WebLog Journal of Biotechnology and Biomedical Engineering

Research Article
Published: 08 Aug, 2026

Machine Learning-Based Classification of Tuberculosis Using Whole-Blood Gene Expression Profiles: A Multi-Cohort Transcriptomic Study with SHAP-Based Interpretability

*Adeel Asghar**

Islamia University of Bahawalpur, Pakistan

Abstract

Background: Tuberculosis diagnosis remains a practical challenge in high-burden settings, where available tests have well-known limitations around sensitivity, cost, and turnaround time. Whole-blood gene expression profiling offers a non-sputum-based alternative approach, but most published machine learning models in this space rely on complex pipelines or proprietary datasets. This study asked a simpler question: using publicly available microarray data and standard classifiers, can we build something that works, generalizes, and makes biological sense.

Methods: We used GSE19491 as the primary dataset, which contained 417 whole-blood transcriptomic samples across three classes: TB, Other Infection, and Control. After median imputation, z-score normalization, and ANOVA-based feature selection retaining the top 500 probes, SMOTE was applied to the training set to address class imbalance. Four models were trained and compared: Random Forest, SVM, Logistic Regression, and XGBoost. Bootstrap resampling with 200 iterations generated 95% confidence intervals for accuracy and weighted F1-score. SHAP values were calculated for the best-performing model. Pathway enrichment was performed using Enrichr. External validation used GSE37250, an independent Malawian cohort of 537 samples.

Results: Logistic Regression achieved the best performance with an accuracy of 0.929 and a weighted F1-score of 0.926, with a 95% confidence interval for accuracy of 0.869 to 0.976 (Table 1). Random Forest and XGBoost tied at 0.881 accuracy. AUC-ROC for the Random Forest was 0.91 for Control, 1.00 for Other Infection, and 0.95 for TB (Figure 2). On external validation in GSE37250, overall accuracy was 0.583, with TB precision of 0.79 and Other Infection recall of 0.73 (Figure 7). The top SHAP predictors included FKSG30, LOC402221, PDCD10, COX15, ARHGEF1, ACO2, IDH1, and ADAR (Table 3, Figures 5 and 6). Pathway enrichment showed consistent enrichment of TCA cycle, oxidative phosphorylation, Rho GTPase signaling, and mRNA editing pathways across multiple gene set libraries (Table 4).

Conclusions: A Random Forest classifier trained on 500 ANOVA-selected genes from whole-blood microarray data can distinguish TB from other infections with high accuracy in the primary cohort and with meaningful but reduced performance in an independent population. The genes driving the model connect to mitochondrial metabolism, macrophage signalling, and innate immune regulation, which are all pathways TB is known to interact with. These findings support the biological validity of the transcriptomic signature and suggest a targeted gene panel built around the top SHAP predictors could be worth developing and testing in a prospective clinical setting.

Keywords: Tuberculosis; Gene Expression; Machine Learning; Random Forest; SHAP, Transcriptomics; GSE19491; GSE37250; Whole Blood; Biomarker

OPEN ACCESS

Correspondence:
Adeel Asghar, Islamia University of Bahawalpur, Pakistan,
E-mail: decentboys867hh@gmail.com

Received Date: *24 Jun 2026*
Accepted Date: *06 Aug 2026*
Published Date: *08 Aug 2026*

Citation:
Adeel Asghar. Machine Learning-Based Classification of Tuberculosis Using Whole-Blood Gene Expression Profiles: A Multi-Cohort Transcriptomic Study with SHAP-Based Interpretability. WebLog J Biotechnol Biomed Eng. wjbtbe.2026.h0807. https://doi.org/10.5281/zenodo.21860321

Introduction

Tuberculosis remains one of the leading infectious causes of death worldwide. In 2023, TB returned to being the world's leading infectious disease killer, surpassing COVID-19, with an estimated 10.8 million people falling ill and 1.25 million deaths recorded [1, 2]. The burden falls heaviest on low- and middle-income countries, particularly in Sub-Saharan Africa and South Asia, where diagnostic infrastructure is often limited and patient access to specialist care is inconsistent

[1].

