TuJNaS

Review Article

Thamar University Journal of Natural & Applied Sciences 11(1) (2026) 21– 28

DOI: https://doi.org/10.59167/g1qw6p23

Deep Learning for Respiratory Sound Analysis: A Systematic Review and Meta-Analysis (2019–2024)

Shaima'a Mohammed Nasser Al-Jabali[1]*, Farhan Nashwan[2], Waleed M. Altalabi[3]

[1]Information Technology Department, Faculty of Engineering and Information Technology, AL-Qalam University, Ibb, Yemen

[2]Electrical Engineering Department, Faculty of Engineering, IBB University, IBB, Yemen

[3]Biomedical Engineering Department, Sana'a Community College, Sana'a, Yemen

***Corresponding Author:** Shaima'a M. N. Al-Jabali, Information Technology Department, Faculty of Engineering and Information Technology, AL-Qalam University, Ibb, Yemen. Email: shaima.aljabali@gmail.com

Received: 18 November 2025. Revised: 12 April 2026. Accepted: 12 April 2026. Published: 29 June 2026.

Abstract

Background: Respiratory diseases remain a significant global health burden, particularly in resource-limited settings. Objective: To review and quantitatively analyze deep learning-based models for interpreting respiratory sounds. **Methodology:** A systematic search was conducted in the PubMed, IEEE Xplore, ScienceDirect, and Scopus databases for the period 2019–2024, using PRISMA 2020 criteria. Of the 678 records accessed, 98 studies met the qualitative criteria, while 42 met the quantitative analysis criteria. **Results:** An exploratory meta-analysis conducted on a subset of studies reporting native accuracy (k = 9) yielded a pooled accuracy of approximately 90.0% (95% CI: 76.0%–97.0%), with substantial heterogeneity (I^2 = 99.2%). **Conclusion:** Deep learning techniques demonstrate promising diagnostic capabilities, but they still face challenges related to reliance on limited databases, poor representation of rare diseases, and difficulty in interpreting their outputs. **Implications:** Future research should focus on diversifying datasets, enhancing the integration of multimodal data, and developing interpretable or federated learning-based models to support their adoption in clinical settings.

Index Terms: Respiratory sounds; lung disease classification; Deep learning; Systematic review; Meta-analysis; Multitask learning; Explainable AI.

1. Introduction

Respiratory diseases are among the most significant global health challenges, including conditions such as asthma, chronic obstructive pulmonary disease (COPD), pneumonia, and bronchiectasis. These diseases continue to cause high rates of morbidity and mortality worldwide, particularly in low- and middle-income countries with limited healthcare resources. Early and accurate diagnosis remains essential for improving clinical outcomes and reducing economic burdens [1].

According to the 2019 Global Burden of Disease report, there were approximately 454.6 million active clinical cases of chronic respiratory diseases, with 4 million deaths in the same year, making these diseases the third among the leading cause of death globally [1]. This underscores the critical need for accurate, low-cost, and widely applicable diagnostic tools.

Unlike previous narrative reviews that merely presented the literature without quantitative analysis, this study applies a PRISMA 2020–guided systematic review framework and includes a structured quantitative synthesis of recent deep learning studies on respiratory sound analysis (2019–2024). It offers a comprehensive synthesis that includes the statistical integration of key indicators such as accuracy, variance, and subgroup analysis. This methodology represents a step toward bridging the gap between scattered evidence and reproducible, unified insights [2].

Traditional diagnostic methods—such as X-rays, computed tomography (CT) scans, and pulmonary function tests—remain effective, but they are often expensive or require specialized equipment. While auscultation with a stethoscope is a widely available and easy-to-use option, its accuracy is influenced by the practitioner's experience and can vary from one physician to another. To overcome these limitations, computational analysis of respiratory sounds has emerged as an objective and non-invasive alternative, transforming the auscultation process into a measurable and repeatable step, thus expanding its applicability in resource-constrained environments [2].

Recent years have witnessed significant advancements in the use of deep learning (DL) techniques for respiratory sound analysis. Convolutional network (CNN) models, recurrent network (RNN/LSTM) models, and newer models based on transformers or hybrid structures have proven their ability to extract accurate and meaningful sound patterns [3-5]. Their applications range from disease detection and classification of various pulmonary conditions to identifying sounds such as wheezing and rattling, and even to estimating disease severity, with studies ranging from analyzing wheezing sounds in children to modeling the severity of COPD [6, 7]. Recent review studies have further confirmed the effectiveness of deep learning–based auscultation systems and intelligent stethoscopes in improving diagnostic performance and automation capabilities [8, 9].

Despite this progress, several challenges remain, most notably:

- **Over-reliance on the ICBHI 2017 database** — which—while contributing to research—limits the generalizability of models to real-world settings [10, 11].

© 2026 The Authors. Published by Thamar University.
This article is published under the terms of the Creative Commons Attribution (CC BY 4.0) License. This license permits unrestricted use, distribution, and reproduction in any medium, provided that the original author(s) and copyright holder(s) are properly credited, and the original publication in this journal is cited in accordance with accepted academic standards. Any use, distribution, or reproduction that does not comply with these terms is prohibited.

- **Data imbalance and scarcity** — with rare diseases being underrepresented, leading to model bias and inconsistent results. In addition, the availability of diverse and clinically annotated lung sound datasets remains limited, as existing datasets are often collected under controlled conditions and include relatively small cohorts, which may not fully reflect real-world variability [12, 13]. Methods such as data augmentation or meta-analysis offer only partial solutions [7, 14].
- **Focus on individual tasks** —as many studies treat each task independently, overlooking the benefits of multitasking learning in enhancing diagnostic performance, as most studies still lack integration of explainable AI (XAI) techniques [15, 16]. As most studies lack the integration of interpretive intelligence (XAI) techniques, which are essential for building clinical confidence [16].
- **The lack of verification across different databases and integration of models in the clinical context,** as models often show a decline in performance when tested on new data, while modern methods such as metadata-guided training or adaptive training seek to narrow this gap [17, 18].

These gaps reveal the need for a comprehensive and systematic evaluation of computational techniques in respiratory sound analysis. Although several recent reviews have discussed artificial intelligence and audio-based respiratory disease diagnosis, few have combined a PRISMA-guided design with a structured quantitative synthesis focused specifically on deep learning-based respiratory sound analysis [19, 20].

