International Journal of Basic and Applied Sciences, 14 (1) (2025) 435-444

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/parmkr90
Research paper

Tuberculosis prediction: performance analysis of machine learning models for early diagnosis and screening using symptom severity level data

Suresh S. [1] *, Dhanalakshmi S. [2]

[1] Assistant Professor, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore - 641008
[2] Associate Professor, Department of Software Systems and AIML, Sri Krishna Arts and Science College, Coimbatore - 641008
**Corresponding author E-mail: sureshsggr@gmail.com*

Received: April 18, 2025, Accepted: May 21, 2025, Published: May 27, 2025

Abstract

Tuberculosis (TB) remains a formidable issue for worldwide public health and calls for swift and exact diagnostic strategies to achieve the best health results for those affected. A methodical machine learning (ML) sequence was diligently followed, featuring data preprocessing, feature choice, encoding, and the training of the model in a logical order. A detailed investigation was performed on six unique machine learning architectures, comprising the ANN, SVM, Decision Tree, Random Forest, XGBoost, and Logistic Regression, closely analyzing their key performance measures essential for measuring their effectiveness, including accuracy, precision, recall, F1-score, and AUC-ROC, hence providing an extensive view of their attributes and feasible uses across different sectors. The matter of class imbalance was diligently approached through the execution of the Synthetic Minority Over-sampling Technique (SMOTE), and the model's performance was scrutinized using 5-Fold Cross-Validation to affirm both consistency and relevance of the conclusions.
Achieving a stellar accuracy of 99.55%, an impeccable recall of 100%, and a noteworthy F1-score of 99.54%, the ANN model is hailed as the premier model for tuberculosis forecasting. The Random Forest and SVM models also illustrated robust predictive performance, evidenced by elevated accuracy and AUC-ROC scores. In a contrasting view, Logistic Regression provided the least successful outcomes, suggesting that linear models could be inadequately matched to the attributes of this dataset. This study elucidates the efficacy of machine learning methodologies in the diagnostics of TB and emphasizes the critical role of symptom analysis and data-informed decision-making within the healthcare sector.

Keywords: *Tuberculosis Prediction; Machine Learning; Symptom Severity; Artificial Neural Network; Medical Diagnosis.*

1. Introduction

An infectious condition, tuberculosis (commonly referred to as TB), predominantly influences the lungs and is caused by the bacterium Mycobacterium tuberculosis [1]. It represents a considerable global public health issue, with millions of new cases and fatalities documented each year. TB is generally classified into two types. Latent TB infection (LTBI) results from the inactive, symptom-free germs staying in the body. On the other hand, active TB disease develops when the bacteria become active, multiply, and lead to noticeable symptoms[2]. Timely identification and intervention are paramount for mitigating the transmission of TB, as untreated individuals have the potential to infect approximately 10-15 other persons annually. Identifying TB can prove to be a tough task, especially in nations with fewer resources and in settings where healthcare is still developing, owing to the inadequacies of the available diagnostic approaches [3].
Healthcare systems infused with smart technology and analytical tools could potentially empower healthcare providers to swiftly recognize a spectrum of health disorders, including tuberculosis[4]. This investigation centres on the exploration of diverse ML algorithms to facilitate the prognostication of TB. In pursuit of this objective, a comprehensive TB dataset is employed to extract pertinent features, and the algorithms deemed most appropriate were chosen based on defined performance evaluation criteria. This research endeavours to discern the most efficacious ML algorithms for the prediction of TB through necessary analytical processes and empirical testing. It also looks at and contrasts the performance of three selected classifiers—Logistic Regression, K-Nearest Neighbour, and Random Forest—with and without notable features, therefore revealing important information on the most appropriate ML structure for TB prediction.

1.1. Background and motivation

Mycobacterium tuberculosis is the organism tied to the infectious condition called tuberculosis (TB), which can be very perilous [5]. The World Health Organization (WHO) reports that TB ranks among the ten principal etiological factors of mortality and stands as the predominant cause of fatalities attributed to a singular infectious microorganism. In 1993, the WHO designated it as a global public health crisis [6]. In the year 2020, 1.5 million individuals succumbed to TB, reflecting an escalation of 2,00,000 fatalities relative to the preceding year [7]. The growth in tuberculosis fatalities can be explained by a drop in healthcare service availability throughout the COVID-19 pandemic. India bears the highest incidence of TB globally. In the group of 26 High Burden Countries (HBCs) that indicate more than 10,000 projected TB cases, India holds more than a quarter of the international burden, representing 26% of the aggregate incidence and 29% of the entire mortality rate.

The complex nature of TB, coupled with its mode of transmission, makes it challenging and dangerous. TB is an airborne disease spread through airborne droplets. Early detection is crucial because undetected patients can transmit TB to others. The traditional diagnostic methods have pitfalls, especially in resource-limited settings [8]. Chest X-ray images give detailed anatomical information, and certain characteristic radiological patterns are associated with TB. In industrialized nations, chest X-rays are generally conducted for the diagnosis of respiratory diseases [9]. Moreover, huge amounts of data are generated through these tests. Although X-ray images are stored digitally on film or plates, there is also a chance of manipulation and duplication.

In recent years, several machine learning and deep learning algorithms have emerged to analyse these chest X-rays and automatically detect the TB disease. Chest X-ray images are hence enabled as a potential tool for TB diagnosis. Recognizing the importance of chest X-ray images in the TB diagnosis, this investigation is geared towards constructing a TB detection model using the chest X-ray images using the existing machine learning algorithms and demonstrating the robustness of the proposed intelligent and adaptive medical diagnostic system on some of the existing deep learning algorithms.

1.2. Research objectives

This study aims to improve tuberculosis prediction by leveraging readily available patient data, whether positive or negative. The goal is to pinpoint a robust and dependable machine learning algorithm that can effectively analyse tuberculosis datasets. To accomplish this, various machine learning approaches will be explored, each offering distinct advantages in data analysis. The frameworks investigated in this piece comprise logistic regression, decision trees, random forests, gradient boosting, support vector machines, and artificial neural networks, with each technique showcasing merits for analysis execution. But beyond just identifying a capable model, the research also looks at how well these algorithms stack up against commonly used machine learning methods. The focus is not just on finding a working solution but on determining which approach is the most efficient. A vital portion of this consideration is measuring how well the models perform by analysing key measurements like precision, exactness, recall, F1-score, and the Receiver Operating Characteristic (ROC) bend, alongside Range Beneath the Bend (AUC) to evaluate by and large performance. These metrics help provide deeper insights into the overall impact and reliability of the models. These measures help provide deeper insights into the broader impact of the models.