The core problem with TB diagnosis is not that we lack tests. We have smear microscopy, culture, GeneXpert, tuberculin skin tests, and interferon-gamma release assays. The problem is that none of this work well in every situation. Sputum smear microscopy has significant limitations because it can only detect TB when sputum has sufficient bacillary load, and sensitivity is substantially reduced in patients with HIV co-infection, where bacillary load in sputum is often lower [3, 4]. GeneXpert, while faster and more accurate, requires a machine costing around US$17,000 and per-test cartridges at US$9.98 each, with additional requirements for reliable power supply and technical support [5]. In high-burden settings with limited resources, this creates real barriers. Meta-analysis data show that GeneXpert sensitivity for smear-negative TB is 67% (95% CrI 60-74%) as an add-on test following negative smear, with reduced performance in people living with HIV compared with HIV-negative patients [6]. Culture is the gold standard but takes weeks. IGRA tests cannot distinguish active from latent TB. In practice, clinicians in high-burden settings are frequently making treatment decisions under genuine diagnostic uncertainty.

This gap has pushed researchers toward blood-based transcriptomic signatures. The concept was established in a landmark 2010 study by Berry et al., who identified a 393-transcript whole-blood signature for active TB that correlated with radiological extent of disease and reverted to healthy control levels after treatment [7]. Since then, multiple groups have identified and validated TB transcriptomic signatures across diverse cohorts, and whole-blood transcriptomics has emerged ahead of proteomics and metabolomics for diagnostic biomarker discovery in TB as a result of established sample-processing pathways [8, 9]. More recently, machine learning approaches have been applied to these datasets, with several studies identifying small gene panels that meet WHO diagnostic targets for triage tests [10, 11].

What has been less explored is whether standard machine learning classifiers, applied to publicly available microarray data without specialized preprocessing pipelines, can produce robust and biologically interpretable results. A 2019 study by Bobak et al. showed that Random Forest had an optimal global sensitivity of 0.8973 across 1164 samples from 4 countries [12], and a study by Maertzdorf *et al*., demonstrated that tree-based methods could generate TB classifier models that generalized across South African, Gambian, and Indian cohorts [13]. However, most of these studies either use complex integration frameworks or focus on binary TB *vs*. healthy classification, leaving the harder three-class problem of distinguishing TB from other active infections less well characterized.

That is what this study set out to do. We used two publicly available GEO datasets, GSE19491 and GSE37250, and trained four classification models, including Random Forest, Support Vector Machine, Logistic Regression, and XGBoost, on whole-blood gene expression profiles from patients with TB, other infections, and healthy controls. We wanted to see which model performed best, whether the genes driving the predictions made biological sense, and whether a model trained on one cohort could generalize meaningfully to a geographically distinct cohort it had never seen. Using SHAP values [14], we looked at which genes were actually driving individual predictions and in which direction, and we followed that up with pathway enrichment to connect the computational findings to known TB biology. The goal was not to overstate what machine learning can do here, but to produce something honest and interpretable that could serve as a foundation for more targeted experimental work.

Materials and Methods

Data sources

Two publicly available gene expression datasets were retrieved from the NCBI Gene Expression Omnibus (GEO). The primary dataset, GSE19491, consisted of 417 whole-blood transcriptomic samples from patients with Pulmonary Tuberculosis (PTB), latent TB infection, other infectious diseases (Staphylococcal, Streptococcal, and autoimmune conditions), and healthy controls, profiled on the Illumina HumanHT-12 V3.0 platform (GPL6947). The validation dataset, GSE37250, comprised 537 whole-blood samples from an independent Malawian cohort including active TB patients, latent TB, and other disease controls, profiled on the Illumina HumanHT-12 V4.0 platform (GPL10558). Characteristics of both datasets are summarized in Table 2 [15].

Data preprocessing

Raw expression values were extracted using the GEOparse library in Python. Sample labels were mapped to three diagnostic classes: TB (including active and latent TB), Other Infection, and Control. Samples without assignable labels were excluded, yielding 417 samples for primary analysis. Missing values were imputed using median imputation per gene *via* scikit-learn's Simple Imputer. Expression values were standardized using z-score normalization (mean=0, standard deviation=1) with Standard Scaler.

