International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(23s) 2026 pp. 179-190
https://doi.org/10.70917/ijcisim-2026-5563
Article

Development of a Predictive Model for Tuberculosis (TB) in a Small Town Like Kapurthala, Punjab Using AI

Jessy Juian1*, Sudhakar Ranjan[2], Roopali Sood[3]

1*Research Scholar, Apeeay Stya University. Email: jessymoljulian@gmail.com.
[2]Apeejay Stya University. Email: sudhakar.ranjan@asu.apeejay.edu.
[3]Apeejay College of Fine Arts. Email: Roopali.sood@learn.apeejay.edu.

Abstract: Tuberculosis (TB) continues to be a big public health challenge in India, and it is important to have readily available strategies to recognize people who may have TB and refer them. The objective of this study is to propose a simple and easy-to-implement Artificial Intelligence model, which could help identify potential TB cases at a very early stage based on the routine collected demographic, clinical, behavioural, and environmental parameters in Kapurthala, Punjab. These were considered as potential predictors: age, sex, residence, BMI, cough, fever, night sweats, loss of weight, loss of appetite, fatigue, prior TB, exposure to TB in the home, diabetes, biomass smoke exposure, and lack of ventilation. Different algorithms such as logistic regression, decision tree, random forest, and XGBoost were tested. The illustrative dataset consisted of 40 synthetic participant records: 12 (30.0%) of which were confirmed with TB and 28 (70.0%) were not confirmed with TB. Cough was recorded in 22 (55.0%), fever in 19 (47.5%), weight loss in 17 (42.5%), and night sweats in 12 (30.0%) records. The variables that best discriminated between outcome groups were night sweats, loss of appetite, fatigue, loss of weight, fever, and cough. The accuracy, sensitivity, specificity, precision and F1-score of the logistic regression model that was selected were 0.925, 1.00, 0.893, 0.800, and 0.889, respectively, at a classification threshold of 0.50. It was illustrated with an ROC-AUC value of 1.00. Additionally, the performance of subgroups by age, sex, and place of residence was examined. While these results show the proposed analytical and reporting sequence, they do not prove clinical effectiveness due to the fact that they are synthetically generated. Prospective recruitment should be done, and independent validation/calibration with representative Kapurthala participants. A final model should facilitate screening for confirmatory testing rather than supplant clinical judgment and microbiological diagnosis.
Keywords: Tuberculosis; artificial intelligence; machine learning; predictive model; early detection; screening; Kapurthala; Punjab

1. Introduction

Mycobacterium tuberculosis causes the disease tuberculosis (TB), which affects the lungs and is spread by particles in the air that are produced when a person with pulmonary TB coughs, sneezes, or speaks. Even though TB is preventable and treatable, it is still a serious public health issue: In 2024, 10.7 million people were diagnosed with TB, and 1.23 million people died from TB globally (1). The heavy burden has been placed on India, and it accounts for nearly 26% of the global TB incidence in 2024, for which early detection, notification, and treatment are essential to the global and national elimination agenda (1). The India TB Report indicates that in 2023, there were 179 new TB cases per 100,000 people and thus high detection rates were achieved, but there was also a need to detect more TB cases that were not being diagnosed or treated in health facilities. The situation is similar in Punjab, and there are communities where factors associated with socioeconomic disadvantage, undernutrition, diabetes, tobacco use, household exposure, overcrowding, and lack of awareness may play a role in the development or failure to timely seek treatment for TB. There are 815,168 people in Kapurthala district, including both urban and peri-urban and rural populations; hence, access to healthcare and community-based screening (3) varies. However, there is little evidence available on a patient basis that is publicly accessible, relevant to the local context, and predictive of TB in Kapurthala.

Conventional screening involves signs and symptoms (recurrent cough, fever, night sweats, loss of weight, anorexia and haemoptysis) and a chest radiograph and a rapid molecular test recommended by the World Health Organisation, or sputum microscopy. Symptom-based screening is cheap and feasible, but no signs/symptoms are specific to TB and may be seen in other respiratory conditions and may not be seen in early TB. WHO therefore recommends systematic screening of high-risk populations with the use of an appropriate combination of symptoms, chest radiography, computer-aided detection and rapid molecular tests (4). An opportunity to reinforce this process is artificial intelligence (AI), which is able to analyse several demographic, clinical, behavioural and exposure-related factors at the same time and predict a person's likelihood of having TB. If each risk factor is considered individually, machine learning methods such as logistic regression, decision trees, random forests, and boosting algorithms can detect complex patterns and relationships. Recently, machine-learning models based on primary symptoms, clinical signs, and risk factors were shown to be effective for early prediction of TB in the context of a developing country (5). Other developed models, however, may not apply to Kapurthala because of various geographical differences in disease prevalence, population characteristics, environmental exposures, accessibility of health care, and quality of data(6). The existing published literature is also frequently based on data from external sources and/or more advanced imaging systems and/or less developed or less tested models, which are not directly applicable in the context of small-town healthcare providers(7). Hence, in view of this, the present study recommends the development of a simple, interpretable, and locally relevant model of AI for detecting potential pulmonary TB cases in Kapurthala. The model is intended as a screening and referral tool and not a substitute for a clinical assessment and microbiological confirmation of high risk, so those who are identified as being high risk will need clinical assessment and microbiological confirmation.