To perform a structured quantitative synthesis of model performance, including an exploratory meta-analysis of studies reporting native accuracy and a descriptive analysis of studies reporting AUC and F1-score. This paper contributes four key points:

1) To conduct a systematic review, aligned with PRISMA 2020, of deep learning models for respiratory sound analysis during the period 2019–2024 [21].
2) To perform a structured quantitative synthesis of model performance, including an exploratory meta-analysis of studies reporting native accuracy and a descriptive analysis of studies reporting AUC and F1-score.
3) To analyze pivotal gaps in database diversity, class imbalances, task design, and the integration of XAI techniques [10, 14-16].
4) To propose future research avenues, including federated learning, self-supervised representational learning, cross-database evaluation, and the integration of multimodal data (e.g., demographic characteristics, medical images, and pulmonary function tests) using adaptive training [5, 17, 18].

By gathering available evidence and formulating clear guidelines, this study seeks to bridge the gap between algorithmic development and clinical application and to support the development of reliable, interpretable, and generalizable AI tools for diagnosing respiratory diseases.

2. Methodology

This survey adopted the *PRISMA 2020* methodology for preferred reporting elements for systematic reviews and meta-analyses to ensure transparency, reproducibility, and adherence to rigorous methodological standards [21]. The workflow comprised four main phases:

(1) Reference search, (2) Screening of studies, (3) Assessment of relevance, and (4) Final selection of studies that met the criteria.

2.1 Search Strategy

We conducted a comprehensive search across four major databases: PubMed (MEDLINE), IEEE Xplore, ScienceDirect, and Scopus. The search included studies published between 2019 and 2024, aiming to identify the latest developments in deep learning-based respiratory sound analysis, in accordance with PRISMA's recommendations for citing sources [21]. The search equation (modified to suit each database) was formulated as follows:

("respiratory sounds" OR "lung sounds" OR "cough sounds" OR "breath sounds")

AND ("deep learning" OR "neural networks" OR "CNN" OR "RNN" OR "transformer" OR "machine learning")

AND ("classification" OR "detection" OR "diagnosis")

We also reviewed the reference lists of included studies and related reviews to ensure comprehensiveness and that no important sources were overlooked [21].

2.2 Inclusion and Exclusion Criteria

2.2.1 Inclusion:

This review includes: (i) Original peer- reviewed studies that used deep learning techniques, machine learning methods, or hybrid models combining both to analyze respiratory sounds; (ii) Studies that addressed one or more of the following aspects: disease detection, disease type classification, sound type recognition, or condition severity assessment; (iii) Studies that provided sufficient information about the datasets used, experimental design, and performance evaluation metrics such as accuracy, F1-score, area under the curve, sensitivity, and specificity to allow structured data extraction; (iv) Articles published in English between 2019 and 2024.

2.2.2 Exclusion:

Analogical reviews, editorials, and commentaries that do not include original experiments are excluded; studies that focus exclusively on speech or non- respiratory signals such as heart sounds or EEG/EKG signals are excluded; research that lacks sufficient methodological detail or extractable evaluation data is excluded; and duplicate publications are excluded. These standards are based on PRISMA best practices [21].

2.3 Screening and Selection

2.3.1 Phase 1 – Study Identification:

The retrieved studies were exported to the reference management software, and duplicates were automatically removed and manually verified.

2.3.2 Phase 2 – Title and Abstract Review:

The titles and abstracts were screened according to the predefined inclusion criteria by a primary reviewer.

2.3.3 Phase 3 – Full Text Evaluation:

The full texts of the studies that initially appeared eligible were independently reviewed by a second reviewer to confirm eligibility and ensure adherence to the predefined inclusion and exclusion criteria. Discrepancies were resolved through discussion and consensus. Formal inter-rater agreement statistics (e.g., Cohen's kappa) were not calculated, as the screening process followed a sequential independent validation approach with consensus resolution. This approach is considered acceptable in systematic reviews when discrepancies are resolved through structured discussion and was adopted to ensure consistency while minimizing reporting complexity.

2.3.4 Phase 4 – Final Inclusion:

The final studies were included after agreement was achieved. This structured screening approach improves transparency and methodological consistency in accordance with PRISMA standards [21].

2.4 Data Extraction

We adopted a standardized data extraction model that included: (i) Study information such as researcher names, publication year, and publication status; (ii) Dataset characteristics, including dataset type (e.g., ICBHI 2017 or clinical datasets), sample size (subjects, recordings, or respiratory cycles), and task definition; (iii) Model characteristics, including architecture type (CNN, RNN, Transformer, hybrid, or classical machine learning); (iv) Evaluation metrics, including native accuracy (when available), AUC, F1-score, sensitivity, and specificity; and (v) Methodological design variables, including validation strategy (e.g., train/test split or cross-validation), external validation, calibration reporting, and availability of uncertainty measures (CI/SE/variance).

All extracted variables were systematically coded into a structured evidence table (master extraction table), ensuring consistency and traceability across studies. Each study was additionally annotated with eligibility indicators for qualitative synthesis, quantitative synthesis, and primary accuracy meta-analysis.

This structured extraction framework enabled reproducible filtering of studies into metric-specific groups (accuracy-based, AUC-based, and F1-based analyses) without enforcing metric harmonization. The approach ensures transparency in how studies were selected, categorized, and subsequently used in the quantitative evidence synthesis.

2.5 Quality Assessment

We used a modified and adapted version of PROBAST-AI (a tool for assessing bias risks in predictive models with AI extensions) to assess bias risks and reproducibility. Instead of applying the full PROBAST-AI scoring system, key domains relevant to machine learning–based respiratory sound analysis were operationalized within the structured data extraction framework. This tool includes four main axes: participant characteristics,

predictor characteristics and processing, study outcomes including labeling quality and comment reliability, and a statistical analysis axis (validation methods, leakage control, and analysis reports). Each axis was qualitatively assessed using the information extracted from each study.

In particular, variables related to split type, external validation, calibration reporting, and availability of uncertainty measures (CI/SE/variance) were used as practical indicators of methodological robustness. These variables were systematically recorded in the master extraction table and used to identify potential sources of bias and heterogeneity across studies.

A summary of these methodological quality indicators is provided in Table I. These indicators are derived from the structured coding of the studies included in the quantitative synthesis.

