[Thesis Title Here]

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of
Master of Science (MSc)
in
[Department Name]

[Your Name]
[Student ID]
[University Name]

Month, Year

Declaration

I hereby declare that the work presented in this thesis is my own and has not been submitted elsewhere for the award of any degree.

[Your Name]
[Date]

Abstract

Acknowledgement

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Antimicrobial resistance (AMR) represents one of the most pressing global health challenges of the 21st century. The World Health Organization has identified AMR as a top-ten global public health threat, with projections estimating that bacterial infections could result in approximately 10 million deaths annually by 2050 if current trends continue [1]. In 2019 alone, antibiotic-resistant bacteria directly caused 1.27 million deaths, with an additional 4.95 million deaths associated with drug-resistant infections [2]. Critical priority pathogens including *Klebsiella pneumoniae*, *Escherichia coli*, *Pseudomonas aeruginosa*, and *Salmonella enterica* account for a substantial proportion of these deaths, with carbapenem-resistant strains posing particular clinical challenges.

Traditional culture-based antimicrobial susceptibility testing (AST) remains the gold standard for determining resistance profiles but suffers from significant limitations. Most culturable bacteria require 24–48 hours of incubation for detection, with pathogen identification adding another 2–4 hours. If AMR is suspected, phenotypic AST extends the timeline by an additional 18–24 hours, resulting in a total turnaround time of 2–4 days [3]. This delay can prove critical in severe infections where rapid administration of appropriate antimicrobials significantly improves patient outcomes—the risk of death doubles if effective antibiotics are not administered within 24 hours in cases of bacteremia [4].

The advent of whole genome sequencing (WGS) and the establishment of comprehensive resistance gene databases have enabled computational approaches to AMR prediction. Machine learning (ML) methods have emerged as powerful tools for predicting resistance phenotypes from genomic data, offering the potential for rapid, accurate predictions that could guide empirical therapy selection before traditional

AST results become available [5]. This thesis presents a resistance gene-based machine learning framework that addresses critical methodological gaps in current AMR prediction approaches, demonstrating that optimized gene-based models can achieve performance competitive with or superior to whole-genome sequence-based methods while maintaining interpretability and computational efficiency.

Figure 1.1 illustrates the difference between traditional culture-based prediction and modern ML-based genomic prediction.

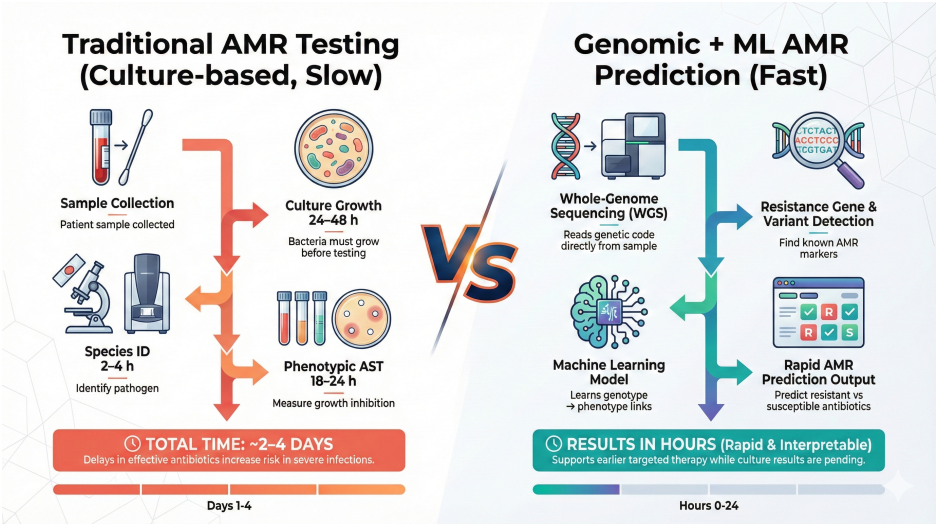

Figure 1.1: Traditional AMR testing vs genomic + ML AMR prediction

1.2 Antimicrobial Resistance: A Global Health Crisis

Antimicrobial resistance occurs when bacteria, viruses, fungi, and parasites evolve mechanisms to survive exposure to antimicrobial drugs that previously killed them or inhibited their growth. For bacteria, resistance mechanisms include enzymatic degradation (e.g., β-lactamases hydrolyzing penicillins and cephalosporins), modification of drug targets (e.g., mutations in DNA gyrase conferring fluoroquinolone resistance), efflux pump overexpression, and reduced membrane permeability [6].

Resistant infections lead to longer hospital stays, higher medical costs, and increased mortality. Patients with drug-resistant infections face a 64% higher risk of death compared to those with susceptible infections [7]. Beyond individual outcomes, AMR threatens the effectiveness of organ transplantation, cancer chemotherapy, and complex surgeries that rely on antimicrobial prophylaxis.

Economically, AMR poses global risks. The World Bank estimates that AMR could reduce global GDP by 1.1% to 3.8% by 2050, with annual costs reaching $3.4 trillion [8]. In the United States alone, resistant infections cost the healthcare system an estimated $20 billion annually, plus $35 billion in productivity losses [9].

1.3 Machine Learning Approaches for AMR Prediction

The application of machine learning to AMR prediction has evolved through distinct methodological paradigms over the past decade. Early approaches relied on rule-based systems that queried databases for known resistance genes and mutations. While these deterministic methods achieve high specificity for well-characterized mechanisms, they cannot identify novel resistance determinants absent from reference databases [10].

Classical ML models such as Random Forest, Support Vector Machines, Gradient Boosted Trees, and Logistic Regression have demonstrated strong predictive performance from genomic features, achieving accuracies ranging from 0.81 to 0.97 across pathogen–antibiotic combinations [11, 12].

Two dominant feature representation paradigms exist:

1. Resistance Gene-Based Features Binary matrices encoding presence/absence of known AMR genes, providing interpretability and computational efficiency but limited by database completeness.

