Infection and Drug Resistance

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Analysis of High-Risk Factors for Tuberculosis Retreatment Based on Machine Learning and Latent Class Analysis

Xilong Du[1], Maiwulajiang Yimamu[2], Yan Na[3], Xiaoxue Li[3], Ziyu Wang[3], Zulimire Z Nuermaihaimaiti[3], Yuxin Wang[3], Liping Zhang[3], Yanling Zheng[3,4]

[1]School of Public Health, Xinjiang Medical University, Urumqi, Xinjiang, People's Republic of China; [2]Tuberculosis and Leprosy Prevention and Control Department, kashgar Prefecture Center for Disease Control and Prevention, Kashgar, Xinjiang, People's Republic of China; [3]College of Medical Engineering and Technology, Xinjiang Medical University, Urumqi, Xinjiang, People's Republic of China; [4]Institute of Medical Engineering Interdisciplinary Research, Xinjiang Medical University, Urumqi, Xinjiang, People's Republic of China

Correspondence: Yanling Zheng, Email zhengyl_math@sina.cn

Object: To identify high-risk factors for tuberculosis retreatment and to provide a scientific basis for developing targeted prevention and control strategies by integrating machine learning with latent class analysis.

Methods: This study retrospectively collected baseline and treatment-related data from 6,821 tuberculosis patients, employing machine learning and latent class analysis (LCA) to investigate the key influencing factors associated with high-risk populations for retreatment.

Results: The XGBoost model achieved an overall accuracy of 84% and an area under the ROC curve (AUC) of 0.938. The analysis identified sputum examination results at month 6 or 8 of treatment, treatment regimen, and diagnostic classification as the most influential factors associated with retreatment. SHAP analysis further revealed that a sputum examination status of "not performed" was strongly linked to increased retreatment risk. Logistic regression confirmed this finding, with "not performed" (OR = 123.47, $P < 0.001$) and a "positive" result (OR = 14.89, P = 0.02) at month 6 or 8 identified as significant risk factors. Latent class analysis stratified patients into four distinct subgroups, among which those characterized by comorbid diabetes or prior treatment failure constituted the highest-risk populations for retreatment.

Conclusion: It is recommended to improve treatment adherence and efficacy monitoring for newly diagnosed patients, strengthen whole-course supervision, and optimize management for elderly patients and those on long-term regimens.

Keywords: tuberculosis, latent class analysis, random forest, xgboost, cramér's v

Introduction

Tuberculosis remains a globally widespread infectious disease with high transmissibility and fatal risks, posing an ongoing threat to human health.[1,2] The World Health Organization's latest *Global Tuberculosis Report 2025*[3] indicates that in 2024, there were an estimated 10.7 million new tuberculosis cases worldwide, with an estimated incidence rate of 131 per 100,000 population. The global tuberculosis incidence rate declined by nearly 2% between 2023 and 2024. In 2024, an estimated 390,000 people developed multidrug-resistant or rifampicin-resistant tuberculosis, accounting for 3.6% of all tuberculosis cases. While the global estimated number of such cases has been declining since 2015–2024, some countries and regions still report localized increases in drug-resistant tuberculosis cases. Regarding drug resistance risk, the 2024 report clearly states that the global rate of multidrug-resistant/rifampicin-resistant tuberculosis among previously treated patients was 16%, compared to only 3.2% among newly treated patients. This highlights that the drug resistance risk in retreatment cases is significantly higher than in newly treated cases.

Tuberculosis retreatment patients refer to individuals with a history of previous tuberculosis treatment who require anti-tuberculosis therapy again after treatment failure or relapse. Tuberculosis recurrence may result from exogenous

https://doi.org/10.2147/IDR.S594300
Received: 28 January 2026
Accepted: 16 April 2026
Published: 25 April 2026

reinfection or endogenous reactivation of the initial infection.[4,5] Currently, the prevention and control of tuberculosis retreatment is one of the key challenges facing China's tuberculosis control system.[6] Neglecting risk factors related to "retreatment" in clinical diagnosis, treatment, and prevention efforts can easily lead to treatment failure or disease recurrence.[7] Therefore, an in-depth exploration of various factors influencing the treatment outcomes of tuberculosis retreatment patients can not only provide a basis for developing personalized treatment plans for different patients,[8] enhance patient treatment confidence, and facilitate recovery, but also offer practical support for the refined optimization of tuberculosis treatment strategies in high-burden regions of western China.[9] This holds significant importance both academically and practically.

In the fields of machine learning and statistical analysis, the synergistic application of multiple methods provides robust support for research on tuberculosis clinical characteristics. Random Forest (RF), as a widely used algorithm in bioinformatics and related fields,[10] serves as an efficient feature selection tool. It can output feature importance scores, exclude irrelevant variables, and capture nonlinear relationships and variable interactions in data, while also offering strong predictive power and intuitive interpretability. XGBoost (eXtreme Gradient Boosting), as a high-performance ensemble learning algorithm, iteratively optimizes loss functions through gradient descent and supports various complex algorithms for precise error fitting,[11] making it suitable for scenarios requiring high prediction accuracy. It has previously been applied in predicting drug resistance of tuberculosis strains.[12] Latent Class Analysis (LCA), on the other hand, can identify potential patient subgroups from complex clinical data through probabilistic models, revealing heterogeneity among different subgroups in terms of clinical characteristics and treatment responses.[13] This effectively addresses the limitations of overall data analysis caused by significant individual differences among tuberculosis patients and provides targeted evidence for the formulation of precision intervention strategies. Additionally, the logistic regression model, as a classical statistical modeling method, offers efficient analysis of associations between binary or multi-class outcomes and clinical characteristics based on its straightforward mathematical logic and clear interpretability. By calculating odds ratios (OR values), it quantifies the impact of features on outcomes such as tuberculosis risk and treatment prognosis. With its low requirements for data distribution assumptions and computational cost, it often serves as a baseline model complementary to machine learning algorithms, playing an irreplaceable foundational role in studies on tuberculosis risk factor screening and prognosis prediction.[14–16]

Based on large-sample baseline and treatment-related data of tuberculosis patients, this study first employs the Random Forest algorithm to screen key features of research value. It then uses Cramér's V coefficient to measure the strength of associations among categorical variables, eliminating highly correlated variables to enhance the performance and interpretability of subsequent models. After completing variable selection, this study applies the XGBoost model to conduct an in-depth analysis of the selected variables. Simultaneously, SHapley Additive exPlanations(SHAP) is incorporated to obtain more precise data on feature contributions, exploring the impact of different factors on the risk of tuberculosis retreatment. Furthermore, LCA is introduced to reveal the heterogeneity of potential subgroups within the patient population. Ultimately, this study aims to clarify the specific effects of different factor categories on tuberculosis retreatment patients, providing theoretical reference for the precision prevention and control of tuberculosis.

While each of these methods can be individually applied in tuberculosis research, their synergistic integration—leveraging Random Forest for efficient feature screening, XGBoost for robust predictive modeling with SHAP for interpretability, and LCA for uncovering hidden patient heterogeneity—offers a more comprehensive analytical framework for identifying high-risk populations and informing precision prevention strategies.

Materials and Methods

Data Source

A retrospective analysis was conducted on data from 6,821 tuberculosis patients in the Kashgar region of China from January 1, 2022 to December 31, 2022. After screening, data from 5,826 tuberculosis patients were included, comprising 4,430 patients undergoing initial treatment and 1,396 patients undergoing retreatment. The dataset encompassed baseline information of the tuberculosis patients (including name, gender, age, patient source, history of previous anti-tuberculosis treatment, diagnostic classification, diagnostic results, comorbidities, sputum smear examination at month 0, sputum

culture results, imaging findings, molecular biology results, etiological results, drug susceptibility testing results, and strain identification). Treatment-related data were also collected (including treatment outcomes, sputum examination, actual medication management method, treatment regimen, sputum smear examination at month 2, sputum smear examination at month 5, treatment protocol, and sputum smear examination at month 6 or 8).

Statistical Analysis Method

Based on the baseline and treatment-related data of tuberculosis patients, this study first calculated the feature importance scores using Random Forest. These scores represent the contribution level of each feature to the model. Features with relatively low contribution levels were excluded. The Cramér's V coefficient was then used to measure the strength of association between two categorical variables, and among highly correlated features, those with lower importance scores were excluded. The XGBoost model was employed to analyze and calculate the impact of different factors on the treatment outcomes of tuberculosis retreatment. The performance of the XGBoost model was evaluated using the confusion matrix, Receiver Operating Characteristic Curve (ROC Curve), and Precision-Recall Curve (PR Curve). The feature contributions from the XGBoost model and the more precise feature contributions provided by SHAP were used to analyze the influence of various factors on tuberculosis treatment outcomes. Additionally, the logistic regression model offered excellent interpretability for each feature. Latent Class Analysis (LCA) revealed hidden heterogeneity within the patient population and provided well-defined target groups for precision intervention. The specific analytical workflow is illustrated in Figure 1.

Statistical analysis was conducted using R version 4.4.1 and Python version 3.10. Random Forest and Cramér's V were implemented in R 4.4.1, while XGBoost and SHAP were implemented in Python.