Table I. Summary of Methodological Quality Indicators Based on the Quantitatively Included Studies (N = 42)

Domain	Indicator	Reporting Level	Common Issues
Validation design	Split type (patient-wise / random/unclear)	Variable	Lack of patient-independent splitting
External validation	Independent dataset validation	Limited	Few studies used true external validation
Calibration	Calibration reporting (e.g., ECE/Brier)	Rare	Calibration almost never reported
Uncertainty	CI / SE / variance reporting	Limited	Missing uncertainty quantification
Data leakage control	Preprocessing before split	Unclear	Potential leakage due to augmentation before the split
Dataset transparency	Sample size and dataset description	Moderate	Incomplete reporting in some studies

Across the quantitatively included studies (N = 42), only a minority explicitly reported patient-wise data splitting, while a considerable proportion relied on random or unclear split designs. External validation was reported in only a small subset of studies, and calibration assessment was rarely performed. These findings indicate that a substantial proportion of studies may be subject to optimistic performance estimation and limited generalizability.

Rather than assigning fixed numerical bias scores, the assessment was used to guide the interpretation of results and the stratification of studies during the quantitative synthesis. Common methodological limitations observed across studies included a lack of patient-independent validation, limited external validation, and incomplete reporting of calibration and uncertainty measures. TRIPOD-AI guidelines were also considered to ensure the clarity and reproducibility of the predictive model reports [22].

2.6 Reported Metrics Cheat-sheet

In order to standardize terminology and facilitate tracking the indicators used, the following studies were included. Table II provides a structured summary of the most commonly reported evaluation metrics, along with their symbols and interpretation notes.

Importantly, performance metrics were not converted across types (e.g., AUC or F1-score into Accuracy), and each metric was analyzed within its native reporting context to avoid introducing methodological bias.

Table II. Metrics used across the included studies.

Metric	Symbol	Notes
Accuracy	Acc	Primary metric for studies reporting native classification accuracy; used only when directly reported.
F1-score	F1	Reported as provided (macro- or weighted-F1); analyzed separately without conversion.
AUC	AUC	ROC-AUC is preferred when available; analyzed independently with corresponding uncertainty when reported.
Sensitivity	Sen	Recall for the positive class, report with Specificity.
Specificity	Spe	True negative rate; clinical interpretability.
Calibration	–	ECE/Brier suggested for clinical readiness.

2.7 PRISMA Flow Diagram

The study selection process is summarized in Figure 1 using a PRISMA 2020–compliant diagram [21]. It reports the number of records identified, screened, excluded (with reasons), assessed for eligibility, and included in qualitative and quantitative syntheses. Counts are reported as follows: *Records identified:* N =678; *after duplicates removed:* N = 578; *screened (title/abstract):* N = 578; *full-text assessed:* N = 155; *full-text excluded with reasons:* N = 57; *included in qualitative synthesis:* N =98; *included in quantitative synthesis (meta-analysis):* N = 42. For completeness, *duplicates removed* amounted to N = 100, and *title/abstract exclusions* to N= 423.

The set of studies included in the quantitative synthesis (N = 42) was further structured according to the type of reported evaluation metric (e.g., Accuracy, AUC, F1-score), and not all studies contributed to the same pooled analysis.

Screening outcomes and exclusion reasons: Of the 578 records screened, 423 were excluded at the title/abstract stage for reasons including non-auscultation audio (e.g., speech-only, heart sounds), out-of-scope tasks (non- classification or non-respiratory), insufficient methodological detail or missing evaluation metrics, and non-English publications outside our 2019–2024 window. Of the 155 full-texts assessed, studies not meeting our predefined inclusion criteria (e.g., lack of clinically meaningful labels, absence of extractable performance metrics, or inadequate validation design) did not progress to qualitative or quantitative synthesis. A detailed breakdown is provided in Table III and Table VII. Table VII summarizes studies included in the qualitative synthesis (N = 98), while studies included in the quantitative synthesis (N = 42) were further filtered based on metric comparability. (Table VI, Table VII, and Table VIII in the APPENDICES).

2.8 Statistical Analysis and Protocol

A structured quantitative synthesis was conducted to summarize reported performance across studies. Due to substantial heterogeneity in metrics, datasets, and validation designs, a fully unified meta-analysis across all included studies was not considered methodologically appropriate. The primary outcome was classification performance as reported in each study, and no metric conversion (e.g., AUC or F1-score to Accuracy) was applied to avoid introducing methodological bias. Studies were grouped and analyzed based on the type of reported metric. A formal meta-analysis was conducted only for studies reporting native accuracy, while studies reporting AUC or F1-score were retained for descriptive synthesis within their respective metric categories. For studies reporting native Accuracy, a restricted subset was defined for quantitative pooling. Proportions were stabilized via a logit transformation and combined using inverse-variance weighting within a random-effects framework. The effect size was defined as the logit-transformed accuracy proportion. Study-specific estimates were weighted using inverse-variance weighting, and when multiple performance results were reported within a single study, only one representative estimate was selected based on predefined criteria to avoid unit-of-analysis errors. Between-study variance (τ^2) and heterogeneity (I^2) were reported together with 95% confidence intervals (CIs) and, where appropriate, 95% prediction intervals (PIs). Given that sample sizes were not consistently reported across studies, a standardized denominator approximation was used for variance estimation; therefore, pooled estimates should be interpreted cautiously. This limitation may affect variance estimation and should be considered when interpreting pooled results.

Sources of heterogeneity were explored through subgroup analysis based on dataset type and model architecture to identify systematic variations across studies. A substantial proportion of studies did not clearly report patient-wise data splitting, and only a limited number of studies included external validation or calibration assessment, indicating potential risks of overestimation and limited generalizability. These methodological limitations were taken into account during the interpretation of pooled and descriptive results. Pre-specified subgroup analyses stratified results by model family (CNN, RNN/LSTM, Transformer/Hybrid), dataset (ICBHI 2017 vs. others), and task granularity (binary vs. multi-class), and evaluation metric type.

Small-study effects were explored using funnel plots as a descriptive tool. Forest and funnel plots were used to visualize pooled estimates within each metric-specific analysis. Formal statistical testing for publication bias was not performed due to the limited number of studies in the pooled subset.

2.8.1 Software and Estimation Method:

All quantitative analyses were performed using the R software environment (version 4.3.2) with the metafor and meta packages. A random-effects model with the restricted maximum likelihood (REML) estimator was used as the primary pooling method, and sensitivity considerations focused on studies with clearly reported native accuracy and more robust validation designs (e.g., patient-wise splits). Heterogeneity statistics (I^2, τ^2) and between-study variance were computed

automatically using these packages.

Sensitivity analyses were conducted by restricting the pooled analysis to studies with clearly reported native accuracy and more robust validation designs (e.g., patient-wise splits) to assess the stability of the pooled estimates.

