Artificial Intelligence Applications And Innovations Vol. 1 No. 4 (2024)

Article history:
Received 5 June 2024
Revised 30 July 2024
Accepted 03 September 2024
Published online 01 October 2024

The Role of Artificial Intelligence in Tuberculosis Diagnosis

Hojatollah. Hamidi[1*], Mohsen. Saffar[1]

[1] Department of Industrial Engineering, Information Technology Group, K. N. Toosi University of Technology, Tehran, Iran

*** Corresponding author email address**: h_hamidi@kntu.ac.ir

A r t i c l e I n f o

Article type:
Review Article

How to cite this article:
Hamidi, H., & Saffar, M. (2024). The Role of AI in Tuberculosis Diagnosis: An Umbrella Review. *Artificial Intelligence Applications and Innovations, 1*(4), 40-54.
https://doi.org/10.61838/jaiai.1.4.4

A B S T R A C T

Tuberculosis remains a major global health challenge, and the World Health Organization has endorsed artificial intelligence tools to support imaging-based screening. This umbrella review mapped AI applications across tuberculosis prevention, diagnosis, and treatment, focusing on data modalities, algorithm types, and validation practices. Following PRISMA guidelines, systematic reviews published between January 2020 and June 2025 were identified through PubMed and Web of Science, and eligible studies were screened, extracted, and appraised in duplicate. Twelve reviews covering 648 primary studies were included from an initial 1,796 records. Accuracy, AUC, sensitivity, and specificity were reported in 54 %, 43 %, 45 %, and 40 % of studies, respectively. Only 33 % conducted internal validation and 4 % performed external validation against independent cohorts or human readers. Deep learning models such as VGG16, ResNet50, and InceptionV3 dominated, while classical algorithms like support vector machines and random forests persisted in contexts requiring interpretability or when data were limited. The findings highlight rapid growth in AI use for tuberculosis but reveal inconsistent reporting, limited external validation, and minimal attention to missing data. Broader clinical adoption will require rigorous, multimodal studies with standardized performance metrics, transparent algorithms, and robust validation strategies.

Keywords: *Tuberculosis, Tuberculosis Diagnosis, Systematic Review, Artificial Intelligence, Deep Learning, Machine Learning*

1. Introduction

Tuberculosis (TB) is one of humanity's oldest and most formidable diseases, with molecular evidence of its existence for over 17,000 years; despite modern advances in diagnosis and treatment, it remains a major global health threat, ranking among the top ten deadliest infectious diseases and causing over 10 million new cases annually [1]. The World Health Organization's Global Tuberculosis Report for 2023 highlights that TB led to almost twice the number of deaths as HIV/AIDS, positioning it as the second deadliest infectious disease behind COVID-19 [1]. Tuberculosis can be effectively cured if identified promptly and treated appropriately [2].

This condition arises from infection with Mycobacterium tuberculosis [3], which predominantly targets the lungs but can also inflict damage on other areas such as the brain, intestines, and spine [4]. Transmission occurs through tiny airborne droplets expelled when someone with the infection coughs or sneezes [5]. TB can manifest in two distinct forms: latent TB, which shows no symptoms and cannot be spread to others, and active TB, which often brings on symptoms including fever, persistent fatigue, chills, night sweats, and a diminished appetite [6]. Even though latent

TB is non-contagious, it has the potential to advance to the active stage, which emphasizes the ongoing need to accurately differentiate between these forms, a task that continues to challenge clinicians in everyday practice [7].

Accurate diagnosis and timely treatment are essential for controlling the disease and minimizing its severity. Conventional approaches for detecting tuberculosis typically include molecular assays, culture methods, sputum smear microscopy, and chest X-ray examinations [8]. Culture tests, which involve growing bacteria from patient samples, are regarded as highly accurate, although the process can be quite lengthy [9]. Sputum smear microscopy works by analyzing sputum samples under a microscope to identify Mycobacterium tuberculosis [10]. Likewise, molecular tests based on PCR techniques can deliver results much more rapidly, but these methods tend to be costly [11].

In the realm of diagnostics, chest imaging particularly chest X-rays (CXR) serves as a cornerstone for identifying intrathoracic TB, thanks to its affordability and widespread availability, especially in children [12]. More sophisticated imaging methods, including CT, MRI, and PET/CT, deliver enhanced precision for spotting both pulmonary and extrapulmonary manifestations of the disease [13]. Over the past few years, the automation of CXR analysis has progressed at a remarkable pace, fueled by deep learning approaches based on convolutional neural networks (CNNs), which underpin the development of AI-assisted diagnostic systems [14-16]. The use of deep learning in radiology is expanding rapidly thanks to its impressive ability to detect diseases. For example, it has shown strong results in identifying pleural effusion and cardiomegaly on chest X-rays [17, 18],as well as spotting mediastinal lymph nodes and lung nodules on CT scans [19, 20]. In particular, researchers have found that AI-powered chest X-rays (CXR) hold great potential for diagnosing tuberculosis, especially in rural areas where medical resources are limited [15].

The advent of computer-based systems has led to the widespread digitization of medical records and the systematic assessment of clinical data within healthcare organizations. Today, healthcare institutions generate vast volumes of data each day, rendering traditional analytical methods increasingly impractical. The adoption of machine learning and deep learning techniques has made it possible to efficiently analyze these large datasets and derive meaningful, actionable insights [21]. Machine Learning is currently one of the most rapidly advancing areas in computer science, posing unique challenges and opportunities in health informatics. The primary goal of ML is to build algorithms capable of continuous learning and improvement, which can be applied for predictive tasks. The healthcare sector, in particular, has experienced significant benefits from machine learning prediction approaches [22]. Essentially, machine learning involves training systems to interpret input data to forecast outcomes or extract valuable information. It is a branch of artificial intelligence (AI), closely related to statistical science [23].

In light of these technological breakthroughs, the World Health Organization in 2021 gave its endorsement to AI tools for interpreting CXRs during TB screening programs, allowing them to supplant human experts in select situations [13]. No matter the diagnostic strategy employed, achieving early and precise detection is essential for advancing global initiatives to curb TB [24]. Ultimately, effective diagnosis hinges on a holistic integration of imaging findings with laboratory results and detailed clinical histories, ensuring that interpretations of images are fully grounded in the patient's broader medical context [25].

To effectively integrate AI into TB care and make its adoption both practical and sustainable, it's essential to gain a deeper understanding of the opportunities and hurdles associated with applying AI to TB prevention, diagnosis, and treatment. With this in mind, our primary aim was to provide a comprehensive overview of the current landscape, focusing on the availability and effectiveness of AI tools designed specifically for addressing TB.

We carried out an umbrella review, which is essentially a systematic synthesis of existing systematic reviews. This involved compiling insights from 12 such reviews that explore the wide range of AI applications in the context of tuberculosis.

Our study had a dual purpose: firstly, to catalog the various AI solutions that have been developed or suggested for managing TB; and secondly, to examine the machine learning and deep learning algorithms used for the detection and diagnosis of TB, primarily focusing on images such as chest X-rays and CT scans.

The next section reviews the articles that were extracted according to the methodology.