For XGBoost modeling, the xgboost package (version 1.7.5) in Python 3.10 was used. Hyperparameters were optimized via grid search with three-fold stratified cross-validation, searching over n_estimators (50, 100), max_depth (3, 6), and learning_rate (0.01, 0.1). The final optimal parameters were n_estimators = 100, max_depth = 6, and learning_rate = 0.1. Random Forest was implemented using the randomForest package (version 4.7–1) in R 4.4.1 with ntree = 500 and mtry = sqrt(p). To address class imbalance between initial treatment (n = 4,430) and retreatment (n

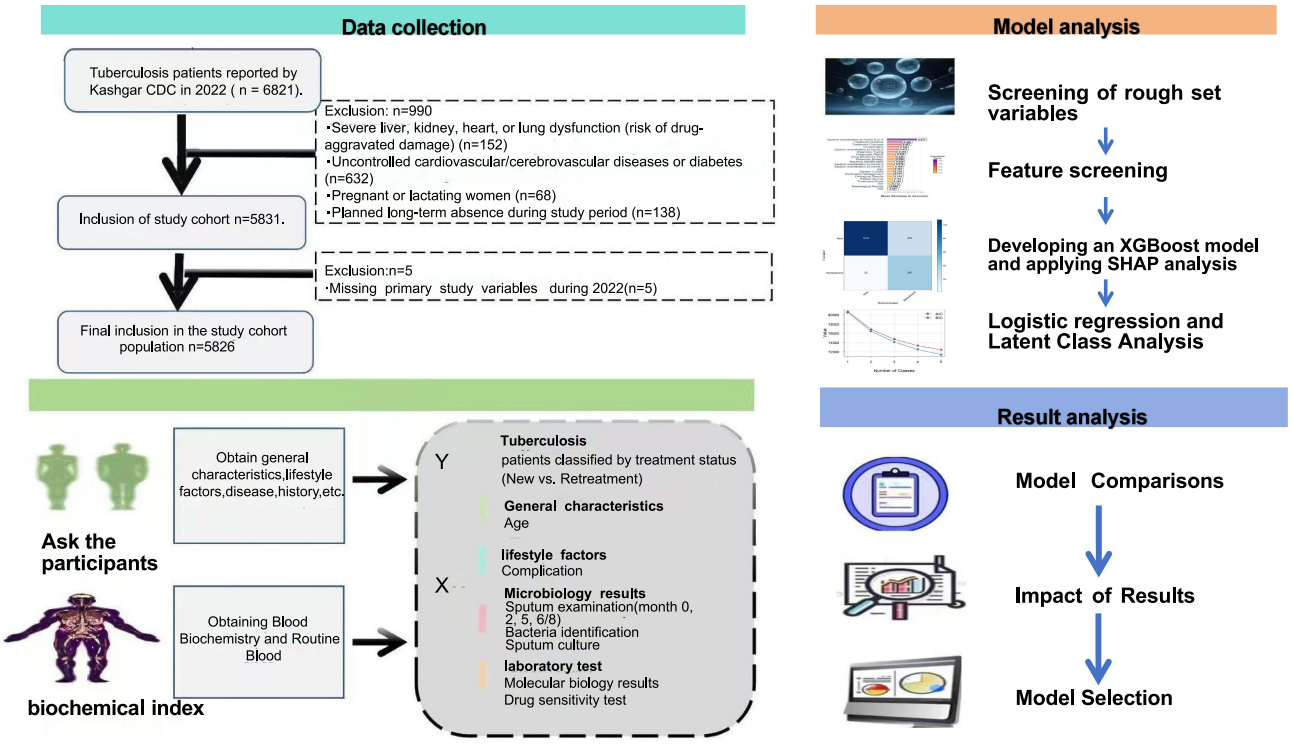

Figure 1 Flowchart of the study. Bold text denotes the three main phases: data collection, model analysis, and result analysis.

https://doi.org/10.2147/IDR.S594300

= 1,396) cases, sample weights were applied using class_weight = "balanced". Model performance was evaluated using precision, recall, F1-score, and the area under the precision-recall curve (AP) in addition to accuracy and ROC-AUC, as these metrics are more robust for imbalanced data.Latent class analysis was performed using the poLCA package (version 1.6.0) in R 4.4.1. Model selection was guided by the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and entropy (ranging from 0 to 1, with higher values indicating better class separation). A maximum of five latent classes was specified, and the final four-class model was selected based on optimal fit indices and interpretability.

Results

Univariate Analysis of Indicators Related to Tuberculosis Retreatment

This study employed the Chi-square test to conduct a univariate analysis of factors associated with retreatment among tuberculosis patients. The results are presented in Table 1. The analysis identified the following as significant risk factors influencing tuberculosis retreatment (all with P-values < 0.05):Sex, Patient Source, HIV, Diagnosis Typing, Diagnostic Result, Complication, Treatment Classification, Medication Management, Treatment Outcome, Treatment Mode, Sputum examination at month 0, Sputum examination at month 2, Sputum examination at month 5, Sputum Culture, Radiological

Table 1 Baseline Characteristics and Clinical Treatment Variables of Tuberculosis Patients, Stratified by Initial Treatment Vs. Retreatment

	Treatment Classification, n (%)			x^2/Fisher's Exact Test	*P*
	New	Retreatment	Total		
Sex				2.9	0.09
Male	2129(48.05%)	708(50.72%)	2837(48.70%)		
Female	2301(51.95%)	688(49.28%)	2989(51.30%)		
Patient Source				6.08	0.2
Health examination	2576(58.14%)	766(54.87%)	3342(57.36%)		
Direct treatment	135(3.05%)	39(2.76%)	174(2.99%)		
Active screening	33(0.74%)	9(0.64%)	42(0.72%)		
Transfer treatment	1053(23.77%)	368(26.36%)	1421(24.39%)		
Tracing	633(14.29%)	214(15.33%)	847(14.54%)		
Diagnosis Typing				10.37	0.02
Secondary	4234(95.58%)	1361(97.49%)	5595(96.04%)		
Tuberculous pleuritis	157(3.54%)	28(2.01%)	185(3.18%)		
Tracheobronchial	19(0.43%)	4(0.29%)	23(0.40%)		
Hematogenous dissemination	20(0.45%)	3(0.21%)	23(0.40%)		
Diagnostic Result				12.83	<0.01
Microbiologically Confirmed	2747(62.01%)	923(66.12%)	3670(62.99%)		
Microbiologically Unconfirmed	1526(34.45%)	445(31.88%)	1971(33.83%)		
Tuberculous pleuritis	157(3.54%)	28(2.00%)	185(3.18%)		
Complication				23.68	<0.01
Unknown	2767(62.46%)	805(57.66%)	3572(61.31%)		
No	1160(26.19%)	423(30.30%)	1583(27.17%)		
Diabetes	380(8.58%)	114(8.17%)	494(8.48%)		
Other	99(2.23%)	52(3.72%)	151(2.59%)		
HIV/AIDS	24(0.54%)	2(0.14%)	26(0.45%)		
Medication Management				10.18	<0.01
Healthcare Workforce Management	3640(82.17%)	1094(78.37%)	4734(81.26%)		
Intelligent tools	769(17.36%)	295(21.13%)	1064(18.26%)		
Unknown	21(0.047%)	7(0.50%)	28(0.48%)		

(*Continued*)

Table 1 (Continued).

	Treatment Classification, n (%)			x^2/Fisher's Exact Test	*P*
	New	**Retreatment**	**Total**		
Treatment Outcome				117.5	<0.01
Other	26(0.59%)	41(2.94%)	67(1.15%)		
Failure	20(0.45%)	12(0.86%)	32(0.55%)		
Death	149(3.36%)	85(6.09%)	234(4.02%)		
Treatment completed	1764(39.41%)	460(32.95%)	2224(38.17%)		
Unknown	179(4.04%)	53(3.80%)	232(3.98%)		
Cure	2233(50.41%)	694(49.71%)	2927(50.24%)		
Switched to MDR-TB regimen	59(1.33%)	51(3.65%)	110(1.89%)		
Treatment Mode				10.32	<0.01
Inpatient Treatment	1264(28.53%)	354(25.36%)	1618(27.77%)		
Unknown	2026(45.74%)	706(50.57%)	2732(46.89%)		
Ambulatory Care	1140(25.73%)	336(24.07%)	1476(25.33%)		
Sputum examination at month 0				1.66	0.4
Not performed	36(0.81%)	13(0.93%)	49(0.84%)		
Positivity	892(20.14%)	260(18.62%)	1152(19.77%)		
Negative	3502(79.05%)	1123(80.45%)	4625(79.39%)		
Sputum examination at month 2				24.97	<0.01
Not performed	183(4.13%)	104(7.45%)	287(4.93%)		
Positivity	33(0.74%)	10(0.72%)	43(0.74%)		
Negative	4214(95.13%)	1282(91.83%)	5496(94.34%)		
Sputum examination at month 5				46.33	<0.01
Not performed	388(8.76%)	209(14.97%)	597(10.25%)		
Positivity	8(0.18%)	5(0.36%)	13(0.22%)		
Negative	4034(91.06%)	1182(84.67%)	5216(89.53%)		
Sputum Culture				4.07	=0.1
Not performed	1052(23.75%)	302(21.63%)	1354(23.24%)		
Positivity	1555(35.10%)	526(37.68%)	2081(35.72%)		
Negative	1823(41.15%)	568(40.69%)	2391(41.04%)		
Radiological Results				12.34	<0.01
No abnormality	2(0.05%)	0(0.00%)	2(0.03%)		
Not performed	2797(63.14%)	952(68.19%)	3749(64.35%)		
Abnormality	1631(36.81%)	444(31.81%)	2075(35.61%)		
Molecular Biology				6.45	0.04
Not performed	68(1.53%)	9(0.64%)	77(1.32%)		
Positivity	2500(56.43%)	795(56.95%)	3295(56.56%)		
Negative	1862(42.03%)	592(42.41%)	2454(42.12%)		
Treatment Scheme				9.62	0.04
2HRZE/10HRE	606(13.68%)	183(13.11%)	789(13.54%)		
2HRZE/4HR	3576(80.72%)	1129(80.87%)	4705(80.76%)		
2HRZE/7-10HRE	199(4.49%)	55(3.94%)	254(4.36%)		
6-9RZELfx	13(0.29%)	5(0.36%)	18(0.31%)		
Other	36(0.82%)	24(1.72%)	60(1.03%)		
Etiological Results				8.98	0.01
Not performed	8(0.18%)	0(%)	8(0.14%)		
Positivity	2764(62.39%)	925(%)	3689(63.32%)		
Negative	1658(37.43%)	471(%)	2129(36.54%)		
Sputum examination at month 6 or 8				2351.3	<0.01
Not performed	923(20.84%)	1300(93.12%)	2223(38.16%)		
Positivity	5(0.11%)	1(0.07%)	6(0.10%)		
Negative	3502(79.05%)	95(6.81%)	3597(61.74%)		

(*Continued*)

https://doi.org/10.2147/IDR.S594300

Table 1 (Continued).