2.8.2 Protocol Registration:

The review protocol was *not pre-registered* (e.g., PROSPERO/OSF), as the study was initiated prior to formal protocol registration planning. Nevertheless, all stages of the review strictly adhered to PRISMA 2020 to ensure methodological transparency and reproducibility [21]. Although not pre-registered, the protocol and the standardized data extraction form are provided in the Supplementary Material to the corresponding author to ensure transparency and reproducibility. To improve transparency, the full review protocol, including eligibility criteria, outcome definitions, and data extraction procedures, is provided in the Supplementary Material. No post-hoc modifications to eligibility criteria, outcome definitions, or analysis plans were made after the screening phase.

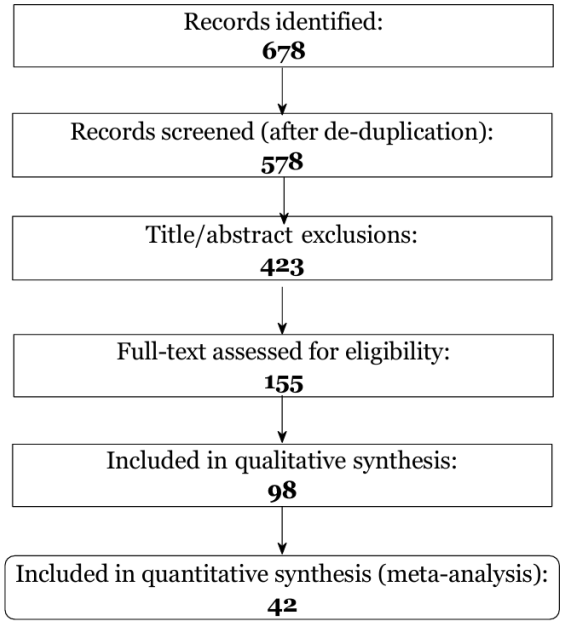

Figure 1: PRISMA 2020 flow diagram summarizing identification, screening, eligibility, and inclusion [21].

The studies included in the quantitative synthesis were further categorized based on metric compatibility for structured analysis. A detailed inspection of Table VIII further confirms that variability in validation design, dataset composition, and reporting transparency remains a key source of methodological heterogeneity across the included studies. This table provides the basis for all subsequent quantitative and descriptive analyses and ensures full transparency of study-level characteristics and methodological variability.

Table III. Full-text articles were excluded for the following reasons (N = 57).

Reason	Count
Not multi-disease / not classification-focused	19
Cough-only scope (outside auscultation inclusion)	10
Insufficient methodological detail/metrics	8
Non-auscultation modality (imaging/PFT-only)	7
Overlapping data / duplicate cohort	6
Outside time window / non-English	4
Other reasons	3
Total	**57**

3. Results and Meta-Analysis

3.1 Study selection

Following the PRISMA protocol [21], we initially identified N = 678 records from all sources. After removing duplicates, N = 578 unique records remained and were screened at the title/abstract level, excluding N = 423. Next, N =155 full-text articles were assessed for eligibility. Finally, N = 98 studies met the inclusion criteria for qualitative synthesis, and N = 42 provided sufficient data for quantitative synthesis.

The subset of studies included in the quantitative synthesis (N =42) was selected based on the availability of extractable and comparable evaluation metrics suitable for structured analysis. The study selection flow is summarized in Figure 1 [21]. A detailed summary of the studies included in the quantitative synthesis (N = 42) is provided in Table VIII.

3.2 Characteristics of Included Studies

The included studies (N = 98, qualitative synthesis) cover 2015–2025, with a noticeable increase after 2020 as interest in respiratory sound analysis and clinically relevant ML/DL grew [3, 4].

3.2.1 Datasets:

The ICBHI 2017 corpus was the most commonly used benchmark. Many studies relied on a single dataset, while only a few performed cross-dataset validation or prospective clinical testing [7, 10, 11]. Some studies used private clinical collections, but reporting practices and public availability varied [3, 4]. Across the included studies, the most frequently used datasets were *ICBHI 2017* (63%), *Clinical/Private* (28%), and *BRACETS* (6%). Only about 7% of studies utilized multiple datasets, highlighting limited dataset diversity and the need for broader benchmark standardization and cross-dataset validation to enhance model generalizability. A structured distribution of these studies is summarized in Table VII.

3.2.2 Model architectures:

CNNs dominated, followed by RNN/LSTM models. More recent work explored Transformers or hybrid CNN–Attention models, showing improved performance in small-data settings or under domain shifts [4, 5, 17, 18]. Some studies used classical ML methods (e.g., SVM RF) with hand-crafted features (MFCC, spectral contrast), achieving competitive results on limited datasets [6, 11, 23].

3.2.3 Tasks:

We observed four main task types: (i) binary disease detection, (ii) multi-class disease classification, (iii) sound-level classification, and (iv) severity estimation. In addition, some studies focused on clinically specific diagnostic targets rather than broad respiratory disease categories, such as the classification of pulmonary sounds for detecting interstitial lung diseases secondary to connective tissue diseases [24].

3.2.4 Evaluation Metrics:

Accuracy, F1-score, and AUC were most commonly reported across studies; however, these metrics were analyzed separately in the quantitative synthesis to avoid cross-metric comparability issues [3, 4, 7].

A comprehensive per-study comparison is provided in Table VIII. A complete per-study table is provided in Appendix A (Table VIII).

3.3 Quantitative Synthesis

We conducted a structured quantitative synthesis of reported performance metrics, grouped by model architecture. To ensure methodological consistency, no cross-metric conversion was applied between different performance measures. Instead, a strict accuracy-only subset was defined for formal pooling, while other metrics such as AUC and F1-score were analyzed descriptively. To provide a structured quantitative summary of model performance, an accuracy-based subset of studies was identified from the 42 studies included in the quantitative synthesis. Only studies reporting native accuracy were considered eligible for pooled analysis, while studies reporting AUC or F1-score were analyzed descriptively without metric conversion. CNN-based models consistently demonstrated high performance, followed by RNN/LSTM and Transformer/hybrid models. Transformer and attention-based pipelines showed robustness, especially with limited labeled data [4, 5, 17]. An exploratory random-effects meta-analysis was conducted on the eligible subset (k = 9). The pooled accuracy was estimated at approximately 90.0% (95% CI: 76.0%–97.0%), indicating generally high reported

performance across studies.

