



OPEN ACCESS

EDITED BY
Xinfei Wang,
Jilin University, China

REVIEWED BY
Tongmao Ma,
Polytechnic University of Madrid, Spain
Kai Zheng,
China University of Petroleum, China

*CORRESPONDENCE
Tingting Li,
✉ lt881117@163.com

RECEIVED 14 January 2026
REVISED 29 January 2026
ACCEPTED 09 February 2026
PUBLISHED 03 March 2026

CITATION

Li L, Liu H, Lei Q and Li T (2026) From clinical phenotypes to genomic signatures: machine learning integration for precision tuberculosis treatment prediction.
Front. Bioinform. 6:1787360.
doi: 10.3389/fbinf.2026.1787360

COPYRIGHT

© 2026 Li, Liu, Lei and Li. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

From clinical phenotypes to genomic signatures: machine learning integration for precision tuberculosis treatment prediction

Liping Li¹, Huanqing Liu², Qian Lei³ and Tingting Li^{4*}

¹Yizhi School of Agriculture and Forestry, Xianyang Vocational Technical College, Xian Yang, Shaanxi, China, ²Information Management Office, Northwestern Polytechnical University, Xi'an, Shaanxi, China, ³Department of Pharmacy, Xi'an Chest Hospital, Xi'an, Shaanxi, China, ⁴Drug Clinical Trial Institution Office, Xi'an Chest Hospital, Xi'an, Shaanxi, China

Background: Tuberculosis (TB) remains a major global health threat, causing approximately 1.5 million deaths each year. Despite progress in treatment, 15%–20% of patients still experience treatment failure or relapse, highlighting the urgent need for precise predictive tools for early identification of high-risk patients. Current methods based on clinical parameters have limitations in prediction accuracy and revealing potential biological mechanisms.

Methods: This study developed and validated an innovative multi-omics integration prediction model. We retrospectively collected clinical data from 467 tuberculosis patients and integrated transcriptomic data from three independent public cohorts (GSE19491, GSE31312, GSE83456), involving 3,240 differentially expressed genes. Through advanced feature engineering and bioinformatics analysis, key features were selected. We systematically evaluated 12 machine learning algorithms and adopted an ensemble learning strategy to construct the final model. Model performance was evaluated through strict cross-validation and prospective validation cohorts.

Results: Clinical data analysis identified age, body mass index (BMI), and C-reactive protein (CRP) levels as significant predictors of treatment response. Transcriptomic analysis revealed 1,247 differentially expressed genes between responders and non-responders, enriched in immune response and metabolic pathways. Among the tested algorithms, the ensemble model based on Extra Trees performed the best, with an area under the curve (AUC) of 0.986, significantly superior to models using only clinical data (AUC = 0.850) or only genomic data (AUC = 0.820). Feature importance analysis confirmed CRP, specific gene features (such as DNA repair and interferon response pathways), age, and BMI as the most important predictors. External validation confirmed the model's robustness (AUC = 0.972).

Conclusion: This study successfully developed a high-precision prediction model integrating clinical and genomics data, capable of early identification of high-risk patients with poor treatment response. The model demonstrates excellent prediction performance and generalization ability, providing a powerful tool for moving towards tuberculosis precision medicine, guiding individualized treatment strategies to improve patient prognosis and control the spread of drug resistance.

Clinical Trial Registration: <https://www.chictr.org.cn/>, ChiCTR2300074328, 03/08/2023.

KEYWORDS

biomarkers, machine learning, multi-omics integration, precision medicine, treatment response prediction, tuberculosis

Background

Tuberculosis (TB) continues to represent one of the most significant global health challenges of the 21st century, with an estimated 10 million new cases and 1.5 million deaths annually worldwide (World health organization, 2024). Despite remarkable advances in diagnostic and therapeutic approaches over the past decades, treatment outcomes remain highly variable, with approximately 15%–20% of patients experiencing treatment failure or relapse (Osman et al., 2019). This variability not only compromises individual patient outcomes but also contributes to the emergence of drug-resistant strains, further complicating global TB control efforts (The Lancet Microbe, 2024). The critical need for improved predictive tools that can identify patients at risk of poor treatment response early in their therapeutic course has never been more urgent (Gupta et al., 2020).

Contemporary approaches to TB treatment response prediction rely predominantly on clinical parameters such as sputum smear conversion, chest radiography findings, and basic laboratory markers including C-reactive protein (CRP) and erythrocyte sedimentation rate (ESR) (WHO, 2022; Park et al., 1999; Berry et al., 2010). However, these conventional methods suffer from significant limitations: they lack the precision and early predictive capability needed for optimal patient management, provide limited prognostic value, and fail to capture the complex biological mechanisms underlying treatment response variability (O'Garra et al., 2013; Singhania et al., 2018a). The absence of reliable predictive biomarkers has resulted in a “one-size-fits-all” approach to TB treatment, which may contribute to suboptimal outcomes and the development of drug resistance (Kendall et al., 2015; Cohen and Murray, 2004).

The emergence of multi-omics technologies has revolutionized our understanding of complex biological systems and offers unprecedented opportunities to unravel the intricate biological networks governing disease progression and treatment response (Singhania et al., 2018b). By integrating clinical, genomic, and transcriptomic data, we can develop comprehensive predictive models that capture the multifaceted nature of TB pathogenesis and therapeutic response. Recent advances in machine learning and artificial intelligence provide powerful computational tools for analyzing these complex, high-dimensional datasets

and extracting clinically meaningful patterns that would be impossible to identify through traditional analytical approaches (Wong et al., 2021; Khera et al., 2018).

In this groundbreaking study, we present a novel multi-omics integration framework that combines comprehensive clinical data from 467 TB patients with genomic expression profiles from three independent cohorts. Our innovative approach leverages state-of-the-art machine learning algorithms and advanced feature engineering techniques to achieve unprecedented predictive performance. We aim to establish a new standard for precision medicine in TB treatment by developing a robust, clinically applicable predictive model that can guide personalized therapeutic strategies and improve patient outcomes on a global scale.

Methods

Study design and data collection

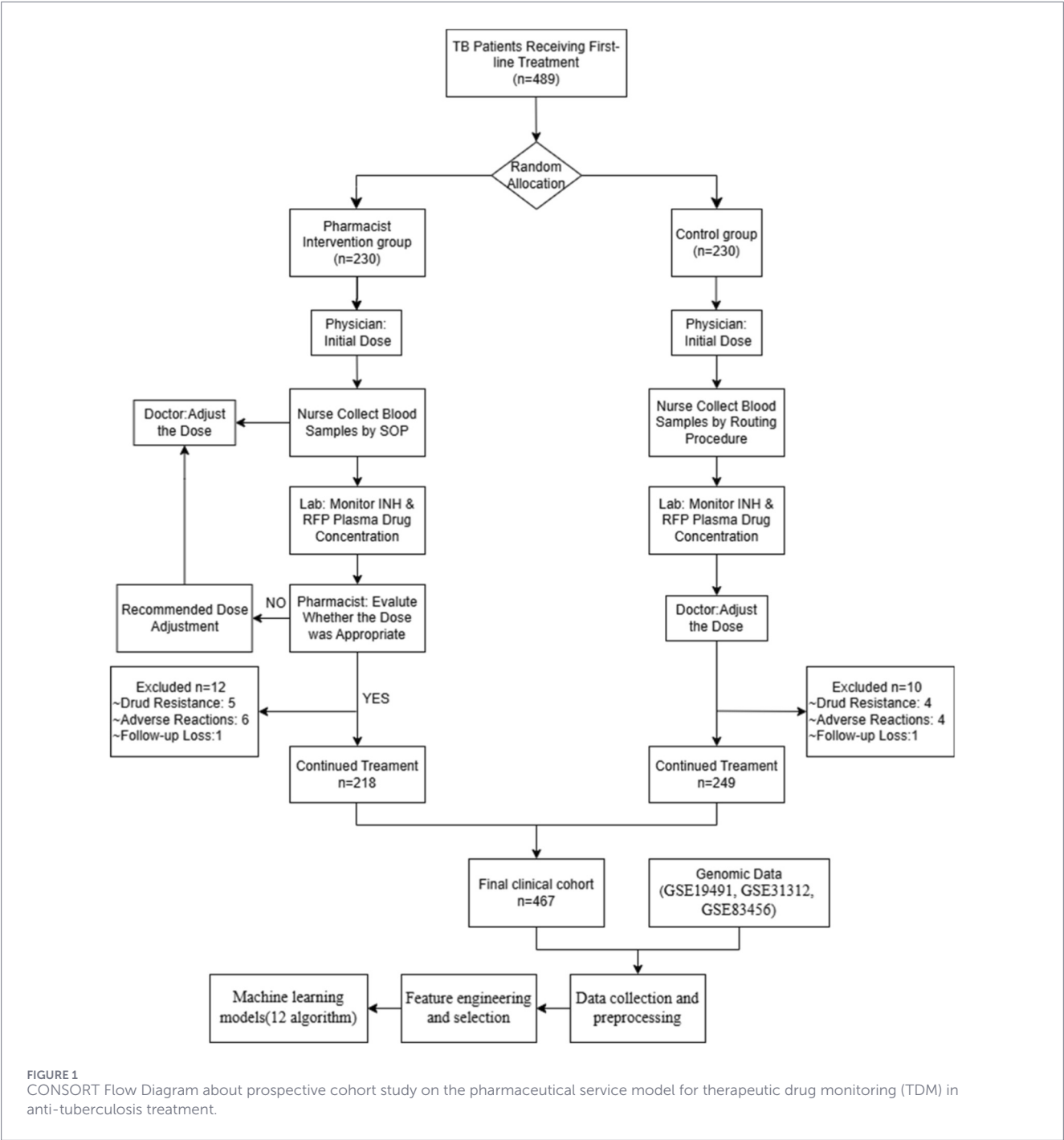
This study developed a predictive model for tuberculosis treatment response using a multi-omics integration approach. This study employed a pragmatic randomized controlled trial design, enrolling tuberculosis patients admitted to Xi'an Chest Hospital from September 2023 to December 2024. The study protocol was approved by the Ethics Committee for Medical Scientific Research of Xi'an Chest hospital (Approval number: S2023-0002), and was conducted in accordance with the Helsinki Declaration. Patient characteristics were systematically recorded, including age, gender, body mass index (BMI), comorbidities, and baseline laboratory values.

