Special Review

Artificial Intelligence for Interstitial Lung Disease Analysis on Chest Computed Tomography: A Systematic Review

Shelly Soffer, MD, Adam S. Morgenthau, MD, Orit Shimon, MD, Yiftach Barash, MD, Eli Konen, MD, Benjamin S Glicksberg, PhD, Eyal Klang, MD

Rationale and Objectives: High-resolution computed tomography (HRCT) is paramount in the assessment of interstitial lung disease (ILD). Yet, HRCT interpretation of ILDs may be hampered by inter- and intra-observer variability. Recently, artificial intelligence (AI) has revolutionized medical image analysis. This technology has the potential to advance patient care in ILD. We aimed to systematically evaluate the application of AI for the analysis of ILD in HRCT.

Materials and Methods: We searched MEDLINE/PubMed databases for original publications of deep learning for ILD analysis on chest CT. The search included studies published up to March 1, 2021. The risk of bias evaluation included tailored Quality Assessment of Diagnostic Accuracy Studies and the modified Joanna Briggs Institute Critical Appraisal checklist.

Results: Data was extracted from 19 retrospective studies. Deep learning techniques included detection, segmentation, and classification of ILD on HRCT. Most studies focused on the classification of ILD into different morphological patterns. Accuracies of 78%-91% were achieved. Two studies demonstrated near-expert performance for the diagnosis of idiopathic pulmonary fibrosis (IPF). The Quality Assessment of Diagnostic Accuracy Studies tool identified a high risk of bias in 15/19 (78.9%) of the studies.

Conclusion: AI has the potential to contribute to the radiologic diagnosis and classification of ILD. However, the accuracy performance is still not satisfactory, and research is limited by a small number of retrospective studies. Hence, the existing published data may not be sufficiently reliable. Only well-designed prospective controlled studies can accurately assess the value of existing AI tools for ILD evaluation.

Key Words: Interstitial Lung Diseases; Computed Tomography, Spiral; Deep Learning; Neural Networks (Computer); Artificial Intelligence.

Abbreviations: ILD Interstitial lung disease, **HRCT** High-resolution computed tomography, **CNN** Convolutional neural networks, **AI** Artificial intelligence, **ANN** Artificial Neural Networks, **HUG** University Hospital of Geneva, **cHP** chronic hypersensitivity pneumonitis, **IPF** Idiopathic pulmonary fibrosis, **UIP** Usual interstitial pneumonia, **ILA** Interstitial lung abnormalities

Acad Radiol 2021; ■:1–10

From the Internal Medicine B, Assuta Medical Center, Ashdod, Israel, and Ben-Gurion University of the Negev, Be'er Sheva, Israel (S.S.); DeepVision Lab, Sheba Medical Center, Tel Hashomer, Israel (S.S., Y.B., E.K.); Department of Medicine, Division of Pulmonary, Critical Care, and Sleep Medicine, Icahn School of Medicine at Mount Sinai, New York, New York (A.S.M.); Department of Anesthesia, Rabin Medical Center, Beilinson Hospital, Petach Tikvah, Israel, and Sackler Medical School, Tel Aviv University, Tel Aviv, Israel (O.S.); Department of Diagnostic Imaging, Sheba Medical Center, Tel Hashomer, Israel, and Sackler Medical School, Tel Aviv University, Tel Aviv, Israel (Y.B., E.K., E.K.); Hasso Plattner Institute for Digital Health at Mount Sinai, Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York (B.S.G.); Institute for Healthcare Delivery Science, Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, New York (E.K.). Received March 24, 2021; revised May 10, 2021; accepted May 11, 2021. **Address correspondence to:** Shelly Soffer, Samson Assuta Ashdod University Hospital, Ha-Refu'a St 7, Ashdod, 7747629, Israel. e-mail: soffer.shelly@gmail.com

https://doi.org/10.1016/j.acra.2021.05.014

INTRODUCTION

Interstitial lung disease (ILD) refers to a group of more than 200 conditions. These conditions exhibit different degrees of lung parenchymal fibrosis and inflammation (1). Among respiratory diseases, ILD represents a major cause of mortality and morbidity. The diagnosis of ILD is partially dependent on medical history, physical examination, and pulmonary function testing. Yet, high-resolution computed tomography (HRCT) is the key to the diagnosis (2,3).

There are several common radiologic features of ILD. These include reticular opacities, cystic lesions, ground-glass opacities, and nodular patterns. The appearance, location, and quantification of these findings differentiate between ILDs and indicate severity. In some cases, diagnosis and classification is relatively straightforward. In other cases, HRCT imaging consists of radiologic features common to multiple

ILDs. Interpretation of these non-typical images may be challenging, even for experienced readers. This may entail inter- and intra-observer variation and ambiguity in the CT assessment (4).

During recent years, artificial intelligence (AI) achieved impressive performance in computer vision tasks (5,6). The breakthrough was made possible due to the emerging technology of deep learning. The prominent deep learning algorithm for image analysis is convolutional neural networks (CNN). A significant turning point for CNN was the ImageNet Challenge of 2012. In this competition, CNN showed revolutionary performance compared with traditional computer vision techniques (7).

CNNs are highly adept at medical image processing (5,8–10). These include radiology images, skin lesions, retinal scans, endoscopic images, and histopathologic specimens (11–15). Therefore, it is not surprising that CNN techniques are being utilized in respiratory medicine (16–18) and are being specifically used for the analysis of HRCT images in ILD.

In this review, we aimed to systematically evaluate the application of AI for the analysis of ILD in HRCT.