2. Whole Genome Sequence (WGS)-Based Features High-dimensional features including k-mers, SNPs, and pan-genome content, capable of capturing novel mechanisms but computationally expensive and less interpretable.

Despite notable progress, gaps remain: simplistic binary encodings, weak feature selection strategies, limited handling of class imbalance, and overreliance on individual classifiers. These limitations motivate the enhanced framework proposed in this thesis.

1.4 Motivation

The motivation for this research stems from several critical observations in the current AMR prediction landscape:

- First, existing resistance gene-based approaches, while interpretable and efficient, often underperform compared to WGS-based methods. The Sunuwar & Azad framework [13] , a foundational gene-based approach, achieves F1-scores

around 0.90 but does not incorporate advanced feature engineering, hybrid selection strategies, or optimized ensemble architectures. The question arises: can methodological enhancements to the gene-based paradigm close the performance gap with WGS approaches while retaining interpretability?

- Second, WGS-based approaches, despite high reported accuracies, often exhibit a concerning accuracy-sensitivity tradeoff. The Noman et al. BioWeka framework [14] reports ≥98% accuracy but sensitivity as low as 62% for some antibiotics, meaning a substantial proportion of resistant isolates are misclassified as susceptible. In clinical contexts, such false negatives can lead to treatment failures with potentially fatal consequences. This raises the question: are high-accuracy WGS models actually superior when balanced performance metrics are considered?

- Third, the cumulative effect of resistance gene carriage remains unexplored as a predictive feature. Biological evidence suggests that isolates harboring multiple resistance genes may exhibit different resistance profiles than those with single genes due to synergistic effects, redundant protection pathways, or co-selection pressures. Yet no existing study incorporates an aggregate measure of resistance gene burden.

- Fourth, class imbalance is a common characteristic of AMR datasets, with many antibiotic-pathogen combinations exhibiting skewed resistant/susceptible ratios. Standard oversampling techniques may generate biologically implausible synthetic samples when applied to sparse binary gene matrices, and advanced hybrid methods combining oversampling with cleaning steps have received limited attention in this domain.

These observations motivate the development of an enhanced resistance gene-based framework that addresses each limitation through targeted methodological innovations: the R-Score for cumulative gene burden quantification, hybrid ANOVA-XGBoost feature selection, SMOTETomek for class imbalance handling, and the R-Blend weighted ensemble architecture.

1.5 Research Objectives

The primary objectives of this thesis are:

1. To develop and evaluate a resistance gene-based ML pipeline (R-Blend) for binary AMR prediction across multiple antibiotic–pathogen datasets.

2. To investigate whether an engineered cumulative gene-burden feature (R-Score) improves prediction performance.

3. To evaluate multiple hybrid feature selection variants (ANOVA + RF/XGB, union/intersection, different thresholds) and select an optimal configuration for resistance gene-based AMR prediction.

4. To compare several resampling strategies for class imbalance in resistance-gene data and select an effective method to improve resistant-class detection in the final model.

1.6 Challenges

The development of an effective resistance gene-based AMR prediction framework faces challenges across data and methodology dimensions.

1.6.1 Data-Related Challenges

- **Class imbalance**: Many antibiotic-pathogen datasets exhibit skewed resistant/susceptible ratios, which can bias naïve models toward the majority (susceptible) class and lead to poor detection of resistant isolates, the clinically critical class.

- **Sparse, high-dimensional gene matrices**: Resistance gene presence/absence data are binary, sparse, and often high-dimensional, making models prone to overfitting and sensitive to feature selection strategies.

- **Small dataset sizes**: Some antibiotic–pathogen combinations contain very few resistant isolates or very small sample sizes overall, leading to unstable estimates and occasional "perfect" scores that must be interpreted cautiously.

- **Genotype–phenotype mismatch**: The presence of a known resistance gene does not always translate into phenotypic resistance due to expression levels, regulatory mutations, or unknown modifiers, introducing label noise into the training data.

1.6.2 Methodological Challenges

- **Sensitivity vs. Overall Accuracy:** Clinically, missing resistant isolates (false negatives) is more serious than overcalling resistance. Designing models that

prioritize recall for the resistant class while maintaining acceptable precision and accuracy is non-trivial.

- **Identifying informative features under sparsity:** Identifying a compact but informative subset of genes from hundreds of candidates, without introducing data leakage or discarding subtle but important signals, is challenging in sparse binary spaces.

- **Handling Imbalance Safely:** Many resampling methods (e.g., SMOTE variants) can generate biologically implausible synthetic gene profiles in high-dimensional 0/1 data. Choosing a method that improves minority-class performance without distorting the data distribution requires careful evaluation.

- **Balancing Performance and Interpretability:** Powerful models like XGBoost can yield high accuracy but may operate as black boxes. The challenge is to design an approach that remains interpretable at the gene level while still being competitive with existing baselines.

1.6.3 Broader Considerations

Beyond the scope of this thesis, there are broader considerations for the eventual clinical deployment of AMR prediction models. These include generalizability across different hospitals and geographic regions, the need for prospective validation, integration with laboratory information systems, and ensuring clinician trust through transparent explanations. Additionally, resistance gene-based models inherently depend on the completeness of underlying gene annotations, meaning emerging or rare mechanisms may be underrepresented.

1.7 Contributions of This Thesis

This thesis makes the following contributions to the field of machine learning for antimicrobial resistance prediction:

- **R-Score**: A Novel Engineered Feature for Resistance Gene Burden. We introduce the R-Score, a normalized measure of cumulative resistance gene content capturing aggregate resistance-gene burden. Across the 12 Sunuwar et al. [13] datasets, adding R-Score improved average F1-score from 0.937 to 0.948 and average recall from 0.940 to 0.952, with larger gains observed on smaller datasets.