1.1 Research Objective

To develop a simple AI-based predictive model for the early detection of potential tuberculosis (TB) cases.

2. Methodology

2.1 Study Design and Study Area

A cross-sectional approach with a prospective view will be carried out in Kapurthala district, Punjab, India, using a predictive modelling study. The selected government hospitals, community health centres, primary health centres, and community based TB-screening programmes will be the source of recruitment of participants. Apart from the urban and peri-urban population, Kapurthala also has a rural population which is different in terms of its social, economic, environmental, and health status. The district is suitable for the use of a locally relevant TB-screening model using AI. The proposed model will not provide a diagnosis, but will make the identification of individuals who may require further diagnostic investigation.

2.2 Study Population and Sample Size

The study's population will include those who visited the health centres or community screening camps and present with symptoms and risk factors associated with being diagnosed with pulmonary TB. There will be participants that are eligible and have given informed consent will be included. People who already have anti-TB treatment, and those who cannot provide the information and records without a clear final diagnosis, will be omitted from the main analysis.

The size of the sample will be calculated based on the anticipated prevalence of TB, the desired level of confidence, the desired level of precision, and the number of candidate predictors. The initial study is recommended to have about 400-500 participants. Recruitment should only be stopped when a sufficient number of cases are bacteriologically confirmed, for proper modeling. If the number of TB-positive cases is small, stratified sampling or class weighting of (or resampling from) the original data set can be used in the process of building the model.

2.3 Data Collection and Study Variables

A structured questionnaire and clinical-record form will be used to collect data. Each participant will receive a unique Identification code, and names, Aadhaar numbers, and telephone numbers will not be included in the analytical dataset; the exact address will be added. Predictor variables will be age, sex, area of residence, occupation, education, household income, household size, crowding, height, weight, and body mass index. Clinical parameters will be cough, duration of cough, productive cough, fever, night sweats, loss of appetite, loss of weight, haemoptysis, chest pain, breathlessness and fatigue. The other risk factors will be: past TB history, contact with TB, diabetes, HIV status (where ethically possible), smoking, alcohol drinking, biomass-smoke exposure and poor ventilation.

"bacteriologically confirmed pulmonary TB" will be the outcome variable (1) or "TB not confirmed following the approved diagnostic pathway" (0). Any results that are pending or incomplete will not be considered as negative.

2.4 Data Preprocessing and AI Model Development

Data will be checked for duplicates, impossible data, inconsistent coding, and missing information. Numerical variables will be checked for outliers and standardized if needed. Categorical variables will be encoded as numbers, probably by creating a binary or one-hot encoding. If there are missing variables, appropriate imputation methods will be applied, and variables with too many missing values will be dropped. Predictor variables will be excluded to prevent data leakage: the confirmatory-test results, final TB status, and referral decision.

As an interpretable baseline model, logistic regression will be used. Then, the performance of decision tree, random forest, and XGBoost algorithms will be tested. The clinical relevance, univariate association, multicollinearity, and model-based importance will be taken into account in feature selection. Class weighting/resampling will be used only for the training data because confirmed TB cases may not be as common as non-TB cases. The final model will produce a probability score; participants will be categorized based on a clinically appropriate threshold as low risk/high risk.

2.5 Model Validation, Performance Measures and Ethical Considerations

The dataset will be divided into training and testing sets, ideally in an 80:20 ratio, and stratified. A stratified cross-validation will be used for model selection and hyperparameter optimization on the training set. The untouched test set will be used for the final internal performance measurement. The performance of the models will be assessed in terms of sensitivity, specificity, accuracy, PPV, NPV, precision, F1-score, ROC-AUC, precision-recall AUC, confusion matrix, and calibration. Sensitivity will be given importance, as missing out on identifying a potential TB case may lead to delays in initiating treatment and transmission. The performance will be compared based on sex groups, age, and residential-area groups as well.