DEEP LEARNING FUNDAMENTALS APPLIED TO ILD

Artificial Intelligence Hierarchy of Terms

Deep learning falls under the umbrella of terms of AI and machine learning (Fig 1). AI is a broad term describing algorithms for problems that require human intelligence. Machine learning is considered a subclass of AI. Machine learning algorithms have the ability to learn without being explicitly programmed. Deep learning is a subclass of AI and machine learning. In deep learning, distinctive features are learned from training data without predefined features (6,19).

Artificial Neural Networks (ANN)

ANNs are the basis of most deep learning methods. ANNs can be loosely compared to biological neural networks (Fig 2). ANNs are composed of multiple layers of interconnected neurons. Each neuron is just a simple mathematical function. The neuron basically weight-averages the outputs of neurons from the previous layer. Then the neuron outputs the results to neurons in the following layer. The final output of the network

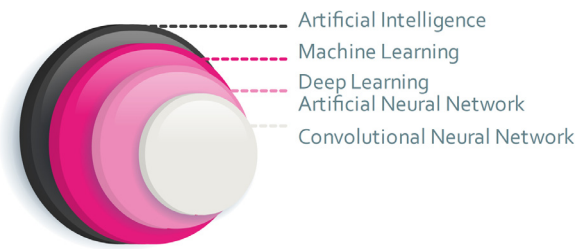

Figure 1. Diagram of the artificial intelligence hierarchy of terms.

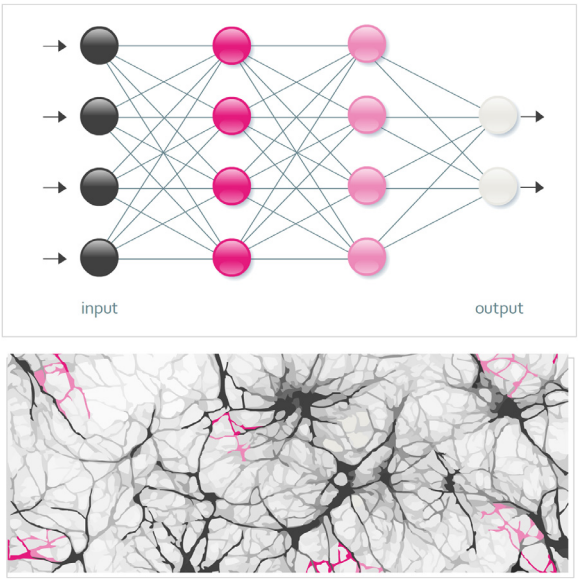

Figure 2. Analogy between biologic and artificial neural networks.

may be a simple logistic regression unit. This unit will ultimately provide a probability regarding the network's input (19). For example, an input may be a chest X-ray image, while the output may be "pneumonia" (1) vs. "no pneumonia" (0). The real power of a neural network is derived from the large number of layers and neurons.

Convolutional Neural Networks (CNN)

CNNs (Fig 3) are a type of ANN designed to handle image data. CNNs are specifically built to exploit repeating patterns (5). This is because images usually have repeating patterns.

Computer Vision

Computer vision is a term describing the use of computers to analyze images. Three main tasks in computer vision include: classification, detection, and segmentation (Fig 4).

Classification labels an entire image or image patch. *Detection* is the localization of an object in the image, using a region of interest, such as a box or a circle. *Segmentation* delineates the pixel-wise borders of an object such as a pathology or an organ.

Understanding these tasks can be exemplified through the analysis of a CT slice. We can classify the whole slices as either pathologic (pneumonia, ILD etc.) or as normal lung. We can further detect pathology such as a nodule using a box or a circle plot. Lastly, we can segment the pixel-wise borders of the lung or honeycombing area.

U-Net is a fully convolutional neural network that was specifically designed for the task of segmentation of biomedical images (5,20). The architecture is able to localize the borders of an abnormality, as the classification is performed on every single pixel. Thus, the input and output share the same

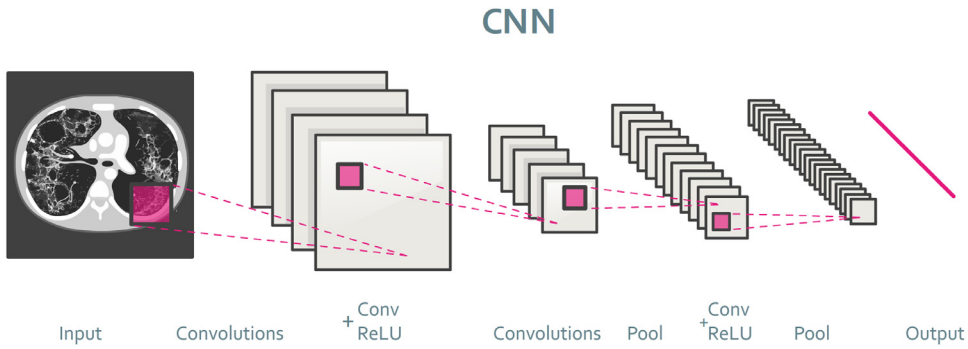

Figure 3. Schematic representation of convolutional neural networks (CNN).

size, as the output is a map of probabilities which corresponds to the map of image pixels.

Patch-based Approach

HRCT, with a thin slice section of 0·625-1·25 mm and wide matrix size of 768 × 768, contains a vast number of voxels. One of the limitations of deep learning is the reliance on very large neural networks. To analyze a full 3D HRCT, a huge amount of computer memory is needed. This is why a common engineering approach is to use a patch-based analysis (Fig 5). In this method, each CT slice is divided into many small patches. The network is then fed with individual patches as inputs.