	Treatment Classification, n (%)			x^2/Fisher's Exact Test	*P*
	New	**Retreatment**	**Total**		
Drug Sensitivity Test				49.72	<0.01
Susceptible	2656(59.95%)	859(61.53%)	3515(60.33%)		
Rifampicin Resistance	41(0.93%)	44(3.15%)	85(1.46%)		
Isoniazid resistance	7(0.16%)	6(0.43%)	13(0.22%)		
Multidrug-resistant	12(0.27%)	9(0.64%)	21(0.36%)		
Not performed	1714(38.69%)	478(34.24%)	2192(37.62%)		
Age	58.48(41.08–75.88)	64.48(50.24–78.72)	59.92 (43.03–76.81)	213.12	<0.01
Bacteria Identification				9.53	<0.01
Mycobacterium tuberculosis	2685(60.61%)	910(65.19%)	3595(61.71%)		
Not performed	1712(38.65%)	478(34.24%)	2190(37.59%)		
No mycobacterium tuberculosis	33(0.74%)	8(0.57%)	41(0.70%)		

Results, Molecular Biology, Treatment Scheme, Etiological Results, Sputum examination at month 6 or 8, Drug Sensitivity Test, Age, Bacteria Identification.

Feature Selection with Random Forest

Due to the large number of independent variables, feature screening was conducted based on their importance scores. Features with relatively low mean decrease in accuracy were excluded. Specifically, seven variables—gender, imaging findings, HIV test results, treatment modality, patient origin, etiological findings, and medication management—were removed.

Selecting key features helps reduce noise from irrelevant variables and lowers the computational complexity of the XGBoost model. Moreover, an excessive number of features may lead to overfitting or slower training in XGBoost (see Figure 2). By pre-screening with Random Forest, the most discriminative features were retained, thereby improving the generalization ability of the subsequent model.

Cramér's V

Highly correlated features can cause XGBoost to repeatedly learn similar information, thereby increasing the risk of overfitting. Furthermore, high correlation can dilute the importance scores among the features. Finally, highly correlated features may introduce bias in gradient updates, affecting the optimization path of the gradient boosting process.

Since the independent variables in this study are categorical, we assessed the correlations between them using Cramér's V coefficient. For pairs of variables exhibiting high correlation, the feature with the lower importance score was excluded (see Figure 3). Consequently, five variables were removed: diagnosis result, May sequential sputum test, molecular biology test, drug susceptibility test, and strain identification.

The Performance of the XGBoost Model

Based on Table 2 and Figure 4, the model demonstrated outstanding performance in identifying "initial treatment" cases, achieving a precision of 0.98, a recall of 0.81, and an F1-score of 0.89. This indicates that the model maintains a low misdiagnosis rate while successfully identifying 81% of patients requiring initial treatment.

For "retreatment" cases, the model achieved a recall of 0.95, demonstrating high sensitivity in detecting actual retreatment patients. This high recall effectively minimizes the risk of delays in initiating second-line or salvage therapy due to missed diagnoses. Although the precision was 0.61, suggesting that some initial treatment cases might be misclassified as retreatment, such instances can be further addressed in clinical practice through secondary evaluations or supplemental examinations. Therefore, after careful consideration of the trade-offs, the model maintains a high level of safety and practical utility.

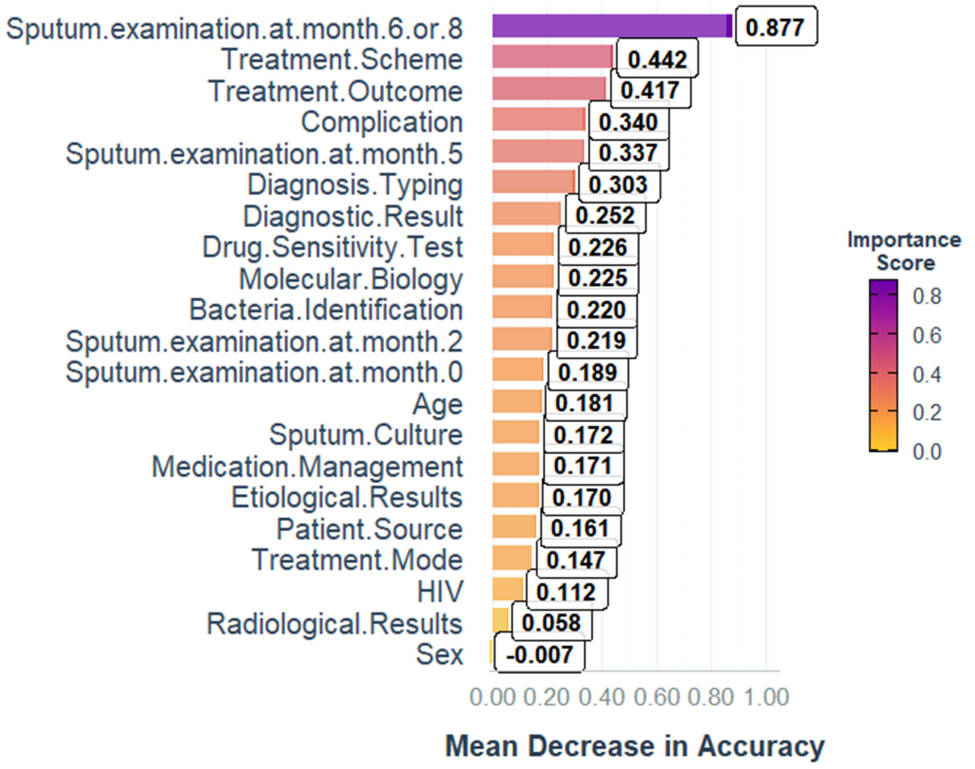

Figure 2 Feature Importance from the Random Forest Model.

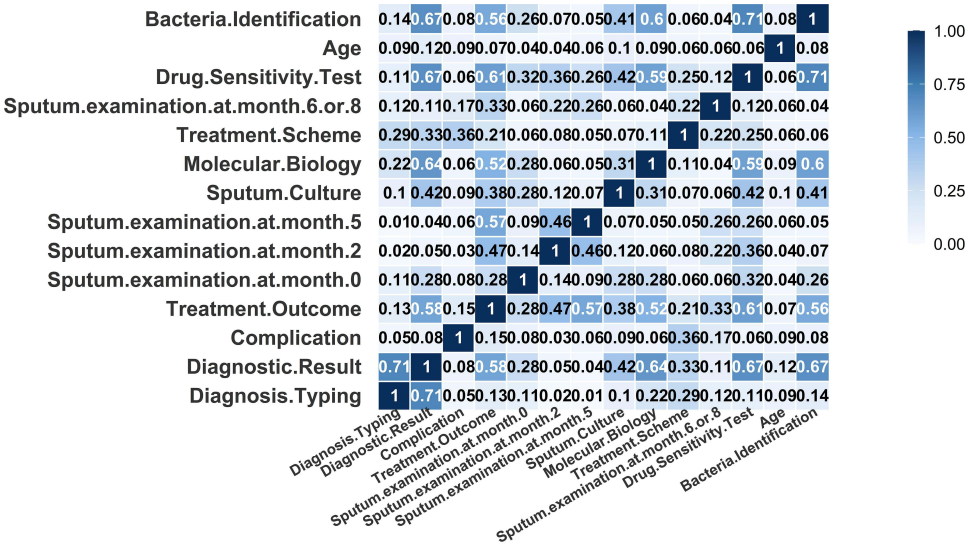

Figure 3 Cramér's V Correlation Matrix.

The model achieved an area under the ROC curve (AUC) of 0.938 and an average precision (AP) of 0.741. Both metrics are substantially higher than the random classifier baseline, further confirming the model's robustness across different classification thresholds and its exceptional ability to distinguish between the two classes.(Please refer to Figure 5 for details).

https://doi.org/10.2147/IDR.S594300

Table 2 Performance Metrics of the XGBoost Model for Distinguishing Initial Treatment from Retreatment Tuberculosis Patients

	Precision	Recall	F1-score	Support
New	0.98	0.81	0.89	1329
Retreatment	0.61	0.95	0.74	419
Accuracy			0.84	1748
Macro Avg	0.80	0.88	0.81	1748
Weighted Avg	0.89	0.84	0.85	1748

Note: Precision = TP/(TP+FP), Recall = TP/(TP+FN), F1-score = 2 × (Precision × Recall)/(Precision + Recall), Accuracy = (TP+TN)/(TP+TN+FP +FN), where TP = true positive (retreatment correctly classified), TN = true negative (initial treatment correctly classified), FP = false positive (initial treatment misclassified as retreatment), and FN = false negative (retreatment misclassified as initial treatment).

Feature Importance and SHAP Analysis of the XGBoost Model

In Figure 6, we employed the XGBoost model combined with SHAP value analysis to deeply explore the characteristics that identify individuals at high risk of tuberculosis retreatment. The research yielded a series of meaningful conclusions.

In Figure 7 and Table 3, the SHAP values are based on the predicted probability of the "initial treatment" outcome. A positive SHAP value indicates that the feature increases the model's prediction of "initial treatment" (ie., a protective factor against retreatment), while a negative SHAP value indicates that the feature increases the prediction of "retreatment" (ie., a risk factor for retreatment).

