Volume 3, Number 3, 2026

Pages 2174-2200

Bioscience Insights

ISSN: 3080-3489 (Print)
ISSN: 3080-3497 (Online)

BioSciInsights.com

DATA LEAKAGE AND REPORTED ACCURACY OF DEEP LEARNING FOR INTERSTITIAL LUNG DISEASE ON COMPUTED TOMOGRAPHY: A METHODOLOGICAL SYSTEMATIC REVIEW

Adyan Ahmad[*1], Arsalan Haleem[2], Asad Khan[3], Dr. Qurratulain Aamir[4], Ahmad Saeed[5]

*[*1,3,5]Department of Computer Science, Bacha Khan University Charsadda, Charsadda, Pakistan*
[2]Department of Electronic Engineering Government Post Graduate Jehanzeb College, University of Swat, Mingora, Pakistan
[4]OTHM Diploma in Health and Social Care Management, UK.

[*1]mradyanahmad@gmail.com, [2]arsalanhaleem444@gmail.com, [3]ak.assad01@gmail.com, [4]qurat.research@gmail.com, [5]ahmadsaeed@bkuc.edu.pk

https://doi.org/10.5281/zenodo.24530646

Article Information

Article History

Received:	September 03, 2026
Revised:	October 31, 2026
Accepted:	November 16, 2026
Available Online:	November 25, 2026

Keywords:
Lung Diseases, Interstitial; Deep Learning; Tomography, X-Ray Computed; Systematic Review; Data leakage

Copyright @Author
Corresponding Author: *
Adyan Ahmad

Abstract

Deep learning (DL) classifiers for interstitial lung disease (ILD) on computed tomography (CT) often report accuracies above 95%, but whether their test data were independent of model development is unclear. This methodological systematic review searched six sources and citation links to 28 October 2026 for studies using DL to classify ILD diagnoses or patterns on CT. Partitioning was graded from quoted source text and tests by whether they used independent patients. Of 35 included studies, 24 (69%) had a high risk of bias, 12 documented sub-patient partitioning and five were suspected of it. Median accuracy was 95.0% with documented sub-patient and 82.9% with patient-level partitioning (p = 0.018), and all six accuracies of 98% or more came from studies without patient-level partitioning. These were all pattern classifiers evaluated on patches, slices or images, so partitioning could not be separated from task and prediction unit. Eleven studies tested on independent patients, and performance fell in five of the seven with an internal comparison. On the Geneva database, accuracy rose from 85.5% to 99.8% while evaluation remained largely patch-level. One study released code; the median reproducibility score was 1 of 6. Patient-level partitioning and testing on independent patients should be minimum standards.

Graphical abstract caption:
Median reported accuracy of deep learning classifiers for interstitial lung disease on CT was 95% in studies that allocated images of the same patients to both training and test sets and 83% in studies that kept each patient in one set (19 of the 23 studies that reported overall accuracy; the remaining four had

suspected or unreported splits). The difference is an association: it is confounded by task and prediction unit

Abbreviations:
AUC, area under the receiver operating characteristic curve; CT, computed tomography; DL, deep learning; HRCT, high-resolution computed tomography; ILD, interstitial lung disease; IPF, idiopathic pulmonary fibrosis; IQR, interquartile range; NSIP, nonspecific interstitial pneumonia; PROBAST+AI, Prediction model Risk Of Bias Assessment Tool for artificial intelligence; UIP, usual interstitial pneumonia.

1 INTRODUCTION

Interstitial lung diseases (ILDs) are a heterogeneous group of parenchymal disorders with very different treatments and outcomes. Idiopathic pulmonary fibrosis (IPF) and other progressive fibrosing ILDs are treated with antifibrotic drugs, whereas many inflammatory ILDs respond to immunosuppression [59]. High-resolution computed tomography (HRCT) is central to diagnosis, and current guidelines accept a usual interstitial pneumonia (UIP) pattern on HRCT as sufficient for a diagnosis of IPF in the appropriate clinical setting [46, 58, 59]. The same guidelines specify how the examination should be acquired, with thin-section volumetric imaging, high-spatial-frequency reconstruction and, where indicated, prone and expiratory series [46, 58]. Agreement between observers on HRCT patterns is only moderate [46], and expert multidisciplinary review is not available everywhere, so automated classification has attracted sustained interest.

Deep learning (DL) has been applied to this task for about ten years, first to small CT patches drawn mostly from the public ILD database of the University Hospitals of Geneva [3, 18, 24], and later to whole scans and patient-level questions such as the presence of UIP or IPF [19, 51, 65]. One CT-based IPF classifier has received marketing authorisation in the United States [69], and many recent papers report accuracies above 95% [38, 66, 74]. When images of one patient fall on both sides of a training and test split, a network can recognise the patient, scanner or reconstruction rather than the disease; this is data leakage, and it makes the internal estimate optimistic without any gain in generalisation to new patients [39, 73]. When classes come from different sources, a network can learn the source [22, 87], and when test data guide model tuning, the estimate is optimistic for the same reason [39]. Performance usually falls on external validation [83], and code is rarely shared [52, 63].

These questions matter in clinical practice. Radiologists and radiographers are increasingly expected to help select, validate, implement and audit artificial intelligence (AI) tools, and the imaging team controls the acquisition and reconstruction parameters on which image-based models depend [32, 50]. Yet formal training in AI remains uncommon among radiologists and radiographers [27], and published accuracy figures are often the only evidence available when a tool is considered for local use.

Previous reviews of AI in ILD catalogued tasks and models, and one pooled the diagnostic accuracy of machine learning for IPF [14, 21, 67]. One noted that missing external validation could overstate accuracy [14], but, to our knowledge, none examined whether studies kept each patient's images out of the test set or measured how evaluation design relates to reported accuracy. These summaries therefore cannot separate genuine diagnostic skill from optimism introduced by the evaluation. In other imaging fields this optimism has been measured directly: splitting brain magnetic resonance images by slice rather than by subject raised accuracy by 29 to 55 percentage points [81], and similar inflation has been shown in optical coherence tomography and chest CT [4, 56, 70]. Appraisal tools such as PROBAST+AI and reporting guidance such as CLAIM 2024 and TRIPOD+AI describe the relevant safeguards [20, 53, 71], but do not show how often a field applies them or what happens to reported accuracy when it does not. This methodological review addressed that gap for DL classification of ILD on CT. It examined (i) how often studies keep patients, tuning and labels independent of the test data and test on independent patients; (ii) whether reported accuracy is associated with how the data were partitioned; (iii) whether progress on the shared Geneva benchmark reflects better models or looser evaluation; and (iv) whether the studies can be reproduced.

2 Methods

This was a methodological systematic review, a design that appraises how primary studies were conducted and reported rather than pooling their results. It was reported according to PRISMA 2020 and its literature search extension PRISMA-S [55, 62] (checklist in Online Resource 1), and the narrative synthesis followed the Synthesis Without Meta-analysis (SWiM) guideline [11]. No protocol was registered, because the review examines study methodology rather than a health outcome; to compensate, the eligibility criteria, appraisal domains and decision rules are reported in full, and the dataset and analysis code are provided.

2.1 Eligibility criteria

Original articles in peer-reviewed journals were eligible if they used a deep neural network on chest CT of human subjects, alone or as a feature extractor, to classify ILD diagnoses, ILD subtypes or radiological or tissue patterns of ILD, and reported a quantitative measure of classification performance. Only articles written in English were eligible, and no date limit was applied.

Studies that only separated ILD from normal lung or from non-ILD disease were excluded, because this detection task is easier and would distort accuracy comparisons. Studies of segmentation, quantification or prognosis without classification, classifiers without a deep network applied to the images, conference papers, preprints, reviews and reports without new results were also excluded, as were articles in journals not indexed in MEDLINE, Scopus or Web of Science, because peer review could not be confirmed.

2.2 Search strategy

PubMed/MEDLINE, Europe PMC, IEEE Xplore, ScienceDirect, OpenAlex and Semantic Scholar were searched without date or language limits and were last searched on 28 October 2026 (Table 1). OpenAlex, Semantic Scholar, IEEE Xplore and ScienceDirect were included because they index engineering and other journals that MEDLINE does not cover. Because OpenAlex ranks records by relevance, the first 100 records of the broadest query and the first 50 of each other query were screened. After full-text selection, the reference lists and citing articles of every included report were retrieved from OpenAlex, compared with the database records and screened in the same way, first by title and then by abstract. Five reports added in the final database update were citation-searched on 29 October 2026, and no searches were run after that date. The complete search string for every source and sub-query, with its date and record count, is given in Online Resource 2.

Table 1 Search strategies and records retrieved (databases last searched on 28 October 2026)

Source	Interface	Strategy	Records (unique)
PubMed/MEDLINE	NCBI PubMed; no filters	ILD-A: "Lung Diseases, Interstitial"[Mesh] OR "interstitial lung disease(s)"[tiab] OR "idiopathic pulmonary fibrosis"[tiab] OR "usual interstitial pneumonia"[tiab] OR "nonspecific interstitial pneumonia"[tiab] OR ILD[tiab]. ILD-B: "Pulmonary Fibrosis"[Mesh] OR "pulmonary fibrosis"[tiab] OR "fibrotic lung disease"[tiab] OR "interstitial lung abnormality/abnormalities"[tiab] OR "diffuse (parenchymal) lung disease"[tiab] OR "hypersensitivity pneumonitis"[tiab]. CT: "Tomography, X-Ray Computed"[Mesh] OR CT[tiab] OR HRCT[tiab] OR "computed tomography"[tiab]. DL-A: "Deep Learning"[Mesh] OR "deep learning"[tiab] OR "convolutional neural network(s)"[tiab] OR CNN[tiab]. DL-B: transformer(s)[tiab] OR "neural network(s)"[tiab] OR "artificial intelligence"[tiab] OR "machine learning"[tiab]. Run as (ILD-A or ILD-B) AND CT AND (DL-A or DL-B) in four sub-queries (254, 354, 150 and 195 records).	953 (495)
OpenAlex	Works API; journal articles; relevance ranked	Eight free-text queries, e.g. "deep learning interstitial lung disease classification CT" (100 screened), "convolutional neural network lung tissue pattern classification high-resolution CT interstitial lung disease", "idiopathic pulmonary fibrosis usual interstitial pneumonia deep learning CT diagnosis", "vision transformer interstitial lung disease HRCT classification", "hypersensitivity pneumonitis nonspecific interstitial pneumonia deep learning CT differentiation", "ILD database MedGIFT Geneva lung CT patches deep learning texture classification", "interstitial lung disease patterns deep neural network HRCT" and "pulmonary fibrosis CT image classification deep learning" (up to 50 screened each).	401 (295)
Semantic Scholar	Academic Graph bulk search; titles and abstracts; journal articles	The PubMed concepts as one Boolean query: (ILD terms, plus "lung tissue pattern" | "lung texture") + (DL terms) + (CT | HRCT | "computed tomography"). All records screened.	545 (542)
Europe PMC	REST API; all sources	TITLE_ABS fields: (ILD terms as for PubMed, plus "lung tissue pattern(s)" OR "lung texture" OR ILD) AND (CT OR HRCT OR "computed tomography") AND (DL terms as for PubMed).	416 (403)
IEEE Xplore	Command search; All Metadata	(ILD terms, plus honeycombing) AND (CT OR HRCT OR "computed tomography") AND ("deep learning" OR "convolutional neural network(s)" OR CNN OR transformer OR "neural network(s)").	184 (184)

