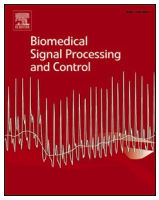
Biomedical Signal Processing and Control 72 (2022) 103325

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



The use of artificial intelligence systems in diagnosis of pneumonia via signs and symptoms: A systematic review

Katy Stokes [a], Rossana Castaldo [b], Carlo Federici [a,c], Silvio Pagliara [a], Alessia Maccaro [a], Francesco Cappuccio [d,e], Giuseppe Fico [f], Marco Salvatore [b], Monica Franzese [b], Leandro Pecchia [a,*]

[a] School of Engineering, University of Warwick, Coventry CV4 7AL, UK
[b] IRCCS SDN, Via E. Gianturco, 113, 80143 Naples, Italy
[c] SDA Bocconi School of Management, Centre of Research on Health and Social Care Management, Milano, Italy
[d] Division of Health Sciences (Mental Health & Wellbeing), Warwick Medical School, University of Warwick, Coventry, UK
[e] University Hospitals Coventry & Warwickshire NHS Trust, Coventry, UK
[f] Life Supporting Technologies, Universidad Politécnica de Madrid, Madrid, Spain

ARTICLE INFO

Keywords:
Systematic review
Pneumonia
Artificial intelligence
Machine learning
Diagnosis
Predictive model

ABSTRACT

Artificial Intelligence (AI) systems using symptoms/signs to detect respiratory diseases may improve diagnosis especially in limited resource settings. Heterogeneity in such AI systems creates an ongoing need to analyse performance to inform future research. This systematic literature review aimed to investigate performance and reporting of diagnostic AI systems using machine learning (ML) for pneumonia detection based on symptoms and signs, and to provide recommendations on best practices for designing and implementing predictive ML algorithms. This article was conducted following the PRISMA protocol, 876 articles were identified by searching PubMed, Scopus, and Ovid[SP] databases (last search 5th May 2021). For inclusion, studies must have differentiated clinically diagnosed pneumonia from controls or other diseases using AI. Risk of Bias was evaluated using The STARD 2015 tool. Information was extracted from 16 included studies regarding study characteristics, ML-model features, reference tests, study population, accuracy measures and ethical aspects. All included studies were highly heterogenous concerning the study design, setting of diagnosis, study population and ML algorithm. Study reporting quality in methodology and results was low. Ethical issues surrounding design and implementation of the AI algorithms were not well explored. Although no single performance measure was used in all studies, most reported an accuracy measure over 90%. There is strong evidence to support further investigations of ML to automatically detect pneumonia based on easily recognisable symptoms and signs. To help improve the efficacy of future research, recommendations for designing and implementing AI tools based on the findings of this study are provided.

1. Introduction

Pneumonia is a form of acute lower respiratory infection. Pneumonia is generally characterized by specific symptoms such as fever, chills, cough with sputum production, chest pain and shortness of breath [1]. Many factors affect how serious pneumonia is, such as the type of pathogen causing the lung infection, age, and overall health status. Pneumonia tends to be more serious for children under the age of five, adults over the age of 65, people with certain conditions such as heart failure, diabetes, or COPD (chronic obstructive pulmonary disease), or people who have weak immune systems due to HIV/AIDS, chemotherapy (a treatment for cancer), or organ or blood and marrow stem cell transplant procedures [2].

When an individual has pneumonia, the alveoli, small sacs within the lungs, are filled with pus and fluid, which makes breathing painful and limits oxygen exchange [3]. There are more than 30 different causes of pneumonia, and they are grouped accordingly: bacterial pneumonia, viral pneumonia, mycoplasma pneumonia and other pneumonias. Moreover, pneumonias can be also categorized as community-acquired (CAP), hospital-acquired (HAP) (excluding ventilator-associated [4],

* Corresponding author.
E-mail address: l.pecchia@warwick.ac.uk (L. Pecchia).

https://doi.org/10.1016/j.bspc.2021.103325
Received 5 July 2021; Received in revised form 17 September 2021; Accepted 2 November 2021
Available online 12 November 2021

which occurs in immunocompromised patients such as patients with human immunodeficiency virus (HIV) infection (see Pneumocystis jirovecii Pneumonia [5], or aspiration pneumonia, which occurs when large volumes of upper airway or gastric secretions enter into the lungs [6–9].

An accurate definition and diagnosis of pneumonia is contentious for several reasons [2]: low specificity of symptoms of lower respiratory tract infections; difficulty in identifying the underlying pathogen in individuals and lack of widespread availability of laboratory tests and imaging. Diagnosis is suggested by a history of cough, dyspnoea, pleuritic pain, or acute functional or cognitive decline, with abnormal vital signs (e.g., fever, tachycardia) and lung examination findings. Diagnosis should be confirmed by chest radiography or ultrasonography.

This uncertainty and the above-mentioned categorizations lead to empirical treatment selection. However, pneumonia is a leading cause of hospitalization in both children and adults. Most cases can be treated successfully, although it can take weeks to fully recover [2]. In many instances, pneumonia is severe, requiring hospitalization and in some cases, people with severe health conditions need to be treated in ICUs (intensive care units). In the last decades, new complications of viral infections by Coronaviruses have been identified. Coronaviruses are a large family of viruses that cause illness ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). SARS-Coronavirus 2 (SARS-CoV-2) is a new strain firstly identified in humans in 2019 and causes Coronavirus Disease (COVID-19) that can spread to the lungs, causing pneumonia. It presents predominantly with fever, persistent cough, fatigue, dyspnoea, loss of smell and taste [10–11]. While many people recover, some develop SARS requiring hospitalisation, with escalation to intensive care support with oxygen, mechanical ventilation and, eventually, death [12]. A novel approach to improve diagnosis and prognosis of pneumonia is the use of biomarkers [13–14]. The diagnostic and prognostic role of procalcitonin (PCT) and mid-regional-pro-adrenomedullin (MR-proADM) were investigated in patients with pneumonia with high positive predictive value.

Confirmation of pneumonia is not trivial and will be by nature context dependent, relying on a combination of what is available from clinical presentation, laboratory tests and diagnostic imaging.

Recent studies push towards the adoption of artificial intelligence (AI) models amplifying diagnostic accuracy in radiology [15]. However, radiography suffers several disadvantages: low sensitivity to early stage pneumonia, lack of standardised interpretation [16], inter-rater variability [17–18], absence of abnormalities in the chest radiographs of children [19] and potential harm due to exposure to x-rays. The biggest shortfall is that radiography is not widely available in low-income settings, which represent the areas with the highest disease burden.

The breadth of challenges to diagnosing pneumonia, especially in low and middle-income countries (LMICs) highlight the potential benefit of a specific, sensitive diagnostic tool for pneumonia. In particular, a major issue is mistaken diagnosis of respiratory diseases due to overlapping symptoms which may offer similar clinical presentation but have differing underlying causes and respond best to different treatments [20], for example pneumonia may be caused by bacteria and require antibiotics, whereas viruses may be the most likely cause of bronchitis [21].

One subset of AI, known as machine learning (ML), which is able to learn, reason, and self-correct without explicit programming, has the potential to provide such a solution. ML could play a major role within the practice of clinical medicine. Moreover, in the last few decades a particular subset of ML, so-called deep learning (DL) based on artificial neural networks (ANNs), is expanding the potential of ML in clinical practise [22].

In the case of pneumonia, ML has been shown to be promising in strengthening diagnostic accuracy when applied to hospitalized patients [23–24]. Despite numerous publications in this field, there are few cases of successful translation of ML techniques to clinical settings across the board [25].

In light of this, it is of great importance that researches consider the clinical setting and end user of their models.

As such, a set of predictors which are easily recognised or even self-reported and a model which is suitable for incorporation into a referral or diagnostic tool such as an APP for mobile phones will be key requirements for assisting diagnosis of pneumonia in low or middle-income settings. Therefore, the use of ML systems to detect respiratory diseases via non-invasive measures such as signs and symptoms is gaining momentum. Indeed, such diagnostic tools are emerging as a route to facilitating successful task redistribution and improving access to accurate diagnosis in araeas with low numbers of qualified clinical staff [26]. However, due to the heterogeneity and diversity of ML systems, there is an ongoing need to assess their performance in order to identify gaps in the research, impact improvements in practices and facilitate future comparative studies. To the best of our knowledge this is the first review of the application of ML to symptom-based detection of pneumonia. The research question we addressed is what symptom-based ML predictive models have been developed and how well do they perform? In this way, the aims of this study were to assess both the performance of published ML methods to diagnose pneumonia based on symptoms or signs, and the reporting quality of these studies.

Therefore, the main contributions of our work were: (1) to show a systematic synthesis of the existing studies which proposed ML algorithms to diagnose pneumonia based on signs and symptoms, (2) to identify common most frequently used symptoms as ML features, (3) the best ML methods and performance.

Based on our findings, we provided a recommended pipeline to design and implement predictive algorithms, with critical steps to follow to achieve a generalised and robust ML model. We anticipate that our findings and recommendations will be constructive in guiding future research and facilitating it is translation into clinical tools.

2. Materials and methods

This systematic review was conducted and reported in accordance with PRISMA guidelines for systematic reviews and meta-analyses (PRISMA checklist) [27–28] and the recommendations from the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy [29]. The methods of the literature review as well as the inclusion and exclusion criteria were specified beforehand in a protocol (available from the authors on request). All types of studies were included if they reported on the use of artificial intelligence (AI) systems such as machine learning (ML) or deep learning (DL) techniques applied to distinguishing pneumonia based on signs and symptoms. The STARD 2015 tool [30], which has been developed to assess the reporting quality of diagnostic accuracy studies, was used to rate the quality of the studies included in the systematic review.

2.1. Search strategy

Potentially relevant studies were identified by searching PubMed, Scopus, and Embase (through Ovid^{sp}) electronic databases published form 2010 to May 2021. It has been shown that searching multiple databases increases the overall recall in systematic reviews, however, there is a limit to the practicality of searching many databases [31]. Therefore, we selected three databases with good evidence of recall, which are appropriate for multidisciplinary research [32–33]. Only studies in English were selected for screening.

A broad search strategy including combinations of search terms for (i) the index test under evaluation and (ii) the target condition of interest was first developed for PubMed, and then adapted to all other databases. The full search strategy is reported in Supplementary Table 1. During the search, no methodology search filters to identify diagnostic test accuracy studies were used to avoid missing relevant records. In addition, in order to identify recent diagnostic accuracy test studies concerning the

diagnosis of pneumonia in patients affected by the recent pandemic of SARS-CoV-2 coronavirus infection, the MedRxiv server of preprints was searched using different combinations of search terms including "COVID-19", "SARS-CoV-2", "diagnosis", "Pneumonia", "signs", "symptoms", "artificial intelligence", and "machine learning". Finally, a linear reference search was conducted by checking the references of the studies identified in the index search that met the review's inclusion criteria. Two researchers (CF and KS), who were blinded to the author information of the articles, independently screened all identified records for inclusion. In case of disagreements, a third author (RC) was consulted.