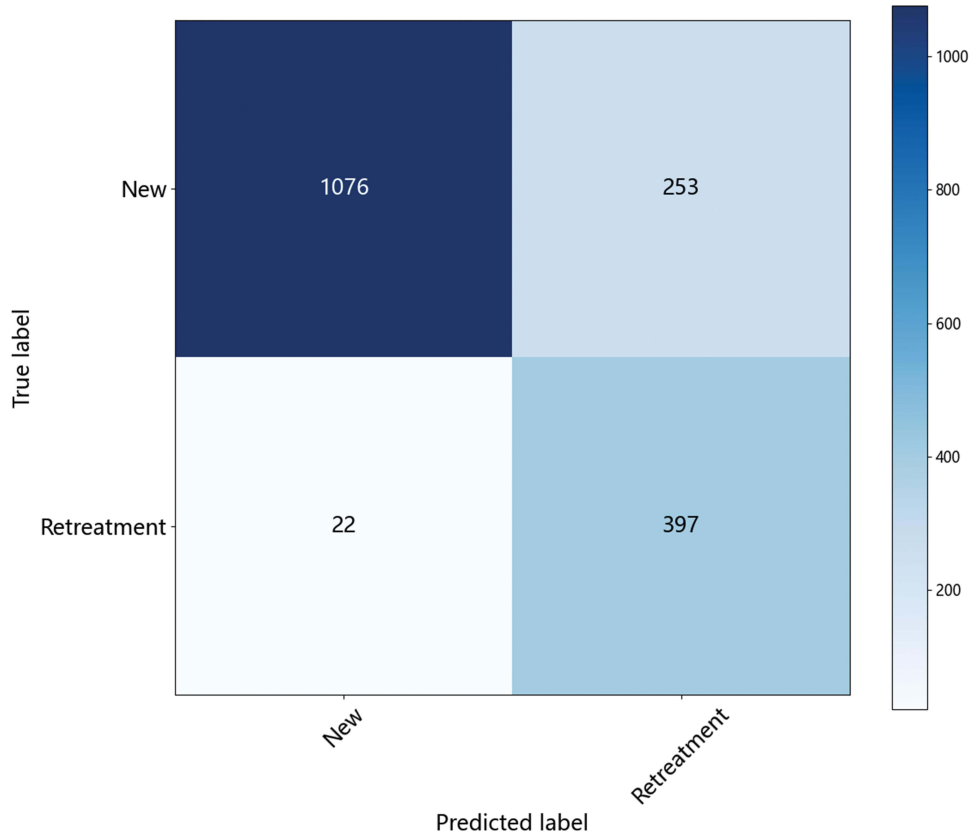

Figure 4 Confusion Matrix.

https://doi.org/10.2147/IDR.S594300

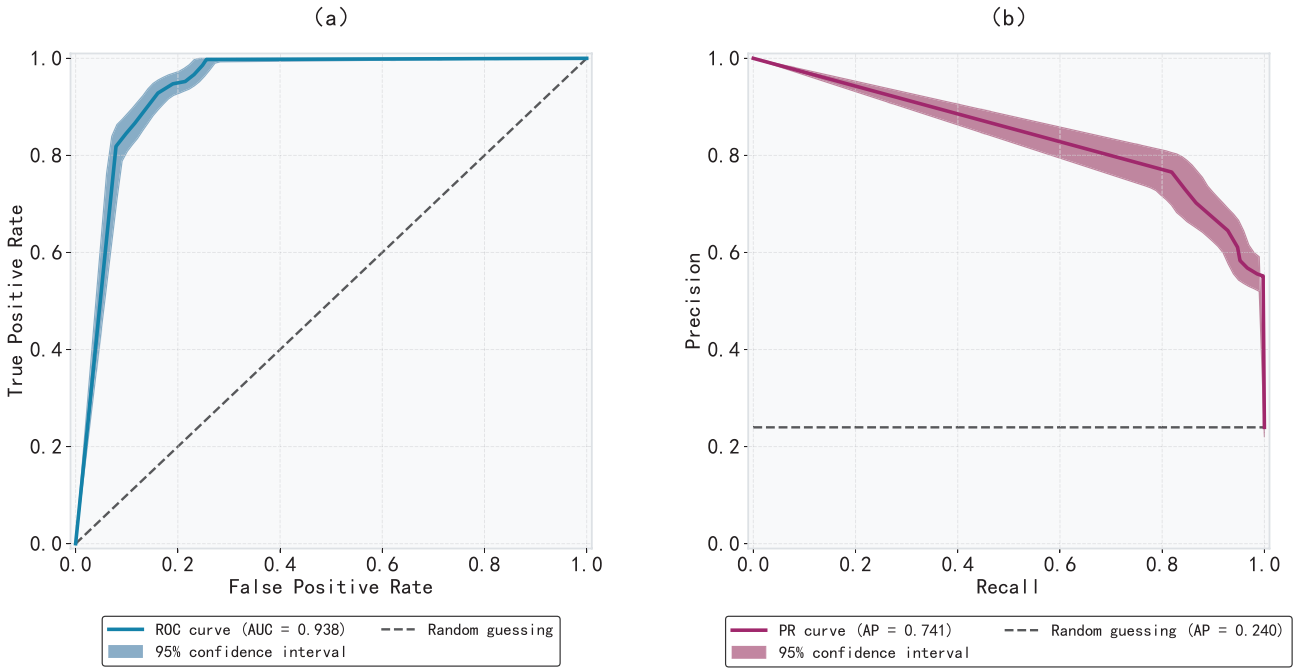

Figure 5 (**a**) Receiver Operating Characteristic curve (**b**) Precision-Recall curve.

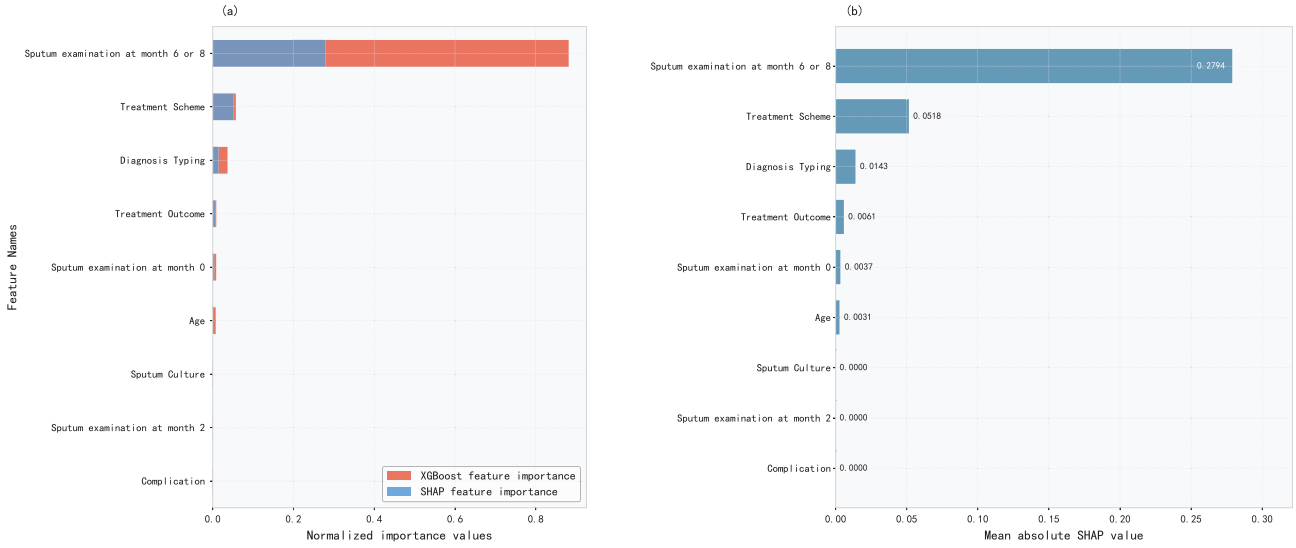

Figure 6 Comparison of global feature importance metrics for the XGBoost model: (**a**) Gain-based importance versus mean |SHAP value|, and (**b**) Top features based on Gain-based importance.

It was found that the intrinsic feature importance metric (Gain) of the XGBoost model showed high concordance with the ranking based on SHAP values. This indicates a consensus between the two methods in identifying critical variables. The alignment in ranking for specific features suggests a stable and reliable assessment of their importance by the XGBoost model, implying that their significance is not an artifact of the evaluation method and is thus highly credible.

Notably, three features were consistently identified as core decision factors influencing the "initial treatment" outcome: 1) sequential sputum test in June, 2) treatment regimen, and 3) diagnostic classification.(Please refer to Figure 6 and 7 and Table 3 for details).