For research, reporting the individual metrics of the network for each patch is perhaps sufficient. Yet, for clinical purposes, a holistic solution is needed, as the clinician is interested in the classification of the entire scan or of regions of the scan as a whole. Thus, some statistical methods may be applied to the results of all the patches. These may include averaging, majority vote or a machine learning analysis.

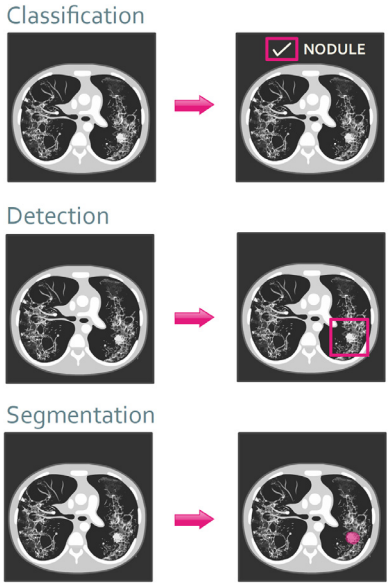

Figure 4. Main computer vision tasks: classification, detection, and segmentation.

METHODS

Search Strategy

A systematic review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement (21), relevant guidelines from the diagnostic test accuracy extension (22), and the recommendations for systematic reviews of prediction models (CHARMS checklist) (23). Modifications were made to suit reporting of machine learning models in biomedical research, as offered by Luo et al. (24).

A systematic search of the published literature was conducted on March 1, 2021. MEDLINE/PubMed were used as databases. The search was performed by using the following terms: ("Lung Diseases, Interstitial"[Mesh] or "interstitial" or "pneumonia" or "pulmonary fibrosis" or "ILD" or "interstitial pneumonia" or "interstitial pneumonitis" or "diffuse parenchymal lung diseases" or "interstitial pneumonitides" or "lung disease" or "Cryptogenic Organizing Pneumonia" or "Sarcoidosis" or "Hypersensitivity Pneumonitis" or "Idiopathic pleuroparenchymal fibroelastosis" or "Lymphangioleiomyomatosis" or "Langerhan's cell histiocytosis") and ("deep learning"[Mesh] or "convolutional neural network" or "CNN" or "Neural Networks, Computer"[Mesh]).

We limited the search to articles in English. Peer-reviewed original publications on the subject of deep learning applications to ILD were included. We excluded non-computer vision articles, non-CNN articles, and non-original articles. Conference abstracts were excluded. To ensure that we did not inadvertently exclude relevant articles, we searched the bibliographies of the articles included in our study. The study is registered with PROSPERO (CRD42020210070).

Study Selection

Two reviewers (SS and EK) independently screened the titles and abstracts to determine whether the studies met the inclusion criteria. In unclear cases, the full text article was reviewed. Disagreements were adjudicated by a third

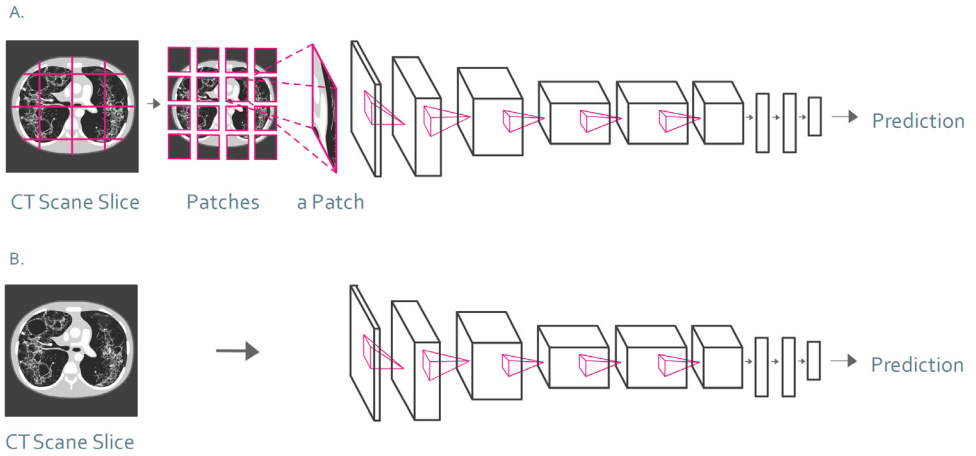

Figure 5. Patch based vs. holistic approach for analyzing CT slices.

reviewer (YB). The two authors (SS and EK) independently assessed the full-texts of the included articles.

Data Extraction

Data from all included studies was collected into a standardized data extraction sheet. Data included publication year, study design, ethical statements, clinical application, inclusion and exclusion criteria, description of the study population, use of an online database, size of the database, use of independent test dataset, whether cross-validation was performed, evaluation metrics, and performance results.

Quality Assessment and Risk of Bias

To evaluate risk of bias, the adapted version of Quality Assessment of Diagnostic Accuracy Studies criteria was used (19,25,26). The studies were also evaluated using the modified Joanna Briggs Institute Critical Appraisal checklist for analytical cross-sectional studies, as was used in Kwong et al. (19,27).

Data Synthesis

We intentionally planned not to perform a formal quantitative synthesis due to heterogeneity between studies (28).

RESULTS

A total of 379 articles were retrieved in the initial search. Nineteen studies applied deep learning to HRCT imaging of ILD. Figure 6 presents a flow diagram of the search.

Descriptive Summary of Results

Table 1 summarizes the characteristics of the articles included in this systematic review. All studies were retrospective. Seven studies appeared in clinical journals, whereas 12 papers appeared in bio-medical papers. Deep learning tasks included detection, segmentation, and classification of ILD on HRCT.