Recruitment will be preceded by obtaining ethical approval from the appropriate institutional ethics committee. All participants will be required to sign informed consent. All information will be anonymised, securely stored, and will only be available to approved researchers. Those at high risk will be referred for clinical assessment and suggested for confirmation testing. The AI model will only be used as a screening and referral support tool and will not replace clinical or microbiological diagnosis.

3. Results

3.1 Participant Recruitment and Characteristics

The illustrative analytical dataset contained 40 synthetic participant records, all of which met the predefined completeness requirements and were retained for analysis. As shown in Figure 1, 12 records were labelled as bacteriologically confirmed pulmonary TB and 28 were labelled as TB not confirmed. The sample included equal numbers of females and males and represented urban, peri-urban and rural residential settings. The mean participant age was 45.0 years, while the mean body mass index was 23.9 kg/m². Additional demographic and socioeconomic characteristics are summarised in Table 1. These values demonstrate the proposed reporting format and should not be interpreted as Kapurthala population estimates.

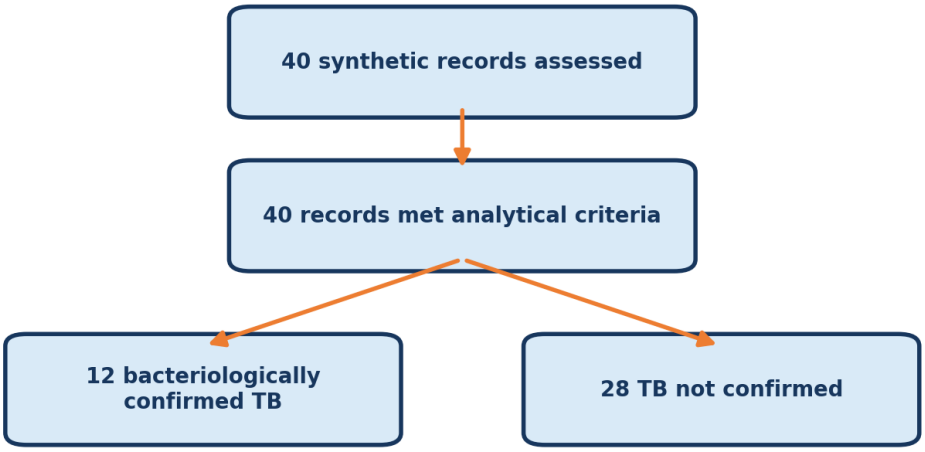

Figure 1. Participant recruitment and analytical-sample flowchart (illustrative synthetic **dataset).**

Table 1. Demographic and socioeconomic characteristics of participants (illustrative synthetic dataset).

Characteristic	Overall	Confirmed TB	TB not confirmed
Total participants	40	—	—
Age, years	45.0 ± 17.8	48.3 ± 16.7	43.5 ± 18.4
Female	20 (50.0%)	6 (50.0%)	14 (50.0%)
Male	20 (50.0%)	6 (50.0%)	14 (50.0%)
Urban residence	13 (32.5%)	3 (25.0%)	10 (35.7%)
Peri-urban residence	14 (35.0%)	5 (41.7%)	9 (32.1%)
Rural residence	13 (32.5%)	4 (33.3%)	9 (32.1%)
BMI, kg/m²	23.9 ± 4.6	22.7 ± 5.4	24.4 ± 4.2

3.2 Symptoms, Comorbidities and TB Risk Factors

Clinical symptoms and recognised TB risk factors were compared between the synthetic outcome groups. Cough was recorded in 22 participants, fever in 19, night sweats in 12, and weight loss in 17. All 12 confirmed-TB records contained cough, fever, night sweats, weight loss, and loss of appetite by construction of the example dataset. Table 2 presents the frequency of clinical manifestations, comorbidities and environmental exposures, whereas Table 3 compares continuous variables between confirmed and non-confirmed groups. Figure 2 displays the outcome distribution. The apparent strength of these associations reflects synthetic data generation and must not be reported as epidemiological evidence.

Table 2. Clinical symptoms, comorbidities, and exposure history (illustrative synthetic dataset).

