease.
(2) The types of studies included case–control study, cohort study, case–control study, and case-cohort study.
(3) A comprehensive ML model was developed to identify the prognosis of IFP or UPF or interstitial lung disease.
(4) Studies lacking external validation are also incorporated.
(5) Various ML studies published in the same dataset.
(6) Studies reported in English.

Exclusion criteria

(1) Research types: Meta-analysis, review, guidelines, expert opinions, and conference abstracts published without peer review.
(2) Only an analysis of risk factors was conducted, and a comprehensive ML model was not constructed.
(3) The following outcome indexes to assess the accuracy of ML model are missing.
(4) Validation of the maturity scale only.

Data sources and search strategy

On August 24, 2024, we systematically searched relevant literature in Cochrane, Embase, and Web of Science databases using a combination of subject terms and free words (with the start time being the establishment time of each database). Detailed information regarding the search materials is available in Table S1.

Study selection and data extraction

We imported the identified literature into Endnote20.0 and, following the removal of duplicates, checked the titles and abstracts. Subsequently, we downloaded the full texts, thoroughly read them, and selected documents that aligned with the objectives of our study. Before data

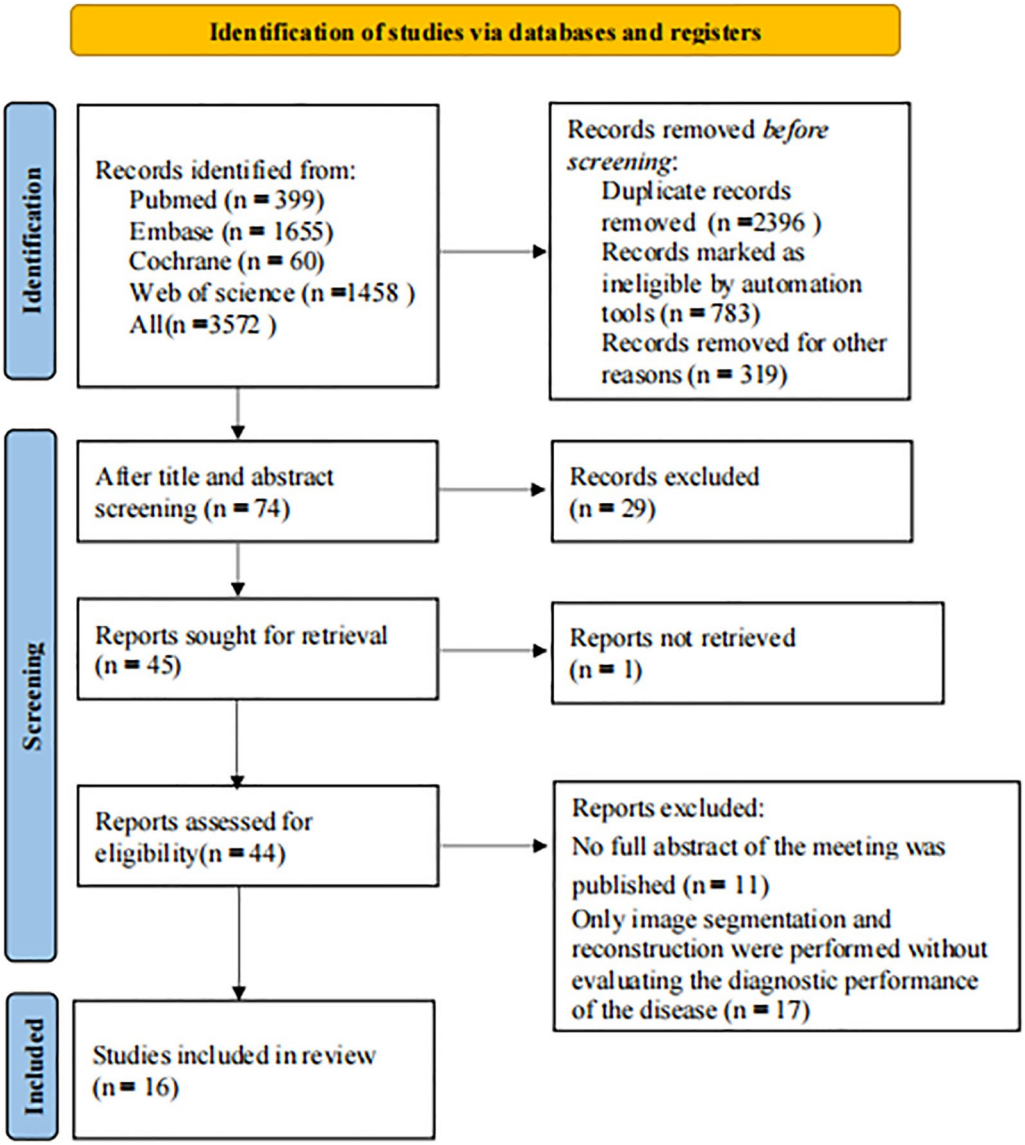

Fig. 1 The literature screening process

extraction, we established a standardized data extraction table to capture essential information such as author details, publication year, study type, diagnostic events, training set, validation set, type of model used, C-index, sensitivity, and specificity. The above literature screening and data extraction were carried out by two researchers (LC & YC) independently, and cross-checking was carried out after completion. In case of any dispute, the involvement of the third researcher (LPC) was consulted to assist in the decision.

Risk of bias in studies

We utilized QUADAS-2 to evaluate the risk of bias in the original studies. This evaluation comprised numerous questions distributed across four distinct areas: participants, predictive variables, results, and statistical analysis. These areas encompassed 2, 3, 6, and 9 specific questions, respectively, each with three response options (yes/possibly yes, no/probably no, and no information). A domain was deemed to be at high risk if at least one question was answered by no or probably no. Conversely, to be classified as low risk, all questions in a domain should be answered by yes or possibly yes. The overall risk of bias was considered low when all areas were deemed low risk and high when at least one domain was categorized as high risk.

