TRENDS IN SCIENCES 2026; 23(1): 11390
https://doi.org/10.48048/tis.2026.11390
REVIEW ARTICLE

Machine Learning for Biomarker-Based Tuberculosis Diagnosis: A Systematic Review and Meta-Analysis

Roy Novri Ramadhan[1], Merita Arini[2,*] and Danendra Rakha Putra Respati[3]

[1]Master of Hospital Administration, Postgraduate Program, Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia
[2]Undergraduate Program of Medicine, Faculty of Medicine, Universitas Airlangga, Surabaya, Indonesia
[3]Undergraduate Program of Medicine, Faculty of Medicine, Universitas Diponegoro, Semarang, Indonesia

([*]Corresponding author's e-mail: merita.arini@umy.ac.id)

Received: 26 July 2025, Revised: 5 September 2025, Accepted: 12 September 2025, Published: 5 October 2025

Abstract

Tuberculosis (TB) is a serious worldwide health concern, demanding reliable and efficient diagnostic techniques. Machine learning (ML) techniques have emerged as viable ways to enhance tuberculosis detection through the analysis of complicated biomarker data. However, the comparative effectiveness of various ML models is unclear, emphasizing the need for a thorough assessment. This study evaluates the diagnostic accuracy of ML algorithms for tuberculosis detection, outlining their strengths, limitations, and clinical applications. A comprehensive search of 6 databases identified studies using biomarker-based ML approaches for TB diagnosis. A 2-level mixed-effects logistic regression model was used to assess pooled sensitivity, specificity, and AUC. The RoB 2.0 tool was used to evaluate quality, whereas RevMan 5.4 was used for meta-analysis. Among the algorithms tested, the Probabilistic Neural Network (PNN) had the greatest pooled sensitivity (96.1%), specificity (89.9%), and AUC (0.94). Decision Tree (DT) models have great sensitivity (95.2%), but poor specificity (58.7%). Naïve Bayes (NB) had the greatest specificity (91.5%) and sensitivity (79.6%), while Random Forest (RF) and Support Vector Machine (SVM) performed similarly, with sensitivities of 83.7% and 84.1%, and accuracies of 82.5% and 83.8%, respectively. Logistic regression (LR) had the lowest sensitivity (54.3%) and accuracy (73.1%). ML algorithms have great potential for improving TB diagnosis using biomarker data, with PNN appearing as the best-performing model in terms of sensitivity, specificity, and AUC. However, the availability and expense of biomarker testing may limit the use of such strategies in resource-constrained environments. The clinical context, diagnostic goals, and infrastructure capabilities should all be considered while selecting algorithms.

Keywords: Tuberculosis, Machine learning, Biomarker-based diagnosis, Early detection, Diagnostic accuracy, Communicable diseases, Hospital management, Artificial intelligence

Introduction

Tuberculosis (TB) is a significant worldwide health concern, especially in areas with high incidence rates such as Southeast Asia and Africa [1]. In 2023, about 10.8 million people contracted TB, and approximately 1.3 million succumbed to the disease, with countries such as India, Indonesia, and the Philippines being the most affected [1]. Sub-Saharan Africa contributes substantially to the worldwide burden of tuberculosis. The issue is exacerbated by the advent and proliferation of multidrug-resistant tuberculosis (MDR-TB), a severe variant of the illness resistant to primary therapies such as rifampicin and isoniazid. MDR-TB presents a significant issue in low- and middle-income countries (LMICs), where healthcare systems often lack the necessary infrastructure, resources, and modern technology for prompt and precise diagnosis [2]. Eight nations represent over 66% of worldwide TB cases, with Indonesia (9.2%) positioned second behind India. The Global TB Report 2022 indicates that Indonesia documented 969,000

tuberculosis infections in 2021, with 44% of these cases concentrated in its most densely populated areas, including East Java, West Java, and Central Java [3].

The concerning data underscore an urgent need for novel and efficient diagnostic instruments to address the TB pandemic. Traditional diagnostic techniques, including sputum microscopy, often exhibit sluggishness, diminished sensitivity, and an inability to identify drug resistance, resulting in postponed or insufficient therapy. This facilitates continuous transmission, extended morbidity, and elevated death rates. Chest X-ray (CXR) is often used as an adjunctive technique for detecting lung abnormalities indicative of tuberculosis, however its interpretation may be subjective and reliant on the radiologist's skill. Alternative non-sputum-based methodologies, such as interferon-gamma release assays (IGRA) and IP-10 tests, have been investigated to enhance tuberculosis detection, especially for latent tuberculosis infection; however, they encounter constraints regarding sensitivity, cost, and accessibility in certain environments [4-6].

Progress in molecular diagnostics and other technologies, such as machine learning (ML) and artificial intelligence (AI), provide intriguing solutions for these difficulties. Researchers have progressively used AI and ML into tuberculosis diagnoses. Machine learning algorithms have shown improved sensitivity and accuracy in identifying tuberculosis from chest X-ray pictures, providing more reliable and expedited diagnosis. Studies show that AI-driven approaches can surpass human diagnostic performance, especially in low-resource environments [7,8]. Machine learning's ability to analyze complex datasets, including radiographic images and biomarkers, has shown great promise in TB diagnostics [9].

Despite these advances, the full potential of ML remains underexplored, especially in regions where healthcare access is limited. While preliminary studies suggest that ML algorithms can offer cost-effective, automated diagnostic solutions, widespread adoption faces barriers such as the need for large-scale clinical validation, data standardization, and integration into existing healthcare systems [10]. Furthermore, there are concerns about the quality and diversity of training datasets, which must be representative of diverse populations to ensure robust and equitable performance. Nonetheless, the initial results are promising, indicating that ML could play a vital role in bridging the gap between traditional diagnostic methods and modern, scalable technologies [10,11]. The integration of ML with biomarker-based approaches and other molecular diagnostics could pave the way for next-generation TB diagnostic tools that are not only accurate but also accessible to underserved populations. As research continues to refine these technologies, their implementation could mark a significant milestone in the global fight against TB, aligning with international health goals to reduce TB incidence and mortality [10,11].

Biomarkers are biological indicators that signify normal or pathological physiological processes in reaction to internal or external exposures or treatments associated with the human immune system. They function as preliminary indicators for certain illnesses and are often categorized or obtained from antigens, proteins, genes, or genomes associated with a disease. Biomarkers have become a crucial area of study for doctors and biomedical scientists. Biomarker-based testing, an essential component of precision medicine, offers critical insights into illness prevention, diagnosis, and therapy by detecting particular biomarkers, including genes inside the human body [12]. These biomarkers are quantifiable and specific to certain disorders, facilitating precise measurement. Biomarkers for tuberculosis (TB) include antibodies such as immunoglobulins (IgG1 and IgG3), mycobacterial antigens such as Mycobacterium tuberculosis DNA and cell wall constituents (e.g., Lipoarabinomannan, LAM), specific genes including RAB20 or INSL3, and cytokines that stimulate T-cell immunity (e.g., interleukins IL-1α and IL-1β) [13]. Notwithstanding the promise of these biomarkers, there exists a paucity of understanding about the utilization of AI methodologies in the diagnosis of TB using these indications.

Previous systematic reviews have examined the role of machine learning in tuberculosis detection by imaging and conventional diagnostic methods. Nevertheless, a thorough meta-analysis on the usefulness and accuracy of machine learning algorithms in biomarker-based tuberculosis detection has yet to be performed. This systematic review and meta-analysis seeks to assess the efficacy of machine learning algorithms in detecting tuberculosis biomarkers,

offering a comprehensive evaluation of their diagnostic performance and quality management. Moreover, most prior systematic reviews focus on imaging or deep learning; our study will focused on biomarker-based ML models. This study will evaluate essential outcomes, including enhanced diagnostic accuracy, sensitivity, specificity, and the quality management of machine learning in tuberculosis detection, which may speed up treatment and perhaps enhance outcomes, particularly in low- and middle-income contexts. This research aims to synthesize current knowledge to provide insights that can improve early diagnosis, inform treatment, and bolster worldwide efforts to eradicate tuberculosis more effectively.

Materials and methods

This meta-analysis was conducted based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement guidelines [13]. This study was registered on PROSPERO with registration number CRD42024593089 as of October 15, 2024.

Search strategy and selection of studies

A comprehensive literature review was conducted on November 12, 2024. The literature search was conducted across 6 databases: PubMed, Cochrane, Wiley, EBSCO, Epistemonikos, and Web of Science, up to October 2024. The literature search used keywords in conjunction with the boolean operators 'AND' and 'OR', as specified in **Table 1**. Three independent evaluators (RNR, DRPR and FAG) conducted the article search, retrieval, and screening. Any disputes that occur throughout this procedure will be handled by consensus. Articles with relevant titles and abstracts will undergo full-text evaluation according to this procedure.

