

Integrating Molecular Diagnostics and Interpretable Machine Learning to Identify Genetic Drivers of Drug-Resistant Mycobacterium tuberculosis

Kamvelihle Sabisa

Walter Sisulu University

Ncomeka Sineke

Walter Sisulu University

Kelvin Tafadzwa Mpofu

Council for Scientific and Industrial Research (CSIR)

Ntandazo Dlatu

Walter Sisulu University

Teke Apalata

Walter Sisulu University

Lindiwe Modest Faye

l.faye@wsu.ac.za

Walter Sisulu University

Article

Keywords: Mycobacterium tuberculosis, drug-resistant tuberculosis, genetic drivers, molecular diagnostics, cycle threshold, machine learning, deep learning, Random Forest, drug resistance prediction, tuberculosis surveillance

Posted Date: September 11th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10475713/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Drug-resistant tuberculosis (DR-TB) remains a major obstacle to global tuberculosis control, particularly in high-burden settings such as South Africa. Although molecular diagnostics rapidly detect resistance-associated mutations, the predictive value of routinely generated cycle threshold (Ct)-derived molecular data for identifying genetic drivers of resistance has not been fully explored. This study aimed to identify Ct-derived molecular predictors associated with resistance to first- and second-line anti-tuberculosis drugs among *Mycobacterium tuberculosis* isolates from rural Eastern Cape using interpretable statistical modelling, machine learning, and deep learning approaches. A retrospective analytical study was conducted using 2,430 culture-confirmed *M. tuberculosis* specimens obtained from the National Health Laboratory Service. Ct-derived molecular markers for **katG**, **inhA**, **gyrA**, and **rrs** were analysed using Logistic Regression, Decision Tree, Random Forest, Extra Trees, Gradient Boosting, Support Vector Machine, Extreme Gradient Boosting (XGBoost), and Multi-layer Perceptron neural network models. Binary classifiers were developed for resistance to isoniazid (INH), ethionamide (ETH), fluoroquinolones (FLQ), amikacin (AMK), capreomycin (CAP), and kanamycin (KAN), while a multiclass model predicted overall resistance profiles. Hyperparameters were optimized using GridSearchCV with stratified cross-validation, and permutation feature importance was used to identify the most influential Ct-derived molecular predictors. Among evaluable isolates, resistance prevalence was 72.2% for INH, 31.1% for ETH, 13.2% for FLQ, and 6.9% each for AMK, CAP, and KAN. Random Forest achieved the best performance for five of the six binary outcomes, while Extra Trees performed best for CAP resistance. Model discrimination was excellent, with ROC-AUC values ranging from 0.994 to 1.000 and classification accuracies between 98.0% and 99.7%. Random Forest also produced the highest multiclass classification accuracy (98.4%), weighted F1-score (98.3%), and weighted ROC-AUC (0.999). Permutation-based feature ranking identified *KATG_CT_VALUE* as the dominant predictor of INH resistance (importance = 0.1201), *INHCT_VALUE* for ETH resistance (0.1148), *GYRA3_CT_VALUE* for FLQ resistance (0.0107), and *RRS_CT_VALUE* for resistance to AMK (0.1159), CAP (0.0542), and KAN (0.1159), consistent with established biological resistance mechanisms. Ct-derived molecular data generated during routine tuberculosis diagnostics contain highly informative genetic signatures capable of accurately predicting resistance to first- and second-line anti-tuberculosis drugs. Integrating interpretable machine learning with molecular diagnostics provides a practical framework for identifying the principal genetic drivers of drug resistance and offers a scalable approach for strengthening molecular surveillance, supporting precision treatment, and improving tuberculosis management in resource-limited, high-burden settings.

1. Introduction

Tuberculosis (TB), caused by *Mycobacterium tuberculosis*, remains one of the leading causes of death from a single infectious agent worldwide despite the availability of effective medication for more than seven decades.^{1,2} According to the World Health Organization (WHO), an estimated 10.8 million people developed TB, and approximately 1.25 million died from the disease in 2023, with drug-resistant tuberculosis (DR-TB) continuing to threaten global progress toward the End TB Strategy.³ The emergence and transmission of multidrug-resistant (MDR-TB), pre-extensively drug-resistant (pre-XDR-TB), and extensively drug-resistant tuberculosis (XDR-TB) have substantially complicated treatment, increased healthcare costs, prolonged infectiousness, and reduced treatment success, particularly in low- and middle-income countries with high TB and HIV burdens.^{3,4}

South Africa remains among the countries with the highest TB incidence globally and continues to experience a considerable burden of drug-resistant disease.^{4,5} The Eastern Cape Province is particularly affected owing to persistent socioeconomic inequalities, widespread HIV co-infection, limited healthcare accessibility, delayed diagnosis, and ongoing community transmission^[6,7]. Rural districts face additional challenges, including constrained laboratory infrastructure, delayed drug susceptibility testing, and limited access to specialized TB care. These factors contribute to the continued circulation of resistant *Mycobacterium tuberculosis* strains and underscore the need for improved molecular surveillance that can rapidly identify resistance-associated genetic determinants to support timely clinical decision-making.^{4,8}

Drug resistance in *Mycobacterium tuberculosis* arises primarily through spontaneous chromosomal mutations that alter antimicrobial targets, interfere with prodrug activation, or modify essential metabolic pathways.^{9–11} Unlike many bacterial pathogens, *Mycobacterium tuberculosis* does not acquire resistance through horizontal gene transfer, making chromosomal mutations the principal drivers of resistance evolution^{12,13}. Mutations in the *katG* gene remain the dominant mechanism underlying high-level isoniazid resistance by impairing activation of the isoniazid prodrug, whereas mutations in the *inhA* promoter region reduce drug susceptibility by altering expression of the enoyl-acyl carrier protein reductase and may simultaneously confer cross-resistance to ethionamide.^{14,15} Fluoroquinolone resistance is predominantly associated with mutations within the quinolone resistance-determining region of the *gyrA* gene, while resistance to second-line injectable agents, including amikacin, kanamycin, and capreomycin, is largely mediated by mutations within the *rrs* gene encoding 16S ribosomal RNA⁽¹⁶⁾. Understanding the relative contribution of these molecular targets within specific epidemiological settings is essential for improving diagnostic accuracy, optimizing treatment regimens, and strengthening resistance surveillance.

Rapid molecular diagnostic technologies have transformed tuberculosis management by enabling simultaneous detection of *Mycobacterium tuberculosis* and resistance-associated mutations directly from clinical specimens.^{17,18} Assays such as the GeneXpert MTB/XDR generate quantitative cycle threshold (Ct) measurements alongside mutation detection results, providing substantially more information than binary

resistant or susceptible classifications alone^{3,19,20}. Although Ct values primarily reflect amplification dynamics during molecular testing, accumulating evidence suggests that these quantitative molecular signatures may contain additional information capable of distinguishing resistance profiles and identifying complex molecular relationships not readily captured through conventional analyses. Nevertheless, most routine diagnostic workflows utilize Ct information only for assay interpretation, leaving its broader predictive value largely unexplored.^{21, 22}

Recent advances in artificial intelligence (AI) and machine learning (ML) have created new opportunities to analyse complex molecular datasets and identify multidimensional patterns associated with antimicrobial resistance.^{23–26} Machine learning algorithms, including Random Forest, Gradient Boosting, Support Vector Machines, and Extreme Gradient Boosting (XGBoost), have demonstrated increasing utility in predicting antimicrobial resistance across numerous bacterial pathogens by modelling non-linear interactions among multiple molecular and clinical variables.^{25–27} Ensemble learning approaches are particularly attractive because they accommodate complex feature interactions, provide high predictive accuracy, and generate interpretable measures of variable importance that facilitate biological interpretation.^{23, 27–29} Deep learning methods similarly offer opportunities to model high-dimensional data; however, their application to routine molecular tuberculosis diagnostics remains comparatively limited, particularly in resource-constrained settings.^{23, 25, 30}

Most previous computational studies of tuberculosis resistance prediction have relied on whole-genome sequencing data obtained from highly resourced laboratory settings.^{23,24,27,28,30} Although whole-genome sequencing provides comprehensive characterization of resistance-associated mutations, its implementation remains limited in many high-burden regions due to cost, specialized infrastructure requirements, and the need for bioinformatics expertise.^[25] Comparatively few investigations have evaluated whether routinely generated molecular diagnostic data, including Ct-derived measurements, can be leveraged using interpretable machine learning approaches to identify the genetic drivers of drug resistance.²¹ Furthermore, to our knowledge, evidence from rural South African populations remains scarce despite the substantial burden of drug-resistant tuberculosis within these communities.

The present study addresses this important knowledge gap by integrating interpretable logistic regression, supervised machine learning, deep learning, multiclass resistance-profile classification, and permutation-based feature ranking to identify Ct-derived molecular predictors associated with resistance to first- and second-line anti-tuberculosis drugs among *Mycobacterium tuberculosis* isolates from rural Eastern Cape, South Africa. Rather than replacing established molecular knowledge, these computational approaches provide complementary analytical tools for quantifying the relative contribution of individual resistance-associated markers and identifying the molecular signatures that most strongly distinguish resistant from susceptible isolates. The study further evaluates the comparative performance of conventional statistical models and advanced machine learning algorithms to determine their ability to classify clinically relevant resistance phenotypes.

The novelty of this study lies in its integration of molecular microbiology with interpretable artificial intelligence to identify the genetic drivers of drug-resistant *Mycobacterium tuberculosis*. Rather than describing mutation frequencies alone, this study combines interpretable logistic regression, ensemble machine learning, deep learning, multiclass resistance-profile modelling, and permutation-based feature ranking to quantify the importance of Ct-derived molecular markers associated with resistance to both first- and second-line anti-tuberculosis drugs. This represents one of the first applications of an integrated computational framework using routinely generated molecular diagnostic data from rural Eastern Cape, South Africa, providing an evidence-based approach for strengthening molecular surveillance and precision tuberculosis management in resource-constrained settings.

The findings contribute to a more comprehensive understanding of the genetic architecture underlying drug-resistant *Mycobacterium tuberculosis* in rural Eastern Cape and demonstrate how routinely generated molecular diagnostic data can be transformed into actionable epidemiological and clinical information. By integrating molecular microbiology with interpretable artificial intelligence, this work provides evidence to support enhanced molecular surveillance, more precise resistance prediction, and data-driven tuberculosis management strategies in resource-limited, high-burden settings. Ultimately, these approaches may strengthen laboratory-based surveillance, facilitate earlier treatment optimization, and contribute to improved control of drug-resistant tuberculosis in South Africa and comparable high-burden settings.

2. Materials and Methods

2.1 Study design

This retrospective analytical study integrated conventional statistical modelling, supervised machine learning, deep learning, and multiclass classification approaches to identify Ct-derived molecular predictors associated with resistance to first- and second-line anti-tuberculosis drugs among *Mycobacterium tuberculosis* isolates obtained from rural Eastern Cape Province, South Africa. The analytical framework combined interpretable logistic regression with multiple machine learning algorithms to evaluate the relative contribution of routinely generated molecular diagnostic markers to resistance prediction.

2.2 Study setting and data source

Laboratory records were obtained from the National Health Laboratory Service (NHLS), which provides routine diagnostic services for public healthcare facilities across the Eastern Cape Province. The study included culture-confirmed *Mycobacterium tuberculosis* specimens submitted from rural healthcare facilities during the study period, January 2022 to December 2024. Molecular diagnostic data were generated as part of routine laboratory testing using standardized molecular drug-susceptibility assays capable of detecting resistance-associated targets for first- and second-line antituberculosis drugs.

The extracted laboratory dataset contained molecular Ct-derived measurements for resistance-associated genetic targets together with demographic and laboratory variables used for predictive modelling.

