Computational Biophysics (CS61060) Term Project: ML and Shapley-Based prediction of resistance mechanisms in pathogenic bacteria

Group 15: Ram Sundaram (20BT10026),
Vishal Ravipati (20CS10076), Giri Nikhil Rajendra (20BT10036)

April 14, 2024

Contents

Acronyms

bp base pairs

CART Classification and Regression Tree

CRyPTIC Comprehensive Resistance Prediction for Tuberculosis: an International Consortium

DT Decision Tree

GBR Gradient Boosting Regression

GPU Graphics Processing Unit

GWAS Genome Wide Association Studies

HHS Hungry Hungry SNPos

INH isoniazid

LGBM Light Gradient Boosting Machine

MAE Mean Absolute Error

MIC Minimum Inhibitory Concentration

ML Machine Learning

MSE Mean Squared Error

MXF moxifloxacin

NCBI National Center for Biotechnology Information

RF Random Forest

SNP Single Nucleotide Polymorphism

WGS Whole Genome Sequencing

XGB eXtreme Gradient Boosting

1 Summary

This report focuses on developing an innovative pipeline for antibiotic resistance prediction. Firstly, it aims to construct a novel pipeline integrating machine learning and statistical analysis leveraging genomic mutation data for Minimum Inhibitory Concentration (MIC) prediction. Secondly, the study emphasizes interpretability through Shapley values, facilitating clinical applicability. These objectives contribute to a comprehensive approach with implications for research decision-making.

The presented pipeline is a comprehensive framework for antibiotic resistance prediction, focusing on *Streptococcus pneumoniae* and *Mycobacterium tuberculosis*. The pipeline involves multiple key stages: data collection, preprocessing, feature selection, model fitting and Shapley value extraction. Two datasets, Maela and Cryptic, are utilized, containing genomic information and Minimum Inhibitory Concentration (MIC) values for the respective pathogens.

Data preprocessing involves standardizing the raw genomic data, variant calling to derive SNP and indel mutations in relation to a reference genome, transforming it into a sparse matrix format, normalizing MIC values and transforming them to the log space. A novel pseudo-ensemble approach is employed for feature selection, combining results from Decision Trees, Random Forests, Gradient Boosting, and XGBoost algorithms. Model fitting involves training the dataset on three selected algorithms: Decision Trees, Random Forests, and XGBoost.

Shapley values, rooted in cooperative game theory, are then calculated for each mutation, providing information about the contribution of each genomic mutation to antibiotic resistance. These Shapley values help us quantify and rank mutations in both genes and non-coding regions in order of their significance towards resistance and help us identify novel mutations that may not otherwise have been correlated with antibiotic resistance. This is particularly useful in the context of rarely-used antibiotics and resistance mechanisms in pathogens that may not be well-researched.

We successfully applied our pipeline to identify key mutations associated with antibiotic resistance in *Mycobacterium tuberculosis* and *Streptococcus pneumoniae*. We then conducted an extensive literature review to validate whether our pipeline could correctly flag genes previously known to be associated with resistance. This review showed promising results. For example, our pipeline's top 10 results for moxifloxacin and isoniazid resistance in *M. tuberculosis* included mutations in genes such as gyrA and katG, aligning with existing literature on their association with antibiotic resistance. The validation process further confirmed the accuracy of our pipeline in flagging known or suspected resistance-associated genes. Additionally, the pipeline was able to detect cross-resistance patterns successfully.

Furthermore, the pipeline's versatility extends beyond antibiotic resistance, holding promise for analyzing diverse genotypic datasets across species. Future improvements may be able to incorpo-

rate advanced feature selection methods and dimensionality reduction techniques, enhancing the pipeline's adaptability and robustness for broader genetic analyses.

2 Introduction

Drug resistance is a critical issue where infectious organisms, or pathogens, develop resistance to commonly used drugs in treatment[22]. This study focuses on leveraging Machine Learning and other statistical methods for the quantifiable and interpretable identification and prediction of resistance mechanisms in pathogenic bacteria. Through this study, we aim to develop a novel pipeline for the identification of resistance-related mutations in pathogenic bacteria.

The emergence of drug resistance poses a severe public health threat globally, affecting not only low and middle-income countries but also high-income countries, particularly within hospital settings. Without the prompt development of new antimicrobial drugs, it is projected that annual mortality due to drug resistance could surpass 10 million people by 2050, exceeding current cancer-related mortality rates[23].

Models for predicting drug resistance from Whole Genome Sequencing (WGS) data can be broadly categorized into two classes. The first, termed "catalogue methods," involves examining WGS data for the presence of specific point mutations, often Single Nucleotide Polymorphism (SNP)s, known to be associated with drug resistance. If at least one mutation that was previously identified to be associated with resistance is detected, the specific isolate is flagged as resistant[28, 2]. While these methods are straightforward, they often lack predictive accuracy[27], especially in identifying new resistance mechanisms or predicting resistance to rarely used drugs.

The second class, known as "machine learning methods," aims to predict drug resistance by training Machine Learning or Deep Learning models directly on WGS and drug susceptibility test data[11]. Although these methods offer high predictive accuracy, they come at the expense of flexibility and interpretability. Typically, they provide limited insights into the drug resistance mechanisms and may lack explicit constraints on model complexity. Deep neural networks, as an example of such machine learning approaches, are highly accurate but represent complex "black-box" models of drug resistance with virtually no interpretability, which inhibits the feasibility of these methods for clinical utility[32].

In this paper, we present a novel solution based on a pipeline encompassing preprocessing to extract features in the form of SNPs from WGS data, feature selection to limit the feature inputs to the model to only those that correspond to resistance, supervised model training and prediction leveraging an ensemble of tree-based regression algorithms and finally analysing a feature/mutation's contribution to resistance through the use of Shapley values. We then analyse the results of our pipeline on WGS data from two common pathogenic bacteria, namely *Streptococcus pneumoniae* and *Mycobacterium tuberculosis* and compare the mutations and genes identified by our pipeline with those already established to be associated with antibiotic resistance by previous literature.

3 Objectives

The following objectives outline the focus areas of the research, emphasizing the development of a comprehensive pipeline for antibiotic resistance prediction:

- *Development of a Novel Pipeline for Antibiotic Resistance Prediction:* This research aims to construct an innovative pipeline integrating machine learning and statistical analysis techniques to predict antibiotic resistance, specifically focusing on MIC prediction through regression analysis.
- *Enhancing Clinical Applicability through Interpretability:* This study endeavours to ensure the interpretability of the developed pipeline. By employing Shapley values, the goal is to enhance the transparency and interpretability of the model's predictions, facilitating its integration into clinical settings.