Table 1 Keywords combinations used in different databases.

Database	Keywords combination	Hits
PubMed	allintitle: (tuberculosis AND (machine learning OR artificial intelligence)) AND (biomarker OR diagnostic OR diagnosis OR detection) AND (hospital OR healthcare setting)	443
Cochrane	TITLE ((tuberculosis AND (machine learning OR artificial intelligence)) AND (biomarker OR diagnostic OR diagnosis OR detection) AND (hospital OR healthcare setting))	7
Wiley	TITLE ((tuberculosis AND (machine learning OR artificial intelligence)) AND (biomarker OR diagnostic OR diagnosis OR detection) AND (hospital OR healthcare setting))	590
WoS	(tuberculosis AND (machine learning OR artificial intelligence)) AND (biomarker OR diagnostic OR diagnosis OR detection) AND (hospital OR healthcare setting)	524
Epistemonikos	(tuberculosis AND (machine learning OR artificial intelligence)) AND (biomarker OR diagnostic OR diagnosis OR detection) AND (hospital OR healthcare setting)	378
EBSCO	(tuberculosis AND (machine learning OR artificial intelligence)) AND (biomarker OR diagnostic OR diagnosis OR detection) AND (hospital OR healthcare setting)	14

Inclusion and exclusion criteria

The inclusion criteria for this review encompassed: Studies involving human subjects with suspected or confirmed tuberculosis, inclusive of both adults and children; clinical trial studies; articles published in English from 2014 to 2024; and studies employing machine learning techniques for tuberculosis detection, diagnosis, classification, or prediction. Included studies were required to include diagnostic performance data, including sensitivity, specificity, accuracy, positive predictive value (PPV), or negative predictive value (NPV). Moreover, only research using particular indicators derived from non-sputum biological specimens - such as blood, urine, or saliva - were deemed suitable. The biomarkers included proteins (such as the ADA enzyme, secretory proteins, MTB H37Rv and T-cell epitopes), antitubercular peptides, and gene expression patterns, which were incorporated

into machine learning models for diagnostic evaluation. The exclusion criteria included: 1) Studies employing methodologies distinct from machine learning, such as deep learning or genetic algorithms, and those utilizing commercial diagnostic software (e.g., Qure.ai); 2) Review articles, conference proceedings, brief communications, research protocols, letters to editors, and preprints; 3) Studies involving non-human sampling; and 4) Non-peer-reviewed studies. The authors assessed the eligibility of each research independently, and any discrepancies were resolved by discussion.

Screening and selection

Following the automated elimination of duplicates using EndNote 19, the screening procedure was executed in 2 phases: An initial evaluation of titles and abstracts, followed by a comprehensive analysis of the entire texts. Each phase was conducted separately by 2 reviewers (RNR and FAG), with discrepancies resolved by consensus. In the absence of anonymity, a third reviewer (MA) was contacted to provide a final conclusion.

Critical appraisal

The quality of the included studies was independently evaluated by 2 reviewers (RNR and FAG) using the Newcastle-Ottawa Scale (NOS) for the assessment of observational studies. A detailed elucidation of the NOS has been previously delineated (8). The NOS assessment assessed 3 principal domains: Participant selection, group comparability, and exposure or outcome assessment. Studies were evaluated on a scale from 0 to 9, with quality classified as low (0 - 2), fair (3 - 5), and good/high (6 - 9). Disagreements were addressed via dialogue or, as required, by engaging a third reviewer (MA).

Data extraction

We retrieved the following data from each included research: First author's name, year of publication, country, study design, and sample size. Demographic information of the study participants was also gathered, including age (in months or years) and gender (male/female). Furthermore, methodological specifics including the used machine learning algorithms, evaluated biomarker characteristics, and performance measures (e.g., sensitivity, specificity and area under the curve (AUC)) were recorded to enable a comprehensive comparison across investigations. The accuracy metric is defined as the ratio of accurate predictions to the total number of forecasts. The specificity parameters are defined as the ratio of genuine negatives accurately detected by the model to the total number of negatives. The sensitivity parameters are defined as the ratio of genuine positives to the total number of actual positives that a model might have identified. Region The Area Under the ROC Curve is a comprehensive metric that evaluates the effectiveness of a binary classifier at various thresholds. All the criteria mentioned span from 0 to 1 (often expressed as a percentage), with an elevated score signifying superior performance. Authors extracted data using a defined manner. The first reviewer (DRPR) conducted data extraction separately, which was then validated for correctness and completeness by the second reviewer (RNR). Discrepancies detected throughout the procedure were addressed via conversation, so confirming the trustworthiness and authenticity of the retrieved material.

Quality assessment

The potential for bias in diagnostic accuracy studies was evaluated using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) instrument [14]. This instrument assesses 4 critical domains: Patient selection, index test, reference standard, and flow and time. The QUADAS-2 evaluation categorizes bias risk into 3 levels: Low (green), moderate (yellow), and high (red). Each research received an overall risk rating, recorded in the "bias" portion of a Microsoft Excel 2021 spreadsheet. The spreadsheet was then uploaded to the ROBVIS platform (https://mcguinlu.shinyapps.io/robvis/ (Accessed on November 14, 2023)) to visualize the assessment results using a traffic light system, enhancing clarity and interpretability.

Quantitative synthesis

A meta-analysis was conducted using Review Manager version 5.4 (The Nordic Cochrane Center, The Cochrane Collaboration, Copenhagen) [15]. The studies used a 2-level mixed-effects logistic regression to compute aggregate diagnostic accuracy measures,

including sensitivity, specificity, and AUC. The measures were used to evaluate the overall efficacy of machine learning models in diagnosing TB using biomarker data. Subgroup studies were performed to investigate the efficacy of various machine learning techniques. Furthermore, the biomarkers used include proteins (e.g., ADA enzyme, secretory proteins, MTB H37Rv, T-cell epitopes), antitubercular peptides, and gene expression patterns, which were incorporated into machine learning models for diagnostic evaluation. To address heterogeneity, we applied random-effects models, subgroup analyses, and sensitivity analyses. Subgrouping by biomarker type was not performed to avoid bias stemming from methodological differences in biomarker measurement and variability across assay platforms, as well as because most studies used multiple biomarkers simultaneously, making it difficult to isolate individual biomarker effects. Heterogeneity among the included studies was evaluated using the I^2 statistic, with thresholds of 0%, 25%, 50%, and 75% indicate negligible, low, moderate, and extreme heterogeneity, respectively. Subgroup analyzes included studies based on established criteria, guaranteeing that only those with relevant and comprehensive data were evaluated. The results of these studies provide significant insights into the variability and reliability of machine learning models for tuberculosis diagnosis, emphasizing critical parameters that affect their incorporation into hospital-based tuberculosis control efforts.

Results and discussion

Study selection and identification

From a total of 1,956 records identified across 6 databases (PubMed: 443, Web of Science: 524, Cochrane: 7, Wiley: 590, EBSCO: 14, and Epistemonikos: 378), 486 records were excluded based on year and type of article filter. Additionally, 119 duplicate records were removed, resulting in 1,351 records for screening. During the screening phase, 983 records were excluded based on their titles, and 330 records were excluded after abstract review. This reduced the list to 38 full-text publications that were evaluated for eligibility and overall analysis. Finally, 12 studies were considered appropriate for inclusion in the meta-analysis. **Figure 1** shows the PRISMA flow diagram, which summarizes the screening and selection procedure.

