Exploring the Role of Transposons in Predicting Antimicrobial Resistance in Tuberculosis and Its Co-infections Using Explainable Machine Learning

Joana S. Tria[1], Zachary B. Lara[2], Geoffrey A. Solano[1]
[1]Department of Physical Sciences and Mathematics, College of Arts and Sciences
University of the Philippines Manila
[2] Microbiology Division, Institute of Biological Sciences
University of the Philippines Los Banos
{jstria1,zblara,gasolano}@up.edu.ph

***Abstract*—Antimicrobial resistance (AMR) presents a critical global health challenge, especially in treating tuberculosis (TB) and its co-infections like *Klebsiella pneumoniae* and *Staphylococcus aureus*. Conventional antimicrobial susceptibility testing (AST) is accurate but time-consuming and often inaccessible. This study proposes a genome-based machine learning (ML) framework that integrates AMR gene and transposon detection to improve resistance prediction. Whole-genome sequences (WGS) from NCBI were analyzed using ABRicate for AMR genes and TnComp_finder for transposons. Engineered features focused on AMR gene–transposon co-occurrence. Five ML models—Logistic Regression, Random Forest, XGBoost, AdaBoost, and Support Vector Machine—were trained with and without SMOTE oversampling with evaluation metrics such as accuracy, precision, recall, ROC AUC, and F2-score. Top models achieved AUC $>$ 0.85 and F2 $>$ 0.80 for key antibiotics. Feature importance analysis revealed critical AMR-transposon interactions driving resistance. This framework demonstrates the predictive power of transposon-informed features and offers a scalable, interpretable solution for clinical decision support in antimicrobial therapy.**

***Index Terms*—Antimicrobial resistance, transposons, whole-genome sequencing, explainable machine learning**

I. Introduction

The rise of antimicrobial resistance (AMR) presents a significant global health threat, with the World Health Organization warning that annual deaths from drug-resistant infections could reach 10 million by 2050 [1], [2]. Tuberculosis (TB), particularly in its multidrug-resistant form (MDR-TB), remains a major contributor to this crisis. Co-infections with pathogens such as *Klebsiella pneumoniae* and *Staphylococcus aureus* further complicate treatment and increase the burden on healthcare systems, particularly in low- and middle-income countries [3].

Traditional antimicrobial susceptibility testing (AST), while accurate, is slow and resource-intensive, often delaying appropriate treatment decisions [4], [5]. In contrast, whole-genome sequencing (WGS) combined with machine learning (ML) offers a rapid, genome-based approach to AMR prediction, leveraging genomic data to detect resistance determinants with greater efficiency [6].

Recent studies highlight the role of mobile genetic elements (MGEs), such as plasmids and transposons, in the spread of AMR genes (ARGs) [7], [8]. Transposons, in particular, facilitate horizontal gene transfer and may act as indicators of resistance patterns. However, few existing ML approaches integrate transposon data into AMR prediction pipelines.

This work proposes a novel ML-based framework that incorporates both ARGs and transposons from WGS data to predict resistance to key antibiotics, including isoniazid, rifampicin, oxacillin, ciprofloxacin, and linezolid. The system is trained on curated bacterial isolates from NCBI, using algorithms such as logistic regression, random forest, support vector machines, and boosting techniques. The models are evaluated using standard metrics (accuracy, precision, recall, F2-score, AUROC), with feature importance analyses applied for interpretability.

To bridge the gap between research and clinical utility, we present a web-based application that enables users to upload FASTA-formatted WGS data, visualize predicted antibiograms, and explore associated AMR genes and transposons. This tool supports faster, cost-effective decision-making for AMR detection in TB and co-infections, aiming to complement conventional AST and improve treatment outcomes.

II. Related Work

Recent advancements in machine learning (ML) and genomic technologies have significantly influenced the landscape of antimicrobial resistance (AMR) detection and prediction. Traditional antimicrobial susceptibility testing (AST) remains the clinical gold standard, but it is time-consuming and often delayed, especially for critical infections like tuberculosis (TB). Consequently, researchers have explored ML as a promising alternative to enable faster and more accurate resistance profiling.

Anahtar et al. [4] presented a foundational framework for applying ML to AMR, highlighting three key domains: prediction of resistance from genomic data, discovery of novel AMR mechanisms, and support for antibiotic stewardship through electronic health record (EHR) analysis. Their work demonstrated how ML could achieve accurate AST predictions from pathogen sequences, discover resistance-driving mutations, and improve treatment decision-making through EHR-based models. However, real-world EHR integration remains underutilized due to data quality concerns.