- **Hybrid ANOVA–XGBoost Feature Selection with Union-Based Integration:** We propose and evaluate hybrid feature selection variants that combine ANOVA F-test filtering ($p \leq 0.30$) with XGBoost importance (85% cumulative cutoff), using union-based integration to retain features supported by either linear or nonlinear relevance. The selected hybrid configuration consistently outperformed single-method selection across datasets.

- **Systematic Evaluation of SMOTETomek for Sparse AMR Gene Matrices:** We systematically compare multiple resampling strategies for imbalanced resistance-gene data and show that SMOTETomek provides the most stable improvement in resistant-class detection, outperforming standard over/undersampling and avoiding degradation seen in aggressive cleaning-based methods such as SMOTE-ENN.

- **R-Blend:** A Weighted Soft-Voting Ensemble Architecture. We develop and optimize R-Blend, a weighted soft-voting ensemble combining Decision Tree (weight=1), Logistic Regression (weight=1.5), and XGBoost (weight=1). The ensemble consistently improved balanced performance over individual base models and produced more stable results across heterogeneous datasets.

- **Comprehensive Benchmarking Against Gene-Based and WGS-Based Approaches:** We provide direct comparative evaluation on two benchmark settings.

    1. On the Sunuwar et al. datasets (12 antibiotic-pathogen combinations) R-Blend achieved average F1-score of 0.948, outperforming Sunuwar et al.'s [15] best classifiers by +6.3 percentage points (baseline F1: 0.885).
    2. On the Noman et al. dataset (12 antibiotics, P. aeruginosa): R-Blend achieved average F1-score of 0.959 and sensitivity of 94.1%, outperforming the WGS-based BioWeka approach by +12.6 pp F1 (BioWeka F1: 0.832) and +9.7 pp sensitivity (BioWeka: 84.4%). These results demonstrate that carefully engineered resistance gene-based models can achieve balanced performance superior to both existing gene-based frameworks and WGS-based approaches while maintaining interpretability and computational efficiency.

1.8 Organization of the Thesis

This thesis is organized into seven chapters as follows:

- **Chapter 1: Introduction**. This chapter provides an overview of the AMR crisis, introduces machine learning approaches for resistance prediction, presents the motivation and objectives of the research, discusses challenges, and outlines the contributions of this thesis.

- **Chapter 2: Theoretical Background**.This chapter presents the foundational concepts underlying this research, including antimicrobial resistance mechanisms, genomic data representation, machine learning algorithms employed, feature selection methods, class imbalance handling techniques, and ensemble learning approaches.

- **Chapter 3: Literature Review**.This chapter reviews the evolution of computational AMR prediction from rule-based detection to machine learning approaches. It covers AMR gene databases and detection tools, resistance gene-based and WGS-based prediction paradigms, and identifies critical gaps in existing methodologies that motivate the present work.

- **Chapter 4: Methodology**. This chapter presents the detailed methodology of the proposed framework, including data acquisition and preprocessing, R-Score feature engineering, hybrid ANOVA-XGBoost feature selection, SMOTETomek class imbalance handling, R-Blend ensemble architecture, experimental design, and evaluation metrics.

- **Chapter 5: Experimental Results**.This chapter presents comprehensive experimental results including dataset-level performance, ablation studies quantifying the contribution of each pipeline component, and comparative analysis against the Sunuwar & Azad and Noman et al. approaches.

- **Chapter 6: Conclusion and Future Work**. This chapter summarizes the key findings, discusses limitations of the current work, and outlines directions for future research.

- **Chapter 7: References**.This chapter provides the complete list of references cited throughout the thesis.

1.9 Summary

Antimicrobial resistance poses a critical threat to global health, with millions of deaths attributed annually to drug-resistant infections. Traditional AST methods, while reliable, are too slow to guide early empirical therapy in severe infections. Machine

learning approaches offer the potential for rapid AMR prediction from genomic data, but existing methods suffer from methodological limitations including lack of aggregate feature engineering, primitive feature selection, inadequate class imbalance handling, and suboptimal ensemble architectures.

This thesis addresses these gaps through a comprehensive resistance gene-based machine learning framework incorporating four key innovations: the R-Score for cumulative gene burden quantification, hybrid ANOVA-XGBoost feature selection with union-based integration, SMOTETomek resampling for class imbalance, and the R-Blend weighted soft-voting ensemble. The framework is evaluated across 12 antibiotic–pathogen combinations spanning carbapenems, aminoglycosides, and clindamycin resistance in K. pneumoniae, E. coli/Shigella, P. aeruginosa, S. enterica, and C. jejuni.

The proposed approach achieves an average F1-score of 0.95, outperforming both the baseline gene-based framework of Sunuwar & Azad (+5.5 percentage points F1) and the WGS-based BioWeka approach of Noman et al. (+12.7 percentage points F1). These results demonstrate that carefully engineered resistance gene-based models can achieve balanced performance superior to whole-genome approaches while maintaining the interpretability and computational efficiency essential for practical application.

Chapter 2

THEORETICAL BACKGROUND

2.1 Overview

Antimicrobial Resistance (AMR) prediction from genomic data represents a critical intersection of machine learning and computational biology. This chapter provides the theoretical foundation for machine learning techniques, feature engineering, ensemble learning, and interpretability frameworks employed in this research for predicting AMR from resistance gene profiles.

2.2 Machine Learning Fundamentals

2.2.1 Supervised Learning

Supervised learning learns a mapping function $f : X \to Y$ from labeled data $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where x_i is the feature vector and $y_i \in \{0, 1\}$ is the class label (1 = resistant, 0 = susceptible). The dataset is partitioned into training and testing sets using stratified splitting to preserve class proportions. Stratified k-fold cross-validation assesses performance by training k times on different partitions, averaging results across iterations.

2.3 Feature Engineering and Selection

2.3.1 Feature Engineering

Feature engineering transforms raw data to improve model performance. Min-Max normalization scales features to $[0, 1]$: $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$. Derived features capture

domain knowledge through aggregation (sum, mean, count), ratios, interactions, or binning.

