# Deep learning for pneumothorax diagnosis: a systematic review and meta-analysis

Takahiro Sugibayashi[1], Shannon L. Walston[1], Toshimasa Matsumoto[1,2], Yasuhito Mitsuyama[1], Yukio Miki[1] and Daiju Ueda [1,2]

[1]Department of Diagnostic and Interventional Radiology, Graduate School of Medicine, Osaka Metropolitan University, Osaka, Japan. [2]Smart Life Science Lab, Center for Health Science Innovation, Osaka Metropolitan University, Osaka, Japan.

Corresponding author: Daiju Ueda (ai.labo.ocu@gmail.com)

## Abstract

*Background* Deep learning (DL), a subset of artificial intelligence (AI), has been applied to pneumothorax diagnosis to aid physician diagnosis, but no meta-analysis has been performed.

*Methods* A search of multiple electronic databases through September 2022 was performed to identify studies that applied DL for pneumothorax diagnosis using imaging. Meta-analysis *via* a hierarchical model to calculate the summary area under the curve (AUC) and pooled sensitivity and specificity for both DL and physicians was performed. Risk of bias was assessed using a modified Prediction Model Study Risk of Bias Assessment Tool.

*Results* In 56 of the 63 primary studies, pneumothorax was identified from chest radiography. The total AUC was 0.97 (95% CI 0.96–0.98) for both DL and physicians. The total pooled sensitivity was 84% (95% CI 79–89%) for DL and 85% (95% CI 73–92%) for physicians and the pooled specificity was 96% (95% CI 94–98%) for DL and 98% (95% CI 95–99%) for physicians. More than half of the original studies (57%) had a high risk of bias.

*Conclusions* Our review found the diagnostic performance of DL models was similar to that of physicians, although the majority of studies had a high risk of bias. Further pneumothorax AI research is needed.

## Introduction

Pneumothorax is defined as the presence of air in the pleural space, *i.e.* the space between the lungs and the chest wall [1, 2]. Pneumothorax is a common disease in the population, with an incidence of primary spontaneous pneumothorax of 7.4/100 000 per year in men and 1.2/100 000 per year in women and an incidence of secondary spontaneous pneumothorax of 6.3/100 000 per year in men and 2.0/100 000 per year in women [3]. In contrast to the benign clinical course of primary spontaneous pneumothorax, secondary spontaneous pneumothorax is a potentially life-threatening event [2]. Additionally, the recurrence rate is high: ~30% in primary spontaneous pneumothorax and ~40% in secondary spontaneous pneumothorax [4–7]. Pneumothorax is one of the conditions the American College of Radiology recommends should be communicated to the physician within minutes to avoid patient decompensation [8].

Chest radiography is the simplest and most common examination [9, 10], and pneumothorax is usually diagnosed in conjunction with the patient's history and clinical presentation [2]. Although errors or delays in diagnosis can harm the patient, the signs of pneumothorax on chest radiography are subtle and up to 20% of occult pneumothoraces are missed on examination [11]. One reason for this is said to be that the workload far exceeds the number of radiologists [12, 13], and computer-based approaches have been developed to assist physicians in their daily work and are expected to be an approach to prevent missed cases.

Deep learning (DL) is one of the fields of artificial intelligence (AI) and it has improved tremendously in the field of medical imaging [14, 15], and with it, the number of certified medical devices that can be used in clinical practice has been increasing [16]. It is formally defined as a computational model that consists of multiple processing layers and learns representations of data with multiple levels of abstraction [17]. In fact, given raw data, DL develops the representations needed for pattern recognition on its own and does not require domain expertise to design data structures or feature extractors [14, 17]. This feature of DL allows it to learn the features important for classification on its own, rather than being directed by a human. Thus, DL requires careful bias assessment and accumulation of original articles for model training and evaluation [18].

This study is a systematic review and meta-analysis of 63 studies on the application of DL to pneumothorax diagnosis, comparing the diagnostic performance of DL and physicians in each modality. Studies in which physicians' pneumothorax diagnosis performance is supported by DL are examined separately. At this time, there are no meta-analyses of DL diagnosis for pneumothorax.

## Methods

### Study registration and guidelines

This systematic review was prospectively registered with PROSPERO (CRD42022351985). Our study followed the guidelines of the Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies (PRISMA-DTA) [19, 20]. Two authors performed all screening, data collection, applicability assessments and bias assessments in duplicate (T.S. and D.U.), and a third independent reviewer was consulted in the event of a disagreement (T.M.).

### Search strategy and study selection

The search strategy for identifying studies that developed and/or validated a DL model for the purposes of pneumothorax diagnosis was developed with an information specialist. The search strategy was as follows: original articles that included the words or variations of "artificial intelligence" or "deep learning" or "neural networks" and also the word "pneumothorax" were included. Peer-reviewed studies in any language from inception to September 2022 from the following databases were evaluated: MEDLINE, Scopus, Web of Science, Cochrane Central Register of Controlled Trials (CENTRAL), and the Institute of Electrical and Electronics Engineers and the Institution of Engineering and Technology IEEE Xplore. The titles and abstracts were screened prior to full-text screening. Studies were included if they were: primary research studies of pneumothorax diagnosis in humans that developed and/or validated a DL model. Any target population, study setting or comparator group was eligible. Studies were excluded if they were: conference abstracts or proceedings, letters to the editor, review articles, or segmentation or detection only studies. Excluded studies, including the reason for exclusion, were recorded in a PRISMA flow diagram (figure 1) [20].

### Data extraction

We extracted information including study design, sample size, comparator groups and numerical results into a predefined data sheet. Contingency tables were constructed using the available diagnostic performance information for each model. These were used to calculate summary area under the curve (AUC), sensitivity and specificity. All available contingency tables were included in the meta-analysis. The datasets involved in the development of a model were defined as the training set (for training the model), tuning set (for tuning hyperparameters) and validation test set (for estimating the performance of the model) [21].

### Statistical analysis

We estimated the diagnostic performance of both the DL model and physicians by carrying out a meta-analysis with random effects of studies providing both internal and external validation contingency tables [22]. These contingency tables were used to construct hierarchical summary receiver operating characteristic (ROC) curves and to calculate pooled sensitivities and specificities, with the anticipation of a high level of heterogeneity [23]. Between-study heterogeneity was represented using the 95% prediction region of the hierarchical summary ROC curves. Statistical significance was defined as a p-value of <0.05. All calculations were performed using R version 4.0.0 with the metafor and meta4diag libraries [24].