However, substantial heterogeneity was observed (I^2 = 99.2%; τ^2 = 1.9571), reflecting significant variability in datasets, task definitions, model architectures, and validation strategies. The wide prediction interval further suggests that model performance is highly context-dependent and may vary considerably across different experimental settings. Forest plots (Figure 2) display study-level effect estimates and their variability, while funnel plots (Figure 3) visualize potential publication bias. Forest plots were used to visualize individual study estimates and pooled effects, while funnel plots were used for descriptive assessment of potential small-study effects. The forest plot (Figure 2) summarizes comparable subsets of studies without cross-metric pooling. The funnel plot (Figure 3) illustrates study-level performance metrics plotted against their standard errors (SE). The vertical dashed line represents the central tendency of the included studies, while the dotted boundaries represent the 95% pseudo-confidence region. Visual symmetry in the funnel plot suggests limited evidence of publication bias, whereas visible asymmetry suggests potential small-study effects or selective reporting [3, 11]. However, due to the limited number of studies included in the pooled subset, formal statistical testing for publication bias was not performed. Overall, these findings indicate that while deep learning models can achieve high performance under controlled conditions, their generalizability remains uncertain due to methodological inconsistencies across studies.

To provide a concise overview of model performance and dataset usage trends across the included studies, Table IV presents a condensed summary of the main model families, reported performance ranges, and qualitative insights. The complete per-study summary table (original Table IV) is provided in Appendix A for detailed reference. The following figures are presented for descriptive visualization purposes only and do not represent formal pooled meta-analytic estimates.

To further quantify performance across comparable studies, an exploratory meta-analysis was conducted on studies reporting native accuracy. The results are visualized in Figure 2.

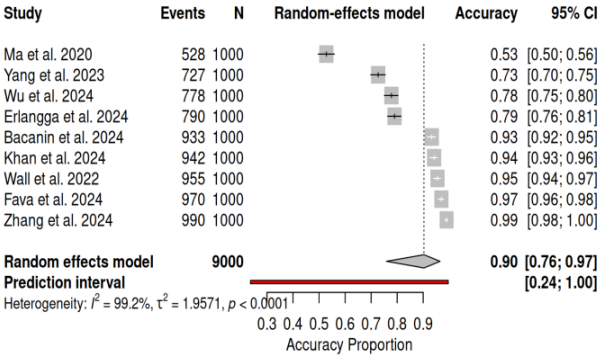

Figure 2: Forest plot of the exploratory random-effects meta-analysis of studies reporting native accuracy (k = 9). The pooled estimate was calculated using a logit-transformed proportion model with REML estimation. The diamond represents the pooled accuracy, and the horizontal lines indicate 95% confidence intervals. The prediction interval reflects between-study variability.

The forest plot illustrates substantial variability across individual study estimates, with reported accuracies ranging from moderate to very high values. The pooled estimate indicates generally strong performance; however, the wide confidence and prediction intervals reflect considerable heterogeneity. This variability is primarily attributed to differences in datasets, task definitions, model architectures, and validation protocols across studies. Table IV provides a descriptive summary of studies with extractable native accuracy values, grouped by model family. These values represent reported accuracy ranges and are presented to complement the exploratory meta-analysis rather than replace it. To explore potential small-study effects, a funnel plot was constructed, as shown in Figure 3.

The funnel plot shows an asymmetric distribution of studies around the pooled estimate, which may reflect heterogeneity in study design rather than true publication bias. Given the limited number of included studies and the use of approximated variance estimates, this plot should be interpreted cautiously and is presented for descriptive purposes only.

Table IV. Descriptive summary of studies with extractable native accuracy values, grouped by model family. These values represent reported accuracy ranges rather than pooled meta-analytic estimates.

Model Type	No. of Studies with Extractable Native Accuracy	Reported Native Accuracy Range (%)	Predominant Dataset (s)	Strength / Interpretation
CNN	8	52.79–97.00	ICBHI 2017 + Clinical/ Private	Predominantly strong performance, but values vary widely according to validation design and dataset composition.
RNN/LSTM	6	84.00–99.01	ICBHI 2017 + Clinical/ Private	Strong temporal modeling capability, with higher performance in task-specific settings.
Transformer/Hybrid	10	72.72–99.94	ICBHI 2017 + Private/ Multimodal	Strong representation learning and robustness, but affected by task heterogeneity and mixed evaluation settings.
Classical (SVM/RF)	4	75.60–99.72	ICBHI 2017 + Clinical/ Private	Competitive in structured or smaller datasets, though highly sensitive to task scope and validation protocol.

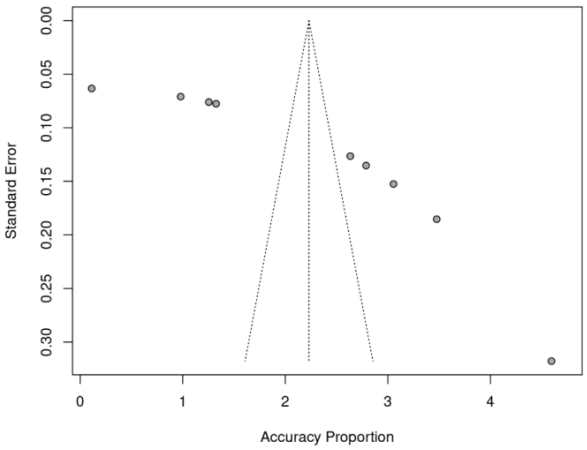

Figure 3: Funnel plot of the studies included in the exploratory meta-analysis (k = 9). Accuracy proportions are plotted against their standard errors. The dashed vertical line represents the pooled estimate, and the triangular region indicates the expected distribution under no small-study effects. This plot is used for descriptive assessment only.

3.4 Subgroup Analysis and Qualitative Insights

3.4.1 Subgroup analysis:

- **Dataset dependency.** Models trained on a single dataset (e.g., ICBHI 2017) tended to report higher internal performance, while studies involving multiple datasets or more heterogeneous data sources generally reported lower but more realistic performance, indicating limited cross-dataset generalization [10, 18].
- **Task complexity.** Studies addressing binary classification tasks generally reported higher accuracy compared to multi-class settings, reflecting the increased complexity and class overlap in multi-class scenarios [4, 11].
- **Imbalance mitigation.** Imbalance handling strategies such as data augmentation and class weighting were commonly employed and were associated with improved model performance; however, the magnitude of improvement varied across studies and could not be consistently quantified due to differences in experimental design [11, 14].
- **Cross-dataset testing.** Only a limited number of studies evaluated models across multiple datasets. Although these studies often reported lower absolute accuracy, they provided stronger evidence of robustness and sensitivity to domain shift effects [10, 18].