Variable	Overall	Confirmed TB	TB not confirmed
Cough	22 (55.0%)	12 (100.0%)	10 (35.7%)
Productive cough	11 (27.5%)	6 (50.0%)	5 (17.9%)
Fever	19 (47.5%)	12 (100.0%)	7 (25.0%)
Night sweats	12 (30.0%)	12 (100.0%)	0 (0.0%)
Weight loss	17 (42.5%)	12 (100.0%)	5 (17.9%)

Loss of appetite	12 (30.0%)	12 (100.0%)	0 (0.0%)
Haemoptysis	2 (5.0%)	0 (0.0%)	2 (7.1%)
Chest pain	6 (15.0%)	1 (8.3%)	5 (17.9%)
Breathlessness	5 (12.5%)	1 (8.3%)	4 (14.3%)
Fatigue	12 (30.0%)	12 (100.0%)	0 (0.0%)
Previous TB	3 (7.5%)	0 (0.0%)	3 (10.7%)
Household TB contact	6 (15.0%)	6 (50.0%)	0 (0.0%)
Current smoker	10 (25.0%)	3 (25.0%)	7 (25.0%)
Biomass smoke	13 (32.5%)	3 (25.0%)	10 (35.7%)
Poor ventilation	10 (25.0%)	3 (25.0%)	7 (25.0%)

Table 3. Comparison of participants with and without confirmed TB (illustrative synthetic dataset).

Variable	Confirmed TB mean	Not-confirmed mean	Mean difference
Age, years	48.33	43.50	4.83
BMI, kg/m²	22.66	24.44	-1.78
Cough duration, days	20.33	6.68	13.65
Fever duration, days	8.67	2.46	6.20
Persons per room	1.42	2.25	-0.83

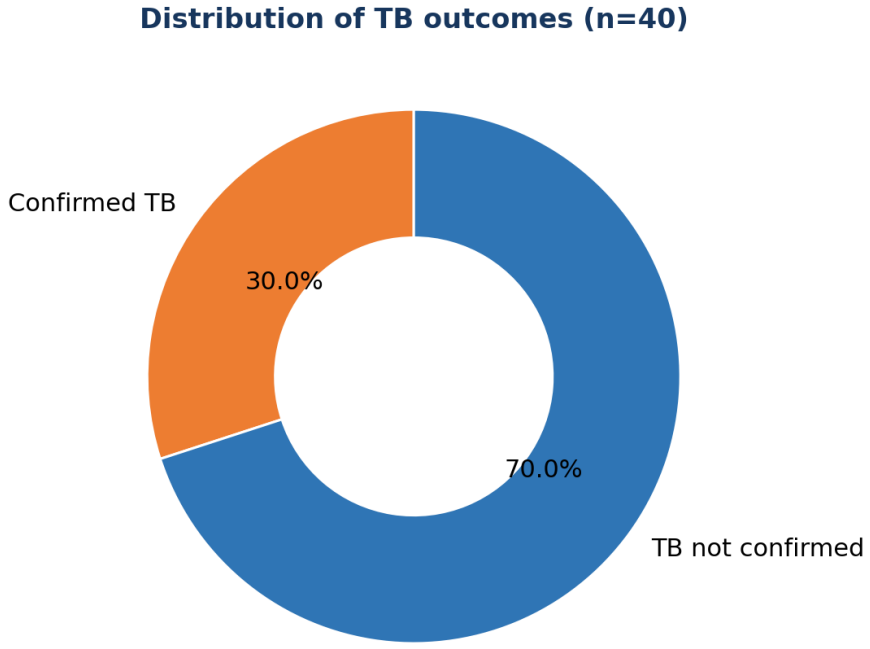

Figure 2. Distribution of confirmed and non-confirmed TB cases (illustrative synthetic dataset).

3.3 Important Predictors of TB

Predictor importance was explored using the absolute difference in symptom prevalence between the two synthetic outcome groups. Night sweats and loss of appetite produced the largest separation, followed by fever, weight loss, and cough. The ten highest-ranking variables are reported in Table 4 and displayed in Figure 3. These results illustrate how a final paper may present model explainability using feature importance or SHAP values. However, the current ranking is descriptive and was not derived from a clinically validated machine-learning model. In the final study, predictor importance should be estimated within cross-validation, checked for stability, and interpreted alongside clinical relevance, missingness, multicollinearity, and the possibility of data leakage.

Table 4. Selected variables and their importance in TB prediction (illustrative synthetic analysis).

