Digital stethoscopes and AI in the detection of respiratory diseases in Primary Care: A Systematic Review

Nassma Elarbi & Sara Ceuppens, Universiteit Antwerpen

Promotor: Prof. Dr. Samuel Coenen, Universiteit Antwerpen

Master of Family Medicine

Masterproef Huisartsgeneeskunde Academiejaar:
2023 – 2026

Deze masterproef is een examendocument dat niet werd gecorrigeerd voor eventueel vastgestelde fouten. Zonder voorafgaande schriftelijke toestemming van zowel de promotor(en) als de auteur(s) is overnemen, kopiëren, gebruiken of realiseren van deze uitgave of gedeelten ervan verboden. Voor aanvragen tot of informatie i.v.m. het overnemen en/of gebruik en/of realisatie van gedeelten uit deze publicatie, wendt u tot de universiteit waaraan de auteur is ingeschreven.

Voorafgaande schriftelijke toestemming van de promotor(en) is eveneens vereist voor het aanwenden van de in dit afstudeerwerk beschreven (originele) methoden, producten, schakelingen en programma's voor industrieel of commercieel nut en voor de inzending van deze publicatie ter deelname aan wetenschappelijke prijzen of wedstrijden.

Table of Contents

1 Acknowledgements 2
2 List of abbreviations 2
3 Glossary 2
4 Abstract 3
5 Introduction 4
6 Method 5
7 Results 13
8 Discussion 19
9 CONCLUSION 23
10 References 24
11 Appendix 26

1 Acknowledgements

We would like to express our sincere gratitude to our promoter, Prof. Dr. Samuel Coenen, for his valuable guidance, critical feedback, and continuous support throughout this research project.

Next, our thanks goes to our clinical instructors from every practice and hospital we worked at. Their support and understanding while working on this thesis made the process feasible.

We also wish to thank the University of Antwerp and ICHO for providing the academic and training framework.

Finally, we would like to thank our family and friends for their support, patience, and encouragement during the completion of this thesis.

2 List of abbreviations

(O)CEBM: (Oxford) Centre for Evidence-Based Medicine
95% CI: 95% Confidence Interval
AI: Artificial Intelligence
AMSTAR: A MeaSurement Tool to Assess systematic Reviews
ANN: Artificial Neural Network
ARI: Acute Respiratory Infection
CASP: Critical Appraisal Skills Program
HRCT: High-Resonance Computed Tomography
ICBHI: International Conference on Biomedical and Health Informatics
MeSH: Medical Subject Headings
ML: Machine Learning
NB: Nota Bene
PEARCH: Pneumonia Etiology Research for Child Health
PRISMA: Preferred Reporting Items for Systematic reviews and Meta-Analyses
PROBAST: Prediction model Risk Of Bias ASsessment Tool
QUADAS: Quality Assessment of Diagnostic Accuracy Studies
RCT: Randomized Controlled Trial
RSV: Respiratory Syncytial Virus
SANRA: Scale for the quality Assessment of Narrative Review Articles
SVM: Support Vector Machine
URTI: Upper Respiratory Tract Infection
WHO: World Health Organization

3 Glossary

AI: Artificial intelligence (AI) refers to systems that display intelligent behaviour by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals.

Artificial Neural Network (ANN): a computational model inspired by the structure and functioning of the human brain. It consists of a network of interconnected nodes, called artificial neurons, that process input information and generate output.

Consolidation: a chest X-ray finding where air in the lungs is replaced by fluid, pus, blood, or cells, appearing as a dense, white (opaque) patch.

Crackles: discontinuous, nonmusical, brief clicking or rattling lung sounds heard on auscultation, typically during inhalation. They represent the explosive opening of small airways or air passing

through secretions. Often described as sounding like Velcro or hair rubbing together, they indicate underlying respiratory conditions. Sometimes called ‘rales’.

Deep Learning: Deep learning is a subset of machine learning driven by multilayered neural networks whose design is inspired by the structure of the human brain.

Digital/electronic stethoscope: a medical device that converts acoustic body sounds—such as heartbeats or lung murmurs—into electrical signals, which are then amplified, filtered, and processed for improved diagnostic accuracy.

Ground glass opacities: hazy, misty areas of increased density in the lungs detected on CT scans where underlying blood vessels and bronchial walls remain visible. They signify partial filling of air sacs (alveoli) with fluid/cells, or thickening of the alveolar walls, indicating conditions like pneumonia, infection, inflammation, or early-stage malignancy.

Machine learning: the subset of artificial intelligence (AI) focused on algorithms that can “learn” the patterns of training data and, subsequently, make accurate inferences about new data. A subset of artificial intelligence (AI) that enables computers to learn from data and make decisions, predictions, or generate content without being explicitly programmed.

Neural networks: a machine learning model that stacks simple "neurons" in layers and learns pattern-recognizing weights and biases from data to map inputs to outputs.

Point-of-care test: tests on patients or patients' specimens like blood, saliva, urine, or faeces, which can be used near the patient, without interference of a laboratory.

P-value: the probability of the observed data given that the null hypothesis is true, which is a probability that measures the consistency between the data and the hypothesis being tested if, and only if, the statistical model used to compute the p-value is correct.

Rhonchi: low-pitched wheezes.

Wheeze: a high-pitched, whistling sound that can occur during breathing when the airways in the lungs become narrowed or blocked.

4 Abstract

Objectives

This systematic review aimed to evaluate the diagnostic value of digital stethoscopes, with and without artificial intelligence (AI), for the assessment of respiratory diseases, with a focus on their potential applicability in primary care. The initial objective was to assess their role in the diagnosis of pneumonia, but this was broadened due to limited available evidence.

Methods

A systematic literature search was conducted in PubMed, Embase, Web of Science, Cochrane Library, and ScienceDirect up to March 2026. Studies were included if they evaluated digital stethoscopes, with or without AI, for the analysis of respiratory sounds. Screening was performed in three stages (title, abstract, full text) according to predefined inclusion and exclusion criteria. Methodological quality was assessed using validated tools (AMSTAR-2, QUADAS-2, PROBAST, CASP, SANRA, and JBI). Studies were categorized by design and level of evidence using the Oxford Centre for Evidence-Based Medicine (OCEBM) framework. Due to heterogeneity, results were synthesized descriptively.

Results
A total of twenty-seven studies were included, comprising systematic reviews, diagnostic accuracy studies, machine learning studies, and observational studies. Evidence regarding the direct clinical benefit of digital stethoscopes compared to conventional auscultation was limited and inconsistent. Some studies suggested improved sound detection and recording capabilities, but no clear diagnostic superiority was demonstrated.
In contrast, AI-based models showed high performance in the classification of respiratory sounds, with reported accuracies frequently exceeding 90%. However, these findings were predominantly derived from controlled datasets and technical environments, with limited external validation. Overall, methodological quality varied, with a generally high risk of bias observed in many studies, particularly in machine learning research.

Discussion
Digital stethoscopes show potential as supportive tools for recording, amplifying, and standardizing respiratory sound assessment. However, current evidence does not demonstrate clear added diagnostic value in clinical practice. AI-based approaches appear promising for sound classification but lack sufficient validation in real-world primary care settings. Importantly, most studies were conducted in hospital or experimental environments, limiting their applicability to primary care. Further high-quality studies in primary care are needed to determine clinical utility, cost-effectiveness, and impact on patient outcomes.

Keywords: Digital stethoscope, artificial intelligence, lung auscultation, respiratory sounds, primary care

5 Introduction

Countries in the northern hemisphere encounter spikes in acute respiratory infections (ARI) caused by various pathogens, such as coronavirus (COVID-19), respiratory syncytial virus (RSV), influenza, and other seasonal viruses and bacteria. The simultaneous presence of these pathogens contributes to outbreaks of respiratory infections, increasing pressure on healthcare systems due to increased primary care visits and hospitalizations, especially for patients with comorbidities and risk factors for severe complications [1]. In clinical practice, however, the challenge often lies not in identifying the causative pathogen, but in accurately diagnosing and managing respiratory diseases, including complications such as pneumonia, which together account for millions of deaths annually worldwide [2].

Traditionally, physicians rely on conventional methods of gathering patient history and clinical examination to diagnose respiratory diseases. While these approaches are invaluable in medical practice, they also have limitations such as subjectivity, discrepancies among clinicians and the need for experience and expertise to recognize subtle signs and symptoms. These limitations can lead to diagnostic delays, over- or underdiagnosis and suboptimal treatment plans [2].
In modern medicine, technological advancement plays a pivotal role in enhancing diagnostic accuracy, optimizing treatment plans, and improving the efficiency of healthcare provision. Particularly promising is its potential in the early detection of respiratory diseases, a common and often challenging aspect of primary care medicine [3][4]. These conditions, ranging from mild upper respiratory illnesses to severe lower respiratory diseases such as pneumonia and chronic obstructive pulmonary disease (COPD), contribute significantly to healthcare demands worldwide.

A promising technological development with potential to improve diagnostic accuracy involves the integration of a digital stethoscope. These stethoscopes capture and interpret bodily sounds such as heart and lung sounds. This enables physicians to analyze sounds with unprecedented precision and to detect subtle variations that may go unnoticed with conventional auscultation. A couple examples

of the most used digital stethoscopes are the 3M Littman Core and Eko Core 500 Digital Stethoscopes, with prices ranging between $300-500 [5, 6].

Through the application of Artificial Intelligence (AI) algorithms to the recorded sound data, digital stethoscopes can identify and interpret patterns characteristic of various respiratory infections. This can support physicians in the decision-making process by providing additional objective information that complements the traditional clinical assessment.
While these devices offer a convenient and accessible tool for frontline healthcare providers, they are not without limitations, including sensitivity to environmental noise, reliance on power sources, and the risk of diagnostic bias if the underlying AI training data do not adequately represent the patient population [7].

The gap between the theoretical promise and the practical implementation of digital stethoscopes enhanced by AI has not yet been fully bridged. Questions remain regarding the effectiveness, accuracy, and cost-effectiveness of this technology [8]. Hence, it remains uncertain whether the digital stethoscope meets the criteria outlined by Cals and Van Weert for the incorporation of new point-of-care tests into general practice (table S1); this study aims to clarify this issue [9].
This systematic literature review focuses on researching the advantages of the digital stethoscope in the early detection of respiratory diseases in primary care. By providing an analysis of the available evidence, this study aims to shed light on the potential benefits, limitations, and clinical applications of this emerging technology, paving the way for future guidelines and practice recommendations in primary care.

6 Method

To ensure transparent and comprehensive reporting, the PRISMA 2020 checklist will be applied. This checklist is based on the PRISMA reporting guideline [10]. A preliminary and incomplete version of the checklist is included in the Appendix (Figure S2, Figure S3). The PRISMA guideline is designed to improve the quality of reporting. It does not primarily focus on the methodological conduct of systematic reviews. To guide the methodological conduct of this review, the Cochrane Handbook for Systematic Reviews of Interventions will be used [11]. This covers study planning, literature searching, data collection, risk of bias assessment, statistical analysis, and interpretation of results.

6.1 Research question and rationale

The original aim for this thesis was to evaluate whether digital stethoscopes combined with AI improve the diagnosis of pneumonia in primary care, translated into the following PICO:
‘Does the use of digital stethoscopes combined with AI (I) lead to a more accurate diagnosis of pneumonia (O) in patients in primary care who present with respiratory symptoms (P) when compared to the traditional diagnosis based on history and clinical examination (C)?’
To conduct thorough literature research this PICO needs to be translated into search terms, specifically (I) and (P).
Since (P) ‘respiratory symptoms’ is a broad term, it was narrowed to ‘cough’ or ‘acute cough’. This is because coughing is one of the most common symptoms of respiratory infections. Primary Care was specified as an inclusion criterion instead of a specific search term. For (I) digital stethoscopes’ will be used for digital stethoscope and ‘AI’ or later ‘Deep Learning’ for artificial intelligence, since in researching it became clear these terms are often used as synonyms.

Preliminary explorations revealed zero or few relevant studies matching this specific focus, highlighting a research gap. Therefore, the research question was revised to assess the diagnostic value of digital stethoscopes (with or without AI) for respiratory diseases through analysis of respiratory sounds. A detailed description of the search strategy, including the adjustments made in terms and scope, is provided in section 5.2.

6.2 Search strategy

Comprehensive literature searches were conducted in PubMed, Embase, Web of Science, Cochrane Library and ScienceDirect, covering all available years until July 2025. The choice of these databases is based on two factors. One is that they are known scientific databases that include a substantial number of studies concluded in the medical field. The second reason is the unavailability of some databases to the students and personnel of the University of Antwerp, which provides access to many databases but not to all databases.

The literature search was initiated in PubMed. Initially, the Clinical Queries tool was applied to conduct two searches (Table 1); however, these did not yield articles meeting the predefined inclusion criteria. Therefore, the traditional search interface was used instead. The search based on terms started out by using combinations of 'digital stethoscope', 'cough', 'AI' and, where possible, Medical Subject Headings (MeSH) Terms were used (Table 2). When trying to combine all three search terms, the search resulted in zero results. When using these same terms in other scientific databases, a few articles were retrieved, but none were relevant to the research subject. This means there is a research gap, there are no existing studies conducting research that matches the original PICO. This led to a re-evaluation of the PICO and research subject of this thesis. Nota bene (NB) when searching PubMed with the following terms: 'pneumonia' and 'digital stethoscope' and 'artificial intelligence', two articles describing the use of AI in diagnosing pneumonia were retrieved. However, none of those matched the PICO.

Clinical Query 1	((digital stethoscope) AND respiratory sounds) AND (Diagnosis/Broad[filter])	59 results
Clinical Query 2	((digital stethoscope) AND "Respiratory Sounds"[MeSH]) AND (Diagnosis/Broad[filter])	44 results

Table 1: Clinical Query searches on PubMed

Search 1	((digital stethoscope) AND ("cough"[MeSH Terms])) AND ("artificial intelligence"[MeSH Terms])	0 results
Search 2	(cough) AND ((digital stethoscope) AND (artificial intelligence [MeSH Terms]))	0 results
Search 3	((digital stethoscope) AND (artificial intelligence)) AND (cough)	0 results

Table 2: Searches on PubMed including 'artificial intelligence'.

The search was broadened, using the same search terms, but adding 'digital stethoscope' OR 'AI'. This led to 107 results. Since 'Deep Learning' was often used as a synonym for 'AI', a new search was conducted using the previous search terms while adding in 'AI' OR 'Deep Learning'. This led to 116 results (Table 3). These results were more promising. However, when scanning the articles based on title and abstract, these articles proved to be too technical and strayed too far from the original research subject. Since this thesis is meant to be relevant for primary care clinicians, technical information about different AI programs will not be the main research goal.

Search 4	("Cough"[Mesh]) AND "Artificial Intelligence"[Mesh]	106 results
Search 5	("Cough"[Mesh]) AND ("Artificial Intelligence"[Mesh] OR (digital stethoscope))	107 results

Search 6	("Cough"[Mesh]) AND (("Artificial Intelligence"[Mesh] OR ("deep learning"[All Fields])) OR (digital stethoscope))	116 results

Table 3: Searches on PubMed including 'artificial intelligence' and/or 'deep learning'.

By searching for 'cough' multiple articles were written about specific AI programs that were used to analyze the sound of a cough. However, that is not the subject of this thesis. Therefore, 'cough' was replaced by 'respiratory sounds' (Table 4). This led to 243 results, all of which were again very technical.

Search 7	("Respiratory Sounds"[Mesh]) AND ("Artificial Intelligence"[Mesh] OR (digital stethoscope))	228 results
Search 8	("Respiratory Sounds"[Mesh]) AND (("Artificial Intelligence"[Mesh] OR (deep learning)) OR (digital stethoscope))	243 results

Table 4: Searches on PubMed including 'respiratory sounds' instead of 'cough'.

The previous searches proved too far off from the proposed thesis subject. This led to a new question: how much is known about the use of digital stethoscopes in primary care? When searching for 'digital stethoscope' and 'primary care, there were three results, but none were relevant. However, searching for 'digital stethoscope' and 'respiratory sounds' led to fifty-one results (Table 5). This means research about using a digital stethoscope in analyzing respiratory sounds has been conducted and this information can be collected and reviewed. The use of digital stethoscopes for other purposes (cardiovascular, abdominal) will not be the scope of this thesis.

Search 9	(digital stethoscope) AND (Primary Care)	3 results
Search 10	(digital stethoscope) AND (Respiratory Sounds [MeSH Terms])	51 results

Table 5: Searches on PubMed including 'primary care' and/or 'respiratory sounds'.

Based on the previously described research, the subject of this thesis was changed to the usefulness of digital stethoscopes (with or without AI) for analyzing respiratory sounds and thereby diagnosing respiratory diseases. Translated into PICO: 'Does the use of digital stethoscopes, whether in combination with AI or not (I), lead to a more accurate diagnosis of respiratory diseases (O) in patients in primary care who present with respiratory symptoms (P) when compared to the traditional diagnosis based on history and clinical examination (C)? No studies that take place in primary care were found; however, the intention of this thesis remains to answer the question of whether these digital stethoscopes would be useful in primary care. In other words, the retained articles will be analyzed to assess whether the available research is applicable in primary care. If not, we can conclude that research about using digital stethoscopes in primary care is missing from multiple databases and depending on the results of this systematic review may or may not be useful to conduct in the future.

During the process of drafting this thesis, the search results will be updated (based on the last search, labelled 'Search 10' (Table 5)) every few weeks until 2 weeks prior to submitting, to make sure the most recent articles are included. The literature found is collected in Endnote.

6.3 Eligibility criteria

The inclusion criteria used are:

- Studies conducted in the medical field
- Studies regarding human subjects

- Articles that are written in Dutch or English
- Articles regarding digital stethoscopes with or without the use of AI
- Articles regarding the use of digital stethoscopes for respiratory sounds

The exclusion criteria used are:
- Articles written in languages other than Dutch or English
- Articles requiring payment for full text access
- Studies focusing on cardiovascular or abdominal auscultation of digital stethoscopes

In addition to the predefined exclusion criteria, studies were considered eligible if they involved the use of a digital stethoscope, with or without AI, for one or more of the following purposes:
- Comparing the diagnostic value of digital stethoscopes with that of conventional stethoscopes
- Recording and digitizing respiratory sounds for diagnostic applications, including:
    - Developing new deep-learning models for respiratory sound classification
    - Validating existing deep-learning models for respiratory sound classification
- Assessing the diagnostic performance of digital stethoscopes and/or AI-based algorithms, either alone or in comparison with established diagnostic tests (e.g., spirometry, imaging)

The main scope of this study is the digital stethoscope as a diagnostic tool. Therefore, ideal sources for this research would be meta-analyses reviewing the diagnostic properties of the digital stethoscope. When following The Centre of Evidence Based Medicine (CEBM) Level of Evidence Guide [12, 13] (Figure S1), since the goal is to conduct research on the diagnostic accuracy of the digital stethoscope, RCT's and cohort studies are less relevant, and cross-sectional studies should be searched for, after meta-analyses of cross-sectional studies. However, since articles containing information on digital stethoscope relevant to this study are limited, the study design was not included as an exclusion criterion.