ILD Segmentation

Among the reviewed studies, several focused on segmentation of diseased lungs (29-31). Lung segmentation is a fundamental step for HRCT image analysis. CNN delineates the anatomic boundaries between normal and diseased lung areas. Park et al. analyzed lung segmentation on HRCT in 647 ILD patients (30). Using state-of-the-art CNN segmentation architecture, U-Net (20), a variety of ILD patterns were examined. U-Net was remarkably accurate, yielding a dice similarity coefficient value of 98·8%. The network was most accurate for NSIP and least accurate for COP. Errors in segmentation were mainly observed in the lung hilum.

Anthimopoulos et al. employed a CNN for the segmentation of normal and diseased lung. With regard to the latter, the CNN evaluated common radiologic features of ILDs including ground-glass opacity, micronodules, consolidation, reticulation, and honeycombing. The dataset included 172 HRCT scans from the University Hospital of Geneva (HUG) and from a private dataset (32). This CNN showed an accuracy of 81·8%. Errors in segmentation resulted from the imaging of normal and diseased lung of similar densities. Emphysematous and consolidative patterns were particularly problematic. In addition, the area between bronchovascular trees was prone to errors in segmentation.

Classifications of ILD Patterns

In most of the reviewed studies, AI was used to classify ILD patterns on HRCT. Typically identified features of ILD included: normal lung, ground-glass opacity, reticular opacity, honeycombing, emphysema, and consolidation (33-42). The majority of studies utilized the HUG public dataset (32), which consists of HRCT scans from 109 ILD patients (33-35,38-41,43).

To maximize the network's performance of this small dataset, both transfer learning (39,40,43) and data augmentation

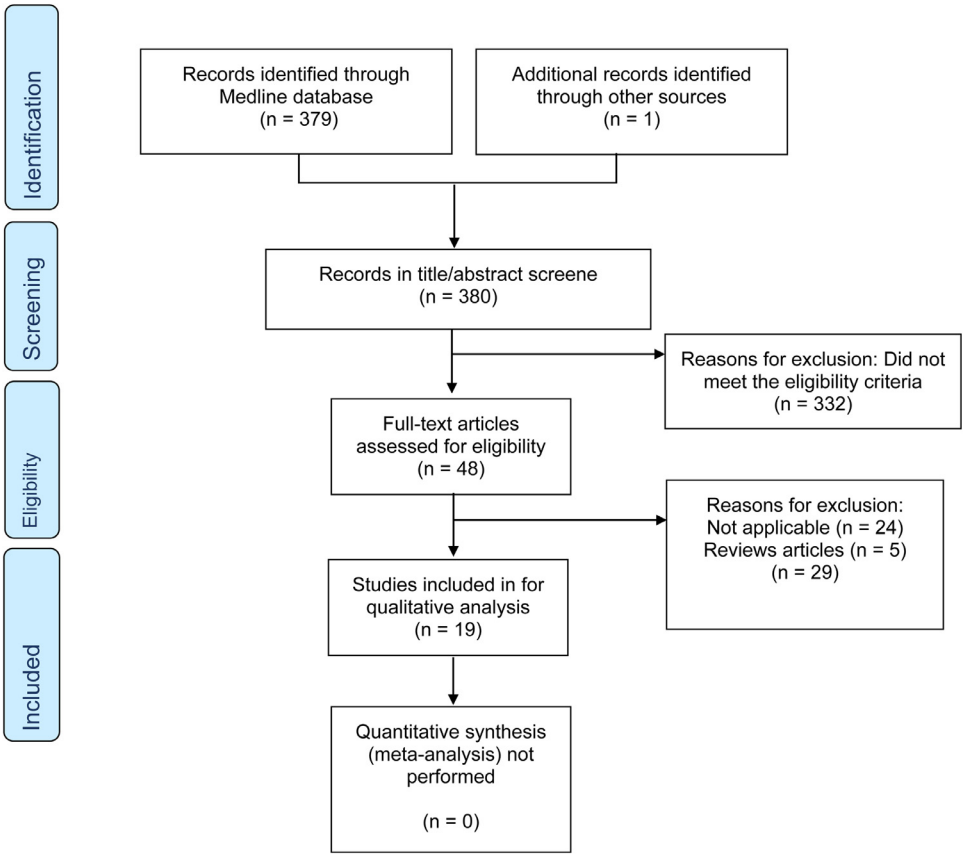

Figure 6. Flow diagram of the search and inclusion process. CNN, convolutional neural networks.

techniques were implemented (34-38,43,44). Transfer learning is a method of pre-training a network on a large unrelated dataset to learn general patterns. Data augmentation uses image manipulation to inflate the amount of available training data. Common manipulations include image flipping, rotating, cropping, and zooming. In combination, transfer learning and data augmentation methodologies yielded accuracies of 78%-91% and an F1 score of 0·85-0·98.

All the studies mentioned above utilized a patch-based classification system of analysis. In contrast, Gao et al. used a holistic approach. This approach evaluated the network's imaging performance of the entire CT slice (38). The holistic approach substantially reduced the network's accuracy to 68·6%. Figure 5 demonstrates the difference between a patch-based approach and a holistic approach. It is important to notice the differences in interpretation between patch-based and holistic metrics. While patch-based metrics provide a localized assessment of the algorithm performance, the holistic metrics are of importance to the clinical usage of the algorithm.

Aliboni et al. developed a CNN algorithm to quantify the different imaging classes in chronic hypersensitivity pneumonitis (cHP) (45). They trained their network using the HUG public dataset. The proposed network was then tested on a prospective data of 27 patients with cHP. Additionally, they demonstrated a correlation between cHP patterns and pulmonary function tests.