### Quality assessment

The Prediction Model Study Risk of Bias Assessment Tool (PROBAST) was used to assess bias and applicability of the included studies [18]. This tool evaluates bias across four domains (participants, predictors, outcomes and analysis) and then these domains are combined into an overall assessment. Our assessment of bias and applicability in the first domain was based on both the images used to develop the
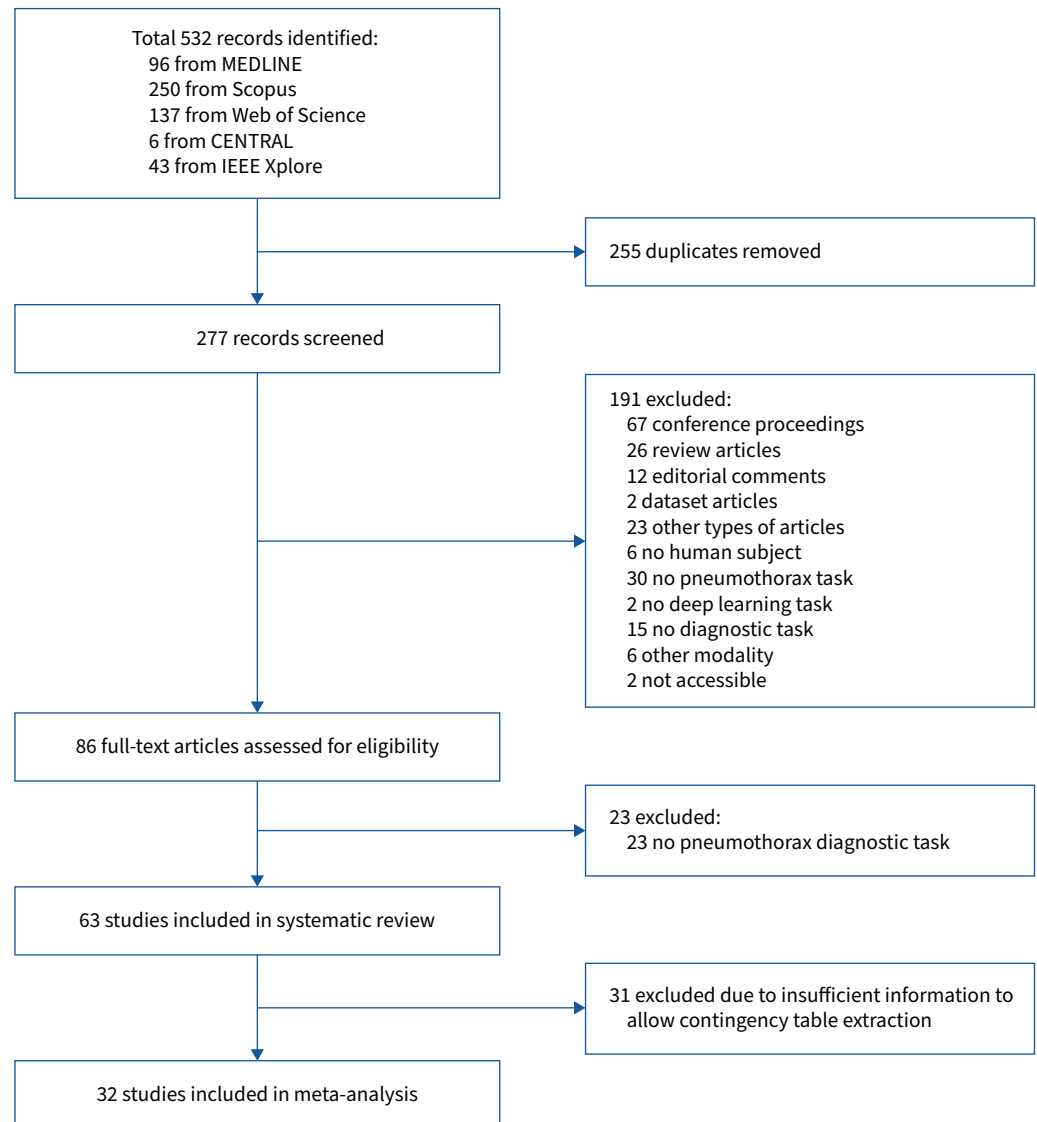
**FIGURE 1** Eligibility criteria. CENTRAL: Cochrane Central Register of Controlled Trials.

models and the patient population the models were tested on. We did not include domain 2 (predictors) in the assessment of bias or applicability. Details of modifications made to PROBAST are provided in supplementary table S1.

### Publication bias
Publication bias was assessed using the effective sample size funnel plot described by EGGER *et al.* [25].

### Results
#### Study selection and characteristics
We identified 532 studies, of which 255 were duplicates. After screening, 63 studies were included in the systematic review and 32 studies were included in the meta-analysis (figure 1 and table 1). Among the 63 studies, 56 studies identified pneumothorax on chest radiography [26–81], four studies on computed tomography [82–85], one study on ECG [86], one study used chest radiography and photography using a smartphone [87], and one study used chest radiography and tabular data [88]. Six studies developed and internally tuned DLs [37, 52, 63, 67, 74, 76], 25 studies also internally tested their DLs [32, 33, 35, 38, 40, 41, 43, 45, 47, 48, 50, 55, 60, 65, 69, 70, 73, 75, 79–83, 85, 86] and 32 studies externally tested the DLs [26–31, 34, 36, 39, 42, 44, 46, 49, 51, 53, 54, 56–59, 61, 62, 64, 66, 68, 71, 72, 77, 78, 84, 87, 88].