4 Methodology

This section encompasses the materials and methodology utilised in this study. The pipeline developed includes data collection, data preprocessing, feature extraction, feature selection, ensemble model fitting and Shapley value extraction. In this section, we will provide a detailed description of each of these steps. The following flowchart shows the full pipeline.

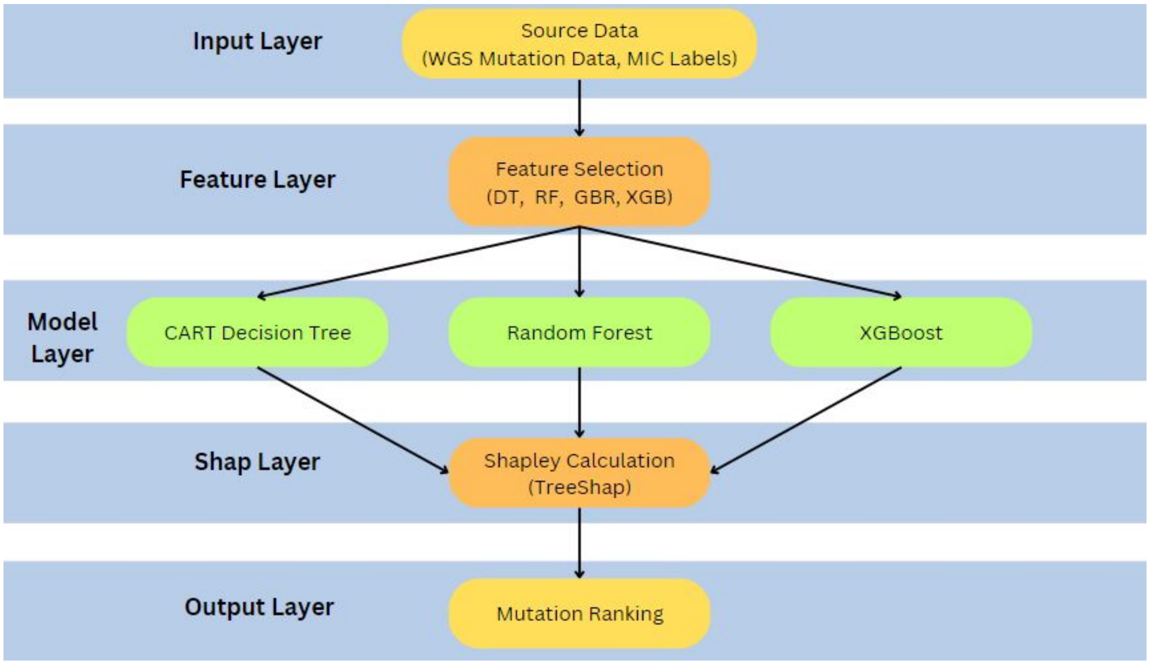

Figure 1: Model Pipeline

4.1 Data sources and collection

4.1.1 Species selection for analysis

For the purposes of this study, we largely used two major datasets corresponding to Minimum Inhibitory Concentrations (MIC) values of the two aforementioned pathogenic bacteria, *Streptococcus pneumoniae* and *Mycobacterium tuberculosis*. These two pathogens were specifically chosen as they represent an ideal combination of characteristics for this study, which is meant to serve as a validation study for the feasibility of the proposed pipeline:

- They are common pathogens within the human population and have a large diversity of mutations within their population.
- They have relatively small genome sizes, easing the computational strain of their analysis. *S. pneumoniae* has a single circular chromosome of 2,038,615 base pairs[16], while *M. tuberculosis* has a genome size of 4.4 million base pairs (bp) encoding approximately 4,400 genes[8].
- Both these species are well-researched in the field of bioinformatics, providing ample data for analysis as well as a literature review to validate results.

4.1.2 Selection of the feature type

Here, we specifically want to analyse WGS to predict resistance mechanisms by identifying mutations associated with resistance patterns. The field of bioinformatics provides standard methods of representing mutations, namely SNPs, k-mers and unitigs. A SNP represents a difference in a single nucleotide, while a k-mer represents a nucleotide substring of length k contained within a biological sequence. Unitigs are defined as the longest sequence(s) that can be obtained given overlap by exactly k-1 nucleotides[6].

Unitigs and k-mers may provide increased prediction accuracy but at the cost of decreased interpretability and an increased number of features, with both k-mers and unitigs providing exponentially more features than representation by SNPs. Given the use of an Machine Learning (ML) ensemble model, generating these many features may lead to overfitting. Since our main aim with this study is to generate interpretable results, we shall limit our analysis to SNPs, using k-mers and unitigs, perhaps in combination with a SNP feature set reserved for future work.

4.1.3 Data sources for analysis

Given the use of SNPs for the analysis of *S. pneumoniae* and *M. tuberculosis* genomes, this study utilised two datasets providing SNPs extracted from WGS data and MIC values for both bacteria.

Maela dataset: The Maela dataset consists of genome sequences from 3,085 *Streptococcus pneumoniae* isolates. Penicillin Minimum Inhibitory Concentration (MIC) values have been measured for each sample, providing resistance characteristics. Each isolate in this dataset is associated with 391,627 SNPs[5].

Cryptic dataset: The Comprehensive Resistance Prediction for Tuberculosis: an International Consortium (CRyPTIC) provides a comprehensive dataset comprising 12,289 *Mycobacterium tuberculosis* clinical isolates. This dataset encompasses whole-genome sequencing and minimum inhibitory concentrations for 13 antitubercular drugs measured in a singular assay. The compendium includes 6,814 isolates resistant to at least one drug, with 2,129 meeting clinical definitions of rifampicin resistance, multidrug resistance, pre-extensively drug-resistant, or extensively drug-resistant. The dataset is rich in rare resistance-associated variants, providing an excellent platform to validate the feasibility of our prediction and identification pipeline on more complex genomes and resistance patterns. The dataset provides for over 2,854,309 independent mutations comprising both SNPs and indels[9].

Dataset	**Species**	**Genome Size (bp)**	**No. of Isolates**	**No. of Mutations (SNPs/Indels)**
Maela	*S. pneumoniae*	2,038,615	3,085	391,627 (SNPs)
Cryptic	*M. tuberculosis*	4.4M	12,289	2,854,309 (SNPs/Indels)

Table 1: Key Characteristics of Chosen Datasets

4.2 Data preprocessing

Our data preprocessing steps included preprocessing on both the WGS data and the MIC labels.