2.3.2 Feature Selection

Feature selection addresses dimensionality, computational efficiency, interpretability, and noise reduction. This research employs a hybrid approach combining filter and embedded methods.

ANOVA F-test: Evaluates features independently using the F-statistic: $F = \frac{\text{Between-group variability}}{\text{Within-group variability}}$. Higher F-statistics indicate greater class separation with statistical significance.

XGBoost Gain-Based Importance: Measures average loss function improvement: Importance(feature) = $\sum$(Gain from splits). Features are ranked and selected until cumulative importance reaches a threshold (e.g., 85%).

Hybrid Strategy: Combining methods (union approach) leverages statistical univariate relationships and multivariate interactions for comprehensive coverage.

2.4 Class Imbalance Handling

2.4.1 The Imbalance Problem

Class imbalance occurs when the majority class significantly outnumbers the minority class. Imbalance ratio = $\frac{\text{Majority samples}}{\text{Minority samples}}$. Ratios $> 1.5 : 1$ are imbalanced, causing models to bias toward the majority class and poorly predict the critical minority class.

2.4.2 SMOTE and SMOTE-Tomek

SMOTE (Synthetic Minority Over-sampling Technique): Generates synthetic minority samples by interpolating between existing instances. For sample x with nearest neighbor x_{nn}: $x_{\text{synthetic}} = x + \lambda \times (x_{nn} - x)$, where $\lambda \in [0, 1]$. SMOTE creates diverse samples but may generate unrealistic instances in sparse spaces.

SMOTE-Tomek: Combines SMOTE with Tomek links removal. A Tomek link is a pair (x_i, x_j) from different classes that are mutual nearest neighbors. The process: (1) Apply SMOTE, (2) Identify Tomek links, (3) Remove linked samples. This balances classes while improving boundary clarity, removing noisy samples, and reducing overfitting—particularly effective for high-dimensional, sparse data.

2.5 Classification Algorithms

2.5.1 Logistic Regression

Models binary outcome probability using the sigmoid function: $P(y = 1 \mid x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}}$. Creates linear decision boundaries, is computationally efficient, produces calibrated probabilities, and offers interpretable coefficients but cannot capture complex non-linear patterns.

2.5.2 Support Vector Machine (SVM)

Finds the optimal separating hyperplane maximizing margin: $f(x) = \text{sign}(w \cdot x + b)$. The kernel trick maps inputs to higher dimensions for non-linear separation. The RBF kernel $K(x, x') = \exp(-\gamma\|x - x'\|^2)$ handles non-linearity with tunable smoothness. Effective in high-dimensional spaces but computationally expensive for large datasets.

2.5.3 Decision Tree

Recursively partitions data using impurity measures. Gini impurity: $\text{Gini}(S) = 1 - \sum p_i^2$. Entropy: $\text{Entropy}(S) = -\sum p_i \times \log_2(p_i)$. Trees are interpretable, require minimal preprocessing, and capture non-linear relationships but are prone to overfitting and instability.

2.5.4 Random Forest

Ensemble of decision trees using bootstrap sampling and random feature selection. Algorithm: (1) Create B bootstrap samples, (2) For each sample, grow a tree selecting m random features at each split, (3) Aggregate predictions via majority voting. Feature importance: Importance $= \frac{1}{B}\sum$ Decrease in impurity. Reduces overfitting, provides robust importance estimates, handles high-dimensional data well but is less interpretable than single trees.

2.5.5 XGBoost (eXtreme Gradient Boosting)

Sequential tree ensemble where each tree corrects previous errors. Optimizes regularized objective: $\text{Obj} = \sum L(y_i, \hat{y}_i) + \sum \Omega(f_t)$, where $\Omega(f_t) = \gamma T + \frac{\lambda}{2}\|w\|^2$ penalizes complexity (T = leaves, w = weights). Key features: L1/L2 regularization, tree pruning, missing value handling, column/row subsampling. Important hyperparameters: n_estimators, max_depth, learning_rate, subsample, colsample_bytree, gamma,

lambda. Achieves state-of-the-art performance with built-in regularization but requires careful tuning.

2.6 Ensemble Learning

2.6.1 Ensemble Principles

Combines multiple models to reduce variance (averaging predictions), reduce bias (capturing diverse patterns), and improve robustness. Effectiveness requires diversity—models making different errors through different algorithms, training subsets, feature subsets, or hyperparameters.

2.6.2 Weighted Soft Voting

Predicts the class with highest average probability: $\hat{y} = \arg\max \sum w_i \times P_i(y = c \mid x)$, where w_i are importance weights ($\sum w_i = 1.0$) based on validation performance (e.g., F1-score). Leverages probability confidence, assigns greater influence to better models, and combines diverse algorithm strengths.

2.7 Model Evaluation Metrics

2.7.1 Confusion Matrix and Basic Metrics

The confusion matrix contains True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ measures overall correctness but misleads with imbalanced data.

Precision = $\frac{TP}{TP+FP}$ measures positive prediction correctness (important when false positives are costly).

Recall = $\frac{TP}{TP+FN}$ measures actual positive identification (critical when missing positives is costly).

F1-Score = $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$ harmonically balances both, penalizing extreme values and proving robust to imbalance.

2.7.2 ROC and Precision-Recall Curves

AUROC (Area Under ROC Curve): Plots True Positive Rate vs. False Positive Rate across thresholds. Represents the probability that a randomly chosen positive

ranks higher than a negative. Provides threshold-independent evaluation but may be overly optimistic for imbalanced data.

AUPRC (Area Under Precision-Recall Curve): Plots Precision vs. Recall across thresholds. Focuses on positive class performance, making it more informative than AUROC for imbalanced datasets where the minority class is critical.