Feature selection

Dimensionality reduction was performed using ANOVA F-test (SelectKBest, scikit-learn), retaining the top 500 most differentially expressed probes across the three diagnostic classes from an initial set of 48,803 probes. Probe identifiers were mapped to HGNC gene symbols using the GPL6947 platform annotation table.

Class balancing

The training set exhibited class imbalance, with the Control class underrepresented relative to TB and Other Infection groups. Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to the training set to generate synthetic minority samples and achieve balanced class distribution prior to model training. SMOTE generates synthetic samples by selecting minority class instances and creating new samples along the line connecting those instances and their k-nearest neighbours in feature space [16].

Machine learning models

Four supervised classification models were trained and evaluated: Random Forest (RF), Support Vector Machine (SVM) with RBF kernel, Logistic Regression (LR), and XGBoost. All models were trained on an 80/20 stratified train-test split with random state fixed at 42 for reproducibility. Class weights were set to balance for RF, SVM, and LR. XGBoost does not natively support a balanced class_ weight parameter in the same manner; class balance for XGBoost was addressed exclusively through SMOTE-based oversampling of the training set prior to model fitting. Model performance was assessed using accuracy, weighted F1-score, and area under the receiver operating characteristic curve (AUC-ROC) in a one-versus-rest framework. Random Forest uses an ensemble of decision trees built on bootstrap samples of the data, with random feature selection at each split, producing both low bias and low variance through averaging [17].

Statistical validation

To quantify uncertainty in model performance estimates, 95% confidence intervals were calculated for accuracy and weighted F1-score using bootstrap resampling with 200 iterations. Five-fold stratified cross-validation was additionally performed on the Random Forest model to provide a robust estimate of generalization performance.

Explainability analysis

SHAP (SHapley Additive exPlanations) values were computed for the best-performing model using the TreeExplainer algorithm to quantify the contribution of each gene to individual predictions [14, 18, 19]. SHAP summary and bar plots were generated to visualize directional feature effects across the test set (Figures 5 and 6).

Pathway enrichment analysis

The top 20 genes identified by SHAP analysis were submitted to Enrichr for pathway enrichment analysis [20] Results were evaluated across KEGG 2026, Reactome 2024, WikiPathways 2024, BioPlanet 2019, and MSigDB Hallmark 2020 gene set libraries.

External validation

The trained Random Forest model was applied without retraining to the independent GSE37250 dataset. Probe alignment was performed by identifying common probes between the two platforms (39,426 of 48,803 probes matched). The 500 feature-selected genes were all present in the validation set. Independent imputation and scaling were applied to the validation data prior to prediction.

Software and reproducibility

All analyses were performed in Python 3.13 using the following libraries: GEOparse 2.0, scikit-learn 1.x, imbalanced-learn, XGBoost, SHAP, pandas, numpy, matplotlib, and seaborn. The complete analysis notebook is available upon request.

Results

Dataset characteristics

The primary dataset GSE19491 had 498 samples total. After removing 81 samples that lacked a clear disease label, 417 samples remained for analysis: 154 TB, 221 Other Infection, and 42 Control. The class imbalance in the Control group was addressed using SMOTE before training. The validation dataset GSE37250 had 537 samples across TB and Other Infection only, with no healthy controls included in that cohort. Full dataset characteristics are shown in Table 2.

Feature selection

Starting from 48,803 probes, ANOVA F-test selection reduced the feature space to 500 genes most differentially expressed across the three classes. This cut removed the majority of low-signal probes and made the models faster and more interpretable without losing discriminative information.

Model performance on primary dataset

All four models were evaluated on a held-out test set of 84 samples. Logistic Regression performed best with an accuracy of 0.929 and a weighted F1-score of 0.926. Random Forest and XGBoost tied at 0.881 accuracy (F1: 0.879 and 0.880 respectively), and SVM performed worst at 0.810. SHAP-based feature interpretation was applied to the Random Forest model, which provides tree-based additive attributions and was retained for its biological interpretability despite being numerically tied for second place [12]. The detailed performance figures with 95% confidence intervals from bootstrap resampling (n=200) are shown in Table 1 and Figure 1.