Source	Interface	Strategy	Records (unique)
ScienceDirect	Title, abstract, keywords; research articles	Five queries combining ILD, tissue-pattern or specific ILD terms with DL terms and CT terms, e.g. ("interstitial lung disease" OR "pulmonary fibrosis" OR "usual interstitial pneumonia" OR "diffuse lung disease") AND ("deep learning" OR "convolutional neural network") AND (CT OR "computed tomography").	148 (93)
Citation searching	OpenAlex; reference lists and citing articles of all 36 included reports	Linked records compared with records already identified by the database searches. Thirty-one reports were searched on 28 October 2026, and the five reports added in the final database update on 29 October 2026.	1,913 (1,649)

Records (unique): records retrieved, with the number left after removal of duplicates within the source in parentheses; for citation searching, the number not already identified by the database searches. Terms written as "disease(s)" or "(parenchymal)" were entered as separate phrases. In Semantic Scholar syntax, | denotes OR and + denotes AND. The complete string for every source and sub-query is given in Online Resource 2.

2.3 Study selection

Records were combined, and duplicates were removed by digital object identifier, PubMed identifier and normalised title. Reviewers screened all titles and abstracts and then the full texts of potentially eligible reports. Screening decisions were agreed jointly, with disagreements resolved by discussion as they arose, rather than made independently in duplicate; a third reviewer, the supervising author, adjudicated disagreements that persisted and checked the reviewers' work. Full texts were obtained from publisher websites, PubMed Central and author or institutional repositories; reports that could not be obtained from these sources were classed as not retrieved. Reports describing the same model or cohort were linked and analysed as one study.

2.4 Data extraction

Reviewer extracted data into a predefined spreadsheet and the other checked every item against the full text, with the third reviewer resolving any differences. Risk-of-bias judgements were made in the same way. Extracted items were the journal, task and classes, data source and availability, sample size, unit of analysis, partitioning method, the way hyperparameters, stopping, features and thresholds were chosen, the strongest test reported, headline internal and independent results, uncertainty, calibration, comparators, and the reporting of code, random seeds and software versions. Tasks were grouped as pattern classification (patches, regions, slices or images assigned to tissue or radiological patterns) or diagnostic classification (a patient or scan assigned a diagnosis or a guideline CT category). Journals were grouped as engineering or clinical. The passage supporting every partitioning, tuning and test judgement was quoted from the source text and recorded with its location (Online Resource 3).

2.5 Risk of bias and reproducibility appraisal

Risk of bias was judged in three domains that apply the leakage-related signalling questions of the participants, data sources and analysis domains of PROBAST+AI [53], and the strength of the test was recorded as a fourth, descriptive domain. This focused adaptation, rather than the full tool, applies equally to patch-level pattern classifiers, which lack the patient-level outcome that the full tool assumes, so all 35 studies could be judged on the same terms. QUADAS-2, the tool most often used in reviews of diagnostic accuracy [78], was not used because it has no item on whether images of one patient reach both training and test sets or whether test data informed model development. For split independence (D1), partitioning was first classified from the methods text into four categories. It was patient-level

when all images of each patient were stated to stay in one partition or when separate patient cohorts, sites or periods were used. It was documented sub-patient when patches, regions, slices or images were stated to be allocated to partitions without regard to patient and the data contained several such units per patient, so that images of one patient could reach both training and test sets. It was suspected when the unit of allocation or the number of units per patient was not stated but the description indicated allocation below the patient level, and not reported when no information was given. Documented and suspected sub-patient partitioning were rated as high concern, not reported as unclear, and patient-level as low concern unless augmentation preceded the split. Tuning independence (D2) asked whether hyperparameters, stopping, features, thresholds and model choice were fixed without the test data. Test strength (D3) recorded the most independent test reported. Tests on independent patients (another site, period, scanner cohort or data source, with no patient in common with development) were separated from three weaker designs: tests that re-examined development patients under other acquisition or reconstruction settings, tests in a hospital whose images had also been used for adaptation, and internal tests. Only tests on independent patients were counted as independent. Label-source confounding (D4) asked whether classes came from different institutions, trials or protocols. Following the PROBAST+AI convention, overall risk of bias was high when any of D1, D2 or D4 was of high concern or no separate test set existed, low when all three were of low concern, and unclear otherwise. The decision rules, with examples, are given in Online Resource 4.

An evaluation rigour score (0 to 4) counted low concern in D1, D2 and D4 and a test on independent patients in D3. A reproducibility score (0 to 6) counted public code, public or controlled-access data, reported random seeds, reported software versions, reported uncertainty and comparators retrained on the same split. Both scores are unweighted descriptions of the literature, not grades of individual studies. GRADE and formal assessment of bias due to missing results were not applied, because the review question concerned how studies were evaluated rather than a pooled effect estimate.

2.6 Reports not retrieved in full text

Every report not obtained at first was sought again on 11 November 2026 through the open-access locations indexed by OpenAlex, which incorporate Unpaywall data, PubMed Central, publisher pages, the institutional repositories listed for each record and an arXiv title search; corresponding authors were not contacted. One report had been deposited in PubMed Central in the meantime and was appraised in full, and the outcome for every report is listed in Online Resource 3. Abstracts of the remaining reports were classed as likely eligible, likely ineligible or lacking information, and the task, data, stated design and headline result of likely eligible reports were extracted. A design was classed as patient-level or external when the abstract stated a patient-level split, patient-level cross-validation, non-overlapping patient cohorts or an external test, and as patch-level or unstated otherwise. Abstracts rarely describe partitioning in enough detail to separate documented from suspected sub-patient partitioning, so this classification is coarser than the full-text appraisal and will misclassify some designs. These reports were therefore kept out of the main synthesis and used only in a sensitivity analysis.

2.7 Data synthesis and statistical analysis

Results were not pooled because tasks, classes and metrics differed widely, and because pooling designs with and without leakage would have reproduced the bias under study. Headline accuracy was the overall accuracy given in the abstract or main results, using the primary internal estimate when several were given. Because tasks, prediction units and numbers of classes differed, accuracy was compared between partitioning categories overall and then within strata of task (pattern or diagnostic classification) and prediction unit (patch, region, slice or image, or patient or scan). Comparisons used two-sided Mann-Whitney U tests. The difference in medians and the rank-biserial correlation (r, from -1 to 1, with the sign following the direction of the difference) are reported with each test, so that conclusions do not rest on p values alone. Accuracies are quoted to the precision reported by each study. A multivariable model was considered but not fitted, because partitioning was almost completely confounded with task and prediction unit and the strata were too small to support adjustment. Associations were examined with Spearman correlation. Studies reporting the same metric on internal

data and on a stricter test were analysed as within-study pairs, grouped by the type of test. All comparisons are descriptive and are interpreted as associations rather than causal effects. Two-sided p values below 0.05 were taken as noteworthy rather than confirmatory, and no adjustment was made for multiple comparisons, because the analyses describe the field rather than test a single primary hypothesis. Analyses were performed in Python 3 with SciPy 1.18.1, and figures were drawn with Matplotlib 3.11.2. The dataset and analysis code are provided as Online Resources 3 and 5.

3 Results

3.1 Study selection

The six database searches returned 2,647 records, of which 1,086 remained after removal of duplicates and 1,003 were excluded on title and abstract (Fig. 1). Of 83 reports sought, 30 could not be retrieved initially; one was later obtained from PubMed Central, so 29 remained unretrieved. Of 54 reports assessed in full, 17 were excluded (Table 2), leaving 35 studies described in 37 reports, because two studies were each described in two linked articles [2, 13, 49, 69]. Citation searching linked 1,913 records, of which 1,649 had not been found by the database searches. After title screening, nine abstracts were assessed; one report was likely eligible [12] but could not be retrieved, so it was appraised with the other unretrieved reports and no further study was included.

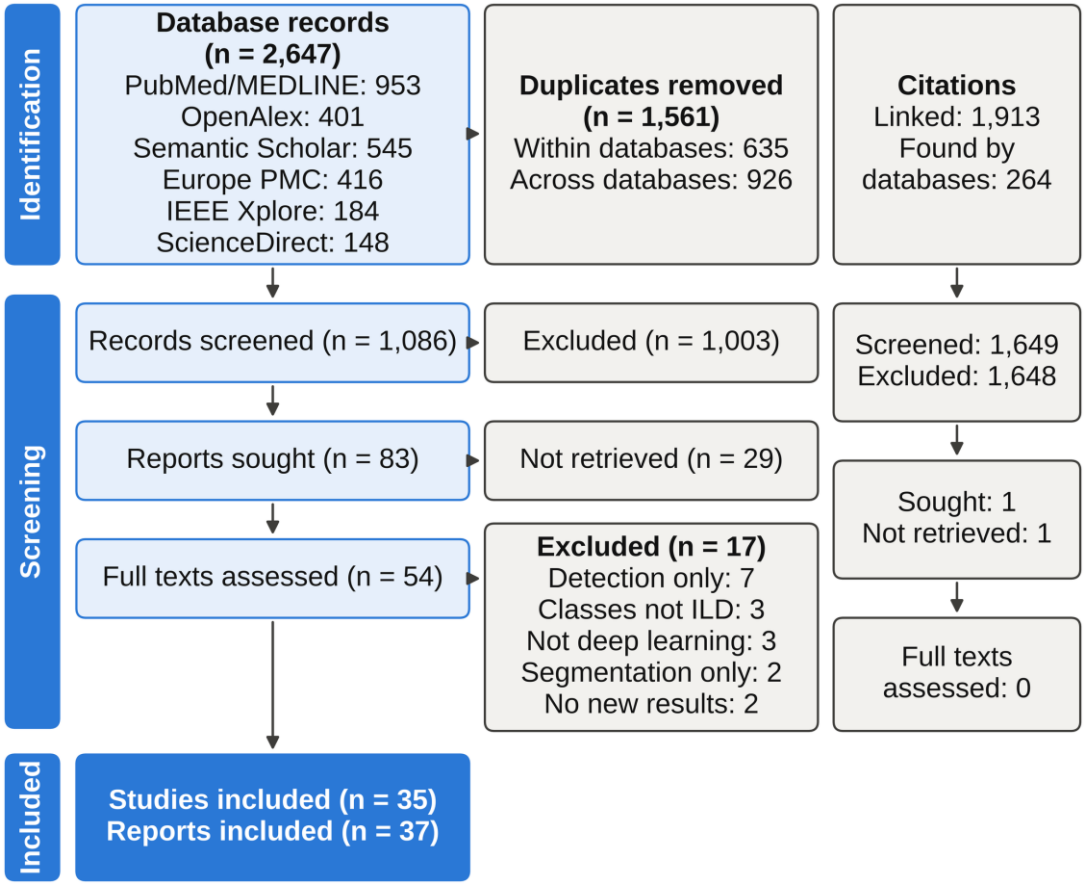

Fig. 1 PRISMA 2020 flow diagram of the search and selection process. Citations, records linked to the included reports through their reference lists and citing articles; found by databases, linked records already identified by the database searches. Of the 30 reports not retrieved initially from the databases, one was obtained later from PubMed Central and assessed in full. The 30 reports that remained unretrieved (29 from the databases and one from citation searching) were appraised from their abstracts (Table 5)