Book chapters, conference abstracts, study protocols, and letters to the editor were excluded from the review.

6.4 Study selection

Duplicates were removed prior to screening. Subsequently, articles that were not relevant to the scope of this review were excluded. The selection process was conducted according to the predefined inclusion and exclusion criteria, focusing on whether each study aligned with the research objectives described above. Screening was performed in three stages: title screening, abstract screening, and full-text review of the remaining articles. The overall selection process is presented in the PRISMA flow diagram (Section 6.1). Both reviewers independently screened titles and abstracts against the predefined criteria.

6.5 Quality assessment

To ensure the methodological rigor and reliability of the findings, a thorough quality assessment of all eligible studies was conducted. This process was performed independently by two reviewers. The methodological quality of the included studies was assessed using validated critical appraisal tools tailored to the study design. For systematic reviews, the AMSTAR-2 (A Measurement Tool to Assess Systematic Reviews) checklist [14] was applied. Diagnostic accuracy and comparison studies were evaluated using QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2) [15]**.**
For prediction model studies involving artificial intelligence, the PROBAST (Prediction model Risk Of Bias ASsessment Tool) [17] was used to assess both risk of bias and applicability. Observational studies were appraised using the CASP (Critical Appraisal Skills Programme) checklist [18], while case reports were evaluated using the Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Case Reports [19].

Narrative reviews were assessed using the SANRA (Scale for the Assessment of Narrative Review Articles) tool [21].
The results of the quality assessment were structured according to domains specific to each tool, including risk of bias and applicability concerns where appropriate. Any discrepancies between reviewers were resolved through discussion and consensus. In cases of persistent uncertainty, methodological decisions were discussed with the supervising researcher.
Studies were also classified according to their methodological design and assigned a level of evidence based on the Oxford Centre for Evidence-Based Medicine (OCEBM) 2011 hierarchy [13].

6.6 Data analysis

During the literature search, it became evident that the broad inclusion criteria resulted in a highly heterogeneous body of literature. The studies varied considerably in terms of outcomes, study designs, and statistical approaches. Consequently, a pooled statistical analysis or synthesis in a single comprehensive table was not feasible. Instead, the results were summarized descriptively, with relevant data highlighted and comparable studies presented together. Where available, diagnostic performance measures such as sensitivity, specificity, and accuracy will be reported, as these metrics were provided in most studies. In some studies, there is mention of odd's ratio, kappa's K, and other outcome metrics. The most relevant metrics are further explained in Box 1. For subsets of studies where statistical comparison is feasible, additional analyses will be conducted using Microsoft Excel. Given the heterogeneity of study populations, interventions, and outcome reporting, a formal meta-analysis was not feasible.

7 Results

7.1 Study selection

The database search retrieved 300 articles. After removal of duplicates, 221 articles were screened based on title and abstract, resulting in the exclusion of 162 articles. Five articles for which full-text access was not available were subsequently removed, leaving 54 articles for full-text assessment. Of these, 27 articles were excluded for the following reasons: inappropriate setting (no diagnostic component or not relevant to the selection criteria; n=17), incorrect intervention (no thoracic auscultation using a digital stethoscope; n=5), or lack of relevant outcomes (no clear conclusion reported or outcome not yet available; n=5). Ultimately, twenty-seven articles were included in the review. The study selection process is depicted in the PRISMA flow diagram (Figure 1).

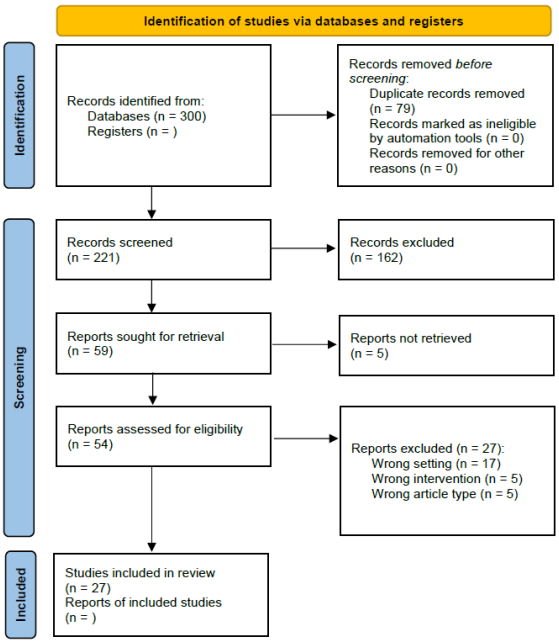


Figure 1: Prisma 2020 flow diagram for new systematic reviews which included searches of databases and registers only. Databases used: Cochrane (n=5), Embase (n=38), PubMed (n=82), ScienceDirect (n=57), Scopus (n=39), Web of Science (n=79)

An overview of the collected article types is provided in Table 6.

Study type	**Main focus**	**Number (n)**	**OCEBM level**
Systematic review	Digital auscultation for childhood pneumonia diagnosis	1	Level 1
Systematic review	Respiratory sound characteristics / AI methodology	2	Level 1
Diagnostic accuracy study	Clinical diagnostic performance	2	Level 2
Observational diagnostic study	Associations with clinically relevant outcomes	1	Level 2
Diagnostic comparison study	Comparison with conventional auscultation	2	Level 3
Diagnostic modelling study	Comparison with expert classification or another diagnostic method	2	Level 3
AI-based diagnostic prediction model study	Respiratory sound classification (mainly technical)	5	Level 2–5
AI-based diagnostic prediction model study	Respiratory disease classification (mainly technical)	9	Level 2–5
Cross-sectional observational study	Descriptive clinical auscultation findings	1	Level 4
Case report	Illustrative clinical use	1	Level 5
Narrative review	Background / technical overview	1	Level 5

Table 6: Overview of included studies by study design, main focus, and corresponding levels of evidence according to the Oxford Centre for Evidence-Based Medicine (OCEBM) 2011 hierarchy [13].

The included studies were highly heterogeneous in terms of design, methodology, and clinical applicability. Only a limited number directly evaluated the diagnostic value of digital stethoscopes or AI-assisted auscultation in clinically relevant settings, whereas a substantial proportion focused mainly on AI-based respiratory sound classification or technical model development. As a result, the included literature only partially addressed the research question, with stronger evidence available for technical

sound classification than for the diagnostic performance and implementation of these tools in primary care.

7.2 Quality assessment

The included studies were systematically assessed for methodological quality according to the methods described in ‘5.4. Quality assessment’.

Detailed quality judgements, including domain-specific risk of bias evaluations, are summarized in the appendix page 69-81.

The methodological quality of the included systematic reviews, assessed using the AMSTAR-2 tool, is presented in Table 7.

		Condition	
		Present	**Not present**
Index test	**Positive**	True Positives (TP)	False Positives (FP)
	Negative	False Negatives (FN)	True Negatives (TN)

Metric	**Formula**	**Explanation**
Sensitivity (or Recall in machine learning, sometimes True Positive Rate (TPR))	$\frac{TP}{TP + FN}$	Ability of the index test to correctly identify true positives, compared to a golden standard
Specificity	$\frac{TN}{FP + TN}$	Ability of the index test to correctly identify false positives, compared to a golden standard
False Negative Rate (FNR)	$1 - Sensitivity$ of $\frac{FN}{FP + TP}$	The proportion of cases with predicted condition present, incorrectly identified as negative by the index test
Positive Likelihood Ratio (LR+)	$\frac{Sensitivity}{1 - (Specificity)}$	The extent to which the predicted condition becomes more likely following a positive index result.
Negative Likelihood Ratio (LR-)	$\frac{1 - Sensitivity}{Specificity}$	The extent to which the predicted condition becomes less likely following a negative index result.
Diagnostic Odds Ratio (OR)	$\frac{TP \times TN}{FP \times FN} = \frac{LR+}{LR-}$	The ability of the index to discriminate between cases with and without the predicted condition present
Positive Predictive Value (PPV) (or Precision in machine learning)	$\frac{TP}{TP + FP}$	The probability that the condition is present if the index test is positive
Negative Predictive Value (NPV)	$\frac{TN}{TN + FN}$	The probability that the condition is not present if the index test is negative?
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Measures how close the results are to the 'correct' values.
F1-Score	$2 \times \frac{Precision \times Recall}{Precision + Recall}$	Harmonic mean of precision and recall, providing a single score that balances both
Matthews Correlation Coefficient (MCC)	$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times TN + FN)}}$	Agreement between the model and the actual outcomes, how well the model predicts the observed outcomes.

Metric	**Explanation**
Prevalence	Total portion of a population that has a specific condition at a given time
Cohen's Kappa	Measuring agreement or concordance between observers or between data.
Percentage agreement (simple agreement)	Proportion of agreement between two independent interpreters
Area under the curve (AUC)/receiver operating characteristic (ROC) curve	Overall measure to compare classifiers' ability to discriminate between those with and those without the condition
Positive Percentage Agreement (PPA)	Same metric as sensitivity, but compared to a non-reference standard
Negative Percentage Agreement (NPA)	Same metric as specificity, but compared to a non-reference standard

Box 1: Explanation of metrics

Study	Protocol	Search	Selection & Extraction	Risk of Bias Assessment	Publication Bias	Funding reported	Overall
Abreu et al. (2021)	🟡	🟢	🟢	🟡	🔴	🔴	🟡
Park et al. (2025)	🟢	🟢	🟢	🟢	🟢	🟡	🟢
Ahmed et al. (2022)	🟢	🟢	🟢	🟢	🔴	🔴	🟡

*Table 7: Summary of methodological quality of included systematic reviews (**AMSTAR-2**).*
Legend: Protocol, Search, Selection and extraction, Risk of Bias assessment, Publication Bias, Funding reported correspond to the six AMSTAR-2 domains. 🟢 = Criterion adequately fulfilled; 🟡 = Partially fulfilled or unclear; 🔴 = Criterion not fulfilled. Overall = Overall methodological confidence based on AMSTAR-2 domain-level assessment, considering the presence of critical and non-critical weaknesses.

The methodological quality of the included systematic reviews showed variability across domains, with several items rated as unclear. Critical domains such as protocol registration, justification of study selection and assessment of risk of bias were not consistently reported across all studies.
While some reviews adequately described their search strategies and data extraction methods, deficiencies were observed in the evaluation and interpretation of bias within the included studies.

As a result, the overall methodological confidence varied between studies, ranging from high to moderate, depending on the completeness of reporting and the presence of weaknesses in key domains.

Although we did not formally classify domains as critical or non-critical in the table, our interpretation and conclusions were primarily based on the key domains that are considered critical in AMSTAR-2, particularly those related to bias assessment and methodological transparency.

The methodological quality of the included studies assessed with the QUADAS-2 tool is summarized in Table 8.

Study	Patient Selection	Index Test	Reference Standard	Flow & Timing	Overall
Andrès et al. (2016)	🟢	🟡	🟢	🟢	🟡
McCollum et al. (2020)	🟡	🟢	🟡	🟢	🟡
Gottlieb et al. (2018)	🔴	🟡	🔴	🟢	🔴
Park DE et al. (2021)	🟡	🟢	🔴	🟡	🔴
Kevat et al. (2020)	🟡	🟡	🔴	🟢	🔴
Gupta et al. (2024)	🟡	🔴	🟢	🔴	🔴

Hoekstra et al. (2025)	🟢	🟢	🟡	🟢	🟡

*Table 8: Summary of methodological quality of included diagnostic studies (**QUADAS-2**).*
Legend: Patient Selection, Index Test, Reference Standard, and Flow and Timing correspond to the four QUADAS-2 domains.
🟢 = Low risk of bias; 🟡 = Unclear risk of bias; 🔴 = High risk of bias. Overall = Overall risk of bias judgment based on domain-level assessment (≥1 high risk domain = high overall risk).

The risk of bias varied across studies and domains, with the highest concerns observed in the domains of patient selection and reference standard. Several studies relied on non-representative populations, convenience-based samples, or selected datasets, which may affect the generalizability of the findings. The index test domain generally showed a lower risk of bias, although in some studies the test thresholds were not clearly predefined. Applicability concerns were most prominent in relation to patient selection and reference standards, reflecting differences between study conditions and real-world clinical practice.
Some studies were conducted in simulated environments rather than real clinical settings, which further limits their applicability to primary care practice.
Based on domain-level assessments, the overall risk of bias was judged to be high in most studies.

The methodological quality of the included studies assessed with the PROBAST tool is summarized in Table 9.

Study	Participants	Predictors	Outcome	Analysis	Overall
Park et al. (2023)	🟡	🟢	🔴	🟡	🔴
Chatterjee et al. (2022)	🔴	🟡	🔴	🔴	🔴
Karaarslan (2024)	🔴	🟡	🔴	🔴	🔴
Ma et al. (2019)	🔴	🟢	🟡	🟡	🟡
Hassan et al. (2024)	🔴	🟡	🔴	🔴	🔴
Gupta et al. (2024)	🟡	🟢	🔴	🟡	🔴
Mahmood et al. (2024)	🔴	🟡	🔴	🔴	🔴
Rady et al. (2015)	🔴	🟡	🔴	🔴	🔴
Sharma et al. (2023)	🔴	🟢	🔴	🔴	🔴
Uppin et al. (2023)	🔴	🟢	🔴	🔴	🔴
Zhang P. et al. (2023)	🔴	🟢	🔴	🔴	🔴
Zhang Y. et al. (2024)	🔴	🟢	🔴	🔴	🔴
Topaloglu et al. (2025)	🟡	🟢	🟡	🟡	🟡
Divakaran et al. (2025)	🔴	🟢	🔴	🔴	🔴

Table 9: Summary of methodological quality of included Machine Learning studies (PROBAST).

Legend: Participants, Predictors, Outcome and Analysis correspond to the four PROBAST domains. ● = Low risk of bias; ● = Moderate/unclear risk of bias; ● = High risk of bias. Overall = Overall risk of bias judgment based on domain-level assessment (≥1 high risk domain = high overall risk).

Across the included prediction model studies, which primarily focused on respiratory sound or disease classification in technical or experimental settings rather than direct clinical diagnosis, the risk of bias was generally high across multiple domains. The domains of participants, outcome, and analysis were most frequently rated as high risk, reflecting limitations in study populations, outcome definitions, and analytical approaches. In contrast, the predictors domain generally showed a lower risk of bias, with many studies using standardized and reproducible feature extraction techniques. Based on the domain-level assessments, the overall risk of bias was judged to be high in most studies, with only one study demonstrating an unclear or moderate overall risk of bias.

These findings highlight important concerns regarding the methodological robustness and clinical applicability of current AI-based prediction models for respiratory sound analysis.

7.3 Study Characteristics

Details of the included studies are presented in Appendix Tables S2-S5.

A brief summary of the characteristics of the studies is available in table 10.

	Study category	**Setting of sound collection/ auscultation**	**Control**	**Diagnosis**	**Outcome measurements**
Abreu et al., 2021	●	/	/	/	/
Park et al., 2025	●	/	/	/	●
Ahmed et al., 2022	●	/	/	/	● ●
Andrès et al., 2026	●	●	● ●	● ●	●
Gottlieb et al., 2018	●	●	●	● ● ●	●
Noda et al., 2020	●	●	●	●	●
Park et al., 2021	●	● ●	●	●	● ● ● ●
Wang et al., 2020	●	●	● ●	●	● ●
McCollum et al., 2020	●	●	●	●	●
Gupta et al., 2024	●	●	●	●	● ● ●
Ahmed et al., 2023	●	●	●	●	● ● ● ●
Park et al., 2023	●	● ●	●	●	● ●

Hoekstra et al., 2025	🟡	🟢	🔴	🟢	🟢🔵🟣🔵
Abreu et al., 2021	🔴	🔴	/	🔵	🟢🟡🩷
Gupta et al., 2024	🔴	🟢	/	🟡	🟡🔴
Karaarslan et al., 2024	🔴	🔴	/	🔵	🟡🔵🔴🩷
Kevat et al., 2020	🔴	🟢	/	🟡	🔴
Ma et al., 2019	🔴	🔴	/	🟡	🟢🟡🩷
Nagvi et al., 2020	🔴	🔴	/	🟢🔵	🟡🔴
Mahmood et al., 2024	🔴	🔴	/	🟢🔵🟡	🟡🩷
Rady et al., 2015	🔴	🟢🔴	/	🟡	🟢🟡
Sharma et al., 2023	🔴	🔴	/	🟢🔵	🟡🔴
Uppin et al., 2021	🔴	🔴	/	🟣	🟡
Zhang et al., 2023	🔴	🔴	/	🟢🔵	🟡🔴
Zhang et al., 2024	🔴	🔴	/	🟢🔵	🟡🔴
Divakaran et al., 2025	🔴	🟢🔴	/	🟡	🟡🔴
Topaloglu et al., 2025	🔴	🟢	/	🟡	🟢🟡🔴

Tabel 10: summary of studies and results
Legend:

Study category: 🟢 = systematic review, 🔵= diagnostic comparison of digital stethoscopes, 🟡 = automatic sound classification in comparison to other methods, 🔴 = automatic sound classification without comparison to other methods
Setting: / = not applicable, 🟢 = in hospital, 🔴 = sounds from pre-existing database used
Control: / = not applicable, 🟢 = acoustic stethoscope, 🔵 = other comparison (lab values, spirometry,...), 🟡 = imaging, 🔴 = panel of clinicians listening to recorded sounds
Diagnosis: / = not applicable, 🟢 = pneumonia, 🔵 = other lung diseases (asthma, COPD,...), 🟡 = lung sounds apart from a diagnosis, 🔴 = non-pneumological diagnosis, 🟣 = normal vs. abnormal
Outcome measurements: / = not applicable, 🟢 = sensitivity and specificity, 🟡 = accuracy, 🔵 = agreement (percentage agreement, Cohen's Kappa), 🔴 = other ML-specific values (precision, recall, F1-score, NPA, PPA, TPR), 🟣 =other non-ML-specific values (NPV, PPV, LR+, LR-), 🔵 = diagnostic odds ratio, 🩷 = other outcome

There were no studies found that gave an answer to the main question of this thesis. The studies that came closest to the objective of this thesis and their results are listed above.

7.3.1 Respiratory sounds and classification

A systematic review by Abreu et al. on respiratory sounds in children revealed that the characteristics of respiratory sounds are affected by dimension of airways, airflow, sex, and chest location. In children, wheeze is the most reported adventitious respiratory sound and is believed to be the most promising parameter as outcome measure in studies regarding respiratory sounds and sound detection [22].