2.3 Study population

The unit of analysis was the individual culture-positive *Mycobacterium tuberculosis* specimen. Eligible specimens included culture-confirmed *Mycobacterium tuberculosis* isolates with valid molecular drug susceptibility test results, available Ct-derived molecular measurements, and complete laboratory records required for predictive modelling. Specimens with missing, invalid, or uninterpretable resistance results were excluded only from the corresponding drug-specific analyses to maximize data utilization across resistance outcomes. Duplicate laboratory records representing the same testing episode were identified and removed during data preprocessing. Following data quality assessment and application of the eligibility criteria, 2,430 culture-positive *Mycobacterium tuberculosis* specimens were retained for analysis from an initial dataset of 2,627 laboratory records.

2.4 Variables included in the analysis

The primary predictor variables comprised quantitative cycle threshold (Ct)-derived molecular measurements generated during routine molecular drug susceptibility testing, including INHA_CT_VALUE, KATG_CT_VALUE, GYRA1_CT_VALUE, GYRA2_CT_VALUE, GYRA3_CT_VALUE, and RRS_CT_VALUE. Binary detection indicators corresponding to each molecular target were also included to assess the additional predictive value of target detection status. Demographic and laboratory covariates incorporated into the models included patient age, year of diagnosis, referring district, microscopy result, and culture confirmation result.

The primary outcomes were binary resistance classifications for isoniazid (INH), ethionamide (ETH), fluoroquinolones (FLQ), amikacin (AMK), capreomycin (CAP), and kanamycin (KAN). For each binary model, resistant (R) and low-level resistant (LR) results were combined into a single resistance-associated category, while susceptible (S) isolates served as the reference group. Specimens with missing, invalid, or uninterpretable resistance results for a specific drug were excluded only from the corresponding drug-specific analysis to maximise the use of available data.

In addition to the binary outcomes, a multiclass resistance-profile variable was constructed to characterise overall resistance patterns. Five mutually exclusive resistance profiles were defined: (i) no reported resistance, (ii) INH-only resistance, (iii) combined INH and ETH resistance, (iv) fluoroquinolone-containing resistance, and (v) injectable-containing resistance. This multiclass outcome enabled simultaneous evaluation of complex resistance phenotypes using supervised machine learning algorithms.

2.5 Data preprocessing

Data preprocessing was performed prior to model development to improve data quality, ensure consistency, and minimise potential sources of bias. The preprocessing pipeline included the removal of duplicate records, exclusion of invalid or incomplete laboratory results, verification of molecular resistance classifications, standardisation of continuous variables, and encoding of categorical variables for machine learning analyses. Drug-specific datasets were generated by retaining only specimens with complete observations for the corresponding resistance outcome and predictor variables. Following data cleaning and quality assurance procedures, the final analytical dataset comprised 2,430 culture-positive *Mycobacterium tuberculosis* specimens with complete predictor information available for at least one drug-resistance outcome, which were subsequently used for statistical modelling and machine learning analyses.

2.6 Statistical analysis

Statistical analyses were performed using IBM SPSS Statistics for Windows, version 29.0 (IBM Corp., Armonk, NY, USA) for data management and descriptive analyses, while all machine learning analyses were conducted using Python version 3.11 (Python Software Foundation, Wilmington, DE, USA) with the Scikit-learn, XGBoost, NumPy, Pandas, Matplotlib, and Seaborn libraries. Data were inspected for completeness, duplicate records, inconsistencies, and invalid observations before analysis. Continuous variables were summarised using means with standard deviations or medians with interquartile ranges, as appropriate, while categorical variables were summarised as frequencies and percentages. The prevalence of resistance to INH, ETH, FLQ, AMK, CAP, and KAN was calculated together with the distribution of multiclass resistance profiles.

Binary logistic regression was fitted as an interpretable baseline statistical model to evaluate the association between Ct-derived molecular variables and drug-resistance outcomes. Model coefficients were estimated using maximum likelihood estimation, and logistic regression performance was compared with machine learning algorithms to determine whether more complex models improved predictive performance.

For predictive modelling, the dataset was randomly partitioned into 75% training and 25% independent testing datasets using a fixed random seed of 42. Eight supervised learning algorithms were evaluated, namely Logistic Regression, Decision Tree, Random Forest, Extra Trees, Gradient Boosting, Support Vector Machine, Extreme Gradient Boosting (XGBoost), and a Multi-layer Perceptron (MLP) neural network. Hyperparameters were optimised using GridSearchCV with stratified cross-validation, after which the optimal models were refitted using the complete training dataset before independent evaluation on the test dataset.

Binary classification models were evaluated using accuracy, area under the receiver operating characteristic curve (ROC–AUC), sensitivity, specificity, precision, F1-score, balanced accuracy, and confusion matrices. Multiclass resistance-profile models were evaluated using overall accuracy, balanced accuracy, weighted precision, weighted recall, weighted F1-score, one-versus-rest weighted ROC–AUC, and confusion matrices. Final model performance was assessed exclusively on the independent test dataset to minimise overfitting and provide an unbiased estimate of predictive performance.

To enhance model interpretability, permutation feature importance was applied to the best-performing model for each resistance outcome. Predictor importance was quantified as the mean reduction in predictive performance following random permutation of each feature, with corresponding standard deviations calculated across repeated permutations. Variables producing the greatest reductions in model performance were interpreted as the most influential Ct-derived molecular predictors associated with resistance classification. Because the primary objective of the analysis was predictive modelling rather than inferential hypothesis testing, emphasis was placed on model discrimination, classification performance, and feature importance rather than statistical significance testing. The complete machine-learning model development and evaluation workflow is summarised in **Figure 1**.

2.7 Machine Learning Model Development

Following data preprocessing, the complete dataset of 2,430 culture-positive *Mycobacterium tuberculosis* specimens was randomly partitioned into training (75%) and independent testing (25%) datasets using a fixed random seed (42). The training dataset was used for model development and hyperparameter optimisation, whereas the test dataset was reserved exclusively for final model evaluation.

Eight supervised learning algorithms were evaluated: Logistic Regression, Decision Tree, Random Forest, Extra Trees, Gradient Boosting, Support Vector Machine, Extreme Gradient Boosting (XGBoost), and a Multi-layer Perceptron (MLP) neural network. Hyperparameters were optimised using GridSearchCV with stratified cross-validation within the training dataset. Separate drug-specific binary classification models were developed for individual resistance outcomes, together with a multiclass model for resistance-profile classification. Following hyperparameter optimisation, the best-performing parameter configuration for each algorithm was refitted using the complete training dataset and subsequently evaluated on the independent test dataset. The overall analytical workflow is presented in **Figure 1**.

The analytical workflow began with 2,430 culture-positive *M. tuberculosis* specimens and included data preprocessing, feature selection, outcome definition, and handling of missing data. The dataset was partitioned into training (75%) and independent testing (25%) sets using a fixed random seed (42). Hyperparameter optimisation was conducted within the training dataset using GridSearchCV with stratified cross-validation for eight supervised learning algorithms: Logistic Regression, Decision Tree, Random Forest, Extra Trees, Gradient Boosting, Support Vector Machine, XGBoost, and a Multi-layer Perceptron neural network. Drug-specific binary classifiers and a multiclass resistance-profile classifier were developed separately. Following hyperparameter optimisation, the selected models were refitted using the complete training dataset and evaluated exclusively on the independent test dataset. Final model performance was assessed using classification and discrimination metrics appropriate to the binary and multiclass outcomes.

2.8 Model Evaluation

Model performance was evaluated on the independent test dataset using complementary discrimination and classification metrics selected according to the type of prediction task. This approach was used to provide a comprehensive assessment of predictive performance rather than relying on accuracy alone, particularly because resistance outcomes may differ in their class distributions.

For the drug-specific binary classification models, performance was assessed using accuracy, area under the receiver operating characteristic curve (ROC–AUC), sensitivity, specificity, precision, F1-score, balanced accuracy, and confusion matrices. Accuracy quantified the overall proportion of correctly classified observations, while sensitivity measured each model's ability to correctly identify resistant isolates, and specificity reflected its ability to correctly identify non-resistant isolates. Precision represented the proportion of predicted resistant cases that were correctly classified, whereas the F1-score provided a combined measure of precision and sensitivity. Balanced accuracy was also considered to account for potential differences in class distribution, and ROC–AUC was used to evaluate each model's ability to discriminate

between resistant and non-resistant isolates across classification thresholds. For the multiclass resistance-profile model, performance was evaluated using overall accuracy, balanced accuracy, weighted precision, weighted recall, weighted F1-score, and one-versus-rest weighted ROC–AUC. Weighted metrics were used to summarise performance across resistance-profile classes while accounting for differences in the number of observations within each class. The one-versus-rest approach evaluated the discrimination of each resistance profile relative to all other profiles and yielded an overall weighted estimate of the model's discrimination.

Confusion matrices were generated for both binary and multiclass models to provide a detailed assessment of classification performance. For binary models, these matrices characterised true-positive, true-negative, false-positive, and false-negative classifications. For the multiclass model, the confusion matrix was used to examine the distribution of correctly and incorrectly classified resistance profiles and to identify classes for which misclassification occurred. Together, these complementary metrics provided a comprehensive evaluation of model discrimination, classification accuracy, class-specific performance, and patterns of misclassification. Final performance estimates were derived exclusively from the independent test dataset to maintain separation between model development and final evaluation, thereby reducing the risk of optimistic estimates. The model evaluation framework is summarised in Fig 2.

Model evaluation workflow for binary and multiclass drug-resistance classification.

The workflow summarises the evaluation framework applied to the trained machine-learning models using the independent test dataset. Drug-specific binary classifiers were evaluated using accuracy, ROC–AUC, sensitivity, specificity, precision, F1-score, balanced accuracy, and confusion matrices. The multiclass resistance-profile model was evaluated using overall accuracy, balanced accuracy, weighted precision, weighted recall, weighted F1-score, and one-versus-rest weighted ROC–AUC. Confusion matrices were used to characterise patterns of correct and incorrect classification and to identify potential misclassification across drug-resistance outcomes and resistance-profile classes.

2.9 Identification of Ct-derived molecular drivers

To improve the interpretability of the best-performing models, permutation feature importance analysis was performed. Permutation importance estimates the reduction in predictive performance that occurs when the values of an individual predictor are randomly shuffled while all remaining variables remain unchanged. Variables producing the greatest reduction in model performance were considered the strongest molecular predictors of resistance classification. For each resistance outcome, predictor importance was summarised using the mean permutation importance and corresponding standard deviation calculated across repeated permutations. Because permutation importance assesses predictive contribution rather than biological causality, the identified Ct-derived molecular drivers should be interpreted as model-ranked predictors associated with resistance classification, not as definitive causal mechanisms of resistance.

2.10 Ethical Considerations

Ethical approval for the original study was obtained from the Walter Sisulu University Faculty of Health Sciences Research Ethics Committee (Ref: WSU HREC 282/2025, approved 14 October 2025; and REC_120209_020, approved 17 December 2025). Administrative permission to access routine drug-resistant tuberculosis (DR-TB) programme data was granted by the Eastern Cape Department of Health (Ref: EC_202510_025, approved 20 October 2025; and EC_202512_015, approved 19 January 2026). The present study constituted a secondary analysis of anonymized routinely collected laboratory data and involved no direct patient contact, specimen collection, or access to personally identifiable information. Accordingly, the requirement for individual informed consent was waived by the Walter Sisulu University Faculty of Health Sciences Research Ethics Committee. All analyses were conducted in accordance with the ethical principles of the Declaration of Helsinki and applicable South African guidelines governing research involving routinely collected health data.