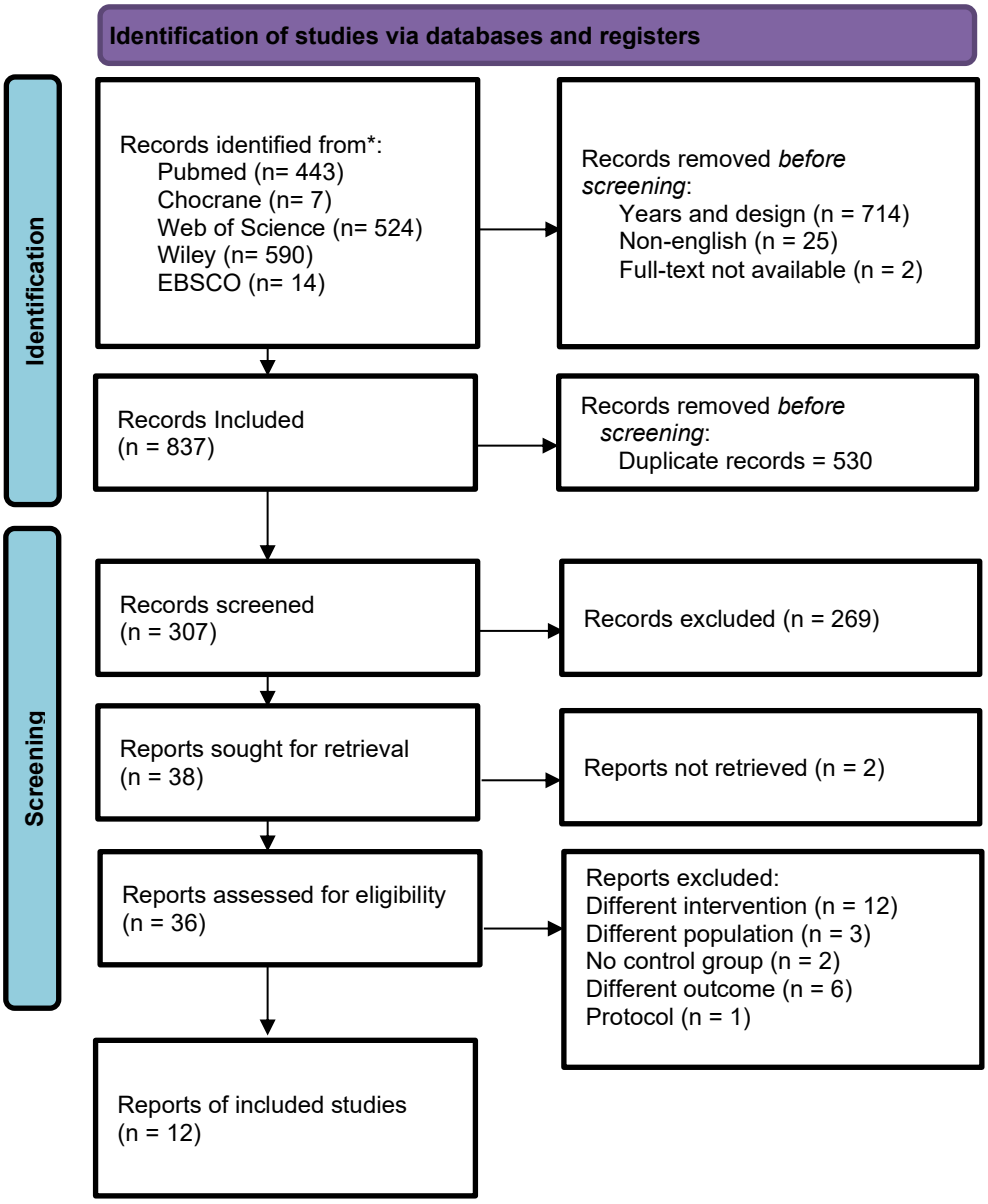

Figure 1 Preferred reporting items for systematic reviews and meta-analyses (PRISMA) flowchart for study identification and selection.

Characteristics of the included studies

Table 2 summarizes the features of the included studies. The table highlights the features of the 12 studies that used ML algorithms for biomarker-based TB diagnosis [17-28]. Sample sizes varied widely, ranging from 64 to 30,574 participants, and studies were conducted in diverse geographic settings including Colombia, Spain, China, and multicenter collaborations [17-28]. A diverse array of machine learning models was utilized, featuring prevalent algorithms such as random forest (RF), support vector machine (SVM), k-nearest neighbors (KNN), logistic regression (LR), gradient boosting machines (GBM/XGB), multilayer perceptron (MLP), and probabilistic neural networks (PNN) [17-28]. The data sources also varied among studies. Some used hospital-based clinical data, such as those from the TB program at Hospital Santa Clara or infectious disease hospitals [16], while others relied on established biological databases such as UniProt [21], STRING [23], NCBI GEO [26], AntiTb RD and MD [22], and the Immune Epitope Database [28]. Feature extraction methods included normalization techniques (such as mean-standard deviation or Pearson correlation), 1-hot encoding, recursive feature elimination, and gene ontology approaches, depending on the nature of the data. In terms of biomarkers, the

studies analyzed various molecular indicators including proteins (e.g., ADA enzyme, secretory proteins, MTB H37Rv, T-cell epitopes), antitubercular peptides, and gene expression profiles. Notably, some studies did not report specific biomarkers [17-28].

Table 2 Included studies characteristic.

No	Author	Algorithms	Sample	Country	Data source	Feature extraction	Biomarker
1	Orjuela-Cañón *et al.* [17]	DT, LR, MLP, RF, SVM	220	Colombia	TB program Hospital Santa Clara	NR	Protein - ADA enzyme
2	Garcia-Zamalloa *et al.* [18]	DT, KNN, MLP, RF	230	Spain	Multicenter Hospital	Normalization using mean-tandard deviation	Protein - ADA enzyme
3	Ghermi *et al.* [19]	GBM, KNN, LR, MLP, NB, RF, SVM, XGB	885	NR	NR	NR	Protein - ADA enzyme
4	Smith *et al.* [20]	GBM, RF, SVM	262	NR	NR	NR	Protein - ADA enzyme
5	Ahmed *et al.* [21]	KNN, PNN, SVM	200	NR	UniProt	NR	Protein - Secretory proteins
6	Akbar *et al.* [22]	KNN, PNN, RF, SVM	398	NR	AntiTb RD and AntiTb MD	One-hot encoding	Protein - Antitubercular peptides
7	Mei *et al.* [23]	LR	30574	NR	STRING database	Multi-instance Gene ontology	Protein - MTB H37Rv
8	Peng *et al.* [24]	LR, RF, SVM, XGB	429	NR	NR	NR	Organic molecule
9	Hu *et al.* [25]	MLP, RF, SVM	64	China	Infectious Disease Hospitals	Pearson correlation coefficient	Organic molecule
10	Osamor and Okezie [26]	NB, SVM	200	NR	NCBI GEO (Transcript Expression Omnibus)	Data normalization and recursive feature elimination	Gene - TB gene expression
11	Chen *et al.* [27]	RF, SVM, XGB	23587	NR	NR	NR	Organic molecule
12	Khanna and Rana [28]	RF, SVM	200	NR	Immune Epitope Database	R-Studio (caret package)	Protein- T-cell epitopes

NR = Not Reported; *DT*= Decision Tree; *GBM* = Gradient Boosting Machine; *KNN* = K-Nearest Neighbors; *LR* = *NR* = Not reported; Logistic Regression; *MLP* = Multi-Layer Perceptron; *NB* = Naive Bayes; *PNN* = Probabilistic Neural Network; *RF* = Random Forest; *SVM* = Support Vector Machine; *XGB* = XGBoost (Extreme Gradient Boosting).

The outcomes of various machine learning techniques applied to biomarker-based TB diagnosis are summarized and presented on **Table 3**. The machine learning techniques identified in the included studies included Decision Trees (DT), Gradient Boosting Machines (GBM), k-Nearest Neighbors (KNN), Logistic Regression (LR), Multilayer Perceptron (MLP), Naive Bayes (NB), Probabilistic Neural

Networks (PNN), Random Forest (RF), and Support Vector Machines (SVM). Decision Trees (DT) exhibited moderate performance, with accuracy ranging from 68.4% to 90%, while Gradient Boosting Machines (GBM) consistently showed high accuracy. K-Nearest Neighbors (KNN) performed well with accuracy between 78.9% and 92.5%, while Logistic Regression (LR) demonstrated reliable results, with accuracies ranging from 70% to 96%. MLP models achieved impressive performance, with accuracy reaching up to 94.7%. NB and PNN showed moderate-to-high accuracy, ranging from 67% to 96% and 79.8% to 91%, respectively.

Table 3 Outcomes reported in the included studies.