Similar studies including adults were not found.

7.3.2 Machine learning and AI detection and classification of respiratory sounds

Machine learning classification of respiratory sounds can be done in multiple ways. There are two-class classifications (normal vs. adventitious/pathological, wheeze vs. crackles), three-class classifications (normal vs. crackles vs. wheeze), four-class classifications (normal vs. crackle vs. wheeze vs. both wheeze and crackle, wheeze vs. rattle vs. stridor vs. normal) [23]

Park et al. wrote a systematic review on the performance of machine learning models in pediatric lung sound analysis reviewing the sensitivity and specificity of different detection models. For wheeze detection models, pooled sensitivity was 90.2%, pooled specificity was 95.5%. A moderate between-study heterogeneity was found. For abnormal lung sound detection models, the pooled sensitivity was 90.7%, pooled specificity was 87.7%. There was low between-study heterogeneity [24].

Another systematic review by Salahuddin et al. focused on machine learning detection and classification of respiratory sounds and pathological conditions. This review showed that accuracy of these models varied from 66.3-100%. Specifically for pneumonia, accuracy was shown to be 86.7%, sensitivity 86.8% and specificity 86.6% [23].

A prediction model study by Park et al. comparing the performance of a machine learning model with physicians on classifying respiratory sounds (normal vs. abnormal, crackle vs. wheezing, normal vs. crackle, normal vs. wheezing) showed that the accuracy of model on external validation ranged between 76.4-91.5%. The model and physicians had comparable outcomes [25].

Hoekstra et al. [26] performed a diagnostic study in which they developed an AI algorithm that was used to classify lung sounds from different chest positions as normal or abnormal, if one or more chest position from the same patient was qualified as 'abnormal', then the patient was qualified as 'abnormal'. The performance of the AI algorithm was compared to a listening panel of three physicians. The adjusted Kappa Statistic, with 95% confidence interval, was 0.662 (0.596, 0.728) for the classification of chest positions and 0.832 (0.718, 0.945) for the classifications of patients.
The AI algorithm was evaluated and gave following values for chest position classification: sensitivity 80.3% (95% CI [75.3%, 84.7%]), specificity 87.2% (95% CI [81.8%, 91.5%] and diagnostic odds ratio 27.7 (95% CI [16.8, 45.7]).
For patient classification they got following results: sensitivity 96.3% (95% CI [89.4%, 99.2%]), specificity 66.7% (95% CI [38.4%, 88.2%]) and diagnostic odds ratio 51.3 (95% CI [11.2, 234]).
Other metrics were used as well; they can be found in the Appendix Tables S2-S5.

Multiple studies were done on machine learning and AI models, either existing models or newly developed models, and their accuracy in classifying different respiratory sounds. Precision, recall and F1-scores were mentioned in a lot of studies, these can be found in tables S2-S5 in the Appendix. The accuracy varied a lot between the different studies, with lowest accuracy 52.79% and highest accuracy 99.5% [27-32].

7.3.3 Machine learning and AI detection and classification of respiratory disorders

Apart from classification of respiratory sounds, some studies showed that these machine learning and AI models can also differentiate between different respiratory disorders. These differentiations include healthy, bronchiectasis, bronchiolitis, COPD, pneumonia, URTI [23]. Some studies focused on healthy vs. non-healthy without specifying the disease.

An observational study by Gupta et al. compared the outcomes of a machine learning model to the outcomes on a spirometry test. This showed that the accuracy of the chosen model was 98.54%. The study claims a strong correlation between the results of the model and spirometry results. [33]

In a prediction model study, Divakaran et al. [34] worked with four different AI models to differentiate between patients with COPD and healthy patients. The best performing model (an Artificial Neural Network) achieved an accuracy of 96.36%, along with precision of 96.23%, recall of 96.10% and F1-score of 96.16%.

Topaloglu et al. [35] created two AI models in a prediction model study. These models were meant to differentiate between patients with asthma and healthy patients. The best performing model (a Quadratic Support Vector Machine) was compared to two existing models and scored higher on accuracy (98.13%), precision (99.88%), sensitivity (99.83%), specificity (99.88%) and F1-score (99.86%).

Multiple studies were done on machine learning and AI models, either existing models or newly developed models, and their accuracy in classifying/diagnosing different respiratory disorders. The accuracies varied a lot between studies, with the lowest accuracy being 90% and the highest accuracy being 100% [30, 36-41].

7.3.4 The digital stethoscope

Andrès et al. compared the eStetho (a Bluetooth electronic stethoscope) with the acoustic 3M Littman® Classic III™ stethoscope in a diagnostic accuracy study. The categories included were COPD, pneumonia, asthma and healthy. The study stated that using the eStetho allowed a correct diagnosis in 10% more cases than when using the acoustic stethoscope [42].

A study by Gottlieb et al. was a diagnostic comparison study comparing the Littmann Cardiology III stethoscope with an Eko Core electronic stethoscope attachment vs. the same stethoscope with a regular microphone in diagnosing cases with respiratory disease showed no significant difference in overall scores between electronic and the acoustic group. It further showed no significant differences in time spent on the diagnosis and in number of times the recordings were played [43].

A case report by Noda et al. about Covid-19 pneumonia compared the auscultatory changes, listened to with a novel wireless stethoscope, to the clinical changes and changes seen on heart-resolution computed tomography (HRCT). This showed that the clearing of clinical signs was accompanied by clear changes in the respiratory sounds [44].

Park et al. conducted a diagnostic comparison study including children with pneumonia that compared commercial digital stethoscopes (ThinkLabs ds32a) with conventional acoustic stethoscopes. This showed a moderate agreement with classification of any crackles and/or wheeze vs. neither crackles nor wheeze and a fair concordance for crackles and wheeze when considered independently. Among the likely pneumococcal pneumonia cases, digital auscultation was more likely to show crackles-only. Furthermore, hearing crackles-only was generally indicative of higher severity and wheeze with reduced severity on both digital and conventional auscultation [45].

A cross-sectional observational study done by Wang et al. on patients with confirmed Covid-19 compared the auscultatory findings by electronic stethoscope (EYA-1A(C)) with radiological findings on chest CT scans and with laboratory parameters. When comparing to chest CT outcomes, these correlations were found: fine and coarse crackles and ground glass opacities (sensitivity 84.6%, specificity 15.2%), fine and coarse crackles and consolidation (sensitivity 88.9%, specificity 15.2%), velcro crackles and ground glass opacities (sensitivity 41%, specificity 81.3%), velcro crackles and

consolidation (sensitivity 44.4%, specificity 67.4%). No statistically significant relation was found between laboratory values and auscultation outcomes [46].

McCollum et al. [47] conducted a diagnostic and prognostic study that looked at the association between digitally recorded lung sounds and radiographic pneumonia, and at the association between digitally recorded lung sounds and mortality in young children hospitalized with pneumonia. This showed that, in children with WHO qualified severe pneumonia, 38.6% had a normal auscultation. The only statistically significant outcome was hearing 'wheeze-only' (in 14.9% of children), with an adjusted (adjusted for sex, age in months, and PERCH site) diagnostic odds ratio for having radiographic pneumonia of 0.35 (95% CI [0.15, 0.82]).
In children with WHO qualified very severe pneumonia, 33.3% had a normal auscultation. The only statistically significant outcome was hearing 'crackles-only' (in 63.6% of children) with an adjusted (adjusted for sex, age in months, and PERCH site) odds ratio for having radiographic pneumonia of 1.45 (95% CI [0.53, 3.93]). No statistically significant relation between auscultation and mortality was found.

8 Discussion

8.1 Principal findings

This systematic review evaluated the diagnostic value of digital stethoscopes, with and without artificial intelligence, for the assessment of respiratory diseases, with a focus on applicability in primary care. Although the initial aim was to evaluate their role in the diagnosis of pneumonia, the scope was broadened due to the limited availability of studies specifically addressing this condition.

In general, the findings suggest that digital stethoscopes have promising capabilities for recording and analyzing respiratory sounds. However, evidence supporting their direct diagnostic superiority over conventional auscultation remains limited. Only a small number of studies directly compared digital and conventional stethoscopes. These studies did not show consistent improvements in diagnostic accuracy. As a result, clear evidence of clinical superiority has not been established.

In contrast, studies focusing on AI-based classification of respiratory sounds reported high performance metrics. In several studies, models achieved accuracy rates above 90% for the classification of specific sound patterns, such as wheezes and crackles. These systems showed strong potential in distinguishing between normal and abnormal lung sounds. However, these results were predominantly derived from controlled datasets and technical environments, raising concerns about their generalizability to real-world primary care settings.

8.2 Interpretation of findings

8.2.1 Clinical utility of digital stethoscopes

The main added value of digital stethoscopes appears to be in sound amplification, recording, and reproducibility. This may help reduce differences between clinicians and improve documentation and follow-up.

However, clinical decision-making in primary care extends beyond sound recognition alone. Diagnosis is based on a combination of patient history, clinical examination, and contextual factors. In this context, improved sound analysis on its own is unlikely to significantly influence clinical outcomes unless it is integrated into broader diagnostic pathways.
Most included studies focused on technical performance rather than clinically relevant endpoints, which further limits their direct applicability to everyday practice. Importantly, none of these studies were conducted in true primary care settings, raising concerns about the external validity of the findings.

Although direct evidence demonstrating superior diagnostic performance compared to conventional stethoscopes remains limited, several features suggest a potential added value of digital stethoscopes in clinical practice. The ability to record and store respiratory sounds may support clinical use by enabling second opinions and improving documentation, while also facilitating education and training. These recordings can contribute to the development of AI models as well.

Digital stethoscopes may also facilitate the expansion of telemedicine. The ability to record and transmit respiratory sounds could support remote assessment, particularly in patients who are unable to attend in-person consultations. However, no studies included in this review specifically evaluated the use of digital stethoscopes for telemedicine in primary care. As such, the clinical value of this application remains unclear.

Overall, these aspects indicate that the current value of digital stethoscopes lies primarily in their supportive role rather than in direct diagnostic superiority.

8.2.2 AI-based classification of respiratory sounds

AI-based models showed high diagnostic performance, particularly in the classification of respiratory sounds such as wheezes and crackles. Several of the included studies reported high accuracy rates, often exceeding 90%, highlighting the technical potential of these approaches. Still, these findings should be interpreted with caution, as most studies relied on pre-recorded datasets with limited diversity, introducing a risk of bias and restricting the generalizability of the results.

In addition, accurate classification of respiratory sounds does not necessarily translate into accurate disease diagnosis. For example, distinguishing crackles from rhonchi does not directly lead to a diagnosis of pneumonia. This highlights an important limitation, as current AI applications mainly operate at the level of pattern recognition of lung sounds rather than clinical diagnosis.

Also, the overall confidence in these findings is tempered by methodological concerns. As reflected in the PROBAST assessment, many prediction model studies were associated with a high risk of bias, particularly in the domains of participant selection and outcome definition. The lack of external validation in heterogeneous, real-world populations further limits the applicability of these models in primary care settings.

8.3 Comparison with existing literature

The findings of this review are largely consistent with previous systematic reviews on the use of digital stethoscopes and AI in respiratory sound analysis [22, 24, 48]. Similar to earlier work, these studies reported high performance of AI models in the classification of respiratory sounds, particularly for detecting wheezes and crackles. For example, Park et al. reported similarly high performance of AI models in respiratory sound classification.

At the same time, previous reviews mainly focused on technical performance and algorithm development, with limited attention to clinical applicability. This is in line with the findings of the present review, where most included studies were conducted in controlled settings, hospital-based populations, or relied on pre-recorded datasets. As a result, the translation of these technologies into routine clinical practice, and more specifically primary care, remains uncertain.

In addition, earlier reviews have highlighted methodological concerns, including heterogeneity between studies and limitations in study design [22, 24, 48]. These issues were also observed in the current review, where risk of bias was frequently high and external validation was often lacking.

A key contribution of this review is the explicit focus on primary care. While previous literature has mainly addressed technical aspects, this review emphasizes the gap between high reported performance of AI models and their limited applicability in real-world primary care settings.

8.4 Implications for primary care practice

The implementation of digital stethoscopes combined with AI in primary care is still in an early stage of the translational research pathway, often described as the transition from "bench to bedside." From a Belgian primary care perspective, this early stage is particularly evident.

Frameworks for the implementation of diagnostic tools in general practice, such as those proposed by Cals and Van Weert [9], emphasize that new technologies should demonstrate clear clinical utility, cost-effectiveness, and a measurable impact on patient outcomes before they can be widely adopted.

From the perspective of this framework, digital stethoscopes combined with AI currently fulfill only a limited number of criteria. While technical performance in terms of sound amplification and classification appears promising, important aspects such as test accuracy in real-world settings, impact on clinical decision-making, and improvement of patient outcomes remain insufficiently demonstrated. In addition, evidence regarding cost-effectiveness and long-term implementation in routine primary care is lacking. As such, these technologies can be considered to be in an early stage of the translational pathway and do not yet meet the requirements for widespread adoption in general practice.
This is also reflected in the findings of this review, where most studies primarily focused on technical performance or sound classification rather than clinically relevant outcomes or real-world applicability [25, 27, 29-32, 36, 37, 40, 41, 49].

In Belgian general practice, the diagnosis of respiratory infections is primarily based on clinical assessment, in line with guidelines such as those from Domus Medica. Additional tests, such as C-reactive protein (CRP) testing, are used selectively when indicated. Within this context, AI-assisted auscultation does not yet have an established role.

Several factors may explain this limited uptake. As highlighted in the included studies, evidence derived from primary care populations is lacking, with most data originating from controlled or highly selected settings [25, 27, 29-32, 36, 37, 40, 41, 49] . Besides that, uncertainty remains regarding cost-effectiveness, and practical challenges related to integration into the consultation workflow should not be underestimated. The need for appropriate training may further complicate implementation in routine practice.

As with most diagnostic tools in primary care, digital stethoscopes combined with AI should be interpreted as adjuncts to clinical reasoning rather than standalone diagnostic instruments. Their potential may lie in supporting clinical practice, for example in education, telemedicine, or the documentation of lung sounds. They cannot replace clinical reasoning or established diagnostic strategies, but rather complement the clinical assessment performed by healthcare professionals.

8.5 Strengths and limitations

This systematic review has several strengths. First, a structured and transparent methodology was applied, supported by the Cochrane Handbook and guided by the PRISMA 2020 reporting guidelines. This ensured both methodological rigor and clear reporting.
Second, a broad inclusion strategy was used, allowing different study designs to be included. Given the limited availability of research on digital stethoscopes in this field, this approach provided a more complete overview of the existing evidence.
Third, the methodological quality of the included studies was systematically assessed using validated

tools tailored to each study design, including AMSTAR-2, QUADAS-2, PROBAST, CASP, SANRA, and JBI [14, 15, 17-19, 21]. This structured appraisal improves transparency and supports a more nuanced interpretation of the findings.

Several limitations should also be considered. The included studies were highly heterogeneous in terms of design, populations, interventions, and outcome measures. This heterogeneity precluded quantitative synthesis and limited comparability between studies. The inclusion of a narrative review, assessed using SANRA, also introduced limitations, as formal risk-of-bias tools are less well established for this type of study [44].
In addition, this review has its own methodological limitations. Although multiple databases were consulted, relevant studies may have been missed, and only articles published in English and Dutch were included, which may have introduced language bias. Access to full-text articles was sometimes restricted and may have influenced study selection.
The review protocol of this study was not registered in advance, which may affect transparency.

This review was conducted within the context of a master's thesis and was therefore subject to practical constraints, including limited time and experience with systematic review methodology and the appraisal of diagnostic studies. The rapid evolution of digital health technologies and AI also means that recent developments may not yet be reflected in the available literature. In addition, no direct hands-on experience with digital stethoscopes or consultation with domain experts was incorporated, which may have limited the practical interpretation of the findings.

The overall methodological quality of the included studies was moderate to low. Many diagnostic and prediction model studies were associated with a high risk of bias, particularly in patient selection, reference standards, and analytical methods. This reduces confidence in the findings.

Another important limitation is the lack of studies conducted in primary care settings. Most studies were performed in hospital or specialized environments, which limits external validity and applicability to general practice. These populations often differ from those seen in primary care, both in disease severity and prevalence. As a result, the findings may not be directly transferable to general practice, which limits their relevance for the research question.

In addition, many studies focused on technical performance rather than clinically relevant outcomes, further limiting their direct translation into clinical practice.

Finally, in relation to the research question, none of the studies specifically addressed the diagnosis of pneumonia in primary care. As a result, the conclusions of this review are based on broader evidence on respiratory sound analysis and may not fully reflect the diagnostic value of digital stethoscopes for pneumonia in real-world primary care settings.

Despite these limitations, this review provides a comprehensive overview of the current evidence and highlights important gaps for future research.

8.6 Future research

There is still considerable room for further research, particularly in primary care settings. Future studies should move beyond technical validation and focus on clinical impact.
Prospective studies in primary care settings are needed to evaluate whether the use of digital stethoscopes combined with AI improves diagnostic strategies for conditions such as pneumonia. In a later stage, diagnostic intervention studies, including randomized designs, may help determine whether these improvements translate into better patient outcomes, including prognosis and clinical management, for example through more appropriate antibiotic prescribing.

Further research is needed to evaluate cost-effectiveness and to ensure external validation of AI models in diverse, real-world populations.

9 CONCLUSION

Digital stethoscopes combined with AI represent a promising innovation in respiratory diagnostics. However, current evidence does not yet support their routine implementation in primary care. Bridging the gap between technical performance and real-world clinical utility remains the key challenge. Further research in primary care settings is needed to determine their true clinical value and potential role in practice.

There is still a lack of evidence specifically addressing the diagnosis of pneumonia, which was the initial focus of this review. Further research in primary care settings is needed to determine their true clinical value and potential role in practice.