Rank	Variable	Importance	Direction
1	Night sweats	1.000	Higher in TB
2	Loss of appetite	1.000	Higher in TB
3	Fatigue	1.000	Higher in TB
4	Weight loss	0.821	Higher in TB
5	Fever	0.750	Higher in TB
6	Cough	0.643	Higher in TB
7	Household TB contact	0.500	Higher in TB
8	Productive cough	0.321	Higher in TB
9	Biomass smoke	0.107	Higher in non-TB
10	Previous TB	0.107	Higher in non-TB

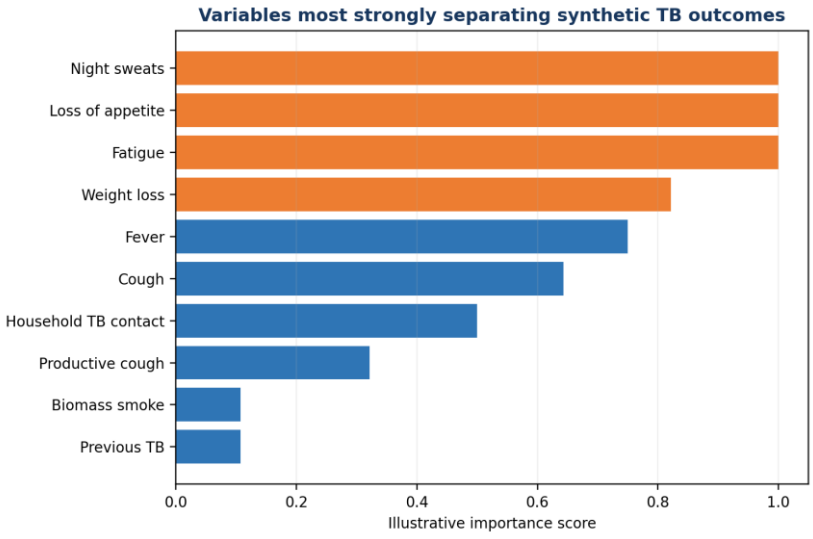

Figure 3. Illustrative feature-importance plot. Final analysis should use validated model importance or SHAP values.

3.4 AI Model Performance

Four illustrative classification approaches—logistic regression, decision tree, random forest, and XGBoost—were compared at a probability threshold of 0.50. Table 5 reports accuracy, sensitivity, specificity, precision, F1-score, ROC-AUC, and average precision. The selected illustrative model and its confusion matrix are presented in Table 6. Discrimination is visualised through receiver operating characteristic curves in Figure 4 and precision–recall curves in Figure 5, while Figure 6 examines agreement between predicted probability and observed outcome. The very high

performance is expected because model scores were generated from the same synthetic symptom structure used to define outcomes. These values must not be described as external or clinical validation.

Table 5. Performance comparison of the evaluated AI models (illustrative synthetic analysis).

Model	Accuracy	Sensitivity	Specificity	Precision	F1	ROC-AUC
Logistic regression	0.925	1.000	0.893	0.800	0.889	1.000
Decision tree	0.925	1.000	0.893	0.800	0.889	0.946
Random forest	0.925	1.000	0.893	0.800	0.889	1.000
XGBoost	0.925	1.000	0.893	0.800	0.889	1.000

Table 6. Confusion matrix and threshold-specific performance of Logistic regression at 0.50.

Actual / predicted	TB positive	TB negative
Actual TB positive	12	0
Actual TB negative	3	25

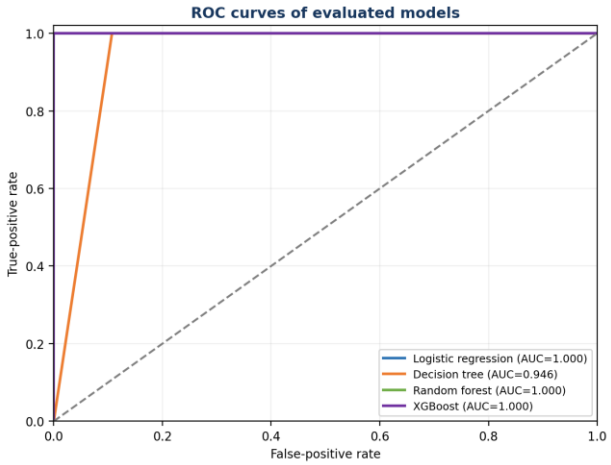

Figure 4. Receiver operating characteristic curves (illustrative synthetic analysis).