https://doi.org/10.2147/IDR.S594300

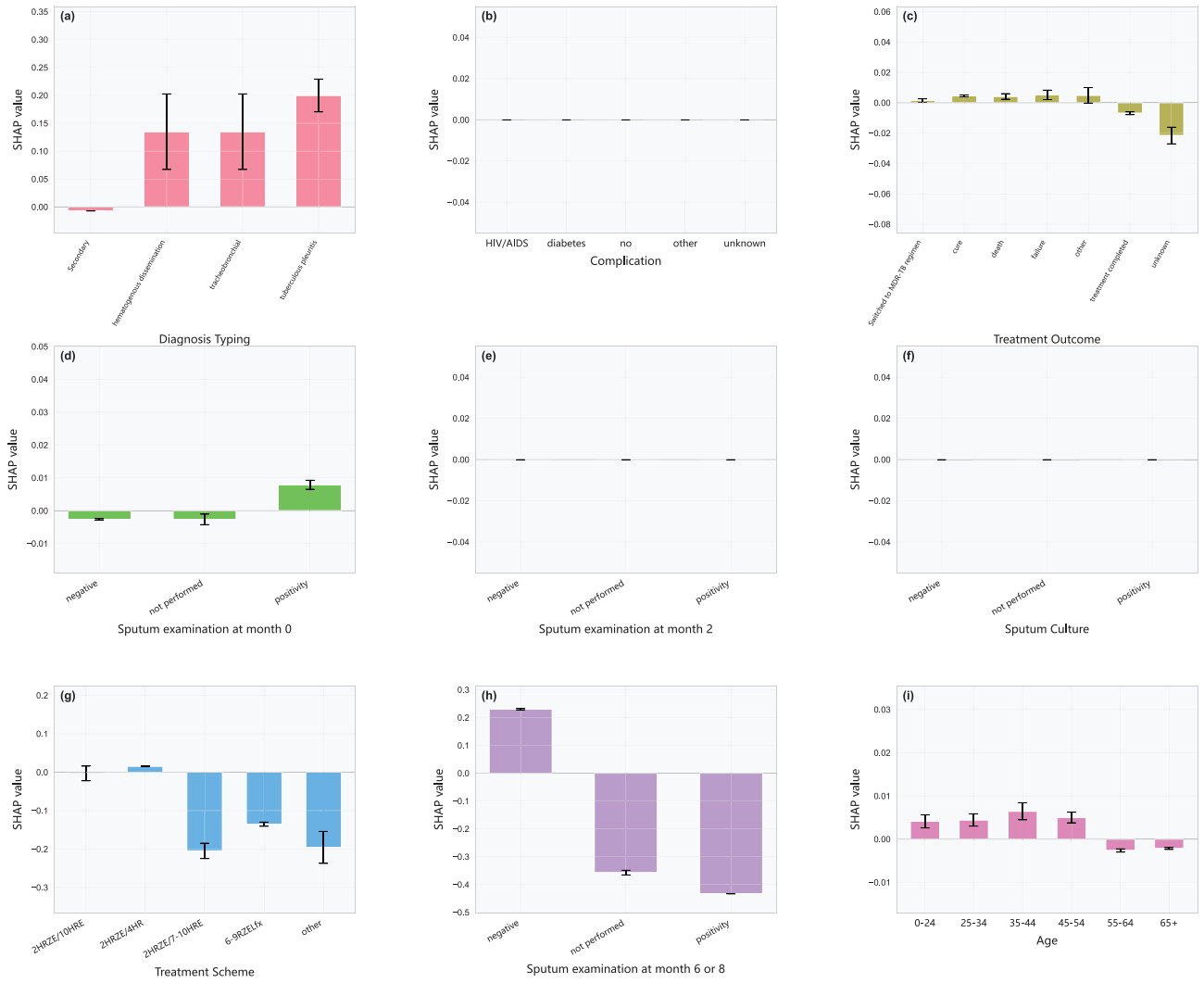

Figure 7 SHAP dependence plots for the "initial treatment" outcome in tuberculosis. (**a**) Diagnosis Typing; (**b**) Complication; (**c**) Treatment Outcome; (**d**) Sputum examination at month 0; (**e**) Sputum examination at month 2; (**f**) Sputum Culture; (**g**) Treatment Scheme; (**h**) Sputum examination at month 6 or 8; (**i**) Age.

Logistic Regression Analysis for Initial Treatment versus Retreatment in Tuberculosis

Multivariable logistic regression analysis identified several factors significantly associated with the risk of tuberculosis retreatment. Age demonstrated a dose-response relationship with increasing risk (45–54 years: *OR* = 2.17; 55–64 years: *OR* = 3.25; ≥65 years: *OR* = 4.09). A diagnosis of secondary tuberculosis was strongly associated with higher odds (*OR* = 5.16), as were various complications (*OR* range: 2.52–3.71). The strongest predictor was sputum smear status at treatment completion, with "not performed" (*OR* = 123.47) and "positive" (*OR* = 14.89) conferring substantially elevated risks. Specific treatment regimens (2HRZE/4HR: *OR* = 2.85) were also risk factors. Conversely, treatment outcome of death (*OR* = 0.40), positive initial sputum smear (*OR* = 0.63), "not performed" sputum smear at month 2 (*OR* = 0.32), and "not performed" sputum culture (*OR* = 0.73) were significantly associated with reduced odds of retreatment. (Please refer to Figure 8 and Table 4 for details).

Classification of Tuberculosis Patients

Latent class analysis identified the optimal subgroup classification for tuberculosis patients. As shown in Table 5 and Figure 9, both AIC and BIC values decreased continuously as the number of latent classes increased. To avoid excessive model complexity and overfitting, a maximum of five subgroups was set for this study. The latent class analysis revealed significant differences among all clinical characteristics across the five subgroups of tuberculosis patients (all *p*-values <

Table 3 SHAP Dependence Analysis for the Prediction of Initial Treatment (Protective) vs. Retreatment (Risk-Increasing) Outcomes: Mean SHAP Values, Standard Deviations, and Sample Sizes by Feature Category

Feature Name	Category	Mean SHAP Value	Std Dev of SHAP	Sample Size (n)
Sputum examination at month 6 or 8	Negative	0.23	0.04	1,073
Sputum examination at month 6 or 8	Not performed	−0.36	0.11	672
Sputum examination at month 6 or 8	Positivity	−0.43	0.00	3
Treatment Scheme	2HRZE/10HRE	0.00	0.15	226
Treatment Scheme	2HRZE/4HR	0.02	0.02	1,412
Treatment Scheme	2HRZE/7-10HRE	−0.20	0.10	89
Treatment Scheme	6-9RZELfx	−0.14	0.01	4
Treatment Scheme	Other	−0.20	0.09	17
Diagnosis Typing	Secondary	−0.01	0.00	1,678
Diagnosis Typing	Hematogenous dissemination	0.13	0.09	7
Diagnosis Typing	Tracheobronchial	0.13	0.09	7
Diagnosis Typing	Tuberculous pleuritis	0.20	0.11	56
Treatment Outcome	MDR-TB regimen	0.00	0.00	24
Treatment Outcome	Cure	0.00	0.01	901
Treatment Outcome	Death	0.00	0.01	69
Treatment Outcome	Failure	0.01	0.01	16
Treatment Outcome	Other	0.00	0.01	20
Treatment Outcome	Treatment completed	−0.01	0.01	645
Treatment Outcome	Unknown	−0.02	0.02	73
Sputum examination at month 0	Negative	0.00	0.00	1,363
Sputum examination at month 0	Not performed	0.00	0.00	14
Sputum examination at month 0	Positivity	0.01	0.01	371
Age	0-24	0.00	0.01	90
Age	25-34	0.00	0.01	116
Age	35-44	0.01	0.01	103
Age	45-54	0.01	0.01	194
Age	55-64	0.00	0.00	396
Age	65+	0.00	0.00	849
Sputum Culture	Negative	0.00	0.00	692
Sputum Culture	Not performed	0.00	0.00	424
Sputum Culture	Positivity	0.00	0.00	632
Complication	HIV/AIDS	0.00	0.00	6
Complication	Diabetes	0.00	0.00	136
Complication	No	0.00	0.00	492
Complication	Other	0.00	0.00	44
Complication	Unknown	0.00	0.00	1,070
Sputum examination at month 2	Negative	0.00	0.00	1,650
Sputum examination at month 2	Not performed	0.00	0.00	87
Sputum examination at month 2	Positivity	0.00	0.00	11

Note: Positive SHAP values indicate contribution toward "initial treatment" (protective), and negative values indicate contribution toward "retreatment" (risk-increasing).

0.001). Although the five-class model showed slightly better absolute fit indices, the four-class model was selected as the optimal and more pragmatic choice due to its higher classification clarity (entropy), better parsimony, and perfect alignment with predefined clinical subgroups. Specifically, Class 1 (Treatment Failure Type) was characterized by poor treatment outcomes, with significantly higher mortality (43.5%) and transition to MDR-TB treatment regimens (26.6%) compared to other classes, along with an extremely high rate of unexamined sputum tests during treatment (63.4% at 2 months, 98.1% at treatment completion). Class 2 (Treatment Success Type) was primarily defined by high success rates under standard treatment regimens (97.9% received the 2HRZE/4HR regimen), achieving a 99.8% sputum culture conversion rate at 2 months and with most patients (98.6%) cured or completing treatment. Class 3 (Diabetes

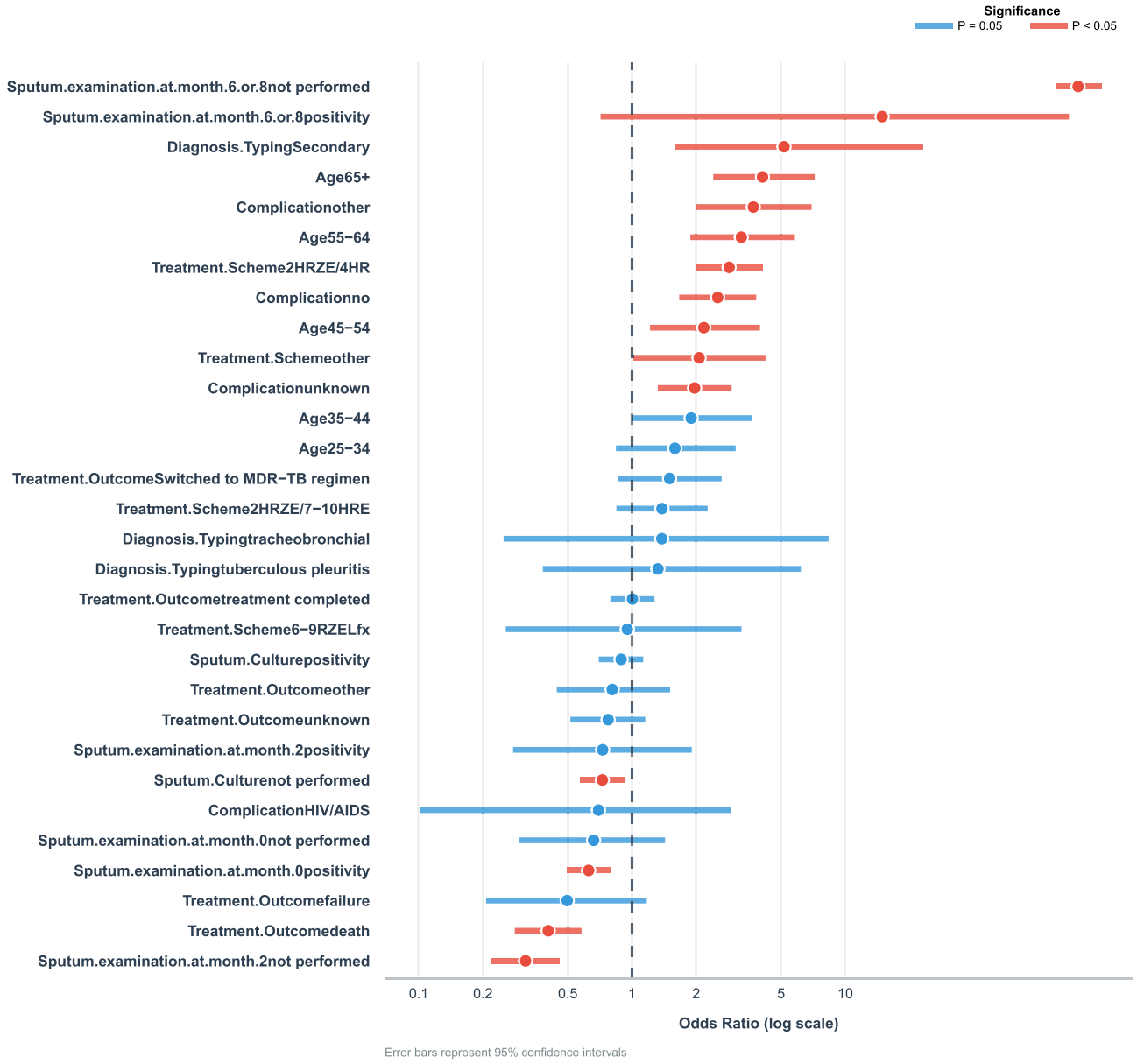

Figure 8 Forest Plot of Risk Factors for Tuberculosis Retreatment.