10 References

1. *Acute respiratory infections in the EU/EEA: epidemiological update and current public health recommendations*. [Epidemiological Web Article] 2023 15/12/2023 [cited 2024 12/05/2024]; Available from: https://www.ecdc.europa.eu/en/news-events/acute-respiratory-infections-eueea-epidemiological-update-and-current-public-health.
2. Zhang, M., M. Li, L. Guo, and J. Liu, *A Low-Cost AI-Empowered Stethoscope and a Lightweight Model for Detecting Cardiac and Respiratory Diseases from Lung and Heart Auscultation Sounds.* Sensors (Basel), 2023. **23**(5).
3. Kapetanidis, P., et al., *Respiratory Diseases Diagnosis Using Audio Analysis and Artificial Intelligence: A Systematic Review.* Sensors (Basel), 2024. **24**(4).
4. Kapetanidis P, Smith J, Brown L, et al. The role of digital health in respiratory diseases management. Front Med (Lausanne). 2025;12:1361667. doi:10.3389/fmed.2025.1361667
5. Landry, J. *The 6+ Best Electronic and Digital Stethoscopes (2024)*. 2024 2/01/2024 [cited 2024; Available from: https://www.respiratorytherapyzone.com/best-electronic-stethoscopes/.
6. Unknown. *The top 5 Stethoscope Innovations in 2024. Explore Now*. 2024 [cited 2024; Available from: https://stethoscope.eu/blog/littmann-stethoscopes-en/the-top-5-stethoscope-innovations-in-2024-explore-now/#:~:text=Digital%20Stethoscopes%3A%20A%20New%20Era&text=They%20offer%20superior%20sound%20quality,Stethoscope%20are%20at%20the%20forefront.
7. *Digital Vs. Traditional Stethoscopes: A Comparative Analysis*. [cited 2024 12/05/2024]; Available from: https://www.stethoscope.com/digital-vs-traditional-stethoscopes/#:~:text=General%20Sound%20Issues&text=For%20example%2C%20digital%20stethoscopes%20aren,Primary%20Care%20Pediatrician%2C%20Dr.
8. Swarup, S. and A.N. Makaryus, *Digital stethoscope: technology update.* Med Devices (Auckl), 2018. **11**: p. 29-36.
9. Cals, J. and H. van Weert, *Point-of-care tests in general practice: hope or hype?* Eur J Gen Pract, 2013. **19**(4): p. 251-6.
10. Page, M.J., et al., *PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews.* BMJ, 2021. **372**: p. n160.
11. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.5 (updated August 2024). Cochrane, 2024. Available from www.cochrane.org/handbook.
12. Jeremy Howick, I.C., Paul Glasziou, Trish Greenhalgh, Carl Heneghan, Alessandro Liberati, Ivan Moschetti, Bob Phillips, and Hazel Thornton. *The 2011 Oxford CEBM Levels of Evidence (Introductory Document)*. Available from: https://www.cebm.ox.ac.uk/resources/levels-of-evidence/ocebm-levels-of-evidence.
13. Jeremy Howick, I.C.J.L.L., Paul Glasziou, Trish Greenhalgh, Carl Heneghan, Alessandro Liberati, Ivan Moschetti, Bob Phillips, Hazel Thornton, Olive Goddard and Mary Hodgkinson *The Oxford Levels of Evidence 2*.
14. Shea BJ, R.B., Wells G, Thuku M, Hamel C, Moran J, Moher D, Tugwell P, Welch V, Kristjansson E, Henry DA, *AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both.* BMJ, 2017. **358**.
15. Whiting, P.F., et al., *QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies.* Ann Intern Med, 2011. **155**(8): p. 529-36.
16. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011 Oct 18;155(8):529-36
17. *PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies.* Annals of Internal Medicine, 2019. **170**(1): p. 51-58.
18. *Critical Appraisal Skills Programme*. CASP (Systematic Reviews with Meta-Analysis of Observational Studies) 2023.
19. Pearson A, J.Z., McArthur A, Florescu S, Cooper A,, Yan H, et al *Systematic reviews of textual evidence: narrative, expert opinion, or policy*. 2024.
20. Riley, D.S., et al., *CARE guidelines for case reports: explanation and elaboration document.* J Clin Epidemiol, 2017. **89**: p. 218-235.
21. Baethge, C., S. Goldbeck-Wood, and S. Mertens, *SANRA-a scale for the quality assessment of narrative review articles.* Res Integr Peer Rev, 2019. **4**: p. 5.
22. Abreu, V., A. Oliveira, J. Alberto Duarte, and A. Marques, *Computerized respiratory sounds in paediatrics: A systematic review.* Respiratory Medicine: X, 2021. **3**: p. 100027.
23. Salahuddin, A., *Community use of digital auscultation to improve diagnosis of childhood pneumonia in Sylhet, Bangladesh*, in *Edinburgh Medical School*. 2024, Edinburgh Medical School: Edinburgh Reseach Archive.
24. Park, J.S., et al., *Artificial Intelligence Models for Pediatric Lung Sound Analysis: Systematic Review and Meta-Analysis.* J Med Internet Res, 2025. **27**: p. e66491.
25. Park, J.S., et al., *A machine learning approach to the development and prospective evaluation of a pediatric lung sound classification model.* Sci Rep, 2023. **13**(1): p. 1289.
26. Hoekstra, N.E., et al., *Performance of an artificial intelligence algorithm for interpreting lung sounds from children hospitalised with pneumonia in Malawi.* J Glob Health, 2025. **15**: p. 04264.
27. Gupta, R., et al., *DeepRespNet: A deep neural network for classification of respiratory sounds.* Biomedical Signal Processing and Control, 2024. **93**: p. 106191.

28. Kevat, A., A. Kalirajah, and R. Roseby, *Artificial intelligence accuracy in detecting pathological breath sounds in children using digital stethoscopes.* Respir Res, 2020. **21**(1): p. 253.
29. Ma, Y., et al. *LungBRN: A Smart Digital Stethoscope for Detecting Respiratory Disease Using bi-ResNet Deep Learning Algorithm*. in *IEEE Biomedical Circuits and Systems Conference (BioCAS)*. 2019. Nara, JAPAN.
30. Mahmood, A.F., A.M. Alkababji, and A. Daood, *Resilient embedded system for classification respiratory diseases in a real time.* Biomedical Signal Processing and Control, 2024. **90**: p. 105876.
31. Rady, R.M., et al. *Respiratory Wheeze Sound Analysis Using Digital Signal Processing Techniques*. in *7th International Conference on Computational Intelligence, Communication Systems and Networks*. 2015. Riga, LATVIA.
32. Uppin, R., S. Ambesange, S. Aralikatti, and M.G. Ve. *Respiratory Sound Abnormality Classification using Multipath Deep Learning Method*. in *Proceedings of the 3rd International Conference on Inventive Research in Computing Applications, ICIRCA 2021*. 2021.
33. Gupta, S., M. Agrawal, and D. Deepak, *Correlating spirometry findings with auscultation sounds for diagnosis of respiratory diseases.* Biomedical Signal Processing and Control, 2024. **87**: p. 105347.
34. Divakaran, S., et al., *AI-Integrated Digital Auscultation System for Early Detection and Monitoring of Chronic Obstructive Pulmonary Disease in Resource-Limited Settings.* Biomedical and Pharmacology Journal, 2025. **18**(3): p. 2061-2069.
35. Topaloglu, I., et al., *Machine Learning-Driven Lung Sound Analysis: Novel Methodology for Asthma Diagnosis.* Advances in Respiratory Medicine, 2025. **93**(5): p. 32.
36. Chatterjee, S., J. Roychowdhury, and A. Dey, *D-Cov19Net: A DNN based COVID-19 detection system using lung sound.* Journal of Computational Science, 2023. **66**: p. 101926.
37. Karaarslan, O., K.D. Belcastro, and O. Ergen, *Respiratory sound-base disease classification and characterization with deep/machine learning techniques.* Biomedical Signal Processing and Control, 2024. **87**: p. 105570.
38. Naqvi, S.Z.H. and M.A. Choudhry, *An Automated System for Classification of Chronic Obstructive Pulmonary Disease and Pneumonia Patients Using Lung Sound Analysis.* Sensors (Basel), 2020. **20**(22).
39. Nct, *Enhancing Respiratory Auscultation Skills Among Medical Students: digital Stethoscopes in Medical Education.* https://clinicaltrials.gov/ct2/show/NCT06359158, 2024.
40. Zhang, P., A. Swaminathan, and A.A. Uddin, *Pulmonary disease detection and classification in patient respiratory audio files using long short-term memory neural networks.* Frontiers in Medicine, 2023. **10**.
41. Zhang, Y., et al., *Research on lung sound classification model based on dual-channel CNN-LSTM algorithm.* Biomedical Signal Processing and Control, 2024. **94**: p. 106257.
42. Andrès, E., et al., *Development and experimentation of a new digital communicating and intelligent stethoscope.* European Research in Telemedicine / La Recherche Européenne en Télémédecine, 2016. **5**(4): p. 145-155.
43. Gottlieb, E.R., J.M. Aliotta, and D. Tammaro, *Comparison of analogue and electronic stethoscopes for pulmonary auscultation by internal medicine residents.* Postgrad Med J, 2018. **94**(1118): p. 700-703.
44. Noda, A., et al., *Evidence of the Sequential Changes of Lung Sounds in COVID-19 Pneumonia Using a Novel Wireless Stethoscope with the Telemedicine System.* Intern Med, 2020. **59**(24): p. 3213-3216.
45. Park, D.E., et al., *Digitally recorded and remotely classified lung auscultation compared with conventional stethoscope classifications among children aged 1-59 months enrolled in the Pneumonia Etiology Research for Child Health (PERCH) case-control study.* Bmj Open Respiratory Research, 2022. **9**(1).
46. Wang, B., et al., *Characteristics of Pulmonary Auscultation in Patients with 2019 Novel Coronavirus in China.* Respiration, 2020. **99**(9): p. 755-763.
47. McCollum, E.D., et al., *Digital auscultation in PERCH: Associations with chest radiography and pneumonia mortality in children.* Pediatr Pulmonol, 2020. **55**(11): p. 3197-3208.
48. Ahmed, S., et al., *Digital auscultation as a diagnostic aid to detect childhood pneumonia: A systematic review.* J Glob Health, 2022. **12**: p. 04033.
49. Hassan Naqvi, S.Z. and M.A. Choudhry, *Embedded system design for classification of COPD and pneumonia patients by lung sound analysis.* Biomed Tech (Berl), 2022. **67**(3): p. 201-218.
50. Ahmed, S., et al., *Evaluation of machine learning to detect adventitious lung sounds using digital auscultation to aid childhood pneumonia diagnosis.* European Respiratory Journal. **60**(suppl 66): p. 272.
51. Park, D.E., et al., *Digitally recorded and remotely classified lung auscultation compared with conventional stethoscope classifications among children aged 1-59 months enrolled in the Pneumonia Etiology Research for Child Health (PERCH) case-control study.* BMJ Open Respir Res, 2022. **9**(1).
52. Ahmed, S., et al., *Digital auscultation as a novel childhood pneumonia diagnostic tool for community clinics in Sylhet, Bangladesh: protocol for a cross-sectional study.* Bmj Open, 2022. **12**(2).
53. Sharma, S., S. Pandey, and D.R. Shah, *Enhancing Medical Diagnosis with AI: A Focus on Respiratory Disease Detection.* Indian Journal of Community Medicine, 2023. **48**(5): p. 709-714.
54. Zhang, P.Z., A. Swaminathan, and A.A. Uddin, *Pulmonary disease detection and classification in patient respiratory audio files using long short-term memory neural networks.* Frontiers in Medicine, 2023. **10**.
55. Kim, Y., et al., *The coming era of a new auscultation system for analyzing respiratory sounds.* BMC Pulmonary Medicine, 2022. **22**(1): p. 119.

11 Appendix

Table of Contents

11 Appendix 26
11.1 Test Accuracy 27
11.2 Levels of Evidence 27
11.3 Prisma Checklists 27
11.4 Overview of Results ..... **Fout! Bladwijzer niet gedefinieerd.**
11.5 Quality Control 71
11.5.1 AMSTAR quality assessment 71
11.5.2 QUADAS assessment 74
11.5.3 PROBAST assessment of included prediction model studies 79
11.5.4 Other critical appraisal tools used 82

11.1 Test Accuracy

1	**Test accuracy: Is the test valid, reliable and robust?**
2	Quality control: Can feasible long-term quality control be assured?
3	Time to decide: Is the test result available quickly enough?
4	Usability: Is the test easy to use?
5	Frequency of use: Will the test be used frequently enough?
6	Indication for testing: Is the indication for using the test sufficiently clear and defined?
7	Test interpretation: Is enough information available for a correct interpretation of the various potential test results?
8	Immediate effect on GP's decision: Will the test result have direct consequences for the GP's decision-making?
9	Effect on patient outcomes: Does the test produce results that under normal practice conditions will benefit patients with the relevant indication?
10	Cost-effectiveness: Has the use of the test been proved to be cost-effective?

Table S 1: Criteria of Cals and Van Weert for the introduction of new point of care tests into general practice

11.2 Levels of Evidence

Question	Step 1 (Level 1*)	Step 2 (Level 2*)	Step 3 (Level 3*)	Step 4 (Level 4*)	Step 5 (Level 5)
How common is the problem?	Local and current random sample surveys (or censuses)	Systematic review of surveys that allow matching to local circumstances**	Local non-random sample**	Case-series**	n/a
Is this diagnostic or monitoring test accurate? (Diagnosis)	Systematic review of cross sectional studies with consistently applied reference standard and blinding	Individual cross sectional studies with consistently applied reference standard and blinding	Non-consecutive studies, or studies without consistently applied reference standards**	Case-control studies, or "poor or non-independent reference standard**	Mechanism-based reasoning
What will happen if we do not add a therapy? (Prognosis)	Systematic review of inception cohort studies	Inception cohort studies	Cohort study or control arm of randomized trial*	Case-series or case-control studies, or poor quality prognostic cohort study**	n/a
Does this intervention help? (Treatment Benefits)	Systematic review of randomized trials or *n*-of-1 trials	Randomized trial or observational study with dramatic effect	Non-randomized controlled cohort/follow-up study**	Case-series, case-control studies, or historically controlled studies**	Mechanism-based reasoning
What are the COMMON harms? (Treatment Harms)	Systematic review of randomized trials, systematic review of nested case-control studies, *n*-of-1 trial with the patient you are raising the question about, or observational study with dramatic effect	Individual randomized trial or (exceptionally) observational study with dramatic effect	Non-randomized controlled cohort/follow-up study (post-marketing surveillance) provided there are sufficient numbers to rule out a common harm. (For long-term harms the duration of follow-up must be sufficient.)**	Case-series, case-control, or historically controlled studies**	Mechanism-based reasoning
What are the RARE harms? (Treatment Harms)	Systematic review of randomized trials or *n*-of-1 trial	Randomized trial or (exceptionally) observational study with dramatic effect			
Is this (early detection) test worthwhile? (Screening)	Systematic review of randomized trials	Randomized trial	Non -randomized controlled cohort/follow-up study**	Case-series, case-control, or historically controlled studies**	Mechanism-based reasoning

*** Level may be graded down on the basis of study quality, imprecision, indirectness (study PICO does not match questions PICO), because of inconsistency between studies, or because the absolute effect size is very small; Level may be graded up if there is a large or very large effect size.**

**** As always, a systematic review is generally better than an individual study.**

Figure S 1: Oxford Centre for Evidence-Based Medicine 2011 Levels of Evidence

11.3 Prisma Checklists

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	Title page
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts Checklist (Figure S3).	P30
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	p1-3
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	P2-3
METHODS			
Eligibility	5	Specify the inclusion and exclusion criteria for the review	P3-5

criteria		and how studies were grouped for the syntheses.	
Information sources	6	Specify all databases, registers, websites, organizations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	P5
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	P3-5
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	P6
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	P3-6
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	P3-5
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	P3-5
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	P6
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	P12
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	P5-9
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	P5-9
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	P31-69
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	P5-9
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	P5-9
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	/

Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	/
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	/
Results			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	P6-7
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	P10
Study characteristics	17	Cite each included study and present its characteristics.	P10-17
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	P11-16 & p70-82
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	P31-69
Results of syntheses	20a	For each synthesis, briefly summarize the characteristics and risk of bias among contributing studies.	P11-16 & & p70-82
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	Not applicable
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	/
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	/
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	/
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	/
Discussion			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	P19-22
	23b	Discuss any limitations of the evidence included in the review.	P21
	23c	Discuss any limitations of the review processes used.	P21
	23d	Discuss implications of the results for practice, policy, and future research.	P22
Other information			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	/
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	/
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	/
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the	/

		review.	
Competing interests	26	Declare any competing interests of review authors.	/
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	P24-25, p31-69

Table S 2: PRISMA 2020 reporting checklist for systematic reviews and meta-analysis

Section and Topic	**Item #**	**Checklist item**	**Reported (Yes/no)**
TITLE			
Title	1	Identify the report as a systematic review.	Yes
BACKGROUND			
Objectives	2	Provide an explicit statement of the main objective(s) or question(s) the review addresses.	Yes
METHODS			
Eligibility criteria	3	Specify the inclusion and exclusion criteria for the review.	No
Information sources	4	Specify the information sources (e.g. databases, registers) used to identify studies and the date when each was last searched.	Yes
Risk of bias	5	Specify the methods used to assess risk of bias in the included studies.	Yes
Synthesis of results	6	Specify the methods used to present and synthesize results.	Yes
RESULTS			
Included studies	7	Give the total number of included studies and participants and summarize relevant characteristics of studies.	Yes
Synthesis of results	8	Present results for main outcomes, preferably indicating the number of studies included and participants for each. If meta-analysis was done, report the summary estimate and confidence/credible interval. If comparing groups, indicate the direction of the effect (i.e. which group is favoured).	Yes
DISCUSSION			
Limitations of evidence	9	Provide a brief summary of the limitations of the evidence included in the review (e.g. study risk of bias, inconsistency and imprecision).	No
Interpretation	10	Provide a general interpretation of the results and important implications.	Yes
OTHER			
Funding	11	Specify the primary source of funding for the review.	Not applicable
Registration	12	Provide the register name and registration number.	Not applicable

Table S 3: PRISMA 2020 abstract checklist

11.4 Overview Of Results

Study	Aim	Included studies	Quality assessment	Outcomes	Findings
Abreu V et al., 2021	Systematically review different acoustic respiratory sound properties in healthy children and in children with different respiratory disorders	N=28 Databases: MEDLINE, EMBASE, AMED, CINHAL Search terms based on combination of: Paediatric OR paediatric OR new-born OR infant OR toddle OR children OR teenager OR adolescent OR baby (for Population) AND healthy OR lower respiratory tract infection OR bronchiolitis (for Comparisons) AND computerized auscultation OR computerized auscultation OR electronic auscultation OR automatic auscultation AND sound AND lung OR pulmonary OR respiratory OR breath* OR wheeze* OR crackle* OR adventitious (for Outcomes) Inclusion: validation, exploring or using computerized respiratory sounds in a paediatric population and written in English, French, Spanish, or Portuguese Exclusion: respiratory sounds assessed with conventional auscultation only or if conducted on animals Additional exclusion: book chapters, abstracts in conferences or meetings, single case studies, letters to the editor, commentaries to articles, unpublished work, study protocols, or reviews	Observational (n=7): • Good 26,7% • Fair 46,6% • Poor 26,7% Before-after (n=12): • Good 50% • Fair 50% • RCT (n=1): • Good 100%	Respiratory sound acquisition Respiratory sound analysis Normal respiratory sound characteristics Adventitious respiratory sounds characteristics	Respiratory sound acquisition: • Flow rate & type of breathing not reported in 21,5% of studies • Microphones were most used but digital stethoscopes are more available and may be first choice for studies conducted in clinical settings with larger samples • Trachea, left posterior and right posterior inferior chest are minimum essential recording locations for lung auscultation. Only 30% of included studies used trachea and 12-23% used right/left posterior chest. 46% of studies used one single location • Characteristics of respiratory sounds are affected by dimension of airways but also by airflow, sex, and chest location • Compliance to sound acquisition guidelines is recommended in future studies to allow interpretation and generalization Respiratory sound analysis: • Sound bandwidth ranged from 40Hz-2,7KHz • Recommended minimum limit is 60-100Hz for wheeze detection • Recommended maximum limit for wheeze detection is 2KHz but may miss detection of other ARS Normal respiratory sound characteristics:

					• Spectral slopes from 600-1200Hz were higher in healthy than asthmatic participants, values increased after treatments – useful outcome measure for bronchoconstriction evaluation • Frequency of normal respiratory sounds F50 increased in asthmatic participants after bronchial provocation tests but not in all studies. Increase in pitch and decrease in intensity is a finding of airway narrowing • Breath sound intensity decreases and F50 increases with maximal bronchoconstriction Adventitious respiratory sounds: • Wheeze is most reported • Wheeze occupation rate is most promising parameter as outcome measure • Duration of wheezing is 24ms in bronchiolitis to 400ms during inspiration in asthma. • American Thoracic Society: 250ms is minimum duration for wheeze, in this review shorter lengths were found in children and teenagers. Respiratory rate is higher in younger individuals; duration of sounds will likely also be shorter. This highlights need for reference values for paediatrics. • Paediatrics are more likely to develop atelectasis than adults, identified by crackles decreasing

					after physiotherapy. Crackles might be related to atelectasis
Park J. et al., 2025 [24]	Assess the performance of machine learning (ML) models in pediatric lung sound analysis, evaluating methodologies, model performance and database characteristics while identifying limitations and future directions for clinical implementation	N = 41 Databases: PubMed, Embase, Web of Science, OVID, IEEE Xplore Search terms based on combination of: “infant,” “child, preschool,” “child,” “adolescent,” “pediatrics,” “respiratory sounds,” “auscultation,” “adventitious sound,” “wheeze,” “crackle,” “rale,” “stridor,” “rhonchus,” “machine learning,” “neural network,” and “artificial intelligence.” Inclusion criteria: Studies developing ML models for pediatric lung sound classification with a defined database, physician-labeled reference standard, and reported performance metrics. Exclusion criteria: Studies focusing on adults, cardiac auscultation, validation of existing models, or lacking performance metrics	QUADAS-2 tool was used. Risk of bias: • Patient selection: high (n=21), low (n=20) • Index test: high (n=16), low (n=21), unknown (n=4) • Reference standard: high (n=10), low (n=24), unknown (n=7) • Flow and timing: high (n=8), low (n=32), unknown (n=1) Applicability: • Patient selection: high (n=0), low (n=41) • Index test: high (n=5), low (n=35), unknown (n=1) Reference standard: high (n=11), low (n=24), unknown (n=6)	Pooled sensitivity and specificity of wheeze detection models Pooled sensitivity and specificity of abnormal lung sound detection models Databases of lung sounds used Extraction methods used ML models used	Pooled sensitivity and specificity of wheeze detection models: • Sensitivity: 0,902 (95% CI 0,726-0,970) • Specificity: 0,955 (95% CI 0,762-0,993) • Moderate between-study heterogeneity Pooled sensitivity and specificity of abnormal lung sound detection models: • Sensitivity: 0,907 (95% CI 0,816-0,956) • Specificity: 0,877 (95% CI 0,813-0,921) • Low heterogeneity Databases of lung sounds used: • Total of 29 databases • Public (n=2) • Private (n=27) • Most used: SPRSound (n=11) Extraction methods used (n=9): • Short-time Fourier transform (STFT) or spectrogram (n=10) • Fourier transform-based spectral feature (n=1) • (log) Mel-spectrogram (n=9) • Mel-frequency cepstral coefficients (MFCC) (n=11) • Statistics (n=7) • Continuous wavelet transformation (CWT) (n=2) • Discrete wavelet transformation (n=2) • Cochleogram (n=1)

					• Time-varying linear predictive coding (TVLPC) (n=1) ML models used (n=10): • Artificial neural network (n=11) • Convolutional neural network (CNN) (n=12) • Residual Network (ResNet) (n=6) • Support vector machine (SVM) (n=9) • CNN + recurrent neural networks (RNNs) (n=2) • Hidden Markov model (n=2) • K-nearest neighbors (KNN) (n=1) • Two-level ensemble model (n=1) • Naïve Bayes (n=1) Transformer (n=1)

Ahmed S. et al., 2022 [50]	Part of a larger PhD which consists of four parts, one part being a systematic review.
	Aim of systematic review: to synthesize the evidence on the performance of digital auscultation compared to human listeners in identifying adventitious lung sounds among children with pneumonia
		N = 10
		Databases: MEDLINE, Embase, CINAHL Plus, Web of Science, Global Health, IEEExplore database, Scopus, ClinicalTrial.gov
		
		Search terms based on combination of: “pneumonia”, “lower respiratory infection”, “auscultation”, “respiratory sound”, “lung sound”, “digital”, “electronic”, “automatic”, “computerized”, “crackles”, “wheeze”, “child”
		
		Inclusion criteria:
		• Participants: Children under five years of age
			QUADAS-2 tool was used.
			
			Risk of bias:
			• Patient selection: high risk 40%, unclear risk 60%
			• Index test: low risk 100%
			• Reference standard: unclear risk 100%
			• Flow and timing: low risk 70%, unclear risk 30%
			
			Applicability concerns:
				Stethoscopes used
				
				Lung sounds classification used
				
				Algorithm used to classify lung sounds
				
				Diagnostic performance of used algorithms/models
					Stethoscopes used:
					• ICBHI data set:
					  ○ 3M Littmann Classic II SE stethoscope
					  ○ 3M Littmann 3200 electronic stethoscope
					  ○ WelchAllyn Meditron Master Elite electronic stethoscope
					  ○ AKG C417 L Microphone
					• Clinicloud DS
					• ThinkLabs ds32a digital stethoscope with a microphone (Sony-ICD-UX71-81) affixed to the

		• Index test: Digital auscultation/computerized analysis of lung sounds • Reference standards: Physicians' diagnosis of adventitious lung sounds (crackles and/or wheeze) by conventional auscultation/acoustic analysis of recorded lung sounds • Target condition: Pneumonia • Outcome: Reported diagnostic accuracy measures of the digital/computerized analysis of adventitious lung sounds • Types of studies: Observational and experimental studies Exclusion criteria: • Digital auscultation or computerized analysis of lung sounds not used in study • Reference standard not a human classification of lung sounds Reports are reviews, conference proceedings, abstracts, case reports, editorials, commentaries	• Patient selection: high risk 10%, low risk 10%, unclear risk 80% • Index test: high risk 60%, low risk 10%, unclear risk 30% Reference standard: high risk 10%, low risk 80%, unclear risk 10%		stethoscope to record ambient noises Lung sounds classification used: • Normal and adventitious/pathological sounds (n=3) • Four-class classification (normal, crackle, wheeze, both wheeze and crackle) (n=2) • Three-class classification (normal, crackle, wheeze) (n=1) • Wheeze and crackles (n=1) • Wheeze, rattle, stridor, normal lung sounds (n=2) • Three-class chronic classification (healthy, non-chronic diseases, chronic diseases) and six-class pathological classification (healthy, bronchiectasis, bronchiolitis, COPD, pneumonia, URTI) (n=1) Algorithm used to classify lung sounds: • Feature extraction methods: ○ Short-time Fourier Transform (STFT) (n=3) ○ Melspectrogram (n=2) ○ Mel Frequency Cepstral Coefficients (MFCC) (n=2) ○ Enhanced distinctive feature

					extraction methods: Wevelt Transform (WT) coefficients, right spectro-temporal feature space, Optimized S-Transform (OST), and hybrid scalogram using empirical mode decomposition (EMD) and continuous wavelet transform (CWT) (n=4) • Sound classification models/algorithms: ○ Dynamic time warping (DTW algorithm) ○ Support Vector Machine (SVM) ○ K-nearest Neighbor (k-NN) based classifier ○ Audio spectrographic analysis ○ Deep Residual Networks (ResNets) ○ Recurrent Neural Networks (RNN) models with Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) ○ Hybrid CNN-RNN model

					○ Convolutional Neural Network (CNN) models ○ Lightweight Convolutional Network (CNN) model Diagnostic performance of used algorithms/models: • Studies that included participants with different target conditions and age groups: overall accuracy on adventitious lung sounds classification or by pathological conditions classification. Accuracies: varied from 66,3%-100% • Accuracy of computerized sound classification system in children aged 1-59 months in pneumonia (n=1): ○ Accuracy 86,7% ○ Sensitivity 86,8% Specificity 86,6%

Table S 4: summary of results of Systematic Reviews (n=3)

Study	Population	Material	Method	Comparison	Outcome	Findings
Andrès E et al., 2016[42]	N=160 • Healthy (n=55) • Heart disorders (n=80) • Lung disorders (n=27) Mean age 45y	• eStetho (Bluetooth electronic stethoscope) connected to dedicated IP phone • VOIP connection through PABX to server for storage and analysis of the auscultation sounds	Diagnostic capacity of eStetho in auscultation of lung and heart is evaluated by having healthcare professionals making a diagnosis using only the eStetho, this is then compared to other diagnostic tests Statistical analysis by	Acoustic auscultation by 3M™ Littmann® Classic III™ Stethoscope/3M United States Final diagnosis based on • Advice of doctors in charge of patient • Analysis of medical files • Results of echocardiography	Concordance between medical file and eStetho system Correct diagnosis of COPD, pneumonia, asthma or healthy by eStetho	Concordance with K value of 0,89 eStetho allowed correct diagnosis in 10% more than the acoustic stethoscope

		• Multicast of auscultation sounds to variable number of devices	Chi^2, Student and Cochrane Q test. P<0,05 is considered significant statistically	• Results of lung CT scan		
Gottlieb E. et al., 2018 [43]	N = 61 • Internal medicine residents • Completed some of the cases (n=20) Completed all six cases (n=41)	Littmann Cardiology III stethoscope chest piece (3M) Eko Core electronic stethoscope attachment (Eko Devices)	Effect of auscultation with electronic stethoscope on clinical decision making is evaluated. Six fictional cases where developed, coupled with related real-life lung sounds recorded by both an electronic and an analogue auscultation. Internal medicine residents were randomly selected to hear the cases with either analogue or electronic lung sounds. Based on auscultation and history, a clinical decision was made. Numbers of correct answers, time spent on each case and numbers of times the recordings were played were compared between the analogue and the electronic group. Total scores and scores by case were analyzed using generalized linear mixed modelling	Littmann Cardiology III stethoscope chest piece (3M) Stethoscope tubing with a lapel microphone (iGear Tech)	Comparison of the case outcomes of the electronic auscultation group with the acoustic auscultation group. Comparison of time spent listening to the recordings between the groups. Comparison of amount of time listened to the recordings between the groups.	No significant difference in overall scores between the electronic group (3.20±0,10 out of 6) and the acoustic group (3.14±0,10 out of 6). P=0,74. No significant differences in time spent on the cases (p=0.67). No significant differences in times the recordings were played (p=0.85).

			(GLMM) with sandwich estimation assuming a binomial distribution (0-6) and (0/1) respectively. Time spent on each case and numbers of times listened to the recordings were compared between the groups using GLMM assuming a negative binomial distribution. Poisson modelling was used. P<0,05 is considered significant statistically			
Noda A. et al., 2020 [44]	N=1 Case Report Woman, 60y/o 2-week history of dry cough, transient fever of 38°C. Covid-19 suspected.	Novel wireless stethoscope (manufactured by Digital Global Systems company)	Sequential changes in lung sounds were recorded in seated position, from posterior right lower lung fields with relaxed relaxation. Lung sounds were recorded directly by connecting systems between a wireless receiver and a cellular phone. Sounds can also be transferred and shared by telemedicine using 3G/4G or Wi-Fi to an outside area.	Heart-resolution computed tomography (HRCT)	Demonstration of the lung sounds that accompany the healing of COVID-19 pneumonia, compared to the improvement of clinical symptoms and HRCT findings.	Clearing of the clinical signs was accompanied by change of timing of inspiratory crackles from “mid to late” to “late” inspiratory crackles with relatively high-pitched sounds higher than 500Hz through the phase (day 1 vs. day 3). Pathognomic radiological findings staged the COVID-19 pneumonia in stage 2.
Park D. et al., 2021 [51]	N = 737 • Collected digitally recorded and conventional auscultation	Commercial digital stethoscopes (ThinkLabs ds32a)	Study from the PERCH project. Digitally recorded lung sounds, conventional	Conventional stethoscopes already owned by included providers.	Concordance between conventional and digital auscultation. Comparison between	Moderate agreement with classification of any crackles and/or wheeze vs. neither crackles nor wheeze

	classifications (n=383) • Non-currently collected assessments (n=354) Children with pneumonia in low-income and middle-income countries.		auscultation classifications and clinical measures and samples from children with pneumonia in low-income and middle-income countries were recorded. Physicians remotely classified recordings as crackles, wheeze or uninterpretable. Conventional and digital auscultation concordance was evaluated among 383 pneumonia cases with concurrently collected conventional and digital auscultation classifications using prevalence-adjusted bias-adjusted kappa (PABAK). An expanded set of 737 cases were evaluated for whether associations between auscultation classifications and clinical or etiological findings differed between conventional or digital auscultation using $\chi 2$ tests and logistic regression adjusted for age, sex, and site.		conventional and digital auscultation classifications among cases with distinct clinical and etiological characteristics. Predictors of concordance between conventional and digital auscultation. Comparison between conventional and digital auscultation classifications and associations with disease severity and clinical outcomes. Digital auscultation classifications among controls and associations of classifications with case-control status.	(75.2% overall agreement, PABAK=0.50, 95% CI 0.41-0.59). Concordance was fair for crackles and wheeze when considered independently. Digital auscultation is more likely to result in wheeze-only or neither crackles nor wheeze classification compared with conventional auscultation, proportion with both crackles and wheeze was similar More children had crackles-only on conventional compared with digital auscultation among 28 likely pneumococcal pneumonia cases (60.7% vs 21.4%, p=0.011) Disagreement between auscultations in classifying normal and abnormal classification were similar across sex, WHO-defined pneumonia severity, time delay between conventional and digital auscultation and

						crying on recording. Crackles-only was generally indicative of higher severity, and wheeze with reduced severity on both digital and conventional auscultation. In both conventional and digital auscultation, children with crackles-only had the highest proportion with CRP ≥40 mg/L (33.7% and 32.7% respectively). Compared with neither cackles nor wheeze, any wheeze was associated with lower CRP (<40 mg/L) on conventional auscultation (aOR=0.50, 95% CI 0.27-0.92) and inversely associated with very severe pneumonia status on digital auscultation (aOR=0.67, 95% CI 0.46-0.97). Crackles-only were associated with abnormal chest X-ray findings on both conventional and digital auscultation (aOR=1.53, 95% CI 0.99-2.36; aOR=2.09,

						95% CI 1.19-3.68, respectively). Mortality is highest among children without crackles nor wheeze on conventional auscultation (17.1%) and highest among children with crackles-only on digital auscultation (17.4%). Mortality was lowest among children with any wheeze on both conventional (3.8%) and digital auscultation (6.4%). Crackles-only were associated with mortality compared with any wheeze classifications on digital auscultation only (aOR=2.70, 95% 1.12-6.25). Using all cases and controls, presence of crackles or wheeze to classify pneumonia case status: • Sensitivity 61.4% • Specificity 84.9% • NPV 46.2% • PPV 91.4%.
Wang B. et al., 2020	N=57	Electronic stethoscope	Clinical information	Chest CT scans	Radiological features	Chest CT showed

[46]	Patients in hospital with lab-confirmed COVID-19. Average age 60.6 years.	(EYZ-1A(C), Exagiga Electric) Cellphone (Huawei Technologies)	obtained and reviewed by physicians. Auscultation and breath sounds recorded by two physicians. Two chest physicians evaluated audio files. Statistical analysis performed by SPSS 25.0.	Laboratory parameters (white blood cell count, neutrophil count, lymphocyte count, CRP)	Audiological characteristics Correlations of auscultation with laboratory and chest CT findings	abnormalities in 55 patients: • Bilateral lesions (96.4%) • Ground-glass opacities (69.1%) • Fibrous stripes (21.8%) • Consolidation (16.4%) • Pleural effusions (23.6%) Abnormal breath sounds: • Coarse breath sound (78.9%) • Wheeze (29.8%) • Coarse crackles (57.9%) • Fine crackles (70.2%) • Velcro crackles (33.3%) • Pleural friction rub (1.8%) No statistically significant relation was found between laboratory values and auscultation outcomes. Correlation of auscultation and CT findings: • Fine and coarse crackles detected and ground-glass opacities

						(sensitivity 84.6%, specificity 15.2%) • Fine and coarse crackles and consolidation (sensitivity 88.9%, specificity 15.2%) • Velcro crackles and ground-glass opacities (sensitivity 41%, specificity 81.3%) • Velcro crackles and consolidation (sensitivity 44.4%, specificity 67.4%)
McCollum E. et al., 2020 [47]	N=742 • WHO severe pneumonia (N=496) • WHO very severe pneumonia (N=246) Lung-recordings of 1–59-month-old children hospitalized with pneumonia at eight African and Asian Pneumonia Etiology Research for Child Health sites in six countries	Commercial digital stethoscope (ThinkLabs ds32a) External microphone (Sony ECM-ES30P) Voice recorder (Sony-ICD-UX71)	Recordings were processed and classified into binary categories positive or negative for adventitial lung sounds. Listening and reading panels classified recordings and radiographs. Associations between lung recordings and radiographic pneumonia or death were evaluated by t-test for continuous variables and Pearson $\chi 2$ or Fisher exact tests for categorical variables.	Chest radiographs with WHO-defined radiographic pneumonia	Association between digitally recorded lung sounds and radiographic pneumonia Association between digitally recorded lung sounds and mortality	Association between digitally recorded lung sounds and radiographic pneumonia: - WHO severe pneumonia: • Normal sounds: 38.6% radiographic pneumonia; OR 1.00 • Crackle only: 54.1% radiographic pneumonia; OR 1.87, p-value 0.10; aOR 2.13, p-value 0.07 • Wheeze only: 14.9% radiographic pneumonia; OR 0.28, p-value