Authors	Algorithms	Accuracy	Sensitivity	Specificity	AUC	N+	N−
Orjuela-Cañón *et al.* [17]	DT	63	90	35	0.60 ± 0.005	184	36
Garcia-Zamalloa *et al.* [18]	DT	96	100	95	96	168	62
Ghermi *et al.* [19]	GBM	NR	83.3	83.3	89	225	660
Smith *et al.* [20]	GBM	NR	88	85	0.88	131	131
Ahmed *et al.* [21]	KNN	88.02	92.59	86.03	NR	100	100
Akbar *et al.* [22]	KNN	78.96	87.23	71.08	NR	199	199
Garcia-Zamalloa *et al.* [18]	KNN	0.91	78	95	96	168	62
Ghermi *et al.* [19]	KNN	76	70.2	82.1	83	225	660
Orjuela-Cañón *et al.* [17]	LR	86	94	66	0.69 ± 0.017	184	36
Ghermi *et al.* [19]	LR	79	71.4	86.9	86	225	660
Mei *et al.* [23]	LR	NR	99.75	99.76	99.78	15,287	15,287
Peng *et al.* [24]	LR	81	68.89	87.95	91	294	135
Orjuela-Cañón *et al.* [17]	MLP	80	82	68	0.71 ± 0.009	184	36
Garcia-Zamalloa *et al.* [18]	MLP	0.93	89	95	98	168	62
Hu *et al.* [25]	MLP	94.7	80	100	NR	27	37
Ghermi *et al.* [19]	NB	77	69	84.5	85	225	660
Osamor and Okezie [26]	NB	96	70	100	NR	100	100
Ahmed *et al.* [21]	PNN	91.02	100	86.46	NR	100	100
Akbar *et al.* [22]	PNN	79.82	93.04	93.62	NR	199	199
Akbar *et al.* [22]	RF	80.85	76.59	85.11	NR	199	199
Orjuela-Cañón *et al.* [17]	RF	66	87	36	0.67 ± 0.008	184	36
Chen *et al.* [27]	RF	85	76	88	NR	21,219	2,368
Garcia-Zamalloa *et al.* [18]	RF	0.93	89	95	98	168	62
Ghermi *et al.* [19]	RF	81	79.8	82.1	88	225	660
Hu *et al.* [18]	RF	89.47	80	85.71	NR	27	37
Khanna and Rana [28]	RF	78	71	87	0.78	100	100
Peng *et al.* [24]	RF	84	93.06	71.11	92	294	135
Smith *et al.* [20]	RF	NR	88	84	0.87	131	131
Ahmed *et al.* [21]	SVM	97	99.24	93.53	NR	100	100
Akbar *et al.* [22]	SVM	84.04	82.98	85.11	NR	199	199
Orjuela-Cañón *et al.* [17]	SVM	81	95	55	0.56 ± 0.013	184	36

Authors	Algorithms	Accuracy	Sensitivity	Specificity	AUC	N+	N−
Chen *et al.* [27]	SVM	79	72	91	NR	21,219	2,368
Ghermi *et al.* [19]	SVM	82	82.1	82.1	88	225	660
Hu *et al.* [25]	SVM	89.47	80	100	NR	27	37
Khanna and Rana [28]	SVM	72	70	76	0.69	100	100
Osamor and Okezie [26]	SVM	95	60	100	NR	100	100
Peng *et al.* [24]	SVM	80	80	80.72	93	294	135
Smith *et al.* [20]	SVM	NR	88	89	0.89	131	131
Chen *et al.* [27]	XGB	87	81	90	NR	21,219	2,368
Ghermi *et al.* [19]	XGB	83	83.3	83.3	89	225	660
Peng *et al.* [24]	XGB	84	71.1	91.67	93	294	135

NR = Not reported*; AUC* = Area Under the Curve; *NR* = Not Reported; *N+* = Number of True Positives; *N−* = Number of True Negatives; *DT*= Decision Tree; *GBM* = Gradient Boosting Machine; *KNN* = K-Nearest Neighbors; *LR* = Logistic Regression; *MLP* = Multi-Layer Perceptron; *NB* = Naive Bayes; *PNN* = Probabilistic Neural Network; *RF* = Random Forest; *SVM* = Support Vector Machine; *XGB* = XGBoost (Extreme Gradient Boosting).

Quality appraisal

Twelve diagnostic studies were evaluated for quality using the QUADAS-2 technique. Among them, one study was classified as having a high risk of bias, 4 were assessed as moderate risk, and the other 7 were considered to have a low risk of bias, as seen in **Figure 2**. In the area of patient selection, Garcia *et al.* [18] were assigned an ambiguous risk due to the study's lack of adequate information on whether participants were recruited sequentially or randomly, as well as the clarity of pre-specified inclusion and exclusion criteria. Meanwhile, Osamor and Okezie [26] were rated unclear due to the apparent use of a retrospective dataset without clarifying if the study design was cohort-based, which can introduce selection bias [23]. In the reference standard domain, Ahmed et al. received a high risk rating as the study failed to explain whether the reference standard was applied independently of the ML algorithm's outcome which might lead to incorporation bias. For the flow and timing domain, 3 studies were rated as unclear [20]. Mei *et al.* [23] did not specify the time interval between biomarker collection and reference testing. Khanna and Rana [28] lacked clarity on whether all participants received both the index and reference tests, and did not state if any patients were excluded from the final analysis [25]. Lastly, Ahmed *et al.* [21] showed discrepancies between the number of participants initially enrolled and those included in the outcome analysis.

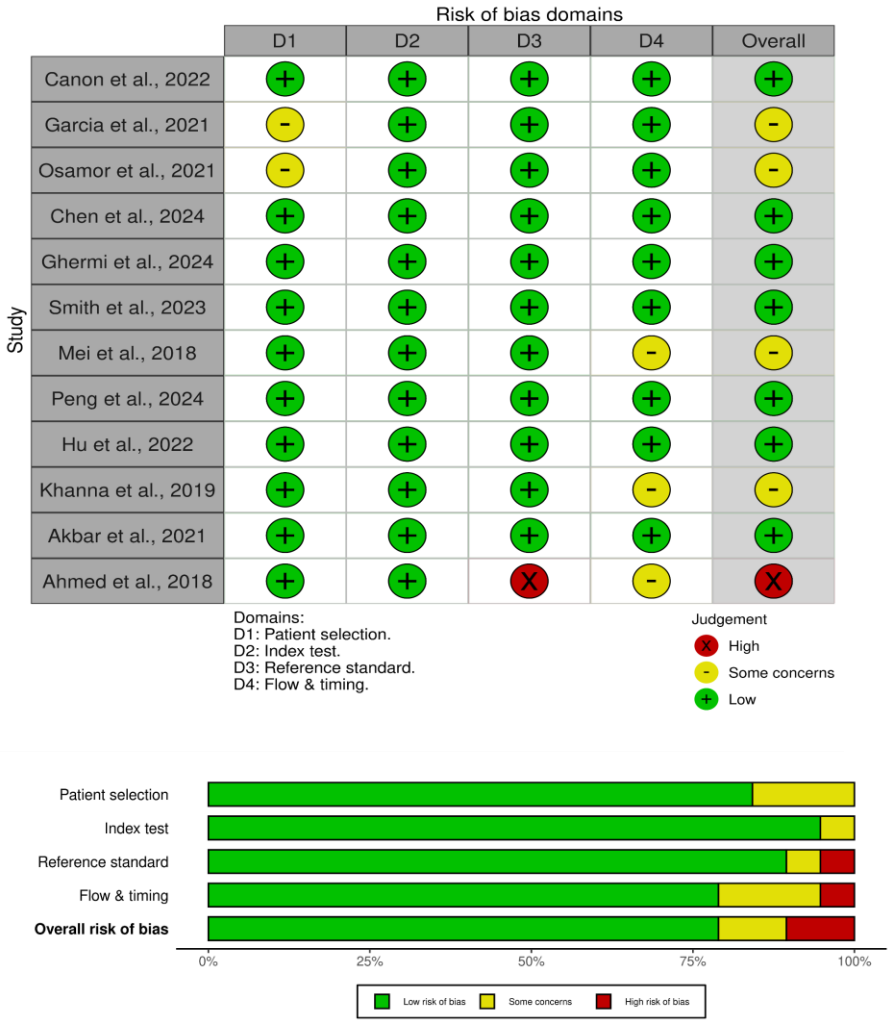

Figure 2 Risk of bias summary using the QUADAS-2 tool for randomized-controlled trial studies. The green region represents studies with a low risk of bias, the yellow region shows studies with unclear risk of bias, and the red region shows studies with a high risk of bias.

Pooled sensitivity, specificity, and accuracy of machine learning algorithms for biomarker-based tuberculosis diagnosis

The table displays the pooled sensitivity, specificity, and accuracy of various machine learning algorithms, along with their respective 95% confidence intervals (CI) and heterogeneity (I^2 values). Algorithms include Decision Tree (DT), Gradient Boosting Machine (GBM), K-Nearest Neighbors (KNN), Logistic Regression (LR), Multilayer Perceptron (MLP), Naive Bayes (NB), Probabilistic Neural Network (PNN), Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGB). PNN demonstrated the highest overall accuracy (0.928, 95% CI: 0.909 - 0.947) and sensitivity (0.961, 95% CI: 0.903 - 1.000), while NB achieved the highest specificity (0.915, 95% CI: 0.784 - 1.000). Variability in performance metrics across algorithms highlights the diversity in diagnostic capabilities.

Table 4 Summary outcomes of each algorithms.