**TABLE 1** Study characteristics

| Study | Study type | | Imaging modality | Comparison group | Images per set, n | | | | Reference standard | Meta-analysis |
|---|---|---|---|---|---|---|---|---|---|---|
| | Development | Test | | | Training | Tuning | Internal testing | External testing | | |
| Lee, 2022 [68] | | External | CXR | Physician, physician+DL | 454 000 | | | 500/550 | Expert consensus | + |
| Rudolph, 2022 [39] | | External | CXR | Physician | | | | 563/6258 | Expert consensus | + |
| Park, 2022 [77] | + | External | CXR | | 12 089 | | | 1220 | Expert consensus | |
| Shin, 2022 [58] | | External | CXR | | | | | 2273 | Expert opinion | + |
| Hallinan, 2022 [51] | + | Internal/external | CXR | | 80 921 | 4260 | 25 596/24 709 | 525 | NLP from CXR report+additional expert opinion/NLP from CXR report/expert consensus | + |
| Gipson, 2022 [56] | | External | CXR | Report | Over 500 000 | | | 1400 | Expert consensus | + |
| Shamrat, 2022 [69] | + | Internal | CXR | | NR | NR | 16 000 | | Various | |
| Hong, 2022 [44] | | External | CXR | Physician, physician+DL | | | | 676 | Expert consensus | + |
| Jin, 2022 [57] | | External | CXR | Physician, physician+DL | | | | 2536/1470/2000 | Expert consensus | |
| Kim, 2022 [45] | + | Internal | CXR | | 8000/29 282 | 2000/2000 | 4000 | | NR, NLP from CXR report/NR | + |
| Wang, 2022 [33] | + | Internal | CXR | | 10 675 | | 1372/3205 | | Expert consensus | + |
| Rudolph, 2022 [28] | + | External | CXR | Physician | 42 942 | 1634 | | 563 | Expert consensus | + |
| Mangalmurti, 2022 [75] | + | Internal | CXR | | 5250 | 750 | 1500 | | NLP from CXR report/NR | + |
| Kakkar, 2022 [67] | + | | CXR | | 84 090 | 28 030 | | | NLP from CXR report | |
| Iqbal, 2022 [65] | + | Internal# | CXR | | 10 675 +2592 | 1372 +664 | 664/664/1372 | | NLP from CXR report+expert consensus | |
| Feng, 2022 [30] | + | Internal/external | CXR | | 15 371 | 1952 | 1914/1607 | 9273 | Expert consensus | + |
| Thian, 2022 [61] | + | External | CXR | | 233 163 | 58 291 | | 525 | NLP from CXR report/additional expert consensus | + |
| Lee, 2022 [86] | + | Internal | ECG | Physician | 99 663 | | 32 464/100 | | Pneumothorax treated invasively | |
| Tian, 2022 [48] | + | Internal | CXR | | 107 442 | 1284 | 1684 | | Expert consensus+NLP from CXR report | + |
| Seah, 2021 [59] | | External | CXR | | | | | 4568 | Expert consensus+NLP from CXR report | + |
| Kuo, 2021 [87] | + | Internal/external | CXR+ photography | | 189 892/1337/1337/ 189 892/248 263/ 1759/1759/248 263 | | 1759/1337 | 1759/1337 | NLP from CXR report | + |
| Hong, 2021 [70] | + | Internal | CXR | | 8000 +45 866 | 2000 +2000 | 4000 | | NLP from CXR report+NR/NR | + |
| Rueckel, 2021 [73] | + | Internal | CXR | | 75 067 | | 3062 | | NLP from CXR report+expert consensus | + |
| Cho, 2021 [71] | + | Internal/external | CXR | | 6683 | 949 | 1902 | 1210 | Expert consensus with CT | + |

Continued

**TABLE 1** Continued

| Study | Study type | | Imaging modality | Comparison group | Images per set, n | | | | Reference standard | Meta-analysis |
|---|---|---|---|---|---|---|---|---|---|---|
| | Development | Test | | | Training | Tuning | Internal testing | External testing | | |
| NIEHUES, 2021 [49] | + | Internal/external | CXR | | 17 265 | 513 | 583 | NR | Expert consensus with CT/NLP from CXR report | |
| KAO, 2021 [31] | + | External | CXR | Physician | 1258 | 314 | | 3739 | Expert consensus (technologist+radiology resident) | + |
| MOSQUERA, 2021 [36] | + | Internal/external | CXR | | 8540 | 2135 | 4376 | 1064 | Expert consensus/NLP from CXR report/expert opinion with CXR report | + |
| WANG, 2021 [62] | + | External | CXR | | 33 696 | 601 | | 442 | NR+NLP from CXR report/NR | |
| NAM, 2021 [53] | + | Internal/external | CXR | Physician, physician+DL | 143 768 | 2523 | 190 | 673/202 | Expert consensus/expert opinion with CT/NLP+additional expert consensus | + |
| CHOI, 2021 [64] | | External | CXR | Physician, physician+DL | | | | 244 | Expert consensus | |
| LYU, 2021 [84] | | External | CT | | | | | 403 | Expert consensus | |
| KIM, 2021 [72] | | External | CXR | | | | | 1694/1858/2335 | Expert consensus | |
| LI, 2021 [88] | + | Internal/external | CXR+tabular | Physician | 6239 | 847 | 1715 | 1804 | NR/NLP from CXR report | |
| THIAN, 2021 [46] | + | External | CXR | | 163 007 | 17 121 | | 511/321/538/519/517/525 | NLP from CXR report/expert consensus+NLP+additional expert consensus | + |
| ABEDALLA, 2021 [37] | + | | CXR | | 10 842 | 1205 | | | Expert consensus | |
| SUNG, 2021 [26] | | External | CXR | Physician, physician+DL | | | | 228 | Expert consensus with CT | |
| DRAELOS, 2021 [82] | + | Internal | CT | | 25 355 | 2085 | 7209 | | NLP (99% agreement with a radiologist) | |
| WANG, 2021 [81] | + | Internal | CXR | | 86 524 | | 26 080 | | NLP from CXR report | |
| RUECKEL, 2020 [66] | | External | CXR | | | | | 6434 | Consensus (medical student+radiology resident) | |
| CHEN, 2020 [55] | + | Internal | CXR | | 2137 | | 531 | | Expert consensus | + |
| RÖHRICH, 2020 [85] | + | Internal | CT | | 27 | 7 | 567 | | Expert opinion | |
| WANG, 2020 [38] | + | Internal | CXR | | 9329 | 1346 | 1372 | | Expert consensus | + |
| YI, 2020 [34] | + | External | CXR | Physician | 111 518 | | | 602 | NLP from CXR report/NLP+additional expert opinion | + |
| HWANG, 2020 [42] | | External | CXR | Report | | 488 | | 602/314/1055/388/170 | Expert consensus+expert opinion | + |
| KITAMURA, 2020 [76] | + | | CXR | | 841 | 1335 | | | Expert opinion | |
| PARK, 2020 [47] | + | Internal | CXR | Physician | 14 048 | 1561 | 200 | | Expert consensus+expert opinion with CT | |
| WANG, 2020 [79] | + | Internal | CXR | | 86 524 | | 25 596 | | NLP from CXR report | |
| ELKINS, 2020 [52] | + | | CXR | | 41 408 | 10 351 | | | NLP from CXR report | |

| Study | Study type | | Imaging modality | Comparison group | Images per set, n | | | | Reference standard | Meta-analysis |
|---|---|---|---|---|---|---|---|---|---|---|
| | Development | Test | | | Training | Tuning | Internal testing | External testing | | |
| Lai, 2023 [78] | + | External | CXR | | 2051/4379 | | | 175 | NLP from CXR report/NR | + |
| Majkowska, 2020 [35] | + | Internal | CXR | Physician | 589 153 +68 801 | 1991 +2412 | 1962/1818 | | Expert consensus/NLP from CXR report | + |
| Li, 2019 [83] | + | Internal | CT | Physician | 76 | 4 | 200 | | Expert consensus | |
| Park, 2019 [27] | + | Internal/ external | CXR | | 12 230 | | 1379 | 1329 | Expert consensus | + |
| Liang, 2019 [50] | + | Internal | CXR | | 86 524 | | 25 596 | | Expert consensus+NLP from CXR report | |
| Hwang, 2019 [54] | + | Internal/ external | CXR | Physician | 87 695 | 1050 | 1089 | 200/245/190/ 184/196 | Expert consensus | + |
| Taylor, 2018 [29] | + | Internal/ external | CXR | | 9309 | 1993 | 1701 | 112 120 | Expert consensus/NLP from CXR report | + |
| Rajpurkar, 2018 [43] | + | Internal | CXR | Physician | 98 637 | 6351 | 420 | | NLP from CXR report/NLP +additional expert consensus | + |
| Cicero, 2017 [80] | + | Internal | CXR | | 34 897 | | 3204 | | NLP+additional expert consensus | + |
| Chen, 2019 [60] | + | Internal | CXR | | 86 524/89 687 | | 25 596/ 22 433 | | NLP from CXR report | |
| Zhou, 2021 [40] | + | Internal | CXR | | NR | NR | NR | | Expert consensus+NLP from CXR report | |
| Luo, 2022 [74] | + | | CXR | | 2530+2530 | | | | NR | |
| Lin, 2020 [63] | + | | CXR | | 270 | | | | Various | |
| Haq, 2021 [41] | + | Internal | CXR | | 12 000 | | 1000 | | NLP from CXR report | |
| Wang, 2020 [32] | + | Internal | CXR | | 7070 | 1768 | 2213 | | Consensus (medical student +radiology resident) | + |