Classification of Fibrotic Lung Disease

Idiopathic pulmonary fibrosis (IPF) is a chronic and progressive ILD. It is histopathologically characterized by a usual interstitial pneumonia (UIP) pattern (46). A UIP pattern on HRCT is sufficient for the diagnosis of IPF and obviates the need for lung biopsy (46,47). Four studies have described automated classification of IPF on HRCT (48,49).

Utilizing the 2011 ATS/ERS/JRS/ALAT criteria and the 2018 Fleischner Society criteria (46,47), Walsh et al. developed a model to classify 1307 HRCT images of fibrotic lung disease (48). They compared the algorithm's performance to that of 91 specialized thoracic radiologists. The algorithm's accuracy was greater than the median accuracy of the radiologists (73·3% vs. 70·7%).

Likewise, Christe et al. reviewed 307 HRCT scans to identify the presence of IPF based on the 2018 Fleischner Society criteria (49). Their network's accuracy was similar to that of a single human reader.

Shaish et al. were the first to develop a deep learning model classifying ILD using histopathology as a reference standard instead of radiologists' interpretation (50). Their CNN algorithm yielded an AUC of 74% for histopathological UIP pattern classification. CNN-predicted UIP and physician-predicted UIP showed a moderate agreement with histopathology-proven UIP.

TABLE 1. A summary of the articles in the literature review that applied CNN techniques for ILD image analysis on chest CT

Paper	Year	Contry of origen	journal	Study design	Use of an online database	No. of patients	Analysis method	Reference standard	Performance scores		
									Accuracy	F1 score	Other
Lung Segmentation for ILD											
Pang et al. (29)	2019	China	Bio-computational	Retrospective	HUG database	128	Patch-based	Two radiologists			DSC-89·4%
Park et al. (28)	2019	Republic of Korea	Bio-computational	Retrospective	Proprietary	647	Patch-based	Thoracic radiologist			DSC - 98·8%
Anthimopoulos et al. (27)	2018	Switzerland	Bio-computational	Retrospective	HUG database+ Proprietary	172	Patch-based	Two experienced radiologists	81·8%		
Classification of ILD pattern											
Huang et al. (42)	2020	China	Bio-computational	Retrospective	HUG database	128	Patch-based	Two radiologists		0·98	
Bae et al. (55)	2018	Republic of Korea	Clinical	Retrospective	Proprietary	106	Patch-based	Two thoracic radiologists	89·5%		
Kim et al. (56)	2017	Republic of Korea	Bio-computational	Retrospective	Proprietary	318	Patch-based	Two radiologists	95·1%.		
Shin et al. (38)	2016	USA	Bio-computational	Retrospective	HUG database	128	Patch-based	Two radiologists	91·1%		
Gao et al. (37)	2018	USA	Bio-computational	Retrospective	HUG database	120	Holistic	Two radiologists	87·9%		
Anthimopoulos et al. (33)	2016	Switzerland	Bio-computational	Retrospective	HUG database+ Proprietary	120	Patch-based	Experienced radiologists		0·85	
Wang et al. (40)	2017	China	Bio-computational	Retrospective	HUG database	113	Patch-based	Two radiologists	90·1%		
Hwang et al. (57)	2021	Korea	Clinical	Retrospective	Proprietary	246	Patch-based	Chest radiologists	81.7%*		
Christodoulidis et al. (39)	2016	Switzerland	Bio-computational	Retrospective	HUG database+ Proprietary	120	Patch-based	Experienced radiologists		0·88	
Guo et al. (32)	2018	China	Bio-computational	Retrospective	HUG database	109	Patch-based	Two experienced radiologists		0·98	
Aliboni et al. (44)	2020	Italy	Clinical	Retrospective**	HUG database+ Proprietary	136	Patch-based	Two experienced radiologists	0.85	0.85	
Classification of Pulmonary Fibrosis											
Christe et al. (48)	2019	Switzerland	Clinical	Retrospective	Proprietary	307	Patch-based	Two chest radiologists		0·80	
Walsh et al. (47)	2018	UK, Italy	Clinical	Retrospective	Proprietary	1307	Patch-based	Single chest radiologists	76·4%,		
Shaish et al. (49)	2021	USA	Clinical	Retrospective	Proprietary	301	Patch-based	Expert thoracic radiologist & expert pulmonologist			AUC 74 %
Joyseeree et al. (34)	2019	Switzerland	Bio-computational	Retrospective	HUG database	128	Patch-based	Two radiologists	78·1%		

(*continued on next page*)

TABLE 1. (Continued)

Paper	Year	Contry of origen	journal	Study design	Use of an online database	No. of patients	Analysis method	Reference standard	Performance scores		
									Accuracy	F1 score	Other
Automated identification of ILA											
Bermejo-Peláez et al. (43)	2020	Spain.	Clinical	Retrospective	Proprietary (part of the COPDGene study)	208	Patch-based	Two pulmonologists			Sensitivity -91·4% specificity - 98·1%

Abbreviation: ILD - Interstitial lung disease; ILA- Interstitial lung abnormalities; DSC -dice similarity coefficient; HUG - Hospitals of Geneva.
* For a content-based image retrieval (CBIR) system that searches for images from the database that depict similar imaging appearances.
** The CNN model was tested on prospective data.

Classification of Interstitial Lung Abnormality

Interstitial lung abnormalities (ILAs) are subtle lung parenchyma changes that may precede ILD radiologic abnormalities (51,52). ILAs have been associated with various clinical outcomes, including mortality (53). Bermejo-Peláez (44) developed a CNN to classify the reviewed images into eight ILA patterns. The patterns included normal parenchyma, ground-glass, reticular, nodular, linear scar, subpleural line, para-septal emphysema, and centrilobular emphysema (43). The authors used a dataset of 208 HRCT scans from the COPDGene study (54). The CNN demonstrated a sensitivity and specificity of 91·4% and 98·2%, respectively.