Table 2 Reports excluded after full-text assessment (n = 17)

Reason (n)	Reports (first author, year, journal)
Detection of ILD or fibrosis only (7)	Groot Lipman 2022 (Eur Radiol); Waseem Sabir 2023 (Front Med); Hassan 2023 (Processes); Mahapackialakshmi 2026 (Front Med); Mahapackialakshmi 2026 (Front Signal Process); Purushothaman 2026 (Premier J Sci); Peltekian 2026 (IEEE Trans Med Imaging)
Classes not ILD types (3)	Qiu 2024 (Diagnostics); Zhong 2024 (Front Med); Kumarganesh 2025 (Syst Soft Comput)
No deep network on CT images (3)	Sun 2024 (J Imaging Inform Med); Colligiani 2025 (Radiol Med); Mortani Barbosa 2026 (Curr Probl Diagn Radiol)
Segmentation only (2)	Yadlapalli 2022 (Scalable Comput Pract Exp); Esnaashari 2026 (J Res Med Sci)
No new classifier results (2)	Fermoyle 2026 (Respirology); Tambe 2026 (Diagnostics)

Database coverage shaped the evidence. Nine of the 35 studies, all in engineering journals, are not indexed in PubMed and were found only through the other sources [1, 26, 28, 31, 38, 60, 66, 68, 74]. These nine include the five highest reported accuracies. A further study indexed in PubMed was missed by the PubMed search, probably because its title used the plural form diffuse lung diseases [47]. A search of PubMed alone would therefore have missed 10 of the 35 included studies.

3.2 Study characteristics

The 35 studies were published from 2016 to 2026, 27 of them from 2021 onwards (Table 3). Twenty classified tissue or radiological patterns and 15 made diagnostic classifications, such as IPF against other ILDs or UIP against non-UIP. Fifteen appeared in engineering and 20 in clinical journals. Eighteen used only private data, eight combined public and private data, eight used only public data and one used controlled-access data. Twelve used the Geneva database: the nine pattern classifiers that reported overall accuracy on it are audited in Sect. 3.6, and the other three reported only F-measures or used it within a diagnostic task. The unit of analysis was a patch, region, slice or image in 19 studies and a patient or scan in 16.

Table 3 Characteristics and overall risk of bias of the 35 included studies, ordered by task and year

Study	Journal (type)	Classes	Data source; sample (availability)	Unit: partitioning	Partitioning category	Headline result	Risk of bias
Pattern classification							
Anthimopoulos 2016 [3]	IEEE Trans Med Imaging (E)	6 patterns and healthy	Geneva database and Bern; 120 scans (mixed)	Patch: Random patch split, 150 patches per class for test	Documented sub-patient	Accuracy 85.5%, mean F1 0.855	High
Christodoulidis 2017 [18]	IEEE J Biomed Health Inform (E)	6 patterns and healthy	Geneva database and Bern; 135 scans (mixed)	Patch: Random patch split	Documented sub-patient	Mean F1 0.88	High
Bae 2018 [6]	Sci Rep (C)	5 patterns and normal	One institution; 106 patients (private)	ROI: ROI-level 80/20 per class	Documented sub-patient	Accuracy 89.5%	High

Study	Journal (type)	Classes	Data source; sample (availability)	Unit: partitioning	Partitioning category	Headline result	Risk of bias
Gao 2018 [30]	Comput Methods Biomech Biomed Eng Imaging Vis (E)	6 patterns	Geneva database; 120 patients (public)	Patch, slice: Patient-level hold-out, 20 patients	Patient-level	Accuracy 87.9% (patch), 68.6% (slice)	Unclear
Kim 2018 [40]	J Digit Imaging (C)	5 patterns and normal	One institution, two scanners; 1,200 ROIs (private)	ROI: ROI-level cross-validation	Documented sub-patient	Accuracy 96.1% within, 86.1% across scanners	High
Bermejo-Peláez 2020 [7]	Sci Rep (C)	8 ILA and emphysema patterns	COPDGene; 208 scans (controlled)	Patch: Sample-level hold-out; 5-patient subject test	Documented sub-patient	Sensitivity 91.4%, specificity 98.2%	High
Raju 2020 [60][a]	Int J Electr Comput Eng (E)	17 categories, multi-label	Geneva-type series plus web images; 3,045 images (mixed)	Image: Image-level 80/20	Documented sub-patient	Accuracy 95%	High
Agarwala 2021 [1][a]	Pattern Recognit Image Anal (E)	5 patterns and normal (localised)	Geneva database; 108 series; 944 slices (public)	Slice: Five-fold CV over ROIs or slices; unit and patient grouping not stated	Suspected	F-score 0.55 to 0.86 by pattern	High
Doddavarapu 2021 [26][a]	IET Image Process (E)	5 patterns including normal	Geneva database; 1,458 ROIs; 24,105 patches (public)	Patch: Patch-level fourfold CV with overlapping patches	Documented sub-patient	Accuracy 93.33%, F-score 93.32%	High
Mabu 2021 [47]	Int J Comput Assist Radiol Surg (E)	5 opacity patterns and normal	Two hospitals, two scanners; 1,139 CT images (private)	ROI: Image-level split; target-hospital images used for adaptation	Suspected	F-measure 0.82 within hospital; 0.73 and 0.75 across hospitals	High
Junayed 2022 [38][a]	Turk J Electr Eng Comput Sci (E)	4 patterns and healthy	Geneva database; 128 patients (reported) (public)	Patch: Random patch split, 150 per class for test	Documented sub-patient	Accuracy 99.09%	High
Pawar 2022 [57]	Biomed Res Int (C)	6 patterns	Geneva database; 108 scans (public)	Slice: 30 random image draws	Suspected	Accuracy about 90%	High

Study	Journal (type)	Classes	Data source; sample (availability)	Unit: partitioning	Partitioning category	Headline result	Risk of bias
Sundaramoorthy 2023 [68][a]	Int J Adv Comput Sci Appl (E)	4 patterns and healthy	Geneva database; 108 series (public)	ROI: Not reported	Not reported	Accuracy 99.8%	High
Kiziloğlu 2024 [41]	Curr Med Imaging (C)	UIP, NSIP and normal patterns	One institution, one scanner; 43 patients; 7,150 images (private)	Image: Image-level split 70/15/15	Documented sub-patient	Accuracy 98.05% (VGG19), 95.02% (VGG16)	High
Vinta 2024 [74][a]	IEEE Access (E)	5 patterns and normal	Geneva database; 108 series (public)	Slice: Slice-level 70/30; loosely described 100-patient clinical set	Suspected	Accuracy 99.10%	High
Sithiq 2025 [66][a]	Comput Mater Contin (E)	Honeycombing, GGO, normal	One institution; 2,400 images (private)	Image: Image-level 70:30, augmentation before split possible	Suspected	Accuracy 99.88%	High
Xu 2025 [79]	Comput Biol Med (E)	UIP, NSIP, OP and normal (slices)	Four hospitals; 180 patients; 27,108 slices (private)	Slice: Random slice split 8:1:1	Documented sub-patient	Accuracy 92.84%	High
Zhang J 2025 [88]	Sci Rep (C)	5 IIM-ILD patterns	One institution; 629 patients (private)	Patient: Random 7:3; temporal 112 patients	Patient-level	AUC 0.885 internal, 0.835 temporal	High
Zhang Q 2025 [89]	IEEE J Biomed Health Inform (E)	6 DLD textures and normal	One institution; 151 patients; 85,603 patches (private)	Patch: Patient-level split 3:1:1	Patient-level	Accuracy 95.65%	Unclear
Godavarthi 2026 [31][a]	Eng Technol Appl Sci Res (E)	4 patterns and healthy	Geneva (Talisman) database; 18,793 patches (public)	Patch: Random patch split 70/15/15	Documented sub-patient	Accuracy 98.09%	High
Diagnostic classification							

Study	Journal (type)	Classes	Data source; sample (availability)	Unit: partitioning	Partitioning category	Headline result	Risk of bias
Christe 2019 [17]	Invest Radiol (C)	4 UIP CT categories	LTRC, Geneva, Bern; 105 patients (mixed)	Patient: Patient-level 5-fold CV (diagnosis stage)	Patient-level	Accuracy 0.56 (4 classes)	High
Shaish 2021 [65]	Ann Am Thorac Soc (C)	Histologic UIP vs. not	One institution; 301 patients (private)	Patient: Subject-level 5-fold CV, later-period hold-out	Patient-level	CV AUC 0.74; temporal sensitivity 74%, specificity 58%	Low
Yu 2021 [86]	Med Phys (C)	IPF vs. non-IPF	Five studies; 1,089 patients (private)	Patient: Patient-level 5-fold CV; leave-one-study-out	Patient-level	Accuracy above 0.95; 25 to 26 point drop in 2 of 8 held-out studies	High
Refaee 2022 [61]	Front Med (C)	IPF vs. non-IPF	Liege and LTRC; 474 patients (mixed)	Patient: 5-fold CV; external 109 patients	Patient-level	DL AUC 0.823 external	Unclear
Maddali 2023; Chang 2024 [13, 49]	Respir Med (C)	IPF vs. other ILD	Multicentre registries and trials; public and proprietary; >2,000 development; 295 tuning; 295 external (mixed)	Patient: Separate development and tuning cohorts; locked model tested on an external registry cohort (295)	Patient-level	AUC 0.87 (tuning); c-statistic 0.87 (external)	High
Mei 2023 [51]	Nat Commun (C)	5 ILD diagnoses	Registry, several hospitals; 449 patients (private)	Patient: Site-based external, 128 patients	Patient-level	AUC 0.74 to 0.85 by class	Low
Yu 2023 [85]	Med Phys (C)	IPF vs. non-IPF	Multicentre trials; 878 patients (private)	Patient: Stratified random subject split 80/20	Patient-level	AUC 0.975 to 0.987	High
Chung 2024 [19]	Chest (C)	UIP vs. non-UIP	Seven sources; 2,907 scans (mixed)	Scan: 70/15/15; external clinical cohort 245	Patient-level	AUC 0.951 internal	Unclear