If there is more than one trial, data are expressed as n/n. CXR: chest radiography; DL: deep learning; NLP: natural language processing; CT: computed tomography; NR: not reported. [#]: data overlap.

Five studies compared the performance of DL with physicians: two studies compared DL with experts (not a resident or technologist) [42, 83], two studies compared DL with both experts and non-experts [43, 68] and one study compared DL with non-experts only [34]. Two studies compared the performance of DL with radiology reports written in daily clinical practice [42, 56]. Two studies included physician performance with and without DL assistance as a comparison group [44, 68]. Detailed physician characteristics are shown in supplementary table S4.

As for model development, to generate a reference standard for image labelling, 18 studies used expert consensus [27–33, 35–38, 49, 53–55, 71, 77, 83], two relied on the opinion of a single expert reader [76, 85], 16 used pre-existing radiological reports or other imaging modalities [34, 41, 43, 45, 46, 52, 60, 61, 67, 75, 78–82, 87], one study defined their reference standard as surgical confirmation (indicated for surgery) [86], 11 studies used mixed methods (any combination of the aforementioned) [40, 47, 48, 50, 51, 62, 63, 65, 69, 70, 73] and two studies did not report how their reference standard was generated [74, 88]. As for model testing, to generate a reference standard for image labelling, 26 studies used expert consensus [26–28, 30–33, 38, 39, 44, 51, 54–57, 61, 64, 66, 68, 71–73, 77, 80, 83, 84], two relied on the opinion of a single expert reader [58, 85], 11 used pre-existing radiological reports or other imaging modalities [35, 40, 41, 48, 50, 60, 79, 81, 82, 87, 88], one study defined their reference standard as surgical confirmation (indicated for surgery) [86], 12 studies used mixed methods (any combination of the aforementioned) [29, 34, 36, 42, 43, 46, 47, 49, 53, 59, 65, 69] and five studies did not report how their reference standard was generated [45, 62, 70, 75, 78].

### Study participants
There was large variation in the number of participants represented by each dataset (median (interquartile range (IQR)) 5288 (516–30 805); range 100–538 390) (supplementary table S2). The proportion of participants with pneumothorax in each dataset also ranged widely (median (IQR) 17.2% (10.8–25.0%)). 23 studies did not describe the sex of the study participants [27, 31–33, 36–38, 55, 59, 62, 65, 69–71, 73, 76–78, 81, 82, 86–88] and 24 studies did not include age information [27, 31–33, 36–38, 55, 59, 62, 65, 69–71, 73, 74, 76–78, 81, 82, 86–88]. Detailed dataset characteristics are shown in supplementary table S2.

### Model development
The size of the training (median (IQR) 17 265 (8540–86 524)), tuning (median (IQR) 1598 (924–3468)) and test (median (IQR) 1684 (575–3107)) datasets at the patient level varied widely (table 1). Two out of 50 (4%) studies that developed a model did not report the size of each dataset separately [40, 69]. In studies that performed external model validation, the median dataset size was 1137 (range 175–112 120). 17 studies included localisation of pneumothorax in model output to improve end-user interpretability [26–28, 30–33, 36–38, 40, 47, 56, 59, 68, 84, 85]. Detailed DL characteristics are shown in supplementary table S3.

### Quality assessment
PROBAST assessment led to an overall rating of 36 (57%) studies as high risk of bias (figure 2). The main contributing factors to this assessment were studies that did not perform external validation or internally validated models with small sample sizes. Five (8%) studies were judged to be at high risk of bias in the participant domain because of inclusion and exclusion criteria.

### Meta-analysis
We extracted 89 contingency tables from 32 studies that provided sufficient information to calculate contingency tables for pneumothorax classification [27–36, 38, 39, 42–46, 48, 51, 53–56, 58, 61, 68, 70, 71, 75, 78, 80, 87]. 68 contingency tables were extracted for reported DL performance and 21 contingency tables were extracted for physician performance. Hierarchical summary ROC curves from the studies evaluating DL or physician performance of all studies are included in figure 3. The total AUC was 0.97 (95% CI 0.96–0.98) for DL and 0.97 (95% CI 0.96–0.98) for physicians. The total pooled sensitivity was 84% (95% CI 79–89%) for DL and 85% (95% CI 73–92%) for physicians and the pooled specificity was 96% (95% CI 94–98%) for DL and 98% (95% CI 95–99%) for physicians (table 2). Two studies reported physician performance with DL assistance and one study showed no significant difference with respect to specificity, but a moderate increase in sensitivity and an increase in accuracy. The other study showed no significant difference with respect to sensitivity or specificity, but a slight increase in accuracy [44, 68]. Accuracy, sensitivity and specificity changed from 92–99% to 97–99%, 67–94% to 85–96% and 100% to 99–100% before and after the use of DL, respectively (table 3).
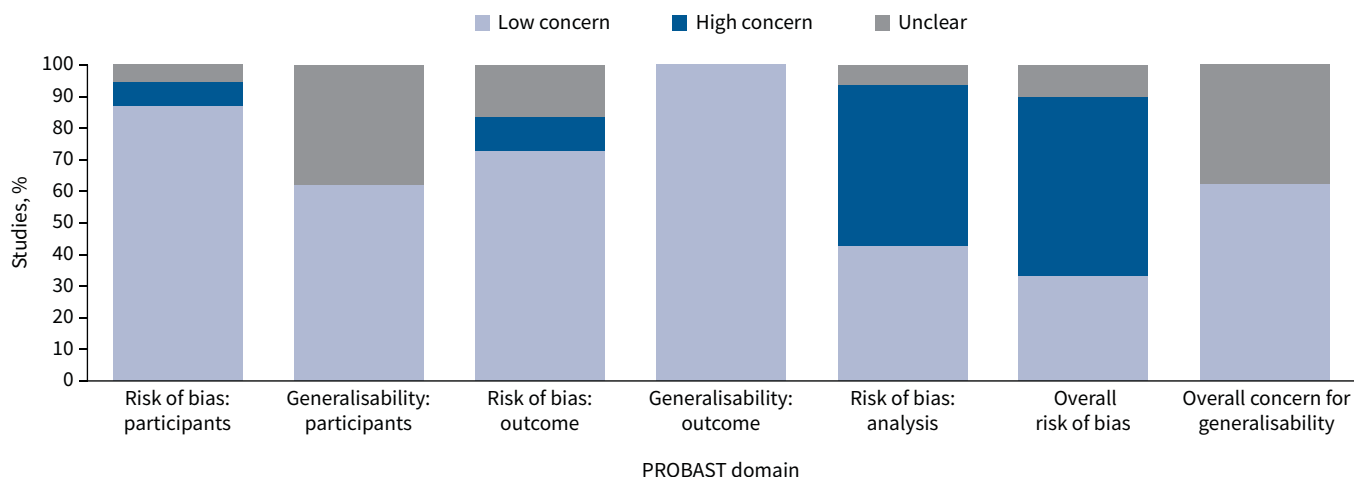
**FIGURE 2** Summary of Prediction Model Study Risk of Bias Assessment Tool (PROBAST) risk of bias.