The confidence intervals for Logistic Regression and SVM do not overlap, providing statistical grounding to their performance difference. Five-fold cross-validation on the Random Forest gave a mean F1 of 0.8931, which is close to the test set weighted F1 of 0.879 and confirms the absence of severe overfitting. ROC-AUC scores for the Random Forest were 0.91 for Control, 1.00 for Other Infection, and 0.95 for TB (Figure 2). The perfect AUC for Other Infection should be interpreted with caution given the small test set size. The confusion matrix for the primary model is shown in Figure 1.

Gene-Level findings

The top 20 genes selected by SHAP analysis for TB classification are shown in Table 3 and Figure 6. What stands out is that many of these are not the classic TB biomarkers cited repeatedly in the literature. FKSG30 and LOC402221 emerged as the two strongest drivers, both with high expression pushing the model away from a TB call, suggesting they may act as exclusion markers rather than positive TB indicators. PDCD10, ACO2, IDH1, and COX15 all connect to mitochondrial metabolism and cell death pathways, consistent with what is known about how TB manipulates host cell biology to survive inside macrophages [22-25]. ADAR appeared in SHAP (Figure 5) but was not among the top 20 probes by basic feature importance (Figure 4), illustrating how SHAP captures genes that contribute meaningfully to individual predictions even when their average importance looks modest [14, 18].

Pathway enrichment

Pathway enrichment analysis of the top 20 SHAP genes using Enrichr [20] showed consistent enrichment of the TCA cycle across KEGG 2026, BioPlanet 2019, WikiPathways, and HumanCyc, driven by ACO2, IDH1, and COX15. Rho GTPase signalling and cytoskeletal regulation pathways appeared through ARHGEF1. The mRNA editing pathway was enriched *via* ADAR. Full pathway results are shown in Table 4.

External validation

When the trained Random Forest was applied to GSE37250 without any retraining, accuracy dropped to 0.583. At the class level, TB precision held at 0.79 and Other Infection recall was 0.73. The confusion matrix for the validation cohort is shown in Figure 7. These numbers suggest the model is picking up on a real signal even in a

Table 1: Model performance with 95% bootstrap confidence intervals on held-out test set (n=84).

Model	Accuracy	95% CI Accuracy	F1 Weighted	95% CI F1
Random Forest	0.881	[0.798 – 0.952]	0.879	[0.804 – 0.952]
Logistic Regression	0.929	[0.869 – 0.976]	0.926	[0.862 – 0.975]
XGBoost	0.881	[0.798 – 0.952]	0.880	[0.796 – 0.952]
SVM	0.810	[0.726 – 0.893]	0.826	[0.743 – 0.902]

Table 3: Top 20 genes by SHAP value with directional effect on TB prediction.

Rank	Probe ID	Gene Symbol	Mean |SHAP|	Direction of Effect on TB
1	ILMN_1814998	FKSG30	0.01923	Negative (high expression excludes TB)
2	ILMN_1811909	LOC402221	0.01567	Negative (high expression excludes TB)
3	ILMN_2365196	PDCD10	0.00701	Mixed
4	ILMN_1718309	COX15	0.00685	Negative
5	ILMN_2405129	ARHGEF1	0.00693	Negative
6	ILMN_1770373	TMEM39A	0.00652	Mixed
7	ILMN_1708204	COPG	0.00641	Mixed
8	ILMN_1791478	MTPN	0.00608	Mixed
9	ILMN_1750088	VRK2	0.00601	Mixed
10	ILMN_2063586	CLIC4	0.00598	Mixed
11	ILMN_1704795	MARK3	0.00591	Mixed
12	ILMN_1654861	ACO2	0.00614	Negative
13	ILMN_1696432	IDH1	0.00604	Negative
14	ILMN_2320964	ADAR	0.00596	Mixed
15	ILMN_1678730	NOMO1	0.00588	Mixed
16	ILMN_1695420	CLTA	0.00586	Mixed
17	ILMN_1785660	SRPR	0.00547	Negative
18	ILMN_1798886	NUDT21	0.00537	Negative
19	ILMN_2394571	FBXW11	0.00512	Mixed
20	ILMN_1796738	SPAST	0.00498	Mixed

Table 4: Top enriched pathways from Enrichr analysis of top 20 SHAP genes.