2. Literature Review

Nansamba et al. (2025) [26] examine the burgeoning role of multimodal learning and explainable artificial

intelligence (XAI) in tuberculosis (TB) diagnosis, addressing clinicians' need to synthesize diverse patient information for accurate decision-making. Motivated by the limitations of single-modality AI models and the opacity of deep-learning "black boxes," the authors conducted a PRISMA-guided systematic review of 31 studies published between 2019 and June 2024. They extracted data on public datasets, modality combinations, fusion strategies, algorithms, and interpretability techniques, comparing diagnostic performance across approaches. Their analysis shows that models integrating imaging with clinical, laboratory, or genomic data consistently outperform unimodal counterparts, yet progress is hindered by scarce multimodal datasets and persistent interpretability challenges. The review advocates for publicly available, well-annotated multimodal TB repositories and for embedding XAI methods within fusion frameworks to enhance diagnostic accuracy, transparency, and clinical trust.

Building on the emphasis on multimodal approaches and interpretability challenges, Santosh et al. (2022) [27] present a systematic appraisal of deep-learning advances in chest-X-ray-based tuberculosis screening published between 2016 and 2021. Motivated by the continuing global TB burden and the rapid maturation of convolutional neural networks, the authors screened PubMed and Web of Science, ultimately meta-analyzing 54 peer-reviewed studies. Applying PRISMA guidelines, they catalogued publicly available CXR datasets, network architectures, transfer-learning strategies, data-augmentation practices and performance metrics. Results reveal an explosive shift from handcrafted features to CNN-driven classifiers—often enhanced by ensemble learning, modality-specific pre-training and explainable activation mapping—yielding AUCs that now approach or exceed expert radiologist levels. Nonetheless, progress is tempered by limited annotated data, class imbalance and inconsistent validation protocols. The review underscores the need for larger, balanced, multi-institutional datasets, standardized evaluation frameworks and integrated explainability to foster clinically reliable AI-assisted TB triage.

Echoing the need for standardized datasets and validation in CXR-based screening, Han et al. (2025) systematically appraise the diagnostic performance of contemporary AI-driven computer-aided detection tools for pulmonary tuberculosis (PTB) on chest X-rays. Motivated by the persistent global TB burden and radiologist shortages, the authors searched four major databases to December 2024, screened 5,651 records, and meta-analyzed 21 eligible studies covering five commercial products (JF CXR-1, qXR, Lunit INSIGHT, CAD4TB, InferRead). Using QUADAS-2 quality assessment and random- or fixed-effects modelling in Stata, they derived pooled sensitivities of 86–91 % and specificities of 59–80 %, noting that recent software versions consistently outperformed predecessors. Nonetheless, no system simultaneously achieved the WHO triage target of 90 % sensitivity and 70 % specificity, and substantial heterogeneity persisted. The review concludes that while AI tools markedly expedite Pulmonary Tuberculosis (PTB) screening, further algorithm optimization, context-specific threshold tuning, and diversified clinical validation are essential before routine deployment [28].

In a similar vein, focusing on deep-learning advances in CXR for TB diagnosis, Oloko-Oba and Viriri (2022) [29]conduct a PRISMA-guided systematic review to map recent deep-learning advances in computer-aided tuberculosis diagnosis from chest radiographs. Searching Scopus, IEEE Xplore, Web of Science and PubMed for 2017–2021, they screened 489 records and analysed 62 full-text studies that used CXR as the sole imaging modality and at least one deep-learning classifier. Convolution-based transfer-learning models especially VGG, ResNet, DenseNet and Inception variants dominate the field, often achieving accuracies above 90 % and AUCs up to 0.99 on public datasets such as Shenzhen, Montgomery and ChestX-ray14. Nevertheless, the authors identify pervasive methodological weaknesses: reuse of the same image subsets for training and testing, class-imbalance, limited external validation and scarce clinical studies. They call for standardized, annotated multimodal repositories and rigorous cross-dataset evaluation to boost generalizability and accelerate safe clinical adoption of CAD tools.

Extending this discussion on AI applications and the call for multimodal data, Hansun et al. (2025) systematically assess the burgeoning literature on artificial-intelligence (AI) applications for tuberculosis detection, addressing the persistent need for rapid, accurate screening tools that complement conventional diagnostics. Adhering to PRISMA-2020 and registered on PROSPERO, the authors searched Scopus, PubMed and ACM Digital Library to July 2023, screening 1146 records and retaining 152 peer-reviewed studies that applied machine- or deep-learning algorithms to radiographic, biochemical or physiological data. QUADAS-2 appraisal revealed generally low

applicability concerns but notable risks of bias in patient selection and reference standards. Convolutional-neural-network models, particularly VGG-16, ResNet-50 and DenseNet-121, dominated the field, often augmented by transfer learning; pooled performance across studies was strong (mean accuracy ≈ 92%, AUC ≈ 93%, sensitivity ≈ 93%, specificity ≈ 92%). Radiographic biomarkers and deep learning outperformed alternative modalities and algorithms, yet only one study explored domain-shift robustness. The review underscores AI's diagnostic promise while urging real-world validation, multimodal integration and rigorous external testing to ensure generalizable clinical impact [30].

Complementing these findings on AI accuracy and validation needs, Zhan et al. (2023) [31] undertake a systematic review and meta-analysis to clarify how accurately artificial-intelligence techniques interpret medical images for pulmonary tuberculosis (PTB) detection, an urgent need given persistent diagnostic gaps in resource-limited settings. Scouring MEDLINE and Embase to November 2022, the authors screened 3,987 records and included 61 eligible studies, 23 clinical evaluations and 38 model-development reports encompassing 124,959 participants. Using QUADAS-2 for quality appraisal and a bivariate random-effects model, they derived pooled sensitivities and specificities of 91 % and 65 % in clinical trials and 94 % and 95 % in development studies, respectively. These findings confirm that AI-based software can equal or surpass human readers, yet highlight heterogeneity in study design and reporting. The authors therefore call for standardized, multicenter trials and clearer AI-specific reporting guidelines to translate promising accuracy into dependable clinical tools.

In another study addressing localization challenges within TB imaging, Feyisa et al. (2023) [32] present a systematic review that foregrounds weakly supervised approaches for localizing radiographic manifestations of pulmonary tuberculosis (PTB) on chest X-rays, a task crucial for automated, clinician-interpretable diagnosis yet hampered by the scarcity of pixel-level annotations. Guided by PRISMA methodology, the authors queried six major databases for studies published between 2017 and 2023, ultimately distilling 35 eligible papers. They catalogue commonly used public datasets, describe prevailing weak-localization pipelines such as class activation mapping, attention-based CNNs and multiple-instance learning and analyze their ability to highlight cavitation, consolidation and other hallmark lesions with minimal labelled data. The review reveals promising qualitative localization but notes inconsistent evaluation metrics, limited dataset diversity and challenges in differentiating overlapping thoracic pathologies. Feyisa and colleagues conclude that future research should standardize benchmarks, incorporate multimodal clinical data and explore hybrid supervision to translate weak localization into reliable, deployable PTB screening tools.

Shifting to applications in clinical practice and multimodal data types, Pongsuwun et al. (2025) [33] systematically evaluate recent machine-learning applications for pulmonary-tuberculosis (PTB) diagnosis and their relevance to nursing practice. Adhering to PRISMA guidance, the authors searched nine biomedical databases for 2019–2024 studies, screening 734 records and retaining 12 original investigations after quality appraisal with JBI tools. Five diagnostic data types emerged—chest radiographs, computed-tomography scans, sputum-smear images, exhaled-breath spectra and demographic–clinical variables—reflecting a shift toward multimodal evidence integration. Among thirteen algorithms, convolutional-neural networks predominated because of their superior performance on imaging, while support-vector machines and ensemble trees showed utility for non-image inputs. Reported accuracies were consistently high, underscoring machine learning's capacity to expedite early PTB detection, particularly in resource-constrained settings. The authors conclude that incorporating AI-driven tools into nursing workflows could enhance assessment speed, reduce transmission and improve outcomes, but emphasize the need for context-specific implementation research and robust validation.