2.2. Inclusion and Exclusion criteria

A set of inclusion and exclusion criteria were defined among the authors before the study.

Studies were considered for inclusion if they were classifiable as accuracy diagnostic test studies and their declared objectives were to differentiate individuals with clinically diagnosed pneumonia from controls or other diseases (e.g., bronchitis). Studies only focusing on determining the severity of pneumonia or its aetiology were considered out of the scope of this review to reduce heterogeneity among studies.

Data collection could be both prospective and retrospective i.e., planned either before, or after the reference tests were performed. All types of study methods to recruit participants were allowed, including studies using a single set of inclusion criteria for patients with and without the target condition (Cohort type accuracy studies) and studies using different set of criteria (Case-control type accuracy studies). Prognostic accuracy studies, such as those using AI systems to identify patients who may develop pneumonia in the future, or experience pneumonia-related adverse events were excluded from the present review.

The target condition had to be defined as pneumonia, without any limitation regarding its pathogenesis (e.g., viral, including SARS-CoV related pneumonia or bacterial pneumonia), or the classification system used (e.g., the WHO IMCI classification). Study participants of all ages and clinical characteristics were admitted. Studies not on human subjects were excluded.

The characteristics of the index test evaluated in the studies were required to (i) use algorithms, defined as machine learning in the study, including appropriate apparatus, such as learning or training, aimed at seeking optimal answers and (ii) include signs and symptoms as predictors in the machine learning algorithms.

Signs and symptoms were defined as any subjective (symptoms) or objective (signs) abnormality that may indicate the presence of pneumonia, such as cough, fever, dyspnoea, chest pain, chest, indrawing, sweating and shivering, breathing rate, etc. No limitations were set concerning other predictors that might have been used alongside signs and symptoms, including epidemiological and demographic parameters (e.g., age, gender, rural/urban site, season, region, etc.), imaging or laboratory test results.

Studies had to report at least one accuracy measure of the index texts, such as sensitivity, specificity, accuracy and the area under the curve (AUC). Lastly, no pre-defined limitations were applied to the type of test used as reference standard (e.g., imaging examinations, microbiological tests), and the spectrum of study participants with and without the target condition. Review articles were not included directly, but references were screened and included individually if they met the review's inclusion criteria.

In summary, the inclusion criteria were as follows:

1. studies classifiable as accuracy diagnostic test studies;
2. the objective of the study was to differentiate individuals with clinically diagnosed pneumonia from controls or other diseases (e.g., bronchitis);
3. algorithms, suggested as machine learning in the study, including appropriate apparatus, such as learning or training, aimed at seeking optimal answers;
4. studies had to report at least one accuracy measure of the index texts, such as sensitivity, specificity, accuracy and the area under the curve;
5. studies had to include signs and symptoms as predictors in the machine learning algorithms.

The following exclusion criteria were also applied:

1. all review articles, letters, comments, abstracts, conference papers and case reports;
2. studies only focusing on determining the severity of pneumonia or its aetiology and/or without diagnostic confirmation;
3. prognostic accuracy studies;
4. non-human subjects (e.g., animals);
5. non-English papers.

2.3. Data extraction and outcomes of interest

Two review authors (KS, CF) performed the title and abstract screening and extracted the data from included studies and a third author (RC) checked the extracted data. For the final set of included records, the following information was retrieved: (i) literature data – title, first author and publication date; (ii) study design; (iii) study participants – kind of pneumonia, mean age and class, clinical setting (e.g., primary, secondary or tertiary care), sample size, sign and symptoms and other diseases; (iv) information regarding the reference standard, i.e., methodologies to distinguish pneumonia patients from control group or other respiratory diseases. In addition, data were also recorded on (v) the specific methodologies used to process and classify data for use in machine learning algorithms, including features selection methods, ML parameters and final predictors. Finally, data were also extracted on (vi) the summary measures for the predictive ability of the identified AI systems, including the systems' sensitivity, specificity, accuracy and AUC measures. In addition, references to relevant ethical issues regarding the studies were also extracted. This study was not meant to estimate an overall measure of the accuracy of ML systems to diagnose pneumonia based on symptoms, but rather to provide a broad overview of the characteristics of the different approaches proposed. Therefore, only a qualitative synthesis of the study results was planned, anticipating a broad heterogeneity in the types of study design, participants, test methods, type of analysis and reported accuracy measures in the included studies.

2.4. Certainty assessment

The reporting study quality of the included studies was rated via the STARD 2015 tool [30], which consists of a checklist of 30 items that should be included in the reports of diagnostic accuracy studies in order to ensure the interpretability of results, enhance the reproducibility of research and improve completeness and transparency.

Given the characteristics of ML tests, items 22 and 25 were considered not applicable and excluded from the quality assessment. When assessing adherence to the STARD 2015 checklist, each reporting requirement was rated as yes, no, maybe, or not applicable, with all disagreements resolved by consensus between the 2 reviewers. If, for each item, information was fully reported in the relevant section of the manuscript or provided in the supplementary material (including online-only material), the item was scored as a "yes". If an item was only reported partially, it was scored as a "maybe", whereas if an item was not applicable to the study was scored as NA. To optimize interobserver agreement, a training session was done for all reviewers using 2 articles.

Three reviewers (KS, CF and RC) completed the study checklist for one third of the included records each. A cross-check by another author

was done for 10% of the studies and any disagreement was resolved by discussion. Results of the quality assessment were analysed qualitatively through a narrative summary of the main reporting issues identified in the studies.

3. Results

3.1. Study selection

According to the search strategy described above, 876 titles were identified in PubMed, Scopus, Ovid[SP], and Pre-print servers. After removing duplicates, 775 titles were considered. Of these, 726 were excluded after reading the title and abstracts as they did not meet the inclusion criteria. From the remaining 49 full-text articles, 34 were removed due to the exclusion criteria. One article was identified through a linear search of the references included in the final studies. Finally, 16 full texts were included in the qualitative analysis. A flow chart of the literature search results is shown in Fig. 1.

3.2. Certainty assessment of included studies

A summary of STARD 2015 adherence by item is presented in Fig. 2. The STARD items reported for each study is listed in the Supplementary Table 2. The STARD items are grouped in macro-categories such as: title or abstract, intro, methods, results, discussion and other info. Each item is coloured in green if information was fully reported in the study ("Yes"); light blues, if an item was only reported partially in the study ("Maybe"); red, if information was not reported ("No"); whereas if an item was not applicable to the study was coloured in grey ("NA").

Overall studies had a moderate reporting quality for all subitems in the sections of the STARD tool concerning the title and abstract, the description of the study design and participants, and discussions, but less so in the sections of the methods concerning the description of the test methods, including the index and reference tests; the analysis of the data; and the results sections including the description of the study participants and the results of the tests.

STARD items were described as frequently reported (if ≥66% of the total studies reported a specific item), moderately reported (33%-66% of the total studies reported a specific item), and infrequently reported (≤33% of the total studies reported a specific item) [34].

Seventeen of the 28 items were frequently reported in whole or in part (i.e., "Yes" or "Maybe") by the included studies. Some of the frequently reported items are of relevance to this study. In the method section related to the test methods, subitem 10.a, which relates to the description of the machine learning method used in the study (i.e., the index test) was fully reported by 11 studies (69%), partially by 1 study (6%), whereas no sufficient information was provided in 4 studies (25%). Similarly, subitem 10.b, related to the description of the reference standard used to calculate the accuracy of the index test was reported fully by 9 studies (56%). Moreover, in the results section, item 24, related to the estimates of diagnostic accuracy and their precision (such as 95% confidence intervals), was reported by 8 studies (50%).

Six of the 28 items were moderately reported, in whole or in part by the included studies. These include for example item 11 in the method macro-area (i.e., rationale for choosing the reference standard) was reported in full by 7 studies (43%). Another important item is 12 (i.e., definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory), which was only reported in full or partially by 6 studies (37%). In the results macro-area, item 20, which regards baseline demographic and clinical characteristics of participants, was reported by 10 studies (62%).

Five of the 28 items were infrequently reported, in whole or in part

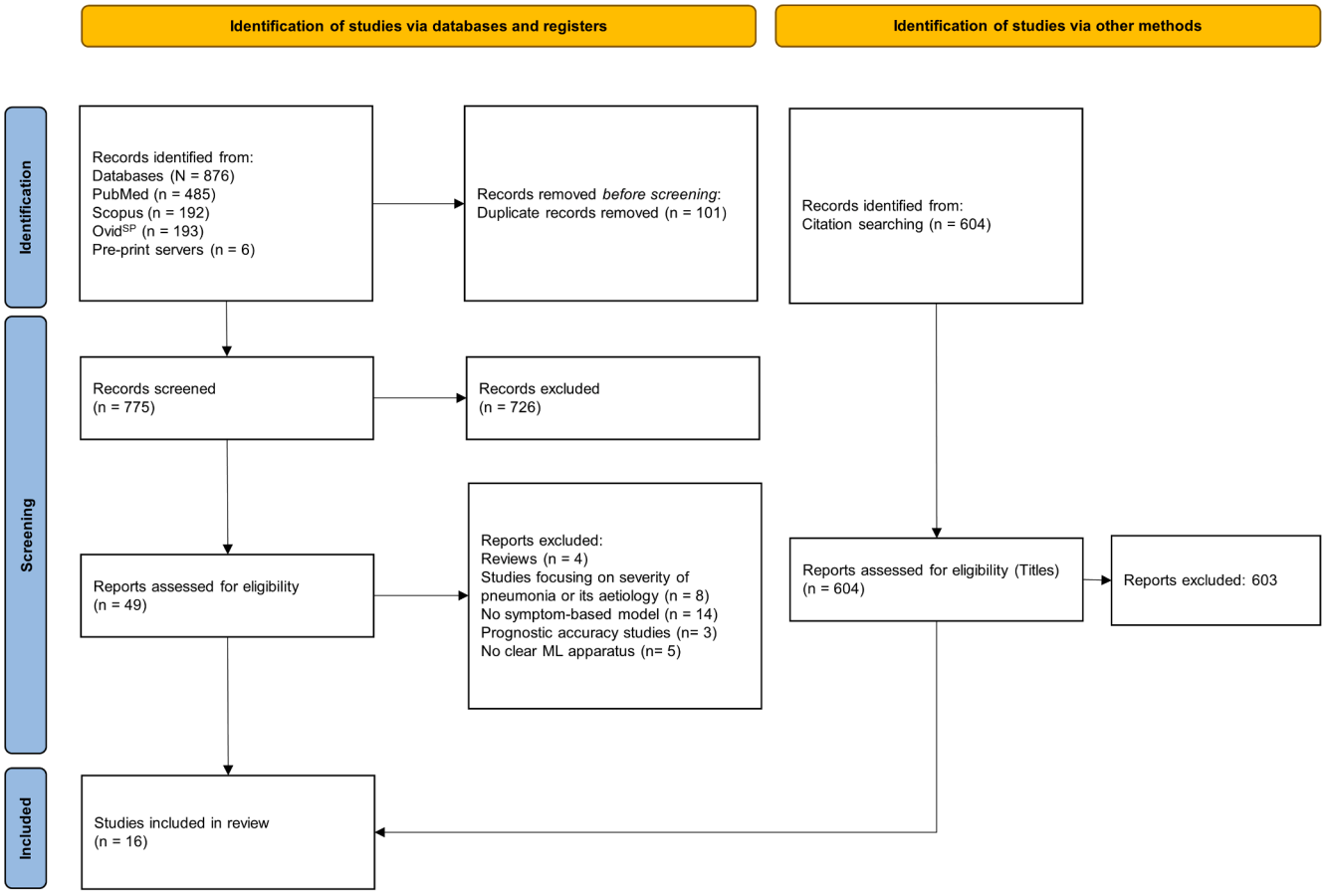


Fig. 1. PRISMA search workflow.