Algorithms	Pooled sensitivity		Pooled specificity		Pooled accuracy	
	Mean (95% CI)	I^2	Mean (95% CI)	I^2	Mean (95% CI)	I^2
DT	0.952 (0.860 - 1.000)	98.50%	0.587 (0.227 - 1.000)	99.24%	0.897 (0.742 - 1.000)	98.00%
GBM	0.852 (0.808 - 0.899)	51.84%	0.835 (0.814 - 0.857)	0.00%	0.844 (0.816 - 0.872)	50.76%
KNN	0.784 (0.717 - 0.858)	88.68%	0.887 (0.826 - 0.954)	90.76%	0.834 (0.786 - 0.885)	90.21%
LR	0.543 (0.284 - 1.000)	99.79%	0.785 (0.687 - 0.898)	94.92%	0.731 (0.599 - 0.893)	98.85%
MLP	0.889 (0.817 - 0.968)	85.04%	0.809 (0.661 - 0.990)	90.41%	0.860 (0.794 - 0.931)	84.07%
NB	0.692 (0.658 - 0.727)	0.00%	0.915 (0.784 - 1.000)	98.81%	0.820 (0.784 - 0.859)	63.42%
PNN	0.961 (0.903 - 1.000)	89.27%	0.899 (0.833 - 0.969)	74.73%	0.928 (0.909 - 0.947)	0.00%
RF	0.824 (0.776 - 0.875)	95.52%	0.773 (0.644 - 0.929)	99.38%	0.825 (0.795 - 0.856)	92.36%
SVM	0.805 (0.734 - 0.882)	98.83%	0.855 (0.782 - 0.936)	98.71%	0.838 (0.793 - 0.887)	97.57%
XGB	0.783 (0.714 - 0.859)	94.89%	0.882 (0.832 - 0.934)	92.74%	0.811 (0.782 - 0.842)	88.43%

CI; Confidence Interval; I^2: I-squared statisti; DT= Decision Tree; GBM = Gradient Boosting Machine; KNN = K-Nearest Neighbors; LR = Logistic Regression; MLP = Multi-Layer Perceptron; NB = Naive Bayes; PNN = Probabilistic Neural Network; RF = Random Forest; SVM = Support Vector Machine; XGB = XGBoost (Extreme Gradient Boosting).

Performance comparison of machine learning algorithms for biomarker-based tuberculosis diagnosis

This figure presents forest plots of sensitivity and specificity, along with their 95% CI, for various machine learning algorithms across different studies. Algorithms include DT, GBM, KNN, LR, MLP, NB, PNN, RF, SVM, and XGB. Each plot includes the TP, FP, FN, and TN values for individual studies. The blue squares represent the point estimates, while the horizontal lines indicate the 95% CI, demonstrating the variability and precision of sensitivity and specificity for each algorithm across different datasets. The consistency of performance varies between algorithms, with some (such as PNN) showing high sensitivity and specificity across studies, while others (such as LR) demonstrate greater variability.

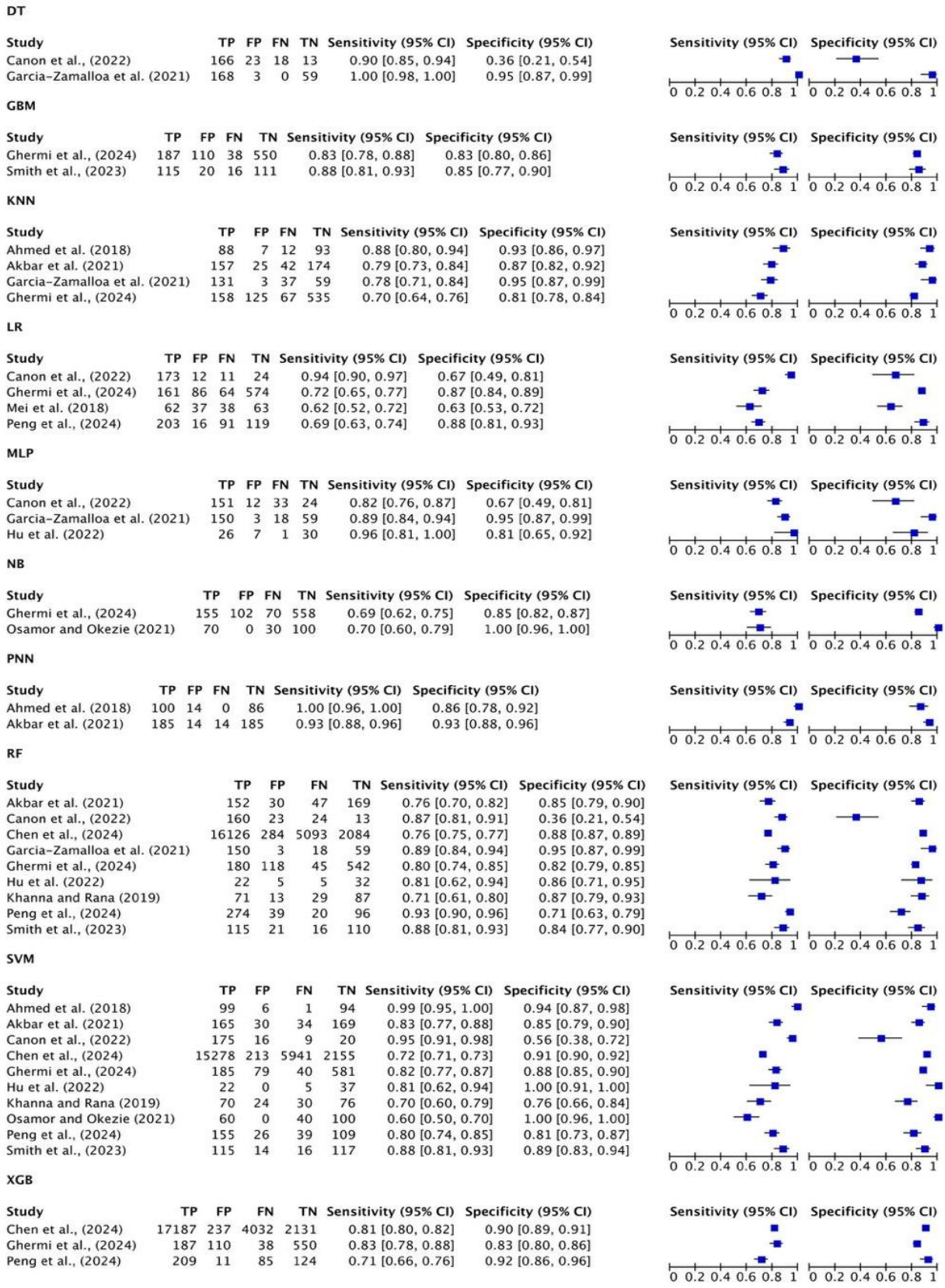

Figure 3 Forest plot summarizing sensitivity and specificity for different machine learning algorithms. The blue square and solid lines represent the odds ratio with 95% confidence intervals. The size of the squares indicates the weight of each study. The black rhombus indicates the pooled estimate with 95% confidence intervals.

Summary receiver operating characteristic (sROC) curves of machine learning algorithms for biomarker-based tuberculosis diagnosis

In this sROC curves, sensitivity is plotted against 1-specificity to evaluate the overall diagnostic performance. Each curve represents the trade-off between sensitivity and specificity for a specific algorithm, summarizing the diagnostic ability across studies. PNN and DT algorithms indicate superior diagnostic capabilities, characterized by higher sensitivity and specificity.