Broadening the scope to diverse lung diseases beyond just TB, Iqbal et al., (2024) present a broad systematic review of deep-learning methods for chest-radiograph interpretation across diverse lung diseases, aiming to synthesize technological trends rather than focus on a single pathology. Screening literature up to November 2024, the authors shortlisted 11 primary studies that leveraged convolutional architectures—often using transfer learning—on publicly available CXR datasets. Reported performance varies by task: EfficientNet and CNN-ELM reached 98–99 % accuracy for COVID-19, VGG19-based pipelines exceeded 96 % for multiclass lung-disease differentiation, and MobileNet V2 achieved 94 % on the NIH CXR-14 corpus. Methodologically, most studies employ extensive data augmentation and class-imbalance handling but seldom undertake external validation or

clinical-workflow assessment. Iqbal and colleagues conclude that while deep learning affords impressive diagnostic precision at low imaging cost, future research must harmonize dataset standards, incorporate radiologist comparison benchmarks and explore explainability to foster clinical trust and adoption [34].

In a related exploration of AI across airway disorders including TB, Koul et al., (2023) systematically map how artificial-intelligence methods are being harnessed to detect and classify a broad spectrum of airway disorders—ranging from cystic fibrosis and emphysema to tuberculosis and COVID-19—whose global burden exceeds three million deaths annually. Searching six major databases for English-language studies published between 2010 and 2022 and filtering via PRISMA criteria, the authors retained 155 peer-reviewed papers that applied machine- or deep-learning algorithms to imaging and clinical data. Convolutional and hybrid neural networks dominated, with reported diagnostic accuracies often surpassing traditional spirometry or plethysmography, especially for lung cancer and pulmonary embolism. Nevertheless, the review underscores persistent hurdles: scarce, imbalanced datasets; variability in evaluation metrics; and limited external validation. The authors advocate for standardized data curation, multicenter benchmarking and clinician-oriented interpretability to translate AI's promising accuracy into reliable, real-world airway-disease screening tools [35].

Continuing with evaluations of ML and DL specifically for TB on CXR, Hansun et al. (2023) [36] conduct a timely systematic literature review to evaluate how machine-learning (ML) and deep-learning (DL) algorithms support tuberculosis detection on chest X-rays (CXR). Motivated by TB's continued global burden and the variability of human radiograph interpretation, the authors searched Scopus, PubMed and IEEE databases, screening 309 records and retaining 47 primary studies that met PRISMA criteria. They extracted methodological details, performance metrics and risk-of-bias information, and complemented narrative synthesis with a meta-analysis of ten studies providing confusion matrices. Convolutional neural networks dominated the field, with ResNet-50, VGG-16/19 and AlexNet most frequent; ML techniques were chiefly support-vector machines, k-nearest neighbors and random forests. Pooled results indicated excellent diagnostic capacity (sensitivity ≈ 0.99, specificity ≈ 0.98), though most models relied on small public datasets and showed unclear reference-standard or timing bias. The authors conclude that both ML and DL hold high potential for automated TB screening, recommending larger, well-curated multimodal datasets and clearer methodological reporting.

Finally, focusing on CT-based DL for PTB to tie back to imaging modalities, Zhang et al. (2025) systematically examine the burgeoning use of deep-learning (DL) algorithms in computed-tomography (CT)–based diagnosis of pulmonary tuberculosis (PTB). Guided by PRISMA standards, the authors searched PubMed and Web of Science, screened 1,643 records and retained seven primary studies that met predefined inclusion criteria. They extracted performance indices—including accuracy, precision, recall, F1-score and AUC—and appraised study quality with QUADAS-2. Across the reviewed literature, convolutional and three-dimensional neural networks achieved impressive diagnostic yields (AUC up to 0.98), yet common limitations emerged: small, heterogeneous datasets, restricted external validation, and opaque model decision pathways. The review underscores the need for larger, multimodal cohorts, interpretable architectures, explicit ethical frameworks and prospective clinical trials to solidify DL's translational impact. Overall, the study positions CT-centric DL as a promising, though not yet mature, adjunct for accelerating and standardizing PTB detection in resource-constrained settings [37].

Table 1 presents all information extracted from the selected articles.

Table 1. Review of the systematic reviews that were included.

Lead author	Year	Country	Total of studies included	Databases searched	Quality assessment	Area of focus (objective)	Reporting method	Missing Data Discussion	Fund
Barbara Nansamba [26]	2025	Uganda	31	PubMed, Scopus, Web of Science	NA	Prediction, Categorization, Discovery	PRISMA	1	No
KC Santosh [27]	2022	United States	54	PubMed, Web of Science	NA	Categorization	PRISMA	No	No
Zhi-Lin Han [28]	2025	China	21	PubMed, Embase, Web of Science, Cochrane Library	QUADAS-2	Categorization	PRISMA-DTA	No	Yes
Mustapha Oloko-Oba [29]	2022	South Africa	62	PubMed, Scopus, IEEE Xplore, Web of Science	NA	Categorization	PRISMA	No	No
Seng Hansun [30]	2025	Australia	152	Scopus, PubMed, Association for Computing Machinery (ACM) Digital Library	QUADAS-2	Categorization	PRISMA	No	No
Yuejuan Zhan [31]	2023	China	61	MEDLINE, Embase	QUADAS-2	Categorization	PRISMA	No	Yes
Degaga Wolde Feyisa [32]	2023	Ethiopia	35	Google Scholar, PubMed, Science Direct, IEEE Xplore, MDPI, Springer Link	NA	Discovery	PRISMA	No	No
Kewalin Pongsuwun [33]	2025	Thailand	12	Scopus, PubMed, Medline, ScienceDirect, CINAHL Plus with Full Text, Clinical Key, Ovid, EMBASE, Web of Science	Joanna Briggs Institute (JBI) critical appraisal tools	Categorization	PRISMA	No	No
Hammad Iqbal [34]	2024	New Zealand	11	MDPI, Elsevier, IEEE, Springer, Google Scholar	NA	Categorization	PRISMA	No	Yes
Apeksha Koul [35]	2023	India	155	Science Direct, Google Scholar, Scopus, Web of Science, PubMed, EMBASE	NA	Categorization	PRISMA	No	No
Seng Hansun [36]	2023	Australia	47	Scopus, PubMed, IEEE	QUADAS-2	Categorization	PRISMA	No	No
Fei Zhang [37]	2025	China	7	PubMed, Web of Science	QUADAS-2	Categorization	PRISMA	No	Yes

3. Methodology

We carried out this review following the guidelines from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). The process of performing this task is illustrated in Figure 1.

3.1. *Literature search*

This review adhered to the formal protocol outlined below.