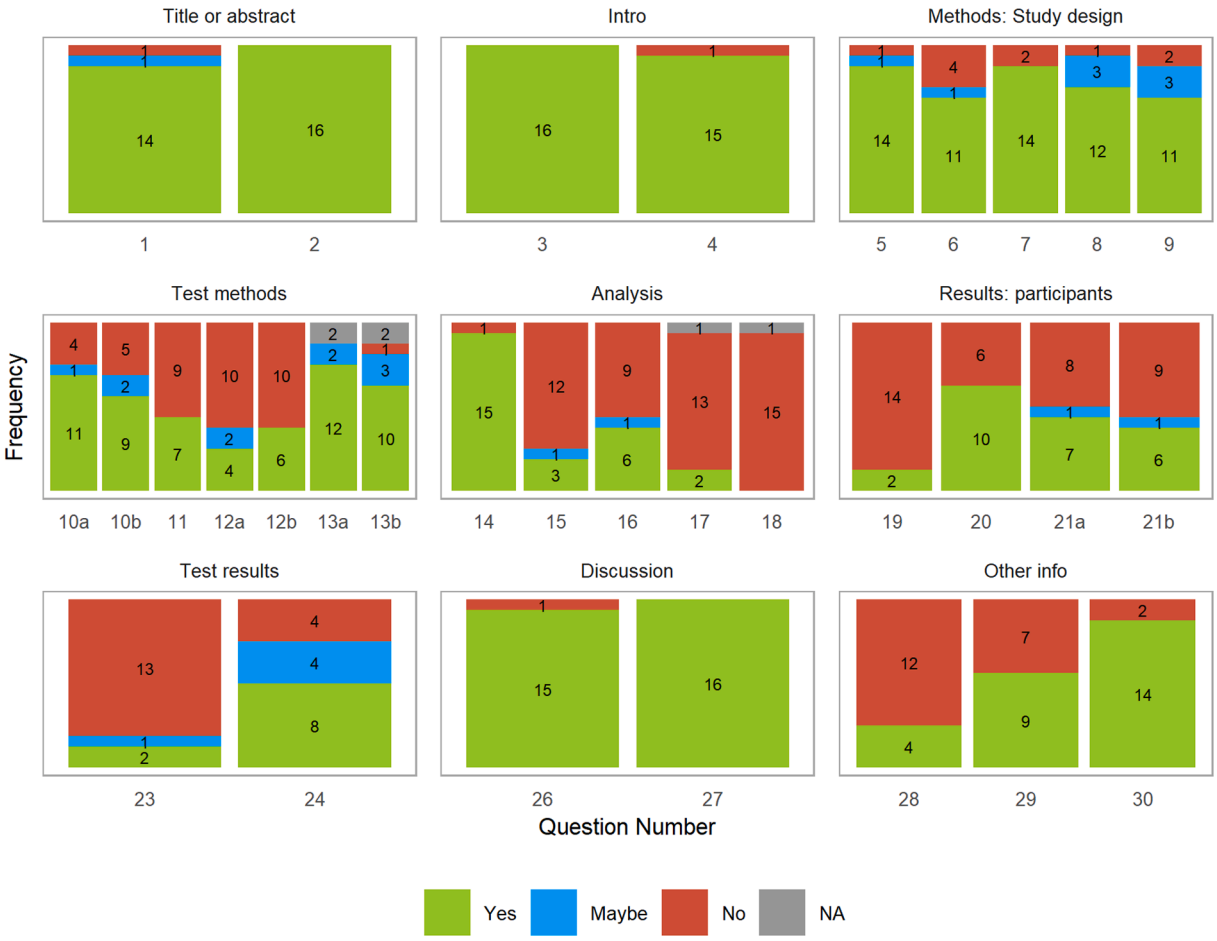


Fig. 2. Frequency of Standards for Reporting of Diagnostic Accuracy 2015 Items. "Yes" was assigned if information was fully reported; "Maybe" was assigned if an item was only reported partially; an item was scored as "No" if information was not reported; whereas if an item was not applicable to the study was scored as "NA".

by the included studies. These include item 15 (i.e., how indeterminate tests were handled; reported by 3 studies (18%)), item 17 (i.e., whether analyses of subgroups and heterogeneity were prespecified or exploratory; reported by 2 studies (12%)), item 18 (i.e., whether intended sample size and how it was determined; reported by none of the studies (0%)), item 19 (i.e., flow of participants, using a diagram; reported by only two studies (12%)) and item 23 (i.e., cross tabulation of the index test results; fully reported by 1 study (6%)).

3.3. *Characteristics of the included studies*

Study characteristics regarding study design and subject population as well as machine learning methods and performance measures were extracted and presented in Tables 1 and 2.

3.3.1. *General Study Characteristics*

A summary of types of pneumonia, reference standards, study populations and other clinical characteristics of the included papers is given in Table 1.

The vast majority of studies identified concerned CAP, with patients presenting symptoms of pneumonia to EDs or other healthcare facilities. In one case (Rother et al. [41] it is unclear whether HAP or CAP is considered. Interestingly, one study by Porter et al. [46] included pneumonia identified through presentation to the ED, inpatient wards and ambulatory care units, suggesting an inclusion of both HAP and CAP. Two studies focused specifically on the detection of COVID-19 pneumonia.

Unsurprisingly, radiography, which is widely considered the gold standard for confirmation of pneumonia, was the most used reference test. In several papers the reference standard was unclear [36–37,41,48–49]. Grigull et al. [36], Yu et al. [48] and Huang et al. [49] gave no mention of how diagnosis of pneumonia was performed, Bejan et al. [37] described diagnosis as being performed by a 'nurse with 6 years of experience' but not the criteria used for classification, Rother et al. [41] mentioned 'standard diagnostic criteria' but offered no further detail. Further, as well as omitting the details of how pneumonia positive cases had been established, Yu et al. [48] and Huang et al. [49] did not make clear how cases of COVID-19 and pneumonia had been confirmed, i.e., whether PCR test results were available.

14 out of 16 studies provided information on study population age. Of these, 8 focused on childhood pneumonia, 5 on pneumonia in adults and one in a mixed age population. Three of the included studies were focused on LMIC settings and the specific challenges regarding diagnosis of childhood pneumonia in such areas [23,42,44], indicative of those most vulnerable to the disease. Of these studies, only Pervaiz et al. [44] included other respiratory diseases which may be difficult to distinguish

Table 1
Information extracted from final selection of papers describing the type of pneumonia considered, population characteristics and study design.

Author, year	Type of study	Reference standard, with criteria (if given)	Population type: Age (average if given), setting	Other diseases	Number of patients	Availability of Data
Steurer et al., 2011 [35]	Prospective cohort study	By set of symptoms and signs associated with radiographic shadowing with 'no other explanation'	Adults (46.7 ± 16.3 years)	Chronic bronchitis	Total: 621. Pneumonia: 127. No pneumonia: 494.	Private dataset (Hospital).
Grigull et al., 2012 [36]	Retrospective case-control study	Diagnosis by clinicians, cross-checked with medical definitions	Children (6.5 ± 2.5 years). ED	Many common diagnoses from ED. Relevant to analysis: Asthma, Bronchitis	Total: 692. Pneumonia: 54.	Private dataset (Hospital).
Bejan et al., 2012 [37]	Retrospective study	Classified by a research study nurse with 6 years of experience. **Positive** if the patient had pneumonia within the first 48 h of ICU admission **Negative** if the patient did not have pneumonia or the pneumonia was detected after the first 48 h of ICU admission	No age specification	Not specified	Total: 426. Pneumonia:66.	Private dataset (Hospital).
DeLisle et al., 2013 [38]	Retrospective study	**'Possible pneumonia'**: non-negative chest imaging report and one or more identified symptom **'Pneumonia-in-plan'**: Non-negative chest imaging report and pneumonia listed as first or second diagnostic possibility by clinician	Adults (61 ± 15 years)	Not specified	Total: 2747. 'Possible pneumonia': 370. 'Pneumonia-in-plan': 250.	Private dataset (Hospital).
Haug et al., 2013 [39]	Retrospective study	Primary discharge diagnosis of pneumonia: Defined by ICD-9 codes	No age specification. ED	Not specified	Total: 48449. Pneumonia: 2413.	Private dataset (Hospital). Available from the corresponding author on reasonable request.
van Vugt et al., 2013 [40]	Retrospective study	Radiographically confirmed clinical pneumonia	Adult (50 years)	Pulmonary Cardiac, Diabetes mellitus.	Total: 2820. Pneumonia: 140.	Private dataset (Hospital).
Rother et al., 2015 [41]	Prospective monocentre study	The diagnosis was confirmed by a paediatric pulmonologist using standard diagnostic criteria.	Children	Not specified	Total: 170. Pneumonia: 21.	Data available in supplementary materials at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4534438/.
Naydenova et al., 2016 [23]	Case-control study	Diagnosed by clinician supported by chest X-rays. Expanded WHO and IMCI guidelines: **Very severe** Central cyanosis/ not able to drink **Severe** Lower chest wall in drawing **non-severe** Fast breaths (>50 breaths per min for 2–11 m, >40 for 12–59 m) **No pneumonia** None of above	Children (2–59 months). Gambia	Not specified	Total: 1581. Pneumonia: 780.	The data supporting this article have been registered on Oxford University Research Archive (ORA) and can be accessed via http://dx.https://doi.org/10.5287/bodleian:ht24wj41z .
Nuzhat et al., 2017 [42]	Unmatched case-control study	Radiographically confirmed clinical pneumonia WHO radiological criteria	Children (0–59 months). Dhaka Hospital	Diarrhoea, Acidosis.	Total: 713. Pneumonia: 267.	Private dataset (Hospital). Available from the corresponding author on reasonable request.
De Santis et al., 2017 [43]	Prospective observational study	Radiographically confirmed clinical pneumonia	Children (2 months-10 years)	Not specified	Total: 1005. Pneumonia: 31.	The database is available from the Zenodo repository (https://zenodo.org/record/166713#.WCr_WU2QyMp).
Pervaiz et al., 2018 [44]	Retrospective	Radiographically confirmed clinical pneumonia. Defined by lobar consolidation (with or without pleural effusion) .	Children (21.3 ± 16.2 months). LMIC Inner-city	Asthma, Bronchitis, Upper respiratory tract infection.	Total: 832. Pneumonia:221.	Private dataset (Hospital).
Groeneveld et al., 2019 [45]	Prospective observational cohort study	Radiographically confirmed clinical pneumonia Defined by consolidation on X-Ray	Adult (56 years)	Comorbidity. No further detail.	Total: 249. Pneumonia: 30.	Private dataset (Hospital). Available from the corresponding author on reasonable request.
Porter et al., 2019 [46]	Prospective multicentre study	At least one feature from both of the following categories:1. History of: (i) fever in prior 48 h	Children (29 days to 12 years)	Not specified	Total: 585. Pneumonia:87.	Private dataset (Hospital). Available from the corresponding author on reasonable request.