Comorbidity Type) was distinguished by a high proportion of diabetes comorbidity (49.3%), with most patients receiving long-term treatment regimens (81.0% on the 2HRZE/10HRE regimen) and exhibiting diverse treatment outcomes. Class 4 (High Infectivity with Rapid Control Type) presented with high pre-treatment sputum bacterial load (83.1% sputum culture positivity) but responded well to standard treatment regimens (96.2% on the 2HRZE/4HR regimen), achieving a 98.7% sputum culture conversion rate at 2 months and a high cure rate of 90.3%, demonstrating rapid and effective disease control.(Please refer to Table 6 for details).

Discussion

Retreatment of tuberculosis is a focal point in tuberculosis prevention and control. Analyzing and identifying high-risk factors for retreatment provides a scientific basis for precision prevention and control strategies, which holds significant importance for tuberculosis containment.Given the retrospective observational design of this study, all findings should be

Table 4 Multivariable Logistic Regression Analysis of Factors Associated with Tuberculosis Retreatment: Odds Ratios (OR), 95% Confidence Intervals, and P-values for Age, Diagnosis Typing, Complications, Treatment Outcomes, Sputum Examinations, and Treatment Regimens

Variable	OR	CI_95	*P*
Reference group:age_group 0–24			
Age25-34	1.59	0.841–3.062	0.16
Age35-44	1.89	1.005–3.641	0.05
Age45-54	2.17	1.215–3.983	0.01
Age55-64	3.25	1.879–5.805	0.00
Age65+	4.09	2.404–7.19	0.00
Reference group:			
Diagnosis.Typing hematogenous dissemination			
Diagnosis.Typing Secondary	5.16	1.598–23.183	0.01
Diagnosis.Typing tracheobronchial	1.38	0.25–8.347	0.71
Diagnosis.Typing tuberculous pleuritis	1.32	0.382–6.191	0.68
Reference group:Complication diabetes			
Complication HIV/AIDS	0.70	0.101–2.921	0.66
Complication no	2.52	1.665–3.823	0.00
Complication other	3.71	1.984–6.95	0.00
Complication unknown	1.97	1.321–2.934	0.00
Reference group:Treatment.Outcome cure			
Treatment.Outcome death	0.40	0.282–0.58	0.00
Treatment.Outcomefailure	0.50	0.207–1.174	0.11
Treatment.Outcomeother	0.81	0.444–1.509	0.49
Treatment.OutcomeSwitched to MDR-TB regimen	1.50	0.862–2.632	0.15
Treatment.Outcometreatment completed	1.00	0.792–1.275	0.97
Treatment.Outcome unknown	0.77	0.514–1.154	0.21
Reference group:			
Sputum.examination.at.month.0 (negative)			
Sputum.examination.at.month.0 (not performed)	0.66	0.296–1.429	0.30
Sputum.examination.at.month.0 (positivity)	0.63	0.496–0.789	0.00
Reference group:			
Sputum.examination.at.month.2 (negative)			
Sputum.examination.at.month.2 (not performed)	0.32	0.218–0.459	0.00
Sputum.examination.at.month.2 (positivity)	0.73	0.277–1.906	0.52
Reference group:Sputum.Culturenot negative			
Sputum.Culture not performed	0.73	0.57–0.925	0.01
Sputum.Culturepositivity	0.89	0.699–1.129	0.34
Reference group:Treatment.Scheme2HRZE/10HRE			
Treatment.Scheme2HRZE/4HR	2.85	1.986–4.103	0.00
Treatment.Scheme2HRZE/7-10HRE	1.38	0.845–2.262	0.20
Treatment.Scheme6-9RZELfx	0.95	0.256–3.262	0.94
Treatment.Schemeother	2.06	1.015–4.225	0.05
Reference group:			
Sputum.examination.at.month.6.or.8 (negative)			
Sputum.examination.at.month.6.or.8 (not performed)	123.47	96.754–159.225	0.00
Sputum.examination.at.month.6.or.8 (positivity)	14.89	0.712–111.851	0.02

https://doi.org/10.2147/IDR.S594300

Table 5 Model Fit Indices (AIC, BIC, Log-Likelihood) and Entropy for Latent Class Analysis with 2 to 5 Classes Among Tuberculosis Patients

Classes	AIC	BIC	LogLikelihood	Nparameters	Entropy
2	76,452.44	76,859.31	−38,165.22	61	0.84
3	74,128.20	74,741.85	−36,972.10	92	0.84
4	72,521.53	73,341.95	−36,137.76	123	0.88
5	71,404.95	72,432.14	−35,548.47	154	0.88

interpreted as associations rather than causal relationships. The identified risk factors are predictive indicators of tuberculosis retreatment, but causality cannot be inferred from this study design.

Figures 4 and 5 and Table 2 show that the overall accuracy of the model reaches 84%, with notable performance differences between categories. The model demonstrates excellent capability in identifying newly diagnosed tuberculosis (TB) patients, with a precision of 0.98 and a recall of 0.81. However, its ability to identify recurrent TB patients is relatively limited, with a precision of 0.61 and a recall of 0.95. This discrepancy is primarily attributed to imbalanced data distribution, obscured subgroup heterogeneity, and biases in the weighting of dynamic monitoring indicators.[17] The area under the ROC curve (AUC) is 0.938, and the area under the Precision-Recall curve (AP) is 0.741, both significantly higher than random classification thresholds. This confirms the model's robustness across different thresholds and its superior discriminative ability, indicating that the model maintains good classification performance under various conditions.

Results from Figure 6 reveal that the XGBoost model, combined with SHAP (SHapley Additive exPlanations), identifies sputum smear results at the 6th or 8th month of treatment, the treatment regimen, and diagnostic type as key influencing factors. Among these, sputum smear results at the 6th or 8th month rank first in importance in both methods. This importance stems from the fact that sputum conversion in recurrent TB patients is typically slower or remains persistently positive, consistent with findings from existing studies.[18] The treatment regimen, as a high-risk influencing factor, not only reflects differences in drug combinations but also implies complex clinical information such as history of prior treatment failure, potential drug resistance, patient adherence, and disease severity.[19] Diagnostic type serves as a risk factor by affecting potential drug resistance risk, disease severity, and the patient's immune baseline, thus indirectly influencing recurrence risk.[20]

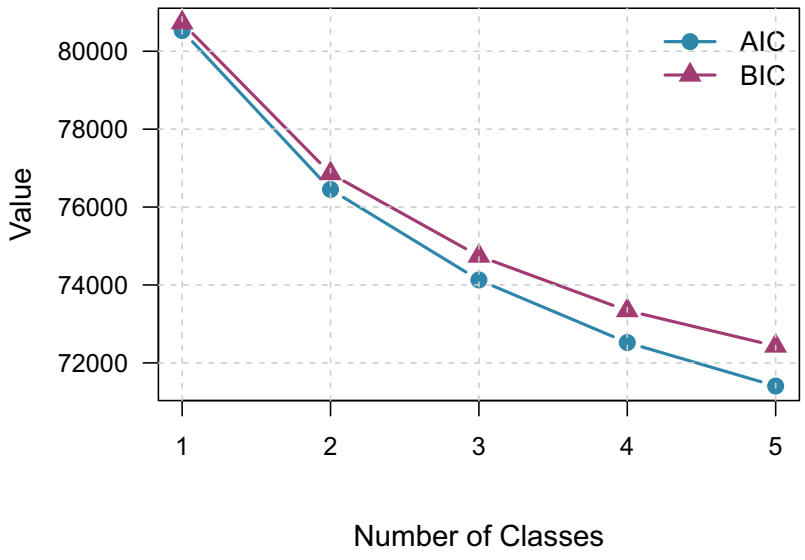

Figure 9 Trends in Model Fit Indices with Increasing Number of Latent Classes in Latent Class Analysis.