Quality Assessment

The Quality Assessment of Diagnostic Accuracy Studies tool identified 15 studies that were prone to a high risk of bias. Eleven papers failed in data management, as ethical approval was not reported. Selection bias was identified in eight of the studies, where the patient population was not described. An independent test set and validation methodology were used in all of the studies. All but one of the studies utilized an experienced radiologist or pulmonologist to annotate the data. One study used the histopathology report as a reference standard. Scores for each study and a summary of the quality assessment are presented in the (Table 2 and Supplementary Table 1).

DISCUSSION

AI is transforming the field of medical imaging and plays an emerging role in ILD analysis. It is used for the classification, segmentation and quantification of various ILD patterns. It also assists with the diagnosis and early detection of the disease. Several studies compared the performance of CNNs to that of experienced radiologists. They found that CNNs performed equally well, if not better than experts (48,55).

Of the 19 articles that were reviewed, seven were recently published in clinical journals. Research published in biocomputational journals focused on improving CNN algorithms. Most of these papers utilized the publicly available HUG dataset for the classification of ILD patterns. In contrast, research in clinical journals focused on clinical applicability. This is reflected in the choice to concentrate on IPF and ILA, two conditions with a potential to modify the course of the disease.

When examining the quality of the manuscripts in this review, the risk of bias was low in only four of these studies. Ethical approval for most of the studies was not obtained. Additionally, almost half of the studies were prone to selection bias as they did not describe the study population or report inclusion criteria. It is critically important to minimize selection bias within the training dataset in order to develop AI algorithms which can be incorporated into comprehensive health care systems. The risk of bias can be reduced by promoting better cooperation between the algorithm designer and the

TABLE 2. QUADAS-2 risk of bias assessment per clinical application

Application	Paper	Risk of bias				
		Patient selection	Index test	Reference standard	Flow and timing	Data management
Classification of ILD patterns	Huang et al. (42)	✓	✓	✓	✓	✗
	Bae et al (55)	✗	✓	✓	✓	✓
	Guo et al. (32)	✓	✓	✓	✓	✗
	Kim et al. (56)	✗	✓	✓	✓	✓
	Anthimopoulos et al. (33)	✗	✓	✓	✓	✗
	Shin et al. (38)	✓	✓	✓	✓	✗
	Wang et al. (40)	✓	✓	✓	✓	✓
	Christodoulidis et al. (39)	✗	✓	✓	✓	✗
	Gao et al. (37)	✓	✓	✓	✓	✗
	Hwang et al. (57)	✗	✓	✓	✓	✓
	Aliboni et al. (44)	✓	✓	✓	✓	✓
Classification of Pulmonary Fibrosis	Christe et al. (48)	✗	✓	✓	✓	✗
	Walsh et al. (47)	✓	✓	✓	✓	✓
	Joyseeree et al. (34)	✓	✓	✓	✓	✗
	Shaish et al. (49)	✓	✓	✓	✓	✓
Identification of ILA	Bermejo-Peláez et al. (43)	✓	✓	✓	✓	✓
Lung Segmentation for ILD	Pang et al. (29)	✓	✓	✓	✓	✗
	Park et al. (28)	✗	✓	✓	✓	✓
	Anthimopoulos et al. (27)	✗	✓	✓	✓	✗

Abbreviation: ILD - Interstitial lung disease; ILA- Interstitial lung abnormalities

clinician. In this way, the AI developers can become more aware of the importance of ethical approval and patient selection. Additionally, uniform standards must be adopted when performing AI research. Recently, novel AI specific research protocol guidelines are being developed that can be utilized by clinicians and developers for designing AI research and presenting the model results in a uniform manner (56,57).

Although CNNs demonstrate impressive capabilities, difficulties in classification were evident among the studies. In particular, several CNNs misclassified patterns of honeycombing and emphysema. Unbalanced training data that did not contain various patterns of ILD was likely the source of these errors (29,30) . In addition, the classification of ILDs, as evident by the significant inter- and intra-observer variability observed among human readers, is quite challenging. It is not surprising that honeycombing was commonly misidentified as reticular opacity, since both radiologic patterns often contain fibrosis. Similarly, CNNs had difficulty differentiating emphysematous tissue from that of a healthy lung tissue of similar density on HRCT (31). In order to cope with these errors, a more balanced dataset is necessary. More samples of patterns prone to misclassification need to be acquired for the training phase. A training on a larger dataset can contribute to a reduction in the misinterpretation of normal tissue as abnormal. Multi-center collaborations and publicly available image sets may help increase available data for training.

Deep learning algorithms require a large number of samples for training. Furthermore, the annotation of medical data is a time-consuming process that requires a specialist. There are several technological solutions which can be utilized to overcome scarcity in data. These include transfer learning, data augmentation, and semi-supervised learning. In addition, clinical solutions, such as collaboration of international data sources, can be used to generate comprehensive datasets. The public HUG dataset, which was used by more than half of the studies, advanced the evaluation of ILDs by CNNs. Nevertheless, this database is relatively small in terms of deep learning.

The power of neural networks is derived from their ability to continuously learn from added data. The ImageNet challenge used more than a million images. In this review, the largest dataset had 1307 images. Other studies utilized only several hundreds images. Still, with these small datasets, near expert performance was achieved. A larger dataset will possibly have a transformative effect on the field of pulmonology diagnostics.