2.8 Model Interpretability

2.8.1 Importance of Interpretability

Interpretability enables trust and adoption, debugging and validation, regulatory compliance, and scientific insight. Models have intrinsic interpretability (linear models, trees) or require post-hoc methods (SHAP, permutation importance).

2.8.2 Permutation Feature Importance

Measures performance degradation when a feature is randomly shuffled, breaking its relationship with the target. Algorithm: (1) Compute baseline performance, (2) For each feature, permute values and compute performance, (3) Importance = Baseline - Permuted performance. Model-agnostic, considers features contextually, but can be unreliable with correlated features.

2.8.3 SHAP (SHapley Additive exPlanations)

Assigns each feature a contribution value based on Shapley values from game theory. The SHAP value ϕ_j for feature j:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(M - |S| - 1)!}{M!} \times [f(S \cup \{j\}) - f(S)] \tag{2.1}$$

SHAP satisfies local accuracy ($f(x) = \phi_0 + \sum \phi_j$), consistency, and additivity. TreeExplainer efficiently computes exact values for tree models.

Global Explanations: Mean absolute SHAP values indicate overall feature importance.

Local Explanations: Instance-level SHAP values show which features contributed to specific predictions.

Ensemble SHAP: For weighted voting with weights w_i: $\phi_{j,\text{ensemble}} = \frac{\sum w_i \times \phi_{j,i}}{\sum w_i}$.

2.9 Hyperparameter Optimization

2.9.1 Grid Search with Cross-Validation

Hyperparameters control the learning process (learning rate, tree depth, regularization). Grid search exhaustively evaluates predefined combinations using cross-validation: (1) Train with cross-validation, (2) Compute average performance, (3) Select best combination. Guaranteed to find optimal grid combination and parallelizable but computationally expensive.

2.9.2 Cross-Validation Protocol

Ensures unbiased selection: (1) Split data into train and test sets, (2) Use cross-validation on training set for hyperparameter selection, (3) Train final model with selected hyperparameters, (4) Evaluate on held-out test set. Prevents overfitting to test set performance.

2.10 Data Leakage Prevention

2.10.1 Definition and Sources

Data leakage occurs when test set information influences training, causing overly optimistic estimates. Sources include: (1) Preprocessing applied before splitting (training "sees" test data), (2) Resampling applied to both sets (synthetic test samples use training information).

2.10.2 Correct Protocol

(1) Split data (stratified), (2) Fit preprocessing on training data only, (3) Apply preprocessing to both sets, (4) Apply resampling only to training set, (5) Train on preprocessed, resampled training set, (6) Evaluate on preprocessed (not resampled) test set. Ensures test set remains unseen and represents real-world conditions.

2.11 Binary Classification Formulation

2.11.1 Problem Definition

Binary classification assigns instances to one of two classes. Input: $x \in \mathbb{R}^n$, Output: $y \in \{0, 1\}$, Objective: Learn $f : \mathbb{R}^n \to \{0, 1\}$ minimizing prediction error.

2.11.2 Decision Threshold and Calibration

Classifiers output probability $P(y = 1 \mid x)$. Predicted class determined by threshold τ: $\hat{y} = 1$ if $P(y = 1 \mid x) \geq \tau$, else 0. Default $\tau = 0.5$, adjustable based on class imbalance and cost trade-offs.

Well-calibrated classifiers produce probabilities reflecting true likelihood. Calibration methods: Platt Scaling (logistic regression on outputs), Isotonic Regression (non-parametric monotonic function). Logistic Regression naturally produces calibrated probabilities; tree models may require calibration.

2.12 Chapter Summary

This chapter established the theoretical foundation for AMR prediction, covering supervised learning, feature engineering and selection (hybrid ANOVA-XGBoost), class imbalance handling (SMOTE-Tomek), classification algorithms (Logistic Regression, SVM, Decision Tree, Random Forest, XGBoost), weighted soft voting ensembles, evaluation metrics (accuracy, precision, recall, F1-score, AUROC, AUPRC), model interpretability (permutation importance, SHAP), hyperparameter optimization, and data leakage prevention. These concepts form the methodological foundation for the subsequent research.

2.13 Chapter Summary

This chapter established the theoretical foundation for AMR prediction, covering supervised learning, feature engineering and selection (hybrid ANOVA-XGBoost), class imbalance handling (SMOTE-Tomek), classification algorithms (Logistic Regression, SVM, Decision Tree, Random Forest, XGBoost), weighted soft voting ensembles, evaluation metrics (accuracy, precision, recall, F1-score, AUROC, AUPRC), model interpretability (permutation importance, SHAP), hyperparameter optimization, and

data leakage prevention. These concepts form the methodological foundation for the subsequent research.

Chapter 3

Literature Review

3.1 Introduction

As established in Chapter 1, the urgent need for rapid antimicrobial resistance (AMR) prediction has driven significant research into computational approaches leveraging genomic data. This chapter reviews the current state of machine learning methods for AMR prediction, tracing the evolution from rule-based systems to advanced ensemble and deep learning architectures.

The review is organized around two primary prediction paradigms: resistance gene-based approaches, which utilize binary presence/absence matrices of known AMR genes, and whole genome sequence (WGS)-based approaches, which extract features such as k-mers, SNPs, or pan-genome content from complete genomic sequences. For each paradigm, we examine representative studies, analyze their methodological choices, and critically evaluate their performance with particular attention to the often-overlooked accuracy-sensitivity tradeoff.

We also review key enabling infrastructure, including AMR gene databases (AMRFinderPlus, ResFinder, CARD) that underpin genomic prediction, as well as cross-cutting methodological considerations: feature engineering strategies, feature selection methods, class imbalance handling techniques, and ensemble architectures. Through this analysis, we identify five critical gaps in the existing literature that motivate the methodological contributions of this thesis: the absence of cumulative gene burden features, limited hybrid feature selection strategies, inadequate evaluation of advanced resampling methods for sparse gene matrices, underexplored weighted ensemble optimization, and insufficient direct comparison between gene-based and WGS-based approaches.