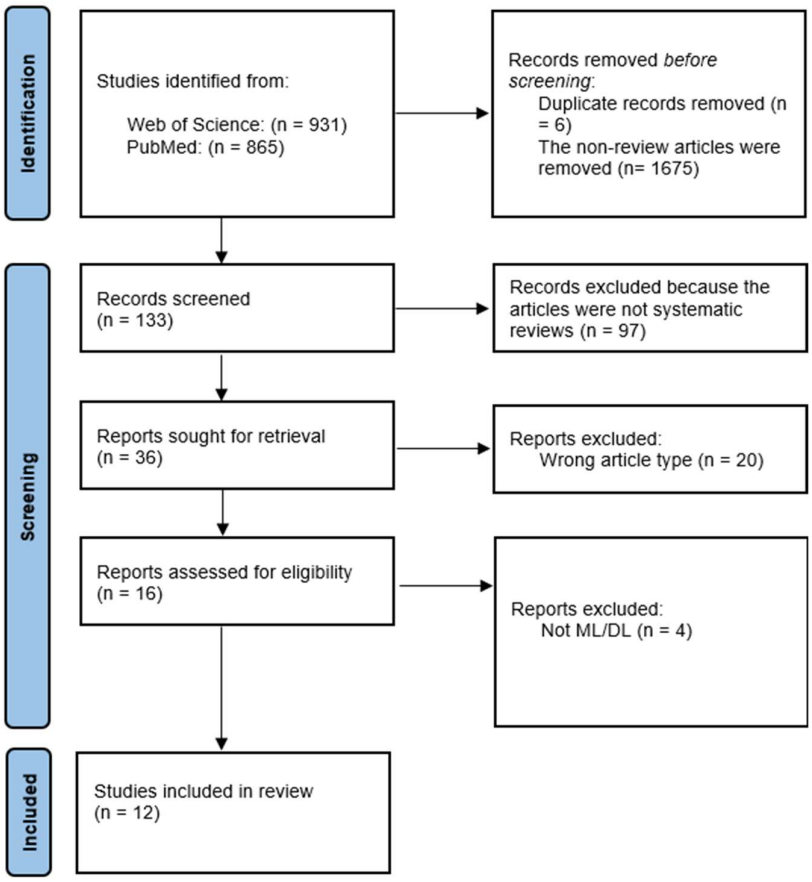

Figure 1. PRISMA process diagram for publication selection.

3.2. *Study Selection*

Systematic literature reviews and systematic reviews documenting the use of AI (machine learning, deep learning, computer-Aided Design) in TB care across any nation, authored in English and published in peer-reviewed journals from 1 January 2020 to 30 June 2025, were included. Investigations were performed in PubMed and Web of Science.

We used the following search query to retrieve data from the Web of Science database:

T S= ("tuberculosis" OR "TB") AND TS = ("artificial intelligence" OR "machine learning" OR "deep learning" OR "ML" OR "DL" OR "AI") AND TS = ("x-ray" OR "chest" OR "CT")

the following search query to retrieve data from the PubMed database:

("tuberculosis"[All Fields] OR "TB"[All Fields]) AND ("artificial intelligence"[All Fields] OR "machine learning"[All Fields] OR "deep learning"[All Fields] OR "ML"[All Fields] OR "DL"[All Fields] OR "AI"[All Fields]) AND ("x-ray"[All Fields] OR "chest"[All Fields] OR "CT"[All Fields])

Two researchers handled the initial assessment of the studies across two phases. To start, two reviewers separately checked the titles, keywords, and abstracts of all

the gathered publications to see if they seemed relevant. Then, in the next phase, two reviewers independently went through the full texts of the picked publications, guided by the inclusion and exclusion criteria along with the study's aims.

3.3. Data extraction

The process of data extraction from systematic reviews was carried out in two main stages. First, two up-to-date checklists were reviewed to identify the relevant evaluation criteria for the reviews [38, 39]. Next, several SLRs were randomly selected and analyzed to pinpoint the most frequently reported details, which informed the design of a data extraction template aimed at promoting consistency, accuracy, and completeness. Once these preparatory steps were completed, the actual extraction of data began, with each SLR analyzed for basic descriptive statistics, approaches to quality assessment, and methods of reporting. Finally, each review was categorized based on its primary focus into one of three groups: categorization (grouping data into clusters), prediction (using historical data to forecast outcomes), or discovery (examining data structures for new insights).

3.4. Data extraction

Key statistics, including data sources, accuracy, AUC, specificity, and sensitivity, were individually extracted from each study presented within the selected systematic reviews.

These performance metrics are frequently used to evaluate categorization techniques, and are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

Accuracy, in predictive modeling, measures how closely a model's outputs match the real, ground-truth values in a dataset. For classification tasks, it's a broad performance gauge showing how often the model gets predictions right across all classes calculated as the number of correct predictions divided by the total predictions. It treats all classes as equally important. While it's straightforward and easy to grasp, use it carefully with imbalanced datasets, where uneven class distributions can inflate scores and hide the model's real ability to distinguish between classes.

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

Sensitivity, also known as recall or the true positive rate, measures how well a model spots actual positives in a dataset it's the proportion of real positive cases correctly labeled as such (calculated as true positives divided by true positives plus false negatives). It's vital in high-stakes areas like medical diagnosis, fraud detection, or fault spotting, where missing a positive can be costly, as high sensitivity cuts down on false negatives. For a well-rounded evaluation, combine it with metrics like specificity or precision to balance catching positives without too many errors.

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

Specificity, or the true negative rate, measures how well a model correctly identifies actual negatives in a dataset—it's the proportion of real negatives labeled as such (calculated as true negatives divided by true negatives plus false positives). It's crucial in scenarios like medical screenings or security where false positives lead to unnecessary costs or risks, as high specificity reduces false alarms. For a balanced view, pair it with sensitivity to weigh avoiding false positives against missing true ones.

AUC, or Area Under the Curve, is a straightforward performance metric pulled from the ROC curve. This curve plots the true positive rate (how well it spots actual positives) against the false positive rate across different decision cutoffs. Essentially, AUC gauges a model's overall skill at distinguishing between positive and negative cases, without tying it to any particular threshold. It's super useful for handling imbalanced data or scenarios where the emphasis on catching true positives versus avoiding false alarms can shift. A higher AUC shows the model reliably assigns stronger predictions to real positives compared to negatives. Thanks to its flexibility and independence from thresholds, it's seen as a reliable, all-in-one tool for evaluating binary classification models in both studies and practical applications.

A true positive (TP) occurs when the predicted value matches the actual value of a data point and that value is positive. For example, in the context of diagnosing tuberculosis, if an individual indeed has the disease (1) and

the model correctly predicts the presence of tuberculosis (1), this instance is classified as a true positive in the given problem.

A true negative (TN) occurs when the predicted outcome is negative and correctly corresponds to the actual negative class of the data point. For example, in the context of tuberculosis diagnosis, if an individual does not have the disease (0) and the model accurately classifies the case as disease-free (0), this instance is categorized as a true negative in the given problem.

A false positive (FP) occurs when the model predicts a positive class for a data point, while in reality the actual class is negative. The term "positive" refers to the model's prediction of the positive class, whereas "false" indicates that this prediction is incorrect. For example, in the context of tuberculosis diagnosis, if an individual does not have the disease (0) but the model incorrectly classifies the case as having tuberculosis (1), this instance is considered a false positive in the given problem.

A false negative (FN) occurs when a data point truly belongs to the positive class (1), but the model incorrectly predicts it as belonging to the negative class (0). The term "false" indicates that the model's prediction is incorrect, while "negative" reflects the predicted classification as the negative class. For example, in the context of tuberculosis diagnosis, a false negative arises when an individual actually has tuberculosis (1) but the model classifies the case as disease-free (0).

In addition, we collected data on the validation methods used and how missing information was handled in the studies. We also documented the different AI techniques that were applied, and for each systematic review included in our analysis, we calculated the number of primary studies that described each specific type of AI algorithm.