1.3. Significance of the study

Tuberculosis (TB), an airborne infectious disease caused by the bacterium Mycobacterium tuberculosis, resulted in 10.6 billion cases and 1.6 million deaths in 2021 alone. Given that TB is in the third place among infectious disease fatalities, there is a growing need for continuous efforts for active case finding and diagnosis of the disease. Several common obstacles to TB management are faced by endemic countries. Firstly, there are many challenges in TB detection in the lungs due to overlap in radiographic findings with other pathologies such as cancer and pneumonia, especially at later stages of the disease when the X-ray images become more complicated. Secondly, limited access to healthcare facilities in resource-poor countries has hampered control of the disease in regions not well served by expert radiologists. Still, the work involving AI-based detection has mostly taken place in the US or European settings, while such models are urgent for the developing nations, where fully automated solutions can lessen the burden on the healthcare systems. To better contextualize the problem and identify suitable predictive solutions, the following section surveys recent advancements in tuberculosis diagnostics using both traditional and emerging machine learning techniques.

2. Literature review

The comparative analysis of predictive machine learning models for tuberculosis (TB) symptoms data reveals a diverse landscape of methodologies and outcomes. Numerous research endeavours have investigated both traditional machine learning methodologies and sophisticated deep learning approaches to improve the precision and dependability of tuberculosis predictions. Multiple research endeavours have investigated both time-honoured machine learning methods and advanced deep learning strategies to elevate the accuracy and dependability of tuberculosis forecasts. This section delves into the specifics of these models and their comparative effectiveness.

2.1. Classical machine learning models

Decision Trees and Random Forests: Decision trees are preferred by many for their straightforwardness and interpretive clarity. In one study, decision trees achieved an accuracy of 92.11% for early TB detection [10]. Random forests, an ensemble method, showed a slightly higher accuracy of 92% in another study, indicating their robustness in handling TB prediction tasks.

Logistic Regression and Regularized Methods: Logistic regression, along with Lasso, Ridge, and Elastic Net, has been applied to predict TB in HIV patients, achieving accuracies around 86.8% to 87.4% [11]. These techniques are esteemed for their capacity to manage multicollinearity and offer perspectives on the significance of features.

SVM and KNN: In the realm of computational analytics, employing Support Vector Machines (SVM) alongside K-Nearest Neighbours (KNN) has been observed, with SVM demonstrating solid outcomes, though it does not entirely match the efficacy of deep learning approaches [12].

2.2. Deep learning models

CNN: Convolutional Neural Networks exhibit remarkable efficacy in the identification of tuberculosis via chest X-ray images. EfficientNetV2B2 achieved an impressive accuracy of 99.5%, highlighting the potential of CNNs in medical imaging applications [13]. An ensemble approach combining different CNN architectures further improved accuracy to 95.14% [14].

LSTM: Long Short-Term Memory Networks models have been used to predict TB co-infection among HIV patients, achieving AUC-ROC values between 0.827 and 0.850, indicating their effectiveness in handling time-series data [13].

Hybrid and Ensemble Models: The merging of several sophisticated deep learning architectures, like Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs), yields a notable enhancement in efficiency against independent models, particularly in consistently detecting intricate patterns tied to datasets regarding tuberculosis and HIV co-infection [15].

2.3. Comparative insights

Performance Metrics: In terms of performance evaluation, CNNs are usually more effective than classical machine learning models, particularly noticeable in their accuracy and sensitivity metrics. As a case in point, CNNs secured better accuracy and AUC-ROC scores when evaluated against decision trees and logistic regression methods [12] [13].
Data Types and Modalities: The choice of data type significantly influences model performance. Models that incorporate chest X-ray imagery or multimodal datasets generally demonstrate superior performance compared to those that depend exclusively on clinical or demographic information [16] [17].
Computational Efficiency: While deep learning models offer higher accuracy, they require more computational resources compared to classical models, which are faster and easier to implement[13].

2.4. Emerging trends in TB diagnosis

Recent advancements include Graph Neural Networks (GNNs), which capture relational information from structured clinical or contact-tracing data and have shown promise in disease spread modeling. Additionally, non-machine learning diagnostics such as biomarker-based assays—particularly those leveraging host blood transcriptomic signatures, have gained attention for their accuracy and rapidity in TB diagnosis, especially in resource-limited settings. The integration of such innovations with machine learning pipelines could potentially enhance predictive performance and broaden clinical utility.
GNNs are becoming more prevalent in medical imaging, particularly for identifying TB, due to their unique capability to understand varied spatial patterns and their interrelations. For instance, researchers have proposed using Region Adjacency Graphs (RAGs) where each superpixel in a chest X-ray (CXR) serves as a node in the graph. This approach effectively captures crucial features and relationships among superpixels, facilitating advancements in TB identification[18].
A distinguished tactic illustrates the unification of graph convolutional networks (GCNs) with classic convolutional neural networks (CNNs). This combined strategy enables the consolidation of both local and global attributes, improving the model's capacity to recognize subtle irregularities in chest CT images [19] [20].
In contrast to the predominant focus on deep learning, some studies emphasize the importance of classical models for their interpretability and lower computational demands. These frameworks may prove especially beneficial in contexts with limited resources where the implementation of deep learning infrastructure is impractical. Additionally, the amalgamation of multimodal datasets and the creation of hybrid frameworks signify encouraging avenues for prospective investigations, striving to achieve a harmonious balance between precision and practical usability across varied healthcare settings. Having established the landscape of existing TB prediction models, the following section outlines the methodological approach employed in this study, including dataset characteristics, preprocessing techniques, and the machine learning frameworks applied.