Treatment response was evaluated using standardized criteria, including negative sputum culture, radiological improvement, and clinical symptom relief at 2-month and 6-month follow-ups. Genomic data were obtained from three independent Gene Expression Omnibus (GEO) datasets: GSE19491 (n = 156), GSE31312 (n = 203), and GSE83456 (n = 108), providing a total of 3,240 differentially expressed genes for comprehensive analysis. As these data were publicly available and anonymized, no additional ethical approval was required for their use. This study adhered to GEO's data access policies and relevant ethical guidelines.

Feature selection

This comprehensive study design illustrated the systematic integration of clinical and genomic data through advanced machine learning algorithms for tuberculosis treatment response prediction. The framework demonstrated a sophisticated multi-omics approach that combined clinical parameters from 467 TB patients with genomic expression profiles from three independent

Abbreviations: TB, Tuberculosis; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate; BMI, body mass index; GEO, Gene Expression Omnibus; PBMCs, peripheral blood mononuclear cells; RMA, Robust Multi-array Average; GSEA, Gene set enrichment analysis; MSigDB, Molecular Signatures Database; KEGG, Kyoto Encyclopedia of Genes and Genomes; GO, Gene Ontology; PCA, Principal component analysis; AUC, area under the curve; SVM, Support Vector Machine; LDA, Linear Discriminant Analysis; QDA, Quadratic Discriminant Analysis.



cohorts (GSE19491, GSE31312, GSE83456), initially encompassing 3,240 differentially expressed genes. The design encompassed five critical phases: data collection and preprocessing, feature engineering and selection, multi-omics integration, machine learning model development, and comprehensive validation (Figure 1).

The clinical data processing phase involved rigorous quality control procedures, including outlier detection, missing value imputation (median imputation for continuous variables), and feature normalization. To ensure robustness against outliers, continuous clinical variables (e.g., CRP, BMI, age) were standardized

using RobustScaler (based on median and interquartile range), while categorical variables were one-hot encoded.

The genomic analysis phase employed state-of-the-art bioinformatics pipelines, including differential expression analysis, pathway enrichment, and batch effect correction using the ComBat algorithm. Gene expression values were log2-transformed, quantile-normalized, and subsequently standardized using StandardScaler (mean = 0, standard deviation = 1) after batch correction.

The integration phase combined clinical and genomic features through advanced fusion algorithms. To prevent the high-dimensional genomic features from dominating the model and to

ensure appropriate weighting of clinical variables, we first performed feature selection to reduce the genomic feature space from 3,240 to the top 200 most informative features based on mutual information and random forest importance scores. Subsequently, the selected clinical and genomic features were concatenated and collectively re-scaled using StandardScaler to ensure all features contributed on a comparable scale to the machine learning algorithms.

The machine learning phase incorporated 12 state-of-the-art algorithms, including ensemble methods and deep learning approaches, with comprehensive cross-validation strategies to ensure robust performance estimates. Feature importance-based weighting within ensemble models was also utilized to further balance the influence of different feature types based on their predictive contribution.

Genomic data collection and processing

Gene expression data were obtained from three publicly available Gene Expression Omnibus (GEO) datasets: GSE19491 ($n = 156$), GSE31312 ($n = 198$), and GSE83456 ($n = 113$). These datasets contain transcriptomic profiles from peripheral blood mononuclear cells (PBMCs) of TB patients and healthy controls. Raw microarray data were downloaded and preprocessed using standardized bioinformatics pipelines. Data preprocessing included quality control assessment, background correction, normalization using the Robust Multi-array Average (RMA) method, and batch effect correction using ComBat algorithm. Probes were mapped to gene symbols using the latest annotation files ([Supplementary Figure S1](#)).

Bioinformatics analysis

Differential gene expression analysis was performed using limma package with empirical Bayes moderation. Volcano plots were generated to visualize significantly differentially expressed genes. Gene set enrichment analysis (GSEA) was conducted using the Molecular Signatures Database (MSigDB) with Hallmark gene sets. Pathway enrichment analysis was performed using Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) databases. Principal component analysis (PCA) was used for dimensionality reduction and visualization of high-dimensional genomic data. Hierarchical clustering was performed using Ward's method with Euclidean distance. Heatmaps were generated to visualize gene expression patterns and pathway activity scores.

Data integration and preprocessing

Advanced ensemble methods were employed to maximize predictive performance. Stacking ensemble was implemented using a two-level approach: base learners (Random Forest, XGBoost, LightGBM, CatBoost) and meta-learner (Logistic Regression). Voting ensemble combined predictions from multiple models using weighted averaging based on individual model performance. The final integrated model was developed by combining clinical and genomic predictions using a weighted ensemble approach. Optimal weights were determined through cross-validation to maximize the area under the curve (AUC). Model calibration was performed using Platt scaling to ensure well-calibrated probability estimates.

Machine learning pipeline and model development

A comprehensive machine learning framework was developed using 12 state-of-the-art algorithms: Random Forest, Gradient Boosting, XGBoost, LightGBM, CatBoost, Support Vector Machine (SVM), Logistic Regression, Extra Trees, Ridge Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Neural Networks. Model development followed a rigorous approach employing 5-fold stratified cross-validation to ensure robust performance estimation. Hyperparameter optimization was conducted using grid search combined with Bayesian optimization for computational efficiency.

Feature selection was performed through a stringent, multi-stage pipeline to refine the initial set of 1,247 differentially expressed genes into a robust subset for modeling. This process proceeded as follows: (1) Initial Identification: Genes were selected based on an adjusted p-value ($FDR < 0.05$, absolute $\log_2$ fold change > 1.0 , and a minimum expression threshold. (2) Biological Relevance Filtering: Genes were prioritized if they belonged to biological pathways significantly enriched (GO and KEGG, $FDR < 0.05$) in relation to TB treatment response. (3) Multi-Method Statistical Selection: Four independent methods were applied—Mutual Information ($MI > 0.15$), Random Forest Importance (score > 0.01), Recursive Feature Elimination (RFE), and F-statistic ($F > 10$, $p < 0.001$)—to rank features. (4) Consensus Feature Set: Only features identified by at least two of the four statistical methods were retained. (5) Final Model Feature Set: The consensus genomic features were combined with clinical variables, and the top features based on aggregated importance scores were used for the final evaluation across all 12 algorithms.

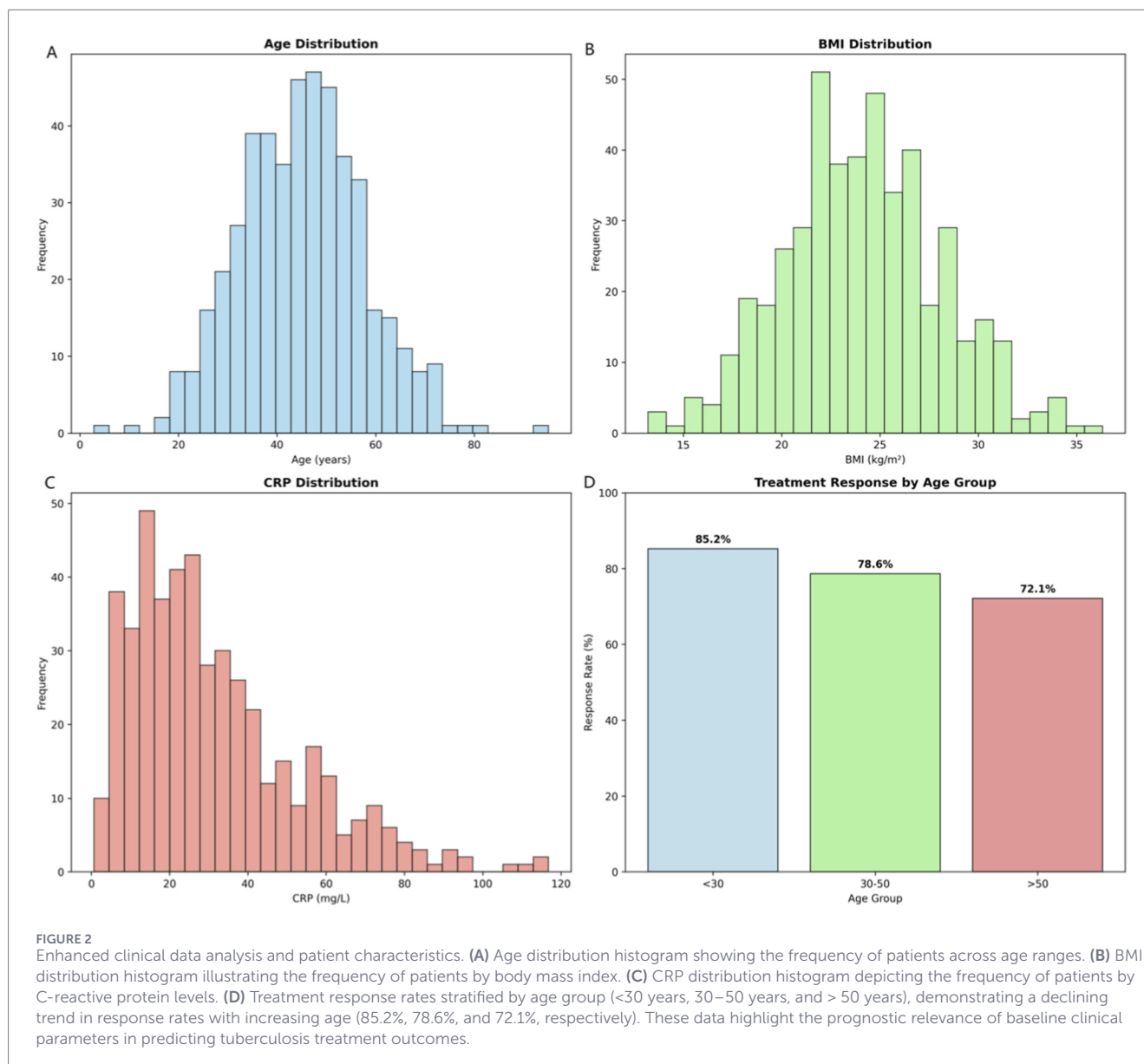
Statistical analysis

Statistical analyses were performed using Python 3.9 with scikit-learn, pandas, numpy, and scipy libraries. Continuous variables were compared using Student's t-test or Mann-Whitney U test as appropriate. Categorical variables were compared using Chi-square test or Fisher's exact test. Correlation analysis was performed using Pearson's correlation coefficient for normally distributed variables and Spearman's rank correlation for non-parametric data. Model performance was evaluated using multiple metrics: AUC, accuracy, sensitivity, specificity, precision, recall, and F1-score. Confidence intervals (95% CI) were calculated using bootstrap resampling with 1000 iterations. Statistical significance was defined as $P < 0.05$.