From the WGS data, the first step was variant calling, which is a bioinformatics process used to identify genetic variations, including single nucleotide polymorphisms (SNPs), from whole genome sequencing (WGS) data. The bioinformatics software tool GATK (Genome Analysis Toolkit) was used. The process involved aligning short reads to a reference genome provided by the dataset originators, detecting variants by comparing reads to the reference and subsequent filtering and annotation of the variants. This method accurately identifies SNP mutations within an individual's genome.

After variant calling, the major requirement with the WGS mutation data was to convert it to a standardised format for analysis within our pipeline. This is a necessary and complicated step due to the varying formats of WGS datasets within the field of bioinformatics and the lack of a single coherent standard.

This standardisation involved representing it as a $n * m$ matrix, where n represents the number of samples and m represents the number of mutations. Each row represents a unique sample, with each column representing a single mutation. Each row is a vector of binary values denoting the presence or absence of a particular mutation associated with that column.

However, a challenge arises from the "fat" nature of genomic data[12], where the number of samples is relatively low (in the order of thousands), while the number of mutations is substantially higher (in the order of millions). This condition provides two challenges. Namely, it escalates computational costs and increases the risk of model overfitting. We shall discuss the solution to the first challenge here while mitigating model overfitting shall be discussed in the next section on feature selection.

To reduce computational costs, we opted to represent the matrix as a sparse matrix, specifically using the Compressed Sparse Row format. We can leverage the use of a sparse matrix as a majority of the binary values within the matrix are zero values. This transformation efficiently stores non-zero values, reducing computational overhead and file size.

In contrast, the preprocessing steps for MIC labels were relatively minimal. A log transformation was applied to bring these values into logarithmic space because MIC values are calculated based on serial dilutions. Additionally, the normalization of MIC values preceded the analysis, ensuring consistency in the dataset.

4.3 Feature selection

As highlighted in the previous section, the "fat" nature of genomic data[12], that is, the relatively large number of features in comparison to the number of samples, increases the risk of overfitting with a Machine Learning model.

This necessitates using feature selection to remove features or other methods, such as Principal Component Analysis, to reduce the feature space. In this case, many mutations were found to be irrelevant to antibiotic resistance and, thus, could be culled from the feature set, thereby reducing the number of features for analysis to the order of 1,000. This feature selection was accomplished by training 4 algorithms of varying complexity, namely Decision Tree (DT), Random Forest (RF), Gradient Boosting Regression (GBR) and eXtreme Gradient Boosting (XGB) on the entire feature space, leveraging MIC values as the dependent variable and setting a threshold cutoff value for feature weights.

These feature weights indicate the relevance of each mutation to the specific algorithm's prediction of the MIC value. So as to not lose any relevant information, the threshold value is sent extremely close to zero, of the order of $1e^{-}6$. Four different feature sets are selected based on the feature weights provided by each of the 4 tree algorithms. An analysis is conducted on each of these datasets to provide a weighted average of each feature space. Thus, this provides a pseudo-ensemble approach to feature selection.

4.4 Model fitting

After the data preprocessing and feature selection steps, we have four different datasets consisting of approximately 1,000-2,000 features/mutations each. Each of these datasets is then fitted to one of three algorithms, namely decision trees, random forests, and XGBoost. This effectively generates 12 different sets of results, which are integrated in the next section. We will first explore the workings of each algorithm and then explore the reasons for selecting these three algorithms.

4.4.1 Decision Trees (CART)

We select the Classification and Regression Tree (CART) model to implement decision trees. It operates by recursively partitioning the training data into smaller subsets through binary splits, creating a binary tree structure. At each node, the algorithm selects a feature and a threshold that optimally separates the data based on information gain or Gini impurity, which measures the effectiveness of the split. This process continues recursively until a stopping criterion, such as a maximum tree depth or minimum instances per leaf node, is met. The prediction is made by traversing the tree from the root to a leaf node, where, for regression, the prediction is the average of target values. The Gini impurity calculation involves determining the impurity of the root node and calculating the impurity for each possible split point, selecting the one that minimizes impurity[18].

The Gini impurity for a set S with classes $C_1, C_2, \ldots, C_k$ is calculated as follows:

$$\text{Gini}(S) = 1 - \sum_{i=1}^{k} P(C_i)^2$$

where $P(C_i)$ is the proportion of instances in class C_i in set S.

The information gain for a split on feature X at threshold t is given by:

$$\text{InfoGain}(S, X, t) = \text{Gini}(S) - \left(\frac{|S_{\text{left}}|}{|S|}\text{Gini}(S_{\text{left}}) + \frac{|S_{\text{right}}|}{|S|}\text{Gini}(S_{\text{right}})\right)$$

where S_{left} and S_{right} are the subsets created by the split[33].

4.4.2 Random Forest

The Random Forest algorithm is an ensemble learning method based on decision trees, combining the predictions of multiple individual trees to enhance overall performance in classification and regression tasks. It operates by constructing a multitude of decision trees during training and outputs the mode of the classes (for classification) or the average prediction (for regression) of the individual trees. Each tree is trained on a random subset of the training data, and at each split, a random subset of features is considered, introducing diversity among the trees. The algorithm then aggregates the predictions from individual trees, mitigating overfitting and improving generalization performance. For classification tasks, the mode of the predicted classes is taken, and for regression, the average prediction is computed[3].

Random Forest aggregates predictions from individual trees, where T denotes the set of trees in the forest and h_t is the prediction of tree t. For regression, the aggregated prediction is the average of individual predictions:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t$$

4.4.3 XGBoost

XGBoost, short for eXtreme Gradient Boosting, is an ensemble learning algorithm useful in both classification and regression tasks. It operates by constructing an ensemble of weak learners, usually decision trees, sequentially, where each tree corrects the errors of its predecessors. The algorithm optimizes the objective function through a combination of gradient descent and the second-order derivative of the loss function, prioritizing the minimization of both bias and variance. Additionally, XGBoost introduces the concept of ”boosting” by assigning weights to misclassified instances, emphasizing learning from errors. The final prediction is a weighted sum of predictions from individual weak learners, with the regularization term contributing to overall model robustness[4].

XGBoost optimizes the objective function $L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$, where l is the loss function measuring the discrepancy between true labels y_i and predicted values $\hat{y}_i$, and $\Omega(f_k)$ is a regularization term penalizing the complexity of each weak learner. The algorithm minimizes this objective by iteratively adding weak learners to the ensemble. The prediction $\hat{y}_i$ is calculated as a weighted sum of predictions from individual weak learners:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

where f_k is the output of the k-th weak learner. The regularization term $\Omega(f_k)$ is designed to enhance model generalization.