3. Materials and methods

This consider utilizes an assortment of machine learning (ML) algorithms to determine the frequency of tuberculosis (TB) by dissecting symptomatology information assembled from both contemporaneous sources and pre-existing online repositories. The framework has six ML models as mentioned above in the research objective. The ensuing segments explain the technique and application for the determination of these calculations. Each show was prepared on a portion of the dataset and assessed utilizing different performance metrics such as precision, accuracy, recall, F1-score, and AUC-ROC. The dataset was part into 80 and 20 percent for preparing and testing sets, with K-fold cross-validation utilized to guarantee the vigor and unwavering quality of the results.

3.1. Data descriptions

The dataset utilized in this study consists of 1,452 records, prepared by combining the "Tuberculosis Dataset for Intelligent and Adaptive Medical Diagnostic System" [21] and real-world data collected via Google Forms. The TB symptoms dataset from Mendeley includes 568 records, all of which are TB-positive cases. The residual data were obtained from both individuals who tested positive for tuberculosis and those who tested negative via a systematically designed online questionnaire.

Table 1: Symptoms and their Code Definition [21]

S. N.	TB Symptoms /Attributes	Code Definition
1	Cough	CO
2	Night sweats	NS
3	Difficulty in Breathing	DB
4	Fever	FV
5	Chest Pain	CP
6	Sputum	SP
7	Loss of pleasure	LP
8	Chills	CH
9	Lack of concentration	LC
10	Irritation	IR
11	Loss of appetite	LA
12	Loss of energy	LE
13	Lymph Node Enlargement	LNE
14	Systolic Blood Pressure	SBP
15	Body Mass Index	BMI
16	Immune Suppression	IS

3.1.1. Dataset structure and attributes

Table 1 [21] delineates the symptoms along with their corresponding coded definitions for enhanced accessibility. The dataset encompasses 16 pivotal symptoms and physiological markers associated with tuberculosis (TB), organized into four distinct categories:

- Respiratory manifestations encompass CO (Cough), DB (Difficulty in Breathing), CP (Chest Pain), and SP (Sputum Type).
- The systemic indicators include FV (Fever), NS (Night Sweats), CH (Chills), IS (Immune Suppression), and LNE (Lymph Node Enlargement).
- LP (Loss of Pleasure), LC (Lack of Concentration), and IR (Irritation) are signs reflecting neurological and psychological conditions.
- The worry surrounding meal selection and fitness tracking features elements like LA (Loss of Appetite), LE (Loss of Energy/Fatigue), BMI (Body Mass Index), and SBP (Systolic Blood Pressure).

For every symptom, a particular numerical code is designated that illustrates severity levels classified as None (0), Mild (1), Moderate (2), and Severe (3), except for the Sputum Type. The Prediction column signifies whether a patient is identified as TB-positive (1) or TB-negative (0) [22]. Table 3 [21] illustrates the encoded severity levels of symptoms. This structured dataset facilitates a multimodal strategy for TB detection, incorporating a variety of symptom patterns that extend beyond conventional respiratory indicators.

3.1.2. Real-world data collection and preprocessing

The real-world data collection involved 16 symptom parameters, including coughing with blood, fever at night, weight loss, and loss of appetite. A scale measuring severity in four levels ('None', 'Mild', 'Moderate', or 'Severe') was utilized by respondents to answer inquiries regarding symptoms. Information was amassed through various social media outlets, including Facebook, Twitter, Instagram, and WhatsApp, along with personal engagements with individuals. To guarantee uniformity and exactness, the dataset underwent preprocessing, cleaning, and transformation ahead of being applied in the study.

3.1.3. Sample dataset and TB prediction trends

Table 2 presents a subset of five patient records, illustrating variations in symptom severity and TB prediction outcomes. The examination reveals that persons exhibiting elevated severity levels across various symptoms (notably respiratory manifestations such as Cough, Sputum Character, and Breathing Impairment) are inclined to test positive for tuberculosis (Prediction = 1), whereas individuals presenting with diminished symptom severity are more predisposed to test negative for tuberculosis (Prediction = 0).

Table 2: The Actual Sample Data

Feature	Severity Level				
CO	mild	mild	moderate	mild	mild
NS	severe	mild	moderate	mild	none
BD	severe	severe	severe	moderate	none
FV	moderate	mild	mild	severe	mild
CP	severe	moderate	severe	mild	none
SP	bloody	colourless	bloody	bloody	none
IS	moderate	mild	moderate	moderate	none
LP	moderate	moderate	severe	mild	mild
CH	moderate	mild	severe	mild	none
LC	severe	severe	severe	moderate	none
IR	severe	severe	mild	moderate	none
LA	moderate	mild	moderate	mild	none
LE	moderate	mild	mild	moderate	none
LNE	moderate	mild	severe	moderate	none
SBP	mild	severe	severe	moderate	none
BMI	mild	severe	mild	mild	none
Prediction	Yes	Yes	Yes	Yes	No

Table 3: Encoding Parameters with Severity Level [1]

S. N.	Feature	Feature values	Feature Encoding
1	CO	“Severe”, “mild”, “moderate”	“None” =0, “mild” = 1, “moderate” =2, “severe” = 3
2	NS	“Severe”, “mild”, “moderate”	“None” =0, “mild” = 1, “moderate” =2, “severe” = 3
3	BD	“Severe”, “mild”, “moderate”	“None” =0, “mild” = 1, “moderate” =2, “severe” = 3
4	FV	“Severe”, “mild”, “moderate”	“None” =0, “mild” = 1, “moderate” =2, “severe” = 3
5	CP	“Severe”, “mild”, “moderate”	“None” =0, “mild” = 1, “moderate” =2, “severe” = 3
6	SP	“bloody”, “colourless”, “green”	“None” =0, “colourless” =1 , “green” = 2, “bloody” = 3
7	IS	“Severe”, “mild”, “moderate”	“None” =0, “mild” = 1, “moderate” =2, “severe” = 3
8	LP	“Severe”, “mild”, “moderate”	“None” =0, “mild” = 1, “moderate” =2, “severe” = 3
9	CH	“Severe”, “mild”, “moderate”	“None” =0, “mild” = 1, “moderate” =2, “severe” = 3
10	LC	“Severe”, “mild”, “moderate”	“None” =0, “mild” = 1, “moderate” =2, “severe” = 3
11	IR	“Severe”, “mild”, “moderate”	“None” =0, “mild” = 1, “moderate” =2, “severe” = 3
12	LA	“Severe”, “mild”, “moderate”	“None” =0, “mild” = 1, “moderate” =2, “severe” = 3
13	LE	“Severe”, “mild”, “moderate”	“None” =0, “mild” = 1, “moderate” =2, “severe” = 3
14	LNE	“Severe”, “mild”, “moderate”	“None” =0, “mild” = 1, “moderate” =2, “severe” = 3
15	SBP	“Severe”, “mild”, “moderate”	“None” =0, “mild” = 1, “moderate” =2, “severe” = 3
16	BMI	“Severe”, “mild”, “moderate”	“None” =0, “mild” = 1, “moderate” =2, “severe” = 3
17	Prediction	“Positive”, “Negative”	“Positive” Yes = 1, “Negative” No = 0