#### Publication bias

We assessed publication bias by regression analysis of the funnel plot to quantify asymmetry (supplementary figure S1) and it suggested a high risk of publication bias (p<0.05).

### Discussion

In our meta-analysis of DL for pneumothorax diagnosis, DL and physician competence were comparable. The total AUC was 0.97 (95% CI 0.96–0.98) for DL and 0.97 (95% CI 0.96–0.98) for physicians. The total pooled sensitivity was 84% (95% CI 79–89%) for DL and 85% (95% CI 73–92%) for physicians and the pooled specificity was 96% (95% CI 94–98%) for DL and 98% (95% CI 95–99%) for physicians. To the best of our knowledge, this article is the first systematic review and meta-analysis of pneumothorax diagnostic DL.

We found data investigating two possible clinical uses of diagnostic DL for pneumothorax during our review: one is to use DL for triage and the other is to use DL as a second opinion. Although pneumothorax is regularly diagnosed in patients presenting to the emergency department, it is detected in a
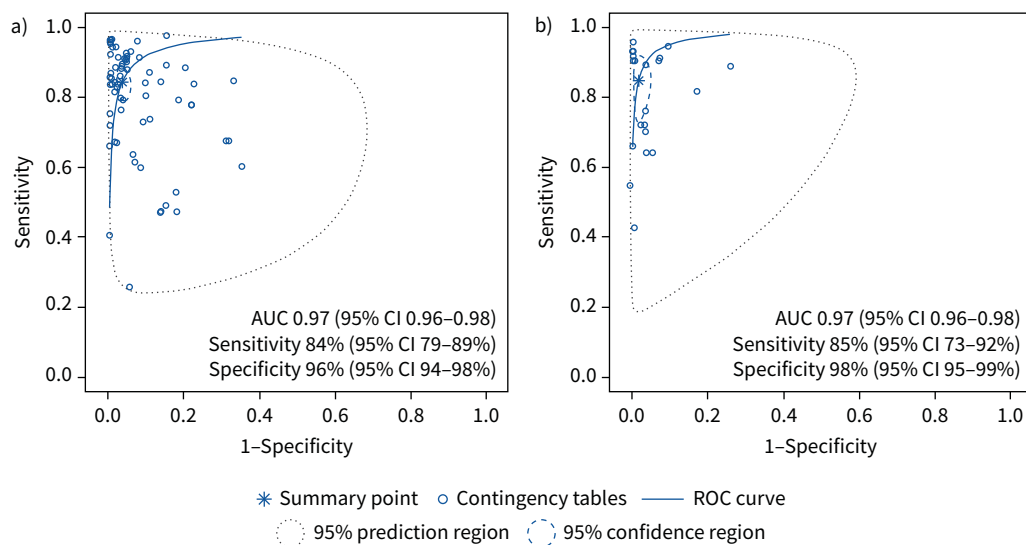


**FIGURE 3** Hierarchical summary receiver operating characteristic (ROC) curves for all studies: a) deep learning models (68 tables) and b) physicians (21 tables). AUC: area under the curve.

**TABLE 2** Pooled metrics in meta-analysis

|  | Sensitivity, % (95% CI) | Specificity, % (95% CI) | AUC (95% CI) | Contingency tables, n |
|---|---|---|---|---|
| **Deep learning, all studies** | 84 (79–89) | 96 (94–98) | 0.97 (0.96–0.98) | 68 |
| **Physician, all studies** | 85 (73–92) | 98 (95–99) | 0.97 (0.96–0.98) | 21 |

AUC: area under the curve.

relatively low number of all radiographs performed. Hence, an easy and accurate screening tool is needed which may help prioritise patients coming to hospitals. In fact, one such study reports a reduction in the reporting delay for pneumothorax [30] and two studies reported that reading times were shorter with AI assisting the physician than with the physician alone [26, 53]. Additionally, in intensive care units, chest radiographs are frequently taken and their reading is often labour intensive. The support of AI is expected to both improve the speed of reading and reduce the total workload [89]. Two papers reported data on the use of DL to complement physicians' decision making [44, 68]. The small number precludes meta-analysis, although one study showed no significant difference with respect to specificity, but a moderate increase in sensitivity and an increase in accuracy; the other study showed no significant difference with respect to sensitivity or specificity (table 3). Although further data related to the performance of physicians supported with DL is required, these two works support the results obtained from this meta-analysis. There is one AI report of chest ultrasound, which is considered more sensitive than chest radiography, and this is an area of promising future research [90]. AI models at present are most useful as a screening tool to determine the presence or absence of pneumothorax. It does not incorporate individual patient's circumstances and other medical conditions in the making of management decisions, especially regarding treatment and follow-up, as physicians currently have to do. Whether additional AIs to measure detailed features (*e.g.* size, evidence of tension, *etc.*) of the pneumothorax are warranted, and how they could best integrate patients' clinical details, will be subjects for future studies.

Confounding factors in images can create bias in diagnostic imaging DL. About 10% (six out of 63) of the articles included in this study mentioned chest tubes as a confounding factor, indicating that DL can recognise tubes and that this can be a strong bias. In other words, a DL may recognise a chest drain as a therapeutic intervention for pneumothorax and diagnose pneumothorax. Although physicians may also use such medical device information to suspect the presence of pneumothorax, a DL with such biases may have poor performance for diagnosing pneumothorax prior to the intervention. Although the impact of such confounding factors may be small or large, their impact should be taken into account when evaluating the model for clinical use to prevent any detriment to the patient. There is a study which reported that the influence of such confounding factors can be avoided when training DL by using annotation of the pneumothorax cavity [73]. External devices (*e.g.* chest tubes, central lines or indwelling pleural catheters) and patient features (*e.g.* skin folds and thickened pleura) may impact the results. These aspects need to be investigated in future research.