Complementing this framework, Sakagianni et al. [6] conducted a comprehensive literature review analyzing 29 studies that employed ML for AMR prediction in clinical settings. The review emphasized the central role of whole-genome sequencing (WGS) in detecting resistance genes and mutations, which ML algorithms such as logistic regression (LR), random forests (RF), support vector machines (SVM), and gradient boosting models leveraged to uncover predictive patterns. These models effectively predicted AMR using structured genomic features, though variations in encoding strategies and clinical data integration introduced model performance differences.

One study that exemplifies the integration of clinical and genomic data is Babirye et al. [9], which developed ML models to predict resistance to four anti-TB drugs using 182 *Mycobacterium tuberculosis* isolates from Uganda. By incorporating both single-nucleotide polymorphisms (SNPs) and patient metadata (e.g., HIV status), the models—especially LR, XGBoost, and gradient boosting—achieved drug-specific predictive accuracy and identified biologically significant resistance markers.

Ren et al. [10] further explored encoding techniques to process raw genomic data for ML-based AMR prediction in *Escherichia coli*. They compared label encoding, one-hot encoding, and Frequency Matrix Chaos Game Representation (FCGR), showing how different representations impact model input and performance. These encodings transformed categorical nucleotide sequences into numerical forms amenable to learning, with one-hot encoding enabling better model interpretability and avoiding ordinal biases.

Other studies also extended ML applications beyond binary classification by predicting minimum inhibitory concentrations (MICs). Gröschel et al. [11] introduced GenTB, a web-based tool combining RF and wide and deep neural networks (WDNN) to predict resistance to 13 anti-TB drugs from raw Illumina sequencing data. It achieved high area under the curve (AUC) values above 91% for first-line drugs like rifampicin and isoniazid. Similarly, Yasir et al. [12] employed unitig-based features and ML models such as RF and CATBoost to predict MICs for *Neisseria gonorrhoeae*, achieving R^2 values up to 0.79.

Finally, researchers have emphasized the importance of mobile genetic elements such as transposons in spreading resistance. Lipszyc et al. [13] demonstrated that transposons can activate silent resistance genes by introducing new promoters. Khezri et al. [14] further revealed that transposon-bearing plasmids often harbor AMR genes, enhancing their horizontal transfer. The use of tools like ResFinder, PlasmidFinder, and hybrid assembly strategies allowed for more precise detection of such mobile genetic elements, improving ARG annotation and supporting ML model inputs.

The integration of ML techniques with genomic data has paved the way for rapid and accurate AMR prediction. Emerging studies are now moving beyond basic classification tasks by incorporating clinical metadata, encoding innovations, and mobile element detection to enhance predictive modeling and resistance mechanism elucidation.

III. Methodology

A. Dataset

The dataset used in this study comprises publicly available whole-genome sequencing data of Mycobacterium tuberculosis, Klebsiella pneumoniae, and Staphylococcus aureus isolates from the NCBI BioSample database. These pathogens are clinically significant due to their involvement in tuberculosis (TB) and common co-infections. As of December 2023, over 612,000 new and relapse TB cases were reported in the Philippines [15], with TB remaining a leading cause of death from a single infectious agent [16]. The study focuses on five key antibiotics: isoniazid, rifampicin, oxacillin, ciprofloxacin, and linezolid, which are widely used to treat infections caused by these bacteria.

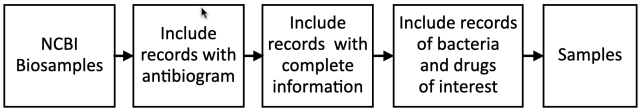
NCBI Biosamples
Include records with antibiogram
Include records with complete information
Include records of bacteria and drugs of interest
Samples

Fig. 1. Data collection.

Figure 1 illustrates the process of data collection from the NCBI BioSample database, where bacterial samples with complete antibiogram are retrieved for analysis. Intermediate susceptibility results were reclassified as resistant to enable binary classification. Final sample counts included 225 *M. tuberculosis*, 125 *S. aureus*, and 125 *K. pneumoniae* isolates. Not all isolates were tested against every antibiotic, and some antibiotics were species-specific (e.g., isoniazid for *M. tuberculosis*). A detailed breakdown is shown in Table I.

TABLE I
Summary of the samples.

Antibiotic	Resistant	Susceptible	Total
Isoniazid	165	56	221
Rifampicin	204	144	348
Oxacillin	76	56	132
Ciprofloxacin	112	102	214
Linezolid	40	291	331

B. Data Preprocessing

Data retrieved from NCBI underwent preprocessing using TnComp_finder for transposon detection and ABRicate for AMR gene identification. The resulting features were used to predict antibiotic resistance, with the binary antibiogram values (0 = susceptible, 1 = resistant) serving as the target variable. Samples lacking target labels were excluded to maintain analytical integrity.