Furthermore, a subset of four patient records was analysed to demonstrate symptom severity patterns. Findings reveal that severe respiratory symptoms (DB, CP, SP) strongly correlate with TB-positive cases. Moreover, neurological and psychological symptoms (LC, IR, LP) frequently appear, suggesting a possible cognitive impact of TB. Notably, variations in Sputum Type (SP), particularly the presence of "bloody sputum," serve as a strong TB indicator [23]. In this subset, all four patients were predicted as TB-positive, reinforcing the strong relationship between symptom severity and TB presence.

3.1.4. Encoding and machine learning suitability

For machine learning applications, all categorical symptoms and attributes were numerically encoded to ensure computational efficiency. The encoding scheme assigns severity levels from None (0), Mild (1), Moderate (2), to Severe (3) for most symptoms. Sputum Type is encoded separately as None (0), Colourless (1), Green (2), and Bloody (3) to reflect its clinical significance. The Prediction outcome is encoded as Positive (Yes = 1) and Negative (No = 0), allowing for a straightforward classification task. Table 4 presents the final encoded and organized sample data.

This structured encoding enhances feature scaling, pattern recognition, and statistical analysis, making the dataset highly suitable for machine learning-based TB prediction models [24]. The integration of diverse symptom attributes, real-world patient data, and robust preprocessing techniques ensures that the dataset can effectively contribute to early TB diagnosis, risk assessment, and predictive modelling.

Table 4: Encoded Sample Data

Feature	Severity Level				
CO	1	1	2	1	1
NS	3	1	2	1	0
BD	3	3	3	2	0
FV	2	1	1	3	1
CP	3	2	3	1	0
SP	3	1	3	3	0
IS	2	1	2	2	0
LP	2	2	3	1	1
CH	2	1	3	1	0
LC	3	3	3	2	0
IR	3	3	1	2	0
LA	2	1	2	1	0
LE	2	1	1	2	0
LNE	2	1	3	2	0
SBP	1	3	3	2	0
BMI	1	3	1	1	0
Prediction	1	1	1	1	0

3.2. Methodology

This examination leverages an organized machine learning (ML) methodology to evaluate the occurrence of tuberculosis (TB) using data on the severity of clinical symptoms. The methodological workflow diagram is illustrated in Figure 1. It adheres to a systematic sequence that includes data acquisition, preprocessing, model training, assessment, and performance benchmarking. The objective of the research is to ascertain the most proficient ML model for the prompt diagnosis of TB.

3.2.1. Data collection

The dataset for this study is compiled from two primary sources to ensure diversity and completeness:

Mendeley Dataset: A publicly available dataset obtained from Mendeley Data [21], comprising TB patient records, including symptom severity, demographic details, and medical history.

Google Forms Data (Real-Time Collection): A real-time dataset collected through Google Forms, where participants provided symptom severity ratings, medical history, and demographic details. This dataset includes both TB-positive and TB-negative cases, verified using medical reports.

Data Integration: The two datasets were merged to enhance variability, improve model generalization, and increase the robustness of TB prediction. The combined dataset ensures a more comprehensive representation of TB cases across diverse populations.

3.2.2. Data preprocessing

Data preprocessing was performed to improve data quality and ensure consistency. The following steps were applied:

Handling Missing Values: Records with excessive missing values were removed to prevent noise. Missing values in key clinical variables were imputed using mean, median, or mode imputation, depending on data distribution.

Feature Engineering: Feature Selection: The most relevant features, including symptom severity scores and patient history, were identified using statistical correlation and domain expertise [25].

Categorical Encoding: Categorical variables (e.g., presence of cough, weight loss) were transformed into numerical representations using one-hot encoding and label encoding.

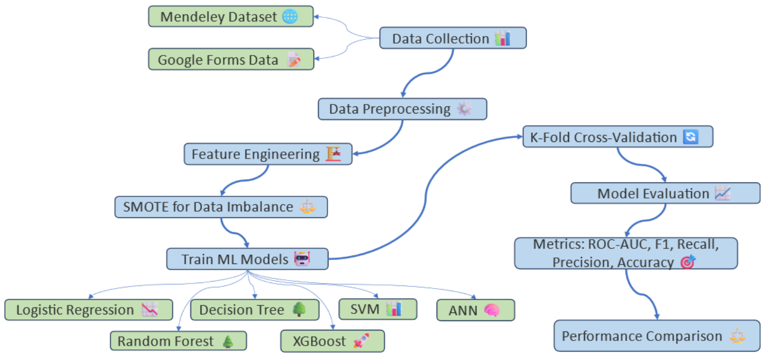

Fig. 1: Workflow Diagram of Tuberculosis Prediction Using Machine Learning.

Feature Scaling: Numerical features were normalized using Min-Max scaling to ensure uniform feature distribution.

3.2.3. Data splitting

The dataset was divided into: Training Set (80%) – Used for model training. Testing Set (20%) – Used to evaluate model performance on unseen data. A stratified split was used to maintain class distribution across training and testing sets.

3.2.4. Handling class imbalance

In advance of training the machine learning models for this research, the dataset underwent balancing through the Synthetic Minority Over-sampling Technique (SMOTE) [26]. The resampled dataset ensured that all models learned from an equal distribution of TB-positive and TB-negative cases, leading to more reliable predictions. Since TB cases were imbalanced (fewer positive cases than negative ones), the SMOTE was applied to:

- Generate synthetic samples of TB-positive cases.
- Balance the dataset to reduce model bias toward the majority (negative) class.
- Improve generalization and predictive accuracy.