Pathway	Database	Key Genes	Biological Relevance
Citrate Cycle (TCA Cycle)	KEGG 2026	ACO2, IDH1, COX15	Mtb reprograms host mitochondrial metabolism for intracellular survival
Tricarboxylic Acid (TCA) Cycle	BioPlanet 2019	ACO2, IDH1, COX15	Metabolic dysregulation in TB-infected host cells
Amino Acid Metabolism	WikiPathways 2024	Multiple	Host metabolic reprogramming during infection
Oxidative Phosphorylation	MSigDB Hallmark 2020	COX15, ACO2	Energy metabolism disruption in infected blood cells
Fatty Acid Metabolism	MSigDB Hallmark 2020	Multiple	TB-driven lipid metabolism alteration in macrophages
Rho GTPase / Cytoskeletal Regulation	NCI-Nature / Panther	ARHGEF1	Macrophage cytoskeletal regulation and phagocytosis mechanics
mRNA Editing	Reactome 2024	ADAR	Post-transcriptional innate immune regulation against intracellular pathogen
Glyoxylate and Dicarboxylate Metabolism	KEGG 2026	ACO2, IDH1	Connected to TCA cycle disruption
TCA Cycle (Human)	HumanCyc 2016	ACO2, IDH1	Consistently enriched across multiple databases confirming metabolic signature

completely new population, despite the overall accuracy drop that is expected in cross-cohort transcriptomic validation [26-28].

Discussion

Overview

This study used whole-blood gene expression data to train and compare four machine learning classifiers for distinguishing TB from other infections and healthy controls. The goal was not to build a clinical tool ready for deployment, but to ask whether publicly available transcriptomic data, processed with standard machine learning methods, can produce something biologically meaningful and externally valid [10, 12]. The answer from this data is mostly yes, with some important caveats.

Why random forest worked best

Logistic Regression performed best, followed by a tie between Random Forest and XGBoost. With 417 samples and significant class imbalance in the Control group, neither RF nor XGBoost had enough data to fully exploit their ensemble mechanisms over the simpler regularised linear model. Random Forest is still more forgiving than XGBoost in this regime because it builds independent trees rather than sequentially correcting errors, which makes it less prone to overfitting on small imbalanced datasets [17]. SVM performed worst among the four, which is somewhat expected when the feature space is high-dimensional even after selection. Logistic Regression outperformed all other models with an accuracy of 0.929, which may reflect that after SMOTE balancing the decision boundary separating TB from the other classes is approximately linear in the 500-dimensional feature space, exactly the condition under which Logistic Regression generalises well. Random Forest and XGBoost tied at 0.881; with only 417 samples, neither boosting method had sufficient data to take full advantage of its sequential error-correction mechanism. Five-fold cross-validation gave a mean weighted F1 of 0.8931 for Random Forest, which is close to its test set F1 of 0.879 and confirms no

severe overfitting. SHAP attribution was applied to Random Forest rather than Logistic Regression because tree-based SHAP values are exact (not approximated), providing more reliable per-feature attribution for the biological interpretation that follows. Nonetheless, multiple independent studies applying machine learning classifiers to TB transcriptomic datasets have consistently demonstrated strong diagnostic performance, with the optimal model varying by dataset size and platform characteristics [12, 13].

What the gene findings actually mean

The two strongest SHAP predictors were FKSG30 and LOC402221. Both have high expression in non-TB samples and push the model away from a TB call when expressed. This pattern makes them more useful as exclusion signals than confirmation signals. In a clinical context, that is actually valuable. A test that reliably rules out TB when these genes are highly expressed could help triage patients faster, even if it cannot confirm TB on its own.

PDCD10, ACO2, IDH1, and COX15 all sit within mitochondrial and metabolic pathways. Mtb infection induces a quiescent energy phenotype in human macrophages with decelerated flux through glycolysis and the TCA cycle, and shifts mitochondrial dependency from endogenous to exogenous fatty acids [23]. Mtb infection reprograms host macrophage carbon metabolism, with active TCA cycle fluxes and significantly altered amino acid profiles observed in infected cells [22]. The fact that the model independently picked up on metabolic gene expression as a discriminating feature, without any biological knowledge built in, suggests it learned something real rather than a statistical artifact.