Heterogeneity across model families is summarized in Table V. observed variability reflects differences in dataset selection, task formulation, and validation strategies rather than purely statistical variation. Transformer-based and hybrid approaches appeared to exhibit

more stable performance across heterogeneous settings, although this observation remains descriptive.

Table V. Descriptive Summary of Reported Accuracy Ranges Across Model Families.

Model Family	Reported Accuracy Range (%)	Approximate Central Tendency (%)	Dataset Context	Interpretation
CNN	~87–95	~91	Mainly ICBHI 2017	High performance under controlled benchmark settings
RNN/LSTM	~83–93	~88	Mainly ICBHI 2017	Effective temporal modeling with moderate variability
Transformer /Hybrid	~90–96	~93	ICBHI + Private	Strong representation capability and robustness across datasets

Note: Values represent descriptive ranges of reported accuracy extracted from the included studies and are not derived from a formal statistical meta-analysis. Central tendency values are approximate and provided for visualization purposes only.

3.5 Qualitative Insights

3.5.1 Strengths:

Deep learning–based approaches consistently demonstrated strong performance across a variety of respiratory sound analysis tasks, particularly when evaluated on benchmark datasets such as ICBHI 2017. Recent deep learning models have shown a strong ability to learn informative acoustic representations from respiratory sounds, particularly when optimized for spectrogram-based or fused-audio feature learning [25, 26]. Their ability to learn complex acoustic patterns from spectrogram-based representations contributed to improved classification capability under controlled experimental settings[3, 4].

3.5.2 Weaknesses:

Despite these advances, many studies relied on single-dataset evaluation and lacked patient-independent validation or external testing, limiting the generalizability of reported results. In addition, limited integration of multimodal information (e.g., clinical metadata, imaging, or physiological signals) and scarce reporting of explainability mechanisms (XAI) remain important limitations for clinical adoption [16, 17].

3.5.3 Gaps and priorities:

Several methodological gaps were identified across the included studies, including: (i) absence of standardized preprocessing and evaluation protocols (e.g., filtering settings, segmentation strategy, patient-wise splitting), (ii) limited cross-dataset validation and inconsistent reporting practices, and (iii) under-exploration of multi-task learning objectives and severity-aware modeling. Addressing these challenges requires more rigorous experimental design and improved transparency to support reproducibility and real-world applicability [5, 15, 18].

4. Discussion

This systematic review and meta-analysis summarize deep learning (DL) approaches for respiratory sound analysis over the period 2019–2024. Our findings confirm that DL models—especially CNN-based and hybrid/attention architectures—can achieve high benchmark accuracy, yet several limitations remain before these models can be reliably used in clinical settings [3, 4].

The extremely high heterogeneity observed (I^2 = 99.2%) indicates that the pooled accuracy estimate should be interpreted with extreme caution. This level of variability suggests that the included studies differ substantially in terms of datasets, validation protocols, and model configurations. Therefore, the pooled estimate does not represent a stable or generalizable performance indicator but rather a broad summary of highly heterogeneous results.

4.1 Generalizability and Dataset Bias

Many studies rely heavily on the ICBHI 2017 corpus as the main benchmark, which raises concerns about overfitting and limited external validity. Models trained on benchmark datasets such as ICBHI may show noticeable performance drops when evaluated on independent datasets due to differences in data distribution and recording conditions [10-13]. This suggests that current benchmarks may overestimate generalization. Multi-institutional datasets, diverse populations, different recording environments, and explicit external validation with patient-wise splits are essential for more reliable assessment [22, 27].

4.2 Class Imbalance and Rare Phenotypes

Class imbalance remains a recurring problem; common diseases such as COPD and pneumonia are well represented while less common diseases—such as asthma or pulmonary fibrosis—remain underrepresented in respiratory sound datasets. Dianat et al., 2023 [24] specifically addressed the classification of pulmonary sounds for diagnosing interstitial lung diseases associated with connective tissue diseases, highlighting the clinical importance of extending deep learning models beyond common respiratory conditions. Techniques such as category balancing, focal loss, or synthetic data generation using GANs may improve performance in some settings, particularly when class imbalance is severe [11, 14]. However, these methods remain partial substitutes for diverse, real-world data. Practical solutions include federated learning across institutions and collecting additional data for underrepresented categories while maintaining privacy requirements [7].

4.3 Task Design and Multi-Task Learning

Disease detection tasks often achieve higher accuracy (over 93%) than multi-category classification tasks (around 80–85%) or sound level classification tasks, due to the complexity of category boundaries and the increasing noise in labels [4, 11]. Multitasking learning (MTL) remains underutilized, despite its potential to leverage common phonological features across tasks (presence, type, intensity, demographics), thereby enhancing robustness and efficiency [15]. Accurate labeling and clear task ontologies remain essential for expanding MTL adoption.

4.4 Explainability, Calibration, and Clinical Utility

The limited use of explainable artificial intelligence (XAI) techniques is a major obstacle to gaining clinical confidence. While some studies have presented attention maps or displayed gradients, clinically relevant approaches and uncertainty reports remain limited [16, 17, 20]. In addition to traditional metrics, future studies should include calibration evaluation, clinical feasibility analysis (such as decision curves), and error studies across different populations and devices, in line with current reporting guidelines [22, 27].

4.5 Ethical, Privacy, and Deployment Considerations

Breath sounds are useful but incomplete signals on their own. Studies have shown that combining them with demographic information, pulmonary function tests, or radiographic images can improve reliability, especially when field conditions vary [8, 10, 26]. Architectures such as multi- source transformer models or graphical networks allow for better integration of these patterns, but standard interfaces and common test metrics are required to ensure fair comparisons [4].

Unlike previous narrative reviews [3, 4], this study adheres to the PRISMA 2020 framework and adds a meta-analysis covering the period 2019–2024 [22, 27] providing quantitative evidence drawn from 42 studies.

From a clinical perspective, deep learning models for analyzing respiratory sounds can be integrated into digital hearing aids, telemedicine platforms, and mobile health applications, enabling scalable decision support tools for medical and community care professionals, particularly in resource-limited settings.