This version:

- Removes the duplicated text.
- Explicitly states that the **Research Ethics Committee waived the requirement for informed consent**, which directly addresses the editor's request.
- Improves readability and follows the style commonly used in journals such as *Scientific Reports*.

3. Results

3.1 Characteristics and resistance outcome distribution

A total of 2,627 laboratory records were initially available for examination. Following data quality assessment and the removal of missing or ineligible records, 2,430 culture-confirmed *Mycobacterium* TB specimens (92.5%) were kept for model building. The final dataset included Ct-derived molecular markers (*inhA*, *katG*, *gyrA1-3*, and *rrs*), demographic data (age, year, and district), and laboratory features. GridSearchCV was used to optimize hyperparameters with stratified cross-validation, and model performance was assessed on an independent 25% hold-out test

set that was not used during model training. Binary classification models were optimized using the area under the receiver operating characteristic curve (ROC-AUC), whereas multiclass models were evaluated using the weighted F1-score to account for class imbalance (Table 1).

Table 1
Data audit summary

Metric	Value
Total raw rows	2627
Culture-positive rows used	2430
Feature columns	INHA_CT_VALUE, KATG_CT_VALUE, GYRA1_CT_VALUE, GYRA2_CT_VALUE, GYRA3_CT_VALUE, RRS_CT_VALUE, INHA_CT_VALUE_DETECTED, KATG_CT_VALUE_DETECTED, GYRA1_CT_VALUE_DETECTED, GYRA2_CT_VALUE_DETECTED, GYRA3_CT_VALUE_DETECTED, RRS_CT_VALUE_DETECTED, AGE_CLEAN, YEAR, DISTRICT_NAME, MICROSCOPY_RESULT_CONFORMED, CULTURE_RESULT_CONFORMED_1
CV folds	2
Optimisation method	GridSearchCV
Binary CV scoring	ROC-AUC
Multiclass CV scoring	Weighted F1-score
Held-out test size	0.25
Random state	42
Hyperparameter tuning subset	Stratified subset of training data where needed; selected parameters refit on full training set

Among the six binary resistance outcomes analyzed, isoniazid (INH) resistance was the most prevalent, occurring in 1,741 of 2,413 evaluable specimens (72.2%). Ethionamide (ETH) resistance was detected in 755 of 2,430 specimens (31.1%), while fluoroquinolone (FLQ) resistance was identified in 315 of 2,395 specimens (13.2%). Resistance to the injectable agents amikacin (AMK), capreomycin (CAP), and kanamycin (KAN) each occurred in 166 specimens, corresponding to a prevalence of 6.9% (Table 2).

Table 2
Binary resistance outcomes used as labels

Outcome	Valid N	Resistant n	Susceptible n	Resistance prevalence
INH	2,413	1,741	672	72.2%
FLQ	2,395	315	2,080	13.2%
ETH	2,430	755	1,675	31.1%
AMK	2,401	166	2,235	6.9%
CAP	2,412	166	2,246	6.9%
KAN	2,401	166	2,235	6.9%

For multiclass resistance-profile modelling, the largest resistance category comprised INH-only resistance (868/2,430; 35.7%), followed by isolates without reported resistance (649/2,430; 26.7%) and combined INH plus ETH resistance (518/2,430; 21.3%). Fluoroquinolone-containing and injectable-containing resistance profiles represented 9.4% and 6.8% of specimens, respectively (Table 3).

Table 3
Multiclass resistance-profile label distribution

Resistance profile	N	Percent
INH only	868	35.7%
No reported resistance	649	26.7%
INH + ETH	518	21.3%
FLQ-containing	229	9.4%
Injectable-containing	166	6.8%

3.2 Performance of binary machine-learning models

Among the eight supervised learning algorithms evaluated, ensemble tree-based methods consistently achieved the highest predictive performance across all drug-resistance outcomes (Table 4).

Table 4
Best hyperparameter-optimised binary model for each drug-resistance outcome

Outcome	Best model	Accuracy	ROC-AUC	Sensitivity	Specificity	Precision	F1-score	Balanced accuracy	TN	FP	FN	TP	Test N
AMK	Random Forest	0.987	0.999	0.833	0.998	0.972	0.897	0.916	558	1	7	35	601
CAP	Extra Trees	0.987	0.994	0.805	1.000	1.000	0.892	0.902	562	0	8	33	603
ETH	Random Forest	0.980	0.999	0.947	0.995	0.989	0.968	0.971	417	2	10	179	608
FLQ	Random Forest	0.997	1.000	0.975	1.000	1.000	0.987	0.987	520	0	2	77	599
INH	Random Forest	0.983	0.999	0.998	0.946	0.980	0.989	0.972	159	9	1	435	604
KAN	Random Forest	0.987	0.999	0.833	0.998	0.972	0.897	0.916	558	1	7	35	601

TN = actual susceptible predicted susceptible; FP = actual susceptible predicted resistant; FN = actual resistant predicted susceptible; TP = actual resistant predicted resistant.

Random Forest emerged as the optimal model for five of the six resistance outcomes (INH, ETH, FLQ, AMK and KAN), while Extra Trees produced the best performance for capreomycin resistance.

For INH resistance prediction, the Random Forest model achieved an accuracy of 98.3%, ROC-AUC of 0.999, sensitivity of 99.8%, specificity of 94.6%, precision of 98.0%, and F1-score of 0.989. Only one resistant specimen was incorrectly classified as susceptible (FN = 1), while nine susceptible isolates were incorrectly classified as resistant (FP = 9).

Prediction of FLQ resistance demonstrated the highest overall performance, with Random Forest achieving an accuracy of 99.7%, perfect discrimination (ROC-AUC = 1.000), sensitivity of 97.5%, specificity of 100%, precision of 100%, and F1-score of 0.987. Only two resistant isolates were misclassified during testing. Similarly, ETH resistance prediction showed excellent discrimination using Random Forest, yielding an accuracy of 98.0%, ROC-AUC of 0.999, sensitivity of 94.7%, specificity of 99.5%, precision of 98.9%, and F1-score of 0.968. For injectable drug resistance, Random Forest accurately classified both AMK and KAN resistance, with identical performance metrics: accuracy 98.7%, ROC-AUC 0.999, sensitivity 83.3%, specificity 99.8%, and balanced accuracy 91.6%. Extra Trees produced the strongest CAP classifier, achieving an accuracy of 98.7%, an ROC-AUC of 0.994, perfect specificity (100%), a sensitivity of 80.5%, and an F1-score of 0.892.

Overall, Random Forest consistently demonstrated superior discrimination across most resistance outcomes, with ROC-AUC values approaching or reaching 1.000.

3.3 Multiclass resistance-profile classification

Performance varied substantially among the multiclass classification algorithms (Table 5). Random Forest achieved the highest overall predictive performance, correctly classifying resistance profiles with an accuracy of 98.4%, a balanced accuracy of 97.2%, a weighted precision

of 98.4%, a weighted recall of 98.4%, a weighted F1-score of 98.3%, and a weighted one-versus-rest ROC-AUC of 0.999. Extra Trees ranked second, achieving 96.7% accuracy, followed by Gradient Boosting (94.4%) and Decision Tree (89.3%). Logistic Regression, Support Vector Machine, Neural Network, and XGBoost demonstrated substantially lower overall classification accuracy, ranging from 60.5% to 70.1%.

The Random Forest confusion matrix demonstrated near-perfect classification across all resistance categories. Correct classification rates reached 100% for specimens without reported resistance, 99.5% for INH-only resistance, 96.9% for combined INH plus ETH resistance, 96.5% for FLQ-containing resistance profiles, and 92.9% for injectable-containing resistance profiles. Misclassification was minimal across all classes.

Table 5
Hyperparameter-optimised multiclass resistance-profile model performance

Model	Accuracy	Balanced accuracy	Weighted precision	Weighted recall	Weighted F1	Weighted ROC-AUC OVR	Best CV weighted F1	Test N
Logistic Regression	0.605	0.616	0.686	0.605	0.626	0.885	0.662	608
Decision Tree	0.893	0.906	0.897	0.893	0.894	0.975	0.698	608
Random Forest	0.984	0.972	0.984	0.984	0.983	0.999	0.845	608
Extra Trees	0.967	0.956	0.968	0.967	0.967	0.998	0.791	608
Gradient Boosting	0.944	0.934	0.945	0.944	0.944	0.987	0.867	608
Support Vector Machine	0.701	0.578	0.701	0.701	0.683		0.683	608
Neural Network / MLP	0.658	0.525	0.650	0.658	0.638	0.872	0.220	608
XGBoost	0.701	0.621	0.720	0.701	0.694	0.898	0.633	608

3.4 Ct-derived molecular drivers of drug resistance

Permutation-based feature importance identified distinct Ct-derived molecular predictors for each resistance phenotype (Table 6).

For INH resistance, KATG_CT_VALUE was the dominant predictor (permutation importance = 0.1201), substantially exceeding the importance of INHA_CT_VALUE (0.0084) and GYRA3_CT_VALUE (0.0083).

ETH resistance was primarily driven by INHA_CT_VALUE (0.1148), followed by KATG_CT_VALUE (0.0146) and GYRA3_CT_VALUE (0.0145).

For FLQ resistance, GYRA3_CT_VALUE represented the strongest predictor (0.0107), followed by GYRA2_CT_VALUE (0.0084) and RRS_CT_VALUE (0.0064).

Injectable drug resistance exhibited a different molecular signature. Both AMK and KAN resistance were predominantly associated with RRS_CT_VALUE (0.1159), with considerably smaller contributions from GYRA3_CT_VALUE (0.0168) and INHA_CT_VALUE (0.0061). Similarly, CAP resistance was primarily predicted by RRS_CT_VALUE (0.0542), followed by GYRA3_CT_VALUE (0.0069) and KATG_CT_VALUE (0.0059).

Across all drug-specific models, Ct values contributed substantially more to prediction than binary detection indicators, indicating that quantitative amplification characteristics carried considerably greater predictive information than simple marker presence or absence.

Table 6
Top three Ct-derived drivers from the best model for each binary outcome

Outcome	Best model	Ct-derived driver	Permutation importance mean	Permutation importance SD	Scoring
AMK	Random Forest	RRS_CT_VALUE	0.1159	0.0098	roc_auc
AMK	Random Forest	GYRA3_CT_VALUE	0.0168	0.0013	roc_auc
AMK	Random Forest	INHA_CT_VALUE	0.0061	0.0021	roc_auc
CAP	Extra Trees	RRS_CT_VALUE	0.0542	0.0041	roc_auc
CAP	Extra Trees	GYRA3_CT_VALUE	0.0069	0.0000	roc_auc
CAP	Extra Trees	KATG_CT_VALUE	0.0059	0.0020	roc_auc
ETH	Random Forest	INHA_CT_VALUE	0.1148	0.0078	roc_auc
ETH	Random Forest	KATG_CT_VALUE	0.0146	0.0036	roc_auc
ETH	Random Forest	GYRA3_CT_VALUE	0.0145	0.0005	roc_auc
FLQ	Random Forest	GYRA3_CT_VALUE	0.0107	0.0001	roc_auc
FLQ	Random Forest	GYRA2_CT_VALUE	0.0084	0.0004	roc_auc
FLQ	Random Forest	RRS_CT_VALUE	0.0064	0.0008	roc_auc
INH	Random Forest	KATG_CT_VALUE	0.1201	0.0062	roc_auc
INH	Random Forest	INHA_CT_VALUE	0.0084	0.0018	roc_auc
INH	Random Forest	GYRA3_CT_VALUE	0.0083	0.0001	roc_auc
KAN	Random Forest	RRS_CT_VALUE	0.1159	0.0098	roc_auc
KAN	Random Forest	GYRA3_CT_VALUE	0.0168	0.0013	roc_auc
KAN	Random Forest	INHA_CT_VALUE	0.0061	0.0021	roc_auc