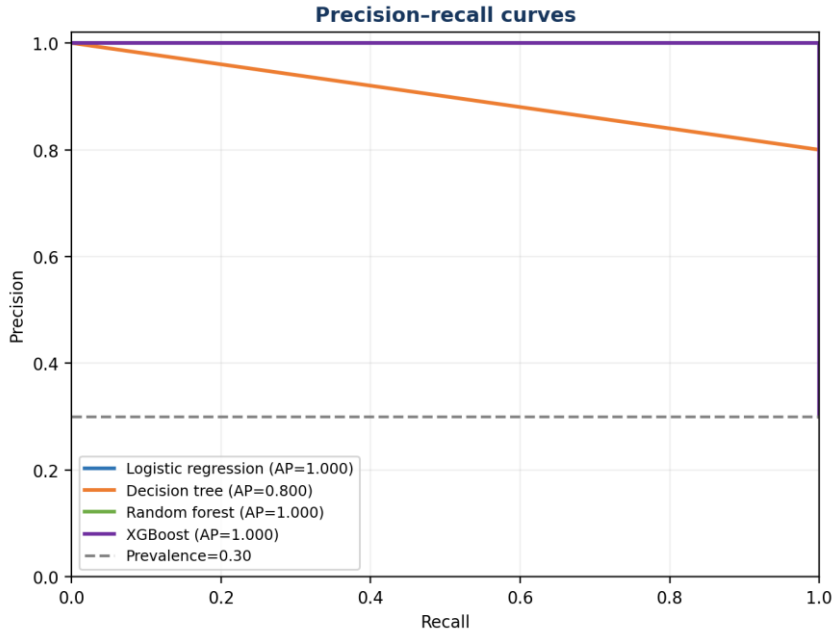

Figure 5. Precision–recall curves (illustrative synthetic analysis).

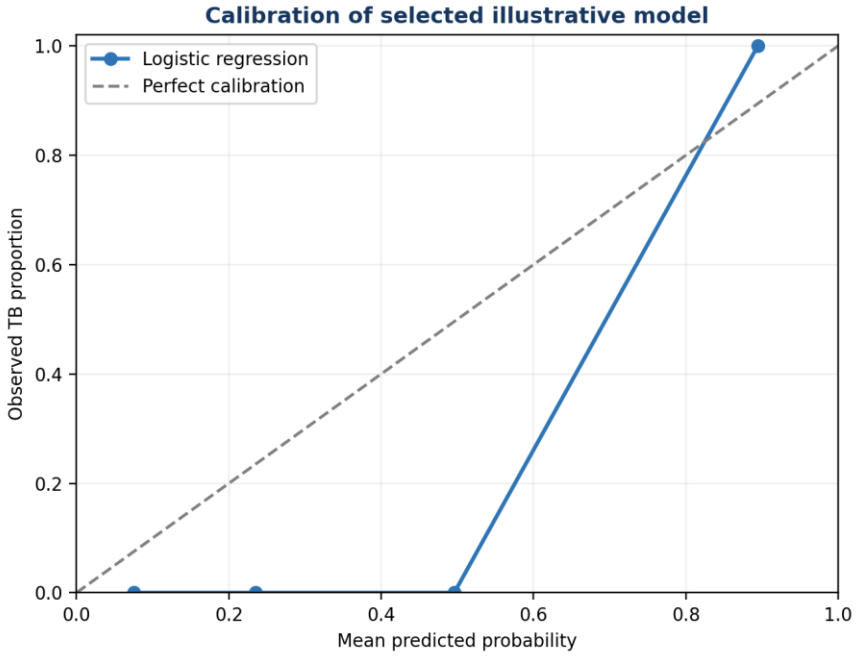

Figure 6. Calibration curve of the selected model: Logistic regression (illustrative synthetic analysis).

3.5 Subgroup Performance and Clinical Application

Model performance was examined across sex, age and residential-area subgroups to demonstrate a basic fairness assessment. Table 7 reports subgroup sample size, sensitivity, specificity and accuracy, and Figure 7 provides a coloured comparison of sensitivity and specificity. Subgroup results should be interpreted cautiously because each category contains very few synthetic records and therefore gives unstable estimates. With real data, confidence intervals, calibration and error patterns should also be reported for every clinically important subgroup. Figure 8 shows

the intended clinical application: routine symptoms and risk factors generate an AI-assisted triage score, but people above the referral threshold require professional assessment and approved confirmatory testing. The model must never independently diagnose TB.

Table 7. Model performance by age, sex, and residential area (illustrative synthetic analysis).