Study	Journal (type)	Classes	Data source; sample (availability)	Unit: partitioning	Partitioning category	Headline result	Risk of bias
Fontanellaz 2024 [28][a]	IEEE Access (E)	4 UIP CT categories; UIP vs. non-UIP	LTRC and Bern; 105 test cases (mixed)	Patient: Leave-one-out on 105 cases	Patient-level	Accuracy 77.2% (2 classes), 58.1% (4 classes)	High
Huang 2024 [34]	BMJ Open Respir Res (C)	AE-IPF, stable IPF, controls	Two hospitals; 306 patients (private)	Patient: Patient-level internal split; later-period test sets including second hospital	Patient-level	AUC 0.96 and 0.92 for AE-IPF in two test sets	Low
Lami 2025 [45]	Respirology (C)	UIP vs. non-UIP	One institution; 93 patients (CT model) (private)	Patient: Random case split 46/47	Patient-level	CT-only accuracy 78.5%, AUC 0.849	Unclear
Yu 2025 [84]	Med Phys (C)	IPF vs. non-IPF (robustness)	One institution; 343 patients (private)	Scan series: Same patients under other acquisitions	Documented sub-patient	Specificity fell to 0.84 to 0.94	High
Kiziloğlu 2026 [43]	Curr Med Imaging (C)	IPF vs. non-IPF with UIP pattern	Single centre; 96 patients (private)	Patient: Patient-level stratified 70/15/15 (seed 42); majority vote over slices	Patient-level	Accuracy 87.2%, AUC 0.91 (30 test patients)	Low
Savadjiev 2026 [64]	Eur J Radiol Open (C)	UIP, 3 groups	Two institutions; 234 patients (private)	Patient: Train site 1, test site 2 (76 patients)	Patient-level	Macro recall 0.83	Unclear
Tahamtani-Omran 2026 [2, 69]	Diagnostics (C)	IPF vs. non-IPF (locked model)	Two centres; 300 patients (private)	Patient: Independent cohort (linked report: Ahmad 2024)	Patient-level	Sensitivity 41.0%, specificity 86.6% (v1)	Low

[a] Not indexed in PubMed. E, engineering journal; C, clinical journal. Availability: public, private, mixed (public and private) or controlled access. Partitioning category: patient-level; documented sub-patient (units below the patient allocated without regard to patient); suspected; or not reported (Online Resource 4). Domain judgements and the strongest test of each study are shown in Fig. 2; the supporting passages are in Online Resource 3. Headline results are given as reported and are not comparable across tasks with different numbers of classes. AE-IPF, acute exacerbation of IPF; CV, cross-validation; DLD, diffuse lung disease; GGO, ground-glass opacity; IIM-ILD, ILD associated with idiopathic inflammatory myopathy; ILA, interstitial lung abnormality; LTRC, Lung Tissue Research

Consortium; OP, organising pneumonia; ROI, region of interest. Other abbreviations are defined in the list after the abstract.

3.3 Risk of bias and evaluation design

Overall risk of bias was high in 24 studies (69%), unclear in six and low in five (Fig. 2). Eighteen of the 20 pattern classifiers had a high risk of bias, compared with six of the 15 diagnostic studies. Partitioning was patient-level in 17 studies. Twelve documented sub-patient partitioning: in 11 pattern classifiers, patches, regions, slices or images were allocated at random although each patient contributed many of them, and in one diagnostic study the evaluation series came from the same patients whose other series had been used to build the models [84]. Sub-patient partitioning was suspected in five pattern classifiers, which divided images or slices without stating the number of patients or grouped them only by image [1, 47, 57, 66, 74], and partitioning was not reported in one [68]. Only three pattern classifiers kept each patient within one partition [30, 88, 89]. In two studies the reported image counts indicate that augmentation preceded the split [1, 66].

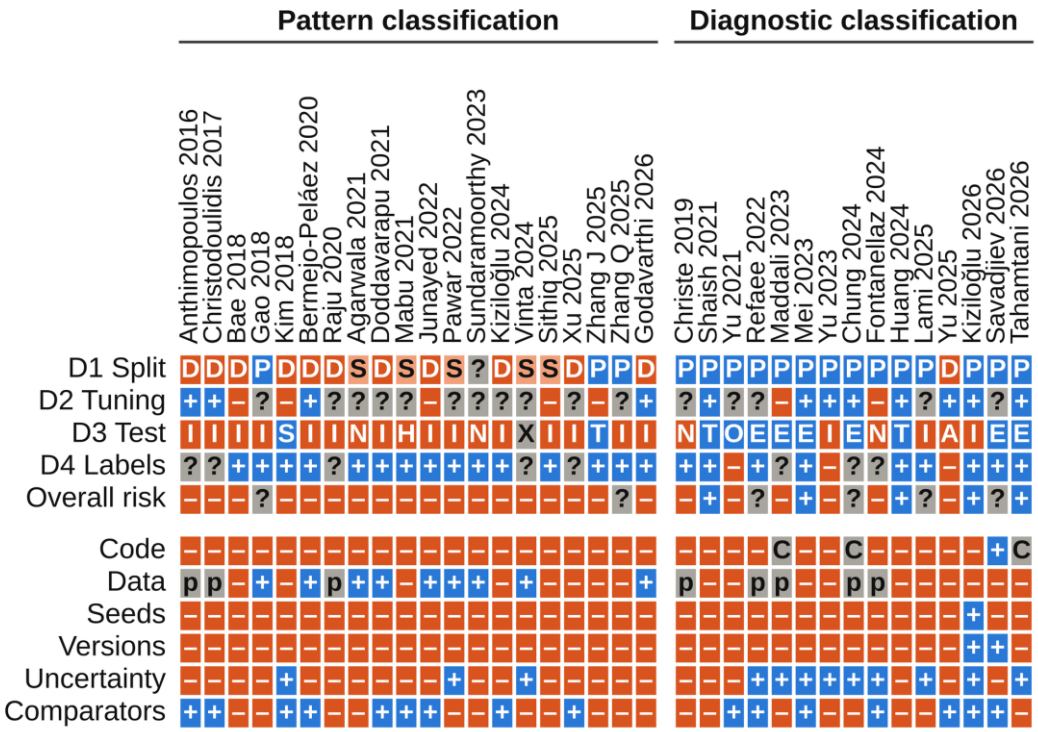

Fig. 2 Risk of bias domains and reproducibility items for each included study, grouped by task and ordered by year. D1, partitioning (P, patient-level; D, documented sub-patient; S, suspected sub-patient; ?, not reported); D2, tuning independence; D3, strongest test (E, other site; T, later period; O, data source held out; S, other scanner cohort; these four are tests on independent patients; A, same patients under other acquisitions; H, second hospital whose images were used for adaptation; X, external set insufficiently described; I, internal; N, none); D4, label-source confounding; Overall risk, overall risk of bias (high if D1, D2 or D4 was of high concern or no separate test set existed, low if all three were of low concern, unclear otherwise). Blue, low concern, test on independent patients or item reported; orange, high concern, other test or item not reported; light orange, suspected; grey, unclear or not reported, partly public data or commercial code. Symbols: +, low concern or reported; dash, high concern or not reported; ?, unclear; p, partly public data; C, commercial code

Tuning independence was of low concern in 12 studies, unclear in 16 and of high concern in seven. High concern arose when test data guided early stopping, model choice or search settings [6, 38, 40, 66, 88], when features were selected on the cases used for evaluation [28], and when one cohort both tuned

a commercial classifier and provided its reported performance [13, 49]. Label-source confounding was of high concern in three studies from one group whose IPF and non-IPF cases came largely from different trials [84–86]. When each of the eight source studies was held out in turn, accuracy fell by 1 to 4 points for six of them but by 25 to 26 points for two [86].

3.4 Performance on independent data

Eleven studies (31%) tested their models on independent patients from another site, period, scanner cohort or data source [19, 34, 40, 49, 51, 61, 64, 65, 69, 86, 88]. Three others reported a shifted test that did not meet this definition: one re-examined development patients under other acquisition settings [84], one tested in a second hospital whose images had also been used to adapt the model [47], and one described an external clinical set too briefly to judge [74]. Four had no separate test set [1, 17, 28, 68] and 17 reported only an internal test. Testing on independent patients was common in diagnostic studies (nine of 15) and clinical journals (11 of 20) but rare in pattern studies (two of 20) and absent from engineering journals. It was reported by one of eight studies from 2016 to 2020, five of 12 from 2021 to 2023 and five of 15 from 2024 to 2026.

Nine studies reported the same metric on internal data and on a stricter test (Fig. 3). Seven tested on independent patients. The value fell in five, was unchanged in one (area under the receiver operating characteristic curve [AUC] 0.87 in both cohorts) [13, 49] and changed in both directions in one, in which the AUC fell from 0.85 to 0.823 while accuracy rose from 77.9% to 85.3% [61]. Accuracy fell from 96.1% to 86.1% in a second scanner cohort [40]. Specificity for histological UIP fell from 66% to 58% in later patients [65], and sensitivity and specificity for UIP fell from 93% and 86% to 81% and 77% in an external cohort [19]. The AUC for ILD patterns associated with idiopathic inflammatory myopathy fell from 0.885 to 0.835 in later patients [88]. At its locked threshold, the sensitivity of the commercial IPF classifier fell from 67% in development data to 41.0% in an independent cohort of 300 patients scanned on four manufacturers' equipment [69]. When development patients were re-examined under other acquisition and reconstruction settings, specificity fell from 99% to 90% and from 97% to 84% [84], which measures robustness to acquisition rather than generalisation to new patients. In the ninth study, a pattern classifier, sensitivity was 91.4% with a sample-level test and 88.9% when five unseen subjects were tested [7], the only direct comparison of partitioning levels within a study. A further pattern classifier fell from an F-measure of 0.82 to 0.73 and 0.75 in a second hospital despite adaptation [47].