PDCD10 connects to apoptosis regulation, which is directly relevant to TB pathogenesis. Virulent Mtb suppresses host cell apoptosis, which is a protective response that would otherwise contain the bacteria, and instead induces necrotic cell death to allow bacterial release and spread [24, 25]. The presence of PDCD10 in the top predictors is consistent with this known TB immune evasion strategy.

ADAR showing up in SHAP but not in the feature importance ranking is worth paying attention to. ADAR1 modulates innate immune response to intracellular double-stranded RNA by affecting the RLR, PKR, OAS, and ZBP1 pathways that sense dsRNA [29]. ADAR1 knockdown in primary macrophages significantly enhances interferon, cytokine, and chemokine production, and its disruption deregulates the RLR-MAVS signalling pathway [30]. Its contribution to individual TB predictions, even with modest average importance, suggests a role in post-transcriptional immune regulation that is detectable at the whole-blood transcriptome level.

CLIC4 and ARHGEF1 both connect to macrophage biology. CLIC4 is involved in inflammatory signalling and ARHGEF1 regulates Rho GTPase activity, which controls how macrophages engulf and process pathogens. These cytoskeletal and phagosomal pathways are implicated in the wider TB cell death literature [24, 31].

Pathway enrichment findings

The most consistently enriched pathway across multiple databases was the TCA cycle, driven by ACO2, IDH1, and COX15. This is biologically meaningful because Mtb remodels host macrophage carbon metabolism in a stage-dependent manner: TCA cycle fluxes and oxidative phosphorylation remain elevated during early infection to maintain mitochondrial integrity, while established infection subsequently induces a more quiescent bioenergetic phenotype as the bacterium adapts to intracellular persistence [22, 23, 32]. Seeing this emerge from a purely data-driven analysis on blood samples from real patients suggests this metabolic signature is detectable at the whole-blood transcriptome level and not just in cell culture or animal models. Rho GTPase and cytoskeletal regulation pathways through ARHGEF1 connect to macrophage phagocytosis mechanics. The mRNA editing pathway *via* ADAR points toward post-transcriptional immune regulation [29, 30]. To be clear, pathway enrichment from 20 genes is hypothesis-generating at best. These findings do not prove causality. What they show is that the genes the model found predictive map onto pathways that TB research has independently implicated through very different experimental methods. That convergence is meaningful.

External validation and what the drop means

The drop from 0.929 to 0.583 accuracy in GSE37250 needs context. GSE37250 is a Malawian cohort; GSE19491 is mostly UK and South African samples. Different populations, different TB strain backgrounds, likely different disease stages at sample collection, and different lab processing protocols all contribute. Differences in gene expression platforms, annotation, preprocessing, and batch effects across studies are well-documented barriers to cross-cohort generalization in transcriptomics [26, 27]. A single TB signature can exhibit poor generalizability in diverse patient populations due to heterogeneity in cohort demographics, disease comorbidities, and profiling technologies [27]. However, there is also evidence that the core immune transcriptional response distinguishing active TB from latent TB is conserved across populations despite differences in genetic background and TB epidemiology [28].

Given all of that, the model still achieved TB precision of 0.79 in a completely new population it had never seen. That means when it called TB, it was right nearly four out of five times. These are not numbers to dismiss. If this model were to be developed further for clinical use, population-specific fine-tuning would be necessary. ComBat normalization or similar batch correction methods applied before validation would likely improve these numbers meaningfully. That is a reasonable next step for future work rather than a flaw in the current analysis.

Limitations

The primary dataset has only 417 samples, with just 42 Control samples even before the train-test split. SMOTE helps with class balance but generates synthetic data points, not real patients [16]. The model has not been tested in a prospective clinical setting. The validation dataset lacked a healthy control class, which meant the three-class performance could not be fully evaluated externally. The generalization drop, while explainable, is real and should not be minimized. This is a computational study using existing public data. The findings are a starting point for hypothesis generation and further experimental validation, not a finished clinical product.