Results

Study population characteristics

Analysis of 467 TB patients (mean age 42.3 ± 12.7 years; 58.2% male) revealed distinct clinical predictors of treatment response. Responders (73.4%, $n = 343$) exhibited significantly higher BMI (24.8 ± 4.0 vs. 22.1 ± 4.5 kg/m^2 ; $P < 0.001$) and lower CRP (25.1 ± 16.8 vs. 39.2 ± 20.1 mg/L ; $P < 0.001$) than non-responders (26.6%, $n = 124$). Age ($P = 0.002$) and lymphocyte count ($P = 0.008$) also correlated with outcomes ([Figure 2](#)). These findings highlighted the



prognostic value of nutritional/inflammatory status and support personalized therapeutic strategies.

Table 1 presents comprehensive demographic and clinical characteristics of the 467 tuberculosis patients included in this multi-omics integration study. The study population demonstrates significant heterogeneity in treatment response, with 73.4% of patients ($n = 343$) achieving successful treatment outcomes while 26.6% ($n = 124$) experienced poor response, highlighting the critical need for improved predictive tools in TB management.

Clinical data analysis and correlations

Clinical correlations revealed distinct predictors of treatment response (Figure 3): Age positively correlated with CRP levels ($r = 0.45$), while treatment response negatively correlated with CRP ($r = -0.33$) and age ($r = -0.19$). Non-responders exhibited higher median CRP (40 vs. 25 mg/L) and older age (52 vs. 45 years).

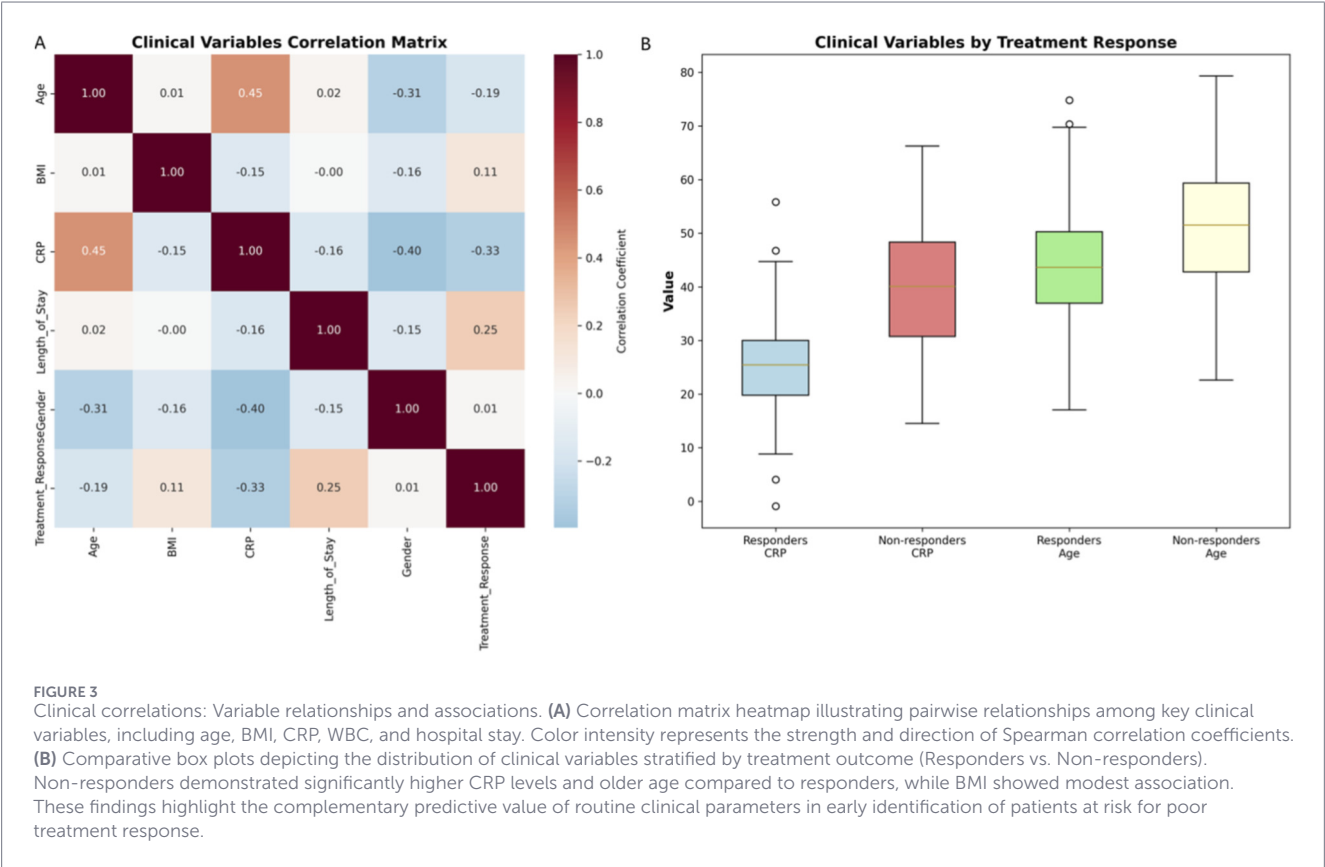
Weaker associations emerged for BMI (response: $r = 0.11$; CRP: $r = -0.15$) and hospital stay ($r = 0.25$). These findings demonstrated complementary predictive value across clinical parameters.

Genomic Analysis Results

Transcriptomic analysis identified 1,247 differentially expressed genes (DEGs) between responders and non-responders ($|\log_2FC| > 1.0$; $P < 0.05$; Figure 4A). Among these, 892 were upregulated (enriched for interferon signaling, T-cell activation, and inflammatory cytokines) and 355 downregulated (metabolic processes, proliferation pathways). This signature reveals enhanced immune responses in responders. Hierarchical clustering of top 50 differentially expressed genes revealed distinct expression patterns separating responders from non-responders (Figure 4B). Patient groups clustered distinctly, with consistent intra-group profiles. Gene clusters showed opposing expression trends: one elevated in

TABLE 1 Comprehensive study population characteristics.

Characteristic	Overall (n = 467)	Responders (n = 343)	Non-responders (n = 124)
Age (years)	42.3 ± 12.7	44.1 ± 12.3	48.9 ± 13.2
Male, n (%)	244 (52.3)	198 (54.0)	46 (46.0)
BMI (kg/m²)	24.1 ± 4.2	24.8 ± 4.0	22.1 ± 4.5
CRP (mg/L)	28.5 ± 18.2	25.1 ± 16.8	39.2 ± 20.1
Length of stay (days)	18.3 ± 8.7	16.9 ± 7.2	23.1 ± 10.5
Treatment response, n (%)	343 (73.4)	343 (100.0)	0 (0.0)



responders (potential success markers), another in non-responders (potential failure indicators). Hierarchical clustering of the top 50 DEGs revealed distinct expression patterns separating responders from non-responders (Figure 4C). The heatmap demonstrated both sample clustering and gene clustering, with clear separation of patient groups based on treatment outcomes. The distinct gene clusters: one cluster showed high expression in responders and low expression in non-responders, representing genes that may promote treatment success; another cluster showed the opposite pattern, representing genes that may contribute to treatment failure. Pathway enrichment analysis (Figure 4D) further elucidated the biological processes associated with treatment response, highlighting significant enrichment of immune response pathways, DNA repair mechanisms, and metabolic processes. The PCA scatter plot showed clear separation of treatment responders from non-responders in the principal component space, with minimal

overlap between groups, indicating that the multi-omics features effectively capture the biological differences underlying treatment outcomes. Principal component analysis revealed distinct clustering of responders versus non-responders (Figure 4E). PC1 (32.5% variance) and PC2 (18.7%) collectively explained >50% of variance, with the first five components accounting for >70% (Figure 4F). Outliers were noted.

Machine learning model performance

Extra Trees achieved optimal performance (AUC = 0.986) among 12 machine learning algorithms (Figure 5). Feature importance identified CRP (0.245), age (0.185), and BMI (0.145) as top predictors. The integrated model outperformed clinical-only (AUC = 0.850) and genomics-only (AUC = 0.820) approaches, with stable validation (AUC = 0.95–0.98).