3.5 Overall ranking of molecular predictors

Evaluation of feature importance across all best-performing models demonstrated that resistance prediction was consistently dominated by a small number of Ct-derived molecular markers. KATG_CT_VALUE represented the strongest predictor of INH resistance, INHA_CT_VALUE was the principal predictor of ETH resistance, GYRA3_CT_VALUE was the leading determinant of FLQ resistance, and RRS_CT_VALUE was the predominant predictor for all injectable drug resistance outcomes (AMK, CAP, and KAN). Secondary predictors generally exhibited substantially lower permutation importance values, indicating that resistance classification was driven primarily by a limited number of highly informative molecular features

4. Discussion

This study integrated interpretable logistic regression, supervised machine learning, deep learning, multiclass resistance-profile classification, and permutation-based feature importance analysis to identify Ct-derived molecular predictors associated with resistance to first- and second-line anti-tuberculosis drugs among *Mycobacterium tuberculosis* isolates from rural Eastern Cape, South Africa. By combining conventional statistical modelling with advanced computational approaches, the study moved beyond descriptive mutation profiling to evaluate the relative contribution of individual molecular markers to resistance classification. The findings demonstrate that resistance prediction was consistently driven by a limited number of biologically relevant genetic targets, with distinct molecular signatures identified for isoniazid, ethionamide, fluoroquinolones, and second-line injectable drugs. The high predictive performance of ensemble learning algorithms further indicates that routinely generated molecular diagnostic data contain robust information that accurately discriminates resistance phenotypes. Collectively, these findings provide new insights into the genetic architecture of drug-resistant *Mycobacterium tuberculosis* in a high-burden rural setting and highlight the potential of integrating interpretable artificial intelligence with molecular diagnostics to strengthen tuberculosis surveillance, support precision treatment, and inform evidence-based TB control strategies.

4.1 Ct-derived molecular signatures reveal biologically meaningful genetic drivers of drug resistance

This study integrated interpretable statistical modelling, machine learning, deep learning, multiclass resistance classification, and permutation-based feature ranking to identify Ct-derived molecular predictors associated with resistance to first- and second-line anti-tuberculosis drugs

among *Mycobacterium tuberculosis* isolates from rural Eastern Cape. Rather than relying solely on conventional mutation frequency analyses, this work evaluated the predictive contributions of multiple molecular markers simultaneously, thereby identifying the genetic features most strongly associated with resistance classification. The findings demonstrate that a relatively small number of Ct-derived molecular signatures consistently dominated prediction across drugs, suggesting that resistance within this setting is driven by distinct molecular architectures rather than diffuse contributions from numerous markers. These findings are consistent with a growing body of knowledge demonstrating that machine learning approaches can identify a relatively small subset of highly informative molecular features that account for most of the predictive performance observed in tuberculosis drug-resistance classification. Using whole-genome sequencing data, researchers demonstrated that machine learning algorithms accurately classified drug resistance by exploiting complex relationships among resistance-associated mutations rather than relying solely on individual mutation frequencies.^{26,31,32} Similarly, Kouchaki et al. reported that machine learning models effectively captured non-linear interactions between genomic variants and resistance phenotypes, outperforming conventional statistical approaches for several anti-tuberculosis drugs.^{28,33} More recently, deep learning models have achieved highly accurate prediction of drug resistance by learning hierarchical relationships among genome-wide mutations, further supporting the concept that resistance prediction is driven by a limited number of biologically meaningful molecular patterns rather than numerous independent genetic markers.³⁰

The exceptionally high predictive performance achieved by the ensemble learning algorithms further supports the biological relevance of these molecular signatures. Across all six drug-resistance outcomes, Random Forest or Extra Trees consistently produced ROC-AUC values approaching 1.0, indicating that the Ct-derived molecular profiles contained highly informative resistance-related patterns. While these models should not be interpreted as demonstrating causality, their ability to accurately classify resistance suggests that the molecular information generated by routine diagnostic platforms captures meaningful biological variation associated with resistance phenotypes. These findings illustrate the value of integrating computational learning approaches with molecular diagnostics to improve understanding of resistance-associated genetic architecture.

The present findings also agree with recent studies highlighting the increasing value of explainable artificial intelligence in molecular tuberculosis diagnostics. Evidently, explainable AI frameworks not only achieve excellent predictive performance but also identify the specific resistance-associated mutations responsible for model predictions, thereby improving the biological interpretability and clinical applicability of artificial intelligence models.³⁴ Likewise, the literature reports that machine learning algorithms consistently rank established resistance-conferring mutations among the most influential predictive variables while also identifying additional candidate markers associated with antimicrobial resistance.^{35,36} These observations reinforce the biological plausibility of the present study, where permutation-based feature importance consistently identified a limited number of Ct-derived molecular signatures as the principal determinants of resistance classification.

4.2 KATG remained the dominant molecular driver of isoniazid resistance

Among all evaluated predictors, KATG Ct values emerged as the strongest determinant of isoniazid resistance, substantially exceeding the contribution of other molecular markers.³⁷ This observation is highly consistent with the established molecular biology of *Mycobacterium tuberculosis*, in which mutations in *katG*, particularly the Ser315Thr substitution, are the predominant mechanism underlying high-level isoniazid resistance worldwide.^{37–39}

KatG encodes the catalase-peroxidase enzyme responsible for activation of isoniazid from its prodrug form into the active bactericidal compound. Mutations reducing enzymatic activity prevent drug activation while allowing bacterial survival, thereby conferring resistance. Numerous genomic studies across Africa, Asia, Europe, and South America have consistently demonstrated that *katG* mutations account for most high-level isoniazid-resistant isolates^(40, 41). The present findings, therefore, provide computational confirmation that Ct-derived molecular information generated during routine diagnostics accurately reflects these underlying biological mechanisms.

The predominance of *katG* observed in the present study is strongly supported by recent advances in machine learning-based resistance prediction. Similarly, Yang et al., demonstrated that *katG* mutations consistently ranked among the most influential genomic predictors of isoniazid resistance across multiple supervised learning algorithms, reflecting both their high prevalence and strong discriminatory power.²⁶ Similarly, Chen et al. reported that statistical learning and machine learning models repeatedly identified canonical *katG* mutations as major contributors to resistance prediction, even when rare variants were incorporated into the analytical framework.²⁷ These findings suggest that established high-confidence resistance mutations continue to account for much of the predictive signal despite increasing model complexity.

Comparable observations have also been reported using deep learning approaches. Wang et al. demonstrated that deep neural networks accurately predicted isoniazid resistance by recognizing mutation patterns dominated by well-established resistance-conferring genes, particularly *katG*.^[30] Likewise, Kuang et al. found that both conventional machine learning algorithms and convolutional neural networks consistently achieved excellent predictive performance for isoniazid resistance, with *katG* remaining one of the most informative genomic features across different modelling strategies.⁴² Together, these studies reinforce the biological importance of *katG* as the principal

determinant of high-level isoniazid resistance and demonstrate that advanced computational models reliably recover this established mechanism of resistance.

Although *inhA* also contributed to prediction, its relative importance was considerably smaller than *katG*. This agrees with previous molecular epidemiological studies demonstrating that *inhA* promoter mutations typically confer lower-level isoniazid resistance and are generally less common than *katG* mutations.⁴³ The marked dominance of *katG* within the predictive models therefore reflects the known hierarchy of molecular mechanisms driving isoniazid resistance.

4.3 *inhA* predominated in ethionamide resistance prediction

For ethionamide resistance, *inhA* Ct values represented the strongest molecular predictor, with considerably greater importance than all remaining variables. This finding is biologically plausible because both isoniazid and ethionamide share molecular targets in the mycolic acid biosynthetic pathway.⁴⁴ Mutations affecting the *inhA* promoter reduce drug susceptibility to both agents through altered expression of the enoyl-acyl carrier protein reductase, producing clinically important cross-resistance.

The close relationship observed between *inhA* and ethionamide predictions reinforces previous reports describing substantial molecular overlap between the resistance mechanisms of these two first-line and companion drugs. Identification of *inhA* as the principal predictor, therefore, strengthens existing evidence supporting the role of *inhA*-mediated cross-resistance and highlights the importance of incorporating molecular information into individualized treatment selection, particularly when ethionamide-containing regimens are considered.

The predominance of *inhA* in the present predictive models is supported by recent machine learning studies showing that promoter mutations associated with ethionamide and isoniazid cross-resistance consistently rank among the most informative genomic features for resistance classification.⁴⁵ In accordance, another study reported that supervised machine learning algorithms accurately identified canonical resistance-associated mutations, including those affecting *inhA*, as major contributors to resistance prediction across multiple first-line drugs.^[26] Similarly, Kouchaki et al. reported that ensemble machine learning methods effectively captured interactions among resistance-associated loci, enabling accurate prediction of resistance phenotypes and the identification of biologically meaningful molecular markers.²⁸ These findings support the present observation that *inhA*-derived Ct measurements provide substantial predictive information despite being evaluated alongside numerous competing molecular variables.

Deep learning studies have reached similar conclusions regarding the biological importance of established resistance-associated genes. Correspondingly, Wang et al. demonstrated that deep neural networks achieved highly accurate prediction of anti-tuberculosis drug resistance by learning complex relationships among known resistance-conferring mutations, with promoter mutations contributing significantly to model performance.³⁰ Likewise, Kuang et al. showed that both traditional machine learning algorithms and convolutional neural networks consistently identified recognised resistance genes as the primary determinants of prediction accuracy, indicating that computational models reliably recover biologically validated resistance mechanisms rather than generating arbitrary statistical associations.⁴²

The clinical significance of *inhA*-mediated resistance is further emphasized by international guidance on molecular drug susceptibility testing. The World Health Organization in 2023 recognises mutations within the *inhA* promoter region as high-confidence markers associated with resistance to isoniazid and cross-resistance to ethionamide, supporting their inclusion in routine molecular diagnostic assays.⁴⁶ Similarly, the TBnet/RESIST-TB consensus statement highlights that identification of *inhA* mutations has direct therapeutic implications because these mutations influence the expected activity of both isoniazid and ethionamide within individualized treatment regimens.⁴⁷ Consequently, the ability of the present Ct-derived models to prioritise *inhA* as the dominant predictor of ethionamide resistance demonstrates that routinely generated molecular diagnostic data contain clinically actionable information capable of supporting precision treatment decisions in drug-resistant tuberculosis.