The two researchers (LC & XLH) independently conducted the bias risk assessment using QUADAS-2 and

Table 1 Basic Information on the Included Literature

First author	Year of publication	Country of the author	Type of study	Patient source	Diagnostic events	Diagnostic criteria for the predictive event	Image source	Number of samples/images for the predicted event	Total number of samples/images	Validation set generation	Overfitting methods	Methods for addressing missing data	Model types	Modeling variables
Wenxi Yu	2022	The U.S	Case–control study	Multi-center study	IPF	Diagnosis of idiopathic pulmonary fibrosis. An official ATS/ERS/JRS/ALAT clinical practice guideline(2018)	CT	349	878	Internal validation	Fivefold cross-validation	None	MSGA + RF	CT image
Wenxi Yu	2022	The U.S	Case–control study	Multi-center study	IPF	Diagnosis of idiopathic pulmonary fibrosis. An official ATS/ERS/JRS/ALAT clinical practice guideline(2018)	CT	279	423	Internal validation	75%/25%	None	CNN	CT image
Refaee, T	2022	The Netherlands	Case–control study	Single-center study, registered database	IPF	The 2018 guidelines for the diagnosis of idiopathic pulmonary fibrosis (IPF) by ATS	CT	139	474	External validation	Fivefold cross-validation	None	HCR + DL	CT image
Hiroyuki Abe	2004	U.S	Case–control study	Registered database	IPF	Not mentioned	CT	37	96	External validation	/	None	ANN	CT image
Aya Fukushima	2004	Japan	Case–control study	Registered database	IPF	Pathological examination	CR	11	130	Internal validation	Leave-one-out method	None	ANN	CR image
Walsh, S. L. F	2022	The UK	Case–control study	Multi-center registered database	UIP	ATS/ERS/JRS/ALAT clinical practice guideline(2011)	CT	282	1157	External validation	/	None	Deep learning/CNN	CT image

Table 1 (continued)

First author	Year of publication	Country of the author	Type of study	Patient source	Diagnostic events	Diagnostic criteria for the predictive event	Image source	Number of samples/images for the predicted event	Total number of samples/images	Validation set generation	Overfitting methods	Methods for addressing missing data	Model types	Modeling variables
Alex Bratt	2022	The U.S	Case–control study	Multi-center registered database	UIP	The 2018 guidelines for the diagnosis of idiopathic pulmonary fibrosis (IPF) by ATS	CT	550	1239	External validation	/	None	Deep learning	CT image
Andreas Christe	2019	Switzerland	Case–control study	Multi-center registered database	UIP	ATS/ERS/JRS/ALAT clinical practice guideline(2011)	CT	51	105	External validation	/	None	CNN	CT image
Adrien Depeursinge	2015	Switzerland	Case–control study	Single-center registered database	UIP	ATS/ERS/JRS/ALAT clinical practice guideline(2011)	CT	15	33	Internal validation	Leave-one-out method	None	3D, SVM	CT image
Chi Wan Koo	2022	The U.S	Case–control study	Registered database	UIP	Guidelines for IPF (2018)	CT	486	911	Internal validation	Tenfold cross-validation	None	ML/DL	CT image
Hiram Shaish	2020	The U.S	Case–control study	Multi-center registered database	UIP	ATS/ERS/JRS/ALAT clinical practice guideline (2018)	CT	93	301	Internal validation	Fivefold cross-validation	None	CNN	CT image
Manoj V. Maddali	2023	The U.S	Case–control study	Multi-center registered database	IPF	ATS/ERS/JRS/ALAT clinical practice guideline (2018)	CT	129	295	External validation	/	None	CNN	CT image

Table 1 (continued)

First author	Year of publication	Country of the author	Type of study	Patient source	Diagnostic events	Diagnostic criteria for the predictive event	Image source	Number of samples/images for the predicted event	Total number of samples/images	Validation set generation	Overfitting methods	Methods for addressing missing data	Model types	Modeling variables
Jonathan H. Chung	2023	The U.S	Case–control study	Multi-center registered database	UIP	ATS/ERS/JRS/ALAT clinical practice guideline (2018)	CT	70	565	Internal validation	70%15%15%	None	CNN + SVM	CT image
Weijia Fan	2023	The U.S	Case–control study	Single-center registered	UIP	ATS/ERS/JRS/ALAT clinical practice guideline (2018)	CT	178	400	Internal validation	Tenfold cross-validation	None	CNN + BART	CT image
Stephen M. Humphries	2024	The U.S	Case–control study	Multi-center registered database	UIP	ATS/ERS/JRS/ALAT clinical practice guideline (2018)	CT	54	127	External validation	/	None	CNN + MIL	CT image
Marcello Chang	2023	The U.S	Case–control study	Multi-center registered database	UIP	ATS/ERS/JRS/ALAT clinical practice guideline (2018)	CT	45	295	Internal validation	Leave-one-out method	None	CNN + VGG	CT image

Fig. 2 (**A**) Risk of Bias Summary for Included Primary Studies, and (**B**) Risk of Bias Graph for Included Primary Studies

cross-checked their results after completion. In case of any dispute, the third researcher (CW) was consulted to assist in the adjudication.

Outcomes

The C-index was used as the primary outcome to depict the prediction precision of the ML models. Meanwhile, the sample size of the case group and the control group was seriously unbalanced, which was insufficient to reflect the prediction accuracy of the case group when the difference was enlarged. Therefore, the sensitivity and specificity of ML are also used as outcome measures.

Synthesis methods

We performed a meta-analysis to evaluate the overall accuracy indicator (c-index) for assessing ML models. In instances where the original studies did not provide a 95% confidence interval and standard error for the c-index, we consulted the research of Debray TP et al. [17] to estimate its standard error. Given the variations in the included variables and the inconsistency of parameters across different ML models, the random effects model was prioritized for the meta-analysis of the c-index.

Additionally, we conducted a meta-analysis of sensitivity and specificity employing a bivariate mixed-effects model. During the meta-analysis process, sensitivity and specificity were evaluated based on the diagnostic fourfold table. However, a majority of original studies did not report the diagnostic fourfold table. In such instances, we utilized two approaches to calculate the diagnostic fourfold table: 1. The diagnostic table was computed based on sensitivity, specificity, precision, and the number of cases; 2. Extraction of sensitivity and specificity were extracted based on the optimal Youden's index, followed by calculation with the number of cases. The meta-analysis for this study was conducted using R 4.2.0 (R development Core Team, Vienna, http://www.R-project.org).

Results

Study selection

A total of 3572 articles were retrieved in the initial search. Following the elimination of duplicates and irrelevant

Table 2 Quality evaluation of the trials

Risk of bias assessment	Case selection: Does the selection of cases produce bias?	Que1	Were consecutive or random cases included?
		Que2	Was the case–control study design avoided?
		Que3	Did the study avoid inappropriate exclusions?
	The trial to be evaluated: Is there any bias in the conduct or interpretation of the trial to be evaluated?	Que1	Was the interpretation of the results of the trial to be evaluated done without knowing the results of the gold standard trial?
		Que2	If a threshold was used, was it predetermined?
	Will there be bias in the implementation and interpretation of the gold standard?	Que1	Can the gold standard correctly distinguish between target disease states?
		Que2	Was blinding used in the interpretation of the gold standard results?
	Is there any bias in the flow of the cases?	Que1	Was there an appropriate time interval between the trial to be evaluated and the gold standard?
		Que2	Did all patients receive the same gold standard?
		Que3	Were all cases included in the analysis?
Evaluation of clinical practicability	que1		Matching of relevant included patients and background with evaluation questions
	que2		Evaluation of the matching between the implementation and interpretation of the trial to be evaluated and the evaluation questions
	que3		Applicability evaluation of gold standard

articles, 74 articles underwent a detailed full-text reading. Among them, we excluded unpublished conference abstracts, studies that only performed image segmentation and reconstruction, and studies that did not assess the diagnostic performance of outcome indicators for the disease. Ultimately, 16 articles [18–33] were included in this study, as illustrated in Fig. 1.