3.2 AMR Gene Databases and Detection Tools

Accurate genomic prediction of AMR relies on high-quality reference databases of resistance determinants. Several major databases underpin modern AMR gene detection and feature engineering for ML models.

AMRFinderPlus (NCBI)

Identifies acquired resistance genes and point mutations using BLAST and HMMs [16]. It covers a broad scope including stress response genes and virulence factors. AMRFinderPlus provides gene coverage and identity metrics to gauge prediction confidence (e.g., distinguishing complete vs. partial hits). The NCBI Pathogen Detection database integrates AMRFinderPlus outputs with phenotypic AST data, yielding labeled datasets suitable for machine learning [17].

ResFinder 4.0

(CGE) uses k-mer based alignment to detect horizontally acquired resistance genes, with PointFinder for known chromosomal mutations [8]. ResFinder offers high specificity for known resistance genes, with concordance ¿96% with phenotypic AST for well-characterized mechanisms [18]. The Comprehensive Antibiotic Resistance Database (CARD) contains over 300,000 resistance gene sequences and their known resistance phenotypes. In 2023, CARD introduced standardized 15-character Short Names for each resistance gene allele to facilitate machine learning feature encoding [19].

CARD (Comprehensive Antibiotic Resistance Database)

While these tools form the foundation for genomic AMR prediction, their coverage is inherently limited to previously known resistance determinants. Rule-based gene detection achieves high accuracy for well-characterized mechanisms but will miss novel resistance genes or mutations absent from databases [20].Additionally, the mere presence of a resistance gene does not guarantee phenotypic resistance. Factors like gene expression level, regulatory mutations, or epistatic interactions modulate the genotype-phenotype relationship [21]. These limitations motivate the use of machine learning to learn resistance patterns from data, potentially capturing signals beyond curated gene lists.

Limitations of Database-Driven Detection:

Database-driven detection captures only known mechanisms, cannot detect novel genes, and gene presence does not guarantee phenotypic resistance due to expression-level or regulatory factors [22]. These limitations motivate ML-based prediction.

3.3 Evolution of Machine Learning Approaches for AMR Prediction

The application of ML to AMR prediction has progressed through distinct stages, from simple rule-based algorithms to complex ensemble and deep learning models.

3.3.1 Rule-Based Detection (2010–2016)

Early efforts relied on deterministic rules leveraging known resistance markers. Tools like ResFinder and ARIBA (Antimicrobial Resistance Identification By Assembly) scan genomes for canonical resistance genes or mutations [23]. Davis et al. (2016) demonstrated an important proof-of-concept using the PATRIC database, encoding 31-mer DNA k-mers and training an AdaBoost classifier to predict resistance phenotypes across A. baumannii, S. aureus, and S. pneumoniae. The k-mer based model achieved 88–99% accuracy [24], highlighting that sequence data contain rich signals of resistance that ML can exploit even without explicit gene annotation.

3.3.2 Classical Machine Learning (2016–2022)

Pan-Genome Models

Her & Wu (2018) pioneered a pan-genome-based approach for E. coli using 59 strains from the PATRIC database. They constructed a pan-genome with 15,950 gene clusters (2,874 core, 13,076 accessory) and found that only 61% of known CARD resistance genes were in the accessory genome—core genes (present in every isolate) provided no discriminative power [25]. Using SVM with genetic algorithm feature selection, they achieved AUC ¿0.90 for most antibiotics [25]. However, their study had a critical limitation: the extremely small sample size (59 strains) serves only as proof-of-concept, and no class imbalance handling was applied.

Large-Scale Studies

Moradigaravand et al. (2018) performed a landmark analysis of 1,936 E. coli isolates, incorporating 90,261 accessory genes and ¿1.4 million SNPs [26]. Gradient boosted decision trees (GBDT) outperformed other classifiers (including deep neural networks), with average accuracy 0.91 (range 0.81 - 0.97) [26]. Critically, they observed a significant accuracy-sensitivity tradeoff: for amoxicillin-clavulanate, accuracy was 81% but recall for resistance was only 64%; for cefuroxime, recall dropped to 74% - meaning 26 - 36% of resistant isolates were misclassified as susceptible [26]. This study applied no class imbalance handling despite average resistance frequency of 0.35 (range 0.15 - 0.63).

Generalization Failure

Nsubuga et al. (2024) tested ML models trained on 1,509 England E. coli isolates against 170 African isolates from Uganda, Nigeria, and Tanzania [27]. They observed dramatic generalization failure: accuracy dropped from 87% to 50% for ciprofloxacin and from 92% to 45% for cefotaxime when validated on African data [27]. Despite high training accuracy, F1-scores were critically low (0.42–0.57), demonstrating the pervasive accuracy-sensitivity tradeoff in AMR prediction. Even with down-sampling for class balancing, the models failed to generalize across populations.

Meta-analysis

A systematic review by Ardila et al. (2025) synthesized findings from 21 studies evaluating ML for AMR in clinical settings [28]. On average, ensemble tree-based models performed best: Random Forest achieved mean AUROC 0.80 (range 0.58–0.98), versus 0.68 (0.50–0.83) for logistic regression [28]. The meta-analysis confirmed that integrating WGS data with AST phenotypes significantly improves prediction performance, but also noted that most models lack prospective validation [28].

3.3.3 Deep Learning Approaches (2022–Present)

Recent advances have explored deep learning and transformer architectures for AMR prediction. López-Cortés et al. (2024) developed MSDeepAMR, a deep neural network with transfer learning for MALDI-TOF mass spectrometry data, achieving AUROC ¿0.83 for predicting resistance in E. coli/K. pneumoniae [29]. He et al. (2025) proposed MCT-ARG, a multi-channel Transformer model achieving AUC-ROC of 99.23% for ARG classification [30].