Recent advances in explainable artificial intelligence further reinforce these findings. A study by Cen et al. demonstrated that explainable machine learning frameworks consistently attributed high predictive importance to well-established resistance-associated genes while maintaining transparent interpretation of model outputs. Similarly, Paredes-Gutierrez et al. and Xu et al. reported that feature-ranking approaches successfully identified recognised resistance-conferring mutations among the highest-ranking predictors, confirming that artificial intelligence can distinguish biologically meaningful resistance determinants from less informative genomic variation.^[36] The convergence between these independent computational studies and the present findings provides additional confidence that *inhA*-derived Ct signatures represent genuine biological markers of ethionamide resistance rather than model-specific artifacts. Collectively, these observations suggest that quantitative molecular information generated during routine diagnostic testing extends beyond simple mutation detection by providing informative signals that reflect the relative biological importance of resistance-associated genes. The consistent prioritization of *inhA* across multiple analytical approaches therefore supports the growing role of interpretable artificial intelligence in complementing molecular diagnostics, strengthening resistance surveillance, and facilitating individualized treatment selection in high-burden tuberculosis settings.^{25, 47}

4.4 Fluoroquinolone resistance was primarily associated with *gyrA*-derived molecular signals

Fluoroquinolone resistance was predominantly driven by GYRA3 and GYRA2 Ct-derived markers, consistent with the central role of *gyrA* mutations in resistance to fluoroquinolone agents. DNA gyrase represents the primary target of fluoroquinolones, and mutations occurring within the quinolone resistance-determining region have long been recognised as the principal molecular mechanism responsible for resistance. Previous whole-genome sequencing studies have consistently shown that mutations involving codons 90, 91 and 94 of *gyrA* account for the majority of fluoroquinolone-resistant *Mycobacterium tuberculosis* isolates globally ⁽⁴⁸⁾. The current machine learning analysis independently identified *gyrA*-derived Ct values as the most informative predictors, despite simultaneously evaluating numerous alternative molecular variables. This convergence between computational prediction and established molecular biology provides additional confidence that the identified markers represent biologically meaningful resistance signatures rather than statistical artefacts. The dominance of *gyrA*-derived molecular signals observed in the present study is strongly supported by recent advances in machine learning-based resistance prediction. Evidently, Yang et al. demonstrated that machine learning algorithms consistently identified mutations within established resistance-determining genes as the most influential predictors of phenotypic drug resistance, with *gyrA* contributing substantially to fluoroquinolone resistance classification.²⁶ Similarly, Kouchaki et al. reported that ensemble machine learning approaches effectively modelled complex interactions among genomic variants while preserving the dominant contribution of canonical resistance genes, demonstrating that biologically validated mutations remain the primary drivers of predictive performance despite the inclusion of thousands of genomic variables.²⁸ These findings closely mirror the present results, where *gyrA*-derived Ct measurements overwhelmingly contributed to fluoroquinolone resistance prediction.

Comparable evidence has emerged from more recent deep learning investigations. Wang et al. demonstrated that deep neural network architectures accurately predicted resistance across multiple anti-tuberculosis drugs by recognizing hierarchical mutation patterns dominated by established resistance-associated loci.³⁰ Likewise, Kuang et al. found that convolutional neural networks and conventional machine learning algorithms achieved excellent predictive performance for fluoroquinolone resistance, with *gyrA* consistently ranking among the highest-contributing genomic features regardless of the computational method employed.⁴² These studies collectively indicate that advanced computational models consistently identify *gyrA* as the principal molecular determinant of fluoroquinolone resistance, supporting the biological validity of the present Ct-derived feature-importance rankings.

The findings are further corroborated by studies on explainable artificial intelligence that demonstrate that biologically meaningful resistance mechanisms remain identifiable even within highly complex predictive models. Cen et al. reported that explainable AI successfully attributed prediction outcomes to recognized resistance-conferring mutations, thereby improving confidence in model interpretability and reducing concerns regarding the 'black-box' nature of artificial intelligence.³⁴ Similarly, Paredes-Gutierrez et al. and Xu et al. demonstrated that feature-ranking approaches consistently prioritized well-established resistance-associated mutations, including those involved in fluoroquinolone resistance, indicating that machine learning algorithms can distinguish clinically relevant genetic determinants from background genomic variation.^{35,36} The present findings, therefore, extend this evidence by demonstrating that explainable computational approaches can identify biologically meaningful *gyrA*-associated resistance signatures using routinely generated Ct-derived molecular diagnostic data rather than whole-genome sequencing alone.

These observations are also highly consistent with international guidance on molecular drug-resistance testing. The World Health Organization Catalogue classifies multiple mutations within the quinolone resistance-determining region of *gyrA*, particularly those affecting codons 90, 91 and 94, as high-confidence markers associated with fluoroquinolone resistance⁴⁶. Likewise, the TBnet/RESIST-TB consensus statement recognises *gyrA* mutations as clinically actionable molecular markers that should directly inform individualized treatment decisions because fluoroquinolones remain cornerstone agents within modern multidrug-resistant tuberculosis treatment regimens.^[49] Consequently, the ability of the present machine learning models to prioritise *gyrA*-derived Ct measurements provides further evidence that routinely generated molecular diagnostic information accurately reflects clinically relevant resistance mechanisms and may support more rapid optimization of treatment regimens.

Given the increasing reliance on fluoroquinolones within multidrug-resistant tuberculosis treatment regimens, early identification of *gyrA*-associated resistance remains critical for preventing ineffective therapy and reducing amplification of drug resistance. From a broader translational perspective, the present findings demonstrate that quantitative Ct-derived molecular information contains substantially more predictive value than conventional binary mutation reporting. Rather than simply confirming the presence or absence of predefined mutations, machine learning algorithms quantified the relative contributions of individual *gyrA*-derived molecular signatures to resistance prediction, thereby providing a more nuanced interpretation of molecular diagnostic data. Sharma et al. concluded that one of the major advantages of artificial intelligence in tuberculosis diagnostics lies in its ability to extract clinically meaningful information from complex molecular datasets that would otherwise remain underutilized.²⁵ Similarly, Jamal et al. proposed that artificial intelligence should be viewed as a complementary

analytical framework that can improve the interpretation of molecular resistance mechanisms rather than replace existing diagnostic technologies.⁴⁷ The present study extends these concepts by demonstrating that routinely generated Ct-derived measurements from molecular diagnostic platforms can be computationally leveraged to improve discrimination of fluoroquinolone resistance, offering a practical and potentially cost-effective strategy for strengthening tuberculosis surveillance and precision treatment in high-burden, resource-limited settings.

4.5 *RRS* dominated prediction of injectable drug resistance

Resistance prediction for amikacin, kanamycin and capreomycin was overwhelmingly driven by *RRS* Ct values, with substantially greater importance than all other evaluated markers. This finding aligns closely with the current understanding of aminoglycoside resistance, in which mutations in the 16S ribosomal RNA gene alter antibiotic-binding sites and reduce susceptibility to aminoglycosides.⁵⁰

Previous molecular investigations have demonstrated that *rrs* mutations, particularly within the 1400 region, confer high-level cross-resistance among second-line injectable drugs.⁵¹ The current study extends these observations by demonstrating that Ct-derived measurements from routine molecular testing contain sufficient discriminatory power to accurately identify injectable resistance with excellent predictive performance.

The predominance of *rrs*-derived molecular signals identified in the present study is strongly supported by recent machine learning investigations demonstrating that established resistance-conferring genes consistently account for most of the predictive performance across multiple computational algorithms.⁵² Yang et al. reported that supervised machine learning models accurately classified resistance phenotypes by assigning the greatest predictive importance to well-characterized resistance-associated loci, including genes responsible for second-line drug resistance.²⁶ Similarly, Kouchaki et al. demonstrated that ensemble machine learning algorithms effectively modelled complex genomic interactions while preserving the dominant contribution of canonical resistance determinants, indicating that biologically validated mutations remain the principal drivers of prediction even within highly complex computational models⁽²⁸⁾. These findings closely align with the present study, in which *rrs* Ct-derived measurements consistently emerged as the strongest predictors of injectable drug resistance despite the simultaneous evaluation of numerous molecular variables.

Comparable findings have been reported using deep learning approaches. Wang et al. demonstrated that deep neural networks accurately predicted resistance by learning hierarchical relationships among established resistance-conferring mutations, while Kuang et al. showed that convolutional neural networks and traditional machine learning algorithms consistently recovered recognized resistance genes as the most influential predictive features across multiple anti-tuberculosis drugs.^{30, 42} These observations indicate that computational models repeatedly prioritise biologically meaningful resistance determinants rather than arbitrary statistical associations, providing additional confidence that the prominence of *rrs* observed in the present study reflects genuine molecular mechanisms underlying aminoglycoside resistance.

Recent explainable artificial intelligence studies further strengthen this interpretation. Cen et al. demonstrated that explainable AI frameworks successfully identified established resistance-associated mutations as the principal contributors to model predictions while maintaining transparency in feature attribution.³⁴ Likewise, Xu et al. and Paredes-Gutierrez et al. reported that feature-ranking approaches consistently prioritized recognized antimicrobial resistance-associated genetic variants, illustrating that interpretable machine learning can distinguish biologically important mutations from less informative genomic variation.^{35,36} The convergence between these independent investigations and the present findings suggests that *rrs*-derived Ct signatures represent robust and biologically meaningful predictors of injectable drug resistance rather than model-specific artifacts.

These findings also suggest that the molecular determinants of injectable resistance are comparatively concentrated within a limited number of highly informative loci, explaining why predictive performance for these drugs remained consistently high across multiple machine learning algorithms. The present findings are further supported by international consensus on molecular drug-resistance testing. The World Health Organization Catalog classifies several mutations within the *rrs* gene, particularly those involving the 1400 region, as high-confidence markers associated with resistance to amikacin, kanamycin, and capreomycin.⁴⁶ Similarly, the TBnet/RESIST-TB consensus statement recognises *rrs* mutations as clinically actionable molecular markers that should guide therapeutic decision-making because resistance to second-line injectable drugs substantially limits treatment options for patients with multidrug-resistant and extensively drug-resistant tuberculosis⁴⁹. The ability of the present Ct-derived machine learning models to identify *rrs* as the dominant predictor demonstrates that routinely generated molecular diagnostic data accurately capture clinically relevant resistance mechanisms that support individualized treatment selection.

An important implication of these findings is that resistance to second-line injectable drugs appears to be governed by a relatively restricted molecular architecture compared with several first-line drugs. This biological concentration of resistance within a limited number of highly informative loci likely explains the consistently superior predictive performance achieved by the ensemble learning models observed in the present study. Similar observations have been reported by Chen et al., who demonstrated that although numerous rare genomic variants contribute incrementally to resistance prediction, a relatively small number of high-confidence resistance mutations account for most of the predictive signal across machine learning models.²⁷ This reinforces the concept that interpretable artificial intelligence not only improves

predictive accuracy but also provides important biological insights into the hierarchical organization of drug-resistance mechanisms in *Mycobacterium tuberculosis*.

4.6 Ensemble machine learning outperformed conventional statistical and deep learning approaches

An important methodological finding of this study was the consistent superiority of ensemble tree-based learning algorithms, particularly Random Forest and Extra Trees, over logistic regression, support vector machines, XGBoost, and neural networks. Ensemble methods achieved near-perfect discrimination across most resistance outcomes while simultaneously maintaining high sensitivity and specificity.

Several factors may explain this performance advantage. First, resistance mechanisms arise through complex, non-linear interactions among multiple molecular features that cannot always be adequately captured by traditional regression approaches. Second, Random Forest models are relatively robust to correlated predictors, missing interactions, and heterogeneous feature distributions that frequently characterize molecular diagnostic datasets. Finally, permutation-based feature importance provides clinically interpretable estimates of predictor contribution, allowing biological interpretation alongside predictive modelling. The superior performance of ensemble learning algorithms observed in the present study is consistent with a growing body of evidence demonstrating that tree-based machine learning approaches are particularly well suited for modelling the complex genetic architecture of *Mycobacterium tuberculosis* drug resistance. Yang et al. showed that Random Forest and gradient boosting algorithms effectively captured non-linear relationships between resistance-conferring mutations and phenotypic drug resistance, outperforming conventional statistical approaches while maintaining excellent classification accuracy.²⁶ Similarly, Kouchaki et al. demonstrated that ensemble machine learning techniques successfully modelled complex interactions among genomic variants that could not readily be represented using linear statistical models, highlighting the ability of tree-based algorithms to accommodate the heterogeneous genetic structure characteristic of tuberculosis resistance.²⁸ These findings closely parallel the present results, where ensemble learning consistently achieved the highest predictive performance across multiple resistance phenotypes.