4. Results

As shown in Figure 1 the systematic review process began with the identification of 1,796 potentially relevant articles through comprehensive searches in PubMed and Web of Science, ensuring a broad and thorough capture of pertinent literature from these two key databases. Following the removal of non-review articles and duplicates, 133 articles advanced for further consideration. Subsequently, after eliminating those that were not systematic reviews, 36 unique systematic reviews (SRs) remained. An initial assessment led to the exclusion of 20 SRs, leaving 16 for more detailed evaluation. Finally, after a thorough final review against all inclusion and exclusion criteria, 12 SRs were selected for the full analysis. Collectively, these 12 SRs incorporated a total of 648 primary studies, with individual reviews varying in scope from as few as 7 studies to as many as 155 (Table 1).

Table 2. Quantity of included studies detailing accuracy, AUC, sensitivity, and specificity in the examined systematic literature reviews (SLRs)

		No. of studies that reported			
Lead author name & year	No. of studies	Accuracy	AUC	Sensitivity	Specificity
Barbara Nansamba,2025 [26]	31	22	19	8	8
KC Santosh,2022 [27]	54	21	25	19	25
Zhi-Lin Han,2025 [28]	21	10	19	21	21
Mustapha Oloko-Oba,2022 [29]	62	43	31	23	27
Seng Hansun,2025 [30]	152	128	87	112	77
Yuejuan Zhan,2023 [31]	61	35	48	44	43
Degaga Wolde Feyisa,2023 [32]	35	2	0	0	0
Kewalin Pongsuwun,2025 [33]	12	2	0	1	1
Hammad Iqbal,2024 [34]	11	11	0	8	0
Apeksha Koul,2023 [35]	155	36	15	26	27
Seng Hansun,2023 [36]	47	35	32	27	23
Fei Zhang,2025 [37]	7	6	4	5	5

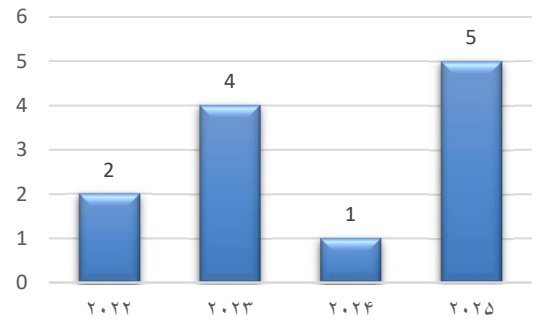

Figure 2. Quantity of systematic reviews used in this analysis published throughout the years.

Characterizing these twelve systematic reviews unveiled notable temporal and geographic trends, highlighting the evolving and international nature of the research landscape in this field. In terms of temporal distribution, none of the reviews predated 2022, underscoring the relatively recent surge in systematic syntheses on the topic; specifically, two reviews (17%) were published in 2022, four (33%) in 2023, one (8%) in 2024, and a substantial five (42%) in 2025, reflecting an accelerating interest and output over these years (Figure 3).

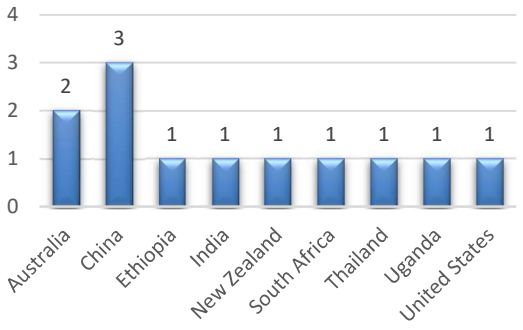

Figure 3. Count of systematic reviews included in this review, classified by country of publication.

Geographically, the reviews originated from nine diverse countries spanning four continents, demonstrating a broad global engagement; they were predominantly led by China with three reviews (25%) and Australia with two (17%), while single contributions (8% each) came from Ethiopia, India, New Zealand, South Africa, Thailand, Uganda, and the United States (Table 1, Figure 3).

Methodologically, adherence to established reporting and validation standards across the twelve systematic reviews (SRs) showed considerable variation, reflecting diverse approaches to ensuring rigor and transparency in evidence synthesis. Notably, eleven SRs (92%) employed the PRISMA guideline to structure their reporting, while one (8%) utilized the specialized PRISMA-DTA framework, which is particularly suited for diagnostic test accuracy studies, thereby highlighting a general commitment to standardized protocols albeit with some adaptation to specific review needs (Table 1).

Reporting of performance metrics was similarly uneven, indicating potential gaps in comprehensive evaluation: accuracy was consistently reported in all twelve SRs (encompassing 351 studies; 54% of the total), AUC appeared in nine SRs (280 studies; 43%), sensitivity was documented in eleven SRs (294 studies; 45%), and specificity in ten SRs (257 studies; 40%), collectively suggesting that while core metrics like accuracy are widely prioritized, others such as specificity receive less uniform attention, which could influence the interpretability and comparability of findings across the reviews (Table 2).

In terms of validation practices across the aggregated 648 primary studies, 214 studies (33%) incorporated internal validation to assess model reliability within the development dataset, and a smaller subset of 29 studies (4%) included external validation to evaluate generalizability across independent datasets—each of these practices was evident in seven of the twelve SRs, underscoring a moderate but inconsistent emphasis on robust validation methods that are crucial for mitigating overfitting and enhancing clinical applicability (Table 3).

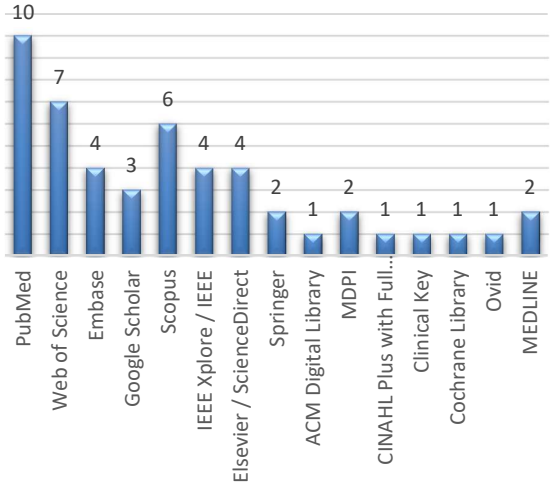

Figure 4. Count of systematic reviews included in this study, based on the databases they used to extract their included articles.

Search strategies across the twelve systematic reviews encompassed a total of 49 database-search instances distributed over 15 distinct platforms, illustrating a multifaceted approach to sourcing relevant literature and minimizing selection bias through varied and extensive querying. The most frequently utilized databases were PubMed (queried in 10 SRs), Web of Science (in 7), and Scopus (in 6), followed by Embase, IEEE Xplore/IEEE, and Elsevier/ScienceDirect each in four; Google Scholar in three; Springer Link, MDPI, and MEDLINE in two; and the ACM Digital Library, CINAHL Plus with Full Text, Clinical Key, Cochrane Library, and Ovid each in one SR, collectively highlighting a reliance on prominent biomedical and multidisciplinary repositories to ensure broad coverage, though with varying degrees of overlap and specificity depending on the review's focus (Table 1, Figure 4).

Regarding quality assessment, explicit reporting was evident in only seven SRs (58%), where five employed the QUADAS-2 tool for evaluating diagnostic accuracy studies and one utilized the Joanna Briggs Institute (JBI) tools for broader critical appraisal, while the remaining six SRs (42%) did not conduct or document any formal quality assessment, which could introduce risks of incorporating lower-quality evidence and underscores the need for more consistent methodological rigor in future syntheses (Table 1).