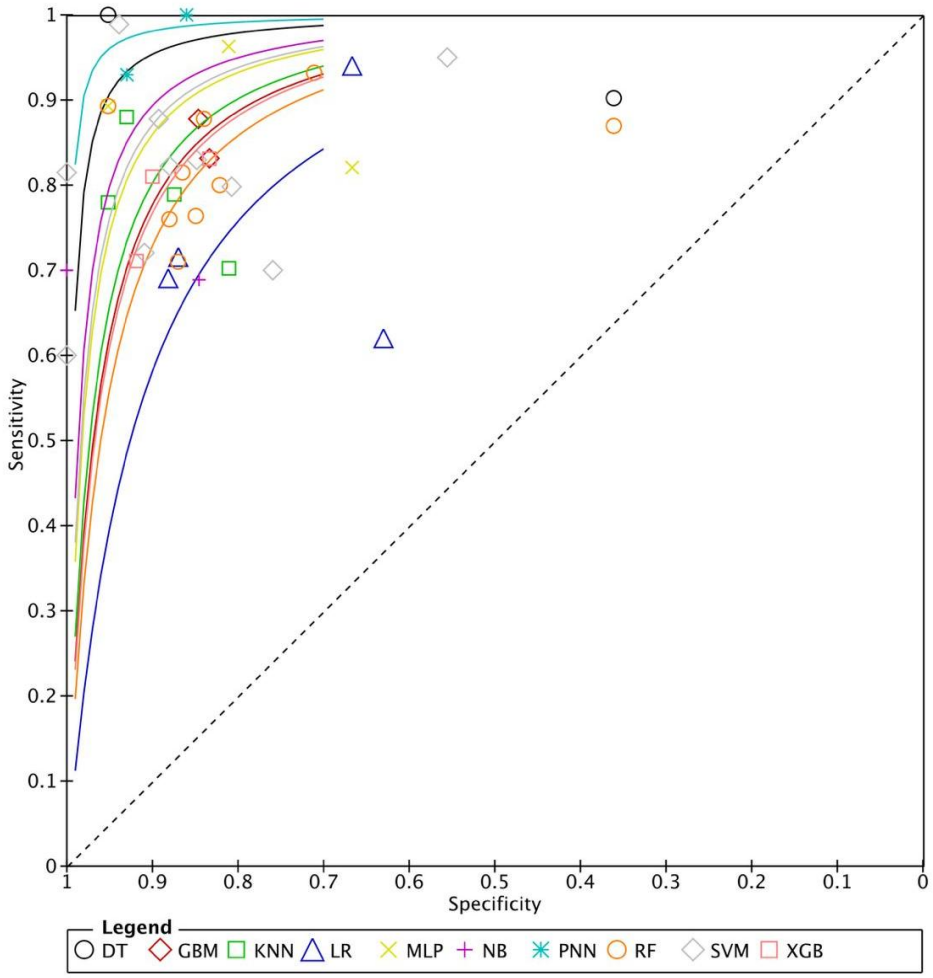

Figure 4 Summary receiver operating characteristic (sROC) curves for different machine learning algorithms. Data points on the plot correspond to individual study results, while the curves reflect the algorithm's aggregated performance. The dashed diagonal line represents random classification performance. Algorithms with curves closer to the upper left corner indicate superior diagnostic capabilities, characterized by higher sensitivity and specificity.

Table 5 Summary of Area Under Curve (sAUC) of each algorithms.

Algoritma	sAUC	95% CI	*p*-value	I^2 (%)
DT	0.8231	0.4876 - 1.1586	<0.0001	98.18
GBM	0.8213	0.7425 - 0.9001	<0.0001	89.61
KNN	0.8053	0.7375 - 0.8731	<0.0001	91.36

Algoritma	sAUC	95% CI	*p*-value	I^2 (%)
LR	0.8342	0.7199 - 0.9485	<0.0001	98.41
MLP	0.8307	0.6750 - 0.9863	<0.0001	95.97
NB	0.8150	0.6790 - 0.9510	<0.0001	96.81
PNN	0.9348	0.9157 - 0.9539	<0.0001	0.00
RF	0.7828	0.7208 - 0.8448	<0.0001	96.81
SVM	0.8194	0.7551 - 0.8837	<0.0001	98.11
XGB	0.7385	0.6642 - 0.8129	<0.0001	96.68

Table 5 shows sAUC for various machine learning algorithms. Among all, PNN achieved the highest AUC (0.9348) with no observed heterogeneity (I^2 = 0%), indicating consistent performance. LR, MLP, and GBM also showed high AUCs (above 0.82), suggesting good discriminative ability, though with high heterogeneity (I^2 > 89%). In contrast, RF and XGB had lower AUCs (0.7828 and 0.7385, respectively) with substantial heterogeneity. SVM, NB, and DT presented statistically significant results, with high I^2 values indicating variability across studies. All models had significant *p*-values (p < 0.0001), confirming the robustness of their pooled performance. Overall, PNN was the most consistent, while other algorithms showed considerable heterogeneity.

Discussion

The present study evaluated the diagnostic performance of various ML algorithms for biomarker-based TB diagnosis. Pooled sensitivity, specificity, and accuracy metrics were provided, complemented by forest plots and sROC curves to visualize their diagnostic capacity across different datasets. Overall, our findings emphasize the increasing importance of ML techniques in TB diagnostics, with significant differences in algorithmic performance. Among these, the PNN demonstrated superior performance with the highest pooled sensitivity (96.1%), specificity (89.9%), and accuracy (92.8%). PNN's superior performance is due to its unique structure as a radial basis function (RBF) network that uses Bayesian decision theory to optimize classification [27]. It estimates probability density functions for each class and classifies samples based on maximum likelihood, which improves accuracy with complex, high-dimensional biomarker data. Its non-parametric nature allows it to adapt well to different datasets without extensive tuning, reducing sensitivity to variations in biomarkers and populations [28]. A previous study also reported that the PNN achieved remarkably high diagnostic performance, with an accuracy of 97.0%, sensitivity of 99.24%, and specificity of 92.53%, indicating its strong potential for detecting true TB cases while minimizing false positives [21]. These results suggest that PNN may be particularly effective in clinical settings where both early identification and precise discrimination between TB and non-TB cases are crucial. This makes PNN promising for TB diagnostics, especially in resource-limited, high-burden areas. PNNs excel in learning complex patterns, adapting to various network architectures, and performing robustly in diverse applications. Their flexibility supports implementation across different systems, which is valuable for biomarker-based TB diagnostics that require efficient handling of high-dimensional data. Additionally, PNNs generalize well across datasets, enhancing their diagnostic accuracy and clinical utility [29]. In the present study, DT algorithms demonstrated high sensitivity (95.2%) and accuracy (89.7%), highlighting their strength in identifying TB cases [30]. By recursively partitioning data based on informative features, DTs create interpretable rule-based models that are easy to implement, especially in low-resource settings. However, their relatively low specificity (58.7%) indicates a tendency for false positives, which may lead to over-diagnosis. Their simplicity,

transparency, and ease of visualization support clinical decision-making. NB algorithms, on the other hand, achieved the highest specificity (91.5%), effectively identifying non-TB cases - particularly valuable in low-prevalence areas [31]. However, with lower sensitivity (69.2%), NB models risk missing true TB cases. Despite these trade-offs, both DT and NB remain practical options for TB diagnostics in settings with limited computational resources due to their fast training, minimal complexity, and interpretability, which fosters clinical trust and usability [32].

Our study also demonstrated that RF and SVM demonstrated balanced diagnostic performance, with pooled accuracy of 82.5% and 83.8%, respectively [21,33]. These models effectively handle complex, non-linear relationships in biomarker data, making them suitable when both sensitivity and specificity are critical. RF showed slightly lower specificity (77.3%) than SVM (85.5%), but its ensemble nature offers robustness to variations in biomarker expression. SVM, with its ability to construct complex decision boundaries, excels in scenarios requiring precise classification. Their adaptability and performance make both models well-suited for TB diagnostics, particularly in resource-limited settings where diagnostic accuracy is vital. In contrast, LR yielded lower pooled sensitivity (54.3%) and accuracy (73.1%), with extremely high heterogeneity in sensitivity (I^2 = 99.8%) [34,35]. This suggests instability across studies, likely due to LR's reliance on linear relationships and its limited capacity to model complex feature interactions [25,34]. Such limitations hinder its performance on heterogeneous, high-dimensional datasets typical of TB biomarkers. While LR may still serve as a baseline comparator or for simpler diagnostic tasks, its reduced generalizability and tendency to underperform in nuanced classification highlight the need for more advanced ML approaches in TB diagnostics. SROC curves also showed that PNN, NB, and DT had the best sensitivity-specificity balance, indicating strong diagnostic performance. In contrast, LR and XGB were farther from the ideal curve, suggesting limited suitability for broad TB diagnostic use.

Prior research assessing machine learning algorithms in pulmonary tuberculosis indicated that decision trees were the most effective algorithm for analyzing clinical trial data, possibly attributable to the dataset's properties and the aims of the analysis [43]. Although DT, RF, SVM, and NB exhibit commendable performance in many scenarios, more sophisticated methodologies such as Gradient Boosting Machines, Deep Learning (Neural Networks), DBN, and advanced ensemble approaches provide enhanced accuracy, adequately manage intricate interactions, and easily scale to extensive datasets [44]. For example, a DBN survival Cox model outperformed other ML algorithms in predicting overall survival in osteosarcoma patients [41]. Additionally, a Chinese study found RF to be the top-performing model among 6 ML approaches for predicting lymph node metastasis in Ewing's sarcoma [45]. This difference could stem from the fact that prior studies dealt with different clinical conditions and prediction tasks (e.g., survival, metastasis), which may favor models designed for continuous outcome prediction or larger, more structured datasets. Therefore, the divergence in findings highlights the importance of matching ML algorithms to the specific data types and clinical questions in TB biomarker diagnostics versus other medical domains. While this study primarily focuses on evaluating ML algorithms for TB diagnostics, it is equally important to consider the baseline diagnostic performance of the biomarkers themselves in the absence of ML. Previous meta-analyses have demonstrated that individual biomarkers such as IFN-γ, IP-10, CRP, and certain miRNAs offer moderate diagnostic accuracy when used alone, typically achieving pooled AUCs ranging from 0.70 to 0.80, with variable sensitivity and specificity depending on the population and setting [43]. However, when these same biomarkers are analyzed using ML algorithms, our findings indicate a substantial improvement in diagnostic performance, with pooled AUCs reaching up to 0.89 and enhanced balance between sensitivity and specificity. This comparison highlights the added value of ML in extracting more patterns from biomarker data, optimizing classification decisions beyond what conventional threshold-based interpretations can achieve. By integrating both biological and computational strengths, ML models amplify the clinical utility of existing biomarkers, enabling more accurate, timely, and context-adaptable TB diagnostics.