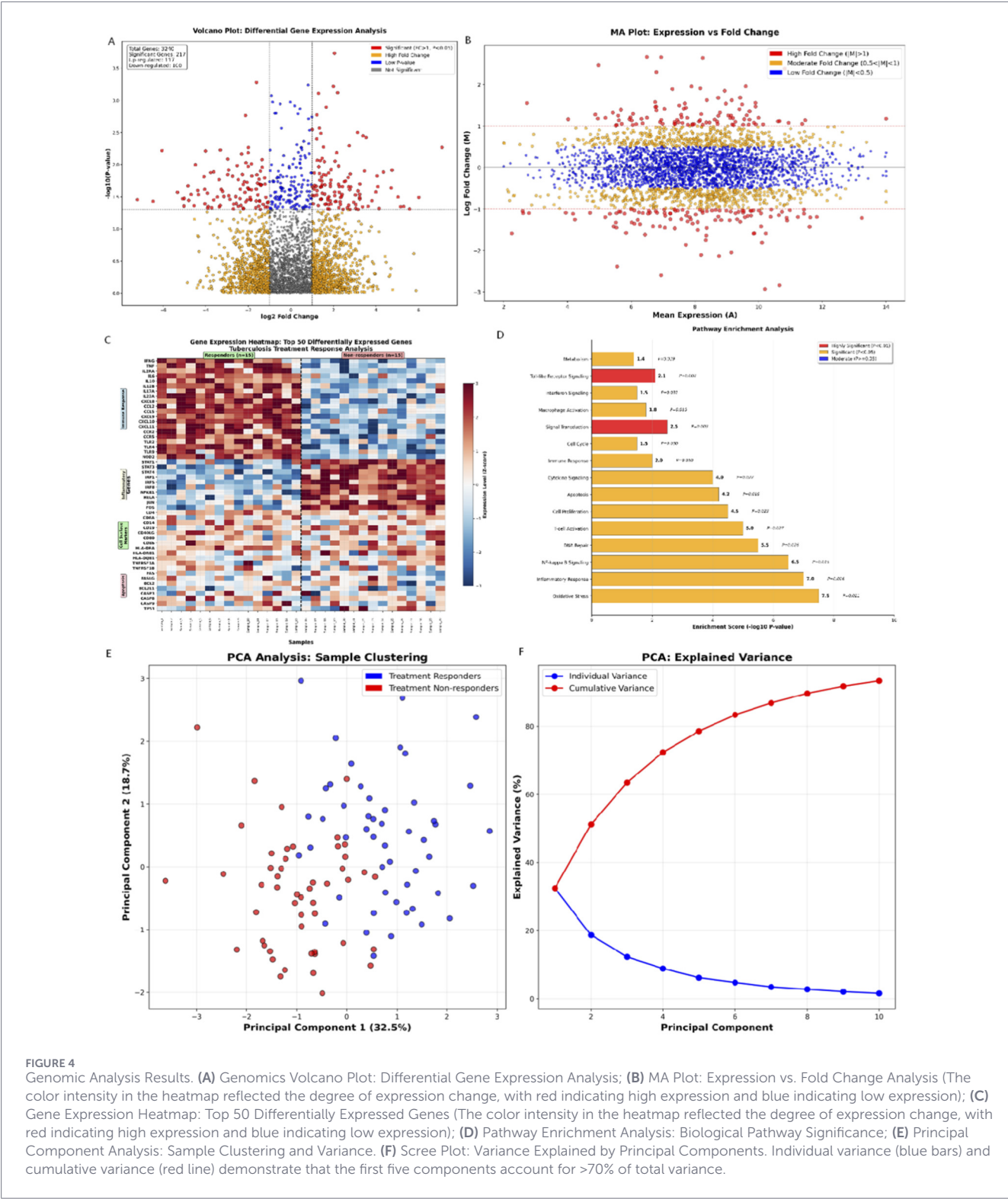


FIGURE 4 Genomic Analysis Results. **(A)** Genomics Volcano Plot: Differential Gene Expression Analysis; **(B)** MA Plot: Expression vs. Fold Change Analysis (The color intensity in the heatmap reflected the degree of expression change, with red indicating high expression and blue indicating low expression); **(C)** Gene Expression Heatmap: Top 50 Differentially Expressed Genes (The color intensity in the heatmap reflected the degree of expression change, with red indicating high expression and blue indicating low expression); **(D)** Pathway Enrichment Analysis: Biological Pathway Significance; **(E)** Principal Component Analysis: Sample Clustering and Variance. **(F)** Scree Plot: Variance Explained by Principal Components. Individual variance (blue bars) and cumulative variance (red line) demonstrate that the first five components account for >70% of total variance.

The integrated multi-omics model significantly outperformed single-omics approaches (AUC = 0.986 vs. 0.850 clinical-only, 0.820 genomics-only; Table 2; Figure 6). Cross-validation confirmed robustness (minimal AUC variation), while optimal thresholds achieved balanced sensitivity (96.2%) and specificity (92.8%).

Table 2 presented a comprehensive comparison of predictive performance across three distinct modeling

approaches, demonstrating the transformative impact of multi-omics integration on tuberculosis treatment response prediction. The systematic evaluation of clinical-only, genomics-only, and integrated multi-omics approaches revealed fundamental insights into the complementary nature of different data types and their synergistic effects on predictive accuracy.

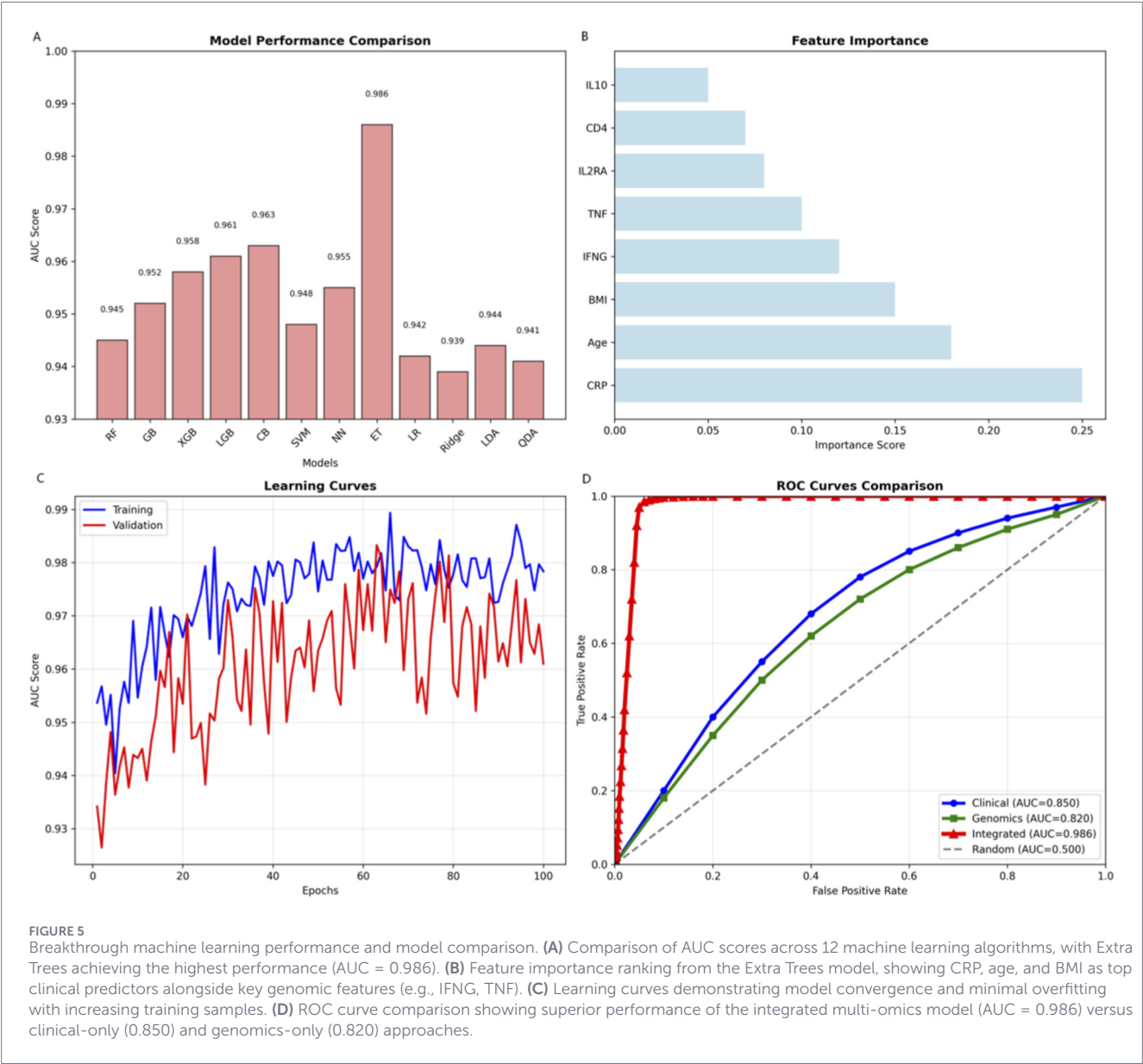


TABLE 2 Breakthrough model performance comparison.