Where this could go

The practical value of this kind of work is in narrowing down the gene list for targeted assay development. A future study could design a targeted panel around the top 20 SHAP genes and test it prospectively in a clinical setting. Whole-blood gene expression tests are already in clinical use for other conditions, and a TB-specific signature based on metabolic and immune regulatory genes like those identified here could in principle be developed into a simpler point-of-care format. Several published signatures already meet WHO triage test targets

using parsimonious gene panels tested with digital PCR platforms [11, 33]. The model also raises an interesting question about latent TB. GSE37250 grouped latent TB with active TB in this analysis. A future study keeping latent and active TB separate would be able to ask whether the transcriptomic signature differs between them, which has real implications for screening programs.

Conclusion

This study shows that machine learning classifiers trained on 500 ANOVA-selected whole-blood gene expression features can distinguish TB from other infections with high accuracy in the primary cohort. Logistic Regression achieved the highest accuracy of 0.929; Random Forest and XGBoost tied at 0.881. SHAP analysis applied to the Random Forest model (AUC 0.95 for TB) provided biologically interpretable gene-level findings. Results in an independent Malawian validation cohort showed meaningful but reduced performance (TB precision 0.79). The gene-level findings from SHAP analysis and pathway enrichment align with known TB host biology, particularly mitochondrial metabolic reprogramming, macrophage apoptosis regulation, and innate immune signalling. The cross-cohort accuracy drop is real but consistent with known challenges in transcriptomic generalization across geographically and demographically distinct populations [26, 27]. These findings support the use of publicly available GEO datasets combined with interpretable machine learning as a valid approach to TB biomarker discovery, and provide a foundation for future prospective validation of the identified gene panel.