Study characteristics

The 16 included studies were published from 2004 to 2024, encompassing a total of 7209 study subjects, predominantly from North America, Europe, and Japan. Among these, 6 articles [18–22] focused on the diagnosis of IPF (Table 1). Notably, one of these articles [21] employed chest X-ray as a modeling variable, while the remaining four utilized high-resolution CT. The remaining 10 articles [23–28] centralized around the diagnosis of UIP. Within this subset, eight articles [23, 25–27] leveraged ML for the diagnosis of UIP patterns under HRCT, and two articles [24, 28] predicted pathological UIP patterns, all employing high-resolution CT as modeling variables. Across the 11 pieces of literature, ML algorithms predominantly included artificial neural network (ANN), convolutional neural network (CNN), and vector space model (SVM). Notably, seven studies [20–26] compared the diagnostic performance between human radiologists/clinical experts and ML models, as illustrated in Fig. 2.

Risk of bias in studies

All included studies adopted a case–control design, and only one study adopted a non-deep learning method to construct a model. In the construction of conventional ML, non-deep learning modeling variables needed to be manually encoded, which may introduce a higher bias risk, especially in the context of case–control studies. The impact of other image-based deep learning on case–control studies was relatively minimal. Consequently, there is a heightened risk of bias in the selection of cases. Various ML methods in the primary studies employed their respective evaluation criteria for the interpretation of the gold standard. This variability in evaluation criteria may introduce bias. However, the impact is considered low when implementing or interpreting the index tests, and we believe there is no high bias risk (Tables 2 and 3). Detailed assessment results are illustrated in Fig. 2.

Meta-analysis

IPF

ML model In the diagnosis of IPF, the combined C-index, sensitivity, and specificity of ML were 0.93 (95% CI 0.89–0.97), 0.79 (95% CI 0.73–0.83), and 0.84 (95% CI 0.79–0.88), respectively. Notably, the model with the highest diagnostic performance was the deep learning model (CT) proposed by Wenxi Yu [21], with a C-index of 0.99 (95% CI 0.97–1), as illustrated in Figs. 3 and 4.

Table 3 Risk of bias assessment results

No	First author	Year of publication	1. Case selection: Does the selection of cases produce bias?			2. The trial to be evaluated: Is there any bias in the conduct or interpretation of the trial to be evaluated?		3. Will there be any bias in the implementation and interpretation of the gold standard?		4. Is there any bias in the flow of the cases?			Evaluation of clinical practicability		
			Que1	Que2	Que3	Que1	Que2	Que1	Que2	Que1	Que2	Que3	Que1	Que2	Que3
1	Wenxi Yu	2022	Yes	No (ML)	Yes	Yes	No	Yes	Not clear	Yes	Yes	Yes	Yes	Yes	Yes
2	Wenxi Yu	2022	Yes	No (DL)	Yes	Yes	No	Yes	Not clear	Yes	Yes	Yes	Yes	Yes	Yes
3	Refaee, T	2022	Yes	No (DL)	Yes	Yes	No	Yes	Not clear	Yes	Yes	Yes	Yes	Yes	Yes
4	Hiroyuki Abe	2004	Yes	No (DL)	Yes	Yes	No	Yes	Not clear	Yes	Yes	Yes	Yes	Yes	Yes
5	Aya Fukushima	2004	Yes	No (DL)	Yes	Yes	No	Yes	Not clear	Yes	Yes	Yes	Yes	Yes	Yes
6	Walsh, S. L. F	2022	Yes	No (DL)	Yes	Yes	No	Yes	Not clear	Yes	Yes	Yes	Yes	Yes	Yes
7	Alex Bratt	2022	Yes	No (DL)	Yes	Yes	No	Yes	Not clear	Yes	Yes	Yes	Yes	Yes	Yes
8	Andreas Christe	2019	Yes	No (DL)	Yes	Yes	No	Yes	Not clear	Yes	Yes	Yes	Yes	Yes	Yes
9	Adrien Depeursinge	2015	Yes	No (DL)	Yes	Yes	No	Yes	Not clear	Yes	Yes	Yes	Yes	Yes	Yes
10	Chi Wan Koo	2022	Yes	No (DL)	Yes	Yes	No	Yes	Not clear	Yes	Yes	Yes	Yes	Yes	Yes
11	Hiram Shaish	2020	Yes	No (DL)	Yes	Yes	No	Yes	Not clear	Yes	Yes	Yes	Yes	Yes	Yes
12	Manoj V. Maddali	2023	Yes	No (DL)	Yes	Yes	No	Yes	Not clea	Yes	Yes	Yes	Yes	Yes	Yes
13	Jonathan H. Chung	2023	Yes	No (DL)	Yes	Yes	No	Yes	Not clea	Yes	Yes	Yes	Yes	Yes	Yes
14	Weijia Fan	2023	Yes	No (DL)	Yes	Yes	No	Yes	Not clea	Yes	Yes	Yes	Yes	Yes	Yes
15	Stephen M. Humphries	2024	Yes	No (DL)	Yes	Yes	No	Yes	Not clea	Yes	Yes	Yes	Yes	Yes	Yes
16	Marcello Chang	2023	Yes	No (DL)	Yes	Yes	No	Yes	Not clea	Yes	Yes	Yes	Yes	Yes	Yes