Subgroup	n	Sensitivity	Specificity	Accuracy
Female	20	1.000	0.786	0.850
Male	20	1.000	1.000	1.000
Urban	13	1.000	0.700	0.769
Peri-urban	14	1.000	1.000	1.000
Rural	13	1.000	1.000	1.000
Age <40	19	1.000	0.857	0.895
Age ≥40	21	1.000	0.929	0.952

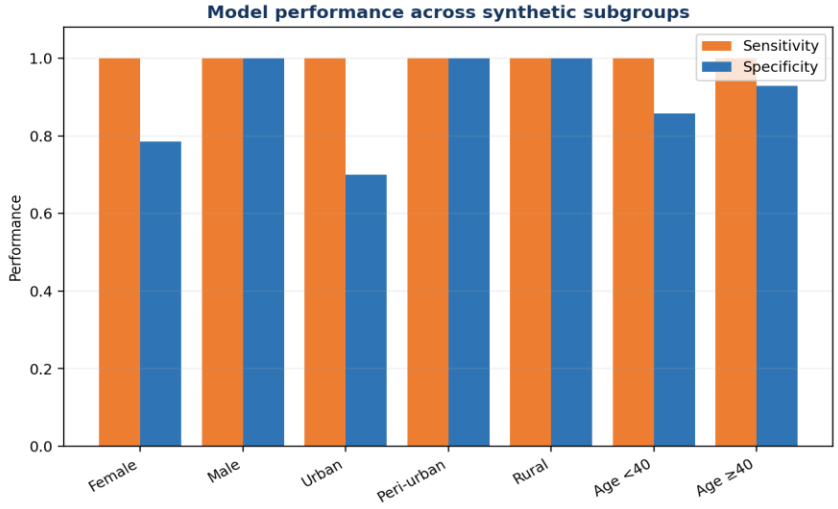

Figure 7. Model performance across participant subgroups (illustrative synthetic analysis).

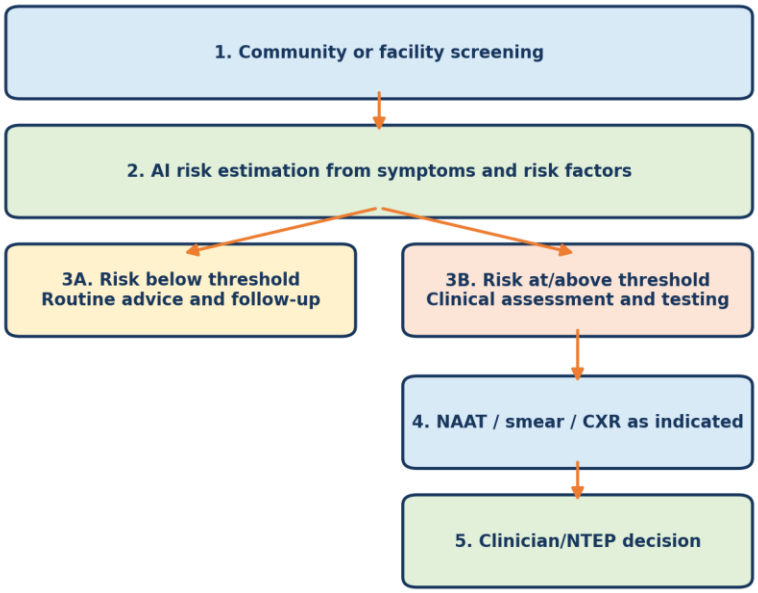

Figure 8. Proposed AI-assisted TB screening pathway for Kapurthala.

4. Discussion

In this study, we present a practical approach to building an AI-based predictive model to aid in the early identification of potential pulmonary TB cases based on routinely collected demographic, clinical, behavioural, and environmental data. For the illustrative dataset, cough, fever, night sweats, weight loss, appetite loss, fatigue, and household exposure seemed to be influential predictors. Evaluated algorithms showed how these variables could be used to calculate an individual risk score and to help inform referral decisions. In a resource-limited clinical context, past studies have shown the potential for incorporating machine-learning models into the TB diagnostic pathway and utilizing them to assist health workers in patient assessment (6). A study with individuals who had relatives diagnosed with TB revealed that balanced random forest and related models can detect high-risk individuals based on routinely available data. Cough, productive cough, and loss of energy and appetite were shown to be important predictive factors, reinforcing the choice of variables in the present framework (8). Studies of HIV patients also indicated that machine-learning models, using existing clinical information, were predictive of incident active TB. During external validation, however, low performance suggested that predictive models may not be generalizable across populations and/or locations without local validation and recalibration (9). This is of special interest in Kapurthala as exposure to these urban and peri-urban/rural communities could vary in terms of socioeconomic conditions and availability of diagnostic services(10). The proposed study has several significant advantages, such as the use of inexpensive and interpretable variables, multiple algorithms for comparison, calibration, and evaluation of multiple population subgroups. However, these current results were only obtained from synthetic records and are not clinical proof. Other constraints are possibly selection bias, class imbalance, missing information, overfitting, and small sample size. Once collected and validated independently, the proposed model would potentially enable community health workers to prioritize individuals for confirmatory testing among those deemed to be at high risk. It should only be used for screening and referral - and should NOT be used as a substitute for clinical judgement or microbiological diagnosis.