Recent advances in deep learning have similarly demonstrated excellent predictive performance for tuberculosis drug-resistance classification; however, several studies have also shown that deep neural networks do not necessarily outperform ensemble learning when datasets are of moderate size or contain structured molecular features. Wang et al. reported that deep learning models accurately predicted resistance using whole-genome mutation profiles but emphasized that model performance remained highly dependent on dataset size, feature representation, and training complexity.³⁰ Likewise, Kuang et al. observed that although convolutional neural networks achieved high predictive accuracy, conventional machine learning algorithms frequently produced comparable or superior performance for several anti-tuberculosis drugs, particularly when resistance prediction relied on a relatively limited number of highly informative genomic features.⁴² These observations provide a plausible explanation for the present findings, where Random Forest and Extra Trees outperformed neural networks despite the greater computational complexity of deep learning architectures. Interestingly, the neural network demonstrated substantially poorer performance than ensemble learning despite its greater computational complexity. This observation likely reflects the moderate sample size and structured feature space of the present dataset, where simpler ensemble approaches are often better suited than deep learning architectures requiring substantially larger datasets to achieve optimal generalisation.

This interpretation is further supported by recent research in explainable artificial intelligence. Cen et al. demonstrated that although complex artificial intelligence models achieved excellent predictive performance, explainability remained essential for translating computational predictions into clinically meaningful knowledge. By identifying the specific mutations responsible for classifying resistance, explainable artificial intelligence reduced the opacity traditionally associated with deep learning while improving confidence in biological interpretation.³⁴ Similarly, Pal and Mohanty reported that machine learning models successfully identified both established and potentially novel resistance-associated mutations while maintaining high interpretability through feature-ranking approaches.⁵³ These studies reinforce the value of the present analytical framework, in which permutation-based feature importance complements high predictive performance by providing biologically interpretable estimates of individual marker contributions.

The present findings also agree with broader systematic evaluations of artificial intelligence applications in tuberculosis. Sharma et al., in a comprehensive review of machine learning approaches for tuberculosis drug-resistance profiling, concluded that ensemble learning algorithms consistently demonstrate strong predictive performance by modeling complex biological interactions while remaining relatively robust to noisy and heterogeneous datasets.^[25] Likewise, Jamal et al. argued that artificial intelligence should be viewed as an extension of molecular diagnostics capable of improving the interpretation of resistance-associated mutations rather than simply replacing existing analytical approaches.^[47] Collectively, these observations support the integration of interpretable machine learning into routine molecular diagnostic workflows to enhance resistance prediction and clinical decision-making.

An additional strength of the present study lies in the integration of predictive modelling with model interpretability. Many previous machine learning investigations have primarily focused on maximizing classification accuracy using whole-genome sequencing data.^{26, 30, 42} In contrast, the present study combined logistic regression, supervised machine learning, deep learning, multiclass classification, and permutation-based

feature importance to determine not only whether resistance could be accurately predicted but also which Ct-derived molecular markers contributed most strongly to prediction. This distinction is important because clinically useful artificial intelligence should generate biologically interpretable outputs that support therapeutic decision-making rather than function solely as highly accurate black-box prediction models.^{34, 49}

Furthermore, the present study extends the current machine learning literature by demonstrating that highly accurate resistance prediction can be achieved using routinely generated Ct-derived molecular diagnostic information rather than relying exclusively on whole-genome sequencing. Most published artificial intelligence studies have been developed using large genomic datasets generated through sequencing technologies.^{26,27,30,36} Although these approaches have substantially advanced resistance prediction, their implementation remains challenging in many high-burden, resource-limited settings because of financial, computational, and bioinformatics requirements. By contrast, the present findings demonstrate that routinely generated molecular diagnostic data contain sufficient biological information to support highly accurate machine learning-based resistance classification without additional laboratory testing. This observation has important implications for strengthening tuberculosis surveillance and precision treatment in resource-constrained health systems, particularly in high-burden rural settings such as the Eastern Cape.

Implications for molecular tuberculosis diagnostics and precision treatment

The findings have important implications for the molecular diagnosis of tuberculosis in high-burden rural settings. Current molecular platforms are primarily used to determine the presence or absence of predefined resistance-associated mutations. However, the present results suggest that quantitative Ct-derived molecular information also contains valuable predictive signals that may contribute to improved resistance classification. Rather than viewing Ct measurements solely as intermediate analytical outputs, these findings support their integration into computational decision-support systems that can identify high-risk resistance profiles in real time. Such approaches may strengthen laboratory surveillance, improve treatment selection, and facilitate earlier recognition of multidrug-resistant and extensively drug-resistant tuberculosis.

Importantly, this approach requires no additional laboratory testing, as the predictive information is derived entirely from routinely generated molecular diagnostic data. Consequently, computational enhancement of existing diagnostic platforms may be a cost-effective strategy to strengthen tuberculosis management in resource-limited health systems, such as those in the rural Eastern Cape.

4.7 Implications for tuberculosis control in rural Eastern Cape

The present findings contribute to the broader objective of deciphering the genetic drivers of drug-resistant *Mycobacterium tuberculosis* within a rural South African population. Identification of dominant molecular signatures provides valuable epidemiological information regarding the resistance mechanisms currently circulating within Eastern Cape and supports ongoing efforts to strengthen molecular surveillance. From a public health perspective, rapid identification of *katG*-, *inhA*-, *gyrA*-, and *rrs*-associated resistance patterns may facilitate earlier optimisation of treatment regimens, reduce prolonged exposure to ineffective therapy, and ultimately contribute to interruption of resistant strain transmission. Integration of interpretable machine learning into routine molecular surveillance therefore offers a promising framework for supporting precision tuberculosis management while complementing existing laboratory diagnostic workflows.

The potential clinical value of integrating computational approaches with routine molecular diagnostics is increasingly recognised within the tuberculosis literature. Dominguez et al. and the World Health Organization emphasize that molecular drug-resistance testing should extend beyond simple mutation detection towards comprehensive interpretation that supports individualized treatment and precision tuberculosis management.^{46, 49} Additional studies further highlighted that combining molecular diagnostic information with computational analytical methods has the potential to improve the accuracy, consistency and clinical utility of resistance prediction, particularly for complex resistance patterns involving multiple anti-tuberculosis drugs.^[54,55] The present findings support this recommendation by demonstrating that routinely generated Ct-derived molecular measurements contain additional predictive information that can be computationally leveraged without requiring changes to existing laboratory workflows.

The existing artificial intelligence literature on tuberculosis drug resistance has been developed using whole-genome sequencing datasets, which require specialized sequencing infrastructure, advanced bioinformatics expertise, and substantial computational resources.^{26, 30, 36, 53} Although these approaches have demonstrated excellent predictive performance, their widespread implementation remains challenging in many high-burden countries where access to sequencing technologies is limited. However, the present study demonstrates that routinely generated Ct-derived molecular diagnostic data can achieve highly accurate resistance classification through interpretable machine learning without additional laboratory investment, thereby providing a more immediately translatable approach for routine clinical practice.

Recent studies further reinforce the translational importance of interpretable artificial intelligence for tuberculosis diagnostics. Cen et al. emphasised that future artificial intelligence systems should prioritise interpretability alongside predictive performance to enhance clinician confidence, facilitate regulatory acceptance, and support clinical implementation.³⁴ Similarly, Paredes-Gutierrez et al. demonstrated that machine learning models capable of identifying biologically meaningful resistance-associated mutations are likely to provide greater clinical utility than prediction models that provide classification without explanatory insight.^[35] Sharma et al. and Jamal et al. further highlighted the

potential of artificial intelligence to complement molecular diagnostics by improving the interpretation of resistance-associated mutations, informing treatment decisions, and strengthening surveillance of drug-resistant tuberculosis. [25, 47] The present findings extend these recommendations by demonstrating that interpretable machine learning applied to routinely generate Ct-derived molecular data can accurately classify resistance while simultaneously identifying the molecular signatures driving prediction through permutation-based feature importance, thereby providing a practical framework for precision tuberculosis management in resource-limited settings.

An important advantage of this framework is its compatibility with existing tuberculosis control programmes. Rather than requiring new laboratory platforms, machine learning-based decision-support systems could be incorporated through software integration with current molecular diagnostic workflows. This would facilitate implementation within established diagnostic infrastructure while maximizing the clinical value of routinely generated molecular data. Such an approach is particularly relevant for resource-constrained national tuberculosis programmes seeking to expand precision medicine without substantial additional investment, consistent with international recommendations to strengthen molecular resistance testing and optimize individualized management of drug-resistant tuberculosis [46, 49]. It is important, however, to recognise that Ct-derived predictors represent model-ranked molecular markers associated with resistance classification rather than direct causal evidence of resistance mechanisms. Future studies integrating whole-genome sequencing, phenotypic drug susceptibility testing, and prospective clinical outcome data will be valuable for validating these computational findings and further elucidating the biological pathways underlying drug resistance in *Mycobacterium tuberculosis*.

5. Strengths and Limitations

Strengths

This study has several important strengths. First, it is among the earliest studies from rural South Africa to integrate interpretable logistic regression, supervised machine learning, deep learning, multiclass resistance-profile classification, and permutation-based feature-importance analysis within a single analytical framework to identify Ct-derived molecular predictors of drug-resistant *Mycobacterium tuberculosis*. Second, the study utilized a large dataset comprising 2,430 culture-confirmed *M. tuberculosis* specimens generated through routine National Health Laboratory Service diagnostics, providing a robust representation of molecular resistance patterns in a high TB-burden rural setting. Third, by comparing eight supervised learning algorithms with standardized hyperparameter optimization and independent test-set validation, the study provides a comprehensive evaluation of predictive performance and demonstrates the superiority of ensemble learning methods for resistance classification. Fourth, permutation-based feature importance enhanced model interpretability by quantifying the relative contribution of individual molecular targets, allowing the computational findings to be directly linked to established biological mechanisms of drug resistance. Finally, the analytical framework relies entirely on routinely generated molecular diagnostic data, making the approach readily scalable and potentially applicable in resource-limited settings where whole-genome sequencing remains inaccessible.

Limitations

Several limitations should be considered when interpreting these findings. First, the study analyzed Ct-derived molecular measurements generated from targeted molecular diagnostic assays rather than whole-genome sequencing data; consequently, the identified predictors represent model-ranked molecular signatures associated with resistance classification and should not be interpreted as direct evidence of causal resistance mutations. Second, the retrospective laboratory-based design precluded the integration of detailed patient-level clinical information, including HIV status, prior tuberculosis treatment, treatment outcomes, and phenotypic drug susceptibility testing, thereby preventing assessment of the clinical implications of the identified molecular signatures. Third, the analysis was restricted to isolates from rural districts in the Eastern Cape Province, which may limit the generalisability of the findings to urban populations or to regions with different epidemiological characteristics. Fourth, resistance classification depended on predefined molecular targets included in the diagnostic assay; therefore, uncommon, novel, or non-targeted resistance-conferring mutations may not have been detected. Finally, although the machine learning models demonstrated excellent internal performance using an independent hold-out test set, external validation using datasets from other provinces, countries, or molecular diagnostic platforms is required before the models can be considered for routine clinical implementation.

6. Recommendations

1. Enhance the use of routine molecular diagnostic data for resistance surveillance. Health facilities should maximise the use of routinely generated molecular diagnostic data, including Ct-derived measurements, to strengthen laboratory surveillance and enable earlier identification of resistance-associated molecular patterns without requiring additional laboratory infrastructure.
2. Prioritise rapid molecular testing for patients at high risk of drug-resistant tuberculosis. Early access to molecular drug susceptibility testing should be prioritised for patients with previous tuberculosis treatment, treatment failure, relapse, or known exposure to drug-resistant TB to facilitate timely initiation of appropriate therapy and reduce ongoing transmission.