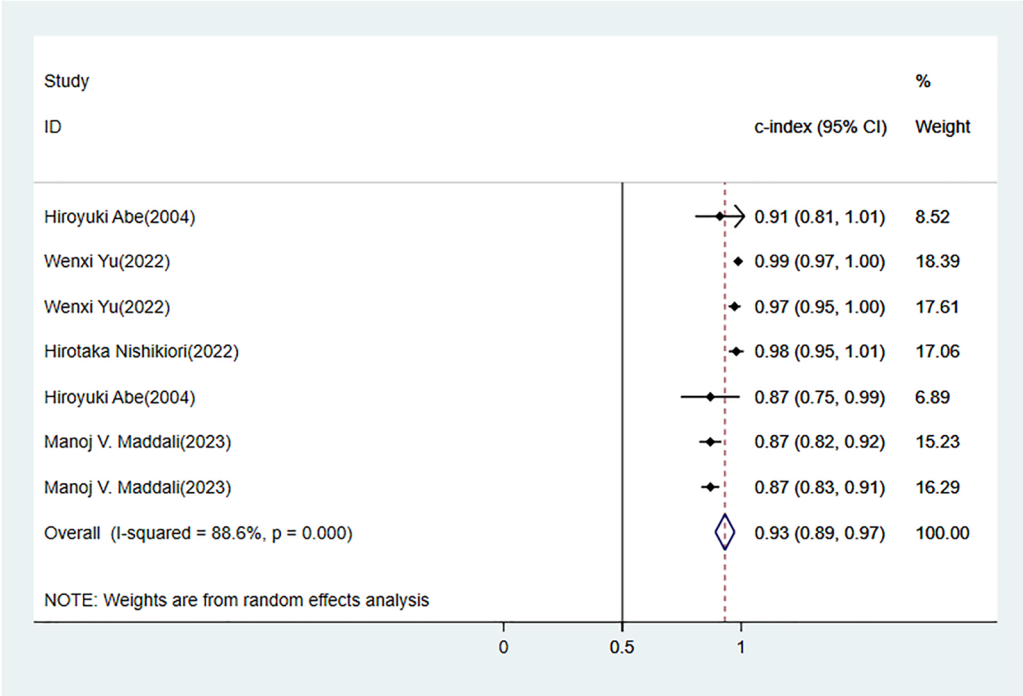

Fig. 3 Forest Map of Meta-Analysis of C-index for ML in Diagnosis of IPF

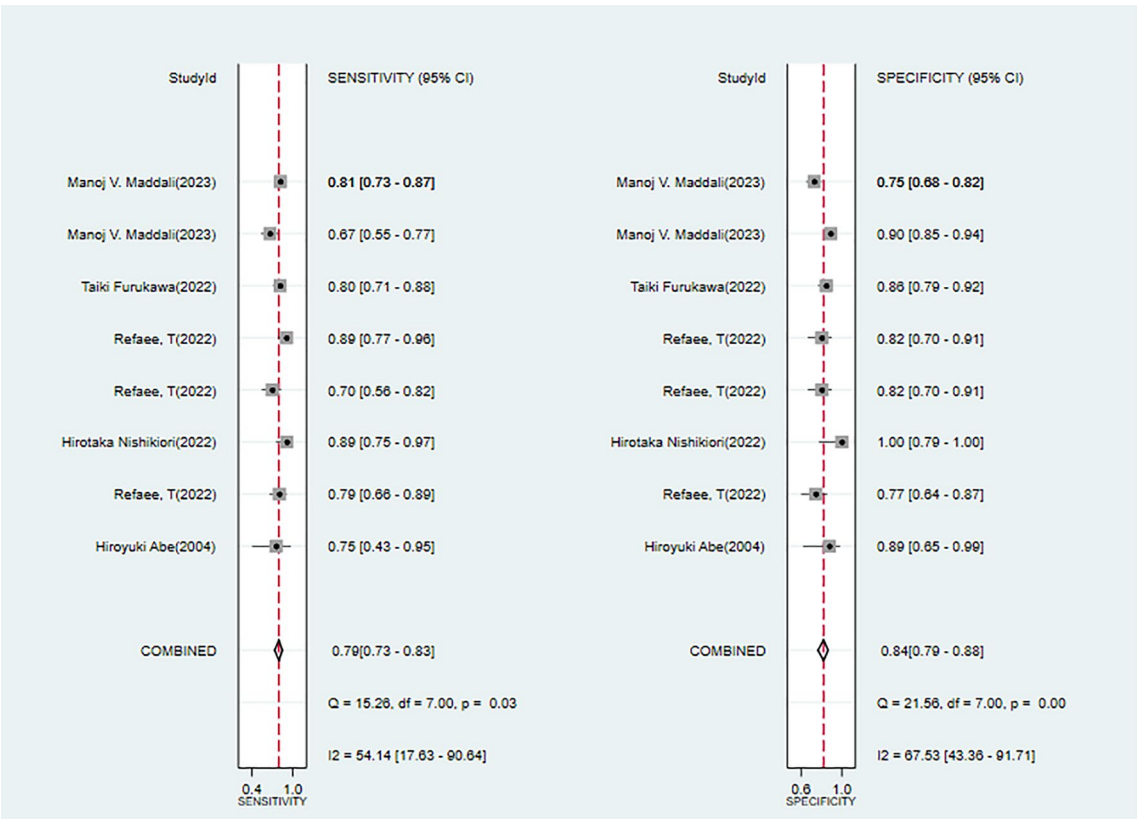

Fig. 4 Forest Map of Meta-analysis of the Sensitivity and Specificity of ML in Diagnosis of IPF