(*continued on next page*)

Table 1 (*continued*)

Author, year	Type of study	Reference standard, with criteria (if given)	Population type: Age (average if given), setting	Other diseases	Number of patients	Availability of Data
		or fever at the time of examination, (ii) cough, (iii) dyspnoea, or (iv) chestpain2. Either focalbexamination findings including crackles, bronchial breath sounds, focal decreased breath sounds; ORA chest radiograph with new consolidation with normal auscultation findings				
Feng et al., 2020 [47]	Retrospective study	**Suspected COVID-19 pneumonia**: - Epidemiological history - CT imaging characteristics of viral pneumonia -Any one of: Fever &/or respiratory symptoms, Total leukocyte count normal or decreased or lymphopenia **Confirmed COVID-19 pneumonia**: Positive test result from throat swab of upper respiratory tract	Adult (34 years, IQR: 29 to 42). ED	Hypertension, Diabetes, Cardiovascular disease, Chronic obstructive pulmonary disease, Malignancy, Chronic kidney disease, Chronic liver disease.	Total: 164. Confirmed COVID-19 pneumonia: 7.	Private dataset (Hospital). Available from the corresponding author on reasonable request.
Yu et al., 2021 [48]	Retrospective study	**Based on 'discharge diagnoses'**	Children (1 day – 18 years, mean age 3.17). Respiratory department of Children's Hospital.	Upper respiratory tract infection, Asthma, Bronchitis.	Total: 14697. Pneumonia: 42.3% (6217) . Some overlap in diseases diagnosed per patient.	Private dataset (Hospital). Available from the authors upon request with Hospital's permission.
Huang et al., 2021 [49]	Retrospective study	**'Confirmed COVID-19' pneumonia**	Children and adults (8–84 years). General People's Hospital.	Not specified	Total: 416. Pneumonia: 209.	The COVID-CT dataset is available at https://github.com/UCSD-AI4H/COVID-CT

ED: Emergency Department.
WHO: World Health Organization.
CT: Chest computerized tomography (radiographic technique).

and indeed are likely to be highly common in presenting patients. Indeed, only 6 studies concerned differential diagnosis of pneumonia from other diseases while the remainder either did not specify or specifically excluded other respiratory conditions from the population. Inability to distinguish between other, similarly presenting respiratory diseases is a clear limitation to the utility of any proposed diagnostic tool.

Finally, in 11 of 16 studies the data used was not provided in the paper nor registered in any opensource database, although in 6 of these studies availability was granted on request to the authors. In 3 cases data was made available in public repositories.

3.3.2. *Artificial intelligence study characteristics*

Summaries of feature selection, machine learning methods, validation process and their performance to automatically detect pneumonia are presented in Table 2, for more details see Supplementary Table 3. The best ML method was chosen as the one presenting a higher AUC, which is a good estimator of sensitivity and specificity. In case a study explored different problems connected to pneumonia such as classification of different grades of severity (e.g., high, moderate and mild), only the methods employed by the included studies to automatically detect pneumonia from controls or other respiratory disease were extracted and tabulated in Table 2.

ML model choice varied from relatively simple methods such as logistic regression (31% of the selected papers) [50] to deep learning algorithms such as artificial Neural Networks (aNNs) [51] or Convolutional Neural Networks (CNNs) [52] (19% of the selected papers) as also shown in Fig. 3.

Logistic regression is a simple technique for binary classification problems, and it is often used to model the probability of a certain class or event existing [50]. Whereas aNNs and CNNs are more advanced machine learning models, which are capable of learning any nonlinear function; in particular CNNs is a type of aNN mainly used in image recognition and processing that is specifically designed to process pixel data [52].

Several studies [40,42,44–45,47] (5 out of 16 studies) selected regression-based models as the method achieving the best overall performance. The regression-based models ranged from simple or multivariate logistic regression [53] to more sophisticated techniques such as LASSO regression [54]. In particular, LASSO is a penalized regression approach that estimates the regression coefficients by maximizing the log-likelihood function (or the sum of squared residuals) and automatically deletes unnecessary covariates.

Five [23,38–39,43,55] out of 16 studies employed a tree-based model to automatically detect pneumonia. The majority of tree-based algorithms were Random Forest (RF) [56] and CART trees [57]. CART algorithm is a decision tree based on Gini's impurity index as splitting criterion. It is a binary classifier built by splitting single nodes into child nodes repeatedly. On the other hand, RF is a bootstrapping algorithm based on the CART tree model. In particular, RF creates multiple CART trees based on "bootstrapped" samples of data and then combines the predictions. The combination is an average of all CART models predictions. Random Forest can achieve better predictive power than a CART model but, RF rules are not easily interpretable.

Table 2
Information extracted from final selected papers describing the feature selection, ML methods and model performance (in the case of multiple ML methods, best performing metrics are given).

Author, year	Feature selection methods	Model development (training, validation and/testing)	Classes (in bold target class)	Final predictors	AI methods (Best performing in bold if specified)	Performance estimates
Steurer et al., 2011 [35]	Manual pre-selection for: Ease of availability Reliability	Leave-one-out cross validation	Binary: patient with pneumonia and without pneumonia	Chronic cough Daily fever Dyspnoea Respiratory rate Pleural friction rub CRP	CART	**Sensitivity:** - **Specificity**: - **Accuracy**: - **AUC:** 90% (95% CI: 87% to 93%).
Grigull et al., 2012 [36]	Manual: identified based on parameters frequently investigated/measured in children in ED who could not be instantly diagnosed	Bootstrap validation and testing on an independent subset of data	Binary: pneumonia and other diseases	14 clinical factors and vital signs including: age, temperature, blood pressure, etc. 12 laboratory parameters including: haemoglobin, leukocyte count, CRP level, etc.	Combination of SVM (parameters set based on data) aNNs (14,400 numeric weights distributed throughout three layers. The input layer included 100 parallel neurons, with each neuron gathering the 26 input signals) fuzzy logics **voting algorithm**	**Sensitivity:** 95% **Specificity**: 92% **Accuracy**: - **AUC:** 99%
Bejan et al., 2012 [37]	Considered as features all possible uni-grams and bi-grams of words and UMLS concepts Ranked variables based on association between feature and category using $\chi 2$ and t statistics	Fivefold cross-validation	Binary: pneumonia versus non pneumonia	Words and concepts (not specified)	SVM	**Sensitivity:** - **Specificity**: 98% **Accuracy**: 86% **AUC:** -
DeLisle et al., 2013 [38]	Not specified	Not clear	Binary: patient with pneumonia and without pneumonia	Pneumonia ICD-9 Code Text of clinical notes Imaging obtained Text of imaging reports	Random fields probabilistic classifier	**Sensitivity:** 58–75% PPV: 20–70% **Specificity**: - **Accuracy**: - **AUC:** -
Haug et al., 2013 [39]	**Two methods (both AUC given):** Fully automated process, based on highest $\chi 2$ value Manual by clinicians	10-fold cross-validation	Binary: positive pneumonia cases and negative pneumonia cases	3 vital signs including temperature, heart rate, respiratory rate 7 laboratory parameters including anion gap, BUN Chloride, Spo2, etc Chest x-ray results Nursing assessment and symptoms including abdominal exam, abnormal breath sounds, pleuritic pain, breath sounds, strong cough, etc.	Bayesian network classifier (tree-augmented naïve Bayes (TAN)5) (Estimation of Bayesian network parameters using expectation maximization)	**Sensitivity:** - **Specificity**: - **Accuracy**: - **Method 1:** **AUC:** 94% (95% CI 94,2% to 0.94,7%) **Method 2:** **AUC:** 92% (95% CI 91,6% to 92,4 %)
van Vugt et al., 2013 [40]	Backward and forward selection	Bootstrapping for internal validation	Binary: patients with pneumonia and without pneumonia	Absence of runny nose Breathlessness Crackles Diminished breath sounds on auscultation tachycardia fever CRP	Multilevel LR	**Sensitivity:** - **Specificity**: - **Accuracy**: - **AUC:** 77% (95% CI 73% to 81%)
Rother et al., 2015 [41]	Not specified	Ten-fold stratified cross validation	Multiclass (pneumonia versus other diseases)	Six questions to evaluate disease history and symptoms such as whistling/ wheezing sound, drowsy, etc	Program consisting of eight classifiers: SVM ANN fuzzy rule-based random forest LR linear discriminant analysis naïve Bayes nearest neighbour ensemble	**Sensitivity:** 90% **Specificity**: - **Accuracy**: - **AUC:** -
Naydenova et al., 2016 [23]	Features must be measurable in point of care setting Features selected if they appear in top 10 of at least three techniques of both: Maximum relevance Majority voting	Fivefold cross-validation and testing on independent folder	Binary: pneumonia and matched-age control group	Respiratory rate Heart rate Temperature Oxygen saturation Age	**RF** (750 decision trees, searching over 2 variables at each tree node)	**Sensitivity:** 98.2% (95% CI 97.9 – 98.8%) **Specificity**: 97.6% (95% CI 97.1 – 98.0%) **Accuracy**: 95.9% (95% CI