AMR gene features were encoded using MultiLabelBinarizer to handle multiple gene entries per sample. The dataset was then split into training and testing sets (75:25 ratio) to facilitate model evaluation. To capture the potential role of mobile genetic elements in resistance, co-occurrence features were engineered by pairing each transposon with every AMR gene per sample. These pairs were weighted by the inverse square root of their frequency to reduce bias from common elements, and averaged to produce the TFxAMR_Encoded feature—representing meaningful gene-mobility interactions.

To mitigate class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied, generating synthetic instances of underrepresented resistance cases. This balanced dataset enabled more robust modeling of the relationships among transposons, AMR genes, and resistance phenotypes.

C. Model Implementation

This study employs five machine learning models—Logistic Regression, XGBoost, AdaBoost, SVM, and Random Forest—with and without SMOTE to evaluate their predictive performance for antibiotic resistance.

Model inputs include transposon profiles from TnComp_finder and AMR genes from ABRicate. Data preprocessing involves cleaning, encoding, handling missing values, and a 75:25 train-test split. Models are initially trained with default hyperparameters, followed by tuning to optimize binary classification of resistance (0 = susceptible, 1 = resistant).

Performance is assessed using accuracy, precision, recall, f2-score, and ROC-AUC. Feature importance analysis identifies key transposons and AMR genes contributing to resistance predictions using model-specific importance methods.

Top-performing models based on AUC and other metrics are integrated into a web-based application for real-time AMR prediction. The study also explores interactions between transposons and AMR genes to understand their combined role in resistance mechanisms.

IV. Results

A. Exploratory Data Analysis

Analysis of 475 samples using TnComp_finder revealed transposons in 431 samples. Table II summarizes their distribution and association with resistance. For antibiotics such as isoniazid, rifampicin, oxacillin, and ciprofloxacin, transposon-positive isolates exhibited markedly higher resistance rates. Notably, oxacillin resistance was observed exclusively in transposon-positive samples. In contrast, linezolid resistance was present in both transposon-positive and -negative samples, indicating variable transposon influence by antibiotic.

Using ABRicate, 303 unique AMR genes were detected across all samples. While many were rare, ten genes dominated the dataset, with *rpoB2*, *ColRNAI_1*, and *Col440I_1* among the most prevalent. The same genes remained frequent among transposon-positive isolates, suggesting these resistance determinants are commonly mobilized via transposons.

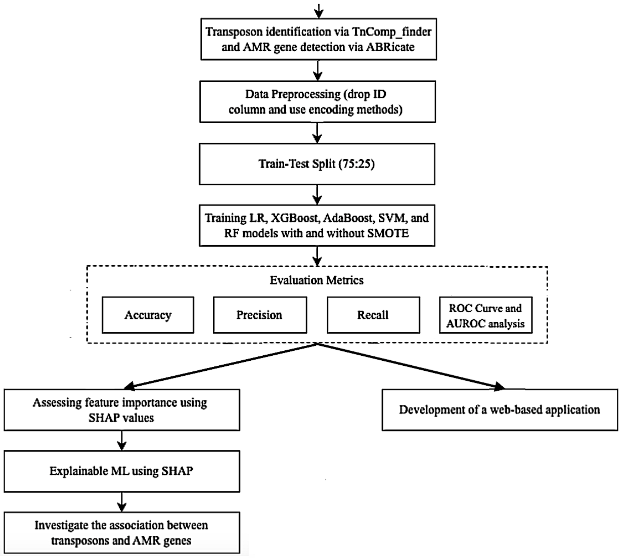
Transposon identification via TnComp_finder and AMR gene detection via ABRicate
Data Preprocessing (drop ID column and use encoding methods)
Train-Test Split (75:25)
Training LR, XGBoost, AdaBoost, SVM, and RF models with and without SMOTE
Evaluation Metrics
Accuracy
Precision
Recall
ROC Curve and AUROC analysis
Assessing feature importance using SHAP values
Development of a web-based application
Explainable ML using SHAP
Investigate the association between transposons and AMR genes

Fig. 2. Workflow of the study.

TABLE II
Summary of the samples with respect to transposons

Antibiotic	Resistant		Susceptible	
	With Transposons	Without Transposons	With Transposons	Without Transposons
Isoniazid	142	23	37	19
Rifampicin	174	30	132	12
Oxacillin	76	0	56	0
Ciprofloxacin	110	2	102	0
Linezolid	39	1	250	41

B. Model Performance

Model performance was primarily evaluated using ROC AUC due to its robustness against class imbalance and its ability to assess discrimination across thresholds. Precision, recall, and F2-score were also considered to reflect clinical priorities: high recall minimizes missed resistant cases, while high precision reduces unnecessary treatments. Accuracy was deemphasized due to dataset imbalance.