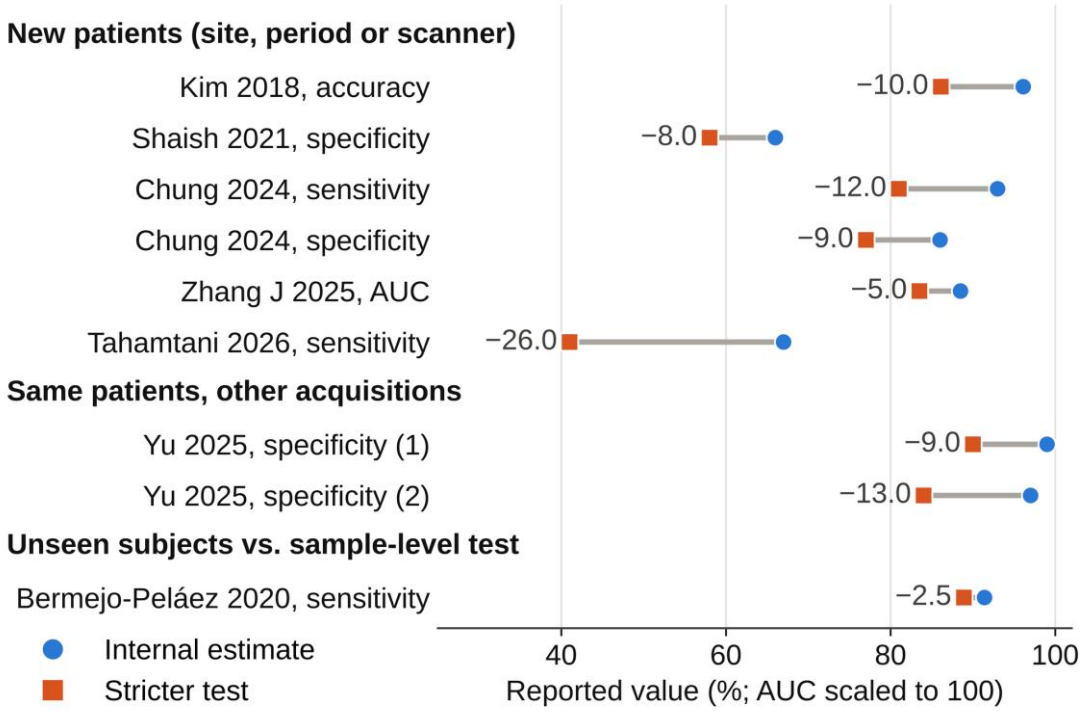

Fig. 3 Internal estimates (circles) and estimates on stricter tests (squares) reported within the same study, grouped by the type of test. Numbers give the difference in percentage points. Two further studies that tested on independent patients are not shown: one reported the same AUC in both cohorts, and in the other the AUC fell slightly while accuracy rose

3.5 Reported accuracy by partitioning, task and prediction unit

Rigour scores were 0 in three studies, 1 in 13, 2 in 10, 3 in five and 4 in four [34, 51, 65, 69]. Engineering journals had a median of 1 (interquartile range [IQR] 1 to 1) compared with 2 (IQR 2 to 3) in clinical journals (Mann-Whitney U = 40.5, $p < 0.001$; r = -0.73), and rigour showed no clear change with publication year (Spearman correlation 0.26, $p = 0.13$). Twenty-three studies reported overall accuracy (Table 4, Fig. 4a). The median was 82.9% (range 56.0% to 95.65%, n = 10) with patient-level partitioning and 95.0% (85.5% to 99.09%, n = 9) with documented sub-patient partitioning, a difference of 12.1 percentage points (U = 74.5, $p = 0.018$; r = 0.66). It was 99.1% (90.0% to 99.88%, n = 3) when sub-patient partitioning was suspected and 99.8% in the study that did not report partitioning. All six accuracies of 98% or more came from studies without patient-level partitioning [31, 38, 41, 66, 68, 74]. Studies not indexed in PubMed reported a median accuracy of 98.6% (n = 8), compared with 89.5% (n = 15) for indexed studies (U = 99.5, $p = 0.012$; r = 0.66).

Table 4 Headline accuracy (%) by partitioning, overall and within strata of task and prediction unit

Stratum	Patient-level: median (range), n	Documented sub-patient: median (range), n	Any sub-patient or not reported: median (range), n	Difference in medians (any sub-patient or not reported minus patient-level); test; effect size
All studies	82.9 (56 to 95.65), 10	95.0 (85.5 to 99.09), 9	96.1 (85.5 to 99.88), 13	13.2 points; U = 111.5, p = 0.004; r = 0.72
Pattern classification	87.9 (72.4 to 95.65), 3	95.0 (85.5 to 99.09), 9	96.1 (85.5 to 99.88), 13	8.2 points; U = 32.0, p = 0.111; r = 0.64
Diagnostic classification	78.5 (56 to 95), 7	None	None	Not estimable
Patch, region, slice or image units	91.8 (87.9 to 95.65), 2	95.0 (85.5 to 99.09), 9	96.1 (85.5 to 99.88), 13	4.3 points; U = 19.0, p = 0.381; r = 0.46
Patient or scan units	77.9 (56 to 95), 8	None	None	Not estimable

Any sub-patient or not reported combines documented (n = 9), suspected (n = 3) and not reported (n = 1) partitioning. For documented sub-patient against patient-level partitioning overall, the difference in medians was 12.1 points (U = 74.5, p = 0.018; r = 0.66). r, rank-biserial correlation, from -1 to 1; positive values indicate higher accuracy without patient-level partitioning. Comparisons are two-sided Mann-Whitney U tests and describe associations; strata without studies in both groups could not be compared. Group summaries of risk of bias, rigour and reproducibility by task, journal type and database coverage are given in Online Resource 3.

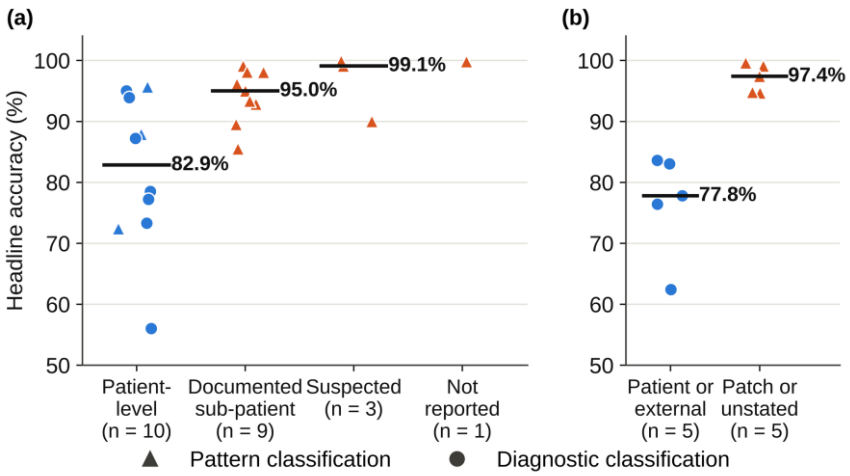

Fig. 4 Headline accuracy by partitioning. (a) The 23 included studies that reported overall accuracy, by partitioning category; triangles, pattern classification; circles, diagnostic classification. Every study without patient-level partitioning was a pattern classifier, so task and partitioning are confounded (Table 4). (b) The ten reports appraised from their abstracts that gave an overall accuracy. Bars show medians for groups of two or more

These comparisons are confounded. All 13 studies without patient-level partitioning that reported accuracy were pattern classifiers evaluated on patches, regions, slices or images, whereas seven of the ten with patient-level partitioning were diagnostic studies and eight reported results per patient or scan. Within pattern classification, the median was 96.1% without patient-level partitioning (n = 13) and

87.9% with it (n = 3; U = 32.0, p = 0.11; r = 0.64). Among studies evaluated on sub-patient units, the corresponding medians were 96.1% (n = 13) and 91.8% (n = 2; U = 19.0, p = 0.38; r = 0.46). No study without patient-level partitioning reported diagnostic or patient-level results, so the other strata could not be compared and adjustment in a multivariable model was not possible. The data therefore show that very high accuracies are concentrated in pattern classifiers evaluated below the patient level without patient-level partitioning, but they cannot separate the contribution of partitioning from that of task and prediction unit.

3.6 Audit of the Geneva benchmark

Nine pattern classifiers reported accuracy on the Geneva database (Fig. 5). Accuracy rose with publication year (Spearman correlation 0.79, p = 0.012), from 85.5% in 2016 [3] to 98.09%, 99.09%, 99.1% and 99.8% in four studies published from 2022 to 2026 [31, 38, 68, 74]. Evaluation did not tighten: only one study split the data by patient, and it reported 87.9% for patches but 68.6% when whole slices were classified [30]. Of the other eight, five documented patch- or image-level allocation, two were suspected of it and one did not report its split. Because the database contains 128 patients [24], a random split of thousands of patches places almost every patient on both sides of the partition. None of four further Geneva reports appraised from their abstracts stated a patient-level split [10, 25, 37, 77].

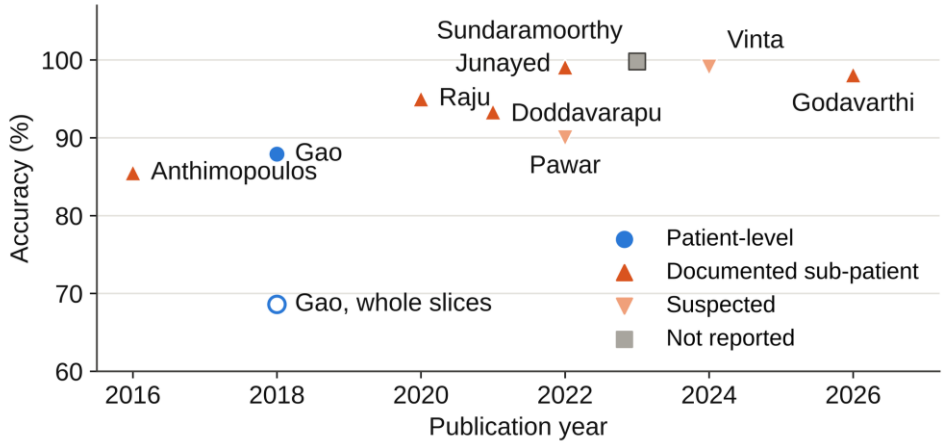

Fig. 5 Accuracy of pattern classifiers on the Geneva ILD database by publication year, by partitioning category. Open circle, whole-slice accuracy of the only study with patient-level partitioning

3.7 Reproducibility

The median reproducibility score was 1 of 6 (IQR 1 to 2), and no study scored more than 4. One study released code [64], three evaluated commercial systems with proprietary code [19, 49, 69], and 31 offered no code, including one that promised release on acceptance [43]. One study reported a random seed and two reported software versions [43, 64]. Nine used only public or controlled-access data, but none released its image selection or split. Twelve reported uncertainty (eight as confidence intervals and four as standard deviations), 16 retrained comparators on their own split, and two of the 15 diagnostic studies reported calibration [19, 65].