The reasoning behind using tree-based algorithms lies in the computational efficiency offered by using TreeShap instead of KernelShap, enabling exponentially faster results. TreeShap requires the usage of tree-based algorithms for reasons we shall explore in the next section. Several tree- and non-tree-based algorithms were tested, including support vector machines, linear regression, Light Gradient Boosting Machine (LGBM), and the aforementioned tree-based algorithms. Deep Learning-based approaches were avoided due to a complete lack of interpretability. The following table summarises the Mean Squared Error (MSE), R^2, and Mean Absolute Error (MAE) values as a result of fitting various ML models on both our Maela and Cryptic datasets after performing 10-fold cross-validation.

Model	**Maela**			**Cryptic**		
	MSE	**R_Squared**	**MAE**	**MSE**	**R_Squared**	**MAE**
Decision Tree	0.585	0.397	0.468	0.843	0.623	0.449
GradientBoost	**0.308**	**0.684**	**0.385**	**0.428**	**0.800**	**0.639**
Linear Regression	0.413	0.574	0.456	0.603	0.890	0.492
LGBM	**0.319**	**0.674**	**0.389**	**0.538**	**0.768**	**0.697**
Random Forest	**0.316**	**0.675**	**0.377**	**0.383**	**0.841**	**0.389**
Support Vector Machine	0.329	0.661	0.388	0.455	0.753	0.392
XGBoost	**0.357**	**0.633**	**0.405**	**0.411**	**0.930**	**0.405**

Table 2: Performance Metrics for ML Models on Maela and Cryptic Datasets.

As we can observe from the results, XGBoost, Random Forest, LGBM and GradientBoost have excellent MSE values. XGBoost, LGBM and GradientBoost are relatively similar in terms of their complexity and implementation. Hence, we select XGBoost among these as a result of observing it to be the most computationally efficient among the three, with Graphics Processing Unit (GPU) compatibility. We additionally selected CART decision trees to be included in our model as their simplistic approach and tendency to underfit would balance our ensemble approach given the complexity of XGBoost and Random Forests and their tendency to overfit.

4.5 Shapley Values

Shapley values, rooted in cooperative game theory, provide a systematic way to attribute a contribution to each contributor among a group of contributors generating some value. In machine learning interpretability, Shapley values offer a principled approach to assign contributions to input features, where each individual feature is essentially a ”contributor” to a ”value” which is represented by the cost function. The Shapley value (ϕ_i) for a feature i is defined as the weighted sum of marginal contributions over all possible feature orderings:

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

Here, N is the set of all features, S is a subset of N and $f(S)$ is the model's prediction on subset S. Shapley values offer insights into the importance of individual features and their interactions.

In the context of interpretability in Machine Learning, here is a detailed explanation of each part of the above equation

- $\phi_i(f)$: Shapley value assigned to feature i in the context of function f.
- $S \subseteq N \setminus \{i\}$: Consideration of all possible subsets S of features in N excluding feature i.
- $\frac{|S|! \cdot (|N|-|S|-1)!}{|N|!}$: Weight factor accounting for the number of ways to form a subset S of size $|S|$ with $|N|$ total features.
- $[f(S \cup \{i\}) - f(S)]$: Marginal contribution of feature i to function f when added to subset S, measuring the change in the function's output.
- $\sum$: Summation over all possible subsets S of features, calculating the average contribution of feature i across all possible orders.

Thus, the Shapley value for feature i is determined by considering its marginal contributions across all possible feature subsets, ensuring fair attribution based on cooperative game theory principles[26].

In practice, the Shapley methodology assigns a single Shapley value (ϕ_i) to each feature in each prediction. The sign of each individual Shapley value indicates the direction and magnitude of its contribution. Positive values signify a positive impact on the prediction, while negative values represent a negative impact. The absolute value of the Shapley value quantifies the strength of the contribution, providing a clear measure of each feature's influence to a single prediction.

The symbol ϕ_{feature} denotes the global Shapley value of the particular feature. It is calculated as the expected value of the absolute Shapley values ($|\phi_i(f)|$) assigned to the feature across all predictions.

$$\phi_{\text{feature}} = \mathbb{E}\left[|\phi_i(f)|\right] = \frac{1}{N} \sum_{i=1}^{N} |\phi_i(f)|$$

This formulation measures the feature's overall influence on the model's predictions, considering both positive and negative contributions. In machine learning, Shapley values are calculated by evaluating the model's predictions for all possible coalitions of features. This process allows for a fair attribution of contributions to each feature.

Shapley values offer versatile insights into model predictions, providing advantages in various aspects:

Two common implementations are Kernel Shap and Tree Shap:

- *Kernel Shap*: Kernel Shap, rooted in cooperative game theory, calculates Shapley values using a weighted average of predictions for all possible permutations of features. It provides a global perspective on feature importance and interactions. However, its time complexity is exponential, making it computationally expensive for large feature sets.
- *Tree Shap*: Tree Shap leverages the structure of tree-based models for efficient computation of Shapley values. Its recursive algorithm allows for linear time complexity, making it exponentially faster than Kernel Shap. The time complexity of Tree Shap is $O(d \cdot 2^d)$, where d is the depth of the tree. This efficiency arises from the additive property of Shapley values for tree models, enabling efficient traversal of the tree structure. This efficiency is especially valuable when interpreting complex models with large feature sets.

In summary, Shapley values offer a theoretically grounded method for interpreting model predictions. While KernelShap provides a global perspective, TreeShap is particularly efficient for tree-based models, allowing for rapid and accurate interpretation.

After iterating over our 3 selected models over the 4 datasets comprising of different feature sets selected by 4 different algorithms, we apply Shap to generate Shapley values for each mutation and take an average of the Shapley value for each feature across the ensemble of 12 trained models to determine the final Shap value for each mutation.

5 Results

5.1 Sample results

The final output from our proposed pipeline is a Shapley value associated with each mutation in regard to its contribution to the pathogen's resistance to a specific antibiotic. By sorting mutations based on their Shapley values, we can find the mutations responsible for the highest contributions to the antibiotic resistance of a specimen. By aggregating results across all our isolates, we can showcase the mutations most commonly responsible for antibiotic resistance across a species. We can further augment our findings by attaching the genes associated with each mutation. While the Cryptic dataset provided metadata regarding the corresponding gene each mutation belongs to, the data for *S. pneumoniae* was obtained from the National Center for Biotechnology Information (NCBI)[10].

The table and figure below show a set of our top 10 sample results for mutations associated with moxifloxacin (MXF) resistance in *M. tuberculosis*, one of the most commonly administered antibiotics for treating tuberculosis. Here, Average Shap refers to the average Shapley value across our ensemble of models across the sample population. Average Rank refers to the average ranks of that specific mutation across the sample population.