Table 6 Comparison of Clinical Characteristics Across Four Latent Classes of Tuberculosis Patients

Variable	Class 1 (n=361)	Class 2 (n=2517)	Class 3 (n=970)	Class 4 (n=1978)	Statistic	*P*
Age						
0-24	8 (2.2%)	150 (6%)	64 (6.6%)	79 (4%)	113.916	<0.001
25-34	22 (6.1%)	172 (6.8%)	73 (7.5%)	99 (5%)		
35-44	25 (6.9%)	159 (6.3%)	68 (7%)	101 (5.1%)		
45-54	27 (7.5%)	336 (13.3%)	133 (13.7%)	189 (9.6%)		
55-64	62 (17.2%)	565 (22.4%)	246 (25.4%)	417 (21.1%)		
65+	217 (60.1%)	1135 (45.1%)	386 (39.8%)	1093 (55.3%)		
Diagnosis typing						
Hematogenous dissemition	0 (0%)	0 (0%)	23 (2.4%)	0 (0%)	1179.343	<0.001
Secondary	361 (100%)	2515 (99.9%)	741 (76.4%)	1978 (100%)		
Tracheobronchial	0 (0%)	0 (0%)	23 (2.4%)	0 (0%)		
Tuberculous pleuritis	0 (0%)	2 (0.1%)	183 (18.9%)	0 (0%)		
Complication						
Diabetes	5 (1.4%)	4 (0.2%)	478 (49.3%)	7 (0.4%)	2665.386	<0.001
HIV/AIDS	0 (0%)	4 (0.2%)	20 (2.1%)	2 (0.1%)		
No	111 (30.7%)	806 (32%)	163 (16.8%)	503 (25.4%)		
Other	11 (3%)	64 (2.5%)	43 (4.4%)	33 (1.7%)		
Unknown	234 (64.8%)	1639 (65.1%)	266 (27.4%)	1433 (72.4%)		
Treatment outcome						
Cure	29 (8%)	745 (29.6%)	366 (37.7%)	1787 (90.3%)	6262.252	<0.001
Death	157 (43.5%)	5 (0.2%)	71 (7.3%)	1 (0.1%)		
Failure	23 (6.4%)	0 (0%)	9 (0.9%)	0 (0%)		
Other	25 (6.9%)	31 (1.2%)	11 (1.1%)	0 (0%)		
Switched to MDR-TB regimen	96 (26.6%)	0 (0%)	14 (1.4%)	0 (0%)		
Treatment completed	14 (3.9%)	1736 (69%)	298 (30.7%)	176 (8.9%)		
Unknown	17 (4.7%)	0 (0%)	201 (20.7%)	14 (0.7%)		
Sputum examition at month 0						
Negative	217 (60.1%)	2509 (99.7%)	733 (75.6%)	1166 (58.9%)	1393.934	<0.001
Not performed	23 (6.4%)	7 (0.3%)	16 (1.6%)	3 (0.2%)		
Positivity	121 (33.5%)	1 (0%)	221 (22.8%)	809 (40.9%)		
Sputum examition at month 2						
Negative	120 (33.2%)	2513 (99.8%)	911 (93.9%)	1952 (98.7%)	2931.237	<0.001
Not performed	229 (63.4%)	3 (0.1%)	53 (5.5%)	2 (0.1%)		
Positivity	12 (3.3%)	1 (0%)	6 (0.6%)	24 (1.2%)		
Sputum culture						
Negative	87 (24.1%)	1863 (74%)	354 (36.5%)	87 (4.4%)	3675.881	<0.001
Not performed	150 (41.6%)	654 (26%)	303 (31.2%)	247 (12.5%)		
Positivity	124 (34.3%)	0 (0%)	313 (32.3%)	1644 (83.1%)		
Treatment scheme						
2HRZE/10HRE	0 (0%)	0 (0%)	786 (81%)	3 (0.2%)	5286.24	<0.001
2HRZE/4HR	339 (93.9%)	2463 (97.9%)	0 (0%)	1903 (96.2%)		
2HRZE/7-10HRE	5 (1.4%)	44 (1.7%)	152 (15.7%)	53 (2.7%)		
6-9RZELfx	4 (1.1%)	4 (0.2%)	3 (0.3%)	7 (0.4%)		
Other	13 (3.6%)	6 (0.2%)	29 (3%)	12 (0.6%)		
Sputum examition at month 6 or 8						
Negative	2 (0.6%)	1975 (78.5%)	265 (27.3%)	1355 (68.5%)	1441.968	<0.001
Not performed	354 (98.1%)	542 (21.5%)	704 (72.6%)	623 (31.5%)		
Positivity	5 (1.4%)	0 (0%)	1 (0.1%)	0 (0%)		

https://doi.org/10.2147/IDR.S594300

Figure 7 and Table 3 further reveals the non-linear impact of these key features on the model's predictions. The sputum smear result at the 6th or 8th month of treatment has a decisive influence on the model's prediction. A "negative" result is a significant protective factor (mean SHAP value = 0.230), while "not examined" (mean SHAP value = −0.357) and "positive" (mean SHAP value = −0.433) are both associated with a very high risk of adverse outcomes. This suggests that standardized sputum examination during the mid-to-late stages of treatment is a critical step in assessing prognosis.[21] In contrast, sputum smear results at the initial stage of treatment have minimal impact, indicating that the model places greater emphasis on continuous monitoring rather than a single baseline measurement.[22] Regarding treatment regimens, the standard "2HRZE/4HR" regimen (mean SHAP value = 0.016) shows a contribution value close to zero and has the largest sample size, reflecting its widespread effectiveness. In contrast, longer regimens for complex cases, such as "2HRZE/7-10HRE" (mean SHAP value = −0.205), are significantly associated with higher risk, primarily because the population receiving such regimens inherently carries a higher baseline risk. For diagnostic types, "tuberculous pleurisy" (mean SHAP value = 0.2) shows a higher risk contribution than the more common "secondary pulmonary tuberculosis." This subtype may represent a patient subgroup with unique clinical characteristics (eg., extrapulmonary lesions, delayed diagnosis), and its independent value as a risk indicator warrants attention.[23]

Logistic regression analysis further quantifies the strength of association between key factors and recurrence risk, providing intuitive validation of the core role of treatment monitoring. Using sputum smear "negative" at the 6th or 8th month of treatment as the reference, the recurrence risk for patients with sputum "not examined" surges (*OR*=123.47), and the risk for those with sputum "positive" also increases significantly (*OR*=14.89). This highlights the decisive impact of obtaining bacteriological evidence during the mid-to-late stages of treatment on prognosis.[24,25] Increasing age shows a clear dose-response relationship with recurrence risk. Using the 0–24 age group as reference, risk begins to rise significantly from the 45–54 age group (*OR*=2.17) and peaks in the 65+ age group (OR=4.09), suggesting that enhanced monitoring and individualized management are needed for elderly patients.[26] Analysis of treatment regimens reveals significant clinical selection bias. Using the "2HRZE/10HRE" regimen as the reference group, the standard "2HRZE/4HR" regimen shows a stronger association with higher risk (*OR*=2.85). This result does not indicate inferior efficacy of the standard regimen but strongly reflects its role as a first-line therapy applied to a broader patient population. In contrast, the risk association for longer regimens like "2HRZE/7-10HRE" did not reach statistical significance, possibly due to their use for more complex cases with smaller sample sizes.[27]

Several findings that appear protective warrant cautious interpretation. First, the negative association between death and retreatment (*OR* = 0.40) reflects a competing risk phenomenon rather than a true protective effect: patients who die during treatment are no longer at risk for retreatment. Second, the lower retreatment risk associated with positive baseline sputum smear (*OR* = 0.63) may seem counterintuitive but likely results from enhanced clinical monitoring and management of patients with higher initial bacterial load, rather than a biological protective effect. Third, the "protective" associations for "not performed" sputum examination at month 2 (*OR* = 0.32) and "not performed" sputum culture (OR = 0.73) should be interpreted with caution; these may be attributable to selection bias (patients with favorable clinical response were less likely to be tested) or coding artifacts (eg., "not performed" may include patients who completed treatment early and were thus no longer under monitoring). Given the retrospective design, these associations should not be misinterpreted as causal protective effects.

This study identified four patient subgroups through Latent Class Analysis (LCA): treatment-failure patients (Class 1), treatment-success patients (Class 2), patients with comorbid diabetes (Class 3), and highly infectious but rapidly controlled patients (Class 4). Classes 2 and 4 represent low-risk groups, exemplifying the successful paradigm of the current TB control system and validating the effectiveness of standardized management. Their characteristics (overall high treatment success rate, minimal comorbidities, excellent sputum conversion rate) represent the successful paradigms within the current TB control system.[28] Identifying these groups allows for more efficient allocation of public health resources rather than a one-size-fits-all approach. Conversely, Classes 1 and 3 constitute high-risk groups, revealing the clinical dilemma that "drug susceptibility does not guarantee treatment success" and the independent impact of comorbid diabetes on treatment outcomes. Patients with comorbid diabetes highlight the critical role of comorbidity management in TB treatment.[29] Even with drug-susceptible strains, the presence of diabetes may complicate treatment independently, possibly by affecting immune response or drug metabolism. Treatment-failure patients (Class 1) reflect weaknesses in

treatment management (slower sputum conversion), systematic gaps in treatment monitoring (extremely high rates of "not examined" at various time points), and a vicious cycle of treatment failure.[30]

This study finds that treatment adherence and efficacy monitoring are key to success or failure. Both LCA and SHAP analysis indicate that the "not examined" status is highly associated with recurrence. This strongly suggests that failure to complete sputum smear examinations at critical time points is itself an independent, more prevalent, and more alarming risk signal than a positive smear result. This directly guides us to prioritize ensuring the completion rate of sputum examinations at key time points as a core performance indicator for assessing healthcare system quality and optimizing patient management strategies. The high proportion of drug resistance in Class 4 identified by LCA corroborates the extremely high risk associated with "rifampicin resistance" and "multidrug resistance" in the logistic regression analysis. The root cause of all these issues points to failures in the management of initial treatment. Therefore, strengthening supervised management for newly diagnosed patients, ensuring they complete the full course of standardized treatment, is the most cost-effective and fundamental strategy for preventing acquired drug resistance. Advanced age and the use of longer treatment regimens indicate more complex patient conditions or the presence of initial drug resistance, alerting healthcare providers to initiate more stringent management and monitoring processes for such patients.

Several limitations of this study should be acknowledged. First, the retrospective observational design precludes causal inference; all identified associations are predictive rather than causal. Second, data were collected from a single region (Kashgar, China), which may limit generalizability to other populations with different epidemiological profiles. Third, the "not performed" category in sputum examinations may introduce information bias, as the reasons for missing tests (eg., clinical improvement, loss to follow-up, or logistical issues) could not be determined from the available data. Fourth, despite the use of sample weights to address class imbalance, the model's precision for retreatment cases remained moderate (0.61), suggesting that additional predictors not captured in this dataset may further improve classification. Fifth, unmeasured confounders (eg., socioeconomic status, patient adherence beyond recorded data, HIV status with CD4 counts) may influence both treatment outcomes and retreatment risk. Future prospective studies with standardized follow-up protocols and larger sample sizes are needed to validate these findings.