References

1. World Health Organization. Tuberculosis. WHO. 2026.
2. Lee H, Kim J, Kim J, Park YJ. Review of the global burden of tuberculosis in 2023: insights from the WHO Global Tuberculosis Report 2024. Jugan Geongang Gwa Jilbyeong. 2025; 18(11) :55-69.
3. Saeed M, Hussain S, Riaz S, Rasheed F, Ahmad M, Iram S, et al. GeneXpert technology for the diagnosis of HIV-associated tuberculosis: is scale-up worth it? Open Life Sci. 2020; 15(1): 458-465.
4. Cuong NK, Ngoc NB, Hoa NB, Dat VQ, Nhung NV. GeneXpert on patients with human immunodeficiency virus and smear-negative pulmonary tuberculosis. PLoS One. 2021; 16(7): e0253961.
5. Piatek AS, Van Cleeff M, Alexander H, Cogwin WL, Rehr M, Van Kampen S, et al. GeneXpert for TB diagnosis: planned and purposeful implementation. Glob Health Sci Pract. 2013;1(1):18-23.
6. Steingart KR, Schiller I, Horne DJ, Pai M, Boehme CC, Dendukuri N. Xpert MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults. Cochrane Database Syst Rev. 2014; (1): CD009593.
7. Berry MPR, Graham CM, McNab FW, Xu Z, Bloch SAA, Oni T, et al. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. Nature. 2010; 466(7309): 973-977.
8. Roe JK, Thomas N, Gil E, Best K, Tsaliki E, Morris-Jones S, et al. Blood transcriptomic diagnosis of pulmonary and extrapulmonary tuberculosis. JCI Insight. 2016; 1(16): e87238.
9. Maertzdorf J, Ota M, Repsilber D, Mollenkopf HJ, Weiner J, Hill PC, et al. Functional correlations of pathogenesis-driven gene expression signatures in tuberculosis. PLoS One. 2011; 6(10): e26938.
10. Omrani M, Ghodousi A, Cirillo DM. An integrative and comprehensive analysis of blood transcriptomes combined with machine learning models reveals key signatures for tuberculosis diagnosis and risk stratification. Front Microbiol. 2025; 16: 1546770.
11. Gliddon HD, Kaforou M, Alikian M, Habgood-Coote D, Zhou C, Oni T, et al. Identification of reduced host transcriptomic signatures for tuberculosis disease and digital PCR-based validation and quantification. Front Immunol. 2021; 12: 637164.
12. Bobak CA, Titus AJ, Hill JE. Comparison of common machine learning models for classification of tuberculosis using transcriptional biomarkers from integrated datasets. Appl Soft Comput. 2019; 74: 264-273.
13. Maertzdorf J, McEwen G, Weiner J, Tian S, Lader E, Schriek U, et al. Concise gene signature for point-of-care classification of tuberculosis. EMBO Mol Med. 2016; 8(2): 86-95.
14. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017; 30: 4765-4774.
15. Chen L, Hua J, He X. Coexpression network analysis-based identification of critical genes differentiating between latent and active tuberculosis. Biomed Res Int. 2022; 2022: 2090560.
16. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002; 16: 321-357.
17. Díaz-Uriarte R, Álvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006; 7: 3.
18. Conard AM, DenAdel A, Crawford L. A spectrum of explainable and interpretable machine learning approaches for genomic studies. WIREs Comput Stat. 2023; 15(5): e1617.
19. Ponce-Bobadilla AV, Schmitt V, Maier CS, Mensing S, Stodtmann S. Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development. Clin Transl Sci. 2024; 17(11): e70056.
20. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016; 44(W1): W90-W97.
21. Hossain MS, Khandocar MP, Riti FA, Ali MY, Dey PR, Haque SMJ, et al. A comprehensive machine learning for high throughput tuberculosis sequence analysis, functional annotation, and visualization. Sci Rep. 2025; 15: 25866.
22. Borah Slater K, Moraes L, Xu Y, Kim D. Metabolic flux reprogramming in Mycobacterium tuberculosis-infected human macrophages. Front Microbiol. 2023; 14: 1289987.
23. Cumming BM, Addicott KW, Adamson JH, Steyn AJ. Mycobacterium tuberculosis induces decelerated bioenergetic metabolism in human macrophages. eLife. 2018; 7: e39169.
24. Nisa A, Kipper FC, Panigrahy D, Tiwari S, Kupz A, Subbian S. Different modalities of host cell death and their impact on Mycobacterium tuberculosis infection. Am J Physiol Cell Physiol. 2022; 323(5): C1444-C1474.
25. Lee J, Repasy T, Papavinasasundaram K, Sassetti C, Kornfeld H. Mycobacterium tuberculosis induces an atypical cell death mode to escape from infected macrophages. PLoS One. 2011; 6(3): e18367.
26. Wang X, Harper K, Sinha P, Johnson WE, Patil P. Analysis of the cross-study replicability of tuberculosis gene signatures using 49 curated human transcriptomic datasets. Tuberculosis. 2025; 153: 102649.
27. Vargas R, Abbott L, Bower D, Frahm N, Shaffer M, Yu WH, et al. Gene signature discovery and systematic validation across diverse clinical cohorts for TB prognosis and response to treatment. PLoS Comput Biol. 2023; 19(7): e1010770.
28. Siddalingaiah HS. Cross-geographic validation demonstrates universal transcriptomic signatures for tuberculosis diagnosis: a machine learning study. Research Square. 2025.
29. Pujantell M, Riveira-Muñoz E, Badia R, Castellví M, Garcia-Vidal E, Sirera G, et al. RNA editing by ADAR1 regulates innate and antiviral immune functions in primary macrophages. Sci Rep. 2017; 7: 13339.

30. Wang Q, Li X, Qi R, Billiar T. RNA editing, ADAR1, and the innate immune response. Genes. 2017; 8(1): 41.

31. Vu A, Glassman I, Campbell G, Yeganyan S, Nguyen J, Shin A, et al. Host cell death and modulation of immune response against *Mycobacterium tuberculosis* infection. Int J Mol Sci. 2024; 25(11): 6255.

32. Liu T, Yang Z, Tong J, Gao M, Pang Y. Glucose metabolic reprogramming of macrophages against Mycobacterium tuberculosis infection. Immunotargets Ther. 2025; 14: 1209-1221.

33. Alsayed AO. Machine-learning-derived, mechanistically informed transcriptomic signature to diagnose active tuberculosis and guide host-directed therapy. Diagnostics. 2026; 16(5): 693.