						<0.01; aOR 0.35, p-value 0.01 • Any wheeze: radiographic pneumonia 20.4%; OR 0.41, p-value <0.01; aOR 0.63, p-value 0.13 - WHO very severe pneumonia: • Normal lung sounds: 33.3% radiographic pneumonia; OR 1.00 • Crackle only: 63.6% radiographic pneumonia; OR 3.50, p-value 0.01; aOR 2.75, p-value 0.08 • Wheeze only: 63.6% radiographic pneumonia; OR 1.47, p-value 0.40; aOR 1.45, p-value 0.46 • Any wheeze: 52.6% radiographic pneumonia; OR 2.22, p-value 0.02; aOR 2.08, p-value 0.05 Association between digitally recorded lung sounds and mortality:

						- WHO severe pneumonia: • Normal lung sounds: 9.6% mortality, OR 1.00 • Crackle only: 7.5% mortality; OR 0.76, p-value 0.68, aOR 1.79, p-value 0.38 • Any crackle: 4.5% mortality; OR 0.45, p-value 0.08; aOR 1.19, p-value 0.72 • Wheeze only: 4.0% mortality; OR 0.39, p-value 0.10; aOR 1.01, p-value 0.98 • Any wheeze: 3.8% mortality; OR 0.37, p-value 0.02; aOR 1.02, p-value 0.97 - WHO very severe pneumonia: • Normal lung sounds: 21.8% mortality; OR 1.00 • Crackle only: 32.0% mortality; OR 1.69, p-value 0.27; aOR 1.51, p-value 0.46 • Any crackle: 21.9% mortality; OR 1.00, p-value 0.99; aOR 1.05, p-value 0.91 • Wheeze only: 17.1% mortality; OR 0.74, p-value

						0.52; aOR 0.93, p-value 0.88 • Any wheeze: 16.3% mortality; OR 0.70, p-value 0.33; aOR 0.75, p-value 0.48

Table S 5: Summary of results from articles describing the diagnostic comparison of digital stethoscopes (n=6)

Study	Population	Data collection	Data augmentation	Classification method	Classification	Comparison	Outcome	Findings
Gupta S et al.,2024[33]	N=137 people • 87 healthy • 33 obstructive lung disease • 17 restrictive lung disease Age: 20-70	Auscultation with Littmann 3200 digital stethoscope, 7 locations during 15s Spirometry with NDD Easy on PC Spirometer	Audio equalization: three parametric IIR filters of different orders, cascaded to obtain desired frequency response Feature extraction: • Gabor filter-bank feature • MFCC • GTCC • Spectral domain-based feature Dimensional reduction using principal component analysis with 94% variance	Four algorithms in 5-fold cross-validation: • Naïve Bayes • KNN • Decision tree • Ensemble classifier	• Healthy • Obstructive • Restrictive	Spirometry test: • Healthy • Obstruction • Restriction	Performance of the algorithms for each of the feature extraction methods: • Precision • Recall • F1 score Performance for correlation with spirometry for each feature set using GBFB: • Accuracy • Precision • Recall • F1 score Overall performance proposed method (KNN with GBFB): • Accuracy	Best outcome using Gabor filter bank feature extraction Precision of GBFB: • Naïve Bayes 100 • KNN 100 • Decision Tree 63,63 • Ensemble Classifier 96 Recall of GBFB: • Naïve Bayes 80 • KNN 96 • Decision Tree 70 • Ensemble Classifier 48 F1 score of GBFB: • Naïve Bayes 88,89 • KNN 97,95 • Decision Tree 66,6 Comparison of different feature sets – accuracy: • GBFB 98,54 • MFCC 96,35 • GTCC 92,70 • Spectral Domain 83,21

							• Sensitivity • Specificity	Comparison of different feature sets – precision: • GBFB 100 • MFCC 100 • GTCC 91,67 • Spectral Domain 84,61 Comparison of different feature sets – recall: • GBFB 96 • MFCC 90 • GTCC 88,7 • Spectral Domain 66 Comparison of different feature sets – F1: • GBFB 97,95 • MFCC 94,73 • GTCC 89,79 • Spectral Domain 74,15 End result of GBFB using KNN: • Accuracy 98,54 • Sensitivity 96 • Specificity 100
Ahmed S. et al., 2023 [52]	N=990 • No pneumonia (N=601) • Pneumonia (N=389) Children aged 2-59 months in community clinics in Bangladesh	Auscultation and recording of lung sounds with prototype Feelix Smartscope device	/	Automated computerized lung sound analysis (CLSA) (N=4) Model C showed highest sensitivity, specificity, and kappa value. This model was chosen to be compared.	Classification • Adventitious sounds (wheeze and/or crackles) No adventitious sounds	Reference panel of pediatricians (N=4)	Evaluation of CLSA to detect adventitious and normal lung sounds compared to listening panel Evaluation	A 95% CI is used for all values. Evaluation of CLSA to detect adventitious and normal lung sounds: • In all children: sens 61.8%; spec 60.7%; PPV 41.8%; NPV

							of CLSA to detect any wheeze compared to listening panel Evaluation of CLSA to detect any crackle compared to listening panel Distribution of lung sounds classification by listening panel and CLSA Normal lung sounds in IMCI-defined pneumonia	77.6%; prev 31.4%; raw percentage agreement 61.0%; Kappa 0.22 • In children with IMCI-defined pneumonia: sens 63.5%; spec 66.8%; PPV 52.6%; NPV 75.9%; prev 36.7%; raw percentage agreement 65.6%; Kappa 0.29 • In children with IMCI-defined pneumonia and 'confident' classification by listening panel: sens 69.8%; spec 66.8%; PPV 45.5%; NPV 84.8%; prev 28.4%; raw percentage agreement 67.7%; Kappa 0.32 Evaluation of CLSA to detect any wheeze: • In all children: sens 55.6%; spec 71.2%; PPV 36.1%; NPV 95%;

								prev 22.6%; Kappa 0.23 • In children with IMCI-defined pneumonia: sens 59.8%; spec 72.1%; PPV 44.0%; NPV 83.0%; prev 26.8%; Kappa 0.29 • In children with IMCI-defined pneumonia and 'confident' classification by listening panel: sens 68.8%, spec 73.2%; PPV 40.7%; NPV 89.7%; prev 21.1%; Kappa 0.34 Evaluation of CLSA to detect any crackle: • Among all children: sens 36.9%; spec 77.2%; PPV 26.8%; NPV 84.4%; prev 18.5%, Kappa 0.12 • In children with IMCI-defined pneumonia: sens 35.6%; spec 82.6%; PPV 35.6%; NPV 82.6%; prev

								21.3%; Kappa 0.18 • In children with IMCI-defined pneumonia and 'confident' classification by listening panel: sens 38.2%; spec 82.7%; PPV 32.8%; NPV 85.8%; prev 18.2%; Kappa 0.20 Distribution of lung sounds classification by listening panel and CLSA: - Any pneumonia • No wheeze and no crackle: listening panel 63.3%; CLSA 55.7%; P-value <0.05 • Wheeze only: listening panel 15.5%; CLSA 23.0%; P-value <0.05 • Crackles only: listening panel 9.9%; CLSA 7.9%; P-value 0.35 • Both wheeze and crackles: listening panel 11.4%; CLSA 13.4%; P-value 0.42

								• Any wheeze: listening panel 26.8%; CLSA 36.4%; P-value <0.01 • Any crackles: listening panel 21.3%; CLSA 21.3%; P-value 1.0 • Adventitious sounds: listening panel 36.7%; CLSA 44.3%; P-value <0.05 • Normal: listening panel 63.3%; CLSA 55.7%; P-value <0.05 - No pneumonia • No wheeze and no crackle: listening panel 72.1%; CLSA 52.3%; P-value <0.01 • Wheeze only: listening panel 11.3%; CLSA 19.7%; P-value <0.01% • Crackles only: listening panel 8.0%; CLSA 13.9%; P-value <0.01 • Both wheeze and crackle: panel 8.6%; CLSA

								14.1%; P-value <0.01 • Any wheeze: listening panel 19.8%; CLSA 33.8%; P-value <0.01 • Any crackles: listening panel 16.6%; CLSA 28.1%; P-value <0.01 • Adventitious sound: listening panel 27.9%; CLSA 47.7%; P-value <0.01 • Normal: listening panel 72.1%; CLSA 52.3%; P-value <0.01 Normal lung sounds in IMCI-defined pneumonia: • Agreed on normal: 42% • Agreed on adventitious: 23% • CLSA normal and listening panel adventitious: 13% • CLSA adventitious and listening panel normal: 21%
Park J. et al., 2023 [25]	Retrospectively collected lung sound clips (N=680)	Digital stethoscope (Thinklabs One Digital Stethoscope;	BayesShrink denoising method Sound clips were	Ensemble model based on support vector machine (SVM) was created	Four classifications: • Normal vs. abnormal	Physician performance tests on prospective	Internal validation (N=680)	Internal validation: • Normal vs. abnormal: accuracy 0.8368;

	• Normal (N=288) • Crackle (N=200) • Wheeze (N=192) Prospectively collected lung sound clips (N=90) – used for prospective validation: • Normal (N=28) • Crackle (N=31) Wheeze (N=31)	Thinklabs Medical LLC) connected to a wired audio recorder (PCM-A10, Sony)	cropped into 6s windows Mel-frequence cepstral coefficients (MFCCs) were used to extract acoustic features. 40 MFCCs extracted by using fast-Fourier-transform window length of 660 samples, hop length of 512 samples and Hann windowing. Mel-spectrogram is used for visualization of the sound clips. Uniform manifold approximation and projection (UMAP) are used for embedding the data.	with MFCCs as input. Training and internal validation by K-fold cross-validation (K=10). Performance was evaluated by using accuracy, precision, recall and F1-score. Prospective validation by using a prospectively collected validation dataset. The same data preprocessing was performed, after K-fold cross-validation model was applied. A nested cross-validation model was applied. External validation was done on ICBHI pediatric data.	• Crackle vs. wheezing • Normal vs. crackle Normal vs. wheezing	validation dataset. Physicians (N=10): • Pediatric specialists (N=5) Non-pediatric specialists (N=5)	Prospective validation (N=90) External validation (N=782) Comparison of accuracies of the model and the physicians for the prospective validation set.	precision 0.8325; recall 0.8340; F1-score 0.8332 • Crackles vs. wheezing: accuracy 0.8367; precision 0.8382; recall 0.8360; F1-score 0.8363 • Normal vs. crackles: accuracy 0.8094; precision 0.8007; recall 0.7938; F1-score 0.7966 • Normal vs. wheeze: accuracy 0.9042; precision 0.9053; recall 0.8863; F1-score 0.8936 Prospective validation: • Normal vs. abnormal: accuracy 0.8222; precision 0/7920; recall 0.8122; F1-score 0.8000 • Crackles vs. wheeze: accuracy 0.6774; precision 0.6804; recall 0.6774; F1-score 0.7966 • Normal vs. crackles: accuracy 0.6780; precision 0.6972; recall

								0.6849; F1-score 0.6746 • Normal vs. wheeze: accuracy 0.8136; precision 0.8298; recall 0.8191; F1-score 0.8127 External validation: • Normal vs. abnormal: accuracy 0.835; precision 0.649; recall 0.717; F1-score 0.793 • Crackles vs. wheeze: accuracy 0.764; precision 0.828; recall 0.707; F1-score 0.764 • Normal vs. crackle: accuracy 0.911; precision 1; recall 0.031; F1-score; 0.871 • Normal vs. wheeze: accuracy 0.915; precision 1; recall 0.187; F1-score 0.888 Comparison of accuracies of model and physicians: • Normal vs. abnormal: model 0.8222; pediatric specialists

								0.8489; other specialties 0.8067; p-value[2] (M vs. P) 0.63; P-value[2] (M vs. O) 0.845 • Crackle vs. wheeze: model 0.6774; pediatric specialists 0.7774; other specialties 0.6613; p-value[2] (M vs. P) 0.128; p-value[2] (M vs. O) 0.922 • Normal vs. wheeze: model 0.6780; pediatric specialists 0.7864; other specialties 0.7390; p-value[2] (M vs. P) 0.102; p-value[2] (M vs. O) 0.423 • Normal vs. crackle: model 0.8134; pediatric specialists 0.8441; other specialties 0.7627; p-value[2] (M vs. P) 0.102; p-value (M vs. O) 0.423 • Normal vs. wheeze: model 0.8136; pediatric specialties 0.8441; other

								specialties 0.7627; p-value[2] (M vs. P) 0.698; p-value[2] (M vs. O) 0.497
Hoekstra et aL, 2025 [26]	N=95 Interpretable recordings from 100 children aged 2-59 months, hospitalized in Malawi with WHO-defined severe pneumonia.	Lung sounds recorded from children by using Zoom H4n Portable Recorder.	Audacity audio editing software segmented the sounds by chest position. Ambient noise filtering was applied.	AI algorithm consisting of a CNN (Convolutional Neural Network) encoder, followed by a recurrent classifier.	Normal vs. abnormal.	Trained physician listening panel: three panelists and two arbitrators. Trained to interpret lung sounds using validated methodology.	Analysis was done by chest position and by patient. Agreement between listening panel and AI was evaluated. Performance of the AI was evaluated.	Agreement between listening panel and AI: - Chest position classification: • Agreement 83.1% • Kappa statistic (95% CI): 0.659 (0.592, 0.725) • Adjusted Kappa statistic (95% CI): 0.662 (0.596, 0.728) - Patient classification: • Agreement: 91.6% • Kappa statistic (95% CI): 0.665 (0.447, 0.884) • Adjusted Kappa statistic (95% CI): 0.832 (0.718, 0.945) Performance of AI in reference to listening panel: - Chest position classification: • Normal/abnormal classification: 52.7% • Sensitivity (95% CI): 80.3% (75.3%, 84.7%) • Specificity (95%

								CI): 87.2% (81.8%, 91.5%) • PPV (95% CI): 90.1% (85.8, 93.4%) • NPV (95% CI): 75.3% (69.3%, 80.7%) • LR+ (95%CI): 6.27 (4.36, 9.01) • LR- (95%CI): 0.23 (0?18, 0.29) • Diagnostic OR (95% CI): 27.7 (16.8, 45.7) - Patient classification: • Normal/abnormal classification: 86.3% • Sensitivity (95% CI): 96.3% (89.4%, 99.2%) • Specificity (95% CI): 66.7% (38.4%, 88.2%) • PPV (95% CI): 93.9% (86.3%, 98.0%) • NPV (95% CI): 76.9% (46.2%, 65.0%) • LR+ (95% CI): 2.89 (1.41, 5.91) • LR- (95% CI): 0.06 (0.02, 0.18) Diagnostic OR (95% CI): 51.3 (11.2, 234)

Table S 6: Summary of results from articles describing the diagnostic qualities of automatic sound classification in comparison to other methods (n=4)

Study	Sound Collection	Data Augmentation	Classification Method	Classification	Training &	Outcome measures	Findings

					Validation		
Abreu V. et al., 2021 [36]	N=920 Kaggle, ERS dataset, SARS-Cov-2 dataset	Time-stretching at multiple levels (audio signal slowed down and sped up by a factor of t = 0,5-1,5), pitch variation (audio signal pitched up and down by a factor of p = -1 to 1)	D-Cov19Net based on tiled Convolution Neural Networks	• Normal • COVID Positive • Others	Training with pre-processed data by backpropagation based on the Adam optimization algorithm on 100 epochs	• Accuracy overall • AUC • Sensitivity	• 97,22% • 0,972 • 98,3%
Gupta R et al., 2024[27]	N = 4095 Recorded in collab with Fortis Hospital in New Delhi India	Signal pre-processing Noise reduction Signal Segmentation Amplitude normalization	• 1D DeepRespNet • 2D DeepRespNet	• Aortic • Bronchial • Crepitation • Normal • Rhonchi • Wheezing	Data separated for the two training systems, each time in a training set and a test set	• Precision • Recall • F1 score • Accuracy	Precision 1D: • Aortic 1,00 • Bronchial 0,75 • Crepitation 0,85 • Normal 0,94 • Ronchi 0,75 • Wheezing 0,85 Recall 1D: • Aortic 1,00 • Bronchial 0,60 • Crepitation 0,65 • Normal 1,00 • Rhonchi 0,60 • Wheezing 0,97 F1 score 1D: • Aortic 1,00 • Bronchial 0,67 • Crepitation 0,73 • Normal 0,97 • Rhonchi 0,67 • Wheezing 0,90 Precision 2D: • Aortic 1,00 • Bronchial 0,88 • Crepitation 0,89 • Normal °,96 • Rhonchi 1,00 • Wheezing 0,97 Recall 2D: • Aortic 1,00

							• Bronchial 0,90 • Crepitation 0,97 • Normal 0,95 • Rhonchi 0,73 • Wheezing 0,98 F1 score 2D: • Aortic 1,00 • Bronchial 0,89 • Crepitation 0,93 • Normal 0,96 • Rhonchi 0,84 • Wheezing 0,98 Accuracy overall: • 1D 85,90% • 2D 95,18%
Karaarslan O.et al., 2024 [37]	N=1692 Respiratory Sound Database (N=761) Coronahack respiratory sound dataset (N=826) Coswara sound dataset (N=105)	Feature extraction from six categories. Preprocessing of audio.	A newly developed Deep Neural Network	Binary classification: • Healthy • Unhealthy Disease characterization: • Healthy • Asthma • COPD Others	80% of audio data was used for training, 20% was used as a test set. Training subset was further split, 80% was used for training and 20% for validation.	Accuracy Precision Recall F1-score ROC-AUC score Cohen's Kappa Coefficient Matthew's Correlation Coefficient Compared against two other classifiers: • Ada Boost classifier Extra Trees classifier	Binary classification - Extra Trees Classifier: • Accuracy 99.71% • Precision 99.85% • Recall 99.56% • F1-score 99.70% • ROC-AUC score 99.72% • Cohen's Kappa Coefficient 0.9942 • Matthew's Correlation Coefficient 0.9942 - Ada Boost Classifier: • Accuracy 99.72% • Precision 99.76%

							• F1-score 99.71% • ROC-AUC score 99.72% • Cohen's Kappa Coefficient 0.9944 • Matthew's Correlation Coefficient 0.9944 - Deep Neural Network: • Accuracy 99.27% • Precision 99.28% • Recall 99.22% • F1-score 99.25% • ROC-AUC score 99.27% • Cohen's Kappa Coefficient 0.9854 • Matthew's Correlation Coefficient 0.9854 Disease characterization: - Extra Trees Classifier: • Accuracy 94.72% • Precision 94.15% • Recall 94.73%