Mutation ID	Avg Shap	Avg Rank	Gene
7582a>g	22.564	1	gyrA
7570c>t	10.977	2	gyrA
7581g>a	5.566	3	gyrA
210624g>a	4.662	4.333	lprO
340372t>c	3.762	5.667	PPE3
7582a>c	3.453	5.333	gyrA
7572t>c	2.823	7.333	gyrA
7582a>o	2.636	8.667	gyrA
7581g>t	2.499	10.333	gyrA

Table 3: Top 10 mutations for moxifloxacin resistance in *M. tuberculosis*

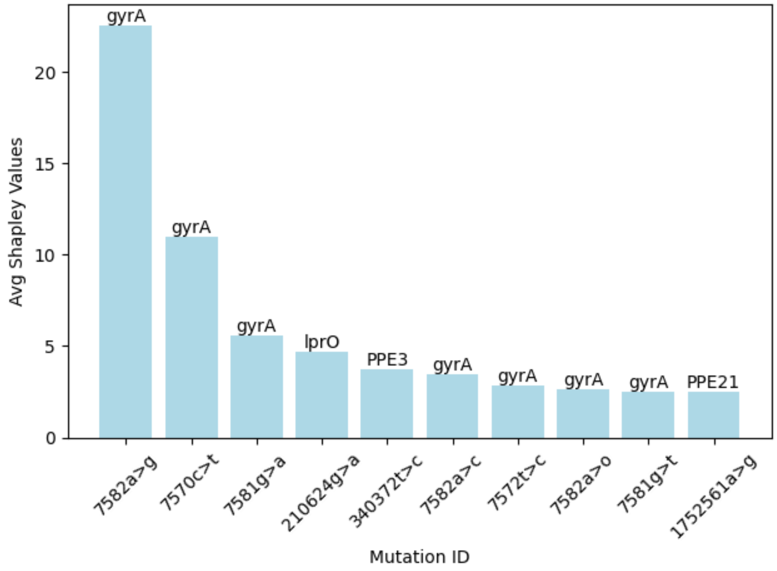

Figure 2: Top 10 mutations for moxifloxacin resistance in *M. tuberculosis*

5.2 Results validation

For the purposes of validation, we must find that our pipeline correctly flags mutations associated with antibiotic resistance. To accomplish this, we limit ourselves to the top 10 mutations identified by our pipeline and conduct an extensive literature review to determine whether the specific gene in which the mutation identified by our pipeline lies is previously known to be associated with antibiotic resistance, suspected to be associated with antibiotic resistance or there exists no conclusive evidence regarding association. We shall first provide individual results for a select few antibiotics within our Maela and Cryptic datasets and then summarise our findings over all 14 antibiotic sets. The below figure summarises our findings for moxifloxacin resistance in *M. tuberculosis*

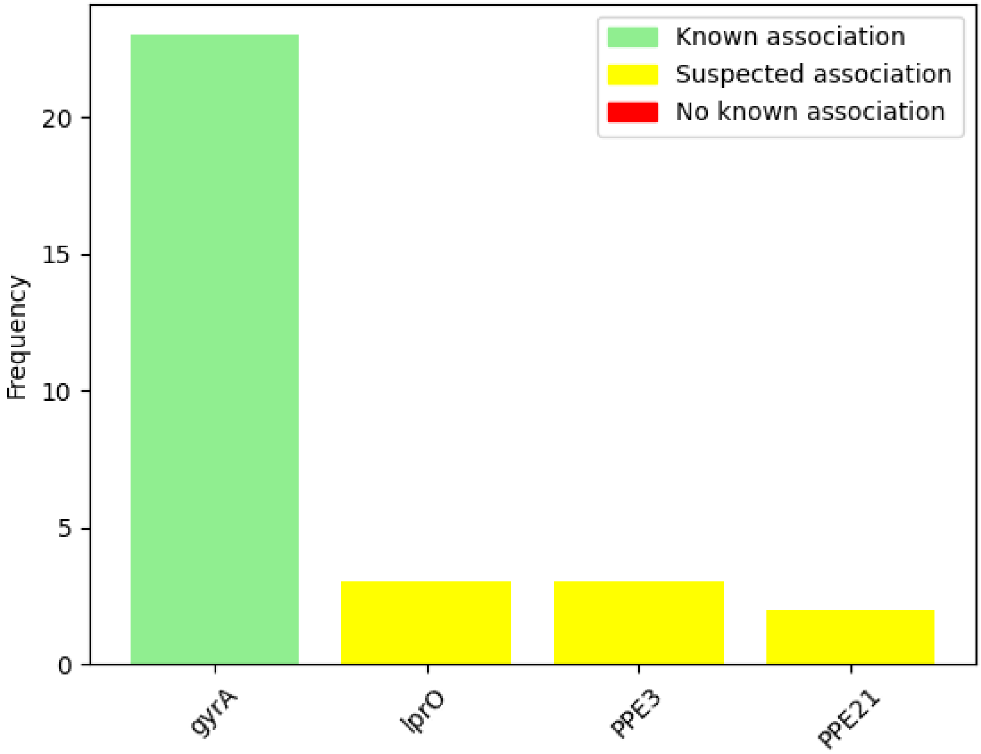

Figure 3: Review of genes identified by pipeline for moxifloxacin resistance in *M. tuberculosis*

In the above figure, we can observe that gyrA, lprO, PPE3 and PPE21 have been flagged by our model for moxifloxacin (MXF) resistance. Through our literature review, we have identified that gyrA has a well-documented association with MXF resistance, altering the enzyme DNA gyrase that helps with DNA replication in *M. tuberculosis*[20]. Meanwhile, lprO is suspected as a possible lipoprotein mutation involved in resistance[30], while PPE[7] and PPE21[15] are both suspected to have an association with resistance

We shall perform the same analysis for isoniazid (INH) resistance in *M. tuberculosis*.