3.8 Sensitivity analysis of unretrieved reports

Of the 29 unretrieved database reports, 19 were likely eligible, four likely ineligible and six lacked the information needed for a decision. Together with the report found by citation searching, the 20 likely eligible reports (Table 5) comprised nine pattern and 11 diagnostic studies. None of the pattern reports stated a patient-level split, whereas ten of the 11 diagnostic reports described a patient-level or external design. Ten reports gave an overall accuracy: the median was 97.4% (n = 5) for patch-level or unstated designs and 77.8% (n = 5) for patient-level or external designs (U = 25.0, p = 0.008; r = 1.00; Fig. 4b). The report later retrieved in full had been classed from its abstract as patient-level, and its full text

confirmed this. Combining full-text and abstract-level evidence gave medians of 96.8% without patient-level designs (n = 18) and 78.5% with them (n = 15; U = 247.5, $p < 0.001$; r = 0.83). Because the abstract-level classes are coarse and subject to the same confounding by task, this analysis supports the direction of the main result but adds little independent evidence.

Table 5 Reports not retrieved in full text and judged likely eligible from their abstracts (n = 20)

Report	Journal	Classes	Data source; sample	Design stated in the abstract	Headline result
Pattern classification					
Cardoso 2018[a] [12]	Methods Inf Med	Diffuse lung disease patterns (regions of interest)	Not stated; 3,252 regions	Not stated	Accuracy 99.60%
Wang 2018 [77]	IEEE J Biomed Health Inform	Lung tissue patterns	Geneva database	Not stated	No value given
Joyseeree 2019 [37]	Med Image Anal	Four diseased tissue types and healthy tissue	Geneva database	Not stated	No value given
Huang 2020 [33]	Med Biol Eng Comput	ILD patterns	Not stated	Not stated	No value given
Doddavarapu 2020 [25]	Imaging Sci J	Normal, ground glass, emphysema, micronodules, fibrosis	Geneva database	Not stated	Accuracy 94.67%
Xu 2020 [80]	IEEE J Biomed Health Inform	Seven pulmonary textures	Source not stated	Not stated	Accuracy 94.78%
Chen 2024 [15]	Phys Eng Sci Med	ILD patterns (3D patches)	Not stated	Not stated	Accuracy 99.1%
Jadhav 2024 [36]	Indian J Pharm Educ Res	Seven classes	Not stated	Whole segmented lung images; split not stated	Class accuracy 84.12% to 94.65%
Büyükpatpat 2025 [10]	Int J Imaging Syst Technol	Five, six and seven patterns (slices)	MedGIFT (Geneva) database	Slice-based; split not stated	F-score 98.07% (five classes)
Diagnostic classification					
Walsh 2018 [75]	Lancet Respir Med	Guideline CT categories for IPF	Two institutions; 1,157 scans	Non-overlapping patient cohorts; test set of 139	Accuracy 76.4%
Bratt 2022 [9]	Chest	Histologic UIP vs. non-UIP	Multicentre; 1,239 patients	Patient-level hold-out test (198)	AUC 0.87
Koo 2022 [44]	Eur Radiol	Histologic UIP, NSIP and CHP	1,085 patients	Patient-level split; 911 for development	AUC 0.798 (three classes), 0.869 (UIP vs. non-UIP)

Report	Journal	Classes	Data source; sample	Design stated in the abstract	Headline result
Furukawa 2022 [29]	Respirology	IPF vs. other ILD	1,068 patients	Five-fold cross-validation over patients	Accuracy 83.6%
Thorat 2022 [72]	Int J Imaging Syst Technol	Three ILDs and healthy lung	Sample not stated	Not stated	Accuracy 97.4%
Bradley 2024 [8]	Am J Med Sci	IPF in non-definite UIP (commercial system)	Two clinical trials; 51 patients	External cohort	Sensitivity 76.5%
Humphries 2024 [35]	Am J Respir Crit Care Med	Histologic UIP	2,143 for training	Three independent cohorts	AUC 0.77 and 0.79
Baba 2025 [5]	Respir Investig	IPF, NSIP, HP and CTD-ILD	630 patients	Five-fold cross-validation over patients	CNN accuracy 62.4%
Kızıloğlu 2026 [42]	J Imaging Inform Med	Fibrotic HP vs. IPF	87 patients	Repeated patient-level five-fold cross-validation	Accuracy 83.03%, AUC 0.92
Wang 2026 [76]	Eng Appl Artif Intell	ILD diagnoses (multiclass)	Three centres; 2,901 patients, 5,213 scans	Two external cohorts	Macro-AUC 0.999 internal; 0.905 and 0.870 external
Ong 2026 [54]	Diagn Interv Imaging	UIP, NSIP and fibrotic bronchiolocentric patterns	Multicentre; 1,097 examinations	External test set (108)	Accuracy 77.8%

[a] Identified by citation searching; all others by the database searches. These reports were not part of the main synthesis. CHP, chronic hypersensitivity pneumonitis; CTD-ILD, ILD associated with connective tissue disease; HP, hypersensitivity pneumonitis.

4 Discussion

Across 35 studies, the accuracy reported for DL classification of ILD on CT was strongly associated with evaluation design. Studies with documented sub-patient partitioning reported a median accuracy 12.1 percentage points higher than studies with patient-level partitioning, studies without patient-level partitioning produced all six results of 98% or more, and evaluation stayed largely patch-level while accuracy on the Geneva benchmark approached 100%. This association cannot be read as an estimate of the effect of leakage, because the studies without patient-level partitioning were all pattern classifiers evaluated on patches, slices or images, and within that stratum the difference was smaller and not statistically significant. Part of the gap is therefore likely to reflect easier tasks and smaller prediction units. Three lines of evidence that are less exposed to this confounding point in the same direction. Performance fell in five of the seven studies that tested their own models on independent patients. The one study that compared a sample-level with a subject-level test found lower sensitivity for unseen subjects [7]. Controlled experiments in other imaging fields that changed only the split found large inflation [56, 70, 81]. Reproducibility was a further weakness: one of the 35 studies released code, none released its image selection or split, and the median reproducibility score was 1 of 6.

4.1 Why image-level partitioning can overstate accuracy

Patches from one CT series share the scanner, reconstruction kernel, slice thickness, noise texture and the patient's own anatomy and disease distribution. When such patches are divided at random, almost every test patch has near-duplicates in training, and a network can succeed by matching them rather than by learning features that transfer to new patients [39]. This is consistent with patch-level accuracies near 99% on the Geneva data, against 87.9% in the one audited study that kept patients apart [30]. The database's developers had evaluated it with leave-one-patient-out cross-validation [23], a practice that later studies largely abandoned. Repeated use of one small benchmark adds a slower bias, because design choices become tuned to it [73].

4.2 Acquisition variability

Several findings concern parameters that the imaging team controls. When the same patients were rescanned or reconstructed differently, tube current-time product was the leading factor associated with inconsistent predictions in two of three models, and 3 mm slices were also associated with disagreement [84]. Performance also fell on a second scanner [40] and in a second hospital despite image adaptation [47]. These shifts are consistent with the known sensitivity of quantitative CT features to scanner model and reconstruction kernel [16, 48]. In the independent cohort of the commercial IPF classifier, which spanned four manufacturers, 23 scanner models and 1 to 5 mm sections, results showed no gross inconsistency across manufacturers or slice thickness, but sensitivity was lowest in patients with preserved lung function [69]. Case mix, as well as acquisition, can therefore affect how a tool performs.

Two practical consequences follow. First, adherence to guideline HRCT protocols [46, 58], together with complete recording of tube current, kernel, slice thickness and patient position in the image metadata, is what allows a department to check whether its own examinations resemble those on which a tool was developed and to detect drift after deployment. Second, the radiologists and radiographers who validate and audit AI tools [32, 50] need to be able to read an evaluation design critically, and training in AI should include this skill [27].

4.3 Implications for developers and clinical users

For developers, the implications are direct. Data should be partitioned by patient before any patch, slice or augmented image is generated; model selection, early stopping and feature selection should use validation data drawn only from the training patients; and the test set should be used once, for the final estimate. A fixed patient-level split of public benchmarks such as the Geneva database, released with code, would make results from different groups comparable [73]. Editors and reviewers can ask for these details as a condition of publication. Engineering journals, where most of the pattern classifiers appeared, are especially well placed to act: asking authors to state the unit of allocation and to split by patient would address most of the concerns raised here.

An accuracy figure from an image-level split says little about how a tool will perform for the next patient referred with fibrotic lung disease. Even the better-designed diagnostic models lost performance outside their development data. Before a tool is procured or used to support multidisciplinary decisions, the imaging team should ask the questions in Table 6, beginning with whether patients were kept together and whether the test set was independent and untouched. Neither check needs specialist knowledge of DL.

Table 6 Questions for readers, reviewers and authors of deep learning studies in ILD, and the number of included studies that met each requirement

No.	Question	Studies meeting it
1	Were all images of each patient kept in one partition, with augmentation applied after the split?	17/35
2	Were hyperparameters, stopping, features and thresholds fixed without the test data?	12/35
3	Was the model tested on independent patients from another site, period or scanner cohort, with no patient in common with development?	11/35
4	Did all classes come from the same sources?	24/35
5	Are results reported per patient or scan?	16/35
6	Is uncertainty reported, preferably as confidence intervals?	12/35
7	For diagnostic models, is calibration reported?	2/15
8	Are code, seeds and software versions released, and were comparators retrained on the same split?	Code 1, seeds 1, versions 2, comparators 16 (of 35)

4.4 What this review adds

Leakage has been quantified in other imaging fields [4, 56, 70, 81], and in Alzheimer's disease reported accuracy rose with leakage risk across studies [82]. To our knowledge, however, no previous review has related evaluation design to reported accuracy for DL classification of ILD on CT. This review linked judgements on the routes by which test information can reach a model to the accuracy each study reported, audited a shared benchmark over ten years, assembled the within-study comparisons in which authors measured their own loss of performance, and showed that the evidence base itself depends on the databases searched. It also measured how rarely the reported results can be checked: one study released code, and none released its data split. Pooled estimates that mix image-level and patient-level designs may therefore overstate performance in practice, and future syntheses should stratify by partitioning, task, prediction unit and test strength.

4.5 Strengths and limitations

Six databases were searched, including engineering sources that MEDLINE does not cover, and citation searching of 1,913 linked records found only one further likely eligible report. Every record was screened by two reviewers with a third adjudicating, every criterion and decision rule is reported in full, and every partitioning, tuning and test judgement is supported by a quoted source passage, with the dataset and code provided so that each judgement and statistic can be checked.