Initial evaluations used default model parameters, followed by tuning with GridSearchCV. All models achieved ROC AUC ≥ 0.75, with rifampicin (0.89), oxacillin (0.88), and linezolid (0.88) performing best.

Precision and recall supported selection of clinically useful models. Isoniazid and rifampicin showed strong recall (0.93 and 1.00) with precision > 0.80. Oxacillin and ciprofloxacin performed well (approximately 0.85 for both metrics). Linezolid had high precision (0.88) but lower recall (0.70), indicating a conservative prediction pattern.

The selected deployment models were: Random Forest + SMOTE (isoniazid), AdaBoost + SMOTE (rifampicin), Random Forest (oxacillin), XGBoost + SMOTE (ciprofloxacin), and AdaBoost (linezolid), chosen for their balance of performance and clinical relevance.

TABLE III
Isoniazid Base Model Performance Without SMOTE.

	Accuracy	Precision	Recall	ROC	F2-Score
Logistic Regression	0.68	0.74	0.88	0.70	0.85
Random Forest	0.71	0.77	0.88	0.69	0.86
XGBoost	0.64	0.73	0.83	0.59	0.81
AdaBoost	0.75	0.75	1.00	0.62	0.94
SVM	0.75	0.75	1.00	0.51	0.94

TABLE IV
Isoniazid Base Model Performance with SMOTE.

	Accuracy	Precision	Recall	ROC	F2-Score
Logistic Regression	0.70	0.82	0.76	0.67	0.77
Random Forest	0.70	0.84	0.74	0.73	0.76
XGBoost	0.61	0.81	0.62	0.69	0.65
AdaBoost	0.75	0.85	0.81	0.72	0.82
SVM	0.77	0.82	0.88	0.61	0.87

TABLE V
Isoniazid without SMOTE: Hyperparameter-Tuned Model Performance.

	Accuracy	Precision	Recall	ROC	F2-Score
Logistic Regression	0.75	0.75	1.00	0.75	0.94
Random Forest	0.75	0.75	1.00	0.74	0.94
XGBoost	0.75	0.75	1.00	0.65	0.94
AdaBoost	0.75	0.75	1.00	0.64	0.94
SVM	0.75	0.75	1.00	0.68	0.94

TABLE VI
Isoniazid with SMOTE: Hyperparameter-Tuned Model Performance.

	Accuracy	Precision	Recall	ROC	F2-Score
Logistic Regression	0.75	0.75	1.00	0.75	0.94
Random Forest	**0.80**	**0.83**	**0.93**	**0.75**	**0.91**
XGBoost	0.75	0.75	1.00	0.67	0.94
AdaBoost	0.75	0.75	1.00	0.56	0.94
SVM	0.77	0.82	0.88	0.71	0.87

TABLE VII
Rifampicin Base Model Performance Without SMOTE.

	Accuracy	Precision	Recall	ROC	F2-Score
Logistic Regression	0.87	0.83	0.98	0.88	0.95
Random Forest	0.82	0.80	0.92	0.89	0.89
XGBoost	0.82	0.82	0.88	0.82	0.87
AdaBoost	0.84	0.78	1.00	0.87	0.95
SVM	0.84	0.79	0.98	0.84	0.94

TABLE VIII
Rifampicin Base Model Performance With SMOTE.

	Accuracy	Precision	Recall	ROC	F2-Score
Logistic Regression	0.85	0.84	0.92	0.89	0.90
Random Forest	0.79	0.79	0.88	0.88	0.86
XGBoost	0.78	0.81	0.82	0.83	0.82
AdaBoost	0.87	0.82	1.00	0.87	0.96
SVM	0.85	0.81	0.98	0.85	0.94

TABLE IX
Rifampicin without SMOTE: Hyperparameter-Tuned Model Performance.

	Accuracy	Precision	Recall	ROC	F2-Score
Logistic Regression	0.89	0.84	1.00	0.87	0.96
Random Forest	0.85	0.81	0.98	0.90	0.94
XGBoost	0.84	0.80	0.96	0.84	0.93
AdaBoost	0.87	0.82	1.00	0.89	0.96
SVM	0.80	0.77	0.96	0.84	0.91

TABLE X
Rifampicin with SMOTE: Hyperparameter-Tuned Model Performance.

	Accuracy	Precision	Recall	ROC	F2-Score
Logistic Regression	0.89	0.84	1.00	0.88	0.96
Random Forest	0.87	0.83	0.98	0.87	0.95
XGBoost	0.83	0.80	0.94	0.86	0.91
AdaBoost	**0.89**	**0.84**	**1.00**	**0.89**	**0.96**
SVM	0.83	0.80	0.94	0.84	0.91

TABLE XI
Oxacillin Base Model Performance Without SMOTE.