							• F1-score 94.73% • Cohen's Kappa Coefficient 0.8346 • Matthew's Correlation Coefficient 0.8375 - Ada Boost Classifier: • Accuracy 94.05% • Precision 93.72% • Recall 94.05% • F1-score 94.73% • Cohen's Kappa Coefficient 0.8157 • Matthew's Correlation Coefficient 0.8177 - Deep Neural Network: • Accuracy 92.81% • Precision 92.54% • Recall 92.81% • F1-score 92.49% • Cohen's Kappa Coefficient 0.7825 • Matthew's Correlation

							Coefficient 0.7853
Kevat A. et al., 2020 [28]	N=192 Auscultation recordings from children Recorded with two digital stethoscopes (Clinicloud™ DS and Littman™) 3200 Electronic Stethoscope)	Processing with normalization and bandwidth filtering. Combined spectrographic, sound wave and audio playback analysis by a pediatric respiratory physician.	Blinded AI algorithm (neural network) (StethoMe AI)	• Wheezes • Crackles • Neither	Neural network was trained and validated on a set of >10.000 real recordings that did not include the recordings collected in the study	True Positive Rate (TPR) including positive percentage agreement (PPA) and negative percentage agreement (NPA) for each crackle and wheeze, for recordings for each stethoscope separately. Afterwards NPA and PPA were calculated with optimized AI detection thresholds.	TPR for crackles: • Clinicloud: 0.95 (PPA 0.86, NPA 0.99) • Littman: 0.75 (PPA 0.6, NPA 0.96) TPR for wheezes: • Clinicloud: 0.93 (PPA 0.90, NPA 0.97) • Littman: 0.8 (PPA 0.76, NPA 0.95) Optimized AI detection thresholds: - Crackle PPA • Clinicloud: 0.95 • Littman: 0.82 - Crackle NPA • Clinicloud: 0.99 • Littman: 0.96 - Wheeze PPA • Clinicloud: 0.90 • Littman: 0.80 - Wheeze NPA • Clinicloud: 0.97 • Littman: 0.95
Ma Y. et al., 2019 [29]	ICBHI 2017 database N=6898 cycles	Pre-processing for ideal frequency range.	LungBRN model: bi-ResNet (BRN) model, developed by writers of the study.	• Normal • Crackle • Wheeze	Adaptive learning rate optimization	Following values were calculated on both the official ICBHI database	Accuracy: - ICBHI 2017:

	• Normal (N=3642) • Crackles (N=1864) • Wheezes (N=886) Crackle plus wheeze (N=506)	Feature extraction via: • Short-time Fourier transform (STFT) Wavelet analysis		Crackle plus wheeze	algorithm (ADAM) was used. Evaluation was done against two types of 'train and test' dataset evaluation methods: • ICBHI official method Random subset method	and a random subset: • Accuracy • Sensitivity • Specificity • Score (official ICBHI 2017 scoring method) Obtained results were compared to existing state-of-the-art works.	• LungBRN: 52.79% - Random subset: • LungBRN: 67.44% Sensitivity: - ICBHI 2017: • Boosted Decisional Tree: 20.81% • LungBRN: 31.12% - Random subset: • NMRNN: 58.43% • LungBRN: 58.54% Specificity: - ICBHI 2017: • Boosted Decisional Tree: 78.05% • LungBRN: 69.20% - Random subset: • NMRNN: 73.00% • LungBRN: 80.06% Score: - ICBHI 2017: • HMM + GMM: 39.56% • Boosted Decisional Tree: 49.86% • SVM: 49.86%

							• LungBRN: 50.16% - Random subset: • NMRNN: 65.70% • LungBRN: 69.30%
Naqvi S. et al., 2020 [38]	ICBHI database N=703 LS data sets Equipment used: • Welch Allyn Meditron Master Elite Plus Stethoscope Model 5079-400 • 3M Littmann 3200 3M Littmann Classic II SE	Preprocessing with empirical mode decomposition (EMD) and discrete wavelet transform (DWT) techniques. Feature extraction by multiple methods, one of which is the adaptive synthetic sampling method (ADASYN).	Various classifiers were evaluated to observe the performance of the proposed method. The quadratic discriminant (QD) has provided maximum accuracy. The proposed method: • preprocessing with EMD and Wavelet Transformation (WT) • Feature extraction: Time Domain Features (N=19), Spectral Domain Features (N=12), Cepstral Features (N=26), Texture Features (N=59) Classifier QD	• COPD • Pneumonia Normal	5-, 10-, 15-and 20-fold cross-validation is performed to authenticate the system performance. 80% of data was used for training, 20% of data was used for testing.	• Accuracy • True positive rate (TPR) False negative rate (FNR)	Accuracy: 99.7% TPR >99% FNR <1%
Mahmood A. et al., 2024 [30]	ICBHI 2017 database N=6898 cycles • Normal (N=3642) • Crackles (N=1864) • Wheezes (N=886) Crackle plus wheeze (N=506)	Preprocessing for ideal frequency. Feature extraction: • Temporal features: Root Mean Square Energy (RMSE), Temporal centroid (TC), Log attack time, Zero Crossing Rate (ZCR), Amplitude Distribution (AD) Spectral features: Spectral centroid (SC), Mel-Frequence Cepstral Coefficients (MFCC)	New proposed system: Random Forest (layer 1-3) + MobileNet	Main classes: • Healthy • Upper respiratory tract infection (URTI) • COPD • Bronchiectasis • Pneumonia Subclasses: • Normal • Crackles • Wheezes Both crackle and wheeze	80% of data was used for training, 20% was used for testing.	Accuracy Power consumption Testing time	Accuracy for main classification: • Healthy: 99.76% • URTI: 99.64% • COPD: 100% • Bronchiectasis: 99.82% • Pneumonia: 99.86% Accuracy for subclassification: • Normal: 99.5% • Crackles: 98.78%

							• Wheezes: 97.5% • Botch crackle and wheeze: 96.9% Savings on power consumption: 63.8% Savings on testing time: 54.7%
Rady R. et al., 2015 [31]	N=206 • Public repositories from online database (N=63) Private repository collected by an electronic stethoscope (3M Littman Electronic Stethoscope Model 3200) (N=143)	Respiration stage: • Respiration detector • Respiratory rate detector • Respiratory onset detector Wheeze stage: • Wheeze detector • Wheeze quantifier Wheeze qualifier	MATLAB (Matrix Laboratory Software)	Wheeze detection Wheeze classification: • Monophonic Polyphonic	/	Wheeze detection: • Sensitivity • Accuracy • Specificity Differentiation between monophonic wheeze and polyphonic wheeze: • Sensitivity Accuracy	Wheeze detection: • Sensitivity: 90% • Accuracy: 91% • Specificity: 90% Differentiation: • Sensitivity: 91% • Accuracy: 70%
Sharma S. et al., 2023 [53]	Opensource Respiratory Database N=6898 • Crackles (N=1864) • Wheezes (N=886) Both wheezes and crackles (N=506)	Analysis of audio waveforms by MFCC's	Convolutional neural network (CNN): four convolution layers of Conv2D with a dense output layer as the final output layer.	• Bronchiectasis • Bronchiolitis • COPD • Healthy • Pneumonia Upper respiratory tract infections (URTI)	80% of the dataset was used for training, 20% of the dataset was used for testing.	For each class: • Precision • Recall • F1 score Accuracy of the model	Precision: • Bronchiectasis: 1.00 • Bronchiolitis: 0.81 • COPD: 0.94 • Healthy: 0.96 • Pneumonia: 0.66 • URTI: 1.00 Recall: • Bronchiectasis: 0.62

							• Bronchiolitis: 1.00 • COPD: 0.97 • Healthy: 0.77 • Pneumonia: 0.84 • URTI: 0.83 F1-score: • Bronchiectasis: 0.771 • Bronchiolitis: 0.90 • COPD: 0.96 • Healthy: 0.86 • Pneumonia: 0.74 • URTI: 0.90 Accuracy: 0.90
Uppin R. et al., 2021 [32]	Kaggle platforms: The Respiratory Sound Database N=6898 • Crackles (N=1864) • Wheezes (N=886) Crackles and wheezes (N=506)	Feature extraction: • MFCC feature extraction • Mel spectrogram • Short-time Fourier transform	Multi Path CNN Model	Undefined 8 classes: normal vs. abnormal (7 classes)	90% of data was used for training. 10% of data was used for testing. Adaptive learning rate method was used.	Accuracy of proposed model	Accuracy: 96.18%
Zhang P. et al., 2023 [54]	Respiratory sound base by Rocha et al., recorded with a digital stethoscope (Littmann 3200, 3M) N=920 recordings: • Asthma (N=1)	Data preprocessing. Feature extracting: • Mel-frequency cepstrum coefficients • Chromogram • Mel-scaled spectrogram • Spectral contrast Tonal centroids	Multiple models were evaluated: • Convolutional neural network (CNN) • Long short-term memory (LSTM) • Convolutional neural network with long short-term memory (CNN-LSTM)	• COPD • Healthy • URTI • Bronchiectasis • Pneumonia Bronchiolitis	80% of data was used for training. 20% of data was used for testing.	Accuracy Precision Recall F1-score	LSTM produced the highest scores: • Accuracy: 98.82% • Precision: 0.96 • Recall: 0.99 • F1-score: 0.97

	• Bronchiectasis (N=16) • Bronchiolitis (N=13) • COPD (N=185) • Healthy (N=35) • LRTI (N=2) • Pneumonia (N=35) UTI (N=23)		• Convolutional neural network with bidirectional long short-term memory (CNN-BLSTM)				
Zhang Y. et al., 2024 [41]	ICBHI 2017 Dataset	Preprocessing: • Time stretching • Pitch shifting • Background noise • Data clipping Feature extraction: • Mel-frequency cepstral coefficients (MFCC) Mel spectrograms	Dual-channel CNN-LSTM model	• Bronchiectasis • URTI • Pneumonia • Healthy • COPD Bronchiolitis	65% of data was used for training. 16% of data was used for verification. 19% of data was used for testing.	K-fold cross validation was used to select the optimal model. Values: • Accuracy • Precision • Recall F1-score	The proposed dual-channel CNN-LSTM performed the best: • Accuracy 99.01% • Precision: 99.17% • Recall: 99.13% • F1-score: 0.9915
Divakaran et aL, 2025 [34]	N=600 Respiratory sound recordings, collected from 220 individuals, including 120 diagnosed COPD patients and 100 healthy controls. Recordings done by conventional stethoscope chest piece, adapted to enable electronic respiratory sound acquisition. Microphone	Preprocessing steps: • Amplitude normalization • De-noising • Segmentation into overlapping time frames Feature extraction by Mel-Frequency Cepstral Coefficients (MFCC). Application of Fast Fourier Transform (FFT), mel filterbank analysis, logarithmic scaling and Discrete Cosine Transform (DCT).	A newly developed ANN (Artificial Neural Network)	COPD vs. healthy	Training and evaluation by using extracted MFCC features. Evaluation through hardware validation, model accuracy benchmarking, user interface implementation.	Outcome measures: • Accuracy • Precision • Recall • F1-score	Accuracy: 96.36% Precision: 96.23% Recall: 96.1% F1-score: 96.16%

	output was interfaced with an Android-based smartphone.						
Topaloglu et al., 2025 [35]	N=3600 Respiratory sound recordings from 120 participants (60 healthy, 60 diagnosed with stable, not exacerbated asthma according to GINA, not currently having). Recorded with Littman CORE Digital Stethoscope in sound-treated rooms.	Samples analyzed by Mel-Frequency Cepstral Coefficients (MFCCs) and Tunable Q-Factor Wavelet Transform (TQWT).	Quadratic Support Vector Machine (SVM) and Narrow Neural Network (NNN) models. After initial analysis, the Quadratic SVM was chosen as the superior model of the two.	Asthma vs. healthy	10-fold cross-validation method was used to validate the machine learning models. 90% of data was used for training.	Outcome measures: • Accuracy • Precision • Sensitivity • Specificity • F1-score In comparison to existing models: • ICBIH2017 Topaloglu et al.	Accuracy: • ICBIH2017: 98.13% • Topaloglu et al.: 93.44% • Proposed model: 99.86% Precision: • ICBIH2017: 98.43% • Topaloglu et al.: 88.52% • Proposed model: 99.88% Sensitivity: • ICBIH2017: 98.65% • Topaloglu et al.: 80.44% • Proposed model: 99.83% Specificity: • ICBIH2017: 97.20% • Topaloglu et al.: 99.44% • Proposed model: 99.88% F1-score: • ICBIH2017: 98.54% • Topaloglu et

							al.: 88.57% Proposed model: 99.86%

Table S 7: Summary of results from articles describing the diagnostic qualities of automatic sound classification without comparison to other methods (n=14)

11.5 Quality Control

11.5.1 AMSTAR quality assessment

1. Abreu et al. — Computerized respiratory sounds in paediatrics: A systematic review [22]

AMSTAR-2 Domain	Rating	Explanation
1. Research question and inclusion criteria include PICO components	Yes	The review clearly defined the population (pediatric patients), intervention (computerized respiratory sound analysis), and outcomes related to diagnostic performance and technology use.
2. Protocol registered before the review	No	No PROSPERO registration or pre-published protocol reported.
3. Explanation of study designs included	Yes	The review justified inclusion of observational and technical studies due to the emerging nature of computerized auscultation research.
4. Comprehensive literature search strategy	Yes	Multiple databases were searched, and keywords and selection strategy were described.
5. Study selection performed in duplicate	Yes	Two independent reviewers screened titles, abstracts, and full texts.
6. Data extraction performed in duplicate	Yes	Data extraction was conducted independently by two reviewers.
7. List of excluded studies provided	No	A full list of excluded studies with reasons was not provided.
8. Description of included studies	Yes	Characteristics of included studies (population, technology, outcomes) were clearly described.
9. Risk of bias assessment of included studies	Partial Yes	Methodological limitations of studies were discussed, but a standardized risk of bias tool was not consistently applied.
10. Sources of funding of included studies reported	No	Funding sources of included primary studies were not systematically reported.
11. Appropriate methods for meta-analysis	No	No meta-analysis was performed; only narrative synthesis.
12. Assessment of impact of risk of bias on results	Partial Yes	The discussion acknowledged methodological limitations but did not formally assess their influence on findings.
13. Consideration of risk of bias when interpreting results	Yes	Authors acknowledged heterogeneity and methodological limitations of included studies.
14. Explanation of heterogeneity	Yes	Differences between studies (technologies, populations, diagnostic methods) were discussed.
15. Investigation of publication bias	No	Publication bias was not formally assessed.
16. Conflict of interest reported	Yes	Authors declared potential conflicts of interest.

Table S 8: AMSTAR quality assessment of study mentioned above.

Conclusion

Overall AMSTAR-2 confidence: Moderate

This systematic review demonstrates several methodological strengths, including a clearly defined research question, a structured search strategy across multiple databases, and an appropriate

description of included studies. However, some methodological limitations were identified, including incomplete reporting of excluded studies, lack of reporting of funding sources of included studies, and no formal assessment of publication bias. These limitations reduce the overall methodological confidence to a moderate level.

2. Park et al. — Artificial Intelligence Models for Pediatric Lung Sound Analysis: Systematic Review and Meta-Analysis [24]

AMSTAR-2 Domain	Rating	Explanation
1. Research question and inclusion criteria include PICO components	Yes	The review clearly defined population (paediatric patients), intervention (AI-based lung sound analysis), and diagnostic outcomes.
2. Protocol registered before the review	Yes	A review protocol was registered prior to study conduct.
3. Explanation of study designs included	Yes	Inclusion criteria for diagnostic accuracy and machine learning studies were justified.
4. Comprehensive literature search strategy	Yes	Multiple databases were searched with clearly reported search strategy.
5. Study selection performed in duplicate	Yes	Two reviewers independently screened studies.
6. Data extraction performed in duplicate	Yes	Data extraction was conducted by two independent reviewers.
7. List of excluded studies provided	Yes	A PRISMA flow diagram and excluded studies were reported.
8. Description of included studies	Yes	Included studies were comprehensively summarized in tables.
9. Risk of bias assessment of included studies	Yes	Risk of bias was assessed using appropriate methodological tools.
10. Sources of funding of included studies reported	Partial Yes	Funding information was reported for several studies but not all.
11. Appropriate methods for meta-analysis	Yes	Meta-analysis was conducted using appropriate statistical models for pooled diagnostic accuracy estimates.
12. Assessment of impact of risk of bias on results	Yes	Sensitivity analyses and subgroup analyses were performed.
13. Consideration of risk of bias when interpreting results	Yes	Authors discussed methodological limitations of included studies when interpreting pooled estimates.
14. Explanation of heterogeneity	Yes	Statistical heterogeneity was evaluated and discussed.
15. Investigation of publication bias	Yes	Publication bias was assessed using funnel plots and statistical tests.
16. Conflict of interest reported	Yes	Conflicts of interest and funding sources were clearly declared.

Table S 9: AMSTAR quality assessment of study mentioned above.

Conclusion
Overall AMSTAR-2 confidence: High

This systematic review and meta-analysis demonstrates several methodological strengths, including a clearly defined research question, a comprehensive literature search strategy, duplicate study selection and data extraction, and the use of appropriate methods for quantitative synthesis. Nevertheless, methodological limitations were identified, including limited reporting of funding sources of included studies and the absence of a detailed list of excluded studies. These limitations result in an overall moderate level of methodological confidence.

3. Ahmed et al. — Evaluation of machine learning to detect adventitious lung sounds using digital auscultation to aid childhood pneumonia diagnosis [50]

AMSTAR-2 Domain	Rating	Explanation
1. Research question and inclusion criteria include PICO components	Yes	The review clearly defined the population (children under 5 years), index test (digital auscultation/computerized analysis), reference standard (physician acoustic analysis), target condition (pneumonia), and diagnostic accuracy outcomes.
2. Protocol registered before the review	Yes	The review protocol was registered in PROSPERO (CRD42020180821).
3. Explanation of study designs included	Yes	The review specified inclusion of observational and experimental studies evaluating diagnostic performance.
4. Comprehensive literature search strategy	Yes	Multiple databases were searched (MEDLINE, Embase, CINAHL Plus, Web of Science, Global Health, IEEE Xplore, Scopus, and ClinicalTrials.gov), with reference list screening and no language restrictions.
5. Study selection performed in duplicate	Yes	Two reviewers independently screened titles/abstracts and full texts, with disagreements resolved by discussion or arbiter.
6. Data extraction performed in duplicate	Yes	Data extraction was independently conducted by two reviewers.
7. List of excluded studies provided	Partial Yes	Reasons for exclusion were documented and a PRISMA flow diagram was provided, but no full itemized list of all excluded full-text studies was presented in the main paper.
8. Description of included studies	Yes	Included studies were described in detail in Table 1, including country, study type, population, number of subjects, pathology analyzed, recording device, feature extraction, classification method, and outcomes.
9. Risk of bias assessment of included studies	Yes	Risk of bias was assessed using QUADAS-2 by two independent reviewers.
10. Sources of funding of included studies reported	No	The review did not systematically report funding sources of the included primary studies.
11. Appropriate methods for meta-analysis	Yes	No meta-analysis was performed because the authors explicitly judged the studies too heterogeneous in methodology and outcome measures; narrative synthesis was appropriate.
12. Assessment of	Partial	The authors discussed methodological limitations and

impact of risk of bias on results	Yes	unclear quality of included studies, but did not perform a formal sensitivity analysis based on risk of bias.
13. Consideration of risk of bias when interpreting results	Yes	The interpretation of findings explicitly acknowledged poor reporting, methodological heterogeneity, and unclear applicability of many included studies.
14. Explanation of heterogeneity	Yes	The review discussed heterogeneity in study populations, diseases, age groups, feature extraction methods, and classification algorithms.
15. Investigation of publication bias	No	Publication bias was not formally assessed.
16. Conflict of interest reported	Yes	Conflicts of interest for the review authors were reported.