Some limitations remain. Subscription databases such as Scopus, Web of Science and Embase were not searched, although the 201 records found only in Europe PMC, IEEE Xplore or ScienceDirect contained no eligible study, which suggests that few studies were missed. Thirty potentially eligible reports could not be read in full despite a second retrieval attempt, and their corresponding authors were not contacted; the abstract-level classification is coarse and confounded by task in the same way, so it supports only the direction of the main result.

The report retrieved on 11 November 2026 was not citation-searched, because no searches were run after 29 October 2026. The comparisons rest on modest numbers, and because partitioning was almost completely confounded with task and prediction unit, they describe an association that the review cannot attribute to leakage alone. For four closed-access studies the full text could not be re-accessed when supporting passages were compiled, so their judgements rest on the original extraction and are flagged in Online Resource 3. The rigour and reproducibility scores weight items equally and reflect what authors reported, and only English-language articles were eligible. Three further process limitations should be noted. Screening decisions were agreed jointly rather than made independently in duplicate, although the supervising author checked every decision (Sect. 2.3). The appraisal domains were a focused adaptation of PROBAST+AI rather than the complete tool, chosen so that patch-level

pattern classifiers and diagnostic models could be judged on the same terms. No protocol was registered, for the reason given in Sect. 2, and the criteria, decision rules, dataset and code are therefore reported in full.

5 Conclusion

Much of the very high accuracy reported for DL classification of ILD on CT comes from pattern classifiers evaluated below the patient level without patient-level partitioning. The review cannot separate the contribution of partitioning from that of task and prediction unit, but in either case such figures do not show how a model will perform for new patients. Patient-level partitioning, an untouched test set of independent patients, testing across scanners and protocols, calibration and released code should be minimum requirements. Developers should adopt these practices by default, and radiologists and radiographers, who control acquisition and increasingly govern the use of AI, are well placed to check them before such tools enter clinical practice.

Declarations

Ethics approval

Not applicable. This systematic review analysed only previously published data and involved no human participants, animals or identifiable personal information.

Data, material and code availability

The PRISMA 2020 checklist (Online Resource 1), the complete search strategies (Online Resource 2), the extraction dataset with the source passage supporting each judgement, the abstract-level appraisal and the retrieval log (Online Resource 3), the appraisal manual (Online Resource 4) and the analysis code (Online Resource 5) are provided as electronic supplementary material in the given link https://github.com/adyan-ahmad-7/Deep-learning-for-ILD-on-computed-tomography-A-methodological-systematic-review. The screening log is available from the corresponding author on reasonable request.

Supplementary Information

Online Resource 1 (ESM1.pdf): completed PRISMA 2020 checklist. Online Resource 2 (ESM2.pdf): complete search strategies for every source and sub-query, with dates, record counts and PRISMA-S items. Online Resource 3 (ESM3.xlsx): extraction dataset for the 35 included studies, with the quoted passage supporting each judgement, within-study comparisons, group summaries, the abstract-level appraisal of unretrieved reports, the retrieval log and a data dictionary. Online Resource 4 (ESM4.pdf): appraisal manual with the decision rules and examples. Online Resource 5 (ESM_5.zip): Python code and data that reproduce every statistic and figure, with a README.

REFERENCES

Agarwala S, Kumar A, Dhara AK, Thakur SB, Sadhu AK, Nandi D (2021) Special convolutional neural network for identification and positioning of interstitial lung disease patterns in computed tomography images. Pattern Recognit Image Anal 31:730–738. https://doi.org/10.1134/S1054661821040027

Ahmad Y, Mooney J, Allen IE, Seaman J, Kalra A, Muelly M, Reicher J (2024) A machine learning system to indicate diagnosis of idiopathic pulmonary fibrosis non-invasively in challenging cases. Diagnostics (Basel) 14:830. https://doi.org/10.3390/diagnostics14080830

Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S (2016) Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. IEEE Trans Med Imaging 35:1207–1216. https://doi.org/10.1109/TMI.2016.2535865

Ashames MMA, Demir A, Gerek ON, Fidan M, Gulmezoglu MB, Ergin S, Edizkan R, Koc M, Barkana A, Calisir C (2024) Are deep learning classification results obtained on CT scans fair and interpretable? Phys Eng Sci Med 47:967–979. https://doi.org/10.1007/s13246-024-01419-8

Bae HJ, Kim CW, Kim N, Park B, Kim N, Seo JB, Lee SM (2018) A Perlin noise-based augmentation strategy for deep learning with small data samples of HRCT images. Sci Rep 8:17687. https://doi.org/10.1038/s41598-018-36047-2

Bermejo-Peláez D, Ash SY, Washko GR, San José Estépar R, Ledesma-Carbayo MJ (2020) Classification of interstitial lung abnormality patterns with an Ensemble of deep convolutional neural networks. Sci Rep 10:338. https://doi.org/10.1038/s41598-019-56989-5

10. Büyükpatpat H, Sezer EA, Güzel MS (2025) Comprehensive experimentation of pretrained models on slice-based classification of interstitial lung disease patterns. Int J Imaging Syst Technol 35:e70232. https://doi.org/10.1002/ima.70232
11. Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S et al (2020) Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ 368:l6890. https://doi.org/10.1136/bmj.l6890
12. Cardoso I, Almeida E, Allende-Cid H, Frery AC, Rangayyan RM, Azevedo-Marques PM, Ramos HS (2018) Analysis of machine learning algorithms for diagnosis of diffuse lung diseases. Methods Inf Med 57:272–279. https://doi.org/10.1055/s-0039-1681086
13. Chang M, Reicher JJ, Kalra A, Muelly M, Ahmad Y (2024) 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 37:297–307. https://doi.org/10.1007/s10278-023-00914-w
14. Chantzi SL, Kosvyra A, Chouvarda I (2025) Radiomics and artificial intelligence in pulmonary fibrosis. J Imaging Inform Med 38:2779–2792. https://doi.org/10.1007/s10278-024-01377-3
16. Choe J, Lee SM, Do KH, Lee G, Lee JG, Lee SM et al (2019) Deep learning-based image conversion of CT reconstruction kernels improves radiomics reproducibility for pulmonary nodules or masses. Radiology 292:365–373. https://doi.org/10.1148/radiol.2019181960
17. Christe A, Peters AA, Drakopoulos D, Heverhagen JT, Geiser T, Stathopoulou T, Christodoulidis S, Anthimopoulos M, Mougiakakou SG, Ebner L (2019) Computer-aided diagnosis of pulmonary fibrosis using deep learning and CT images. Invest Radiol 54:627–632. https://doi.org/10.1097/RLI.0000000000000574
18. Christodoulidis S, Anthimopoulos M, Ebner L, Christe A, Mougiakakou S (2017) Multisource transfer learning with convolutional neural networks for lung pattern analysis. IEEE J Biomed Health Inform 21:76–84. https://doi.org/10.1109/JBHI.2016.2636929
19. Chung JH, Chelala L, Pugashetti JV, Wang JM, Adegunsoye A, Matyga AW et al (2024) A deep learning-based radiomic classifier for usual interstitial pneumonia. Chest 165:371–380. https://doi.org/10.1016/j.chest.2023.10.012
20. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B et al (2024) TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 385:e078378. https://doi.org/10.1136/bmj-2023-078378
21. Cong L, Chen Y, He X, KuErBan M, Wu C, Chen L (2025) Diagnosis accuracy of machine learning for idiopathic pulmonary fibrosis: a systematic review and meta-analysis. Eur J Med Res 30:288. https://doi.org/10.1186/s40001-025-02501-x
22. DeGrave AJ, Janizek JD, Lee SI (2021) AI for radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell 3:610–619. https://doi.org/10.1038/s42256-021-00338-7
23. Depeursinge A, Van de Ville D, Platon A, Geissbuhler A, Poletti PA, Müller H (2012) Near-affine-invariant texture learning for lung tissue analysis using isotropic wavelet frames. IEEE Trans Inf Technol Biomed 16:665–675. https://doi.org/10.1109/TITB.2012.2198829
24. Depeursinge A, Vargas A, Platon A, Geissbuhler A, Poletti PA, Müller H (2012) Building a reference multimedia database for interstitial lung diseases. Comput Med Imaging Graph 36:227–238. https://doi.org/10.1016/j.compmedimag.2011.07.003
25. Doddavarapu VNS, Kande GB, Rao BP (2020) Differential diagnosis of interstitial lung diseases using deep learning networks. Imaging Sci J 68:170–178. https://doi.org/10.1080/13682199.2020.1781394
26. Doddavarapu VNS, Kande GB, Rao BP (2021) Rotational invariant fractional derivative filters for lung tissue classification. IET Image Process 15:2202–2212. https://doi.org/10.1049/IPR2.12188