Our study findings carry important clinical implications, especially for resource-limited and high-

burden settings. Algorithms with high sensitivity, such as PNN and DT, are ideal in such contexts as they enhance early detection of TB cases, enabling timely treatment and reducing community transmission [36]. In contrast, high-specificity algorithms such as NB are better suited for low-prevalence areas, where minimizing false positives helps conserve limited resources and reduce patient anxiety. For more balanced needs, RF and SVM offer strong overall performance, adaptable across various clinical settings [37]. Beyond diagnostics, ML integration supports hospital management through early case detection, contact tracing, and infection control [46]. In high-burden hospitals, sensitive algorithms can improve screening among patients and staff, aiding in early isolation and preventing nosocomial spread. In low-burden settings, specific models can prevent unnecessary isolation, ensuring efficient use of resources. This targeted application of ML enhances diagnostic workflows and infection control strategies across diverse healthcare environments [38,39]. This targeted ML approach can improve patient flow and reduce costs by minimizing overcrowding and optimizing bed use. ML-based risk assessment tools also help hospital administrators identify high-risk zones, guiding tailored infection control strategies like improved ventilation, triage systems, and staff rotation. However, given performance variability across studies, algorithm selection should be tailored to local data, population characteristics, and available resources [40]. Successful implementation requires both technical and strategic planning to enhance diagnostic accuracy, reduce TB burden, and support better patient care. As hospitals advance digitally, integrating ML into TB diagnostics can strengthen public health efforts, support clinical decisions, and reduce transmission in healthcare settings [41,42].

A major strength of this study lies in its ability to synthesize diverse data to highlight the potential of ML models in enhancing diagnostic accuracy of TB on a global scale. However, several limitations should be acknowledged. Although biomarker type is likely to influence the diagnostic accuracy of ML models, our review did not perform subgroup analyses based on biomarker categories. This decision was made because most studies assessed multiple biomarkers simultaneously, and measurement platforms varied widely across investigations. Attempting to isolate individual biomarker effects in this context risked introducing methodological bias. Future studies using standardized biomarker panels and harmonized assay platforms are needed to systematically evaluate the relative contributions of different biomarker types and combinations to ML diagnostic performance. The differences in algorithm implementation in real-settings of each study can also contribute to the observed heterogeneity in results. Additionally, the lack of standardized evaluation metrics and reporting formats across studies limits the generalizability of the findings and hinders direct comparison. Despite these limitations, the study emphasizes the potential of machine learning algorithms in increasing tuberculosis diagnosis, as well as the necessity for standardized methodology to improve future research and clinical applications. Future studies might possibly fill these gaps. Finally, this systematic review intends to shed light on the application of machine learning algorithms in tuberculosis diagnoses, emphasizing their strengths, limits, and clinical significance.

Conclusions

This systematic review and meta-analysis demonstrate that machine learning algorithms hold significant promise for biomarker-based tuberculosis diagnosis, with probabilistic neural networks (PNN) showing the most consistent performance, while decision trees and naïve Bayes offer interpretability and feasibility in resource-limited settings. However, the findings must be interpreted with caution given the high heterogeneity across studies, variability in biomarker types and assay methods, potential publication bias, and dataset imbalance. These limitations highlight the need for standardized biomarker panels, harmonized evaluation protocols, and balanced datasets to strengthen the reliability of future models. Moreover, deep learning approaches, though excluded from the present analysis, represent an important avenue for biomarker-based diagnostics that should be systematically assessed as the field evolves. Future research should integrate accuracy, explainability, and clinical feasibility to develop ML-driven diagnostic tools that are robust, equitable, and adaptable to diverse healthcare settings.

Acknowledgements

The authors would like to express their sincere gratitude to Dr. Merita Arini for her invaluable guidance, support, and insightful feedback throughout the development of this study. We also extend our appreciation to the Master of Hospital Administration Program at Universitas Muhammadiyah Yogyakarta, Indonesia for providing the academic environment and resources necessary to complete this work.

Declaration of Generative AI in Scientific Writing

This research used artificial intelligence (AI) technologies and approaches to assist with article composition. AI-driven language models, like ChatGPT, were used for linguistic enhancement, including the refinement of grammar and sentence structure. We affirm that all AI-assisted procedures underwent rigorous evaluation by the authors to guarantee the integrity and trustworthiness of the outcomes. The conclusions and interpretations articulated in this article were exclusively determined by the writers.

CRediT Author Statement

Roy Novri Ramadhan: Conceptualization; Methodology; Software; Data curation; Writing - Original draft preparation. **Danendra Rakha Putra Respati:** Visualization; Investigation; Writing - Original draft preparation. **Merita Arini:** Supervision; Validation.

References

[1] World Health Organization. *Global tuberculosis report 2024*. World Health Organization, Geneva, Switzerland, 2024.

[2] DN Ardiansyah, C Suryawati and MS Adi. Provision of resources in the implementation of tuberculosis-multi drug resistance treatment service in "X" Hospital. *Jurnal Medicoeticolegal dan Manajemen Rumah Sakit* 2019; **8(3)**, 123-130.

[3] Kementerian Kesehatan Republik Indonesia. *Laporan unit pelayanan Kesehatan* (*in Indonesia*n). Kementerian Kesehatan RI Jakarta, Jakarta, Indonesia, 2021.

[4] R Guo, K Passi and CK Jain. Tuberculosis diagnostics and localization in chest X-rays via deep learning models. *Frontiers in Artificial Intelligence* 2020; **3**, 583427.

[5] B Wijiseno, M Arini and E Listiowati. Healthcare workers' acceptance of the integrated tuberculosis-COVID-19 screening in central Java Private Hospitals, Indonesia. *Journal of Taibah University Medical Sciences* 2023; **18(6)**, 1311-1320.

[6] DA Prakoso, W Istiono, Y Mahendradhata and M Arini. Acceptability and feasibility of tuberculosis-diabetes mellitus screening implementation in private primary care clinics in Yogyakarta, Indonesia: A qualitative study. *BMC Public Health* 2023; **23(1)**, 1908.

[7] K Murphy, SS Habib, SMA Zaidi, S Khowaja, A Khan, J Melendez, ET Scholten, F Amad, S Schalekamp, M Verhagen, RHHM Philipsen, A Meijers and BV Ginneken. Computer aided detection of tuberculosis on chest radiographs: An evaluation of the CAD4TB v6 system. *Scientific Reports* 2020; **10(1)**, 5492.

[8] N Tang, M Yuan, Z Chen, J Ma, R Sun, Y Yang, Q He, X Guo, S Hu and J Zhou. Machine learning prediction model of tuberculosis incidence based on meteorological factors and air pollutants. *International Journal of Environmental Research and Public Health* 2023; **20(5)**, 3910.

[9] S Memon, S Bibi and G He. Integration of AI and ML in tuberculosis (TB) management: From diagnosis to drug discovery. *Diseases* 2025; **13(6)**, 184.

[10] J Salcedo, M Rosales, JS Kim, D Nuno, S Suen and AH Chang. Cost-effectiveness of artificial intelligence monitoring for active tuberculosis treatment: A modeling study. *Plos One* 2021; **16(7)**, e0254950.

[11] MMS Rodrigues, B Barreto-Duarte, CL Vinhaes, M Araújo-Pereira, ER Fukutani, KB Bergamaschi, A Kristki, M Cordeiro-Santos, VC Rolla, TR Sterling, ATL Queiroz and BB Andrade. Machine learning algorithms using national registry data to predict loss to follow-up during tuberculosis treatment. *BMC Public Health* 2024; **24(1)**, 1385.

[12] DS Reddy and HN Abeygunaratne. Experimental and clinical biomarkers for progressive evaluation of neuropathology and therapeutic interventions for acute and chronic neurological

disorders. *International Journal of Molecular Sciences* 2022; **23(19)**, 11734.

[13] V Kulkarni, ATL Queiroz, S Sangle, A Kagal, S Salvi, A Gupta, J Ellner, D Kadam, VC Rolla, BB Andrade, P Salgame and V Mave. A 2-gene signature for tuberculosis diagnosis in persons with advanced HIV. *Frontiers in Immunology* 2021; **12**, 631165.

[14] MJ Page, JE McKenzie, PM Bossuyt, I Boutron, TC Hoffmann, CD Mulrow, L Shamseer, JM Tetzlaff, EA Akl, SE Brennan, R Chou, J Glanville, JM Grimshaw, A Hróbjartsson, MM Lalu, T Li, EW Loder, E Mayo-Wilson, S McDonald, …, D Moher. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *The BMJ* 2021; **372**, n71.