Challenges: While deep learning approaches show promise, they require substantially larger training datasets and offer reduced interpretability compared to classical ML methods [31]. These are important considerations for clinical deployment where clinicians must understand which genetic factors drive predictions. Wang et al. (2025) caution that feature importance reflects correlation with predictions, not causal associations [31].

3.4 Resistance Gene-Based Prediction Methods

Resistance gene-based approaches use binary presence/absence features of known AMR genes as inputs to ML models. This paradigm offers biological interpretability (predictions can be traced to specific genes), computational efficiency (reduced feature dimensionality), and alignment with established microbiological practiced.

3.4.1 Sunuwar & Azad Framework

Sunuwar & Azad (2021) developed a foundational machine learning framework for resistance gene-based AMR prediction [13]. Their approach was unbiased in considering all protein-coding genes, not just those annotated in resistance databases. They compiled datasets from the NCBI Pathogen Detection database, including K. pneumoniae, E. coli, P. aeruginosa, Salmonella enterica, and Campylobacter jejuni, with resistance profiles for carbapenems (doripenem, ertapenem, imipenem, meropenem), aminoglycosides (kanamycin, streptomycin), and clindamycin [13]. Using multiple classifiers (LDA, SVM, Naive Bayes, Decision Trees, XGBoost), they established baseline performance around F1 ≈ 0.90 for many pathogen-antibiotic combinations [13].

3.4.2 Limitations and Research Gaps

While the Sunuwar & Azad framework provides a solid foundation, several methodological limitations present opportunities for improvement:

- **Feature Engineering Gap**: Each resistance gene is treated as a separate binary feature (present/absent), ignoring the cumulative effect of carrying multiple resistance genes [13]. An isolate with 10 resistance genes likely has a more robust resistance phenotype than one with a single gene, yet no aggregate "gene load" metric is incorporated. No existing study incorporates a resistance gene burden score despite biological rationale for its predictive value.

- **Feature Selection Gap**: Standard approaches use either univariate filtering or single-method embedded selection. Hybrid strategies combining filter and embedded methods to capture both linear and non-linear feature relevance remain unexplored in AMR contexts [13].
- **Ensemble Architecture Gap**: Most studies employ single classifiers or simple voting ensembles with equal weights. Weighted soft voting ensembles that leverage complementary strengths of different model families and account for probability calibration quality have not been systematically explored [13].
- **Class Imbalance Handling Gap**: While class imbalance is acknowledged, systematic evaluation of advanced resampling strategies specifically suited to sparse, high-dimensional binary gene matrices is limited. Hybrid methods combining synthetic sample generation with cleaning steps (e.g., SMOTE-Tomek) remain underexplored for AMR prediction [13].

3.5 Whole Genome Sequence-Based Prediction Methods

While class imbalance is acknowledged, systematic evaluation of advanced resampling strategies specifically suited to sparse, high-dimensional binary gene matrices is limited. Hybrid methods combining synthetic sample generation with cleaning steps (e.g., SMOTE-Tomek) remain underexplored for AMR prediction [13].

3.5.1 K-mer Models

Nguyen et al. (2018) built the first complete in silico MIC prediction panel for K. pneumoniae. They counted 10-mer frequencies from each genome and trained XGBoost models to predict susceptibility for 20 antibiotics, achieving average accuracy 92% with prediction time of 2 minutes per genome [32]. ValizadehAslani et al. (2020) explored amino acid k-mers as an alternative, demonstrating that 5-mer amino acid features provide comparable accuracy while producing interpretable alignments to known AMR genes [33].

3.5.2 The Noman et al. BioWeka Framework: A Critical Analysis

Noman et al. (2023) applied WGS-based ML to P. aeruginosa across 12 antimicrobial agents using the BioWeka framework [14]. Using Random Forest with 10-fold cross-validation on complete genomic sequences, they reported impressive mean accuracy ≥98% [14]. However, a critical examination of their results reveals a fundamental flaw: despite high accuracy, the model exhibits severely low sensitivity and precision for several antibiotics.

Despite reporting ≥98% accuracy, Noman et al.'s BioWeka approach exhibits: (1) Very low sensitivity for amoxicillin (62%)—meaning 38% of resistant isolates are missed, which could lead to treatment failure; (2) Suboptimal average sensitivity (84.4%) across all antibiotics; (3) Low average F1-score (83.2%) indicating poor balance between precision and recall. No class imbalance handling methods were applied.

These limitations are likely attributable to the high-dimensional feature space (millions of k-mers) causing the model to overfit to the majority class (susceptible), achieving high overall accuracy at the expense of sensitivity for the clinically critical resistant class. As Kim et al. (2022) stated: "False-negative diagnosis leads to treatment failure" [36]—making low sensitivity clinically unacceptable despite high accuracy. This raises a critical question: Can resistance gene-based approaches, using only a fraction of the genomic information with appropriate methodological enhancements, achieve superior balanced performance?

3.5.3 Comparative Considerations:

Table 2.2 summarizes the trade-offs between resistance gene-based and WGS-based approaches:

Table 3.1: Comparison of Gene-Based vs. WGS-Based AMR Prediction

Aspect	Gene-Based	WGS-Based
Feature Dimensionality	10–500 genes	Millions of k-mers
Interpretability	High	Low
Novel Mechanisms	Limited	High
Computational Cost	Low	Very High
Clinical Acceptance	High	Moderate
Sensitivity Risk	Moderate	High (overfitting)

3.6 Feature Engineering and Selection Strategies

3.6.1 Feature Engineering

Most resistance gene-based studies employ simple binary encoding: each known resistance gene is a feature set to 1 if present, 0 if absent [34]. This treats all genes independently and equally, which may not reflect biology. There is evidence that cumulative gene content affects phenotype—an isolate with multiple β-lactamase genes may exhibit higher resistance due to synergistic effects or redundant protection pathways [35].

Surprisingly, the literature reveals no exploration of aggregate gene burden metrics as engineered features. While gene load has been studied epidemiologically, its incorporation as a predictive feature for ML models remains unexplored [35]. This represents a significant gap: a normalized resistance gene count could capture important predictive signal not available from individual gene indicators alone.