Conclusions

This study utilized clinical data from tuberculosis patients and applied Random Forest and Cramér's V for feature selection. An XGBoost predictive model was constructed, and SHAP was employed to interpret feature contributions. The research identified several factors associated with tuberculosis retreatment. The methodology adopted in this study demonstrates potential clinical value: it may help address nonlinear relationships and sample imbalance in tuberculosis data, and SHAP improves model interpretability. Given the retrospective observational design, these findings should be interpreted as associations rather than causal relationships. Based on these associations, the following considerations may inform tuberculosis control programs: optimizing treatment for patients with comorbid diabetes, enhancing treatment adherence and efficacy monitoring for newly diagnosed patients, strengthening whole-course supervised management, and implementing further optimized management for elderly patients and those on long-term treatment regimens.

Abbreviations

AIC, Akaike Information Criterion; AP, Average Precision; AUC, Area Under the ROC Curve; BIC, Bayesian Information Criterion; LCA, Latent Class Analysis; MDR-TB, Multidrug-Resistant Tuberculosis; OR, Odds Ratio; PR Curve, Precision-Recall Curve; RF, Random Forest; ROC Curve, Receiver Operating Characteristic Curve; SHAP, SHapley Additive exPlanationsTB, Tuberculosis; XGBoost, eXtreme Gradient Boosting.

Ethical Approval and Consent to Participate

This retrospective observational study was conducted in accordance with the principles of the Declaration of Helsinki. The study protocol was approved by the Ethics Committee of Xinjiang Medical University (Approval No. XJYKDXR20240724011). The requirement for informed consent was waived by the ethics committee because: (1) this study utilized only existing, fully de-identified historical medical records and did not involve any patient intervention; (2) obtaining consent from each patient was impracticable and would have rendered the retrospective study

unfeasible; and (3) the research posed no more than minimal risk to participants. All data were handled with strict confidentiality, and researchers had no access to personally identifiable information.

Acknowledgments

The authors appreciate the works by the Kashgar CDC.

Funding

Project of Top-notch Talents of Technological Youth of Xinjiang [Grant No. 2024TSYCCX0080]. This study was funded by the grants from the National Natural Science Foundation of China (72174175, 72064036, 72163033) and the College Student Innovation and Entrepreneurship Training Program (Grant No. S202510760111).

Disclosure

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.

References

1. Dheda K, Mirzayev F, M CD, et al. Multidrug-resistant Tuberculosis. *Nat. Rev. Dis. Primers*. 2024;10(1). doi:10.1038/s41572-024-00504-2.
2. Janssen S, Murphy M, Upton C, Allwood B, Diacon AH. Tuberculosis: an Update for the Clinician. *Respirology*. 2025;30(3):196–205. doi:10.1111/resp.14887
3. World Health Organization. Global tuberculosis report 2025. Geneva:World Health Organization,2025.
4. Vega V, Cabrera-Sanchez J, Rodríguez S, et al. Risk factors for pulmonary tuberculosis recurrence, relapse and reinfection: a systematic review and meta-analysis. *BMJ Open Respir Res*. 2024;11(1):e002281. doi:10.1136/bmjresp-2023-002281
5. Hermans SM, Akkerman OW, Meintjes G, Grobusch MP. Post-tuberculosis treatment paradoxical reactions. *Infection*. 2024;52(5):2083–2095. doi:10.1007/s15010-024-02310-0
6. Vo TTB, Nguyen DT, Nguyen TC, et al. Exploring gene mutations and multidrug resistance in Mycobacterium tuberculosis: a study from the Lung Hospital in Vietnam. *Mol. Biol. Rep*. 2024;51(1). doi:10.1007/s11033-024-10015-8.
7. Lv H, Zhang X, Zhang X, et al. Global prevalence and burden of Multidrug-resistant tuberculosis from 1990 to 2019. *BMC Infect Dis*. 2024;24(1). doi:10.1186/s12879-024-09079-5.
8. Naidoo K, Perumal R, Cox H, et al. The epidemiology, transmission, diagnosis, and management of drug-resistant tuberculosis—lessons from the South African Experience. *Lancet Infect Dis*. 2024;24(9):e559–e575. doi:10.1016/S1473-3099(24)00144-0
9. Jin C, Wu Y, Chen J, et al. Prevalence and patterns of Drug-resistant Mycobacterium tuberculosis in newly diagnosed patients in China: a systematic review and meta-Analysis. *J Global Antimicrob Resist*. 2024;38:292–301. doi:10.1016/j.jgar.2024.05.018
10. Sambarey A, Smith K, Chung C, et al. Integrative analysis of multimodal patient data identifies personalized predictors of tuberculosis treatment Prognosis. *IScience*. 2024;27(2):109025. doi:10.1016/j.isci.2024.109025
11. Liang D, Wang L, Zhong P, et al. Perspective: global burden of iodine deficiency: insights and projections to 2050 using xgboost and SHAP. *Adv Nutr*. 2025;16(3):100384. doi:10.1016/j.advnut.2025.100384
12. M RS, Shiddik B. A.Utilizing artificial intelligence to predict and analyze socioeconomic, environmental, and healthcare factors driving tuberculosis globally. *Sci Rep*. 2025;15(1):13619. doi:10.1038/s41598-025-96973-w
13. Wang S, Li Z, Zhang T, et al. An interpretable machine learning approach reveals the interaction between air pollutants and climate factors on tuberculosis. *Urban Climate*. 2025;102420.
14. Wang Z, Guo Z, Wang W, et al. Prediction of tuberculosis treatment outcomes using biochemical makers with machine learning. *BMC Infect. Dis*. 2025;25(1):229.
15. Pal A, Mohanty D. Pal A,Mohanty D.Machine learning-based approach for identification of new resistance associated mutations from whole genome sequences of Mycobacterium tuberculosis. *Bioinform Adv*. 2025;5(1):vbaf050. doi:10.1093/bioadv/vbaf050
16. Kong H, Li Y, Shen Y, et al. Predicting the risk of pulmonary embolism in patients with tuberculosis using machine learning algorithms. *Eur. J. Med. Res*. 2024;29(1):618. doi:10.1186/s40001-024-02218-3
17. Regan M, Barham T, Li Y, et al. Risk factors underlying racial and ethnic disparities in tuberculosis diagnosis and treatment outcomes, 2011-19: a multiple mediation analysis of national surveillance data. *Lancet Public Health*. 2024;9(8):e564–e572. doi:10.1016/S2468-2667(24)00151-8
18. Yue X, Yanfei C, Ruijian H, et al. Interpretable machine learning in predicting drug-induced liver injury among tuberculosis patients: model development and validation study. *BMC Med. Res. Method*. 2024;24(1):92. doi:10.1186/s12874-024-02214-5
19. Xu R, Zhang Y, Li Z, et al. Breathomics for diagnosing tuberculosis in diabetes mellitus patients. *Front Mol Biosci*. 2024;1436135.
20. Srinivasan S, H D, S HRR, et al. Evaluating factors influencing tuberculosis treatment outcomes and the impact of COVID-19 on TB incidence in Bengaluru, India (2017-2023). *Infectious Diseases*. 2025;1–9.
21. Guo K, Xu X, Zhan Q, et al. Study on influencing factors of tuberculosis based on logistic regression and decision tree model. *Soc Med Health Manage*. 2025;6(1):1235–1245.
22. Zhou F, Sun Q, Huang S, et al. Trends and delays in pulmonary tuberculosis diagnosis among elderly patients (≥ 60 Years) in Southern China: a 13-year surveillance data analysis (2010–2022). *BMC Public Health*. 2025;25(1):1854. doi:10.1186/s12889-025-23031-5
23. Zhang W, Chen J, Chen Z, et al. Differentiating nontuberculous mycobacterial pulmonary disease from pulmonary tuberculosis in resource-limited settings: a pragmatic model for reducing misguided antitubercular treatment. *Healthcare*. 2025;13(9):1065. doi:10.3390/healthcare13091065

24. L FM, Magwaza C, Dlatu N, et al. Exploring determinants and predictive models of latent tuberculosis infection outcomes in rural areas of the eastern cape: a pilot comparative analysis of logistic regression and machine learning approaches. *Information*. 2025;16(3)):239.
25. Mok J, Jeong D, Sohn H, et al. Nationwide coverage of molecular drug susceptibility testing in patients with pulmonary multidrug/rifampicin-resistant tuberculosis in South Korea: a retrospective cohort study (2015-2021). *BMJ Open Respir. Res*. 2025;12(1):e003307. doi:10.1136/bmjresp-2025-003307
26. Xue D, Chen X, Shao L, et al. Risk factors for the progression from pulmonary tuberculosis to spinal tuberculosis: a logistic regression analysis. *J. Orthop. Surg. Res*. 2025;20(1):422. doi:10.1186/s13018-025-05848-3
27. Ma Z, Liu X, Zhang M, et al. Differences analysis between spinal tuberculosis and brucella spondylitis with preoperative non-invasive differential diagnosis. *European Spine Journal*. 2025;34(2):1–9. doi:10.1007/s00586-025-08647-w
28. Zhang L, Ma X, Gao H, et al. Analysis of care-seeking and diagnosis delay among pulmonary tuberculosis patients in Beijing, China. *Front Public Health*. 2024;1369541.
29. Rupani PM. Silicosis predicts drug resistance and retreatment among tuberculosis patients in India: a secondary data analysis from Khambhat, Gujarat (2006–2022). *BMC Pulm Med*. 2024;24(1):522. doi:10.1186/s12890-024-03338-6
30. J LY, Myong J, Kim Y, et al. Identifying predictors of unfavorable treatment outcomes in tuberculosis patients. *Int J Environ Res Public Health*. 2024;21(11):1454. doi:10.3390/ijerph21111454