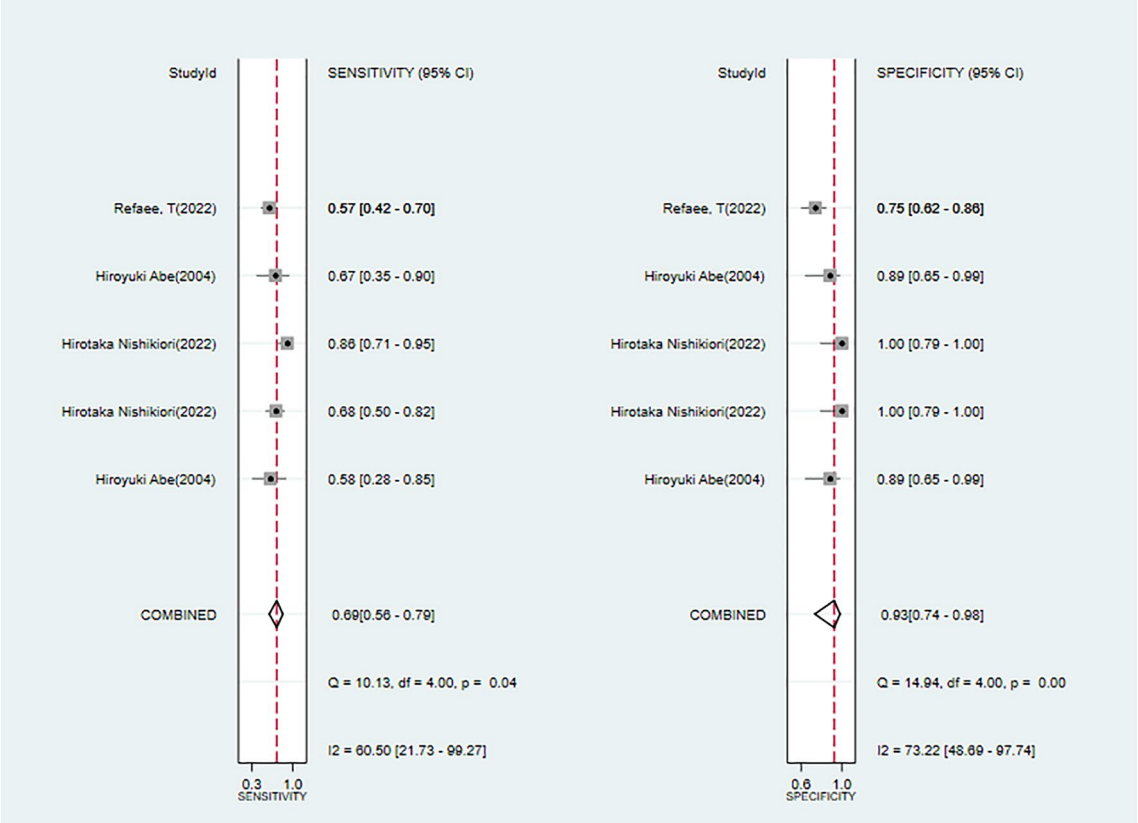

Fig. 5 Sensitivity and Specificity of Human Expert Diagnosis of IPF

Radiologists/clinicians Among the six included pieces of literature on IPF [18–22], three of them [20–22] compared ML with human experts. In these studies, the sensitivity of radiologists/clinicians in the diagnosis of IPF was notably lower compared to that of ML. However, it was essential to note that due to the small sample size, only observational analyses were performed. In these three pieces of literature, the sensitivity of radiologists/clinicians diagnosing without the assistance of ML models ranged from 57 to 85%, while the specificity ranged from 75 to 98% (Fig. 5).

UIP

ML model In the diagnosis of UIP, the combined C-index, sensitivity, and specificity of ML were 0.91 (95% CI 0.87–0.94), 0.92 (95% CI 0.80–0.97) and 0.92 (95% CI 0.82–0.97), respectively. Notably, Aya Fukushima's deep learning model exhibits the highest diagnostic performance, boasting a C-index of 0.99 (95% CI 0.97 ~ 1.02), as illustrated in Figs. 6 and 7A.

We utilized funnel plot analysis to investigate publication bias concerning machine learning diagnoses of UIP, and the findings from the funnel plot along with Egger's test reveal a significant publication bias within each validation set of the models ($P < 0.05$), as illustrated in Fig. 8.

Radiologists/clinicians Ten pieces of literature did not explore the joint use of doctors and ML models in the diagnosis of UIP. The diagnostic sensitivity of radiologists/clinicians based solely on HRCT was reported to be 0.69 (95% CI 0.50 ~ 0.84), which was comparable to the diagnostic sensitivity of ML models, as depicted in Fig. 7B.

Discussion

Summary of the main findings

We have observed that the primary source of IPF for ML remains HRCT, with only a limited number of studies utilizing chest X-rays. In this study, specifically, only one study employed chest X-rays as a variable [21], making direct comparisons with CT challenging. The ML methods employed include convolutional neural networks, deep learning, and random forest. In the diagnosis of IPF, ML demonstrates superior sensitivity and specificity compared to clinicians/radiologists. For the diagnosis of UIP, whether examining the UIP pattern on HRCT or the pathological UIP pattern, variables are consistently derived from HRCT. The primary ML methods include

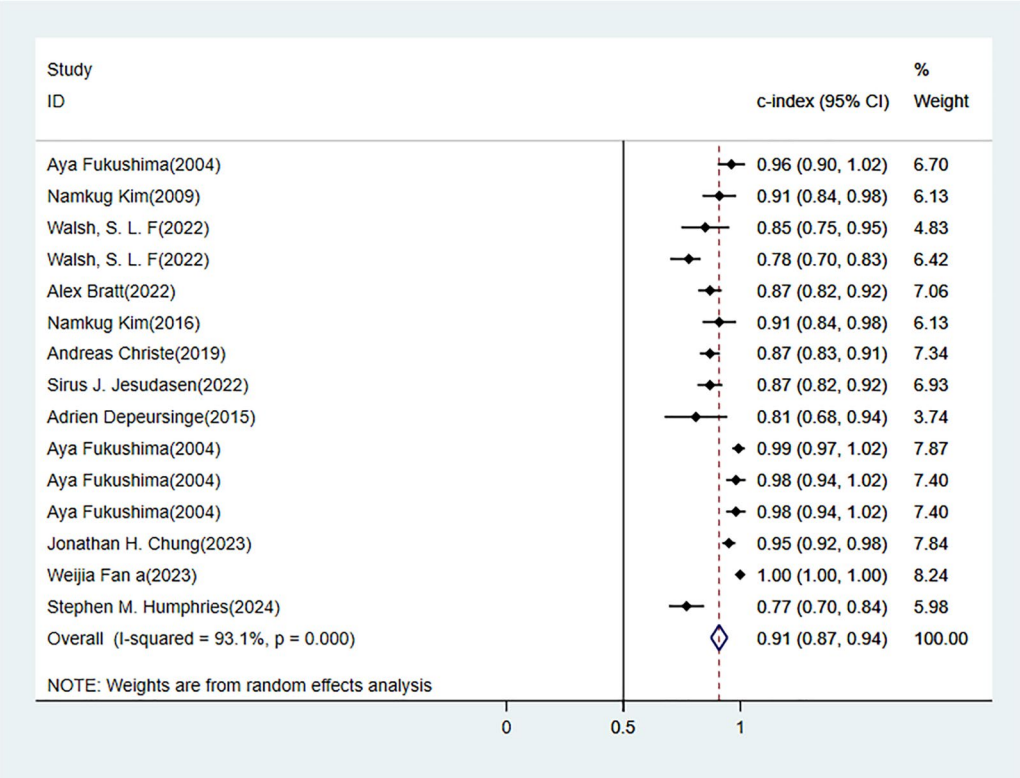

Fig. 6 Forest Map of Meta-Analysis of C-index for Machine Learning in Diagnosis of UIP

convolutional neural networks, and deep learning, with a smaller representation of extreme gradient boosting (XGboost). In UIP diagnosis, the diagnostic sensitivity and specificity of ML align closely with those of clinicians/radiologists.

Comparison with previous studies

IPF needs to be distinguished not only from various interstitial lung disease (ILD) but also from specific types of IIP. Existing research indicates that IPF comprises 47–71% of IIP cases, whereas non-specific Interstitial Pneumonia (NSIP) accounts for 13–30% of IIP cases [34, 35]. The subjects in five included studies are patients with ILD or diffuse lung lesions. The primary outcome indicators focus on the sensitivity, specificity, and diagnostic performance of ML in the detection of IPF among various ILDs. Among these studies, three [20–22] compared ML with human experts, revealing significantly lower diagnostic sensitivity of human experts compared to ML (Fig. 7B). The study by Raghu [12] found that, despite all 95 cases being confirmed as IPF by surgical lung biopsy, interstitial disease experts' sensitivity in diagnosing IPF based on CT features was 78.5%. Another study [36, 37] reported low sensitivity of human experts in the diagnosis of interstitial pneumonia or interstitial pulmonary fibrosis combined with emphysema. Therefore, HRCT-based deep learning is deemed meaningful for diagnosing IPF. Its pooled diagnostic sensitivity is higher than that of clinical/radiological experts, potentially reducing interobserver variability and shortening the time from suspicion to diagnosis. It proves advantageous in the detection of IPF among various ILDs.