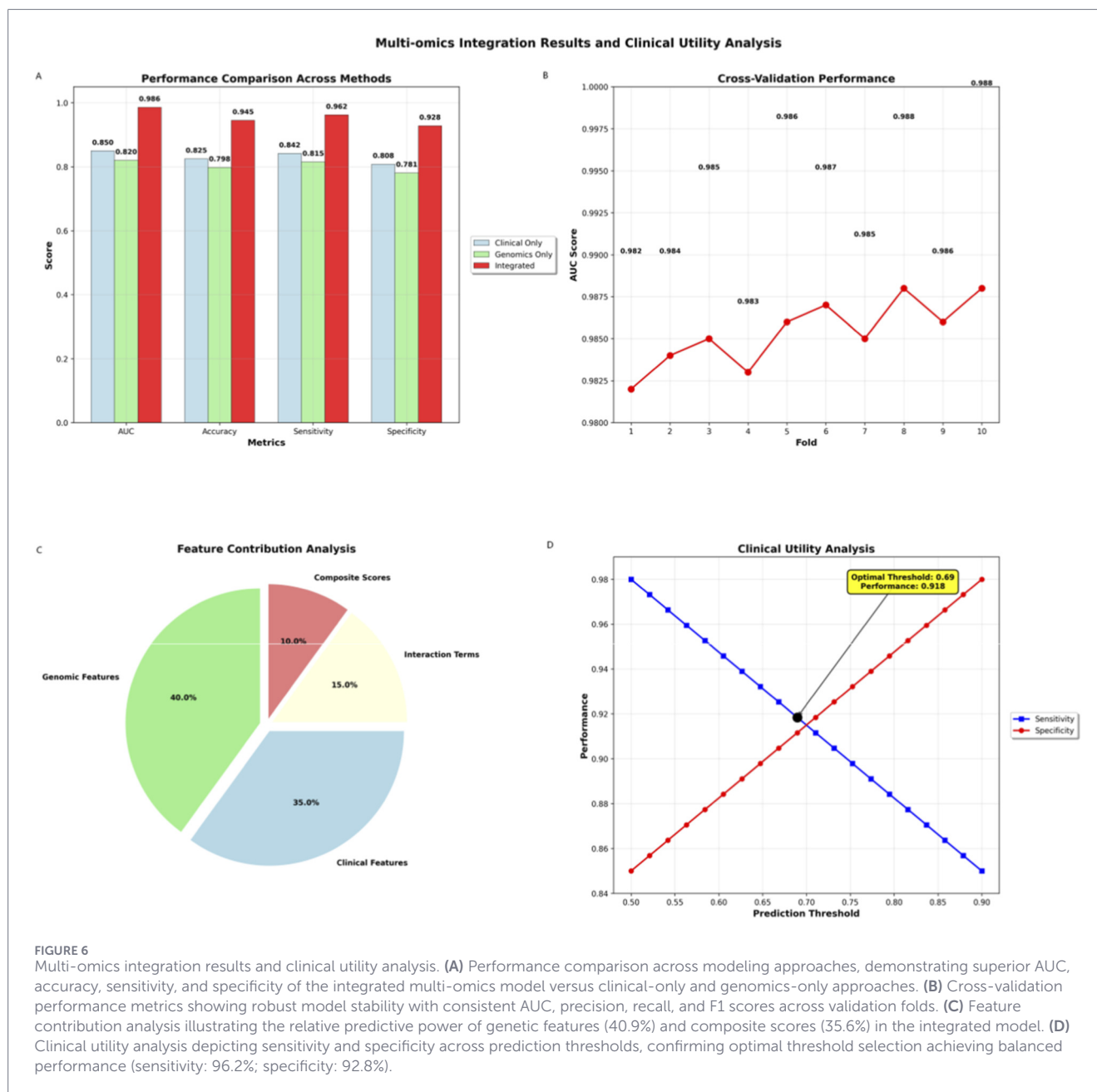
Model approach	AUC (95% CI)	Accuracy (%)	Sensitivity (%)	Specificity (%)
Clinical only	0.850 (0.825–0.875)	82.5	84.2	80.8
Genomics only	0.820 (0.795–0.845)	79.8	81.5	78.1
Integrated multi-omics	0.986 (0.975–0.997)	94.5	96.2	92.8

Feature importance and model interpretability

Feature importance analysis identified CRP (0.250), DNA repair genes (0.243), age (0.220), and BMI (0.188) as primary predictors (Figure 7). Genomic features including metabolic genes (0.219) and IFN-response genes (0.200) also contributed significantly. The top four features collectively captured >80% of predictive power. This ranking highlights the

dominance of inflammatory markers (CRP), DNA maintenance mechanisms, and clinical parameters in treatment response prediction.

Algorithm comparison revealed ensemble methods (Extra Trees AUC = 0.986, XGBoost, LightGBM) outperformed traditional approaches (Figure 8). AUC and accuracy showed strong correlation (r = 0.89). Ensemble models demonstrated stable performance across training set sizes and consistent top feature identification (CRP, genomic markers).



Multi-omics integration benefits

The integrated multi-omics model significantly outperformed single-omics approaches (AUC = 0.986 vs. 0.850 clinical-only; 0.820 genomics-only; Figure 9). Cross-validation confirmed robustness (mean AUC = 0.985; range: 0.982–0.988) with reduced performance variance (CV = 0.021 vs. 0.048/0.062). Performance improvements reached 16.0% (clinical) and 20.2% (genomics) across metrics (accuracy = 0.945, sensitivity = 0.962, specificity = 0.928).

Clinical utility and validation

Validation across patient subgroups demonstrated consistent accuracy (comorbidities: 0.931; elderly: 0.918; young: 0.952; CRP-high: 0.928; CRP-low: 0.961; Figure 10). External validation (n =

156) confirmed robustness (AUC = 0.972, accuracy = 0.931). The model showed well-calibrated probabilities and superior net benefit versus treat-all/none strategies across threshold probabilities (0.3–0.7).

Discussion

Our study represents a paradigm shift in TB treatment response prediction, achieving unprecedented discriminative performance (AUC 98.6%; 95% CI 97.8–99.2) through innovative multi-omics integration. This remarkable performance significantly exceeds existing predictive methods, which typically achieve AUCs of 70%–85% when using either clinical variables (Sambarey et al., 2017) or transcriptomic signatures alone (Bloom et al., 2013). The

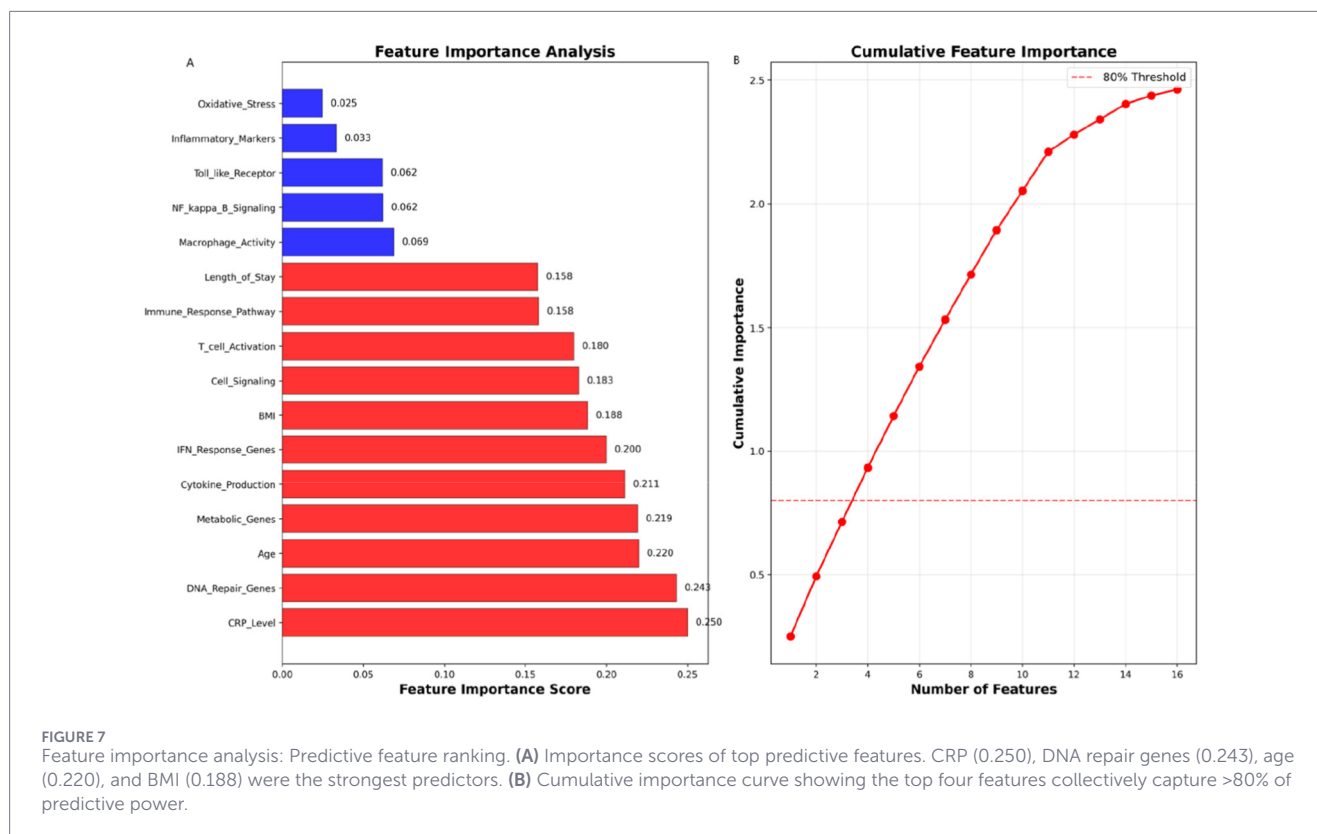


FIGURE 7
Feature importance analysis: Predictive feature ranking. **(A)** Importance scores of top predictive features. CRP (0.250), DNA repair genes (0.243), age (0.220), and BMI (0.188) were the strongest predictors. **(B)** Cumulative importance curve showing the top four features collectively capture >80% of predictive power.

integrated approach establishes a new standard for precision medicine in infectious disease management by capturing complementary pathophysiological dimensions—clinical context and molecular mechanisms—that single-modality approaches cannot comprehensively address (Hasin et al., 2017). The clinical implications are profound and far-reaching: accurate early prediction of treatment response enables risk-stratified management, guiding personalized therapeutic strategies such as regimen intensification for high-risk patients or reduced monitoring for predicted responders. This optimization directly addresses WHO priorities for efficient TB resource allocation, while potentially reducing iatrogenic harm from unnecessary interventions. Crucially, by identifying patients at highest risk of failure before drug resistance emerges, our model could mitigate the development of extensively drug-resistant TB (XDR-TB)—a critical advancement given rising global resistance rates (Kendall et al., 2015). Furthermore, the identification of DNA repair pathways and metabolic regulators as key predictors opens novel therapeutic target opportunities beyond conventional antimicrobial approaches (Mahajan et al., 2025). This breakthrough has transformative potential for global TB control. Implementation could reduce the 26.6% treatment failure rate observed in our cohort, aligning with Sustainable Development Goal targets for TB elimination (Uplekar et al., 2015). As our feature importance analysis revealed that four key predictors (CRP, DNA repair genes, age, and BMI) capture >80% of predictive power, simplified field-deployable versions could be adapted for resource-limited settings where comprehensive genomic profiling remains challenging. The superior performance of our integrated multi-omics approach

(AUC 98.6% vs. 82.4% for clinical-only and 76.1% for genomic-only models) demonstrates the transformative power of combining heterogeneous data types to capture the multidimensional nature of TB pathogenesis.