Finally, despite the critical role of addressing missing data to maintain analytical integrity and reduce bias, only a single primary study (0.15% of the 648 total) referenced the use of "null value filling" as a handling technique, albeit without providing further methodological details; the vast majority—647 studies (99.85%)—made no mention whatsoever of missing-data procedures, potentially overlooking a key source of uncertainty and highlighting an area for improvement in primary research reporting. Additionally, funding status indicated that four of the twelve SRs (33%) reported receiving financial support, which may have facilitated more comprehensive searches or analyses, while the remaining eight (67%) indicated no such funding, suggesting that a significant portion of this research was conducted without dedicated external backing (Table 1).

Table 3. Summary of included studies by data type and validation methods.

Lead author name & year	No. of studies	Type of Data	Internal Validation	External Validation
Barbara Nansamba,2025 [26]	31	Chest X-ray (CXR), Computed Tomography (CT) Scan, Immunological, Microbiology, Molecular, Biochemical, Clinical/Behavioral, Human exhaled breath, Demographic/Clinical Data	4	4
KC Santosh,2022 [27]	54	Chest X-ray (CXR)	28	2
Zhi-Lin Han,2025 [28]	21	Chest X-ray (CXR)	0	0
Mustapha Oloko-Oba,2022 [29]	62	Chest X-ray (CXR)	0	0
Seng Hansun,2025 [30]	152	Chest X-ray (CXR), Computed Tomography (CT) Scan, Biochemical, Physiological/Clinical	143	9
Yuejuan Zhan,2023 [31]	61	Chest X-ray (CXR), Computed Tomography (CT) Scan	29	9
Degaga Wolde Feyisa,2023 [32]	35	Chest X-ray (CXR)	0	0
Kewalin Pongsuwun,2025 [33]	12	Chest X-ray (CXR), Computed Tomography (CT) Scan, Microbiology, Human exhaled breath, Demographic/Clinical Data	0	0
Hammad Iqbal,2024 [34]	11	Chest X-ray (CXR)	2	0
Apeksha Koul,2023 [35]	155	Chest X-ray (CXR), Computed Tomography (CT) Scan, Biochemical, Physiological/Clinical	3	1
Seng Hansun,2023 [36]	47	Chest X-ray (CXR)	0	1
Fei Zhang,2025 [37]	7	Computed Tomography (CT) Scan	5	3

Across 12 systematic review articles examining 353 different algorithms and totaling 1,247 mentions, deep learning methods led with about 44% of mentions mostly convolutional neural networks (CNNs) like VGG-16, ResNet-50, and InceptionV3. Support vector machines appeared in roughly 5% of mentions, followed by random forests (≈2.5%) and decision trees (≈1%). Boosting techniques such as AdaBoost and XGBoost showed up in about 1.3%, while linear and logistic regression methods made up around 2%. K-nearest neighbors and linear

discriminant analysis were rare (under 1%). The remaining 44% of mentions covered a mix of niche or newer algorithms. Overall, deep learning dominates medical imaging research, but simpler and ensemble methods remain relevant for various needs like speed, clarity, or smaller datasets.

Table 4 shows that among the 15 most-cited algorithms across the 12 studies totaling 586 mentions (excluding the "Others" category) convolutional neural networks dominate with approximately 85% of these mentions, led by VGG-16 (64 mentions), ResNet-50 (63 mentions), InceptionV3 (48 mentions), Custom CNN (49 mentions), and DenseNet-121 (39 mentions). Classical methods include support vector machines (≈9.6%, 56 mentions) and random forests (≈5.3%, 31 mentions). Specialized tools like CAD4TB (42 mentions) and qXR (30 mentions) also appear, while the "Others" category (661 mentions, or ≈53% of the total) reflects a diverse range of niche or emerging approaches.

Across 12 systematic review articles examining 353 different algorithms and totaling 1,247 mentions, deep learning methods led with about 44% of mentions—mostly convolutional neural networks (CNNs) like VGG-16, ResNet-50, and InceptionV3. Support vector machines appeared in roughly 5% of mentions, followed by random forests (≈2.5%) and decision trees (≈1%). Boosting techniques such as AdaBoost and XGBoost showed up in about 1.3%, while linear and logistic regression methods made up around 2%. K-nearest neighbors and linear discriminant analysis were rare (under 1%). The remaining 44% of mentions covered a mix of niche or newer algorithms. Overall, deep learning dominates medical imaging research, but simpler and ensemble methods remain relevant for various needs like speed, clarity, or smaller datasets.

Table 4 shows that among the 15 most-cited algorithms across the 12 studies—totaling 586 mentions (excluding the "Others" category)—convolutional neural networks dominate with approximately 85% of these mentions, led by VGG-16 (64 mentions), ResNet-50 (63 mentions), InceptionV3 (48 mentions), Custom CNN (49 mentions), and DenseNet-121 (39 mentions). Classical methods include support vector machines (≈9.6%, 56 mentions) and random forests (≈5.3%, 31 mentions). Specialized tools like CAD4TB (42 mentions) and qXR (30 mentions) also appear, while the "Others" category (661 mentions, or ≈53% of the total) reflects a diverse range of niche or emerging approaches.

Table 4. Distribution of the 15 Most Frequent Deep Learning and Machine Learning Algorithms Across Studies.

	References											
Algorithms	[26]	[27]	[28]	[29]	[30]	[31]	[32]	[33]	[34]	[35]	[36]	[37]
VGG-16	0	4	0	8	37	5	0	0	1	1	8	0
ResNet-50	6	2	0	7	33	1	0	0	0	3	11	0
InceptionV3	2	5	0	12	18	3	0	0	1	1	6	0
CAD4TB	0	7	16	0	0	18	0	1	0	0	0	0
Custom CNN	0	0	0	0	39	0	0	0	1	0	9	0
DenseNet-121	2	2	0	9	19	1	0	0	1	0	5	0
Support Vector Machine (SVM)	2	1	0	0	33	3	0	4	0	8	5	0
VGG-19	0	3	0	7	16	1	0	0	2	1	7	0
AlexNet	0	1	0	8	10	1	0	0	1	1	6	0
CNN	2	4	0	6	0	4	0	5	0	18	0	0
Xception	2	0	0	1	13	2	0	0	0	0	4	0
qXR	0	2	11	0	0	8	0	1	0	8	0	0
GoogLeNet	0	2	0	4	9	1	0	0	0	0	5	0
Random Forest (RF)	4	0	0	0	17	0	0	2	0	6	2	0
InceptionResNetV2	1	2	0	2	9	1	0	0	0	0	2	0
Others	55	25	15	59	251	51	47	9	13	105	24	7

5. Discussion and Conclusion

The discovery of 12 systematic reviews published since 2022 that aggregate 648 primary studies and catalogue 353 distinct AI models most of them convolutional-neural-network solutions built around chest-X-ray data highlights the rapid and growing global commitment to harnessing machine-learning for tuberculosis diagnosis and management.

Key findings revealed:

- Chest radiography is the dominant data source.
- CNNs eclipse other algorithms.

- Severe scarcity of external validation.