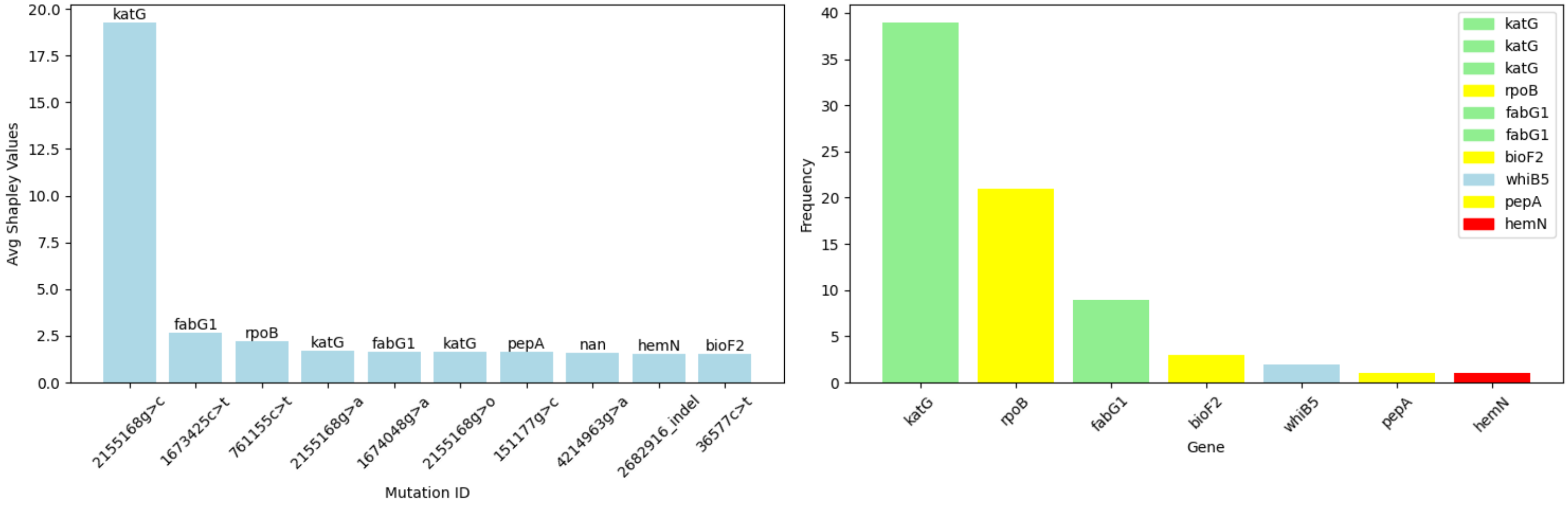

Figure 4: Review of genes identified by pipeline for isoniazid resistance in *M. tuberculosis*

Our literature review has identified that KatG (catalase-peroxidase) encodes a peroxidase responsible for the activation of INH, which is otherwise a pro-drug. Thus, katG mutations are commonly associated with INH resistance[29]. Mutations in the fabG1 gene are commonly associated with those in the katG region and are themselves commonly associated with INH resistance[17]. Mutations in rpoB and bioF2 are associated with rifampicin[19] and streptomycin[25], respectively, potentially indicating an isolate's cross-resistance.

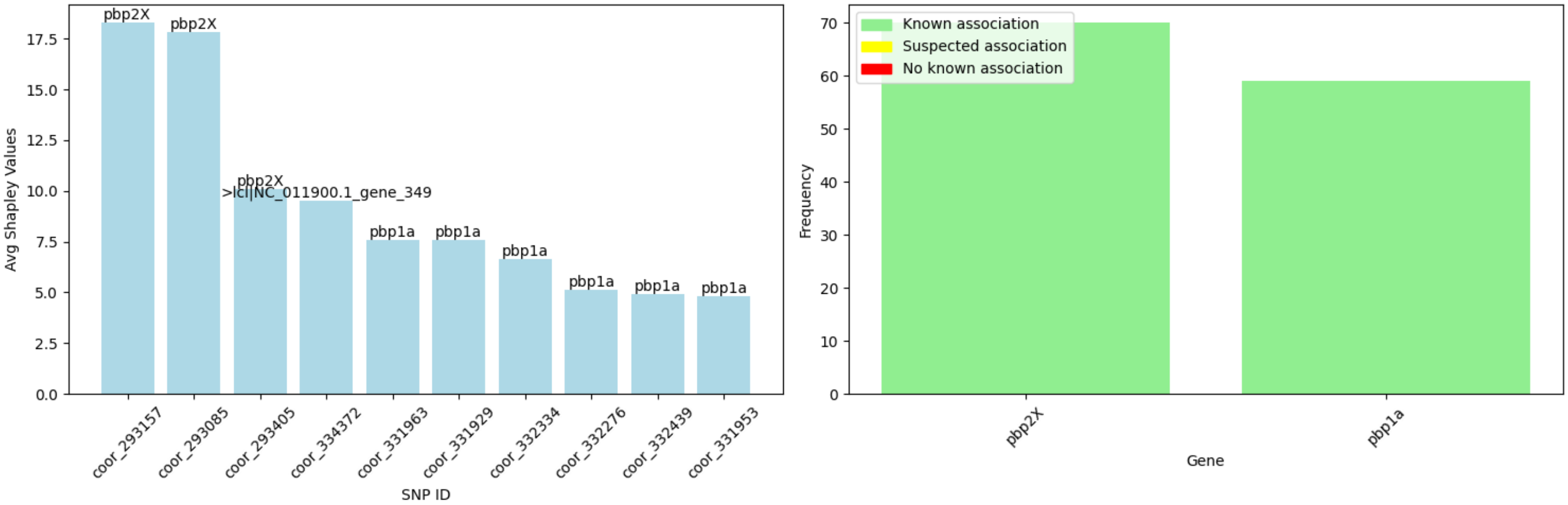

Figure 5: Review of genes identified by pipeline for penicillin resistance in *S. pneumoniae*

All the mutations identified in the top 10 for the Maela *S. pneumoniae* dataset correspond to either pbp1a or pbp2X, both of which correspond to mutations in the penicillin-binding protein, which are well-known to be associated with resistance to penicillin[13]. Thus, across both *Streptococcus pneumonia* and *Mycobacterium tuberculosis*, our pipeline is able to identify mutations within genes known to be or suspected to be associated with antibiotic resistance. Further, our pipeline is additionally able to detect cross-resistance patterns.

The following table and scatterplot summarise our results for the top 5 identified mutations across all 13 antibiotics in the cryptic dataset.