27. Doherty G, McLaughlin L, Hughes C, McConnell J, Bond R, McFadden S (2024) Radiographer education and learning in artificial intelligence (REAL-AI): A survey of radiographers, radiologists, and students' knowledge of and attitude to education on AI. Radiography 30(Suppl 2):79–87. https://doi.org/10.1016/j.radi.2024.10.010
28. Fontanellaz M, Christe A, Christodoulidis S, Dack E, Roos JE, Drakopoulos D et al (2024) Computer-aided diagnosis system for lung fibrosis: from the effect of radiomic features and multi-layer-perceptron mixers to pre-clinical evaluation. IEEE Access 12:25642–25656. https://doi.org/10.1109/ACCESS.2024.3350430
30. Gao M, Bagci U, Lu L, Wu A, Buty M, Shin HC et al (2018) Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks. Comput Methods Biomech Biomed Eng Imaging Vis 6:1–6. https://doi.org/10.1080/21681163.2015.1124249
31. Godavarthi AT, Nakkeeran R (2026) A hybrid CNN with feature fusion for patch-level interstitial lung disease classification. Eng Technol Appl Sci Res 16:39270–39277. https://doi.org/10.48084/etasr.19461
32. Hardy M, Harvey H (2020) Artificial intelligence in diagnostic imaging: impact on the radiography profession. Br J Radiol 93:20190840. https://doi.org/10.1259/bjr.20190840
34. Huang X, Si W, Ye X, Zhao Y, Gu H, Zhang M et al (2024) Novel 3D-based deep learning for classification of acute exacerbation of idiopathic pulmonary fibrosis using high-resolution CT. BMJ Open Respir Res 11:e002226. https://doi.org/10.1136/bmjresp-2023-002226
37. Joyseeree R, Otálora S, Müller H, Depeursinge A (2019) Fusing learned representations from Riesz filters and deep CNN for lung tissue classification. Med Image Anal 56:172–183. https://doi.org/10.1016/j.media.2019.06.006
38. Junayed MS, Jeny AA, Islam MB, Ahmed I, Shah AFMS (2022) An efficient end-to-end deep neural network for interstitial lung disease recognition and classification. Turk J Electr Eng Comput Sci 30:1235–1250. https://doi.org/10.55730/1300-0632.3846
39. Kapoor S, Narayanan A (2023) Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y) 4:100804. https://doi.org/10.1016/j.patter.2023.100804
40. Kim GB, Jung KH, Lee Y, Kim HJ, Kim N, Jun S, Seo JB, Lynch DA (2018) Comparison of shallow and deep learning methods on classifying the regional pattern of diffuse lung disease. J Digit Imaging 31:415–424. https://doi.org/10.1007/s10278-017-0028-9
41. Kiziloğlu HA, Çevik E, Zengin K (2024) Evaluation of interstitial lung diseases with deep learning method of two major computed tomography patterns. Curr Med Imaging 20:e15734056279295. https://doi.org/10.2174/0115734056279295231229095436
43. Kiziloğlu HA, Zengin K (2026) Attention mechanism-enhanced deep learning for the differential diagnosis of idiopathic pulmonary fibrosis on computed tomography. Curr Med Imaging 22:e15734056487503. https://doi.org/10.2174/0115734056487503260805094403
46. Lynch DA, Sverzellati N, Travis WD, Brown KK, Colby TV, Galvin JR et al (2018) Diagnostic criteria for idiopathic pulmonary fibrosis: a Fleischner Society white paper. Lancet Respir Med 6:138–153. https://doi.org/10.1016/S2213-2600(17)30433-2
47. Mabu S, Miyake M, Kuremoto T, Kido S (2021) Semi-supervised CycleGAN for domain transformation of chest CT images and its application to opacity classification of diffuse lung diseases. Int J Comput Assist Radiol Surg 16:1925–1935. https://doi.org/10.1007/s11548-021-02490-2
48. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B et al (2015) Measuring computed tomography scanner variability of radiomics features. Invest Radiol 50:757–765. https://doi.org/10.1097/RLI.0000000000000180
49. Maddali MV, Kalra A, Muelly M, Reicher JJ (2023) Development and validation of a CT-based deep learning algorithm to augment non-invasive diagnosis of idiopathic pulmonary fibrosis. Respir Med 219:107428. https://doi.org/10.1016/j.rmed.2023.107428
50. Malamateniou C, McFadden S, McQuinlan Y, England A, Woznitza N, Goldsworthy S et al (2021) Artificial intelligence: Guidance for clinical imaging and therapeutic radiography

professionals, a summary by the Society of Radiographers AI working group. Radiography 27:1192–1202. https://doi.org/10.1016/j.radi.2021.07.028
51. Mei X, Liu Z, Singh A, Lange M, Boddu P, Gong JQX et al (2023) Interstitial lung disease diagnosis and prognosis using an AI system integrating longitudinal data. Nat Commun 14:2272. https://doi.org/10.1038/s41467-023-37720-5
52. Moassefi M, Rouzrokh P, Conte GM, Vahdati S, Fu T, Tahmasebi A et al (2023) Reproducibility of deep learning algorithms developed for medical imaging analysis: A systematic review. J Digit Imaging 36:2306–2312. https://doi.org/10.1007/s10278-023-00870-5
53. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P et al (2025) PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ 388:e082505. https://doi.org/10.1136/bmj-2024-082505
55. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372:n71. https://doi.org/10.1136/bmj.n71
56. Parsarad S, Saeedizadeh N, Soufi GJ, Shafieyoon S, Hekmatnia F, Zarei AP et al (2023) Biased deep learning methods in detection of COVID-19 using CT images: A challenge mounted by subject-wise-split ISFCT dataset. J Imaging 9:159. https://doi.org/10.3390/jimaging9080159
57. Pawar SP, Talbar SN (2022) Two-stage hybrid approach of deep learning networks for interstitial lung disease classification. Biomed Res Int 2022:7340902. https://doi.org/10.1155/2022/7340902
58. Raghu G, Remy-Jardin M, Myers JL, Richeldi L, Ryerson CJ, Lederer DJ et al (2018) Diagnosis of idiopathic pulmonary fibrosis. An official ATS/ERS/JRS/ALAT clinical practice guideline. Am J Respir Crit Care Med 198:e44–e68. https://doi.org/10.1164/rccm.201807-1255ST
59. Raghu G, Remy-Jardin M, Richeldi L, Thomson CC, Inoue Y, Johkoh T et al (2022) Idiopathic pulmonary fibrosis (an update) and progressive pulmonary fibrosis in adults: An official ATS/ERS/JRS/ALAT clinical practice guideline. Am J Respir Crit Care Med 205:e18–e47. https://doi.org/10.1164/rccm.202202-0399ST
60. Raju N, Anita HB, Augustine DP (2020) Identification of interstitial lung diseases using deep learning. Int J Electr Comput Eng 10:6283–6291. https://doi.org/10.11591/ijece.v10i6.pp6283-6291
61. Refaee T, Salahuddin Z, Frix AN, Yan C, Wu G, Woodruff HC et al (2022) Diagnosis of idiopathic pulmonary fibrosis in high-resolution computed tomography scans using a combination of handcrafted radiomics and deep learning. Front Med (Lausanne) 9:915243. https://doi.org/10.3389/fmed.2022.915243
62. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, Koffel JB (2021) PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 10:39. https://doi.org/10.1186/s13643-020-01542-z
63. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S et al (2021) Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell 3:199–217. https://doi.org/10.1038/s42256-021-00307-0
64. Savadjiev P, Rezanejad M, Bhatnagar S, Camirand D, Kauffmann C, Siddiqi K, Dandurand RJ, Bourgouin P, Chartrand-Lefebvre C, Semionov A (2026) Automated diagnosis of usual interstitial pneumonia on chest CT via the mean curvature of isophotes. Eur J Radiol Open 16:100723. https://doi.org/10.1016/j.ejro.2025.100723
65. Shaish H, Ahmed FS, Lederer D, D'Souza B, Armenta P, Salvatore M, Saqi A, Huang S, Jambawalikar S, Mutasa S (2021) Deep learning of computed tomography virtual wedge resection for prediction of histologic usual interstitial pneumonitis. Ann Am Thorac Soc 18:51–59. https://doi.org/10.1513/AnnalsATS.202001-068OC

66. Sithiq HAAH, Shuib LM, Ahmad M, Antony CD (2025) HCL Net: deep learning for accurate classification of honeycombing lung and ground glass opacity in CT images. Comput Mater Contin 86:1–25. https://doi.org/10.32604/cmc.2025.067781
67. Soffer S, Morgenthau AS, Shimon O, Barash Y, Konen E, Glicksberg BS, Klang E (2022) Artificial intelligence for interstitial lung disease analysis on chest computed tomography: A systematic review. Acad Radiol 29(Suppl 2):S226–S235. https://doi.org/10.1016/j.acra.2021.05.014
68. Sundaramoorthy K, Anitha R, Kayalvili S, Ghazala AFA, El-Ebiary YAB, Al-Ashmawy S (2023) Hybrid optimization with recurrent neural network-based medical image processing for predicting interstitial lung disease. Int J Adv Comput Sci Appl 14. https://doi.org/10.14569/IJACSA.2023.0140462
69. Tahamtani-Omran O, Kalra A, Muelly M, Reicher J, Ahmed R (2026) Paired validation of a multimodal vision-transformer system for the non-invasive diagnosis of idiopathic pulmonary fibrosis. Diagnostics (Basel) 16:2398. https://doi.org/10.3390/diagnostics16152398
70. Tampu IE, Eklund A, Haj-Hosseini N (2022) Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images. Sci Data 9:580. https://doi.org/10.1038/s41597-022-01618-6
71. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, Kahn CE (2024) Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. Radiol Artif Intell 6:e240300. https://doi.org/10.1148/ryai.240300
73. Varoquaux G, Cheplygina V (2022) Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digit Med 5:48. https://doi.org/10.1038/s41746-022-00592-y
74. Vinta SR, Lakshmi B, Safali MA, Kumar GSC (2024) Segmentation and classification of interstitial lung diseases based on hybrid deep learning network model. IEEE Access 12:50444–50458. https://doi.org/10.1109/ACCESS.2024.3383144
77. Wang Q, Zheng Y, Yang G, Jin W, Chen X, Yin Y (2018) Multiscale rotation-invariant convolutional neural networks for lung texture classification. IEEE J Biomed Health Inform 22:184–195. https://doi.org/10.1109/JBHI.2017.2685586
78. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM, QUADAS-2 Group (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 155:529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009
81. Yagis E, Atnafu SW, García Seco de Herrera A, Marzi C, Scheda R, Giannelli M, Tessa C, Citi L, Diciotti S (2021) Effect of data leakage in brain MRI classification using 2D convolutional neural networks. Sci Rep 11:22544. https://doi.org/10.1038/s41598-021-01681-w
82. Young VM, Gates S, Garcia LY, Salardini A (2025) Data leakage in deep learning for Alzheimer's disease diagnosis: A scoping review of methodological rigor and performance inflation. Diagnostics (Basel) 15:2348. https://doi.org/10.3390/diagnostics15182348
83. Yu AC, Mohajer B, Eng J (2022) External validation of deep learning algorithms for radiologic diagnosis: A systematic review. Radiol Artif Intell 4:e210064. https://doi.org/10.1148/ryai.210064
84. Yu W, McNitt-Gray MF, Goldin JG, Song JW, Kim GHJ (2025) Evaluating the robustness of deep learning models trained to diagnose idiopathic pulmonary fibrosis using a retrospective study. Med Phys 52:4239–4249. https://doi.org/10.1002/mp.17752
85. Yu W, Zhou H, Choi Y, Goldin JG, Teng P, Wong WK, McNitt-Gray MF, Brown MS, Kim GHJ (2023) 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 50:894–905. https://doi.org/10.1002/mp.16053
86. Yu W, Zhou H, Goldin JG, Wong WK, Kim GHJ (2021) End-to-end domain knowledge-assisted automatic diagnosis of idiopathic pulmonary fibrosis (IPF) using computed tomography (CT). Med Phys 48:2458–2467. https://doi.org/10.1002/mp.14754

87. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK (2018) Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 15:e1002683. https://doi.org/10.1371/journal.pmed.1002683
88. Zhang J, He L, Wei Y, Tong J, Yang K, Wu J, Guo Y, Shi F, Jin C (2025) Deep learning for classifying imaging patterns of interstitial lung disease associated with idiopathic inflammatory myopathies. Sci Rep 15:31655. https://doi.org/10.1038/s41598-025-15960-3
89. Zhang Q, Shao D, Lin L, Gong G, Xu R, Kido S, Cui H (2025) Feature separation in diffuse lung disease image classification by using evolutionary algorithm-based NAS. IEEE J Biomed Health Inform 29:2706–2717. https://doi.org/10.1109/JBHI.2024.3481012