Table S 10: AMSTAR quality assessment of study mentioned above.

Conclusion
Overall AMSTAR-2 confidence: Moderate
This systematic review demonstrates several methodological strengths, including protocol registration, a comprehensive literature search, duplicate study selection and data extraction, and appropriate risk of bias assessment of included studies using QUADAS-2. However, some methodological limitations were identified, including lack of reporting of funding sources of included studies and absence of formal publication bias assessment. These limitations result in an overall moderate level of methodological confidence.

11.5.2 QUADAS assessment

Article. Andrès E, Reichert S, Brandt C, Hill N, Gass R. Development and experimentation of a new digital communicating and intelligent stethoscope.
Study type. Diagnostic accuracy study
OCEBM level. Level 2

Domain	Risk of Bias	Applicability Concerns	Explanation
Patient Selection	Low	Low	Consecutive hospitalized patients in a real clinical setting were included, reducing selection bias; however, this may limit applicability to primary care due to differences in disease spectrum and severity.
Index Test	Unclear	Low	The digital stethoscope (eStetho) was applied in clinical practice, but it is unclear whether interpretation was blinded or whether diagnostic thresholds were predefined.
Reference Standard	Low	Low	Final diagnosis was based on clinical evaluation, medical file review, echocardiography, and lung CT scan,

			constituting a valid clinical reference standard.
Flow and Timing	Low	Low	Patients appear to have received both index test and reference standard within the same clinical setting, with no evidence of differential verification or exclusion.

Table S 11: QUADAS-2 quality assessment of article mentioned above.

Conclusion. Overall methodological quality was moderate to good, with low risk of bias in most domains. Some uncertainty remains regarding the conduct and interpretation of the index test, particularly concerning blinding and pre-specification.

Article. McCollum ED, Park DE, Watson NL, et al. Digital auscultation in PERCH: Associations with chest radiography and pneumonia mortality in children.
Study type. Observational diagnostic study
OCEBM level. Level 2

Domain	Risk of Bias	Applicability Concerns	Explanation
Patient Selection	Unclear	Low	A large multicenter cohort of hospitalized children with pneumonia was included. however, partial convenience sampling may introduce selection bias and the hospital-based pediatric population may limit applicability to primary care settings.
Index Test	Low	Low	Digital lung sounds were recorded using a standardized protocol and interpreted by trained blinded listening panels, reducing measurement bias.
Reference Standard	Unclear	Moderate	Chest radiography using World Health Organization criteria was used as the reference standard, but this method has recognized limitations and is not a perfect gold standard for pneumonia.
Flow and Timing	Low	Low	Most patients had both lung recordings and chest radiography with a short median time interval, supporting appropriate temporal alignment.

Table S 12: QUADAS-2 quality assessment of article mentioned above.

Conclusion. Overall methodological quality was moderate to good, with strengths including the large multicenter prospective design and standardized assessment procedures. However,

limitations arise from the imperfect reference standard and potential selection bias at some sites.

Article. Gottlieb ER, Aliotta JM, Tammaro D. Comparison of analogue and electronic stethoscopes for pulmonary auscultation by internal medicine residents.
Study type. Diagnostic comparison study
OCEBM level. Level 3

Domain	Risk of Bias	Applicability Concerns	Explanation
Patient Selection	High	High	Participants were internal medicine residents evaluating simulated cases rather than real patients, resulting in a non-representative study population and high risk of selection bias; this design also limits applicability to real-world primary care.
Index Test	Unclear	Low	The index test was applied in a standardized way, but it is unclear whether interpretation was blinded and whether diagnostic criteria were predefined.
Reference Standard	High	High	The reference standard consisted of predefined correct answers based on fictional cases rather than an independent clinical gold standard.
Flow and Timing	Low	Low	All participants completed the same structured cases under similar conditions, with consistent comparison between groups and no evidence of differential verification.

Table S 13: QUADAS-2 quality assessment of article mentioned above.

Conclusion. Overall methodological quality was low, with a high risk of bias due to the use of simulated cases and the absence of a true clinical reference standard. While the randomized design strengthens internal comparison, the findings have limited applicability to real clinical practice.

Article. Park DE, Watson NL, Focht C, et al. Digitally recorded and remotely classified lung auscultation compared with conventional stethoscope classifications among children aged 1–59 months enrolled in the PERCH case–control study.
Study type. Diagnostic comparison study
OCEBM level. Level 3

Domain	Risk of Bias	Applicability Concerns	Explanation
Patient Selection	Unclear	Low	A large multicenter sample of pneumonia cases and controls

			was used, but the digital auscultation sub study consisted of a convenience subset of the PERCH population, introducing potential selection bias.
Index Test	Low	Low	Digital auscultation recordings were standardized, denoised, and interpreted by a blinded expert listening panel using predefined criteria.
Reference Standard	High	High	No true gold standard was used; the study compared digital auscultation to conventional auscultation, which itself has known interobserver variability and limited diagnostic accuracy.
Flow and Timing	Unclear	Moderate	Not all auscultation measurements were performed concurrently, which may introduce timing bias and affect the validity of the diagnostic accuracy results.

Table S 14: QUADAS-2 quality assessment of article mentioned above.

Conclusion. Overall methodological quality was moderate, with strengths including the large multicenter design and blinded assessment of digital recordings. However, the absence of a true reference standard and non-concurrent measurements limit the interpretation of concordance and diagnostic validity.

Article. Kevat A, Kalirajah A, Roseby R. Artificial intelligence accuracy in detecting pathological breath sounds in children using digital stethoscopes.
Study type. Diagnostic modelling study
OCEBM level. Level 3

Domain	Risk of Bias	Applicability Concerns	Explanation
Patient Selection	Unclear	Moderate	Recordings were obtained from 25 pediatric patients recruited sequentially in a clinical setting, however, the small sample size may limit the reliability of the diagnostic accuracy estimates.
Index Test	Unclear	Low	The artificial intelligence algorithm was applied in a blinded manner to recordings, but diagnostic thresholds were optimized post hoc using receiver operating characteristic curves, which may introduce bias.

Reference Standard	High	High	The reference standard consisted of expert clinician labelling based on audio playback and spectrogram analysis, without a true gold standard for lung sound classification.
Flow and Timing	Low	Low	All recordings were analyzed using both the reference standard and the artificial intelligence algorithm, with no exclusions after analysis and consistent processing of all samples, reducing the risk of bias in the flow and timing domain.

Table S 15: QUADAS-2 quality assessment of article mentioned above.

Conclusion. Overall methodological quality was moderate to low, with major concerns related to the absence of a true reference standard and small sample size. Although the study used blinding and real-world recordings, limitations in labelling and validation reduce confidence in the reported diagnostic performance.

Article. Gupta S, Agrawal M, Deepak D. Correlating spirometry findings with auscultation sounds for diagnosis of respiratory diseases.
Study type. Diagnostic modelling study
OCEBM level. Level 3

Domain	Risk of Bias	Applicability Concerns	Explanation
Patient Selection	Unclear	Moderate	A dataset of 137 participants was used, but the recruitment strategy and representativeness are not clearly described. The data were collected in different environments, introducing potential selection bias, which may affect the reliability and generalizability of the diagnostic accuracy results possibly leading to over- or underestimation of diagnostic performance.
Index Test	High	Moderate	The machine learning model was developed and evaluated on the same dataset using cross-validation, without external validation, increasing risk of overfitting and optimistic performance estimates.
Reference Standard	Low	Low	Spirometry, including forced expiratory volume in one second,

			forced vital capacity, and their ratio, was used as the reference standard, which is appropriate for classification of obstructive and restrictive disease.
Flow and Timing	High	Moderate	The study used retrospective data with feature extraction and model training pipelines; it is unclear whether separation between training and testing was fully independent, raising concerns about data leakage and potential overestimation of model performance, while the lack of external validation limits the reliability and applicability of the results to clinical practice.

Table S 16: QUADAS-2 quality assessment of article mentioned above.

Conclusion. Overall methodological quality was low to moderate, with major concerns related to the use of internal validation only and risk of overfitting of the machine learning model. While the reference standard is appropriate, limitations in study design and validation reduce confidence in the reported diagnostic performance.

11.5.3 PROBAST assessment of included prediction model studies

S. Chatterjee, J. Roychowdhury and A. Dey, D-Cov19Net: A DNN based COVID-19 detection system using lung sound [36]

Domain	Risk of Bias	Applicability	Explanation
Participants	High	High	Small dataset, unclear selection
Predictors	Moderate	Moderate	Preprocessed audio features
Outcome	High	High	No clinical gold standard
Analysis	High	High	No external validation

Table S 17: PROBAST quality assessment of article mentioned above.

Conclusion: High risk of bias due to small datasets, lack of external validation, and non-clinical outcome definitions.

R. Gupta, R. Singh, C. M. Travieso-González, R. Burget and M. Kishore Dutta, DeepRespNet: A deep neural network for classification of respiratory sounds [27]

Domain	Risk of Bias	Applicability	Explanation
Participants	Moderate	High	Single-center dataset
Predictors	Low	Moderate	Standard spectrogram input
Outcome	High	High	Non-clinical labels
Analysis	Moderate	Moderate	Internal validation only

Table S 18: PROBAST quality assessment of article mentioned above.

Conclusion: High risk of bias due to small datasets, lack of external validation, and non-clinical

outcome definitions.

O. Karaarslan, K. D. Belcastro and O. Ergen, Respiratory sound-base disease classification and characterization with deep/machine learning techniques [37]

Domain	Risk of Bias	Applicability	Explanation
Participants	High	High	Public dataset bias
Predictors	Moderate	Moderate	Complex feature engineering
Outcome	High	High	No clinical confirmation
Analysis	High	High	No external validation

Table S 19: PROBAST quality assessment of article mentioned above.

Conclusion: High risk of bias due to small datasets, lack of external validation, and non-clinical outcome definitions.

Y. Ma, X. Z. Xu, Q. Yu, Y. H. Zhang, Y. F. Li, J. Zhao, et al., LungBRN: A Smart Digital Stethoscope for Detecting Respiratory Disease Using bi-ResNet Deep Learning Algorithm [29]

Domain	Risk of Bias	Applicability	Explanation
Participants	Moderate	Moderate	More balanced dataset
Predictors	Low	Moderate	Deep learning features
Outcome	Moderate	Moderate	Partial clinical relevance
Analysis	Moderate	Moderate	Internal validation

Table S 20: PROBAST quality assessment of article mentioned above.

Conclusion: High risk of bias due to small datasets, lack of external validation, and non-clinical outcome definitions.

S. Z. Hassan Naqvi and M. A. Choudhry, An Automated System for Classification of Chronic Obstructive Pulmonary Disease and Pneumonia Patients Using Lung Sound Analysis [49]

Domain	Risk of Bias	Applicability	Explanation
Participants	High	High	Very small dataset
Predictors	Moderate	Moderate	Extensive features
Outcome	High	High	Dataset labels only
Analysis	High	High	Overfitting risk

Table S 21: PROBAST quality assessment of article mentioned above.

Conclusion: High risk of bias due to small datasets, lack of external validation, and non-clinical outcome definitions.

A. F. Mahmood, A. M. Alkababji and A. Daood, Resilient embedded system for classification respiratory diseases in a real time [30]

Domain	Risk of Bias	Applicability	Explanation
Participants	High	High	Imbalanced dataset
Predictors	Moderate	Moderate	Complex features
Outcome	High	High	No gold standard
Analysis	High	High	Augmentation + no validation

Table S 22: PROBAST quality assessment of article mentioned above.

Conclusion: High risk of bias due to small datasets, lack of external validation, and non-clinical outcome definitions.

S. Sharma, S. Pandey and D. R. Shah, Enhancing Medical Diagnosis with AI: A Focus on Respiratory Disease Detection [53]

Domain	Risk of Bias	Applicability	Explanation
Participants	High	High	Kaggle dataset bias
Predictors	Low	Moderate	Standard CNN input
Outcome	High	High	Non-clinical labels
Analysis	High	High	Augmentation + no external validation

Table S 23: PROBAST quality assessment of article mentioned above.

Conclusion: High risk of bias due to small datasets, lack of external validation, and non-clinical outcome definitions.

R. Uppin, S. Ambesange, S. Aralikatti and M. G. Ve, Respiratory Sound Abnormality Classification using Multipath Deep Learning Method [32]

Domain	Risk of Bias	Applicability	Explanation
Participants	High	High	Small dataset
Predictors	Low	Moderate	MFCC + spectrogram
Outcome	High	High	Annotated labels
Analysis	High	High	Overfitting risk

Table S 24: PROBAST quality assessment of article mentioned above.

Conclusion: High risk of bias due to small datasets, lack of external validation, and non-clinical outcome definitions.

P. Zhang, A. Swaminathan and A. A. Uddin, Pulmonary disease detection and classification in patient respiratory audio files using long short-term memory neural networks [54]

Domain	Risk of Bias	Applicability	Explanation
Participants	High	High	ICBHI dataset
Predictors	Low	Moderate	MFCC features
Outcome	High	High	Non-clinical labels
Analysis	High	High	No external validation

Table S 25: PROBAST quality assessment of article mentioned above.

Conclusion: High risk of bias due to small datasets, lack of external validation, and non-clinical outcome definitions.

Y. Zhang, Q. Huang, W. Sun, F. Chen, D. Lin and F. Chen , Research on lung sound classification model based on dual-channel CNN-LSTM algorithm [41]

Domain	Risk of Bias	Applicability	Explanation
Participants	High	High	ICBHI dataset
Predictors	Low	Moderate	CNN-LSTM features
Outcome	High	High	Dataset labels
Analysis	High	High	No external validation

Table S 26: PROBAST quality assessment of article mentioned above.

Conclusion: High risk of bias due to small datasets, lack of external validation, and non-clinical outcome definitions.

Park et al. – Pediatric lung sound classification model (PERCH) case-control study [54]

Domain	Risk of Bias	Applicability	Explanation
Participants	Moderate	Moderate	Single-center pediatric outpatient dataset with limited sample size
Predictors	Low	Moderate	Standardized digital recordings and feature extraction
Outcome	High	High	Clinician-labelled lung sounds without gold standard
Analysis	Moderate	Moderate	Internal + small prospective validation, no external validation

Table S 27: PROBAST quality assessment of article mentioned above.

Conclusion: High risk of bias due to reliance on subjective outcome labels and lack of external validation.

Rady et al. – Wheeze sound analysis using DSP techniques [54]

Domain	Risk of Bias	Applicability	Explanation
Participants	High	High	Mixed datasets with unclear selection and representativeness
Predictors	Moderate	Moderate	DSP features without standardization
Outcome	High	High	Algorithm-based classification without clinical gold standard
Analysis	High	High	No external validation, small datasets

Table S 28: PROBAST quality assessment of article mentioned above.

Conclusion: High risk of bias due to non-clinical outcomes, small datasets, and lack of validation.

11.5.4 Other critical appraisal tools used

Study: Wang et al.[46] – Observational study (cross-sectional, CASP)
OCEBM level. Level 4

CASP Domain	Assessment	Explanation
Clear focus	Yes	Aims to describe auscultation characteristics in COVID-19
Appropriate design	Yes	Cross-sectional design appropriate
Recruitment strategy	Yes	Confirmed COVID-19 patients included
Measurement of exposure	Yes	Standardized electronic auscultation
Measurement of outcome	Moderate	Subjective interpretation of sounds
Confounding factors	Moderate	Limited adjustment for patient variability
Follow-up	Not applicable	Cross-sectional
Precision of results	Moderate	Descriptive statistics only
Applicability	Moderate	Hospital setting limits generalizability

Table S 29: CASP quality assessment of study mentioned above.

Conclusion. This study demonstrates the feasibility and clinical applicability of electronic

auscultation but lacks objective diagnostic validation and control for confounding factors.

Study: Noda et al.[44] – Case report (JBI checklist)
OCEBM level. Level 5

Domain	Assessment	Explanation
Patient demographics clearly described	Yes	Age, clinical presentation, and relevant background characteristics were adequately reported.
Patient history clearly described	Yes	Clinical history and presenting symptoms were sufficiently detailed.
Clinical condition on presentation	Yes	The initial clinical status and examination findings were clearly documented.
Diagnostic tests and results clearly described	Yes	Lung auscultation findings using both digital and conventional stethoscopes were reported.
Intervention(s) clearly described	Yes	Use of digital auscultation and remote classification was adequately explained.
Post-intervention clinical condition described	Yes	Follow-up findings and diagnostic interpretation were reported.
Adverse events identified	No	No adverse events were reported or discussed.
Take-away lessons provided	Yes	The study provided clinically relevant insights into digital auscultation feasibility.

Table S 30: CASP quality assessment of study mentioned above

This case report demonstrates moderate methodological quality, with clear and structured reporting of clinical presentation and diagnostic procedures. However, its evidentiary value remains limited due to the inherent constraints of a single-case design and the absence of systematic validation.

Study: Kim et al.[55] – Narrative review (SANRA)
OCEBM level. Level 5

SANRA Item	**Rating**	**Explanation**
Justification of article importance	Yes	Clearly explains relevance of AI in auscultation
Statement of aims	Yes	Objectives well defined
Literature search	Partial	Not systematic
Referencing	Yes	Relevant references included
Scientific reasoning	Yes	Logical synthesis of evidence
Appropriate presentation of data	Yes	Well-structured discussion

Table S 31: SANRA quality assessment of study mentioned above.

Conclusion. This narrative review provides a structured overview of AI in respiratory sound analysis, but lacks systematic methodology, limiting its methodological rigor.