AMI	**BDQ**	**CFZ**	**DLM**	**EMB**	**ETH**
rrs	rpoB	Rv0383c	PPE21	katG	fabG1
rrs	PE_PGRS59	Rv1205	Rv0383c	embB	rpoB
Rv0383c	Rv0383c	Rv0064	katG	rpoB	Rv1729c
fadD30	kdpE	PE_PGRS59	rpoA	embB	fabG1
nadD	PE_PGRS22	mpt53	pstP	embB	fabG1

Table 4: Genes for top 5 mutations identified across 13 antibiotics for *M. tuberculosis*

INH	KAN	LEV	LZD	MXF	RFB
katG	rrs	gyrA	rpoB	gyrA	rpoB
fabG1	rpoB	PPE3	gyrA	katG	katG
katG	rpoB	Rv0064	gyrA	rpoB	rpoB
katG	rpsL	gyrA	Rv0197	lprO	rpoB
fabG1	eis	gyrA	PPE21	PPE3	rpoB

Table 5: Genes for top 5 mutations identified across 13 antibiotics for *M. tuberculosis*

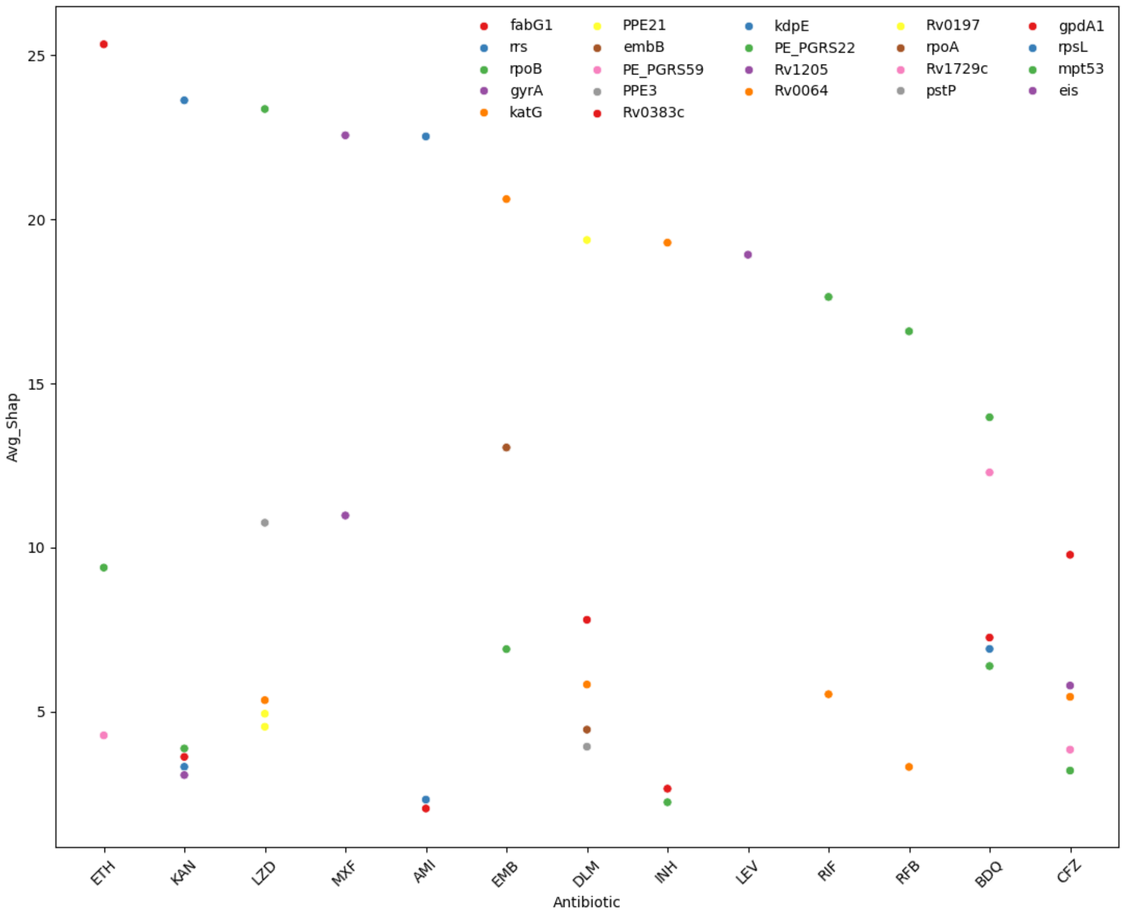

Figure 6: Top 5 mutations and associated genes identified across all antibiotics for *M. tuberculosis*

6 Discussion

The developed pipeline exhibits proficiency in identifying genes and mutations closely linked to antibiotic resistance within the context of a WGS-MIC dataset. It effectively prioritizes these mutations based on their relative contributions to the conferred resistance. This pipeline holds substantial potential for applications, particularly in the identification of resistance mechanisms pertaining to rarely used antibiotics and bacterial species that are less extensively studied. While not intended as a definitive mechanism for identifying novel resistance-related mutations in pathogenic bacteria, the pipeline may serve as a valuable guide, highlighting high-priority genes for future exploration through Genome Wide Association Studies (GWAS) and wet lab studies.

Despite the pipeline's success in its current form, it is noteworthy that the computational cost is relatively high, particularly in the calculation of Shapley values. On average, the pipeline requires around 2 hours to run on the Cryptic dataset, encompassing *M. tuberculosis* data with approximately 4.4 million base pairs. Challenges may arise when applying the pipeline to larger genomes, such as the human genome. However, the pipeline exhibits potential for optimization through parallelization and the proposed implementation of GPU-accelerated TreeShap[21], promising significant acceleration of processing times.

While our focus has been on analysing mutations associated with antibiotic resistance, the pipeline holds broader applicability to any genotypic dataset with phenotypic labels. Within the antibiotic resistance phenotype domain, the pipeline can be tested with different representations of genotypes and mutations, including k-mers, unitigs, and gene expression profiles. Moreover, the pipeline's versatility extends beyond bacterial datasets, offering the potential to analyze WGS data from diverse species, including humans. This provides an avenue for investigating genetic mechanisms underlying various phenotypic characteristics, including genetic diseases and cancer susceptibility.

To enhance the pipeline's capabilities, future iterations could incorporate more advanced feature selection or dimensionality reduction methods, such as principal component analysis analysis[31] and other techniques addressing the high levels of linkage disequilibrium in bacterial populations. These methods may include the selection of ”top-correlated” SNPs[14], iterative approaches like ABESS and Hungry Hungry SNPos (HHS)[24], median markers and haplotype blocks[1], or statistical methods like F-scores and Kendall's Rank Coefficient. Such enhancements would contribute to the robustness and adaptability of the pipeline for a broader range of genetic analyses.

7 Conclusion

In conclusion, this research introduces a novel pipeline at the intersection of machine learning and statistical analysis, offering a comprehensive framework for antibiotic resistance prediction in *Streptococcus pneumoniae* and *Mycobacterium tuberculosis*. By integrating interpretability through Shapley values, our approach not only demonstrates predictive accuracy but also ensures clinical relevance, providing a nuanced understanding of the genomic contributions to antibiotic resistance. The successful application of our pipeline to identify key mutations, validated through an extensive literature review, underscores its efficacy in flagging known resistance-associated genes and uncovering potential novel mutations. The ability to detect cross-resistance patterns further highlights the pipeline's robustness. Beyond antibiotic resistance, the versatility of our pipeline opens avenues for analyzing diverse genotypic datasets across species, promising broader applications in genetic research. As we move forward, the integration of advanced feature selection and dimensionality reduction methods stands as a potential enhancement, reinforcing the adaptability and reliability of our pipeline for a spectrum of genetic analyses.