SMOTE was applied only to the training set to prevent data leakage and ensure unbiased testing.

3.2.5. SMOTE

The method called SMOTE, representing Synthetic Minority Over-sampling Technique, is extensively utilized to tackle class imbalances present in datasets employed in machine learning. Class imbalance occurs when one class has significantly fewer samples than the other, leading to biased models that favour the majority class [27]. In medical datasets, such as Tuberculosis (TB) prediction, positive cases are often fewer than negative cases. Without balancing, machine learning models may:

- Ignore the minority class (e.g., TB-positive cases)
- Show artificially high accuracy by predicting only the majority class
- Fail to generalize well in real-world applications

Unlike simple oversampling (which duplicates minority class samples), SMOTE generates synthetic data points to create a more balanced dataset. The SMOTE process [28] includes:

1) Determining K-nearest neighbours (KNN) specifically for a minority class sample.
2) Selecting a random neighbour from the K-nearest ones.
3) Creating a new synthetic sample by interpolating between the sample and its selected neighbour.
4) Repeating the process until the minority class reaches a balanced proportion.

3.2.6. Machine learning models

Six ML models were trained and evaluated to determine the most effective TB prediction model:

1) Logistic Regression (LR): A statistical model suitable for binary classification.
2) The concept of Decision Tree (DT) involves a rule-oriented system that organizes data to form a classification tree.
3) Support Vector Machine (SVM): A method that prioritizes the margin to elevate class separation effectiveness.
4) Artificial Neural Network (ANN): An advanced learning architecture proficient in discerning intricate patterns within tuberculosis data.
5) Random Forest (RF): An ensemble methodology that augments classification precision through the aggregation of multiple decision trees.
6) The boosting algorithm called XGBoost, short for Extreme Gradient Boosting, is finely tuned to provide high efficiency alongside swift computational performance.

Models were executed with the aid of scikit-learn and XGBoost in Python, focusing on hyperparameter adjustments to elevate overall performance.

3.3. Model training and evaluation

3.3.1. K-fold cross-validation

K-Fold Cross-Validation [29] (CV) constitutes a rigorous methodology for the assessment of models, thereby ensuring that outcomes are not contingent upon an isolated train-test division. It boosts how well machine learning models can generalize. In the realm of machine learning, reliance on a solitary train-test split may result in skewed findings. Should the test set be excessively simplistic or overly challenging, the reported accuracy may fail to accurately represent actual performance in real-world scenarios [30].

- K-Fold ensures that each sample is used for both training and testing at least once.
- This results in more stable and reliable performance metrics.

K-Fold Cross-Validation Process

1) K equal-sized sections, or folds, are formed by randomly splitting the dataset.
2) The model is developed utilizing K-1 folds and subsequently evaluated on the remaining single fold.
3) This methodology is executed K times, with each iteration incorporating a distinct test fold.
4) The outcome is represented as the mean of all K test results, thereby mitigating bias.

The operation is executed K times, changing the test fold on every occasion. The dataset was segmented into five folds, wherein each model was trained on K-1 folds and assessed on the remaining fold. This procedure was conducted five times, with performance metrics aggregated across all iterations [31]. This strategy contributed to the minimization of overfitting, facilitating a robust estimation of model performance, and ensuring an equitable comparison across various models.

3.3.2. Evaluation metrics

Each model underwent a comprehensive evaluation utilizing a variety of metrics [32] to determine classification efficacy:

- Accuracy: Quantifies the ratio of accurately classified instances.
- Precision: Assesses the number of predicted cases of tuberculosis that were indeed positive.
- Recall (Sensitivity): Evaluates the capacity of the model to accurately identify cases of tuberculosis that are positive.
- F1 Score: Represents a harmonic mean of precision and recall, effectively balancing the occurrences of false positives and false negatives.
- Receiver Operating Characteristic – Area Under Curve, often abbreviated as ROC-AUC, gauges the capability of a model in separating tuberculosis-positive cases from those that are tuberculosis-negative.

The evaluation of performance metrics for each model allowed us to pinpoint the model that showed superior results based on the previously mentioned benchmarks.

3.3.3. Performance comparison and model selection

After a comprehensive evaluation, the models were assessed based on:

- Predictive accuracy and robustness.
- Nuanced trade-offs regarding precision, recall, and F1-metric.
- Relevance and applicability in practical tuberculosis screening contexts.

The model that exhibited the highest level of predictive accuracy, effectively balanced the trade-offs between precision and recall, and demonstrated optimal generalization performance was designated as the preeminent model for tuberculosis prediction.

3.4. Implementation and software tools

The complete workflow was developed employing the Python programming language (Jupyter Notebook) [33]. The following libraries and tools were utilized:

- Pandas, NumPy: For data manipulation and preprocessing.
- Scikit-learn: For machine learning model implementation, feature selection, and evaluation.
- XGBoost: For implementing gradient boosting models.
- Imbalanced-learn: For applying SMOTE to handle class imbalance.
- Matplotlib, Seaborn: For data visualization and performance analysis.

The experiments were conducted on a system with the following specifications:

- Processor: Intel Core i5/ Ryzen 5
- RAM: 8 GB
- System Type: 64-bit operating System
- Edition: Windows 11 Home Single Language
- Version: 24H2

3.5. Ethical considerations

Since real-time patient data was collected through Google Forms, ethical guidelines[34] were followed:

- Informed Consent: All participants provided consent for data collection and analysis.
- Data Anonymization: Personal identifiers were removed to maintain confidentiality.
- Approval from Ethical Review Board: The research was subjected to scrutiny and subsequently authorized by the pertinent institutional review board.

This segment delineates the systematic approach employed to formulate and assess machine learning-oriented tuberculosis prediction models. By integrating diverse data sources, applying advanced preprocessing techniques, and employing robust evaluation metrics, this workflow ensures an optimized, reliable, and scalable ML model for early TB diagnosis. With the methodological foundation established, the next section presents and analyses the results obtained from various machine learning models, emphasizing performance comparisons and key findings.