Regarding the diagnosis of UIP, the IPF Guidelines published in 2011 emphasized the role of identifying HRCT manifestations, including UIP patterns, as one of the independent diagnostic criteria. Studies comparing pathology with HRCT confirmed the accuracy of HRCT diagnosis of classic UIP at 80–90%. However, in some early UIP diagnosis studies, the diagnostic specificity was relatively low, at 43–78%. This decrease in specificity is related to patients with no honeycombing or atypical features on CT, limiting radiologists' ability to diagnose UIP based solely on CT. Among the studies included in this analysis, the pooled sensitivity of ML for diagnosing UIP is 92% (95% CI 0.80 ~ 0.97), with a specificity of 92% (95% CI 0.82 ~ 0.97). The sensitivity and diagnostic performance of ML in distinguishing UIP from non-UIP are comparable to those of human

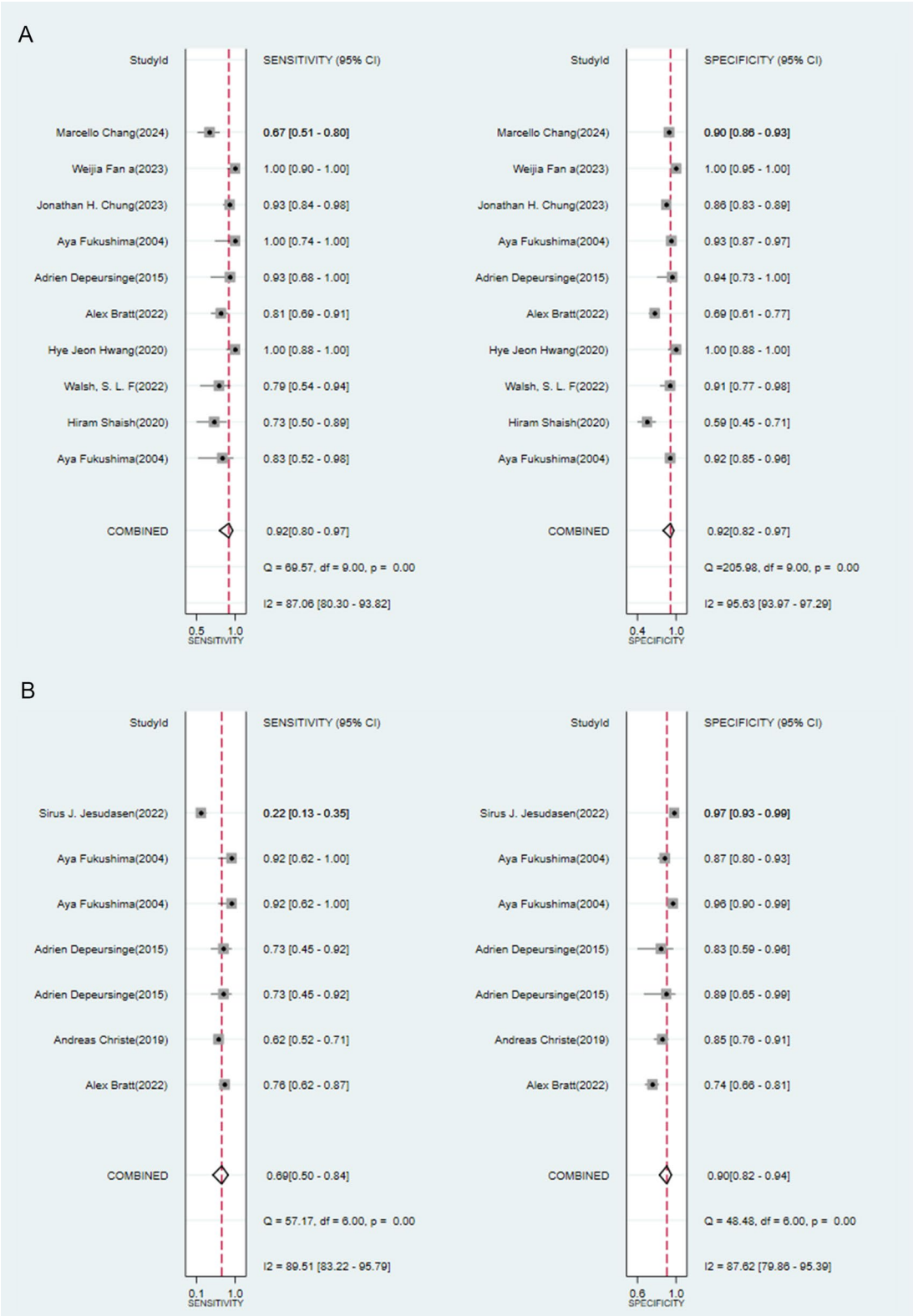

Fig. 7 Forest Map of Meta-Analysis of Sensitivity and Specificity for Machine Learning in Diagnosis of UIP. (**A** Meta-Analysis of the Sensitivity and Specificity of Machine Learning in Diagnosis of UIP. **B** Sensitivity and Specificity of Human Expert Diagnosis of UIP)

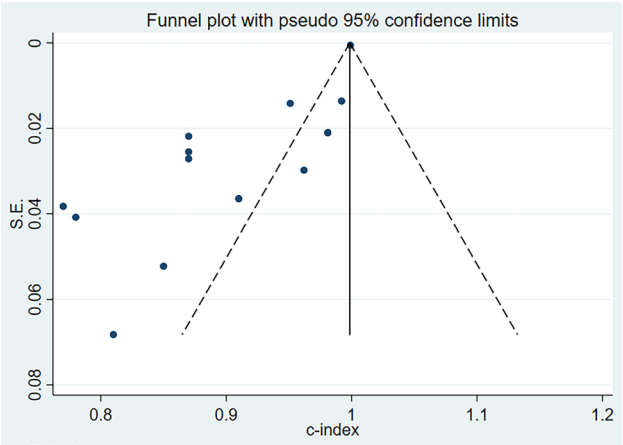

Fig. 8 Funnel plot of Meta-Analysis of C-index for Machine Learning in Diagnosis of UIP

experts. In the future, ML algorithms may have an advantage in diagnosing potential UIP and non-UIP patterns.