[15] PF Whiting. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine* 2011; **155(8)**, 529-536.

[16] The Cochrane Collaboration. *Review manager (RevMan)*. The Cochrane Collaboration, London, 2020.

[17] AD Orjuela-Cañón, AL Jutinico, C Awad, E Vergara and A Palencia. Machine learning in the loop for tuberculosis diagnosis support. *Frontiers in Public Health* 2022; **10**, 876949.

[18] A Garcia-Zamalloa, D Vicente, R Arnay, A Arrospide, J Taboada, I Castilla-Rodríguez, U Aguirre, N Múgica, L Aldama, B Aguinagalde, M Jimenez, E Bikuña, MB Basauri, M Alonso and E Perez-Trallero. Diagnostic accuracy of adenosine deaminase for pleural tuberculosis in a low prevalence setting: A machine learning approach within a 7-year prospective multi-center study. *Plos One* 2021; **16(11)**, e0259203.

[19] M Ghermi, M Messedi, C Adida, K Belarbi, MEA Djazouli, ZI Berrazeg, M Kallel Sellami, Y Ghezini and M Louati. TubIAgnosis: A machine learning-based web application for active tuberculosis diagnosis using complete blood count data. *Digital Health* 2024; **10**, 1-12.

[20] JP Smith, K Milligan, KD McCarthy, W Mchembere, E Okeyo, SK Musau, A Okumu, R Song, ES Click and KP Cain. Machine learning to predict bacteriologic confirmation of *Mycobacterium tuberculosis* in infants and very young children. *Plos Digital Health* 2023; **2(5)**, e0000249.

[21] S Ahmed, M Kabir, M Arif, Z Ali, F Ali and ZNK Swati. Improving secretory proteins prediction in *Mycobacterium tuberculosis* using the unbiased dipeptide composition with support vector machine. *International Journal of Data Mining and Bioinformatics* 2018; **21(3)**, 212-229.

[22] S Akbar, A Ahmad, M Hayat, AU Rehman, S Khan and F Ali. iAtbP-Hyb-EnC: Prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model. *Computers in Biology and Medicine* 2021; **137**, 104778.

[23] S Mei, EK Flemington and K Zhang. Transferring knowledge of bacterial protein interaction networks to predict pathogen targeted human genes and immune signaling pathways: A case study on *M. tuberculosis*. *BMC Genomics* 2018; **19(1)**, 505.

[24] A Peng, XH Kong, S Liu, H Zhang, L Xie, L Ma, Q Zhang and Y Chen. Explainable machine learning for early predicting treatment failure risk among patients with TB-diabetes comorbidity. *Scientific Reports* 2024; **14(1)**, 6814.

[25] X Hu, J Wang, Y Ju, X Zhang, W Qimanguli, C Li, L Yue, B Tuohetaerbaike, Y Li, H Wen, W Zhang, C Chen, Y Yang, J Wang and F Chen. Combining metabolome and clinical indicators with machine learning provides some promising diagnostic markers to precisely detect smear-positive/negative pulmonary tuberculosis. *BMC Infectious Diseases* 2022; **22(1)**, 707.

[26] VC Osamor and AF Okezie. Enhancing the weighted voting ensemble algorithm for tuberculosis predictive diagnosis. *Scientific Reports* 2021; **11(1)**, 14806.

[27] CF Chen, CH Hsu, YC Jiang, WR Lin, WC Hong, IY Chen, MH Lin, KA Chu, CH Lee, DL Lee and PF Chen. A deep learning-based algorithm for pulmonary tuberculosis detection in chest radiography. *Scientific Reports* 2024; **14(1)**, 14917.

[28] D Khanna and PS Rana. Ensemble technique for prediction of T-cell *Mycobacterium tuberculosis* epitopes. *Interdisciplinary Sciences:*

Computational Life Sciences 2019; **11(4)**, 611-627.

[29] J Peurifoy, Y Shen, L Jing, Y Yang, F Cano-Renteria, BG DeLacy, JD Joannopoulos, M Tegmark and M Soljačić. Nanophotonic particle simulation and inverse design using artificial neural networks. *Science Advances* 2018; **4(6)**, eaar4206.

[30] A Paszke, S Gross, F Massa, A Lerer, J Bradbury and G Chanan. *PyTorch: An imperative style, high-performance deep learning library*. *In*: H Wallach, H Larochelle, A Beygelzimer, F D'Alché-Buc, E Fox and R Garnett (Eds.). Advances in neural information processing systems. Curran Associates, New York, 2019.

[31] H Blockeel, L Devos, B Frénay, G Nanfack and S Nijssen. Decision trees: From efficient prediction to responsible AI. *Frontiers in Artificial Intelligence* 2023; **6**, 1223426.

[32] R Ahmad, L Xie, M Pyle, MF Suarez, T Broger, D Steinberg, SM Ame, MG Lucero, MJ Szucs, M MacMullan, FS Berven, A Dutta, DM Sanvictores, VL Tallo, R Bencher, DP Eisinger, U Dhingra, S Deb, SM Ali, …, MA Gillette. A rapid triage test for active pulmonary tuberculosis in adult patients with persistent cough. *Science Translational Medicine* 2019; **11(515)**, eaaw8287.

[33] O Estévez, L Anibarro, E Garet, Á Pallares, L Barcia, L Calviño, C Maueia, T Mussá, F Fdez-Riverola, D Glez-Peña, M Reboiro-Jato, H López-Fernández, NA Fonseca, R Reljic and Á González-Fernández. An RNA-seq based machine learning approach identifies latent tuberculosis patients with an active tuberculosis profile. *Frontiers in Immunology* 2020; **11**, 1470.

[34] T Domaszewska, J Zyla, R Otto, SHE Kaufmann and J Weiner. Gene set enrichment analysis reveals individual variability in host responses in tuberculosis patients. *Frontiers in Immunology* 2021; **12**, 703941.

[35] K Zou, W Ren, S Huang, J Jiang, H Xu, X Zeng, H Zhang, Y Peng, M Lü and X Tang. The role of artificial neural networks in prediction of severe acute pancreatitis associated acute respiratory distress syndrome: A retrospective study. *Medicine* 2023; **102(29)**, e34399.

[36] R Sutradhar and L Barbera. Comparing an artificial neural network to logistic regression for predicting ED visit risk among patients with cancer: A population-based cohort study. *Journal of Pain and Symptom Management* 2020; **60(1)**, 1-9.

[37] Y Xiao, Y Chen, R Huang, F Jiang, J Zhou and T Yang. Interpretable machine learning in predicting drug-induced liver injury among tuberculosis patients: Model development and validation study. *BMC Medical Research Methodology* 2024; **24(1)**, 92.

[38] RM Carrillo-Larco, LT Car, J Pearson-Stuttard, T Panch, JJ Miranda and R Atun. Machine learning health-related applications in low-income and middle-income countries: A scoping review protocol. *BMJ Open* 2020; **10(5)**, e035983.

[39] B Wahl, A Cossy-Gantner, S Germann and NR Schwalbe. Artificial intelligence and global health: How can AI contribute to health in resource-poor settings? *BMJ Global Health* 2018; **3(4)**, e000798.

[40] J Wu and Y Zhao. Machine learning technology in the application of genome analysis: A systematic review. *Gene* 2019; **705**, 149-156.

[41] H Habehh and S Gohel. Machine learning in healthcare. *Current Genomics* 2021; **22(4)**, 291-300.

[42] E Johns, A Alkanj, M Beck, L Dal Mas, B Gourieux, EA Sauleau and B Michel. Using machine learning or deep learning models in a hospital setting to detect inappropriate prescriptions: A systematic review. *European Journal of Hospital Pharmacy* 2024; **31(4)**, 289-294.

[43] S Ahamed Fayaz, L Babu, L Paridayal, M Vasantha, P Paramasivam, K Sundarakumar and C Ponnuraja. Machine learning algorithms to predict treatment success for patients with pulmonary tuberculosis. *Plos One* 2024; **19(10)**, e0309151.

[44] IH Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science* 2021; **2(3)**, 160.

[45] W Li, Q Zhou, W Liu, C Xu, ZR Tang, S Dong, H Wang, W Li, K Zhang, R Li, W Zhang, Z Hu, S Shibin, Q Liu, S Kuang and C Yin. A machine learning-based predictive model for predicting

lymph node metastasis in patients with ewing's sarcoma. *Frontiers in Medicine* 2022; **9**, 832108.

[46] M Arini, D Sugiyo and I Permana. Challenges, opportunities, and potential roles of the private primary care providers in tuberculosis and diabetes mellitus collaborative care and control: A qualitative study. *BMC Health Services Research* 2022; **22(1)**, 215.