3. Strengthen laboratory capacity through workforce development. Continuous training of laboratory personnel in molecular diagnostics, data interpretation, and quality assurance should be expanded to improve the accuracy, consistency, and utilisation of molecular resistance data within routine tuberculosis programmes.
4. Integrate simple machine learning decision-support tools into existing laboratory information systems. Interpretable predictive models based on routinely collected molecular data should be evaluated for incorporation into existing laboratory information systems to assist in identifying specimens with a high probability of drug resistance and support clinical decision-making without additional diagnostic costs.
5. Undertake prospective multicentre validation studies. Future research should validate these findings across multiple provinces and healthcare settings by integrating routine molecular diagnostics with phenotypic drug susceptibility testing, clinical outcomes, and, where feasible, whole-genome sequencing to confirm the generalisability and clinical utility of the identified molecular predictors.

7. Conclusion

This study demonstrates that routinely generated Ct-derived molecular diagnostic data can be effectively leveraged to identify the principal genetic drivers of resistance to first- and second-line anti-tuberculosis drugs among *Mycobacterium tuberculosis* isolates from rural Eastern Cape, South Africa. By integrating interpretable logistic regression, supervised machine learning, deep learning, multiclass resistance-profile classification, and permutation-based feature importance analysis, the study identified biologically meaningful molecular signatures associated with resistance, with *katG*, *inhA*, *gyrA*, and *rrs* emerging as the dominant Ct-derived predictors for isoniazid, ethionamide, fluoroquinolone, and second-line injectable drug resistance, respectively. The excellent predictive performance achieved by ensemble learning algorithms further demonstrates that routinely generated molecular diagnostic data contain sufficient discriminatory information to accurately classify resistance phenotypes.

Beyond identifying molecular predictors, this study illustrates the potential of combining molecular microbiology with interpretable artificial intelligence to enhance tuberculosis surveillance and support evidence-based clinical decision-making in resource-limited settings. Although external validation and integration with whole-genome sequencing and phenotypic drug susceptibility testing remain necessary, the findings provide a practical framework for strengthening molecular resistance surveillance using existing diagnostic infrastructure. Leveraging routinely collected molecular data through interpretable predictive models may facilitate earlier recognition of drug-resistant tuberculosis, improve treatment optimisation, and contribute to more effective tuberculosis control in high-burden rural settings.

Declarations

Conflict of Interest

The authors declare that the research was conducted without any commercial, financial, or personal relationships that could be construed as a potential conflict of interest.

Generative AI Statement

The authors declare that no generative AI tools or artificial intelligence technologies were used in the writing, data analysis, or preparation of this manuscript.

Publisher's Note

All claims, opinions, and data expressed in this article are solely those of the authors and do not necessarily reflect the official views or policies of their affiliated institutions or the publisher.

Consent for Publication

Not applicable.

Funding

The authors state that no external funding was received for the conduct of this research. This study was supported by internal operational funds from the School of Pathology, and Article Processing Charges (APC) were covered by the School of Public Health at Walter Sisulu University.

Author Contribution

Author Contributions Kamvelihle Sabisa: Conceptualization, Methodology, Validation, Investigation, Writing – Original Draft. Ncomeka Sineke: Conceptualization, Methodology, Validation, Writing – Review & Editing Kelvin Tafadzwa Mpofu: Formal Analysis, Data Curation, Validation, Writing – Review & Editing. Ntandazo Dlatu: Conceptualization, Methodology, Validation, Writing – Review & Editing Teke Apalata: Funding Acquisition, Writing – Review & Editing. Lindiwe Modest Faye: Conceptualization, Methodology, Supervision, Writing – Review & Editing, Final Approval. All authors have read and approved the final version of the manuscript.

Acknowledgments

The authors express their sincere gratitude to the Honors and Master's students at the School of Pathology and Laboratory Medicine, CSIR, as well as the Walter Sisulu University TB Research Group, for their valuable assistance with data collection. We also thank all colleagues and collaborators for their ongoing support and contributions, which facilitated the completion of this study.

Data Availability

Data Availability Statement To protect patient privacy and comply with national data protection regulations, individual-level health records used in this study are restricted by the Walter Sisulu University Research Ethics and Biosafety Committee and the Eastern Cape Department of Health. De-identified data may be made available to qualified researchers upon reasonable request to the corresponding author. Data release is contingent upon evaluation by the relevant ethics review boards, execution of a standard data-sharing agreement, and adherence to established legal and ethical frameworks.

References

1. World Health Organization. Global tuberculosis report 2020: executive summary. (2020).
2. Floyd, K., Glaziou, P., Zumla, A. & Raviglione, M. The global tuberculosis epidemic and progress in care, prevention, and research: an overview in year 3 of the End TB era. *Lancet Respiratory Med.* **6** (4), 299–314 (2018).
3. World Health Organization. *WHO consolidated guidelines on tuberculosis. Module 6: tuberculosis and comorbidities* (World Health Organization, 2024).
4. Naidoo, K. et al. The epidemiology, transmission, diagnosis, and management of drug-resistant tuberculosis—lessons from the South African experience. *Lancet. Infect. Dis.* **24** (9), e559–e575 (2024).
5. Cox, H. et al. Drug-resistant tuberculosis in South Africa: history, progress and opportunities for achieving universal access to diagnosis and effective treatment. *South. Afr. Health Rev.* **2017** (1), 157–167 (2017).
6. Faye, L. M., Hosu, M. C. & Apalata, T. Drug-resistant tuberculosis in rural Eastern Cape, South Africa: a study of patients' characteristics in selected healthcare facilities. *International Journal of Environmental Research and Public Health*, 21(12), p.1594. (2024).
7. Mokhehi, N. P. Exploring barriers to TB contact tracing and screening: health service providers' perspective (Doctoral dissertation, University of the Western Cape). (2023).
8. Hosu, M. C., Tsuro, U., Dlatu, N., Faye, L. M. & Apalata, T. August. Strengthening clinical governance and public health interventions to improve drug-resistant tuberculosis outcomes in rural South Africa. In *Healthcare* (Vol. 13, No. 17, p. 2093). MDPI. (2025).
9. Gygli, S. M., Borrell, S., Trauner, A. & Gagneux, S. Antimicrobial resistance in *Mycobacterium tuberculosis*: mechanistic and evolutionary perspectives. *FEMS Microbiol. Rev.* **41** (3), 354–373 (2017).
10. Hameed, H. A. et al. Molecular targets related drug resistance mechanisms in MDR-, XDR-, and TDR-*Mycobacterium tuberculosis* strains. *Front. Cell. Infect. Microbiol.* **8**, 114 (2018).
11. Omoteso, O. A., Fadaka, A. O., Walker, R. B. & Khamanga, S. M. Innovative strategies for combating multidrug-resistant tuberculosis: advances in drug delivery systems and treatment. *Microorganisms*, 13(4), p.722. (2025).
12. Gagneux, S. Ecology and evolution of *Mycobacterium tuberculosis*. *Nat. Rev. Microbiol.* **16** (4), 202–213 (2018).
13. Gopalaswamy, R. & Subbian, S. *The power of resistance: mechanisms of antimicrobial resistance in Mycobacterium tuberculosis and its impact on tuberculosis management* pp.e00194–e00125 (Clinical Microbiology Reviews, 2026).
14. Dookie, N., Rambaran, S., Padayatchi, N., Mahomed, S. & Naidoo, K. Evolution of drug resistance in *Mycobacterium tuberculosis*: a review on the molecular determinants of resistance and implications for personalized care. *J. Antimicrob. Chemother.* **73** (5), 1138–1151 (2018).
15. Sheikh, B. A. et al. Development of new therapeutics to meet the current challenge of drug resistant tuberculosis. *Curr. Pharm. Biotechnol.* **22** (4), 480–500 (2021).

16. Deshpande, A., Likhar, R., Khan, T. & Omri, A. Decoding drug resistance in *Mycobacterium tuberculosis* complex: genetic insights and future challenges. *Expert Rev. Anti-infective Therapy*. **22** (7), 511–527 (2024).
17. Bartolomeu-Gonçalves, G. et al. Tuberculosis diagnosis: Current, ongoing, and future approaches. *Diseases*, 12(9), p.202. (2024).
18. Wu, X. et al. Targeted next-generation sequencing-a promising approach in the diagnosis of *Mycobacterium tuberculosis* and drug resistance. *Infection* **53** (3), 967–979 (2025).
19. Pillay, S. et al. and G. Theron. Xpert MTB/XDR for detection of pulmonary tuberculosis and resistance to isoniazid, fluoroquinolones, ethionamide, and amikacin (Protocol). status and date: New, published in 6 (2021).
20. Martin-Higuera, M. C., Rivas, G., Rolo, M., Muñoz-Gallego, I. & Lopez-Roa, P. Xpert MTB/RIF Ultra CT value provides a rapid measure of sputum bacillary burden and predicts smear status in patients with pulmonary tuberculosis. *Scientific Reports*, 13(1), p.1591. (2023).
21. Lange, B. et al. Diagnostic accuracy of the Xpert® MTB/RIF cycle threshold level to predict smear positivity: a meta-analysis. *Int. J. Tuberc Lung Dis*. **21** (5), 493–502 (2017).
22. Singh, M. et al. Evolution of machine learning in tuberculosis diagnosis: a review of deep learning-based medical applications. *Electronics*, 11(17), p.2634. (2022).
23. Kavvas, E. S. et al. Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance. *Nature communications*, 9(1), p.4306. (2018).
24. Ezewudo, M. et al. Integrating standardized whole genome sequence analysis with a global *Mycobacterium tuberculosis* antibiotic resistance knowledgebase. *Scientific reports*, 8(1), p.15382. (2018).
25. Sharma, A., Machado, E., Lima, K. V. B., Suffys, P. N. & Conceição, E. C. Tuberculosis drug resistance profiling based on machine learning: A literature review. *The Brazilian Journal of Infectious Diseases*, 26(1), p.102332. (2022).
26. Yang, Y. et al. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics* **34** (10), 1666–1671 (2018).
27. Chen, M. L. et al. *Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction* Vol. 43, pp.356–369 (EBioMedicine, 2019).
28. Kouchaki, S. et al. CRyPTIC Consortium and Clifton, D.A., Application of machine learning techniques to tuberculosis drug resistance analysis. *Bioinformatics*, 35(13), pp.2276–2282. (2019).
29. Alqudah, A. M. & Moussavi, Z. Bridging Signal Intelligence and Clinical Insight: A Comprehensive Review of Feature Engineering, Model Interpretability, and Machine Learning in Biomedical Signal Analysis. *Applied Sciences*, 15(22), p.12036. (2025).
30. Wang, Y. et al. TB-DROP: deep learning-based drug resistance prediction of *Mycobacterium tuberculosis* utilizing whole genome mutations. *BMC genomics*, 25(1), p.167. (2024).
31. Saliba, J. G. et al. Enhanced diagnosis of multi-drug-resistant microbes using group association modeling and machine learning. *Nature Communications*, 16(1), p.2933. (2025).
32. Li, L., Liu, H., Lei, Q. & Li, T. From clinical phenotypes to genomic signatures: machine learning integration for precision tuberculosis treatment prediction. *Frontiers in Bioinformatics*, 6, p.1787360. (2026).
33. Arulsamy, S. & Sharma, S. Machine learning in tuberculosis: Advancements in diagnostics, drug resistance prediction, and prognosis. *Disease Prev. Epidemiol.* **1** (1), 48–62 (2026).
34. Cen, H., Zhang, P., Ling, Y., Zhao, G. & Zhang, G. An explainable artificial intelligence framework reveals mutations associated with drug resistance in *Mycobacterium tuberculosis*. *Biosaf. Health*. **7** (06), 384–394 (2025).
35. Paredes-Gutierrez, G. et al. Predicting drug resistance in *mycobacterium tuberculosis*: A machine learning approach to genomic mutation analysis. *Diagnostics*, 15(3), p.279. (2025).
36. Xu, Y. et al. Machine learning-based prediction of antimicrobial resistance and identification of AMR-related SNPs in *Mycobacterium tuberculosis*. *BMC Genomic Data*, 26(1), p.48. (2025).
37. Valafar, S. J. Systematic review of mutations associated with isoniazid resistance points to continuing evolution and subsequent evasion of molecular detection, and potential for emergence of multidrug resistance in clinical strains of *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* **65** (3), 10–1128 (2021).
38. Napier, G., Campino, S., Phelan, J. E. & Clark, T. G. Large-scale genomic analysis of *Mycobacterium tuberculosis* reveals extent of target and compensatory mutations linked to multi-drug resistant tuberculosis. *Scientific Reports*, 13(1), p.623. (2023).
39. Bodaghi, A. B. et al. Mechanisms of Antibiotic Resistance and Novel Therapeutic Approaches for *Mycobacterium tuberculosis*: A Narrative Review With a Focus on Tuberculosis Mutations in Iran. *Health Sci. Rep.* **8** (10), e71313 (2025).
40. Faye, L. M. et al. The detection of mutations and genotyping of drug-resistant *Mycobacterium tuberculosis* strains isolated from patients in the Rural Eastern Cape Province. *Infect. disease Rep.* **15** (4), 403–416 (2023).