(*continued on next page*)

Table 2 (*continued*)

Author, year	Feature selection methods	Model development (training, validation and/testing)	Classes (in bold target class)	Final predictors	AI methods (Best performing in bold if specified)	Performance estimates
						95.3 – 96.5%) **AUC:** 99.7% (95% CI 99.3 – 99.8%)
Nuzhat et al., 2017 [42]	Two methods Odds ratio Backward stepwise LR (controlled for covariates)	No internal or external validation	Binary: pneumonia and unmatched-control group	Cough and lower chest wall in drawing (combined)	LR	**Sensitivity:** 94% (95% CI 89& to 97%) **Specificity**: 99% (95% CI 97% to 100%) **Accuracy**: - **AUC:** -
De Santis et al., 2017 [43]	Multivariate analyses	No internal or external validation	Binary: pneumonia versus other diseases	For radiological pneumonia: abnormal chest auscultation For acute HHV6 infection: Dehydration For bacterial disease (any) : Chest in drawing For viral disease (any) : Jaundice	CART	**Sensitivity:** 38% **Specificity**: 97% **Accuracy**: - **AUC:** -
Pervaiz et al., 2018 [44]	None: Features set based on WHO criteria	No internal or external validation	Binary: pneumonia and acute respiratory illnesses group	WHO pneumonia predictors	**LR**	**Sensitivity:** 66% (95% CI, 59%-73%) **Specificity**: 53% (95% CI, 49%-57%) **Accuracy**: - **AUC:** 62%; 95% CI, 0.58–0.67
Groeneveld et al., 2019 [45]	Univariate analysis of clinical risk factors Multivariate analysis of signs and symptoms and variables/ biomarkers.	No internal or external validation	Binary: Patients with pneumonia and without pneumonia	Runny nose absent Feel ill CRP > 30 mg/l	**LR** Runny nose absent, B = 1.230 Feel ill, B = 2.378 CRP > 30 mg/l B = 1.572 Intercept: −4.797	**Sensitivity:** - **Specificity**: - **Accuracy**: - **AUC:** 75% (95% CI 65% to 85%)
Porter et al., 2019 [46]	Not specified	Leave-one-out cross-validation	Binary: pneumonia versus other diseases	Not specified	SoftMax neural network	PPA 87%, NPA 85%
Feng et al., 2020 [47]	Candidate features based on expert opinion and availability in medical records LASSO	Testing on an independent subset of data	Binary: Covid19 pneumonia versus suspected patients	7 laboratory parameters including basophil count, platelet count, interleukin-6, etc. 7 symptoms including: 5 clinical factors and vital signs including age, heart rate, etc., fever, shiver, shortness of breath, etc.	LR (LASSO)	**Sensitivity:** 100% **Specificity**: 78% **Accuracy**: - **AUC:** 93%
Yu et al., 2021 [48]	adaptive feature infusion	internal validation and testing on independent subset of data.	Binary: pneumonia versus other diseases	unstructured clinical notes including chief complaints, physical examinations, and clinical test results	Deep learning with adaptive feature infusion module.	**Sensitivity:** - **Specificity**: - **Accuracy**: - **AUC:** 87.8%
Huang et al., 2021 [49]	Not specified	testing on independent subset of data.	Binary: Covid19 pneumonia versus healthy patients	CT image information 11 symptoms including fever, cough, muscle ache, fatigue, headache, nausea, diarrhoea, stomach-ache and dyspnoea	Deep learning FaNet	**Sensitivity:** - **Specificity**: - **Accuracy**: 98.28% **AUC:** -

CI: Confidence Interval, LR: Logistic Regression, SVM: Support Vector Machine, CART: Classification and Regression Tree, aNN: artificial Neural Network, LASSO: Least Absolute Shrinkage and Selection Operation, CT: Computed Tomography imaging technique.

Two studies investigated the combination of several ML methods via voting [36,41]. Voting is one of the easiest ensemble methods. In particular, ensemble methods are techniques that create multiple models and then combine them to produce improved results [58]. Grigull et al., [36] and Rother et al., [41] showed that the combination of different ML methods such as SVM, aNNs, fuzzy logics and more traditional ML (RF, LR, etc.) achieved higher accuracy to discriminate pneumonia versus other diseases.

Two studies [38–39] employed probabilistic ML methods to detect patients with pneumonia and without pneumonia. DeLisle et al. [38] reproduced a previously reported model presented in [59], whereas Haug et al. [39] developed Bayesian networks, built around directed

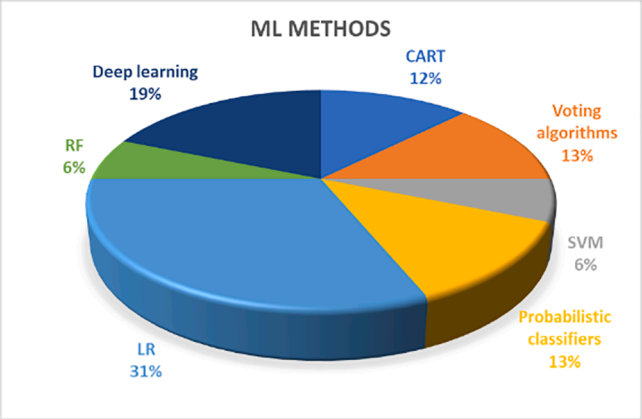


Fig. 3. ML methods Information extracted from final selected papers. SVM: Support Vector Machine; LR: Logistic Regression; RF: Random Forest.

links reflecting mathematical relationships between variables.

Support vector machine (SVM) was frequently employed in several studies but only one study [37] reported SVM as the best performing model. SVM belongs to a general field of kernel-based machine learning methods and are used to efficiently classify both linearly and non-linearly separable data [60].

Only one study achieved the best performance using a deep learning algorithm (i.e., artificial neural networks) compared to other traditional machine learning methods [46]. Two recent studies [48–49] also employed deep learning methods achieving high accuracy and AUC value. Yu et al. [48] presented a novel deep learning algorithm for the disease identification stage, including adaptive feature infusion and multi-modal attentive fusion in order to fuse structured and text data together. Huang et al. [49] explored a deep learning based dual-tasks network, named FaNet [61], performing both diagnosis and severity assessments for COVID-19 based on the combination of CT imaging and clinical symptoms.

The reason for the use of mostly linear classifiers may be due to the fact that those models are mainly developed to be implemented in a Decision support system or CAD system. In such systems less-complex, more interpretable models to clinicians and non-AI experts are preferred over more advanced AI methods such as deep learning, which are often referred as “black boxes” along with more complex ML algorithms such as SVM.

Among the selected studies, two studies [47,49] investigated patients with suspected COVID-19 pneumonia. Feng et al. [47] employed a Logistic regression (LASSO) method to discriminate among COVID-19 pneumonia and healthy patients using combinations of symptoms and laboratory parameters. Whereas, Huang et al. [49] detected COVID-19 pneumonia patients from healthy patients using a state-of-the-art deep learning algorithm (FaNet) by using CT images and symptoms.

3.3.3. *Ethical aspects*

Several ethical issues were addressed by the included studies. Some of them were not fully investigated, such as informed consent, reference to minors or, generally, to the age of patients (considering that that the youngest and the oldest are the most affected by pneumonia) or to gender issue [36,41–44,46,48].

Aspects that were frequently mentioned were related to the allocation of resources [23,43,62], in particular for limited resource settings (LRSs) [42–44,48], in which morbidity or infant mortality for pneumonia [42] is high, the lack of resource [42,48], but also the need for specific training of staff [44], doctors and/or nurses [42,48] on more advanced tools. This raises an important ethical question that is a global challenge [47]: the difficulty of LRSs in complying with medical and technological international standards.

Another recurring theme was the man–machine relationship [36,39], still controversial from an ethical point of view. There is unanimous recognition that technology is a support tool for doctors [36,39,41,43] that adds objectivity, precision [44–45,48–49] and speed to the diagnosis. However, the entry of technology in the moment of diagnosis somehow changes the doctor-patient relationship, objectifying the care relationship [36–37,40], which on one hand leads to its depersonalization and on the other to greater rigor (for sensitivity and specificity) which could limit medical malpractice and the consequent litigation [36]. The question of a possible replacement of man by machine for the doctor decision making is also mentioned in [23,36,39], but it is made clear that doctors themselves are aware of the urgent clinical need for algorithms and that they do not perceive this as competition [41,45].

In addition, some ethical problems relating to data management come forward, in particular about collection, anonymization/deidentification and sharing of data [36,38,42,62]. Those issues were addressed transversally with reference to some ethical principles involved, not only in data management, but more generally in the use of artificial intelligence in the field of healthcare. Principles mostly addressed were: affordability [23], accessibility [44,46,49], accuracy [23,39–41,46,48], appropriateness [46], timing and efficiency [23,37–38,42,46,49,62] and reproducibility [23,38]. In general, there is an overall need for ethical guidelines and references (including ethics committees [41,46] that can be a guarantee for the approach to technology.

4. Discussion

This systematic literature review provided a comprehensive overview of the existing studies which proposed ML algorithms to diagnose pneumonia based on signs and symptoms. The use of AI for image-based detection of pneumonia and particularly COVID-19 has been reviewed [63–64], but to the best of our knowledge systematic review of symptom-based models is lacking in the existing literature. This is particularly timely as AI based diagnostic tools begin to appear in medical devices. However, the practicality of AI in current medical practice is still not fully understood by clinicians. AI could help to reduce mistaken diagnosis. In fact, respiratory diseases can present overlapping symptoms which may offer similar clinical presentation but have differing underlying causes and respond best to different treatments. Therefore, the advances made in machine learning models could assist clinicians in diagnosing pneumonia in rapid time by considering a high number of variables related to patient care and medical history. To address the difficulties effectively and efficiently, it may be worth considering the inclusion of AI in medical practice. This could positively contribute to the patients’ condition by analysing treatment personalization strategies as a result of predicting clinical situations that could deteriorate patients’ health. With the dramatically fast spread of COVID-19, analysing complex medical datasets based on machine learning can provide opportunities for developing a simple and efficient COVID-19 diagnostic system. Nevertheless, several issues, such as poor realised performance in clinical settings, as discussed by van Schalkwyk et al. [65], may be alleviated by proper ethical, contextual and performance evlaution during their conception and design.

Of the hits retrieved in the systematic review, many studies were published from 2020 to 2021 and concerned detection of COVID-19. However, the majority of these articles did not meet the inclusion criteria as they either focused on symptomatic detection of early disease (not associated with pneumonia) or imaging-based detection with no input from symptoms or signs. The included studies were highly heterogeneous concerning the study design, the healthcare setting, the study population and the ML algorithm employed. Specifically, three papers focused on diagnosing childhood pneumonia in LMIC settings. This is a very relevant context which warrants more research, as application of AI algorithms in countries with highly constrained healthcare settings and deprived populations may be of even higher

value compared to higher income countries.