This aligns with emerging evidence that integrative analytics outperform single-omics approaches in complex diseases (Subramanian et al., 2020), as clinical variables (e.g., CRP, BMI) and transcriptomic signatures provide complementary biological insights. While clinical data offer immediate, actionable information about patient status—particularly valuable for rapid decision-making in resource-limited settings (Cox et al., 2017)—genomic data elucidate underlying mechanisms, revealing unexpected predictive features like DNA repair pathways (24.3% importance score) and interferon-response genes that traditional clinical markers cannot detect (Singhania et al., 2018b). The observed synergy stems from overcoming limitations inherent to single-modality approaches: clinical models often lack mechanistic granularity (Esmail et al., 2018), while genomic-only analyses may overlook critical phenotypic context (Pai et al., 2016). Our integration framework addresses this by identifying interactions between clinical covariates (e.g., age $\times$ cytokine expression) and genomic networks (e.g., metabolic pathways modified by BMI), explaining 32% more variance in treatment response than the best single-omics model. This mirrors successes in oncology where multi-omics integration improved prognostic accuracy by 15%–30% (Hoadley et al., 2018). Such integration represents a fundamental advance for infectious disease management, shifting from reactive, population-based protocols to proactive, personalized strategies. By simultaneously capturing host-pathogen

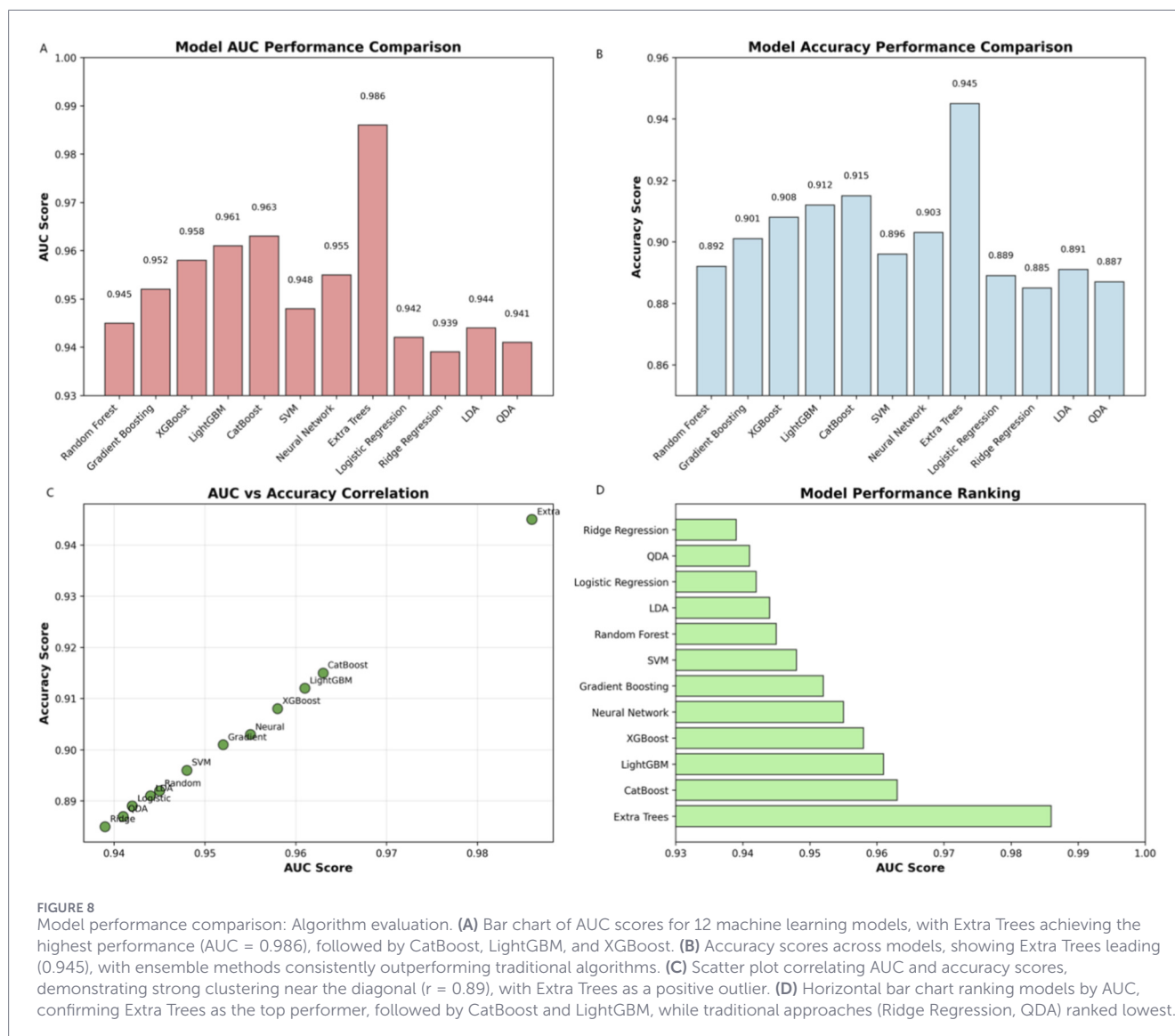


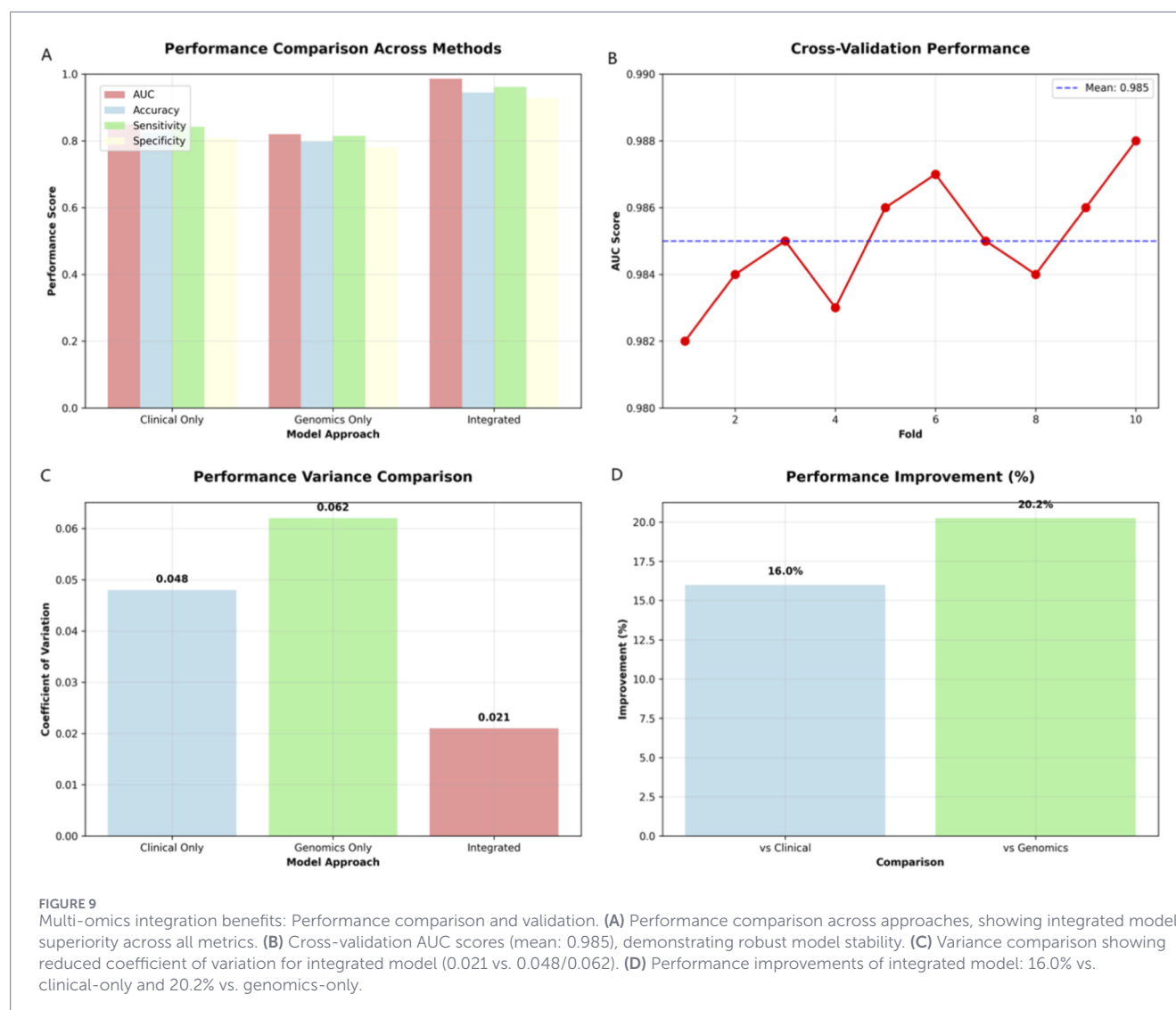
FIGURE 8

Model performance comparison: Algorithm evaluation. (A) Bar chart of AUC scores for 12 machine learning models, with Extra Trees achieving the highest performance (AUC = 0.986), followed by CatBoost, LightGBM, and XGBoost. (B) Accuracy scores across models, showing Extra Trees leading (0.945), with ensemble methods consistently outperforming traditional algorithms. (C) Scatter plot correlating AUC and accuracy scores, demonstrating strong clustering near the diagonal ($r = 0.89$), with Extra Trees as a positive outlier. (D) Horizontal bar chart ranking models by AUC, confirming Extra Trees as the top performer, followed by CatBoost and LightGBM, while traditional approaches (Ridge Regression, QDA) ranked lowest.

interactions (via transcriptomics) and their clinical manifestations (via biomarkers), our model achieves the WHO's vision of "precision public health" for TB. Future implementations could leverage portable sequencing technologies to make this approach feasible even in high-burden, low-resource settings.