4.6 Ethical And Data-Privacy Considerations

Most of the datasets used in the studies were public and de-identified—such as ICBHI 2017—significantly reducing privacy risks. However, large-scale data collection in the future will require adherence to ethical standards, including informed consent, anonymization, and approval by ethical authorities. Fairness and bias reduction must also be addressed when training models on different populations to prevent discrepancies in diagnostic performance. Transparent documentation of data sources and adherence to responsible AI principles remain crucial.

5. Strengths and Limitations of this Review

This systematic review and meta-analysis combine several important strengths with some limitations that should be noted.

5.1 Strengths

Strengths. First, this review is, to the best of our knowledge, among the few recent studies that combine a PRISMA-guided design with a structured quantitative synthesis focused specifically on deep learning–based respiratory sound analysis, extending beyond prior review-oriented summaries in the field [19, 20].

Second, the review adhered to the PROBAST-AI and TRIPOD-AI standards, which allowed for a transparent assessment of bias sources and better assurance of the reproducibility of the AI-based diagnostic models. Third, the analysis encompassed multiple families of deep learning models, such as CNNs, RNN/LSTMs, transformers, and hybrids, and evaluated them within a standardized set of metrics, providing a broad comparative framework for researchers and practitioners.

Finally, the use of subcategory analysis, meta- regression, and sensitivity tests helped identify consistent performance patterns while highlighting methodological gaps that warrant future attention.

5.2 Limitations

Despite these advantages, some limitations should be noted. First, the protocol has not been previously registered on platforms such as PROSPERO or OSF, which may limit the traceability and reliability of the process. Second, the funnel plot results showed potential asymmetry, suggesting that small studies with positive results may be overrepresented. However, formal statistical testing for publication bias was not performed due to the limited number of studies included in the pooled analysis. Therefore, the possibility of small effects from small studies cannot be entirely ruled out. Third, some of the included studies relied on small, homogeneous datasets with weak external validation, which may inflate performance estimates and reduce generalizability. Furthermore, publication bias could not be entirely ruled out due to the limited number of studies within some subcategories. Finally, some studies suffered from a lack of detailed methodological information, which affected the consistency of data extraction and analytical comparisons.

However, these limitations do not weaken the overall validity of the findings. Rather, they underscore the strength of larger, multicenter studies based on standard protocols such as TRIPOD-AI and PROBAST-AI to strengthen the evidence and support the future clinical use of these models.

6. Conclusion and Future Directions

This systematic review and meta-analysis highlight the significant progress made in the analysis of respiratory sounds and the classification of lung diseases using deep learning techniques over the past five years. Convolutional CNN (CNN) models and attentional hybrid constructs have demonstrated high performance, often reporting high accuracy under controlled settings in several studies when evaluated using a single dataset within controlled experimental environments. Newer transducer-based models have also shown greater ability to handle data scarcity and domain changes, indicating a shift from feature-based to representation-based learning.

Overall, the results confirm the maturity of deep learning techniques in automated listening, although methodological and application gaps still exist that preclude their immediate clinical use.

Importantly, the prediction interval derived from the meta-analysis ranged from 0.24 to 1.00, indicating that future studies may report substantially lower or higher accuracy depending on the dataset and experimental conditions. This finding reinforces that model performance is highly context-dependent and should not be interpreted as universally reliable.

Key limitations and ongoing challenges. Despite significant progress, several challenges still limit the clinical applicability of these models: (1) *Reliance on limited datasets and poor generalizability:* Over-reliance on the 2017 ICBHI dataset and small, homogeneous groups still inflates internal performance but limits external validity; cross-testing on other groups shows a 10–15% decrease in performance. (2) *Imbalance of categories and rarity of certain disease patterns:* Underrepresentation of diseases such as asthma and pulmonary fibrosis reduces model robustness; although reweighting, artificial incrementing, and focus loss techniques improve performance by 5–7%, they are not a true substitute for data diversity. (3) *Narrow scope of tasks:* Most studies focus on binary tasks, with limited attention to multi-category classification, intensity estimation, or temporal segment identification. (4) *Weak interpretability and clinical validation:* Explainable AI (XAI) applications, model calibration, and bias reviews remain underdeveloped, limiting clinicians' confidence and the readiness of these models for regulatory evaluation.

Future research priorities. To achieve tangible progress in this field, six key research directions stand out:

1) *Large, multi-institutional datasets with clear, standardized reporting:* Development of diverse, patient- based, and standardized databases with transparent metadata, aligned with TRIPOD-AI and PROBAST-AI guidelines.
2) *Multitask Learning (MTL): The adoption* of models capable of predicting disease type, severity, and acoustic events within a single framework based on shared representations, reducing over-personalization and increasing efficiency.
3) *Multimodal Integration:* Linking respiratory sounds with complementary data such as clinical characteristics, medical imaging, and lung function using transducers or graph neural networks to enhance understanding of multidimensional relationships.
4) *Generalization-oriented training:* Adopt interdisciplinary approaches, self-directed or semi-supervisory learning, and federated learning to mitigate biases and reduce data-sharing constraints across medical centers.
5) *Enhance interpretability and clinical value:* Incorporate XAI tools, uncertainty measurements, and decision curve analysis alongside traditional metrics to facilitate clinician adoption of models.
6) *Conduct large-scale prospective field validation:* Conduct multisite studies based on real-world clinical settings, encompassing a variety of equipment, patients, and environments, to assess equity, scalability, and cost-effectiveness.

Final perspective. Addressing these gaps—by expanding data diversity, improving interpretation, and integrating models into multimodal and federated clinical systems—will help transform deep learning-based auscultation from research models into reliable and widely applicable diagnostic tools. As this field develops, these technologies will enable greater opportunities for the early detection of respiratory diseases, improved patient follow-up, and the provision of high-quality respiratory care in resource-limited settings.

Data Availability

The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare no conflicts of interest.

References

[1] Momtazmanesh, S., Moghaddam, S.S., Ghamari, S.-H., Rad, E.M., Rezaei, N., Shobeiri, P., Aali, A., Abbasi-Kangevari, M., Abbasi-Kangevari, Z., Abdelmasseh, M. (2023) Global burden of chronic respiratory diseases and risk factors, 1990–2019: an update from the Global Burden of Disease Study 2019, *EClinicalMedicine* **59**: 101936.

[2] Kim, Y., Hyon, Y., Lee, S., Woo, S.-D., Ha, T., Chung, C. (2022) The coming era of a new auscultation system for analyzing respiratory sounds, *BMC Pulmonary Medicine* **22**: 119.