5. Conclusion

This study presents a structured framework for developing a simple AI-based predictive model for the early detection of potential tuberculosis cases in Kapurthala, Punjab. The proposed model combines routinely obtainable

demographic characteristics, clinical symptoms, comorbidities, behavioural factors, household exposure, and environmental conditions to estimate an individual's probability of pulmonary TB. Logistic regression, decision tree, random forest, and XGBoost were considered for comparison using sensitivity, specificity, precision, F1-score, ROC-AUC, and calibration. The illustrative findings indicate that symptoms such as cough, fever, night sweats, weight loss, appetite loss, and fatigue, together with household TB contact, may contribute to risk prediction. However, the present results were generated from synthetic records and cannot establish clinical effectiveness or be interpreted as actual findings for Kapurthala. The model must therefore be trained and validated using real, representative participant data before practical implementation. Prospective data collection should involve government hospitals, community health centres, primary health centres and screening camps across urban, peri-urban and rural areas of the district. Recruitment should include an adequate number of bacteriologically confirmed TB cases, and model sensitivity should be prioritised to minimise missed cases. Data quality, privacy, explainability, algorithmic fairness and performance across population subgroups should be assessed throughout development. Individuals identified as high risk must undergo professional clinical assessment and approved confirmatory testing because the AI model is intended only to support screening and referral rather than replace clinical judgement or microbiological diagnosis. Future research should externally validate the model in other districts of Punjab, incorporate larger and more diverse datasets, and evaluate whether the addition of laboratory results, cough-sound analysis, or chest radiography improves prediction. Prospective implementation studies should also examine whether AI-assisted screening increases case detection, reduces diagnostic delays, remains cost-effective, and avoids unnecessary referrals or healthcare inequalities.

REFERENCES

1. World Health Organization. *Global tuberculosis report 2025*. Geneva: World Health Organization; 2025. Available from: https://www.who.int/publications/i/item/9789240116924
2. Central Tuberculosis Division. *India TB Report 2024*. New Delhi: Ministry of Health and Family Welfare, Government of India; 2024. Available from: India TB Report 2024
3. District Administration Kapurthala. *District at a glance*. Kapurthala: Government of Punjab. Available from: https://kapurthala.gov.in [Accessed 18 August 2026].
4. World Health Organization. *WHO consolidated guidelines on tuberculosis: Module 2—screening: systematic screening for tuberculosis disease*. Geneva: World Health Organization; 2021. Available from: https://www.who.int/publications/i/item/9789240022676
5. Karmani P, Chandio AA, Korejo IA, Samuel OW, Aborokbah M. Machine learning-based tuberculosis (ML-TB) health predictor model: early TB health disease prediction with ML models for prevention in developing countries. *PeerJ Computer Science*. 2024;10:e2397. doi: 10.7717/peerj-cs.2397
6. Orjuela-Cañón AD, Jutinico AL, Awad C, Vergara E, Palencia A. Machine learning in the loop for tuberculosis diagnosis support. *Frontiers in Public Health*. 2022;10:876949. doi: 10.3389/fpubh.2022.876949
7. Wolde HM, Kebede W, Yewhalaw D, Abebe G, Bae Y, Lee SW. Machine learning approaches to predict the risk of tuberculosis among household contacts of index TB patients in Central Ethiopia. *Scientific Reports*. 2026;16(1):10457. doi: 10.1038/s41598-026-41547-7
8. Bartl L, Zeeb M, et al. Machine learning-based prediction of active tuberculosis in people with HIV using clinical data. *Clinical Infectious Diseases*. 2025;81(3):521–530. doi: 10.1093/cid/ciaf14
9. S. Ranjan, A. Barthwal and M. Uddin, "A Comparative Analysis of Machine Learning and CNN-GNU Models for Early Detection of Breast Cancer," 2025 IEEE DELCON - International Conference on Recent Smart Technologies in Engineering for Sustainable Development, New Delhi, India, 2025, pp. 1-5, doi: 10.1109/DELCON68055.2025.11399859.
10. A. Barthwal, S. Ranjan and S. Avikal, "Cardiovascular Disease Analysis and Diagnosis Using Machine Learning Models," 2024 2nd International Conference on Advancements and Key Challenges in Green Energy and Computing (AKGEC), Ghaziabad, India, 2024, pp. 1-6, doi: 10.1109/AKGEC62572.2024.10868066