In this study, we found that the majority of the included articles were at high risk of bias according to PROBAST [18]. One reason for this is that medical DL research is at the intersection between medicine and engineering, each with different concepts. There were both medical and engineering papers in our included articles. PROBAST is only one method of evaluating bias from a medical perspective. Of course, for medical researchers, medical DL must first adhere to the "do no harm" principle for patients [91]. Therefore, it is important to evaluate medical DL in various validation settings and this should be an

**TABLE 3** Performance summary of the study physicians with deep learning (DL)

| Study | Sensitivity, % (95% CI) | Specificity, % (95% CI) | Accuracy, % (95% CI) |
|---|---|---|---|
| LEE, 2022 [68] | | | |
| Alone | 94.3 (91.4–96.4) | 99.5 (99.1–99.7) | 98.8 (98.3–99.1) |
| With DL | 95.6 (93.0–97.5) | 99.7 (99.4–99.9) | 99.2 (98.8–99.5) |
| HONG, 2022 [44] | | | |
| Alone | 67.1 (59.6–74.6) | 99.6 (99.1–100.0) | 92.3 (90.3–94.3) |
| With DL | 85.4 (79.1–91.6) | 99.3 (98.6–100.0) | 96.8 (95.4–98.1) |

important factor to prevent bias. On the other hand, DL, which benefits not only medicine but also many other fields, is a product of the accumulated knowledge of engineering researchers. For engineering researchers, one of the key factors is that the DL must perform better relative to prior research and the ingenuity to achieve this can be the novelty. From this aspect, it is reasonable to develop and validate the DL using the same open dataset as prior studies to demonstrate improved performance. Open datasets have contributed greatly to the development of DL because of their ease of use, the results are highly reproducible and comparisons with previous studies are easy to make even if there is bias in a clinical sense. In addition, patient privacy issues may make it difficult to access each hospital's data. In order to better use DL created by engineering research in medical practice, medical researchers must verify biases from various perspectives, understand the characteristics of DL, and conduct research that will benefit patients and reduce the daily clinical burden on physicians. Medical and engineering researchers should cooperate and share roles to advance medical care. About half (32 out of 63) of the papers included in this study were externally validated, which is the most important factor in the evaluation of AI. The high risk of bias in this study was largely influenced by the fact that 35 out of the 63 papers included risks of bias in the analysis portion of the study. It is likely that a more refined design of the analysis would allow for a low risk of bias study and also a better understanding of the pneumothorax diagnosis AI model.

The present study has several limitations. More than half (57%) of the included studies were classified as high risk of bias by PROBAST, limiting the conclusions that could be drawn from the meta-analysis. In addition, some papers lacked training and validation details, which contributed to the high risk of bias. Also, in terms of comparing model performance, it is reasonable that multiple studies used the same large open database for training and validation, but actual clinical practice will have a variety of different cohorts, which reduces the applicability of our conclusions to significantly different cohorts. Furthermore, publication bias also affected the results of this study.

In order to provide better medical care to patients and reduce the burden on physicians, pneumothorax diagnosis DL and physicians may complement each other to improve the accuracy of pneumothorax diagnosis in clinical practice. DL will be used in various medical fields in the future. Therefore, it is important to build up evidence by integrating individual original research and capturing overall characteristics through systematic review and meta-analysis.

### Points for clinical practice

- Use of AI as an adjunct to physicians' diagnosis of pneumothorax may have potential benefits and deserves further exploration.

### Questions for future research

- How much improvement has been made in the performance of pneumothorax diagnosis by physicians with AI assistance?
- To what extent do confounding factors inherent in chest radiographs impact pneumothorax diagnostic AIs?

### References

1   O'Connor AR, Morgan WE. Radiological review of pneumothorax. *BMJ* 2005; 330: 1493–1497.
2   Sahn SA, Heffner JE. Spontaneous pneumothorax. *N Engl J Med* 2000; 342: 868–874.
3   Melton LJ  3rd, Hepper NG, Offord KP. Incidence of spontaneous pneumothorax in Olmsted County, Minnesota: 1950 to 1974. *Am Rev Respir Dis* 1979; 120: 1379–1382.
4   Schramel FM, Postmus PE, Vanderschueren RG. Current aspects of spontaneous pneumothorax. *Eur Respir J* 1997; 10: 1372–1379.
5   Light RW, O'Hara VS, Moritz TE, *et al.* Intrapleural tetracycline for the prevention of recurrent spontaneous pneumothorax. Results of a Department of Veterans Affairs cooperative study. *JAMA* 1990; 264: 2224–2230.

6    Lippert HL, Lund O, Blegvad S, *et al.* Independent risk factors for cumulative recurrence rate after first spontaneous pneumothorax. *Eur Respir J* 1991; 4: 324–331.

7    Videm V, Pillgram-Larsen J, Ellingsen O, *et al.* Spontaneous pneumothorax in chronic obstructive pulmonary disease: complications, treatment and recurrences. *Eur J Respir Dis* 1987; 71: 365–371.

8    Larson PA, Berland LL, Griffith B, *et al.* Actionable findings and the role of IT support: report of the ACR Actionable Reporting Work Group. *J Am Coll Radiol* 2014; 11: 552–558.

9    United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR). Effects of ionizing radiation. 2008. https://digitallibrary.un.org/record/692844 Date last accessed: 25 March 2023.

10   Mettler FA Jr, Bhargavan M, Faulkner K, *et al.* Radiologic and nuclear medicine studies in the United States and worldwide: frequency, radiation dose, and comparison with other radiation sources – 1950–2007. *Radiology* 2009; 253: 520–531.

11   Brar MS, Bains I, Brunet G, *et al.* Occult pneumothoraces truly occult or simply missed: redux. *J Trauma* 2010; 69: 1335–1337.

12   Nakajima Y, Yamada K, Imamura K, *et al.* Radiologist supply and workload: international comparison – Working Group of Japanese College of Radiology. *Radiat Med* 2008; 26: 455–465.

13   Rimmer A. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ* 2017; 359: j4683.

14   Esteva A, Robicquet A, Ramsundar B, *et al.* A guide to deep learning in healthcare. *Nat Med* 2019; 25: 24–29.

15   Ueda D, Shimazaki A, Miki Y. Technical and clinical overview of deep learning in radiology. *Jpn J Radiol* 2019; 37: 15–33.

16   Tadavarthi Y, Vey B, Krupinski E, *et al.* The state of radiology AI: considerations for purchase decisions and current market offerings. *Radiol Artif Intell* 2020; 2: e200004.

17   LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; 521: 436–444.

18   Wolff RF, Moons KGM, Riley RD, *et al.* PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019; 170: 51–58.

19   McInnes MDF, Moher D, Thombs BD, *et al.* Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. *JAMA* 2018; 319: 388–396.