	Accuracy	Precision	Recall	ROC	F2-Score
Logistic Regression	0.76	0.87	0.68	0.79	0.71
Random Forest	0.82	0.88	0.79	0.91	0.81
XGBoost	0.61	0.69	0.58	0.76	0.60
AdaBoost	0.73	0.73	0.84	0.74	0.82
SVM	0.73	0.71	0.89	0.83	0.85

TABLE XII
Oxacillin Base Model Performance With SMOTE.

	Accuracy	Precision	Recall	ROC	F2-Score
Logistic Regression	0.76	0.87	0.68	0.78	0.71
Random Forest	0.82	0.88	0.79	0.93	0.81
XGBoost	0.73	0.86	0.63	0.80	0.67
AdaBoost	0.73	0.92	0.58	0.80	0.63
SVM	0.79	0.83	0.79	0.86	0.80

TABLE XIII
Oxacillin without SMOTE: Hyperparameter-Tuned Model Performance.

	Accuracy	Precision	Recall	ROC	F2-Score
Logistic Regression	0.79	0.93	0.68	0.76	0.72
Random Forest	**0.85**	**0.89**	**0.84**	**0.88**	**0.85**
XGBoost	0.76	0.74	0.89	0.83	0.86
AdaBoost	0.73	0.73	0.84	0.79	0.82
SVM	0.76	0.74	0.89	0.85	0.86

TABLE XIV
Oxacillin with SMOTE: Hyperparameter-Tuned Model Performance.

	Accuracy	**Precision**	**Recall**	**ROC**	**F2-Score**
Logistic Regression	0.79	0.93	0.68	0.77	0.72
Random Forest	0.85	0.89	0.84	0.87	0.85
XGBoost	0.79	0.93	0.68	0.87	0.72
AdaBoost	0.76	0.92	0.63	0.80	0.67
SVM	0.82	0.88	0.79	0.87	0.81

TABLE XV
Ciprofloxacin Base Model Performance Without SMOTE.

	Accuracy	**Precision**	**Recall**	**ROC**	**F2-Score**
Logistic Regression	0.70	0.71	0.71	0.71	0.71
Random Forest	0.67	0.67	0.71	0.74	0.71
XGBoost	0.65	0.66	0.68	0.70	0.67
AdaBoost	0.70	0.71	0.71	0.77	0.71
SVM	0.72	0.70	0.82	0.76	0.79

TABLE XVI
Ciprofloxacin Base Model Performance With SMOTE.

	Accuracy	**Precision**	**Recall**	**ROC**	**F2-Score**
Logistic Regression	0.70	0.71	0.71	0.72	0.71
Random Forest	0.67	0.67	0.71	0.75	0.70
XGBoost	0.65	0.66	0.68	0.70	0.67
AdaBoost	0.74	0.73	0.79	0.78	0.77
SVM	0.72	0.70	0.82	0.75	0.79

TABLE XVII
Ciprofloxacin without SMOTE: Hyperparameter-Tuned Model Performance.

	Accuracy	**Precision**	**Recall**	**ROC**	**F2-Score**
Logistic Regression	0.76	0.76	0.79	0.77	0.78
Random Forest	0.76	0.76	0.79	0.76	0.78
XGBoost	0.78	0.75	0.86	0.78	0.83
AdaBoost	0.76	0.76	0.79	0.80	0.78
SVM	0.70	0.70	0.75	0.75	0.74

TABLE XVIII
Ciprofloxacin with SMOTE: Hyperparameter-Tuned Model Performance.

	Accuracy	**Precision**	**Recall**	**ROC**	**F2-Score**
Logistic Regression	0.76	0.76	0.79	0.77	0.78
Random Forest	0.76	0.76	0.79	0.77	0.78
XGBoost	**0.80**	**0.77**	**0.86**	**0.78**	**0.84**
AdaBoost	0.76	0.76	0.79	0.81	0.78
SVM	0.69	0.68	0.75	0.74	0.73

TABLE XIX
Linezolid Base Model Performance Without SMOTE.

	Accuracy	**Precision**	**Recall**	**ROC**	**F2-Score**
Logistic Regression	0.93	0.75	0.60	0.75	0.63
Random Forest	0.94	0.86	0.60	0.83	0.64
XGBoost	0.94	0.86	0.60	0.82	0.64
AdaBoost	0.94	0.86	0.60	0.89	0.64
SVM	0.94	1.00	0.50	0.89	0.56

TABLE XX
Linezolid Base Model Performance With SMOTE.

	Accuracy	**Precision**	**Recall**	**ROC**	**F2-Score**
Logistic Regression	0.90	0.60	0.60	0.74	0.60
Random Forest	0.94	0.86	0.60	0.86	0.64
XGBoost	0.94	0.76	0.70	0.82	0.71
AdaBoost	0.93	0.70	0.70	0.87	0.70
SVM	0.93	0.83	0.50	0.84	0.54

TABLE XXI
Linezolid without SMOTE: Hyperparameter-Tuned Model Performance.