References

[1] Mairead L Bermingham et al. "Application of high-dimensional feature selection: evaluation for genomic prediction in man". In: *Scientific reports* 5.1 (2015), p. 10312.

[2] Phelim Bradley et al. "Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis". In: *Nature communications* 6.1 (2015), p. 10063.

[3] Leo Breiman. "Random forests". In: *Machine learning* 45 (2001), pp. 5–32.

[4] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* 2016, pp. 785–794.

[5] Claire Chewapreecha et al. "Dense genomic sampling identifies highways of pneumococcal recombination". In: *Nature genetics* 46.3 (2014), pp. 305–309.

[6] Rayan Chikhi, Antoine Limasset, and Paul Medvedev. "Compacting de Bruijn graphs from sequencing data quickly and in low memory". In: *Bioinformatics* 32.12 (2016), pp. i201–i208.

[7] Keira A Cohen et al. "Paradoxical Hypersusceptibility of Drug-resistant Mycobacteriumtuberculosis to β-lactam Antibiotics". In: *EBioMedicine* 9 (2016), pp. 170–179.

[8] ST Cole et al. "Barry 3rd". In: *CE, Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, MA, Rajandream, MA, Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, JE, Taylor, K., Whitehead, S., Barrell, BG* (1998), pp. 537–544.

[9] CRyPTIC Consortium. "A data compendium associating the genomes of 12,289 Mycobacterium tuberculosis isolates with quantitative resistance phenotypes to 13 antibiotics". In: *PLoS biology* 20.8 (2022), e3001721.

[10] Nicholas J Croucher et al. "Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone Streptococcus pneumoniae Spain23F ST81". In: *Journal of bacteriology* 191.5 (2009), pp. 1480–1489.

[11] Sorin Drăghici and R Brian Potter. "Predicting HIV drug resistance with neural networks". In: *Bioinformatics* 19.1 (2003), pp. 98–107.

[12] Alexandre Drouin et al. "Interpretable genotype-to-phenotype classifiers with performance guarantees". In: *Scientific reports* 9.1 (2019), p. 4071.

[13] Thorsten Grebe and Regine Hakenbeck. "Penicillin-binding proteins 2b and 2x of Streptococcus pneumoniae are primary resistance determinants for different classes of beta-lactam antibiotics". In: *Antimicrobial agents and chemotherapy* 40.4 (1996), pp. 829–834.

[14] Giorgio Guzzetta, Giuseppe Jurman, and Cesare Furlanello. "A machine learning pipeline for quantitative phenotype prediction from genotype data". In: *BMC bioinformatics* 11.8 (2010), pp. 1–9.

[15] Nguyen Thi Le Hang et al. "Whole genome sequencing, analyses of drug resistance-conferring mutations, and correlation with transmission of Mycobacterium tuberculosis carrying katG-S315T in Hanoi, Vietnam". In: *Scientific reports* 9.1 (2019), p. 15354.

[16] JoAnn Hoskins et al. "Genome of the bacterium Streptococcus pneumoniae strain R6". In: *Journal of bacteriology* 183.19 (2001), pp. 5709–5717.

[17] Caroline Lavender et al. "Molecular characterization of isoniazid-resistant Mycobacterium tuberculosis isolates collected in Australia". In: *Antimicrobial agents and chemotherapy* 49.10 (2005), pp. 4068–4074.

[18] Wei-Yin Loh. "Classification and regression trees". In: *Wiley interdisciplinary reviews: data mining and knowledge discovery* 1.1 (2011), pp. 14–23.

[19] Deneke H Mariam et al. "Effect of rpoB mutations conferring rifampin resistance on fitness of Mycobacterium tuberculosis". In: *Antimicrobial agents and chemotherapy* 48.4 (2004), pp. 1289–1294.

[20] Fernanda Maruri et al. "A systematic review of gyrase mutations associated with fluoroquinolone-resistant Mycobacterium tuberculosis and a proposed gyrase numbering system". In: *Journal of Antimicrobial Chemotherapy* 67.4 (2012), pp. 819–831.

[21] Rory Mitchell, Eibe Frank, and Geoffrey Holmes. "GPUTreeShap: massively parallel exact calculation of SHAP scores for tree ensembles". In: *PeerJ Computer Science* 8 (2022), e880.

[22] World Health Organization et al. *Antimicrobial resistance: global report on surveillance.* World Health Organization, 2014.

[23] Mario C Raviglione and Ian M Smith. "XDR tuberculosis—implications for global public health". In: *New England Journal of Medicine* 356.7 (2007), pp. 656–659.

[24] KO Reshetnikov et al. "Feature selection and aggregation for antibiotic resistance GWAS in Mycobacterium tuberculosis: a comparative study". In: *bioRxiv* (2022), pp. 2022–03.

[25] Deisy MGC Rocha et al. "Heterogeneous streptomycin resistance level among Mycobacterium tuberculosis strains from the same transmission cluster". In: *Frontiers in Microbiology* 12 (2021), p. 659545.

[26] Benedek Rozemberczki et al. "The shapley value in machine learning". In: *arXiv preprint arXiv:2202.05594* (2022).

[27] Viola Schleusener et al. "Mycobacterium tuberculosis resistance prediction and lineage classification from genome sequencing: comparison of automated analysis tools". In: *Scientific reports* 7.1 (2017), p. 46327.

[28] Andreas Steiner et al. "KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes". In: *BMC genomics* 15 (2014), pp. 1–12.

[29] Javier Suarez et al. “Antibiotic resistance in Mycobacterium tuberculosis: peroxidase intermediate bypass causes poor isoniazid activation by the S315G mutant of M. tuberculosis catalase-peroxidase (KatG)”. In: *Journal of Biological Chemistry* 284.24 (2009), pp. 16146–16155.

[30] Gang Sun et al. “Dynamic population changes in Mycobacterium tuberculosis during acquisition and fixation of drug resistance in patients”. In: *The Journal of infectious diseases* 206.11 (2012), pp. 1724–1733.

[31] Svante Wold, Kim Esbensen, and Paul Geladi. “Principal component analysis”. In: *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), pp. 37–52.

[32] Hooman Zabeti et al. “INGOT-DR: an interpretable classifier for predicting drug resistance in M. tuberculosis”. In: *Algorithms for Molecular Biology* 16.1 (2021), p. 17.

[33] Guangyi Zhang and Aristides Gionis. “Regularized impurity reduction: accurate decision trees with complexity guarantees”. In: *Data mining and knowledge discovery* 37.1 (2023), pp. 434–475.