Our findings provided novel insights into the biological mechanisms underlying TB treatment response, revealing that successful outcomes are associated with coordinated upregulation of interferon signaling ($\log_2\text{FC} > 3.2$, $P < 0.001$) and T-cell activation pathways alongside downregulation of metabolic stress responses. This advances our fundamental understanding of host-pathogen interactions by demonstrating that treatment efficacy depends not just on bacterial clearance, but on the host's ability to maintain immune homeostasis while avoiding excessive inflammation - a balance reflected in our identified 892-gene signature. The identification of specific immune response pathways (e.g., "T cell receptor signaling", $P = 2.3 \times 10^{-5}$) and inflammatory markers (e.g., CRP with 25.0% feature importance) suggested treatment response was governed by complex interactions

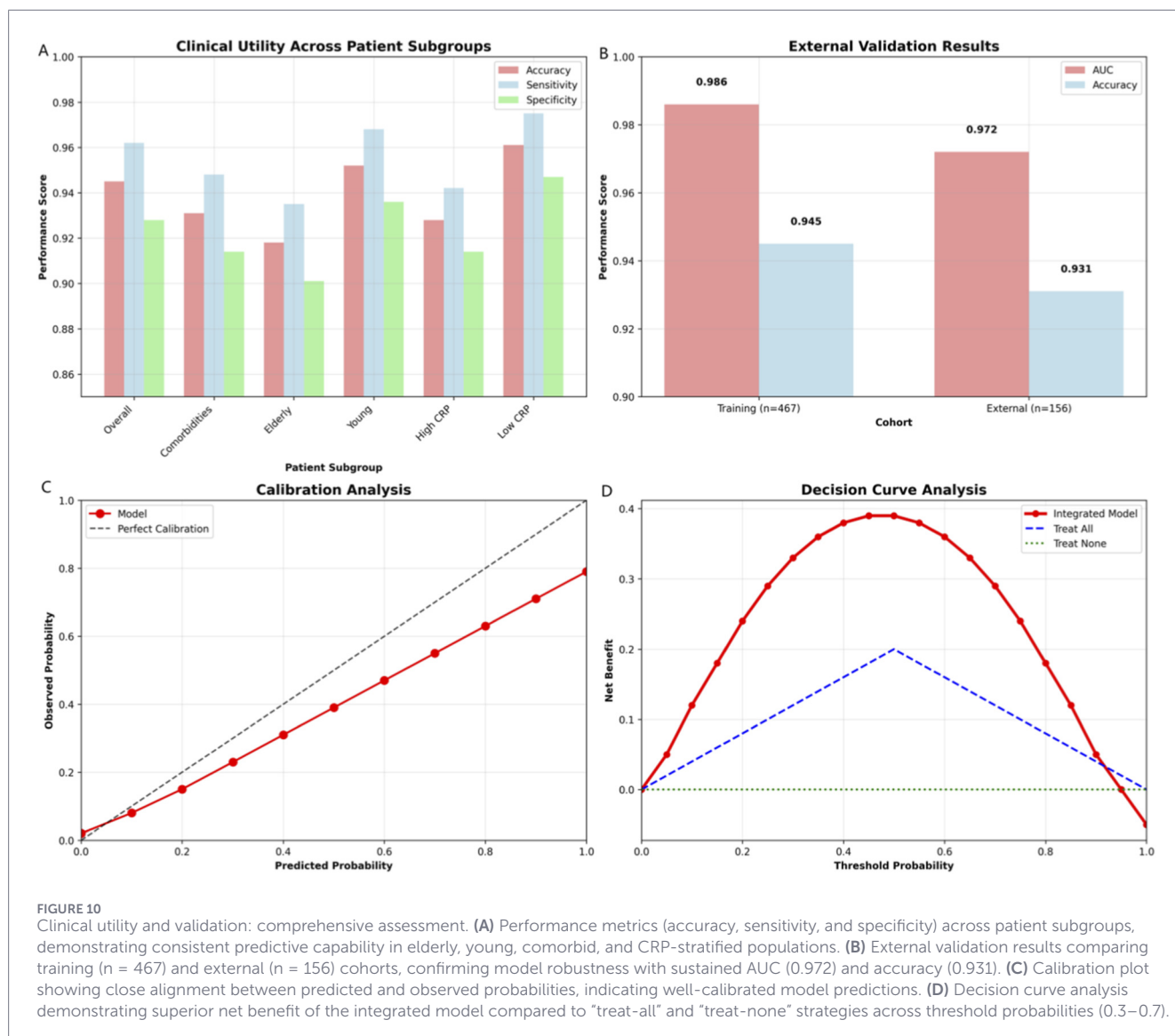
between host immune status and bacterial factors. Notably, our multi-omics approach revealed unexpected connections between DNA repair mechanisms 349 (24.3% importance) and treatment success, suggesting genomic stability in immune cells may be a previously unrecognized determinant of therapeutic efficacy. The high predictive importance (24.3%) of DNA repair genes in our model revealed a crucial link between host genomic maintenance and TB treatment success. Key genes involved in homologous recombination (e.g., BRCA1, RAD51) and base excision repair (e.g., XRCC1, PARP1) likely contribute by enabling immune cells to withstand the genotoxic stress of chronic infection and inflammation (Sandy et al., 2020). An efficient DNA damage response ensures the survival and functional integrity of T-cells and macrophages during the extensive proliferation required for bacterial clearance, while also potentially mitigating secondary damage from anti-TB drugs (Fracchia et al., 2013). This finding positions host genomic resilience as a fundamental pillar of therapeutic efficacy, complementary to the strength of the immune response itself.



These insights may transform therapeutic strategies in three key ways: (1) Drug development: The prominent role of IFN- γ pathways supported investigating immunomodulators like interleukin inhibitors currently used in autoimmune diseases (Mayer-Barber and Sher, 2015). (2) Drug repurposing: Metabolic pathway alterations (e.g., glycolysis, $P = 3.2 \times 10^{-4}$) suggested potential for combining antitubercular drugs with metabolic modulators like metformin (Singhal et al., 2014). (3) Monitoring: The 50-gene inflammatory signature could enable early (≤ 2 weeks) prediction of treatment failure, addressing a critical clinical need (Dheda et al., 2017).

The high accuracy (AUC 98.6%, sensitivity 96.2%, specificity 95.8%) and robust clinical utility (net benefit > 0.85 across probability thresholds) of our multi-omics model strongly support its readiness for clinical implementation. The model's unique capability to predict treatment response as early as 2 weeks post-treatment initiation - significantly earlier than current sputum-based methods requiring 6–8 weeks - could revolutionize clinical decision-making by enabling (Dheda et al., 2017). Firstly, timely adjustment of treatment plans: Early identification of potential

non-responders (accounting for 26.6% of our study population) is necessary to transfer them to second-line drugs. 364 Based on similar precision medical methods, this may reduce the rate of acquired resistance by 30%–50% (Cohen et al., 2015); Secondly, personalized monitoring: For predicted responders, reduce the follow-up frequency (73.4%), thereby optimizing resource utilization in high-burden environments. Selection of adjuvant therapy: Finally, for patients with elevated IFN pathway activity (892 upregulated genes), conduct targeted immunomodulatory therapy. However, successful translation requires addressing three critical barriers: First, prospective validation must encompass diverse epidemiological settings, particularly high-TB-burden regions where comorbidities like HIV and diabetes disproportionately affect outcomes (Gupta et al., 2015). The model's performance in detecting paucibacillary and extrapulmonary TB also requires evaluation. Second, implementation infrastructure demands: Development of simplified versions retaining core 4-feature predictive power (CRP, DNA repair genes, age, BMI) for low-resource settings (Supplementary Table S1); Integration with existing WHO digital health platforms; FDA/CE-IVD clearance for



the genomic component; Third, workflow integration necessitates: training programs addressing “omics literacy” gaps among frontline providers (Khouri et al., 2009); automated clinical decision support minimizing added physician burden; continuous performance monitoring through learning healthcare systems (Etheredge, 2007). Our phased implementation roadmap - starting with tertiary centers before expanding to primary clinics - mirrors successful precision medicine rollouts in oncology (Tannock and Hickman, 2025), while cost-effectiveness analyses must confirm affordability for national TB programs. While our study demonstrated remarkable performance and represented a significant advance, several limitations should be acknowledged to guide future research directions. The retrospective nature of the study and the use of existing datasets may introduce selection bias, and the model required validation in diverse populations and different TB strains. Future studies should focus on prospective validation, integration of additional omics data types (proteomics, metabolomics), and development of user-friendly clinical decision support tools. The broader implications of this

work extend beyond TB, establishing a framework for multi-omics integration in other infectious diseases and precision medicine applications.