[3] Nguyen, T., Pernkopf, F. (2022) Lung sound classification using co-tuning and stochastic normalization, *IEEE Transactions on Biomedical Engineering* **69**: 2872-2882.

[4] Wall, C., Zhang, L., Yu, Y., Kumar, A., Gao, R. (2022) A deep ensemble neural network with attention mechanisms for lung abnormality classification using audio inputs, *Sensors* **22**: 5566.

[5] He, W., Yan, Y., Ren, J., Bai, R., Jiang, X. (2024) Multi-view spectrogram transformer for respiratory sound classification. *In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE*, Seoul, South Korea, 14 – 19 April 2024, pp. 8626-8630.

[6] Moon, H.J., Ji, H., Kim, B.S., Kim, B.J., Kim, K. (2025) Machine learning-driven strategies for enhanced pediatric wheezing detection, *Frontiers in Pediatrics* **13**: 1428862.

[7] Albiges, T., Sabeur, Z., Arbab-Zavar, B. (2025) Features and eigenspectral densities analyses for machine learning and classification of severities in chronic obstructive pulmonary diseases, *Intelligence-Based Medicine* **11**: 100217.

[8] Albakaa, Z.H., Alb-Salih, A.T. (2024) An Evolutionary Deep Learning for Respiratory Sounds Analysis: A Survey. *In Proceedings of the International Conference on Forthcoming Networks and Sustainability in the AIoT Era*, *Springer*, Istanbul, Türkiye, 27–29 January 2024, pp. 217-235.

[9] Sabry, A.H., Bashi, O.I.D., Ali, N.N., Al Kubaisi, Y.M. (2024) Lung disease recognition methods using audio-based analysis with machine learning, *Heliyon* **10**: e26218.

[10] Zhang, Y., Huang, Q., Sun, W., Chen, F., Lin, D., Chen, F. (2024) Research on lung sound classification model based on dual-channel CNN-LSTM algorithm, *Biomedical Signal Processing and Control* **94**: 106257.

[11] Erlangga, M.D., Faisal, M.R., Muliadi, M., Indriani, F., Kartini, D., Satou, K. (2025) Incorporating MFCC Feature Extraction to the Classification of Respiratory Sounds by Machine Learning Algorithms. *In Proceedings of the 2025 International Conference on Computer Sciences, Engineering, and Technology Innovation (ICoCSETI)*, *IEEE*, Jakarta, Indonesia, January 21, 2025, pp. 301-306.

[12] Ruchonnet-Métrailler, I., Siebert, J.N., Hartley, M.-A., Lacroix, L. (2024) Automated interpretation of lung sounds by deep learning in children with asthma: scoping review and strengths, weaknesses, opportunities, and threats analysis, *Journal of Medical Internet Research* **26**: e53662.

[13] Fraiwan, M., Fraiwan, L., Khassawneh, B., Ibnian, A. (2021) A dataset of lung sounds recorded from the chest wall using an electronic stethoscope, *Data in Brief* **35**: 106913.

[14] Bacanin, N., Jovanovic, L., Stoean, R., Stoean, C., Zivkovic, M., Antonijevic, M., Dobrojevic, M. (2024) Respiratory condition detection using audio analysis and convolutional neural networks optimized by modified metaheuristics, *Axioms* **13**: 335.

[15] Suma, K., Koppad, D., Kumar, P., Kantikar, N.A., Ramesh, S. (2024) Multi-task learning for lung sound and lung disease classification, *SN Computer Science* **6**: 51.

[16] Fava, A., Dianat, B., Bertacchini, A., Manfredi, A., Sebastiani, M., Modena, M., Pancaldi, F. (2024) Pre-processing techniques to enhance the classification of lung sounds based on deep learning, *Biomedical Signal Processing and Control* **92**: 106009.

[17] Kim, J.-W., Toikkanen, M., Choi, Y., Moon, S.-E., Jung, H.-Y. (2024) Bts: Bridging text and sound modalities for metadata-aided respiratory sound classification. *In Proceedings of the 25th Annual Conference of the International Speech Communication Association (INTERSPEECH 2024)*, *International Speech Communication Association (ISCA)*, Kos Island, Greece, 1–5 September 2024, pp. 690–1694.

[18] Kim, J.-W., Toikkanen, M., Jalali, A., Kim, M., Han, H.-J., Kim, H., Shin, W., Jung, H.-Y., Kim, K. (2025) Adaptive metadata-guided supervised contrastive learning for domain adaptation on respiratory sound classification, *IEEE Journal of Biomedical and Health Informatics* **29**: 5381-5393.

[19] Chen, J., Guo, Z., Xu, X., Jeon, G., Camacho, D. (2024) Artificial intelligence for heart sound classification: A review, *Expert Systems* **41**: e13535.

[20] Wang, Y., Wahab, M., Hong, T., Molinari, K., Gauvreau, G.M., Cusack, R.P., Gao, Z., Satia, I., Fang, Q. (2024) Automated Cough Analysis with Convolutional Recurrent Neural Network, *Bioengineering* **11**: 1105.

[21] Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E. (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, *BMJ* **372**: n71.

[22] Collins, G.S., Moons, K.G., Dhiman, P., Riley, R.D., Beam, A.L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J.B., Van Smeden, M. (2024) TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods, *BMJ* **385**: e078378.

[23] Taloba, A.I., Matoog, R. (2025) Detecting respiratory diseases using machine learning-based pattern recognition on spirometry data, *Alexandria Engineering Journal* **113**: 44-59.

[24] Dianat, B., La Torraca, P., Manfredi, A., Cassone, G., Vacchi, C., Sebastiani, M., Pancaldi, F. (2023) Classification of pulmonary sounds through deep learning for the diagnosis of interstitial lung diseases secondary to connective tissue diseases, *Computers in Biology and Medicine* **160**: 106928.

[25] Gupta, R., Singh, R., Travieso-González, C.M., Burget, R., Dutta, M.K. (2024) DeepRespNet: A deep neural network for classification of respiratory sounds, *Biomedical Signal Processing and Control* **93**: 106191.

[26] Truong, T., Lenga, M., Serrurier, A., Mohammadi, S. (2024) Fused audio instance and representation for respiratory disease detection, *Sensors* **24**: 6176.

[27] Moons, K.G., Damen, J.A., Kaul, T., Hooft, L., Navarro, C.A., Dhiman, P., Beam, A.L., Van Calster, B., Celi, L.A., Denaxas, S. (2025) PROBAST+ AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods, *BMJ* **388**: e082505.