20   Moher D, Liberati A, Tetzlaff J, *et al.* Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009; 339: b2535.

21   Liu X, Faes L, Kale AU, *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019; 1: e271–e297.

22   Jackson D, Turner R. Power analysis for random-effects meta-analysis. *Res Synth Methods* 2017; 8: 290–302.

23   Macaskill P, Gatsonis C, Deeks J, *et al.* Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. London, Cochrane, 2010.

24   R Core Team. R: A Language and Environment for Statistical Computing. Vienna, R Foundation for Statistical Computing, 2013.

25   Egger M, Davey Smith G, Schneider M, *et al.* Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997; 315: 629–634.

26   Sung J, Park S, Lee SM, *et al.* Added value of deep learning-based detection system for multiple major findings on chest radiographs: a randomized crossover study. *Radiology* 2021; 299: 450–459.

27   Park S, Lee SM, Kim N, *et al.* Application of deep learning-based computer-aided detection system: detecting pneumothorax on chest radiograph after biopsy. *Eur Radiol* 2019; 29: 5341–5348.

28   Rudolph J, Huemmer C, Ghesu FC, *et al.* Artificial intelligence in chest radiography reporting accuracy: added clinical value in the emergency unit setting without 24/7 radiology coverage. *Invest Radiol* 2022; 57: 90–98.

29   Taylor AG, Mielke C, Mongan J. Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: a retrospective study. *PLoS Med* 2018; 15: e1002697.

30   Feng S, Liu Q, Patel A, *et al.* Automated pneumothorax triaging in chest X-rays in the New Zealand population using deep-learning algorithms. *J Med Imaging Radiat Oncol* 2022; 66: 1035–1043.

31   Kao CY, Lin CY, Chao CC, *et al.* Automated radiology alert system for pneumothorax detection on chest radiographs improves efficiency and diagnostic performance. *Diagnostics* 2021; 11: 1182.

32   Wang Q, Liu Q, Luo G, *et al.* Automated segmentation and diagnosis of pneumothorax on chest X-rays with fully convolutional multi-scale ScSE-DenseNet: a retrospective study. *BMC Med Inform Decis Mak* 2020; 20: Suppl. 14, 317.

33   Wang X, Yang S, Lan J, *et al.* Automatic segmentation of pneumothorax in chest radiographs based on a two-stage deep learning method. *IEEE Trans Cognit Dev Syst* 2022; 14: 205–218.

34   Yi PH, Kim TK, Yu AC, *et al.* Can AI outperform a junior resident? Comparison of deep neural network to first-year radiology residents for identification of pneumothorax. *Emerg Radiol* 2020; 27: 367–375.

35   Majkowska A, Mittal S, Steiner DF, *et al.* Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology* 2020; 294: 421–431.

36    Mosquera C, Diaz FN, Binder F, *et al.* Chest x-ray automated triage: a semiologic approach designed for clinical implementation, exploiting different types of labels through a combination of four deep learning architectures. *Comput Methods Programs Biomed* 2021; 206: 106130.

37    Abedalla A, Abdullah M, Al-Ayyoub M, *et al.* Chest X-ray pneumothorax segmentation using U-Net with EfficientNet and ResNet architectures. *PeerJ Comput Sci* 2021; 7: e607.

38    Wang H, Gu H, Qin P, *et al.* CheXLocNet: automatic localization of pneumothorax in chest radiographs using deep convolutional neural networks. *PLoS One* 2020; 15: e0242013.

39    Rudolph J, Schachtner B, Fink N, *et al.* Clinically focused multi-cohort benchmarking as a tool for external validation of artificial intelligence algorithm performance in basic chest radiography analysis. *Sci Rep* 2022; 12: 12764.

40    Zhou Y, Zhou T, Zhou T, *et al.* Contrast-attentive thoracic disease recognition with dual-weighting graph reasoning. *IEEE Trans Med Imaging* 2021; 40: 1196–1206.

41    Haq NF, Moradi M, Wang ZJ. A deep community based approach for large scale content based X-ray image retrieval. *Med Image Anal* 2021; 68: 101847.

42    Hwang EJ, Hong JH, Lee KH, *et al.* Deep learning algorithm for surveillance of pneumothorax after lung biopsy: a multicenter diagnostic cohort study. *Eur Radiol* 2020; 30: 3660–3671.

43    Rajpurkar P, Irvin J, Ball RL, *et al.* Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med* 2018; 15: e1002686.

44    Hong W, Hwang EJ, Lee JH, *et al.* Deep learning for detecting pneumothorax on chest radiographs after needle biopsy: clinical implementation. *Radiology* 2022; 303: 433–441.

45    Kim S, Rim B, Choi S, *et al.* Deep learning in multi-class lung diseases' classification on chest X-ray images. *Diagnostics* 2022; 12: 915.

46    Thian YL, Ng D, Hallinan J, *et al.* Deep learning systems for pneumothorax detection on chest radiographs: a multicenter external validation study. *Radiol Artif Intell* 2021; 3: e200190.

47    Park S, Lee SM, Lee KH, *et al.* Deep learning-based detection system for multiclass lesions on chest radiographs: comparison with observer readings. *Eur Radiol* 2020; 30: 1359–1368.

48    Tian Y, Wang J, Yang W, *et al.* Deep multi-instance transfer learning for pneumothorax classification in chest X-ray images. *Med Phys* 2022; 49: 231–243.

49    Niehues SM, Adams LC, Gaudin RA, *et al.* Deep-learning-based diagnosis of bedside chest X-ray in intensive care and emergency medicine. *Invest Radiol* 2021; 56: 525–534.

50    Liang X, Peng C, Qiu B, *et al.* Dense networks with relative location awareness for thorax disease identification. *Med Phys* 2019; 46: 2064–2073.

51    Hallinan J, Feng M, Ng D, *et al.* Detection of pneumothorax with deep learning models: learning from radiologist labels vs natural language processing model generated labels. *Acad Radiol* 2022; 29: 1350–1358.

52    Elkins A, Freitas FF, Sanz V. Developing an app to interpret chest X-rays to support the diagnosis of respiratory pathology with artificial intelligence. *J Med Artif Intell* 2020; 3: 8.

53    Nam JG, Kim M, Park J, *et al.* Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs. *Eur Respir J* 2021; 57: 2003961.

54    Hwang EJ, Park S, Jin KN, *et al.* Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Netw Open* 2019; 2: e191095.

55    Chen KC, Yu HR, Chen WS, *et al.* Diagnosis of common pulmonary diseases in children by X-ray images and deep learning. *Sci Rep* 2020; 10: 17374.

56    Gipson J, Tang V, Seah J, *et al.* Diagnostic accuracy of a commercially available deep-learning algorithm in supine chest radiographs following trauma. *Br J Radiol* 2022; 95: 20210979.

57    Jin KN, Kim EY, Kim YJ, *et al.* Diagnostic effect of artificial intelligence solution for referable thoracic abnormalities on chest radiography: a multicenter respiratory outpatient diagnostic cohort study. *Eur Radiol* 2022; 32: 3469–3479.