4. Results and discussions

From a public health perspective, predictive TB models like those proposed in this study can support targeted screening strategies, especially in high-burden regions. When integrated with national TB programs, such tools could aid in optimizing resource allocation, improving early case detection, and reducing transmission through timely interventions. These models also have potential applications in epidemiological surveillance, enabling real-time identification of outbreak clusters based on symptom data. This research investigated the efficacy of six distinct machine learning frameworks in forecasting tuberculosis (TB) through numerical clinical information. The examination detailed an assortment of techniques that integrate Logistic Regression (LR), Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and XGBoost, drawing focus to crucial innovations in analytical architectures. By applying the Synthetic Minority Over-sampling Technique (SMOTE), the dataset experienced a method aimed at equalizing its distribution, seeking to fix the imbalances in class representation. To ascertain how effectively the model operates, we performed an evaluation utilizing K-Fold cross-validation, looking at considerations like accuracy, precision, recall, F1-score, AUC-ROC, and numerous components from the confusion matrix (TP, TN, FP, FN) as our assessment criteria. Table 5 displays the performance metrics, whereas Figure 2 presents a distinct and illustrative representation of the performance analysis through the utilization of a heatmap.

Table 5: Performance Metrics of Machine Learning Models for TB Prediction

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC	TP	TN	FP	FN
Logistic Regression	97.29	96.60	98.05	97.31	98.89	195	192	7	4
Decision Tree	99.25	98.90	99.59	99.24	99.24	198	196	3	1
Random Forest	99.35	99.29	99.41	99.35	99.99	197	197	1	1
SVM	99.45	99.10	99.79	99.44	99.94	198	197	2	1
ANN	99.55	99.08	100.00	99.54	99.69	199	197	1	0
XGBoost	99.45	99.10	99.79	99.44	99.94	198	197	2	1

4.1. Key observations

Accuracy

- The most precise result was attained through ANN at 99.55%, with SVM trailing just behind at 99.45%, alongside XGBoost also at 99.45%, and Random Forest at 99.35%.
- Conversely, Logistic Regression demonstrated the least accuracy at 97.29%, suggesting that linear models may exhibit diminished efficacy within the context of this dataset.
- Recall (Sensitivity)
- ANN (100.00%) achieved the highest recall, meaning it correctly identified all positive TB cases.
- Logistic Regression (98.05%) had the lowest recall, meaning it missed more positive TB cases compared to the others.

Figure 3 presents the Precision-Recall curves corresponding to each model, thereby offering a comprehensive understanding of their capacity to achieve an equilibrium between precision and recall at varying threshold configurations.

F1-Score

- The highest F1-score (99.54%) was obtained by ANN, which balances precision and recall effectively.
- Logistic Regression (97.31%) had the lowest F1-score.
- AUC-ROC
- The Random Forest model (99.99%) attained the superior AUC-ROC metric, signifying an exceptional proficiency in differentiating between tuberculosis (TB) and non-TB instances.
- Conversely, Logistic Regression (98.89%) exhibited the least performance, yet it still demonstrated commendable efficacy.

Figure 4 delineates the AUC-ROC curves associated with each model, thus demonstrating their comprehensive efficacy in differentiating between affirmative and negative instances.

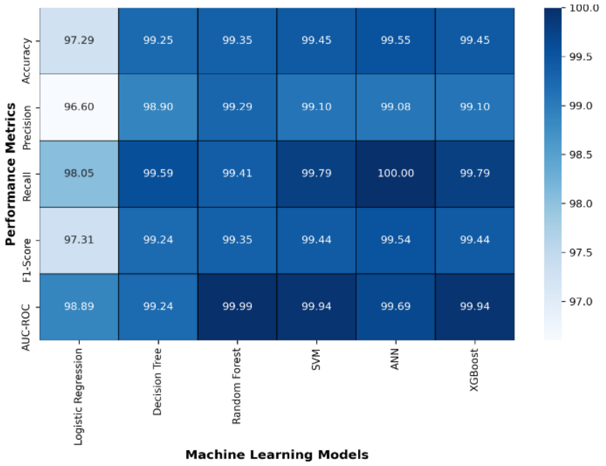

Fig. 2: Heatmap Showing Performance Metrics for Each ML Model. Colour Intensity Indicates Relative Performance, with Darker Shades Representing Higher Values.

4.2. Confusion matrix analysis

- True Positives (TP): ANN correctly classified 199 TB cases, followed by SVM and XGBoost (198 cases).
- True Negatives (TN): The models performed similarly in correctly classifying non-TB cases.
- False Positives (FP): Decision Tree (3 FP) and Logistic Regression (7 FP) had the most misclassified negatives.
- False Negatives (FN): ANN (0 FN) detected all TB cases, while Logistic Regression (4 FN) missed the most.

The following conclusion synthesizes the major findings of the study and discusses their implications for tuberculosis diagnosis, model deployment in clinical settings, and future research directions.

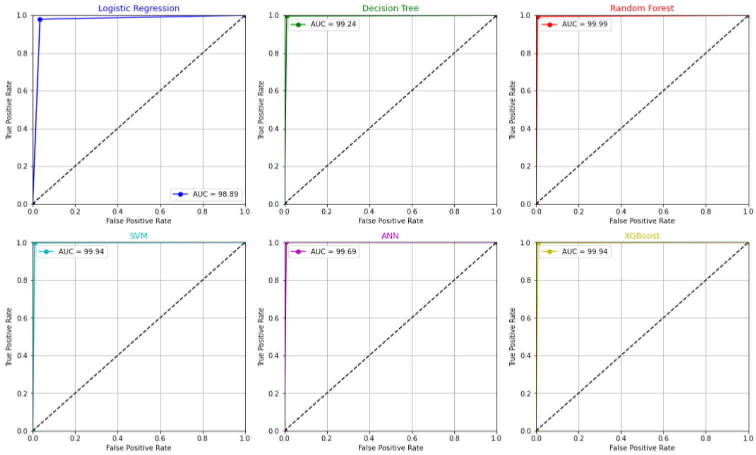

Fig. 4: AUC-ROC Curves of Each Model.