It should be noted that the described developments are still in the research stage. They need to be further refined in their ability before entering the clinic. Most of the reviewed articles used a patch-based approach. This approach does not accurately reflect the real-world clinical setting. In the clinical setting, the exam interpretation is based on the entire CT scan as well as on the clinical context. This was emphasized in a study that investigated the analysis of an entire CT slice. The authors demonstrated a decrease in the algorithm's performance (39). Future studies should concentrate on this clinically relevant holistic approach.

Most of the reviewed studies did not evaluate their AI solution in the radiology clinical setting. In a recent study, Yacoub et al. demonstrated the importance of assessing the AI-based system in comparison to the radiologist's performance (58). Their results showed that the AI support

platform significantly contributes to clinical radiology reporting. For example, they reported a similar diagnostic accuracy between the AI system and radiology performance in detecting pulmonary emphysema. Future studies should compare AI assessment with radiologists.

It can also be noted that the reviewed studies were retrospective. Collaboration between medical centers should be made in order to design prospective research. Specifically, future research should determine whether CNN can change the course of the disease by affecting treatment decisions and by reducing the need for invasive diagnostic tests.

Our review has several limitations. This study is limited by a small number of samples. High heterogeneity and variability in evaluation metrics were seen between investigated studies. Thus, we could not perform a meta-analysis. Most studies had a high risk of bias. Furthermore, all of the studies were retrospective and assessed technical feasibility of AI for ILD analysis. None of the studies evaluated their algorithm in the clinical setting.

In conclusion, AI has the potential to contribute to the radiologic diagnosis and classification of ILD. However, the accuracy performance is still not satisfactory, and research is limited by a small number of retrospective studies. Hence, the existing published data may not be sufficiently reliable. Only well-designed prospective controlled studies can accurately assess the value of existing AI tools for ILD evaluation.

RESEARCH ETHICS APPROVAL

Human Participants - Does this study involve human participants? No

FUNDING

The authors have not received a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

REFERENCES

1. Crystal RG, Gadek JE, Ferrans VJ, Fulmer JD, Line BR, Hunninghake GW. Interstitial lung disease: current concepts of pathogenesis, staging and therapy. Am J Med 1981; 70(3):542–568.
2. SOCIETY BT, COMMITTEE S. The diagnosis, assessment and treatment of diffuse parenchymal lung disease in adults. Thorax 1999; 54(Suppl 1):S1.
3. Scatarige JC, Diette GB, Haponik EF, Merriman B, Fishman EK. Utility of high-resolution CT for management of diffuse lung disease: results of a survey of US pulmonary physicians. Acad Radiol 2003; 10(2):167–175.
4. Aziz ZA, Wells AU, Hansell DM, et al. HRCT diagnosis of diffuse parenchymal lung disease: inter-observer variation. Thorax 2004; 59(6):506–511.
5. Soffer S, Ben-Cohen A, Shimon O, Amitai MM, Greenspan H, Klang E. Convolutional neural networks for radiologic images: a radiologist's guide. Radiology 2019; 290(3):590–606.
6. Klang E. Deep learning and medical imaging. J Thorac Dis 2018; 10 (3):1325–1328.
7. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 2012; 25:1097–1105.
8. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015; 521 (7553):436.
9. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017; 42:60–88.
10. McBee MP, Awan OA, Colucci AT, et al. Deep learning in radiology. Acad Radiol 2018; 25(11):1472–1480.
11. Klang E, Barash Y, Yehuda Margalit R, et al. Deep learning algorithms for automated detection of Crohn's disease ulcers by video capsule endoscopy. Gastrointest Endosc 2019; 11:606–613.e2.
12. Barash Y, Klang E. Automated quantitative assessment of oncological disease progression using deep learning. Ann Transl Med 2019; 7(Suppl 8): S379.
13. Christopher M, Belghith A, Bowd C, et al. Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. Sci Rep 2018; 8(1):1–13.
14. Yiftach B, Liran A, Shelly S, et al. Ulcer severity grading in video-capsule images of Crohn's disease patients: an ordinal neural network solution. Gastrointest Endosc 2020; 93:187–192.
15. Hosseinzadeh Kassani S, Hosseinzadeh Kassani P. A comparative study of deep learning architectures on melanoma detection. Tissue Cell 2019; 58:76–83.
16. Gonem S, Janssens W, Das N, Topalovic M. Applications of artificial intelligence and machine learning in respiratory medicine. Thorax 2020; 75:695–701.
17. Kulkarni S, Jha S. Artificial intelligence, radiology, and tuberculosis: a review. Acad Radiol 2020; 27(1):71–75.
18. Guo Y, Song Q, Jiang M, et al. Histological subtypes classification of lung cancers on CT images using 3D deep learning and radiomics. Acad Radiol 2020. S1076-6332(20)30360-3.
19. Soffer S, Klang E, Shimon O, et al. Deep learning for wireless capsule endoscopy: a systematic review and meta-analysis. Gastrointest Endosc 2020; 92:831–839.e8.
20. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer; 2015:234–241.
21. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009; 151(4):264–269.
22. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM. Group atP-D. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: The PRISMA-DTA statement. JAMA 2018; 319(4):388–396.
23. Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014; 11(10).
24. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016; 18(12):e323.
25. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011; 155(8):529–536.
26. Fleuren LM, Klausch TLT, Zwager CL, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 2020; 46(3):383–400.
27. Kwong MT, Colopy GW, Weber AM, Ercole A, Bergmann JH. The efficacy and effectiveness of machine learning for weaning in mechanically ventilated patients at the intensive care unit: a systematic review. Bio-Design and Manufacturing 2019; 2(1):31–40.
28. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020: 368. elocation-id = {m689}.
29. Anthimopoulos M, Christodoulidis S, Ebner L, Geiser T, Christe A, Mougiakakou S. Semantic segmentation of pathological lung tissue with dilated fully convolutional networks. IEEE J Biomed Health Inform 2019; 23(2):714–722.
30. Park B, Park H, Lee SM, Seo JB, Kim N. Lung segmentation on HRCT and volumetric CT for diffuse interstitial lung disease using deep convolutional neural networks. 2019; 32(6):1019-26.
31. Pang T, Guo S, Zhang X. Automatic lung segmentation based on texture and deep features of HRCT images with interstitial lung disease. 2019; 2019:2045432.
32. Depeursinge A, Vargas A, Platon A, Geissbuhler A, Poletti P-A, Müller H. Building a reference multimedia database for interstitial lung diseases. Comput Med Imaging Graph 2012; 36(3):227–238.