The reporting quality was satisfactory for some sections of the STARD checklist, but less so for relevant sections such as the description of the index (ML model) and reference tests and the analysis of the data in the methods section as well as the description of the study participants and the results of the tests in the results section.

For example, items which were less frequently reported included the details of the reference and index tests, such as a clear description of the reference standard used as benchmark, the definition of rationale for test positivity cut-offs or result categories, or the way that indeterminate results of the reference test were handled. Noteworthy and concerning was the fact that such details remained absent even in the most recent publications, which focused on providing improved detection of COVID-19 pneumonia, with only one study providing any details on either diagnosis of pneumonia or method for confirmation of SARS-CoV-2 infection. In addition, the characteristics of the study participants, such as the distribution of severity of disease in those with the target condition, or the distribution of alternative diagnoses in those without the target condition were also less frequently reported. All these aspects warrant more careful consideration and higher reporting standards to allow a clear judgment on the risk of bias in the accuracy estimates and to allow replication and validation of the proposed ML-algorithm in other settings or populations. Similar issues and deviations from best practice have been highlighted concerning the relative explosion in the publication of ML algorithms for diagnosis and management in response to the spread of COVID-19, the result of which clouds the most clinically beneficial routes and prevents the realisation of benefits to patients [66].

The references to ethics found in the selected texts suggest that there is an overall awareness of the importance of ethical principles and guidelines to guarantee the protection of people's rights. However, the urgency of adopting a shared ethical reference framework emerges (i.e., European Commission, *Ethics guidelines for trustworthy AI*). Furthermore, in order to make ethics a real tool of concrete support and not just a humanitarian embellishment, it should be considered a decisive reference also in the design and implementation phase of AI algorithms, to better guarantee users' rights.

Concerning the ML algorithms, there was a huge heterogeneity among studies and many pitfalls were identified in the development of a reliable and generalisable ML model to diagnosis pneumonia via symptoms and signs. Few studies employed a feature selection step in the development of the ML model. Nevertheless, feature selection is a critical step to develop a robust classifier in medical and health applications. In fact, in order to minimize the over-fitting risk in a ML model, the number of features used in the model and its cardinality should be limited by the number of subjects presenting the event to detect (i.e., pneumonia) in the training folder or in a separate folder specifically designed to conduct the feature selection process [67–69]. The splitting of the dataset in subfolders is crucial in order to avoid bias and over-fitting problems and increase the external validity of the model. If data availability is not a problem, the dataset could be split into three different folders, where folder 1 is designed for feature selection via several existing techniques [70–71]; folder 2 to train and validate the model; an independent dataset (folder 3) to test the final model and assess the overall performance [67–69]. However, although the best approach is to select the minimum set of features using a different folder from the one adopted to train the machine learning model [67–69], in case the dataset is small, feature selection and model training can be performed on the same folder (folder 1). As reported in Table 2, some studies did not employ a clear feature selection method and in case they did, they performed the feature selection on the whole dataset or during the training of the algorithm. It is important to bear in mind that a significant small set of clinical features strongly simplifies the physiological interpretation of results, by directing attention only on the most informative features [67]. In the detection of pneumonia, the identification of symptoms that can be used as final predictors is of extreme importance to the physicians. Therefore, hand-crafted features and the use of PCA is not recommended. As reported in Table 2, there is a mixture of manual and automated approaches to feature selection process in the selected studies. Manual methods have a clear focus on clinical utility and application. Some key criteria used were: (i) measurable in a point of care setting [23]; (ii) parameters frequently investigated [36]; (iii) ease of availability [35,62] and (iv) reliability [35]. Haug et al. [39] make an interesting comparison between a fully automated ML model, from feature selection to performance, and a semi-automated model in which features are chosen manually based on medical relevance by clinicians. The large dataset available in this study allowed selection of 40 features by both methods. Of these features there was considerable overlap, notably certain symptoms picked up by both methods were: heart rate, respiratory rate, temperature, abnormal breath sounds, moderate cough, wheezes, productive cough and rales breath sounds. It seems certain features such as 'Not oriented to place', which are selected in the automated process are absent in the manual, perhaps due to a lack of direct clinical/biological relevance to pneumonia. Interestingly slightly better performance was achieved using the manually created model, which may highlight the motivation for a firm evidence basis in ML design. Other popular methods were uni/multi variate analysis and logistic regression. One technique appearing in the most recent publication [47] was Least Absolute Shrinkage and Selection Operator (LASSO). LASSO builds on classic regression models and is emerging as a more interpretable clinically useful method for selecting predictors, as by nature it strives to create sparse models (fewer predictors) [72]. Five studies [38,41,44,46,49] did not employ any feature selection process.

As far as the validation process is regarded, the training dataset is not known to have a sub-category, whereas the validation dataset can be further divided by types: (i) internal validation, whose sample originates from the same sample as the training dataset, (ii) external validation, whose sample is composed of independently sampled data, (iii) internal-split validation, which uses a sample that has been separated from the original dataset for the purpose of validation, and (iv) internal-cross validation, which repeats validation process over a sample that is left out of the training dataset. Five studies [38,42–45] did not employ either internal or external validation techniques, making the developed models difficult to generalize and compare with other diagnostic tools. Only three out of 16 identified studies [23,36,48] employed both cross-validation and testing on an independent set of data. Two studies [47,49] tested the models on an independent subset of data. The remaining studies developed the ML models using training and internal validation techniques.

The majority of the included studies employed big datasets which were highly unbalanced. In medicine, a well-balanced dataset is vital to develop a good prediction model [73]. In fact, when the imbalance is large, it is hard to build a good classifier using conventional learning algorithms. The cost in miss predicting minority classes is higher than that of the majority class for imbalanced datasets; this is particularly relevant in medical datasets where high risk patients tend to be the minority class (e.g., pneumonia cases). Therefore, there is a need of a good sampling technique for medical datasets. Among the selected studies, only four out of the 16 studies [23,36,40,43] adopted a bootstrapping or oversampling technique to address the problem of unbalanced datasets.

Among the selected studies, there are a variety of predictors used to develop the machine learning algorithms. Eight studies used a combination of laboratory results and symptoms as their final predictors, with only 5 papers using symptoms alone. Symptoms/signs which occurred often included: fever (5 studies), temperature (5), abnormal breathing (4), cough (3), productive cough (2), dyspnoea (2), absence of runny nose (2) and chest in drawing (2). Other population differences are also reflected in the final predictors, for example chest in drawing is only used in studies concerning childhood pneumonia. This is consistent with the known age-specific presentations of the disease [74] and thus highlights a potential challenge in production of a general model. The

utility of C-reactive protein (CRP) level as a biomarker in classifying pneumonia was addressed by 4 studies. As it has been recognised already in the literature [13–14], there were some contradictions between studies of its performance as a pneumonia predictor. Naydenova et al. [23] and Groeneveld et al. [45] specifically investigated addition of CRP to models based on symptoms, vital signs and age and find that CRP worsened model performance in diagnosis of pneumonia. It is worth noting, however, that both authors described the utility of CRP as a beneficial predictor for pneumonia severity and aetiology. Interestingly the reference standard of pneumonia in these studies is not the same, Naydenova et al. [23] had a subject population of children and use clinical evaluation based on WHO and IMCI guidelines, whereas the subject population in Groeneveld et al. [45] was adult and their reference was consolidation on X-ray. In contrast to this, Steurer et al. [35] and van Vugt et al. [40] found CRP to be a useful predictor. Indeed, Steurer et al. [35] found CRP to be the strongest indicator of radiographically confirmed pneumonia in adults from a set of mostly symptomatic predictors. Together, this highlights the need for further investigation of biomarkers as candidate features for diagnostic classifiers to gain further understanding of the seemingly complex presentation of these levels. Only one study [49] used a combination of 3D CT imaging and clinical symptoms via deep learning model (FaNet) to detect patients affected by COVID-19 pneumonia. Their experimental results illustrated that FaNet achieves fast clinical assessment for COVID-19 with an accuracy of 98.28%. The proposed framework consisted of 4 modules: encoding for symptoms, feature extraction from CT image sequences, fusion, and prediction. They developed a Symptom-fused Channel Attention Module to fuse the clinical symptoms and the CT image sequences. Finally, the prediction module predicts the clinical assessment based on the fused feature.

Some studies used a combination of many signs or symptoms, even though they employed large datasets, the class to predict (i.e., pneumonia) is often the minority class. According to Foster at al. [67], as rule of thumb, for each predictor at least 10 observations and/or patients are needed for the event to detect. In the case of some studies, the number of predictors overcome the number of patients included in the target classes, incrementing the risk of overfitting of the model.

Comparison of predictors and ML model performance across all studies is strictly limited for several reasons: (i) variation in pneumonia type/reference standard; (ii) variation in subject population and (iii) differential reporting of performance metrics. The overall performance AUC varied to 75–99%. However, not all the included studied reported AUC measure. Moreover, there is great heterogeneity in performance reporting of the diagnostic tools used in the included studied. The reference standard to report performance of ML methods is described by [75]. The lack of homogeneity among the selected studies in ML development and performance reporting were the main reasons of conducting a qualitative systematic review as meta-analysis was not possible with the available gathered data.

a. Recommendations when designing and implementing AI tools

In light of this scenario, recommendations of on how to develop a ML method is given to researchers to improve the efficacy of AI tools to automatically detect pneumonia or any other respiratory diseases. The recommended pipeline is formed by:

1. *Pre-Processing step.* For building any ML model, it is important to have a sufficient amount of data to train the model. The data is often collected from various resources and might be available in different formats. Due to this reason, data cleaning and pre-processing become a crucial step, which include: impute the missing values, encode categorical variables (in case of symptoms), normalize and/or scale the data if required. Moreover, clinical information and reference standard results should be available to the performer of the ML model. More important, explanations on how indeterminate reference standard results were handled should be provided.
2. *Dataset splitting.* In the case where the dataset presents an adequate number of instances, the whole dataset can randomly be split per subjects and or instances into two or more sub-folders. For instance, one folder (usually the 20 % of the total data) can be used for feature selection; a second folder 2 (usually the majority of the data, 60%) can be used for training and validating the classification models; finally, a third folder (e.g., 20 % of the data) is adopted to evaluate the performance of the developed classification models. In the case of a highly imbalanced dataset, each folder should contain the same proportional percentage of minority instances and techniques to address this problem should be employed (e.g., SMOTE, over-sampling, under sampling or boosting).
3. *Identifying features to predict the target.* The number of features used in a machine learning algorithm should be strongly limited by the number of subjects and or instances presenting the event to detect in each folder, in order to minimise the risk of over-fitting. However, selecting the minimum set of features using the same folder utilised to train the machine learning algorithm can reduce the generalisability of the final decisional algorithm. Researchers using manual feature selection based on clinical usage or more advanced techniques to reduce the number of features, should always bear in mind that the maximum number of features that can be used in the classification process is strongly limited to the number of subjects (i.e., belonging to the minority class) presenting the event to detect or predict.
4. *Designing the ML Pipeline using the best model.* Different ML methods can be used to develop classifiers aiming to automatically detect the event based on the selected combinations of features. Regarding algorithm parameters, they should be tuned during training and carefully reported in the study to guarantee the reproductivity of the results. The training of the ML methods should be performed using cross-validation procedure, which needs to be repeated K times, with K equal to or greater than the number of instances belonging to the minority class. This procedure needs to be performed for each machine learning method used to develop predictive algorithms.
5. *Predict the target on the unseen data.*
6. *Reporting performance according to standards.* Moreover, researchers are highly encouraged to define the rationale for test positivity cut-offs or result categories of the ML method, distinguishing pre-specified from exploratory results.

b. Limitations of the study

This study has provided several new insights on the existing approaches to predicting pneumonia based on signs and symptoms and the aspects that warrant consideration both in the design and implementation phases of the tests and in the reporting of the findings. However, several limitations are also outlined. To the authors' knowledge, there is no available and reliable tool for the quality assessment of studies incorporating ML, and as a result, the quality of the studies that have been found in this area could not be systematically assessed. Second, while the medRxiv preprints database was included in the search strategy, in order to capture all possible recent contributions addressing SARS-COV related pneumonia, the search was conducted using simple combinations of search terms due to the limited flexibility of the available search options. Therefore, relevant records in this dataset may have been missed. Lastly, we limited our search in the bibliographic databases to the last 10 years. This choice was driven a motivation to evaluate the recent use of ML techniques.

5. Conclusion

This systematic literature review found huge heterogeneity among studies using ML to detect pneumonia based on symptoms and signs. Many differing study designs, healthcare settings, populations and ML

algorithms were used. The most frequently used symptoms as ML features were fever, temperature, abnormal breathing, and cough. Several studies did not follow best practice in their ML methodology, in particular during feature selection, model validation, and handling of imbalanced datasets. In addition, reporting quality was low for details of the reference tests used and characteristics of the study participants (severity of disease and alternative diagnoses). However, overall performance was high, suggesting there is strong motivation for further investigations using ML to improve diagnostic capability of existing CAD systems for automatic pneumonia detection. Such systems may contribute to improving access to diagnosis for respiratory disease in limited-resource settings. This review is limited to recent applications of ML (within the past 10 years). In addition, quality assessment of ML could not be assessed due to lack of an appropriate validated tool. Despite the limitations, this study provides insights on existing approaches to ML based pneumonia detection and provides recommendations for future research in how best to develop a ML method for effective automatic detection tools with clinical utility.

Funding

This work was supported by EPSRC Impact Accelerator Award (EP/K503848/1 and EP/R511808/1).

AM is supported by WIRL COFUND (Marie Sklodowska Curie Actions) Institute of Advanced Study - University of Warwick (UK).

KS is funded by the MRC Doctoral Training Partnership [grant number MR/N014294/1].

RC, MS and MF are supported by “Progetti di Ricerca Corrente” funded by the Italian Ministry of Health.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bspc.2021.103325.

References

[1] J. BiscevicTokic, N. Tokic, A. Musanovic, Pneumonia as the most common lower respiratory tract infection, Med. Archiv. 67 (6) (2013) 442.
[2] Zanfardino, Pane, Mirabelli, Salvatore, Franzese, TCGA-TCIA Impact on Radiogenomics Cancer Research: A Systematic Review, Int. J. Mol. Sci. 20 (23) (2019) 6033, https://doi.org/10.3390/ijms20236033.
[3] J.G. Bartlett, L.M. Mundy, Community-acquired pneumonia, N. Engl. J. Med. 333 (24) (1995) 1618–1624.
[4] S. Visscher, P.J.F. Lucas, C.A.M. Schurink, M.J.M. Bonten, Modelling treatment effects in a clinical Bayesian network using Boolean threshold functions, Artif. Intell. Med. 46 (3) (2009) 251–266.
[5] E. Catherinot, F. Lanternier, M.-E. Bougnoux, M. Lecuit, L.-J. Couderc, O. Lortholary, Pneumocystis jirovecii pneumonia, Infectious Disease Clinics 24 (1) (2010) 107–138.
[6] E.S. Kim, K.U. Park, S.H. Lee, Y.J. Lee, J.S. Park, Y.-J. Cho, H.I. Yoon, C.-T. Lee, J. H. Lee, C.D. Russell, Comparison of viral infection in healthcare-associated pneumonia (HCAP) and community-acquired pneumonia (CAP), PLoS ONE 13 (2) (2018) e0192893.
[7] S.T. Micek, K.E. Kollef, R.M. Reichley, N. Roubinian, M.H. Kollef, Health Care-Associated Pneumonia and Community-Acquired Pneumonia: a Single-Center Experience, Antimicrob. Agents Chemother. 51 (10) (2007) 3568–3573.
[8] T.M. Wardlaw, et al., Pneumonia : the forgotten killer of children, World Health Organization, Geneva, 2006.
[9] J.M. Galván, O. Rajas, J. Aspa, Review of Non-Bacterial Infections in Respiratory Medicine: Viral Pneumonia, Arch. Bronconeumol. 51 (11) (2015) 590–597.
[10] W.-J. Guan, Z.-y. Ni, Y.u. Hu, W.-H. Liang, C.-Q. Ou, J.-X. He, L. Liu, H. Shan, C.-L. Lei, D.S.C. Hui, B. Du, L.-J. Li, G. Zeng, K.-Y. Yuen, R.-C. Chen, C.-l. Tang, T. Wang, P.-Y. Chen, J. Xiang, S.-Y. Li, J.-L. Wang, Z.-J. Liang, Y.-X. Peng, L.i. Wei, Y. Liu, Y.-H. Hu, P. Peng, J.-M. Wang, J.-Y. Liu, Z. Chen, G. Li, Z.-J. Zheng, S.-Q. Qiu, J. Luo, C.-J. Ye, S.-Y. Zhu, N.-S. Zhong, Clinical characteristics of coronavirus disease 2019 in China, N. Engl. J. Med. 382 (18) (2020) 1708–1720.
[11] J. Yang, Y.a. Zheng, X.i. Gou, K.e. Pu, Z. Chen, Q. Guo, R. Ji, H. Wang, Y. Wang, Y. Zhou, Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis, Int. J. Infect. Dis. 94 (2020) 91–95.
[12] L. Gattinoni, et al., COVID-19 pneumonia: different respiratory treatments for different phenotypes? Springer, 2020.
[13] S. Spoto, J.M. Legramante, M. Minieri, M. Fogolari, A. Terrinoni, E. Valeriani, C. Sebastiano, S. Bernardini, M. Ciccozzi, P.S. Angeletti, How biomarkers can improve pneumonia diagnosis and prognosis: procalcitonin and mid-regional-pro-adrenomedullin, Biomarkers Med. 14 (7) (2020) 549–562.
[14] M. Christ-Crain, P. Schuetz, B. Müller, Biomarkers in the management of pneumonia, Expert review of respiratory medicine 2 (5) (2008) 565–572.
[15] L. Rosenberg, et al., Artificial swarm intelligence employed to amplify diagnostic accuracy in radiology. 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), IEEE, 2018.
[16] World Health Organization. Pneumonia Vaccine Trial Investigators, G. and O. World Health, Standardization of interpretation of chest radiographs for the diagnosis of pneumonia in children / World Health Organization Pneumonia Vaccine Trial Investigators’ Group. 2001, World Health Organization: Geneva.
[17] S. Ben Shimol, R. Dagan, N. Givon-Lavi, A. Tal, M. Aviram, J. Bar-Ziv, V. Zodicov, D. Greenberg, Evaluation of the World Health Organization criteria for chest radiographs for pneumonia diagnosis in children, Eur. J. Pediatr. 171 (2) (2012) 369–374.
[18] M.A. Elemraid, M. Muller, D.A. Spencer, S.P. Rushton, R. Gorton, M.F. Thomas, K. M. Eastham, F. Hampton, A.R. Gennery, J.E. Clark, S. Worgall, Accuracy of the interpretation of chest radiographs for the diagnosis of paediatric pneumonia, PLoS ONE 9 (8) (2014) e106051.
[19] M.D. Garber, R.A. Quinonez, Chest Radiograph for Childhood Pneumonia: Good, but Not Good Enough, Pediatrics 142 (3) (2018) e20182025, https://doi.org/10.1542/peds.2018-2025.
[20] M. Miravitlles, Diagnosis of asthma–COPD overlap: the five commandments, Eur. Respir. J. 49 (5) (2017) 1700506, https://doi.org/10.1183/13993003.00506-2017.
[21] S. Kinkade, N.A. Long, Acute Bronchitis, Am Fam Physician 94 (7) (2016) 560–565.
[22] D. Ben-Israel, et al., The impact of machine learning on patient care: A systematic review, Artif. Intell. Med. 103 (2020), 101785.
[23] E. Naydenova, A. Tsanas, S. Howie, C. Casals-Pascual, M. De Vos, The power of data mining in diagnosis of childhood pneumonia, J. R. Soc. Interface 13 (120) (2016) 20160266, https://doi.org/10.1098/rsif.2016.0266.
[24] G.F. Cooper, C.F. Aliferis, R. Ambrosino, J. Aronis, B.G. Buchanan, R. Caruana, M. J. Fine, C. Glymour, G. Gordon, B.H. Hanusa, J.E. Janosky, C. Meek, T. Mitchell, T. Richardson, P. Spirtes, An evaluation of machine-learning methods for predicting pneumonia mortality, Artif. Intell. Med. 9 (2) (1997) 107–138.
[25] C.J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, D. King, Key challenges for delivering clinical impact with artificial intelligence, BMC Medicine 17 (1) (2019), https://doi.org/10.1186/s12916-019-1426-2.
[26] A. Sabet Sarvestani, M. Coulentianos, K.H. Sienko, Defining and characterizing task-shifting medical devices, Global Health 17 (1) (2021) 60.
[27] M.J. Page, et al., PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews, bmj 372 (2021).
[28] D. Moher, A. Liberati, J. Tetzlaff, D.G. Altman, Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement, PLoS Med 6 (7) (2009) e1000097.
[29] P. Macaskill et al., Cochrane handbook for systematic reviews of diagnostic test accuracy. Version 0.9. 0. London: The Cochrane Collaboration, 2010.
[30] P.M. Bossuyt, et al., STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies, Clin. Chem. 61 (12) (2015) 1446–1452.
[31] W.M. Bramer, M.L. Rethlefsen, J. Kleijnen, O.H. Franco, Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study, System. Rev. 6 (1) (2017), https://doi.org/10.1186/s13643-017-0644-y.
[32] A.-W. Harzing, S. Alakangas, Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison, Scientometrics 106 (2) (2016) 787–804.
[33] A. Martín-Martín, M. Thelwall, E. Orduna-Malea, E. Delgado López-Cózar, Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations’ COCI: a multidisciplinary comparison of coverage via citations, Scientometrics 126 (1) (2021) 871–906.
[34] R. Prager, J. Bowdridge, H. Kareemi, C. Wright, T.A. McGrath, M.D.F. McInnes, Adherence to the Standards for Reporting of Diagnostic Accuracy (STARD) 2015 Guidelines in Acute Point-of-Care Ultrasound Research, JAMA Network Open 3 (5) (2020) e203871, https://doi.org/10.1001/jamanetworkopen.2020.3871.
[35] J. Steurer, U. Held, A. Spaar, B. Bausch, M. Zoller, R. Hunziker, L.M. Bachmann, A decision aid to rule out pneumonia and reduce unnecessary prescriptions of antibiotics in primary care patients with cough and fever, BMC Medicine 9 (1) (2011), https://doi.org/10.1186/1741-7015-9-56.
[36] L. Grigull, W.M. Lechner, Supporting diagnostic decisions using hybrid and complementary data mining applications: a pilot study in the pediatric emergency department, Pediatr. Res. 71 (6) (2012) 725–731.
[37] C.A. Bejan, F. Xia, L. Vanderwende, M.M. Wurfel, M. Yetisgen-Yildiz, Pneumonia identification using statistical feature selection, J. Am. Med. Inf. Assoc.: JAMIA 19 (5) (2012) 817–823.
[38] S. DeLisle, B. Kim, J. Deepak, T. Siddiqui, A. Gundlapalli, M. Samore, L. D’Avolio, M.G. Semple, Using the electronic medical record to identify community-acquired pneumonia: toward a replicable automated strategy, PLoS ONE 8 (8) (2013) e70944.