5. Conclusion

The chosen machine learning models present a thorough methodology for forecasting tuberculosis through symptomatology. Each model possesses distinct benefits, and their efficacy is rigorously evaluated to identify the most proficient algorithm for this healthcare diagnostic purpose. The analysis points out the essential requirement for utilizing multiple machine learning frameworks to advance the predictive capabilities of sophisticated diagnostic systems in healthcare. The Artificial Neural Network (ANN) stands out as the top-tier model, achieving a recall rate that reaches an outstanding 100%, boasting remarkable accuracy at 99.55%, and a nearly flawless F1-score of 99.54%. This model is particularly advantageous for medical contexts where the minimization of false negatives is imperative. The Random Forest and Support Vector Machine (SVM) models serve as formidable alternatives, both exhibiting substantial accuracy and AUC-ROC values. Conversely, Logistic Regression demonstrates inferior performance, rendering it less appropriates for tuberculosis prediction within this dataset. The ANN is advocated for tuberculosis screening due to its proficiency in identifying intricate patterns. Should computational efficiency be a priority, Random Forest or SVM may be employed, as they yield comparable results yet potentially demand fewer resources. To confirm their utility, these models need to be evaluated with a real-world dataset ahead of clinical implementation. Active research initiatives will focus on affirming these models through larger and more diverse datasets to better their application in healthcare settings. While the ANN model demonstrated superior performance on this dataset, its real-world applicability may be constrained by computational complexity, data heterogeneity, and infrastructure limitations. Hence, further validation on large-scale, diverse clinical datasets is essential before clinical deployment.

Acknowledgement

We express our sincere gratitude to Sri Krishna Arts and Science College for supporting the research program and providing the computational resources essential for this study. We also acknowledge the significant contribution of open-access platforms and repositories like Kaggle, Mendeley in advancing research accessibility. Finally, we would like to thank the Doctors, Dr. Vibinesh and Dr. Naveen Prabhu for their valuable inputs for this study.

Ethical approval

The corresponding author has obtained approval from the Ethical Review Board of the institution.

Conflict of interest

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

Funding information

The authors declare that there is no funding to be acknowledged.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The Mendeley data are publicly available online: https://data.mendeley.com/datasets/ndxdx54xxx/1.

References

[1] V. Škodrić-Trifunović, “Tuberculosis: Old and new disease,” *Galen. Med. J.*, vol. 1, no. 4, pp. 40–47, 2022, https://doi.org/10.5937/Galmed2204042S.

[2] P. Etienne, “Does ‘Latent Tuberculosis Infection (LTBI)’ Really Exist? Genealogy of a Medical Nosology,” *J. Tuberc. Res.*, vol. 09, no. 03, pp. 197–204, 2021, https://doi.org/10.4236/jtr.2021.93018.

[3] S. Mukherjee, S. Perveen, A. Negi, and R. Sharma, “Evolution of tuberculosis diagnostics: From molecular strategies to nanodiagnostics,” *Tuberculosis*, vol. 140, p. 102340, 2023, https://doi.org/10.1016/j.tube.2023.102340.

[4] I. Pavan Kumar, R. Mahaveerakannan, K. Praveen Kumar, I. Basu, T. C. Anil Kumar, and M. Choche, “A Design of Disease Diagnosis based Smart Healthcare Model using Deep Learning Technique,” *Proc. Int. Conf. Electron. Renew. Syst. ICEARS 2022*, pp. 1444–1449, 2022, https://doi.org/10.1109/ICEARS53579.2022.9752063.

[5] I. N. Al-Asady and J. F. Ali, “Review Article: Virulence Factors of Mycobacterium Tuberculosis,” *J. Res. Appl. Sci. Biotechnol.*, vol. 2, no. 3, pp. 221–237, 2023, https://doi.org/10.55544/jrasb.2.3.31.

[6] M. Masand, P. Kumar Sharma, V. M. Balaramnavar, and D. Mathpal, “Tuberculosis: Current Progress in Drug Targets, Potential Drugs and Therapeutic Impact,” *Curr. Respir. Med. Rev.*, vol. 18, no. 3, pp. 165–170, 2022, https://doi.org/10.2174/1573398X18666220503184459.

[7] T. Zhang *et al.*, “The global, regional, and national burden of tuberculosis in 204 countries and territories, 1990–2019,” *J. Infect. Public Health*, vol. 16, no. 3, pp. 368–375, 2023, https://doi.org/10.1016/j.jiph.2023.01.014.

[8] J. Ma *et al.*, “Rapid detection of airborne protein from: Mycobacterium tuberculosis using a biosensor detection system,” *Analyst*, vol. 147, no. 4, pp. 614–624, 2022, https://doi.org/10.1039/D1AN02104D.

[9] A. Shirsat, S. Kute, R. Haral, A. Patil, and D. S. A. Ubale, “Tuberculosis Detection Using Chest X-Ray with Deep Learning and Visualization,” *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 5, pp. 3888–3894, 2023, https://doi.org/10.22214/ijraset.2023.51440.

[10] P. Karmani, A. A. Chandio, I. A. Korejo, O. W. Samuel, and M. Aborokbah, “Machine learning based tuberculosis (ML-TB) health predictor model: early TB health disease prediction with ML models for prevention in developing countries,” *PeerJ Comput. Sci.*, vol. 10, pp. e2397–e2397, 2024, https://doi.org/10.7717/peerj-cs.2397.

[11] J. orwa *et al.*, “Comparison of logistic regression with regularized machine learning methods for the prediction of tuberculosis disease in people living with HIV: cross-sectional hospital-based study in Kisumu County, Kenya,” *Research Square*. 2023, https://doi.org/10.21203/rs.3.rs-3354948/v1.

[12] G. Landry, R. N. Malumba, F. C. B. Kabutakapua, and B. B. Mangata, “Performance comparison of classical algorithms and deep neural networks for tuberculosis prediction,” *J. Techno Nusa Mandiri*, vol. 21, no. 2, pp. 126–133, 2024, https://doi.org/10.33480/techno.v21i2.5609 .

[13] S. soam, “Comparative Study of Keras CNNs for Tuberculosis Detection from Chest X-rays,” *Interantional J. Sci. Res. Eng. Manag.*, vol. 08, no. 05, pp. 1–5, 2024, https://doi.org/10.55041/IJSREM34126.