Types of ML

Common ML models are applicable to interpretable clinical features, and some researchers have conducted ML studies based on imageological data. However, in the implementation process, manual image region segmentation and texture feature extraction are required, which may introduce bias due to the influence of the observer's prior knowledge. Deep learning is primarily applied to the identification of medical images, diagnosis, and prognosis of diseases, demonstrating a higher diagnostic rate. In our study, we included studies on both ML and deep learning. Due to the limited number of included studies, we did not conduct separate meta-analyses for ML and deep learning, but only briefly reviewed them. Therefore, more studies are desired to explore whether ML methods constructed based on conventional interpretable clinical features and images will outperform deep learning. Another advantage of deep learning is its capacity to intelligently diagnose diseases, providing crucial insights for the development of intelligent tools.

Advantages and limitations of the study

The strength of our study lies in being the first systematic review of the diagnostic value of ML in IPF and UIP, providing a comprehensive summary of diagnostic results compared with clinical/radiological findings. However, certain limitations should be acknowledged. Firstly, although a comprehensive systematic search was conducted, the number of studies included remains relatively low, which calls for careful interpretation of the findings. Secondly, the sensitivity analysis indicated that the predictive performance was not significantly influenced by chest radiography, which was one of the variables alongside others from high-resolution CT. Thirdly, during the image acquisition process, the image parameters were not modified, and preliminary experiments were carried out with varying image parameters. Consequently, it was not feasible to eliminate the heterogeneity resulting from the transitional setup of the equipment. Lastly, high heterogeneity presents a considerable challenge in machine learning-based meta-analysis. Given the restricted number of included studies, we were unable to further investigate this concern.

Conclusions

This study systematically evaluated the diagnostic performance of machine learning (ML) based on high-resolution CT imaging for idiopathic pulmonary fibrosis (IPF) and usual interstitial pneumonia (UIP). The results demonstrate that ML models leverage their strengths in data processing and pattern recognition to achieve rapid and efficient disease classification with high sensitivity and specificity. This highlights the potential of AI-based imaging diagnostic tools in enhancing early disease screening efficiency and reducing human error.

From a clinical standpoint, machine learning models can proficiently tackle the observer variability that is inherent in conventional diagnostic methods. Specifically, they can improve both consistency and accuracy in multidisciplinary collaborative diagnoses. Furthermore, this study conducts a systematic comparison between machine learning and traditional clinical radiological diagnostic approaches, validating the capability of machine learning to enhance diagnostic efficiency for IPF and UIP. This could serve as a foundation for its incorporation into future clinical practice. Given these findings, prospective clinical applications might include the integration of machine learning algorithms into hospital imaging analysis systems, facilitating expedited and more precise early screening for pulmonary fibrosis, minimizing diagnostic cycles, and ultimately enhancing patient outcomes.

Nevertheless, the utilization of these technologies necessitates additional extensive models and multicenter clinical trials to further confirm their applicability and effectiveness in diverse clinical environments. Consequently, although the existing findings underscore the considerable promise of machine learning, its broad implementation in clinical practice still demands thoughtful advancement.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40001-025-02501-x.

Additional file 1

Acknowledgements
Not applicable.

Author contributions
LC: Data curation; Writing—original draft. YC: Data curation; Writing—original draft. XH: Formal analysis; Validation. Malika: Data curation; Formal analysis. CW: Data curation; Validation. LC: Conceptualization; Project administration; Supervision.

Funding
This work was supported by the Immunological mechanism of stem cell therapy for fibrotic hypersensitivity pneumonia (2022TSYCCX0036). The funding had no role in the study's design, conduct, or reporting and could neither approve nor disapprove the submitted manuscript.

Availability of data and materials
The data supporting the results in this study are available on request from the first author (Li Cong).

Declarations

Ethics approval and consent to participate
Institutional Review Board approval was not required because this study is based exclusively on published literature.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Respiratory and Critical Care Medical Center, People's Hospital of Xinjiang Uygur Autonomous Region, Urumqi 830000, China. [2]Radiographic Center, People's Hospital of Xinjiang Uygur Autonomous Region, Urumqi, China. [3]Department of Respiratory and Critical Care Medicine, The First Affiliated Hospital of Shihezi University, Shihezi, China.

Received: 17 December 2024 Accepted: 24 March 2025
Published online: 15 April 2025