	Accuracy	**Precision**	**Recall**	**ROC**	**F2-Score**
Logistic Regression	0.90	0.58	0.70	0.77	0.67
Random Forest	0.94	0.78	0.70	0.87	0.71
XGBoost	0.94	0.86	0.60	0.85	0.64
AdaBoost	**0.95**	**0.88**	**0.70**	**0.88**	**0.73**
SVM	0.93	0.75	0.60	0.86	0.63

C. Feature Importance

Since decision tree-based models performed best, this study utilized the built-in feature importance functionality of AdaBoost, XGBoost and Random Forest to identify the key predictive features for determining drug resistance in the models.

For isoniazid, the most important feature—TFxAMR_Encoded—captures weighted co-occurrence between transposons and AMR genes, highlighting the role of horizontal gene transfer. Other top features include RareGeneCount, presence of transposons, and specific AMR genes such as *rpoB2*, *msr(D)*, and *hasC*.

Further analysis of TFxAMR_Encoded revealed frequent co-occurrence of genes like *RbpA* with transposons such as *IS256*, *IS21*, and *IS110*. Notably, *IS256* often appeared alongside virulence-associated genes (e.g., *esxH*, *mbtH*, *PE35*), suggesting gene mobility significantly contributes to isoniazid resistance.

TABLE XXII
Linezolid with SMOTE: Hyperparameter-Tuned Model Performance.

	Accuracy	**Precision**	**Recall**	**ROC**	**F2-Score**
Logistic Regression	0.90	0.58	0.70	0.78	0.67
Random Forest	0.94	0.86	0.60	0.83	0.64
XGBoost	0.94	0.86	0.60	0.84	0.64
AdaBoost	0.93	0.70	0.70	0.85	0.70
SVM	0.92	0.62	0.80	0.82	0.75

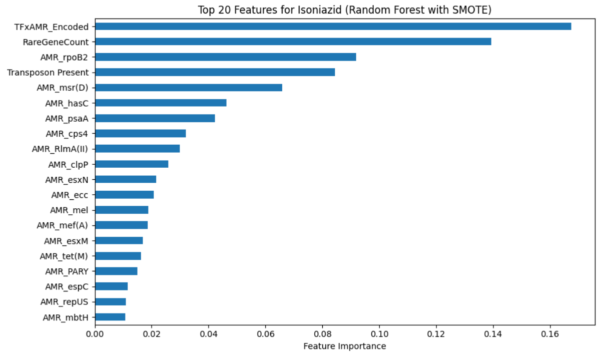

Fig. 3. Top 20 Features for Isoniazid

For rifampicin, TFxAMR_Encoded remains the top predictor, emphasizing the importance of transposon-AMR gene co-occurrence. Key AMR genes such as *mgrA*, *esxA*, *msr(D)*, and *mepR* also contribute significantly. RareGeneCount ranks 10th, while Transposon Present does not appear in the top 20.

Notable co-occurrences include *esaB* and *hld* with *IS6* and *IS1182*, and *RbpA* with *IS21*, *IS110*, and *IS256*, highlighting gene mobilization's role in resistance.

For oxacillin, TFxAMR_Encoded remains the top feature, followed by key AMR genes *(Bla)mecI*, *hld*, *dfrC*, and RareGeneCount (ranked third). Frequent co-occurrences include transposons like *Tn3*, *IS6*, and *IS1182* with AMR genes *FosA6*, *KpnF*, *KpnE*, and predictors *esaB* and *hld*.

For ciprofloxacin, AMR genes dominated the top features, led by *dfrC*, followed by *marA*, *KpnF*, *ColRNA_1*, and *mepR*. TFxAMR_Encoded ranked 20th. The transposon *IS607* frequently co-occurred with genes such as *RbpA*, *MMR*, and several *esx* family members, suggesting potential mobilization links.

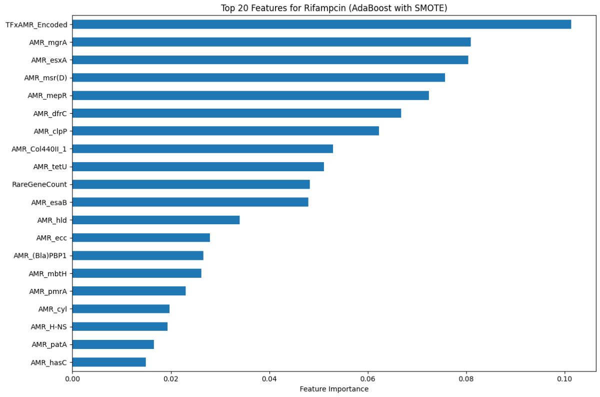

Fig. 4. Top 20 Features for Rifampicin

For linezolid, TFxAMR_Encoded is the top predictor, followed by AMR genes *fosD_1*, *dfrC*, *ColRNAI_1*, and *patB*. Transposon Present and RareGeneCount also rank within the top 20. Key transposon-gene associations include *Tn3* with *KpnE*, *marA*, *ramA*, and *FosA6*; *IS6* with *hld* and *esaB*; and *IS1182* with *esaB*.