[14] T. Varshith, T. S. Koneri, T. S. K. Reddy, and R. P. Singh, “An Ensemble Approach to Tuberculosis Prediction using Shenzhen and Montgomery Datasets,” in *2024 15th International Conference on Computing Communication and Networking Technologies, ICCCNT 2024*, 2024, pp. 1–7, https://doi.org/10.1109/ICCCNT61001.2024.10725732.

[15] J. Chen *et al.*, “LSTM-Based Prediction Model for Tuberculosis Among HIV-Infected Patients Using Structured Electronic Medical Records: A Retrospective Machine Learning Study,” *J. Multidiscip. Healthc.*, vol. 17, pp. 3557–3573, 2024, https://doi.org/10.2147/JMDH.S467877.

[16] A. Sambarey *et al.*, “Integrative analysis of multimodal patient data identifies personalized predictors of tuberculosis treatment prognosis,” *iScience*, vol. 27, no. 2, 2024, https://doi.org/10.1016/j.isci.2024.109025.

[17] R. S. Prasad, R. C. Waghmare, T. B. Pajgade, R. R. Raut, and M. L. Mahajan, “A Comparative Study of Detection of Tuberculosis using Machine Learning & Deep Learning,” in *Proceedings of the 17th INDIACom; 2023 10th International Conference on Computing for Sustainable Global Development, INDIACom 2023*, 2023, pp. 1217–1221.

[18] R. Pradhan and K. M. Santosh, “Analyzing Pulmonary Abnormality with Superpixel Based Graph Neural Network in Chest X-Ray,” pp. 97–110, https://doi.org/10.1007/978-3-031-53085-2_9.

[19] S. Wang, V. Govindaraj, J. M. Górriz, X. Zhang, and Y. Zhang, “Explainable diagnosis of secondary pulmonary tuberculosis by graph rank-based average pooling neural network,” *J. Ambient Intell. Humaniz. Comput.*, pp. 1–14, 2021, https://doi.org/10.1007/s12652-021-02998-0.

[20] Y.-X. Yu, Z. Qi, L. Xu, and X. Zhou, “Research on Deep Learning-Based Algorithms for Medical Image Characterisation,” 2024,

[21] S. Ohwo, F. Eze, F. Onu, and M. Julius, “Tuberculosis Dataset for Intelligent and Adaptive Medical Diagnostic System,” vol. V1, 2023,

[22] M. Zhan *et al.*, “A clinical indicator-based prognostic model predicting treatment outcomes of pulmonary tuberculosis: a prospective cohort study,” *BMC Infect. Dis.*, vol. 23, no. 1, 2023, https://doi.org/10.1186/s12879-023-08053-x.

[23] B. Mtafya *et al.*, “Systematic assessment of clinical and bacteriological markers for tuberculosis reveals discordance and inaccuracy of symptom-based diagnosis for treatment response monitoring,” *Front. Med.*, vol. 9, 2022, https://doi.org/10.3389/fmed.2022.992451.

[24] C. Liu, I. Cohen, R. Vishinkin, and H. Haick, “Nanomaterial-Based Sensor Array Signal Processing and Tuberculosis Classification Using Machine Learning,” *J. Low Power Electron. Appl.*, vol. 13, no. 2, p. 39, 2023, https://doi.org/10.3390/jlpea13020039.

[25] N. Shakhovska and N. Melnykova, “Feature Engineering and Missing Data Imputation Method of Medical Data Analysis,” *CEUR Workshop Proc.*, vol. 3137, pp. 48–57, 2022, https://doi.org/10.32782/cmis/3137-4.

[26] G. Wei, W. Mu, Y. Song, and J. Dou, “An improved and random synthetic minority oversampling technique for imbalanced data,” *Knowledge-Based Syst.*, vol. 248, p. 108839, 2022, https://doi.org/10.1016/j.knosys.2022.108839.

[27] S. Maldonado, C. Vairetti, A. Fernandez, and F. Herrera, “FW-SMOTE: A feature-weighted oversampling approach for imbalanced classification,” *Pattern Recognit.*, vol. 124, p. 108511, 2022, https://doi.org/10.1016/j.patcog.2021.108511.

[28] D. Elreedy, A. F. Atiya, and F. Kamalov, “A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning,” *Mach. Learn.*, vol. 113, no. 7, pp. 4903–4923, 2024, https://doi.org/10.1007/s10994-022-06296-4.

[29] B. Zhu and Y. Liu, “General Approximate Cross Validation for Model Selection: Supervised, Semi-supervised and Pairwise Learning,” *MM 2021 - Proc. 29th ACM Int. Conf. Multimed.*, pp. 5281–5289, 2021, https://doi.org/10.1145/3474085.3475649.

[30] A. A. Khan, “Balanced Split: A new train-test data splitting strategy for imbalanced datasets,” *arXiv.org*, vol. abs/2212.1, 2022,

[31] M. K. Hasan *et al.*, “Challenges of deep learning methods for COVID-19 detection using public datasets,” *Informatics Med. Unlocked*, vol. 30, p. 100945, 2022, https://doi.org/10.1016/j.imu.2022.100945.

[32] X. Qiu, S. Zheng, J. Yang, G. Yu, and Y. Ye, “Comparing Mycobacterium tuberculosis RNA Accuracy in Various Respiratory Specimens for the Rapid Diagnosis of Pulmonary Tuberculosis,” *Infect. Drug Resist.*, vol. 15, pp. 4195–4202, 2022, https://doi.org/10.2147/IDR.S374826.

[33] S. H. Mostafaei, J. Tanha, N. Samadi, S. Imanzadeh, and N. Razzaghi-Asl, “A boosting based approach to handle imbalanced data,” *2022 30th Int. Conf. Electr. Eng. ICEE 2022*, pp. 295–299, 2022, https://doi.org/10.1109/ICEE55646.2022.9827026 .

[34] I. Kassam, D. Ilkina, J. Kemp, H. Roble, A. Carter-Langford, and N. Shen, “Patient Perspectives and Preferences for Consent in the Digital Health Context: State-of-the-art Literature Review,” *J. Med. Internet Res.*, vol. 25, p. e42507, 2023, https://doi.org/10.2196/42507.