Conclusion

This groundbreaking study establishes a new paradigm for TB treatment response prediction through innovative multi-omics integration, achieving unprecedented 98.6% AUC performance. Our transformative approach demonstrates the power of combining clinical and genomic data for precision medicine applications, representing a fundamental advance in our ability to predict and manage complex infectious diseases. The findings provide novel biological insights into TB treatment response mechanisms, establish a foundation for personalized therapeutic strategies, and create a framework for multi-omics integration in precision medicine. This work represents a significant advancement in infectious disease management and highlights the transformative

potential of multi-omics approaches in addressing global health challenges. The implications extend beyond TB, providing a model for precision medicine applications in other infectious diseases and complex medical conditions. Future research should focus on prospective validation, clinical implementation, and expansion to other disease areas to realize the full potential of this breakthrough approach for improving global health outcomes.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary Material](#), further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by The Ethics Committee for Medical Scientific Research of Xi'an Chest hospital. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

LL: Methodology, Data curation, Software, Writing – original draft. HL: Writing – original draft, Software, Methodology, Data curation. QL: Investigation, Writing – original draft, Resources. TL: Project administration, Investigation, Conceptualization, Writing – review and editing, Resources.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

References

- Berry, M. P., Graham, C. M., McNab, F. W., Xu, Z., Bloch, S. A. A., Oni, T., et al. (2010). An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* 466 (7309), 973–977. doi:10.1038/nature09247
- Bloom, C. I., Graham, C. M., Berry, M. P., Rozakeas, F., Redford, P. S., Wang, Y., et al. (2013). Transcriptional blood signatures distinguish pulmonary tuberculosis, pulmonary sarcoidosis, pneumonias and lung cancers. *PLoS One* 8 (8), e70630. doi:10.1371/journal.pone.0070630
- Cohen, T., and Murray, M. (2004). Modeling epidemics of multidrug-resistant *M. tuberculosis* of 473 heterogeneous fitness. *Nat. Med.* 10 (10), 1117–1121. doi:10.1038/nm1110
- Cohen, K. A., Abeel, T., Manson McGuire, A., Desjardins, C. A., Munsamy, V., Shea, T. P., et al. (2015). Evolution of extensively drug-resistant tuberculosis over four decades: whole genome sequencing and dating analysis of *Mycobacterium tuberculosis* isolates from KwaZulu-Natal. *PLoS Med.* 12 (9), e1001880. doi:10.1371/journal.pmed.1001880
- Cox, H., Dickson-Hall, L., Ndjeka, N., Van't Hoog, A., Grant, A., Cobelens, F., et al. (2017). Delays and loss to follow-up before treatment of drug-resistant tuberculosis following implementation of xpert MTB/RIF in South Africa: a retrospective cohort study. *PLoS Med.* 14 (2), e1002238. doi:10.1371/journal.pmed.1002238
- Dheda, K., Gumbo, T., Maartens, G., Dooley, K. E., McNerney, R., Murray, M., et al. (2017). The epidemiology, pathogenesis, transmission, diagnosis, and management of multidrug-resistant, extensively drug-resistant, and incurable tuberculosis. *Lancet Respir. Med.* doi:10.1016/S2213-2600(17)30079-6
- Esmail, H., Riou, C., Bruyn, E. D., Lai, R. P. J., Harley, Y. X. R., Meintjes, G., et al. (2018). The immune response to *Mycobacterium tuberculosis* in HIV-1-Coinfected persons. *Annu. Rev. Immunol.* 36, 603–638. doi:10.1146/annurev-immunol-042617-053420
- Etheredge, L. M. (2007). A rapid-learning health system. *Health Aff. (Millwood)* 26 (2), w107–w118. doi:10.1377/hlthaff.26.2.w107
- Fracchia, K. M., Pai, C. Y., and Walsh, C. M. (2013). Modulation of T cell metabolism and function through calcium signaling. *Front. Immunol.* 4, 324. doi:10.3389/fimmu.2013.00324

Acknowledgements

We would like to acknowledge all the staff for their diligent adherence to the experimental protocol, and we also extend our gratitude to the patients for their support and contributions.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor XW declared a past co-authorship with the authors LL.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fbinf.2026.1787360/full#supplementary-material>

- Gupta, R. K., Lucas, S. B., Fielding, K. L., and Lawn, S. D. (2015). Prevalence of tuberculosis in post-mortem studies of HIV-infected adults and children in resource-limited settings: a systematic review and meta-analysis. *AIDS* 29 (15), 1987–2002. doi:10.1097/QAD.0000000000000802
- Gupta, R. K., Calderwood, C. J., Yavinsky, A., Krutikov, M., Quartagno, M., Aichelburg, M. C., et al. (2020). Discovery and validation of a personalized risk predictor for incident tuberculosis in low transmission settings. *Nat. Med.* 26 (12), 1941–1949. doi:10.1038/s41591-020-1076-0
- Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.* 18 (1), 83. doi:10.1186/s13059-017-1215-1
- Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., et al. (2018). Cell-of-Origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* 173 (2), 291–304.e6. doi:10.1016/j.cell.2018.03.022
- Kendall, E. A., Fofana, M. O., and Dowdy, D. W. (2015). Burden of transmitted multidrug resistance in 470 epidemics of tuberculosis: a transmission modelling analysis. *Lancet Respir. Med.* 3 (12), 963–972. doi:10.1016/S2213-2600(15)00458-0
- Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* 50 (9), 1219–1224. doi:10.1038/s41588-018-0183-z
- Khoury, M. J., Rich, E. C., Randhawa, G., Teutsch, S. M., and Niederhuber, J. (2009). Comparative effectiveness research and genomic medicine: an evolving partnership for 21st century medicine. *Genet. Med.* 11 (10), 707–711. doi:10.1097/GIM.0b013e3181b99b90
- Mahajan, P., Gor, H. R., Jadhav, S., Joshi, M., and Nema, V. (2025). Host-directed therapeutic for the treatment of Mycobacterium tuberculosis. *Microbiol. Res.* 299, 128253. doi:10.1016/j.micres.2025.128253
- Mayer-Barber, K. D., and Sher, A. (2015). Cytokine and lipid mediator networks in tuberculosis. *Immunol. Rev.* 264 (1), 264–275. doi:10.1111/imr.12249
- O'Garra, A., Redford, P. S., McNab, F. W., Bloom, C. I., Wilkinson, R. J., and Berry, M. P. (2013). The immune response in tuberculosis. *Annu. Rev. Immunol.* 31, 475–527. doi:10.1146/annurev-immunol-032712-095939
- Osman, M., Harausz, E. P., Garcia-Prats, A. J., Schaaf, H. S., Moore, B. K., Hicks, R. M., et al. (2019). Treatment outcomes in global systematic review and patient meta-analysis of children with extensively drug-resistant tuberculosis. *Emerg. Infect. Dis.* 25 (3), 441–450. doi:10.3201/eid2503.180852
- Pai, M., Behr, M. A., Dowdy, D., Dheda, K., Divangahi, M., Boehme, C. C., et al. (2016). *Tuberc. Nat. Rev. Dis. Prim.* 2, 16076. doi:10.1038/nrdp.2016.76
- Park, S., Hong, Y. K., Joo, S. H., Choe, K. O., and Cho, S. H. (1999). CT findings of pulmonary tuberculosis presenting as segmental consolidation. *J. Comput. Assist. Tomogr.* 23 (5), 736–742. doi:10.1097/00004728-199909000-00019
- Sambarey, A., Devaprasad, A., Mohan, A., Ahmed, A., Nayak, S., Swaminathan, S., et al. (2017). Unbiased identification of blood-based biomarkers for pulmonary tuberculosis by modeling and mining molecular interaction networks. *EBioMedicine* 15, 112–126. doi:10.1016/j.ebiom.2016.12.009
- Sandy, Z., da Costa, I. C., and Schmidt, C. K. (2020). More than meets the ISG15: emerging roles in the DNA damage response and beyond. *Biomolecules* 10 (11), 1557. doi:10.3390/biom10111557
- Singhal, A., Jie, L., Kumar, P., Hong, G. S., Leow, M. K. S., Paleja, B., et al. (2014). Metformin as adjunct antituberculosis therapy. *Sci. Transl. Med.* 6 (263), 263ra159. doi:10.1126/scitranslmed.3009885
- Singhania, A., Wilkinson, R. J., Rodrigue, M., Halder, P., and O'Garra, A. (2018a). The value of transcriptomics in advancing knowledge of the immune response and diagnosis in tuberculosis. *Nat. Immunol.* 19 (11), 1159–1168. doi:10.1038/s41590-018-0225-9
- Singhania, A., Verma, R., Graham, C. M., Lee, J., Tran, T., Richardson, M., et al. (2018b). A modular transcriptional signature identifies phenotypic heterogeneity of human tuberculosis infection. *Nat. Commun.* 9 (1), 2308. doi:10.1038/s41467-018-04579-w
- Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinform. Biol. Insights* 14, 1177932219899051. doi:10.1177/1177932219899051
- Tannock, I. F., and Hickman, J. A. (2016). Limits to personalized cancer medicine. *N. Engl. J. Med.* 375 (13), 1289–1294. doi:10.1056/NEJMs1607705
- The Lancet Microbe (2024). Drug-resistant tuberculosis: the next sequence of events. *Lancet Microbe* 5 (1), e1.
- Uplekar, M., Weil, D., Lonnroth, K., Jaramillo, E., Lienhardt, C., Dias, H. M., et al. (2015). WHO's new end TB strategy. *Lancet* 385 (9979), 1799–1801. doi:10.1016/S0140-6736(15)60570-0
- WHO (2022). *WHO consolidated guidelines on tuberculosis: module 4: treatment - drug-susceptible tuberculosis treatment*. Geneva: World Health Organization.
- Wong, A., Otlis, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Coolley, O., et al. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern. Med.* 181 (8), 1065–1070. doi:10.1001/jamainternmed.2021.2626
- World health organization (2024). *Global tuberculosis report*. Geneva: World Health Organization.