5.1. Strength and weakness

Our study has several notable limitations, primarily stemming from inherent human errors in research documentation and reporting that are common across scientific literature. First, by relying solely on existing systematic reviews, we were constrained to the details those reviews elected to include, which could introduce human-induced oversights—such as incomplete nuances or misinterpretations—and potentially lead to an overestimation of distinct machine learning (ML) algorithms, as it remains unclear whether different reviews overlapped in their analysis of datasets due to variable reporting practices. Second, the exclusion of primary studies not captured in these systematic reviews may skew our overview of ML applications in healthcare, reflecting potential gaps in human-curated review processes that inadvertently omit relevant works. Third, limiting our search to English-language publications introduces a geographic selection bias, often a byproduct of human tendencies toward linguistic familiarity in global research aggregation. Fourth, publication bias is a persistent concern, as human decisions in the publishing process tend to favor positive or conclusive results, potentially resulting in an overly optimistic portrayal of ML model performance in our findings. Finally, variations in terminology and reporting across studies arising from human inconsistencies in classification and documentation could mean that certain techniques were overlooked or misclassified, further complicating a comprehensive synthesis.

Additionally, we evaluated the performance of our ML and deep DL models using standard metrics for classification tasks, including accuracy, area under the curve (AUC), sensitivity, and specificity. However, for assessing predictive models on quantitative variables, alternative metrics would be more appropriate, highlighting the need for tailored evaluation approaches depending on the model's objectives.

Furthermore, some of the reviewed articles encompassed other pulmonary diseases beyond tuberculosis. In line with our inclusive methodology, we did not exclude these instances and incorporated them into our review and overall counts to maintain a broader perspective on related applications.

5.2. Conclusion

This umbrella review synthesizes the burgeoning yet methodologically heterogeneous field of artificial intelligence (AI) applications in tuberculosis (TB) research, drawing from 12 systematic reviews encompassing 648 primary studies published between 2022 and mid-2025. Predominantly, deep-learning architectures, particularly convolutional neural network (CNN) variants such as VGG-16, ResNet-50, and InceptionV3, have taken center stage in diagnostic tasks, often yielding remarkable internal validation accuracies that suggest substantial potential for enhancing TB detection and management. These models frequently leverage imaging data, with chest X-rays serving as the primary modality, and demonstrate high performance in controlled settings, highlighting AI's promise as a scalable tool for addressing the global TB epidemic, which claims over 1.5 million lives annually according to World Health Organization estimates. However, the review uncovers significant shortcomings that undermine the reliability and generalizability of these findings. Notably, only about one-third of the underlying primary studies incorporate any form of internal validation, while a scant 4% extend their evaluations to independent external cohorts or direct comparisons with human radiologists or clinicians. This paucity of rigorous testing raises concerns about overfitting and real-world applicability, especially in diverse clinical environments. Furthermore, reporting of performance metrics is inconsistent and incomplete; for instance, accuracy is detailed in 54% of studies, area under the curve (AUC) in 43%, sensitivity in 45%, and specificity in 40%, making cross-study comparisons challenging. Quality appraisal tools, essential for assessing methodological robustness, are employed in just over half of the reviews, and strategies for handling missing data are virtually nonexistent, potentially introducing biases. The evidence base is further compromised by language-restricted literature searches, which may exclude valuable non-English contributions, and a likely publication bias favoring positive results, skewing the overall narrative toward optimism without sufficient scrutiny. Despite these limitations, several positive trajectories offer hope for advancement. The geographic scope of research has broadened beyond high-income countries, increasingly incorporating perspectives from low- and middle-income settings where TB prevalence is highest. Innovations such as multimodal data integration combining computed tomography (CT) scans, microbiological assays, and

demographic variables with traditional X-rays are emerging, enriching model inputs and potentially improving diagnostic precision. Additionally, the adoption of explainable AI techniques is on the rise, fostering greater transparency and trust in algorithmic decisions, which is crucial for clinical adoption.

Authors' Contributions

All authors equally contributed to this study.

Declaration

None.

Transparency Statement

None.

Acknowledgments

None.

Declaration of Interest

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

Funding

According to the authors, this article has no financial support.

Ethical Considerations

Not applicable.

References

[1] Who, "2024 Global tuberculosis report," 2024.
[2] S. K. Sharma and A. Mohan, "Tuberculosis: From an incurable scourge to a curable disease - journey over a millennium," *Indian Journal of Medical Research,* vol. 137, pp. 455-493, 2013, doi: 10.23640554.
[3] Who, "Tuberculosis Report," vol. XLIX, 2020.
[4] S. I. Nafisah and G. Muhammad, "Tuberculosis detection in chest radiograph using convolutional neural network architecture and explainable artificial intelligence," *Neural Computing and Applications,* vol. 36, pp. 111-131, 2024, doi: 10.1007/s00521-022-07258-6.
[5] M. F. Alcantara *et al.*, "Improving tuberculosis diagnostics using deep learning and mobile health technologies among resource-poor communities in Perú," *Smart Health,* vol. 1-2, pp. 66-76, 2017, doi: 10.1016/j.smhl.2017.04.003.
[6] L. Richeldi, "An update on the diagnosis of tuberculosis infection," *American Journal of Respiratory and Critical Care Medicine,* vol. 174, pp. 736-742, 2006, doi: 10.1164/rccm.200509-1516PP.
[7] Y. Luo *et al.*, "Development of diagnostic algorithm using machine learning for distinguishing between active tuberculosis and latent tuberculosis infection," *BMC Infectious Diseases,* vol. 22, p. 965, 2022, doi: 10.1186/s12879-022-07954-7.
[8] G. M. M. Alshmrani, Q. Ni, R. Jiang, H. Pervaiz, and N. M. Elshennawy, "A deep learning architecture for multi-class lung diseases classification using chest X-ray (CXR) images," *Alexandria Engineering Journal,* vol. 64, pp. 923-935, 2023. [Online]. Available: https://doi.org/10.1016/j.aej.2022.10.053.
[9] V. Palani, T. Thanarajan, A. Krishnamurthy, and S. Rajendran, "Deep Learning Based Compression with Classification Model on CMOS Image Sensors," *Trait du Signal,* vol. 40, pp. 1163-1170, 2023, doi: 10.18280/ts.400332.
[10] T. Rahman *et al.*, "TB-CXRNet: Tuberculosis and Drug-Resistant Tuberculosis Detection Technique Using Chest X-ray Images," *Cognitive Computation,* vol. 16, pp. 1393-1412, 2024, doi: 10.1007/s12559-024-10259-3.
[11] V. Ravi, V. Acharya, and M. Alazab, "A multichannel EfficientNet deep learning-based stacking ensemble approach for lung disease detection using chest X-ray images," *Cluster Computing,* vol. 26, pp. 1181-1203, 2023, doi: 10.1007/s10586-022-03664-6.
[12] S. K. Jain *et al.*, "Advanced imaging tools for childhood tuberculosis: potential applications and research needs," *Lancet Infectious Diseases,* vol. 20, pp. e289-e297, 2020, doi: 10.1016/S1473-3099(20)30177-8.
[13] E. Skoura, A. Zumla, and J. Bomanji, "Imaging in tuberculosis," *International Journal of Infectious Diseases,* vol. 32, pp. 87-93, 2015, doi: 10.1016/j.ijid.2014.12.007.
[14] T. K. Kim *et al.*, "Deep Learning Method for Automated Classification of Anteroposterior and Posteroanterior Chest Radiographs," *Journal of Digital Imaging,* vol. 32, pp. 925-930, 2019, doi: 10.1007/s10278-019-00208-0.
[15] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology,* vol. 284, pp. 574-582, 2017, doi: 10.1148/radiol.2017162326.
[16] P. H. Yi *et al.*, "Automated semantic labeling of pediatric musculoskeletal radiographs using deep learning," *Pediatric Radiology,* vol. 49, pp. 1066-1070, 2019, doi: 10.1007/s00247-019-04408-2.
[17] K. Dimopoulos *et al.*, "Cardiothoracic ratio from postero-anterior chest radiographs: A simple, reproducible and independent marker of disease severity and outcome in adults with congenital heart disease," *International Journal of Cardiology,* vol. 166, pp. 453-457, 2013, doi: 10.1016/j.ijcard.2011.10.125.
[18] D. Rai, R. Kirti, S. Kumar, S. Karmakar, and S. Thakur, "Radiological difference between new sputum-positive and sputum-negative pulmonary tuberculosis," *Journal of Family Medicine and Primary Care,* vol. 8, p. 2810, 2019, doi: 10.4103/jfmpc.jfmpc_652_19.
[19] J. Z. Cheng *et al.*, "Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans," *Scientific Reports,* vol. 6, p. 24454, 2016, doi: 10.1038/srep24454.