33. Guo W, Xu Z, Zhang H. Interstitial lung disease classification using improved DenseNet. Multimed Tools Appl 2019; 78(21):30615–30626.
34. Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. IEEE Trans Med Imaging 2016; 35(5):1207–1216.
35. Joyseeree R, Otálora S, Müller H, Depeursinge A. Fusing learned representations from Riesz filters and deep CNN for lung tissue classification. Med Image Anal 2019; 56:172–183.
36. Kim GB, Jung KH, Lee Y, et al. Comparison of shallow and deep learning methods on classifying the regional pattern of diffuse lung disease. J Digit Imaging 2018; 31(4):415–424.
37. Bae HJ, Kim CW, Kim N, et al. A Perlin noise-based augmentation strategy for deep learning with small data samples of HRCT images. Sci Rep 2018; 8(1):17687.
38. Gao M, Bagci U, Lu L, et al. Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks. Comput Methods Biomech Biomed Eng Imaging Vis 2018; 6(1):1–6.
39. Shin HC, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016; 35(5):1285–1298.
40. Christodoulidis S, Anthimopoulos M, Ebner L, Christe A, Mougiakakou S. Multisource transfer learning with convolutional neural networks for lung pattern analysis. IEEE J Biomed Health Inform 2017; 21(1):76–84.
41. Wang Q, Zheng Y, Yang G, Jin W, Chen X, Yin Y. Multiscale rotation-invariant convolutional neural networks for lung texture classification. IEEE J Biomed Health Inform 2018; 22(1):184–195.
42. Huang S, Lee F, Miao R, Si Q, Lu C, Chen Q. A deep convolutional neural network architecture for interstitial lung disease pattern classification. Med Biol Eng Comput 2020; 58(4):725–737.
43. Huang S, Lee F, Miao R, Si Q, Lu C, Chen Q. A deep convolutional neural network architecture for interstitial lung disease pattern classification. Med Biol Eng Comput 2020; 58(4):725–737.
44. Bermejo-Peláez D, Ash SY, Washko GR, San José Estépar R, Ledesma-Carbayo MJ. Classification of interstitial lung abnormality patterns with an ensemble of deep convolutional neural networks. Sci Rep 2020; 10 (1):338.
45. Aliboni L, Dias OM, Pennati F, et al. Quantitative CT analysis in chronic hypersensitivity pneumonitis: a convolutional neural network approach. Acad Radiol 2020. S1076-6332(20)30596-1.
46. Lynch DA, Sverzellati N, Travis WD, et al. Diagnostic criteria for idiopathic pulmonary fibrosis: a Fleischner Society White Paper. Lancet Respir Med 2018; 6(2):138–153.
47. Raghu G, Remy-Jardin M, Myers JL, et al. Diagnosis of idiopathic pulmonary fibrosis. An official ATS/ERS/JRS/ALAT clinical practice guideline. Am J Respir Crit Care Med 2018; 198(5):e44–e68.
48. Walsh SLF, Calandriello L, Silva M, Sverzellati N. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. Lancet Respir Med 2018; 6(11):837–845.
49. Christe A, Peters AA, Drakopoulos D, et al. Computer-aided diagnosis of pulmonary fibrosis using deep learning and CT images. Invest Radiol 2019; 54(10):627–632.
50. 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(1):51–59.
51. Washko GR, Hunninghake GM, Fernandez IE, et al. Lung volumes and emphysema in smokers with interstitial lung abnormalities. N Engl J Med 2011; 364(10):897–906.
52. Lederer DJ, Enright PL, Kawut SM, et al. Cigarette smoking is associated with subclinical parenchymal lung disease: the Multi-Ethnic Study of Atherosclerosis (MESA)–lung study. Am J Respir Crit Care Med 2009; 180 (5):407–414.
53. Putman RK, Hatabu H, Araki T, et al. Association between interstitial lung abnormalities and all-cause mortality. JAMA 2016; 315(7):672–681.
54. Regan EA, Hokanson JE, Murphy JR, et al. Genetic epidemiology of COPD (COPDGene) study design. COPD 2011; 7(1):32–43.
55. Christe A, Christodoulidis S, Stathopoulou T, Mougiakakou S, Ebner L. Computer-aided diagnosis of pulmonary fibrosis using deep learning and CT images. J Thorac Imaging 2019; 34(4):W78.
56. Watkinson P, Clifton D, Collins G, McCulloch P, Morgan L, Group D-AS. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. Nat Med 2021; 27 (2021):186–187.
57. Mongan J, Moy L, Kahn Jr CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020; 2:e200029.
58. Yacoub B, Kabakus IM, Schoepf UJ, et al. Performance of an artificial intelligence-based platform against clinical radiology reports for the evaluation of noncontrast chest CT. Acad Radiol 2021. S1076-6332(21) 00070-2.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.acra.2021.05.014.