[39] P.J. Haug, J.P. Ferraro, J. Holmen, X. Wu, K. Mynam, M. Ebert, N. Dean, J. Jones, An ontology-driven, diagnostic modeling system, J. Am. Med. Inf. Assoc.: JAMIA 20 (e1) (2013) e102–e110.
[40] S.F. van Vugt, B.D.L. Broekhuizen, C. Lammens, N.P.A. Zuithoff, P.A. de Jong, S. Coenen, M. Ieven, C.C. Butler, H. Goossens, P. Little, T.J.M. Verheij, Use of serum C reactive protein and procalcitonin concentrations in addition to symptoms and signs to predict pneumonia in patients presenting to primary care with acute cough: diagnostic study, BMJ (Clinical Research ed.) 346 (apr30 1) (2013) 2450.
[41] A.-K. Rother, N. Schwerk, F. Brinkmann, F. Klawonn, W. Lechner, L. Grigull, D. Hartl, Diagnostic Support for Selected Paediatric Pulmonary Diseases Using Answer-Pattern Recognition in Questionnaires Based on Combined Data Mining Applications–A Monocentric Observational Pilot Study, PLoS ONE 10 (8) (2015) e0135180.
[42] S. Nuzhat, T. Ahmed, C.A. Kawser, A.I. Khan, S.M.R. Islam, L. Shahrin, K. M. Shahunja, A.S.M.S.B. Shahid, A. Al Imran, M.J. Chisti, O. Schildgen, Age specific fast breathing in under-five diarrheal children in an urban hospital: Acidosis or pneumonia? PLoS ONE 12 (9) (2017) e0185414.
[43] O. De Santis, M. Kilowoko, E. Kyungu, W. Sangu, P. Cherpillod, L. Kaiser, B. Genton, V. D'Acremont, K. Mortimer, Predictive value of clinical and laboratory features for the main febrile diseases in children living in Tanzania: A prospective observational study, PLoS ONE 12 (5) (2017) e0173314.
[44] F. Pervaiz, M.A. Chavez, L.E. Ellington, M. Grigsby, R.H. Gilman, C.H. Miele, D. Figueroa-Quintanilla, P. Compen-Chang, J. Marin-Concha, E.D. McCollum, W. Checkley, Building a Prediction Model for Radiographically Confirmed Pneumonia in Peruvian Children: From Symptoms to Imaging, Chest 154 (6) (2018) 1385–1394.
[45] G.H. Groeneveld, J.W. van 't Wout, N.J. Aarts, C.J. van Rooden, T.J.M. Verheij, C. M. Cobbaert, E.J. Kuijper, J.J.C. de Vries, J.T. van Dissel, Prediction model for pneumonia in primary care patients with an acute respiratory tract infection: role of symptoms, signs, and biomarkers, BMC Infect. Dis. 19 (1) (2019), https://doi.org/10.1186/s12879-019-4611-1.
[46] P. Porter, U. Abeyratne, V. Swarnkar, J. Tan, T.-W. Ng, J.M. Brisbane, D. Speldewinde, J. Choveaux, R. Sharan, K. Kosasih, P. Della, A prospective multicentre study testing the diagnostic accuracy of an automated cough sound centred analytic system for the identification of common respiratory disorders in children, Respir. Res. 20 (1) (2019), https://doi.org/10.1186/s12931-019-1046-6.
[47] C. Feng, et al., A Novel Triage Tool of Artificial Intelligence Assisted Diagnosis Aid System for Suspected COVID-19 pneumonia In Fever Clinics, Ann. Translational Med. 9 (3) (2021).
[48] G. Yu, et al., Identification of pediatric respiratory diseases using a fine-grained diagnosis system, J. Biomed. Inform. 117 (2021), 103754.
[49] Z. Huang, X. Liu, R. Wang, M. Zhang, X. Zeng, J. Liu, Y. Yang, X. Liu, H. Zheng, D. Liang, Z. Hu, FaNet: fast assessment network for the novel coronavirus (COVID-19) pneumonia based on 3D CT imaging and clinical symptoms, Appl. Intell. 51 (5) (2021) 2838–2849.
[50] R.E. Wright, Logist. Regress. (1995).
[51] D. Graupe, Principles of Artificial Neural Networks. Vol. 7., World Scientific, 2013.
[52] J. Gu, et al., Recent advances in convolutional neural networks, Pattern Recogn. 77 (2018) 354–377.
[53] G.F. Glonek, P. McCullagh, Multivariate logistic models, J. Roy. Stat. Soc.: Ser. B (Methodol.) 57 (3) (1995) 533–546.
[54] J. Ranstam, J. Cook, LASSO regression, J. Br. Surg. 105 (10) (2018) 1348.
[55] Y. Zoabi,N. Shomron, COVID-19 diagnosis prediction by symptoms of tested individuals: a machine learning approach. medRxiv, 2020: p. 2020.05.07.20093948.
[56] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[57] D. Steinberg, CART: Classification and Regression Trees, in: The top ten algorithms in data mining, Chapman and Hall/CRC, 2006. pp. 193-216.
[58] T.G. Dietterich, Ensemble methods in machine learning. International workshop on multiple classifier systems, Springer, 2000.
[59] J. Lafferty, A. McCallum, F.C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.
[60] S.R. Gunn, Support vector machines for classification and regression, ISIS Tech. Rep. 14 (1) (1998) 5–16.
[61] J. Huang, C. Yuan, Fanet: factor analysis neural network. International Conference on Neural Information Processing, Springer, 2015.
[62] C. Feng et al., A Novel Triage Tool of Artificial Intelligence Assisted Diagnosis Aid System for Suspected COVID-19 pneumonia In Fever Clinics. medRxiv, 2020, p. 2020.03.19.20039099.
[63] L. Wynants, et al., Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal, BMJ 369 (2020), m1328.
[64] I. Ozsahin, et al., Review on Diagnosis of COVID-19 from Chest CT Images Using Artificial Intelligence, Comput. Math. Methods Med. 2020 (2020) 9756518.
[65] M.CI. van Schalkwyk, A. Bourek, D.S. Kringos, L. Siciliani, M.M. Barry, J. De Maeseneer, M. McKee, The best person (or machine) for the job: Rethinking task shifting in healthcare, Health Policy 124 (12) (2020) 1379–1386.
[66] D. Driggs, I. Selby, M. Roberts, E. Gkrania-Klotsas, J.H.F. Rudd, G. Yang, J. Babar, E. Sala, C.-B. Schönlieb, Machine Learning for COVID-19 Diagnosis and Prognostication: Lessons for Amplifying the Signal While Reducing the Noise, Radiol. Artif. Intell. 3 (4) (2021) e210011, https://doi.org/10.1148/ryai.2021210011.
[67] K.R. Foster, R. Koprowski, J.D. Skufca, Machine learning, medical diagnosis, and biomedical engineering research-commentary, Biomed. Eng. Online 13 (1) (2014) 94, https://doi.org/10.1186/1475-925X-13-94.
[68] R. Castaldo, P. Melillo, R. Izzo, N. De Luca, L. Pecchia, Fall prediction in hypertensive patients via short-term HRV Analysis, IEEE J. Biomed. Health. Inf. 21 (2) (2017) 399–406.
[69] R. Castaldo, L. Montesinos, P. Melillo, C. James, L. Pecchia, Ultra-short term HRV features as surrogates of short term HRV: a case study on mental stress detection in real life, BMC Med. Inf. Decis. Making 19 (1) (2019), https://doi.org/10.1186/s12911-019-0742-y.
[70] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (1) (2014) 16–28.
[71] J. Tang, S. Alelyani, H. Liu, Feature selection for classification: A review, in: Data classification: Algorithms and applications, 2014, p. 37.
[72] T. Goto, C.A. Camargo, M.K. Faridi, R.J. Freishtat, K. Hasegawa, Machine Learning-Based Prediction of Clinical Outcomes for Children During Emergency Department Triage, JAMA Netw. Open 2 (1) (2019) e186937.
[73] M.M. Rahman, D.N. Davis, Addressing the class imbalance problem in medical datasets, Int. J. Mach. Learn. Comput. 3 (2) (2013) 224.
[74] E. Prina, O.T. Ranzani, A. Torres, Community-acquired pneumonia, Lancet (London, England) 386 (9998) (2015) 1097–1108.
[75] C. Parker, An analysis of performance measures for binary classifiers. 2011 IEEE 11th International Conference on Data Mining, IEEE, 2011.