[20] K. Li *et al.*, "Assessing the predictive accuracy of lung cancer, metastases, and benign lesions using an artificial intelligence-driven computer aided diagnosis system," *Quantitative Imaging in Medicine and Surgery,* vol. 11, pp. 3629-3642, 2021, doi: 10.21037/qims-20-1314.

[21] M. Badawy, N. Ramadan, and H. A. Hefny, "Healthcare predictive analytics using machine learning and deep learning techniques: a survey," *Journal of Electrical Systems and Information Technology,* vol. 10, p. 40, 2023, doi: 10.1186/s43067-023-00108-y.

[22] B. Nithya and V. Ilango, "Predictive analytics in health care using machine learning tools and techniques," *2017 International Conference on Intelligent Computing, Control Systems, IEEE,* pp. 492-499, 2017, doi: 10.1109/ICCONS.2017.8250771.

[23] P. L. Bokonda, K. Ouazzani-Touhami, and N. Souissi, "Predictive analysis using machine learning: Review of trends and methods," *2020 International Symposium on Advanced Electrical and Communication Technologies, IEEE,* pp. 1-6, 2020, doi: 10.1109/ISAECT50560.2020.9523703.

[24] M. Breuninger *et al.*, "Diagnostic Accuracy of Computer-Aided Detection of Pulmonary Tuberculosis in Chest Radiographs: A Validation Study from Sub-Saharan Africa," *PLOS ONE,* vol. 9, p. e106381, 2014, doi: 10.1371/journal.pone.0106381.

[25] S. C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines," *Npj Digital Medicine,* vol. 3, p. 136, 2020, doi: 10.1038/s41746-020-00341-z.

[26] B. Nansamba, J. Nakatumba-Nabende, A. Katumba, and D. P. Kateete, "A Systematic Review on Application of Multimodal Learning and Explainable AI in Tuberculosis Detection," *IEEE Access,* vol. 13, pp. 62198-62221, 2025, doi: 10.1109/ACCESS.2025.3558878.

[27] K. C. Santosh, S. Allu, S. Rajaraman, and S. Antani, "Advances in Deep Learning for Tuberculosis Screening using Chest X-rays: The Last 5 Years Review," *Journal of Medical Systems,* vol. 46, 2022, doi: 10.1007/s10916-022-01870-8.

[28] Z. Han, Y. Zhang, J. Li, S. Gao, W. Liu, and W. Yang, "A systematic review and meta-analysis of artificial intelligence software for tuberculosis diagnosis using chest X-ray imaging," *Journal of Thoracic Disease,* vol. 17, pp. 3223-3237, 2025, doi: 10.21037/jtd-2025-604.

[29] M. Oloko-oba and S. Viriri, "A Systematic Review of Deep Learning Techniques for Tuberculosis Detection From Chest Radiograph," *Frontiers in Medicine,* vol. 9, p. 830515, 2022, doi: 10.3389/fmed.2022.830515.

[30] S. Hansun *et al.*, "Diagnostic Performance of Artificial Intelligence-Based Methods for Tuberculosis Detection: Systematic Review," *Journal of Medical Internet Research,* vol. 27, p. e69068, 2025, doi: 10.2196/69068.

[31] Y. Zhan, Y. Wang, W. Zhang, B. Ying, and C. Wang, "Diagnostic Accuracy of the Artificial Intelligence Methods in Medical Imaging for Pulmonary Tuberculosis: A Systematic Review and Meta-Analysis," *Journal of Clinical Medicine,* vol. 12, 2023, doi: 10.3390/jcm12010303.

[32] D. W. Feyisa, Y. M. Ayano, T. G. Debelee, and F. Schwenker, "Weak Localization of Radiographic Manifestations in Pulmonary Tuberculosis from Chest X-ray: A Systematic Review," *Sensors,* vol. 23, p. 6781, 2023, doi: 10.3390/s23156781.

[33] K. Pongsuwun, W. Puwarawuttipanit, S. Nguantad, B. Samart, U. Pollayut, and P. T. T. Phuong, "A Systematic Review of the Accuracy of Machine Learning Models for Diagnosing Pulmonary Tuberculosis: Implications for Nursing Practice and Implementation," *Nursing and Health Sciences,* vol. 27, pp. 1-10, 2025, doi: 10.1111/nhs.70077.

[34] H. Iqbal, A. Khan, N. Nepal, F. Khan, and Y. K. Moon, "Deep Learning Approaches for Chest Radiograph Interpretation: A Systematic Review," *Electronics,* vol. 13, p. 4688, 2024, doi: 10.3390/electronics13234688.

[35] A. Koul, R. K. Bawa, and Y. Kumar, "Artificial Intelligence Techniques to Predict the Airway Disorders Illness: A Systematic Review," *Health Informatics Journal,* vol. 30, 2023, doi: 10.1007/s11831-022-09818-4.

[36] S. Hansun, A. Argha, S. T. Liaw, B. G. Celler, and G. B. Marks, "Machine and Deep Learning for Tuberculosis Detection on Chest X-Rays: Systematic Literature Review," *Journal of Medical Internet Research,* vol. 25, p. e43154, 2023, doi: 10.2196/43154.

[37] F. Zhang, H. Han, M. Li, T. Tian, G. Zhang, and Z. Yang, "Revolutionizing diagnosis of pulmonary Mycobacterium tuberculosis based on CT: a systematic review of imaging analysis through deep learning," *Frontiers in Microbiology,* vol. 15, p. 1510026, 2024, doi: 10.3389/fmicb.2024.1510026.

[38] T. J. Loftus, P. J. Tighe, T. Ozrazgat-Baslanti, J. P. Davis, M. M. Ruppert, and Y. Ren, "Ideal algorithms in healthcare: Explainable, dynamic, precise, autonomous, fair, and reproducible," *PLOS Digital Health,* vol. 1, p. e0000006, 2022, doi: 10.1371/journal.pdig.0000006.

[39] W. V. Padula, N. Kreif, D. J. Vanness, B. Adamson, J. D. Rueda, and F. Felizzi, "Machine Learning Methods in Health Economics and Outcomes Research-The PALISADE Checklist: A Good Practices Report of an ISPOR Task Force," *Value in Health,* vol. 25, pp. 1063-1080, 2022, doi: 10.1016/j.jval.2022.03.022.