References

1. Maher TM, Bendstrup E, Dron L, et al. Global incidence and prevalence of idiopathic pulmonary fibrosis. Respir Res. 2021;22:197. https://doi.org/10.1186/s12931-021-01791-z.
2. Riddell P, Kleinerova J, Eaton D, et al. Meaningful survival benefit for single lung transplantation in idiopathic pulmonary fibrosis patients over 65 years of age. Eur Respir J. 2020. https://doi.org/10.1183/13993003.02413-2019.
3. Majewski S, Królikowska M, Costabel U, et al. Double lung transplantation for idiopathic pulmonary fibrosis in a patient with a history of liver transplantation and prolonged journey for disease-specific antifibrotic therapy. Case Rep Pulmonol. 2022;2022:4054339. https://doi.org/10.1155/2022/4054339.
4. Hyldgaard C, Møller J, Bendstrup E. Changes in management of idiopathic pulmonary fibrosis: impact on disease severity and mortality. Eur Clin Respir J. 2020;7:1807682. https://doi.org/10.1080/20018525.2020.1807682.
5. Spencer LG, Loughenbury M, Chaudhuri N, et al. Idiopathic pulmonary fibrosis in the UK: analysis of the British thoracic society electronic registry between 2013 and 2019. ERJ Open Res. 2021. https://doi.org/10.1183/23120541.00187-2020.
6. Aono Y, Nakamura Y, Kono M, et al. Prognostic significance of forced vital capacity decline prior to and following antifibrotic therapy in idiopathic pulmonary fibrosis. Ther Adv Respir Dis. 2020. https://doi.org/10.1177/1753466620953783.
7. Raghu G, Collard HR, Egan JJ, et al. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. Am J Respir Crit Care Med. 2011;183:788–824. https://doi.org/10.1164/rccm.2009-040GL.
8. Widell J, Lidén M. Interobserver variability in high-resolution CT of the lungs. Eur J Radiol Open. 2020;7:100228. https://doi.org/10.1016/j.ejro.2020.100228.
9. Camp R, Smith ML, Larsen BT, et al. Reliability of histopathologic diagnosis of fibrotic interstitial lung disease: an international collaborative standardization project. BMC Pulm Med. 2021;21:184. https://doi.org/10.1186/s12890-021-01522-6.
10. Guiot J, Vaidyanathan A, Deprez L, et al. A review in radiomics: making personalized medicine a reality via routine imaging. Med Res Rev. 2022;42:426–40. https://doi.org/10.1002/med.21846.
11. Cho YH, Seo JB, Lee SM, et al. Quantitative CT imaging in chronic obstructive pulmonary disease: review of current status and future challenges. J Korean Soc Radiol. 2018;78:1–12. https://doi.org/10.3348/jksr.2018.78.1.1.
12. Tominaga J, Sakai F, Johkoh T, et al. Diagnostic certainty of idiopathic pulmonary fibrosis/usual interstitial pneumonia: the effect of the integrated clinico-radiological assessment. Eur J Radiol. 2015;84:2640–5. https://doi.org/10.1016/j.ejrad.2015.08.016.
13. Hochhegger B, Marchiori E, Zanon M, et al. Imaging in idiopathic pulmonary fibrosis: diagnosis and mimics. Clinics. 2019;74:e225. https://doi.org/10.6061/clinics/2019/e225.
14. Takahashi R, Kajikawa Y. Computer-aided diagnosis: a survey with bibliometric analysis. Int J Med Inform. 2017;101:58–67. https://doi.org/10.1016/j.ijmedinf.2017.02.004.
15. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism. 2017;69:S36-s40. https://doi.org/10.1016/j.metabol.2017.01.011.
16. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44. https://doi.org/10.1038/nature14539.
17. Debray TP, Damen JA, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Stat Methods Med Res. 2019. https://doi.org/10.1177/0962280218785504.
18. Yu W, Zhou H, Goldin JG, et al. End-to-end domain knowledge-assisted automatic diagnosis of idiopathic pulmonary fibrosis (IPF) using computed tomography (CT). Med Phys. 2021;48:2458–67. https://doi.org/10.1002/mp.14754.
19. Yu W, Zhou H, Choi Y, et al. Multi-scale, domain knowledge-guided attention + random forest: a two-stage deep learning-based multi-scale guided attention models to diagnose idiopathic pulmonary fibrosis from computed tomography images. Med Phys. 2023;50:894–905. https://doi.org/10.1002/mp.16053.
20. Refaee T, Salahuddin Z, Frix AN, et al. Diagnosis of idiopathic pulmonary fibrosis in high-resolution computed tomography scans using a combination of handcrafted radiomics and deep learning. Front Med. 2022;9:915243. https://doi.org/10.3389/fmed.2022.915243.
21. Abe H, Ashizawa K, Li F, et al. Artificial neural networks (ANNs) for differential diagnosis of interstitial lung disease: results of a simulation test with actual clinical cases. Acad Radiol. 2004;11:29–37. https://doi.org/10.1016/s1076-6332(03)00572-5.
22. Fukushima A, Ashizawa K, Yamaguchi T, et al. Application of an artificial neural network to high-resolution CT: usefulness in differential diagnosis of diffuse lung disease. AJR Am J Roentgenol. 2004;183:297–305. https://doi.org/10.2214/ajr.183.2.1830297.
23. Walsh SLF, Calandriello L, Silva M, et al. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. Lancet Respir Med. 2018;6:837–45. https://doi.org/10.1016/s2213-2600(18)30286-8.
24. Bratt A, Williams JM, Liu G, et al. Predicting usual interstitial pneumonia histopathology from chest CT imaging with deep learning. Chest. 2022;162:815–23. https://doi.org/10.1016/j.chest.2022.03.044.

25. Christe A, Peters AA, Drakopoulos D, et al. Computer-aided diagnosis of pulmonary fibrosis using deep learning and CT images. Invest Radiol. 2019;54:627–32. https://doi.org/10.1097/rli.0000000000000574.
26. Depeursinge A, Chin AS, Leung AN, et al. Automated classification of usual interstitial pneumonia using regional volumetric texture analysis in high-resolution computed tomography. Invest Radiol. 2015;50:261–7. https://doi.org/10.1097/rli.0000000000000127.
27. Koo CW, Williams JM, Liu G, et al. Quantitative CT and machine learning classification of fibrotic interstitial lung diseases. Eur Radiol. 2022;32:8152–61. https://doi.org/10.1007/s00330-022-08875-4.
28. Shaish H, Ahmed FS, Lederer D, et al. Deep learning of computed tomography virtual wedge resection for prediction of histologic usual interstitial pneumonitis. Ann Am Thorac Soc. 2021;18:51–9. https://doi.org/10.1513/AnnalsATS.202001-068OC.
29. Humphries SM, Thieke D, Baraghoshi D, et al. Deep learning classification of usual interstitial pneumonia predicts outcomes. Am J Respir Crit Care Med. 2024;209:1121–31. https://doi.org/10.1164/rccm.202307-1191OC.
30. Chang M, Reicher JJ, Kalra A, et al. Analysis of validation performance of a machine learning classifier in interstitial lung disease cases without definite or probable usual interstitial pneumonia pattern on CT using clinical and pathology-supported diagnostic labels. J Imaging Inform Med. 2024;37:297–307. https://doi.org/10.1007/s10278-023-00914-w.
31. Fan W, Chen Q, Maccarrone V, et al. Developing radiology diagnostic tools for pulmonary fibrosis using machine learning methods. Clin Imaging. 2024;106:110047. https://doi.org/10.1016/j.clinimag.2023.110047.
32. Chung JH, Chelala L, Pugashetti JV, et al. A deep learning-based radiomic classifier for usual interstitial pneumonia. Chest. 2024;165:371–80. https://doi.org/10.1016/j.chest.2023.10.012.
33. Maddali MV, Kalra A, Muelly M, et al. Development and validation of a CT-based deep learning algorithm to augment non-invasive diagnosis of idiopathic pulmonary fibrosis. Respir Med. 2023;219:107428. https://doi.org/10.1016/j.rmed.2023.107428.
34. Travis WD, Hunninghake G, King TE Jr, et al. Idiopathic nonspecific interstitial pneumonia: report of an American thoracic society project. Am J Respir Crit Care Med. 2008;177:1338–47. https://doi.org/10.1164/rccm.200611-1685OC.
35. Monaghan H, Wells AU, Colby TV, et al. Prognostic implications of histologic patterns in multiple surgical lung biopsies from patients with idiopathic interstitial pneumonias. Chest. 2004;125:522–6. https://doi.org/10.1378/chest.125.2.522.
36. Reddy TL, Tominaga M, Hansell DM, et al. Pleuroparenchymal fibroelastosis: a spectrum of histopathological and imaging phenotypes. Eur Respir J. 2012;40:377–85. https://doi.org/10.1183/09031936.00165111.
37. Kheir F, Uribe Becerra JP, Bissell B, et al. Use of a genomic classifier in patients with interstitial lung disease: a systematic review and meta-analysis. Ann Am Thorac Soc. 2022;19:827–32. https://doi.org/10.1513/AnnalsATS.202102-197OC.

Publisher's Note