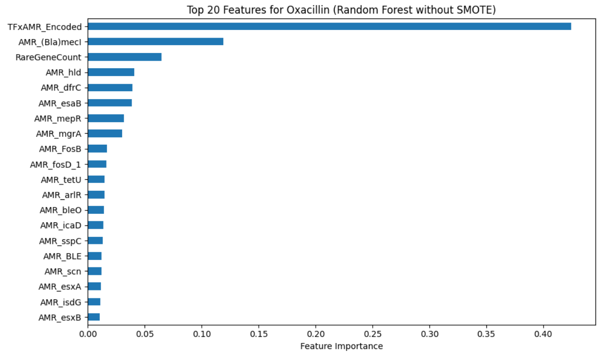

Fig. 5. Top 20 Features for Oxacillin

V. Discussion

This study evaluated five machine learning models—Logistic Regression, Random Forest, XGBoost, AdaBoost, and SVM—for predicting antibiotic resistance in *Mycobacterium tuberculosis*, *Klebsiella pneumoniae*, and *Staphylococcus aureus* using whole-genome sequencing (WGS) data. A key focus was assessing transposons as predictive features. A web application was developed to integrate these models for rapid AMR detection to support timely clinical decisions.

WGS, once limited by time and cost, now offers rapid pathogen profiling, often within 6-24 hours, matching or surpassing traditional culture-based methods, especially for slow-growing bacteria. The web tool leverages genomic inputs to generate early antibiograms, aiding empirical treatment decisions—crucial in resource-limited settings.

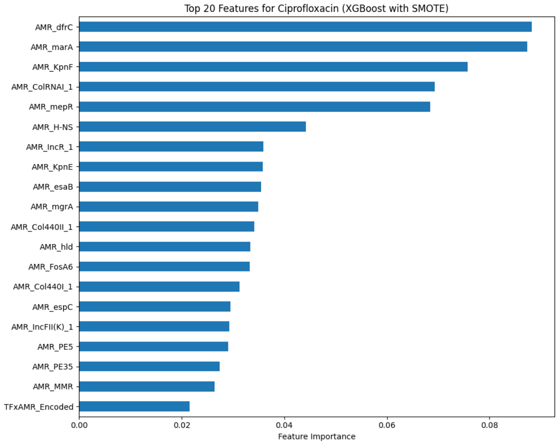

Fig. 6. Top 20 Features for Ciprofloxacin

Decision tree-based models (AdaBoost, XGBoost, Random Forest) outperformed others and provided interpretable insights. Key AMR genes like *dfrC*, *esaB*, *msrD*, *clpP*, and *ecc* were consistently important across antibiotics, aligning with known resistance and virulence roles.

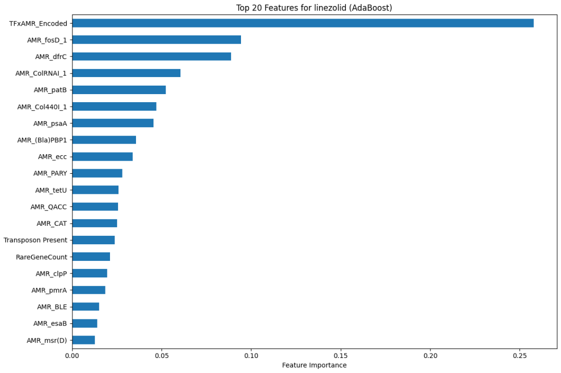

Fig. 7. Top 20 Features for Linezolid

The top feature, TFxAMR_Encoded, representing transposon-AMR gene co-occurrence, highlighted horizontal gene transfer as a major resistance mechanism. Co-occurrence patterns involving transposons (e.g., *IS256*, *IS6*, *Tn3*) and genes such as *RbpA*, *esaB*, and *hld* were consistent across antibiotics, emphasizing the biological relevance of gene mobility in resistance spread.

Regulatory genes like *ramA*, *marA*, and *H-NS*—known to modulate efflux pumps and gene expression—also frequently co-occurred with transposons, linking gene regulation and mobility. Though transposon presence alone was not always a top predictor, their role as vehicles for AMR gene dissemination under selective pressure is fundamental to resistance evolution.