41. Singh, A. K., Singh, N., Kumar, S., Mishra, A. K. & Singh, N. P. Molecular insights of drug-resistant tuberculosis: genetic mutations and their profile. *Frontiers in Microbiology*, 16, p.1669327. (2025).
42. Kuang, X., Wang, F., Hernandez, K. M., Zhang, Z. & Grossman, R. L. Accurate and rapid prediction of tuberculosis drug resistance from genome sequence data using traditional machine learning algorithms and CNN. *Scientific reports*, 12(1), p.2427. (2022).
43. Patel, K. et al. Predicting isoniazid resistance in *Mycobacterium tuberculosis* complex in New York State using whole-genome sequencing. *Journal of Clinical Microbiology*, 64(4), pp.e01609-25. (2026).
44. Diab, A., Dickerson, H. & Al Musaimi, O. Targeting the heart of *Mycobacterium*: advances in anti-tubercular agents disrupting cell wall biosynthesis. *Pharmaceuticals*, 18(1), p.70. (2025).
45. Pruthi, S. S. et al. Leveraging large-scale *Mycobacterium tuberculosis* whole genome sequence data to characterise drug-resistant mutations using machine learning and statistical approaches. *Scientific reports*, 14(1), p.27091. (2024).
46. World Health Organization. *Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance* (World Health Organization, 2023).
47. Jamal, S. et al. Artificial Intelligence and Machine learning based prediction of resistant and susceptible mutations in *Mycobacterium tuberculosis*. *Scientific reports*, 10(1), p.5487. (2020).
48. Nehru, V. J. et al. Risk assessment and transmission of fluoroquinolone resistance in drug-resistant pulmonary tuberculosis: a retrospective genomic epidemiology study. *Scientific Reports*, 14(1), p.19719. (2024).
49. Domínguez, J. et al. Clinical implications of molecular drug resistance testing for *Mycobacterium tuberculosis*: a 2023 TBnet/RESIST-TB consensus statement. *Lancet. Infect. Dis.* **23** (4), e122–e137 (2023).
50. Deshpande, A., Likhar, R., Khan, T. & Abdelwahab Omri. Decoding drug resistance in *Mycobacterium tuberculosis* complex: genetic insights and future challenges. *Expert Rev. Anti-infective Therapy*. **22** (7), 511–527 (2024).
51. Roshdi Maleki Mehdi. A Comprehensive Review of Genetic Mechanisms of *Mycobacterium tuberculosis* Resistance to Injectable Agents, Second-Line Drugs, and Novel Therapeutics. *Infect. Drug Resist.* : 1–22. (2026).
52. Serajian, M., Lotfollahi, M., Green, O., Smith, K. & Marini, S. Mattia Prosperi, and Christina Boucher. Scalable machine learning improves resistance prediction and identifies novel determinants in *Mycobacterium tuberculosis*. *bioRxiv* : 2026–2004. (2026).
53. Pal, A. & Mohanty, D. Machine learning-based approach for identification of new resistance associated mutations from whole genome sequences of *Mycobacterium tuberculosis*. *Bioinf. Adv.* **5** (1), 050 (2025). .vba
54. Merchant, S. A., Shaikh, M. J. S. & Nadkarni, P. Tuberculosis conundrum-current and future scenarios: a proposed comprehensive approach combining laboratory, imaging, and computing advances. *World Journal of Radiology*, 14(6), p.114. (2022).
55. Akwaronwu, B. G., Ayankoya, F. Y., Kuyoro, S. O., Akwaronwu, I. U. & Adeniyi, O. J. Machine Learning for Predicting Drug Resistance in Tuberculosis: A Systematic Review of Model Performance and Clinical Applications. *Big Data Min. Analytics.* **9** (1), 314–337 (2025).

Figures

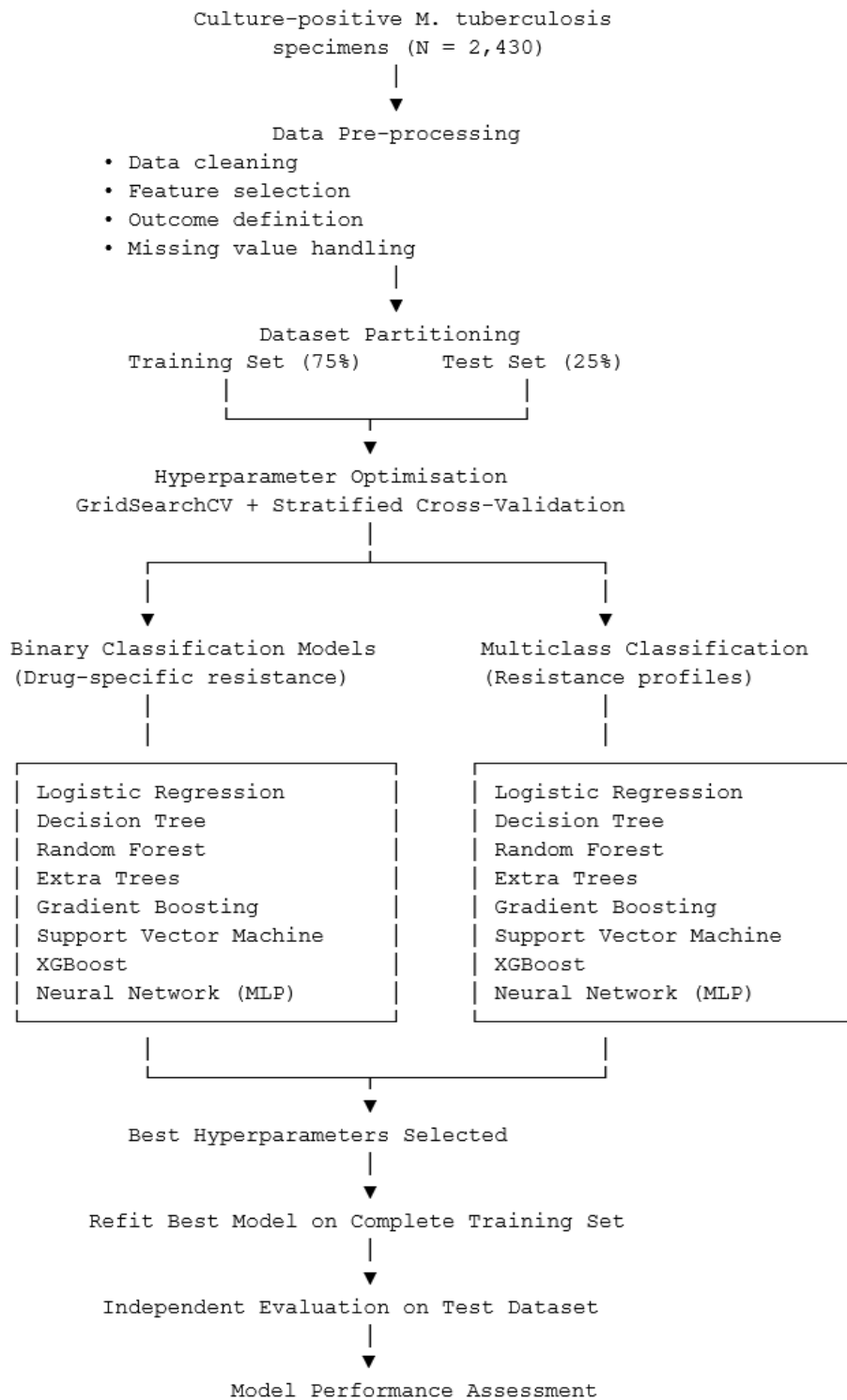


Figure 1

Machine-learning model development and evaluation workflow for predicting drug-resistant *Mycobacterium tuberculosis*.

Fig. 2

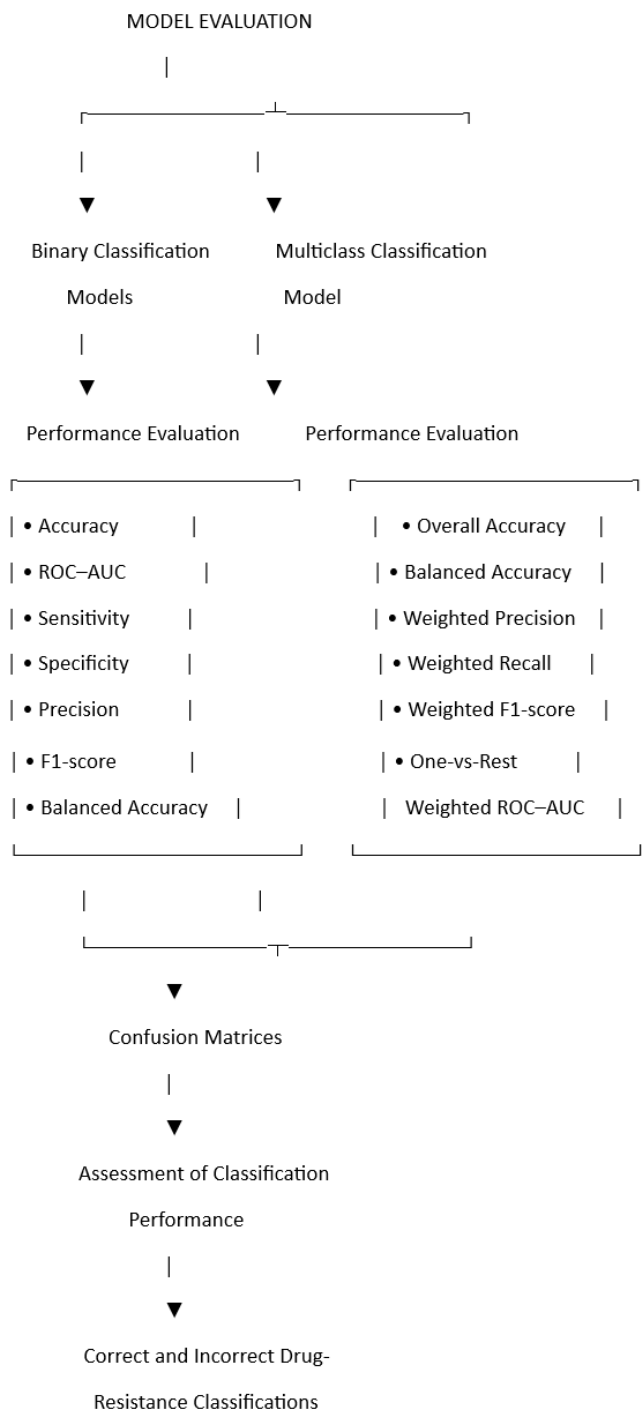


Figure 2
Legend not included with this version.