58    Shin HJ, Son NH, Kim MJ, *et al.* Diagnostic performance of artificial intelligence approved for adults for the interpretation of pediatric chest radiographs. *Sci Rep* 2022; 12: 10215.

59    Seah J, Tang C, Buchlak QD, *et al.* Do comprehensive deep learning algorithms suffer from hidden stratification? A retrospective study on pneumothorax detection in chest radiography. *BMJ Open* 2021; 11: e053024.

60    Chen B, Li J, Guo X, *et al.* DualCheXNet: dual asymmetric feature learning for thoracic disease classification in chest X-rays. *Biomed Signal Process Control* 2019; 53: 101554.

61    Thian YL, Ng DW, Hallinan J, *et al.* Effect of training data volume on performance of convolutional neural network pneumothorax classifiers. *J Digit Imaging* 2022; 35: 881–892.

62    Wang Y, Sun L, Jin Q. Enhanced diagnosis of pneumothorax with an improved real-time augmentation for imbalanced chest X-rays data based on DCNN. *IEEE/ACM Trans Comput Biol Bioinform* 2021; 18: 951–962.

63    Lin CH, Wu JX, Li CM, *et al.* Enhancement of chest X-ray images to improve screening accuracy rate using iterated function system and multilayer fractional-order machine learning classifier. *IEEE Photonics J* 2020; 12: 4100218.

64    Choi SY, Park S, Kim M, *et al.* Evaluation of a deep learning-based computer-aided detection algorithm on chest radiographs: case–control study. *Medicine* 2021; 100: e25663.

65 Iqbal T, Shaukat A, Akram MU, *et al.* A hybrid VDV model for automatic diagnosis of pneumothorax using class-imbalanced chest X-rays dataset. *IEEE Access* 2022; 10: 27670–27683.

66 Rueckel J, Trappmann L, Schachtner B, *et al.* Impact of confounding thoracic tubes and pleural dehiscence extent on artificial intelligence pneumothorax detection in chest radiographs. *Invest Radiol* 2020; 55: 792–798.

67 Kakkar B, Johri P, Kumar Y, *et al.* An IoMT-based federated and deep transfer learning approach to the detection of diverse chest diseases using chest X-rays. *Hum Centric Comput Inf Sci* 2022; 12: 24.

68 Lee SY, Ha S, Jeon MG, *et al.* Localization-adjusted diagnostic performance and assistance effect of a computer-aided detection system for pneumothorax and consolidation. *NPJ Digit Med* 2022; 5: 107.

69 Shamrat F, Azam S, Karim A, *et al.* LungNet22: a fine-tuned model for multiclass classification and prediction of lung disease using X-ray images. *J Pers Med* 2022; 12: 680.

70 Hong M, Rim B, Lee H, *et al.* Multi-class classification of lung diseases using CNN models. *Appl Sci* 2021; 11: 9289.

71 Cho Y, Park B, Lee SM, *et al.* Optimal number of strong labels for curriculum learning with convolutional neural network to classify pulmonary abnormalities in chest radiographs. *Comput Biol Med* 2021; 136: 104750.

72 Kim EY, Kim YJ, Choi WJ, *et al.* Performance of a deep-learning algorithm for referable thoracic abnormalities on chest radiographs: a multicenter study of a health screening cohort. *PLoS One* 2021; 16: e0246472.

73 Rueckel J, Huemmer C, Fieselmann A, *et al.* Pneumothorax detection in chest radiographs: optimizing artificial intelligence system for accuracy and confounding bias reduction using in-image annotations in algorithm training. *Eur Radiol* 2021; 31: 7888–7900.

74 Luo JX, Liu WF, Yu L. Pneumothorax recognition neural network based on feature fusion of frontal and lateral chest X-ray images. *IEEE Access* 2022; 10: 53175–53187.

75 Mangalmurti Y, Wattanapongsakorn N. Practical machine learning techniques for COVID-19 detection using chest. *Intelligent Automat Soft Comput* 2022; 34: 733–752.

76 Kitamura G, Deible C. Retraining an open-source pneumothorax detecting machine learning algorithm for improved performance to medical images. *Clin Imaging* 2020; 61: 15–19.

77 Park S, Kim G, Oh Y, *et al.* Self-evolving vision transformer for chest X-ray diagnosis through knowledge distillation. *Nat Commun* 2022; 13: 3848.

78 Lai KHA, Ma SK. Sensitivity and specificity of artificial intelligence with Microsoft Azure in detecting pneumothorax in emergency department: a pilot study. *Hong Kong J Emerg Med* 2023; 30: 8–15.

79 Wang H, Jia H, Lu L, *et al.* Thorax-Net: an attention regularized deep neural network for classification of thoracic diseases on chest radiography. *IEEE J Biomed Health Inform* 2020; 24: 475–485.

80 Cicero M, Bilbily A, Colak E, *et al.* Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. *Invest Radiol* 2017; 52: 281–287.

81 Wang H, Wang S, Qin Z, *et al.* Triple attention learning for classification of 14 thoracic diseases using chest radiography. *Med Image Anal* 2021; 67: 101846.

82 Draelos RL, Dov D, Mazurowski MA, *et al.* Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Med Image Anal* 2021; 67: 101857.

83 Li X, Thrall JH, Digumarthy SR, *et al.* Deep learning-enabled system for rapid pneumothorax screening on chest CT. *Eur J Radiol* 2019; 120: 108692.

84 Lyu W, Xia F, Zhou C, *et al.* [Application of deep learning-based chest CT auxiliary diagnosis system in emergency trauma patients]. *Zhonghua Yi Xue Za Zhi* 2021; 101: 481–486.

85 Röhrich S, Schlegl T, Bardach C, *et al.* Deep learning detection and quantification of pneumothorax in heterogeneous routine chest computed tomography. *Eur Radiol Exp* 2020; 4: 26.

86 Lee CC, Lin CS, Tsai CS, *et al.* A deep learning-based system capable of detecting pneumothorax via electrocardiogram. *Eur J Trauma Emerg Surg* 2022; 48: 3317–3326.

87 Kuo PC, Tsai CC, López DM, *et al.* Recalibration of deep learning models for abnormality detection in smartphone-captured chest radiograph. *NPJ Digit Med* 2021; 4: 25.

88 Li F, Shi JX, Yan L, *et al.* Lesion-aware convolutional neural network for chest radiograph classification. *Clin Radiol* 2021; 76: 155.e1–155.e14.

89 Cho Y, Kim JS, Lim TH, *et al.* Detection of the location of pneumothorax in chest X-rays using small artificial neural networks and a simple training process. *Sci Rep* 2021; 11: 13054.

90 Chan KK, Joo DA, McRae AD, *et al.* Chest ultrasonography versus supine chest radiography for diagnosis of pneumothorax in trauma patients in the emergency department. *Cochrane Database Syst Rev* 2020; 7: CD013031.

91 Wiens J, Saria S, Sendak M, *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019; 25: 1337–1340.