Clinically, identifying key resistance genes (e.g., *dfrC*) aids targeted therapy decisions, while highlighting transposons underscores the importance of monitoring mobile genetic elements to curb AMR spread.

VI. Conclusion and Future Work

This study demonstrated the potential of whole-genome sequencing (WGS) combined with explainable machine learning to rapidly predict antibiotic resistance in *Mycobacterium tuberculosis*, *Klebsiella pneumoniae*, and *Staphylococcus aureus*. Despite challenges like limited sample sizes and class imbalance, decision tree–based models achieved robust performance and interpretability, revealing key AMR genes and the crucial role of transposon-mediated gene mobility.

To enhance the study's accuracy and applicability, future work should focus on expanding and diversifying training datasets to improve generalization and capture complex resistance patterns. Incorporating additional biologically informed features—such as operon structures, plasmid types, and functional annotations—and employing feature selection techniques will refine model performance. Extending analysis to a broader range of antibiotics will increase clinical relevance.

Given the importance of mobile genetic elements, future research should explore mobility scores and deeper transposon-gene co-occurrence analyses across species to better understand horizontal gene transfer. Finally, improving scalability through periodic retraining with new genomic data, and integrating metagenomic or real-time surveillance inputs, will keep the platform current with emerging resistance threats, enhancing its clinical impact.

References

[1] World Health Organization, “Antimicrobial resistance,” 2019.

[2] G. Mancuso, A. Midiri, E. Gerace, and C. Biondo, “Bacterial antibiotic resistance: The most critical pathogens,” *Pathogens*, vol. 10, no. 10, p. 1310, 2021.

[3] World Bank, “Drug-Resistant Infections: A Threat to Our Economic Future,” 2017.

[4] M. N. Anahtar, J. H. Yang, and S. Kanjilal, “Applications of machine learning to the problem of antimicrobial resistance: An emerging model for translational research,” *J. Clin. Microbiol.*, vol. 59, no. 7, 2021.

[5] M. A. Salam, M. Y. Al-Amin, J. S. Pawar, N. Akhter, and I. B. Lucy, “Conventional methods and future trends in antimicrobial susceptibility testing,” *Saudi J. Biol. Sci.*, vol. 30, no. 3, p. 103582, 2023.

[6] A. Sakagianni, C. Koufopoulou, G. Feretzakis, D. Kalles, V. S. Verykios, P. Myrianthefs, and G. Fildisis, “Using machine learning to predict antimicrobial resistance—A literature review,” *Antibiotics*, vol. 12, no. 3, p. 452, 2023.

[7] J. Botelho, A. Cazares, and H. Schulenburg, “The ESKAPE mobilome contributes to the spread of antimicrobial resistance and CRISPR-mediated conflict between mobile genetic elements,” *Nucleic Acids Res.*, vol. 51, no. 1, pp. 236–252, 2023.

[8] S. Babakhani and M. Oloomi, “Transposons: The agents of antibiotic resistance in bacteria,” *J. Basic Microbiol.*, vol. 58, no. 11, pp. 905–917, 2018.

[9] S. R. Babirye, M. Nsubuga, G. Mboowa, C. Batte, R. Galiwango, and D. P. Kateete, “Machine learning-based prediction of antibiotic resistance in *Mycobacterium tuberculosis* clinical isolates from Uganda,” *BMC Infect. Dis.*, vol. 24, no. 1, 2024.

[10] Y. Ren *et al.*, “Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning,” *Bioinformatics*, vol. 38, no. 2, pp. 325–334, 2021.

[11] M. I. Gröschel *et al.*, “GenTB: A user-friendly genome-based predictor for tuberculosis resistance powered by machine learning,” *Genome Med.*, vol. 13, no. 1, 2021.

[12] M. Yasir, A. M. Karim, S. K. Malik, A. A. Bajaffer, and E. I. Azhar, “Prediction of antimicrobial minimal inhibitory concentrations for *Neisseria gonorrhoeae* using machine learning models,” *Saudi J. Biol. Sci.*, vol. 29, no. 5, pp. 3687–3693, 2022.

[13] A. Lipszyc, M. Szuplewska, and D. Bartosik, “How do transposable elements activate expression of transcriptionally silent antibiotic resistance genes?,” *Int. J. Mol. Sci.*, vol. 23, no. 15, p. 8063, 2022.

[14] A. Khezri, E. Avershina, and R. Ahmad, “Plasmid identification and plasmid-mediated antimicrobial gene detection in Norwegian isolates,” *Microorganisms*, vol. 9, no. 1, p. 52, 2020.

[15] G. Ng, “Responding to the alarming rise of tuberculosis cases in the Philippines,” *Medical Channel Asia*, Nov. 2024.

[16] World Health Organization, ”Tuberculosis,” 2024.