3.6.2 Feature Selection Methods:

Feature selection methods fall into three categories: filter methods (univariate statistical tests like ANOVA or chi-square), wrapper methods (search-based selection), and embedded methods (selection integrated into model training such as tree-based importance) [36]. For AMR prediction, commonly employed approaches include chi-square tests for univariate association and Random Forest or XGBoost importance for non-linear relationships.

A hybrid approach combining filter and embedded methods can leverage complementary strengths. For instance, taking the union of top features by ANOVA p-value (filter) and XGBoost importance (embedded) captures genes that are either linearly associated or non-linearly predictive. However, systematic evaluation of such hybrid strategies for AMR gene matrices is absent from the literature [36].

3.7 Class Imbalance Handling in AMR Prediction

Class imbalance is endemic in AMR datasets, with resistant isolates often underrepresented. Imbalance ratios can range from nearly balanced to severely skewed (¿10:1), biasing classifiers toward majority class prediction and reducing sensitivity for resistance detection [35].

Resampling strategies include Random Oversampling, Random Undersampling, and SMOTE (Synthetic Minority Over-sampling Technique) [37]. Hybrid methods

combining oversampling with cleaning steps have shown promise: SMOTE-Tomek combines SMOTE with Tomek links removal (eliminating nearest-neighbor pairs from different classes), while SMOTE-ENN combines SMOTE with Edited Nearest Neighbors cleaning [38]. For sparse, high-dimensional binary matrices characteristic of AMR gene data, simple SMOTE may generate biologically implausible synthetic samples; cleaning steps can mitigate this risk but have not been systematically evaluated in the AMR context [38].

3.8 Ensemble Methods and Model Interpretability

3.8.1 Ensemble Architecture

Ensemble methods combining multiple base classifiers have demonstrated improved robustness. Yang & Wu (2022) showed that stacked generalization improves performance by 1.77–3.20% compared to individual models [39]. However, most studies employ simple averaging or majority voting without exploring weighted combination strategies.

Weighted soft voting ensembles allow differential contribution from models with varying reliability. Weight optimization strategies specifically for AMR prediction remain underexplored. Considerations include model calibration quality (well-calibrated probabilities are essential for effective soft voting), complementarity of model families (linear vs. tree-based), and dataset-specific performance characteristics [39].

3.8.2 Model Interpretability

Clinical acceptance requires interpretability—clinicians must understand which genetic factors drive predictions. SHAP (SHapley Additive exPlanations) has emerged as the standard for feature attribution [40]. Khaledi et al. (2020) demonstrated integrated genomic-transcriptomic prediction for P. aeruginosa with sensitivity of 0.81–0.95; SHAP analysis identified known resistance determinants (gyrA, ampC, oprD) alongside novel markers [41]. However, Wang et al. (2025) caution that feature importance reflects correlation, not causation [31].

3.9 Research Gaps and Motivation for Present Study

This literature review identifies several critical methodological gaps that motivate the present study:

- **Gap 1 – Feature Engineering:** While binary gene presence/absence matrices are standard, aggregate measures of resistance gene burden have not been explored as engineered features despite biological rationale. We introduce the R-Score, a normalized resistance gene load score to capture cumulative resistance effects.

- **Gap 2 – Hybrid Feature Selection:** Systematic evaluation of hybrid strategies combining filter methods (ANOVA) with embedded methods (XGBoost importance) is lacking. We propose union-based integration that retains features important under either modeling assumption.

- **Gap 3 – Advanced Class Imbalance Handling:** While SMOTE is occasionally applied, hybrid resampling methods combining synthesis with cleaning have not been systematically evaluated for sparse binary AMR gene matrices. We evaluate SMOTETomek specifically tailored for this data type.

- **Gap 4 – Weighted Ensemble Architectures:** Weighted soft voting ensembles optimized for AMR prediction, leveraging complementary strengths of linear and tree-based models, remain unexplored. We develop the R-Blend ensemble with optimized model weights.

- **Gap 5 – Gene-Based vs. WGS Performance Comparison:** Direct comparison of optimized resistance gene-based approaches against WGS-based methods on identical datasets is needed. We benchmark our approach against both Sunuwar & Azad (gene-based) and Noman et al. (WGS-based) to quantify trade-offs.

3.10 Chapter Summary

Table 2.3 provides a comprehensive comparison of key machine learning studies for AMR prediction, highlighting methodological approaches, performance metrics, and limitations that the present work addresses.

3.11 Chapter Summary

This chapter reviewed the landscape of machine learning approaches for antimicrobial resistance prediction, tracing the evolution from rule-based detection to sophisticated ensemble and deep learning methods. Key databases (AMRFinderPlus, ResFinder,

Table 3.2: Comparison of Machine Learning Approaches for AMR Prediction

Study	Isolates	Feature Type	Best Model	Best F1/ACC	Key Limitation
Her & Wu (2018)	59	Pan-genome	SVM+GA	AUC > 0.90	Only 59 strains
Moradigaravand (2018)	1,936	Pan-genome + SNPs	GBDT	Acc: 0.91	Recall 64–74% for some
Nguyen (2018)	≈ 400	10-mer k-mers	XGBoost	Acc: $\sim$92%	Low interpretability
Sunuwar & Azad (2021)	Varies	Resistance Genes	LDA/SVM	F1 ≈ 0.90	No single proposed model
Noman et al. (2023)	1,200	WGS (BioWeka)	RF	Acc $\geq 98\%$	Sens: 62–84%, F1: 83%
Nsubuga (2024)	1,509	WGS	SVM/LGB	Acc: 87–92%	F1: 0.42–0.57; poor generalization
Ardila (2025)	Meta	Various	RF/GBDT	AUC: 0.80	No prospective validation
Present Study	15,000+	R-Score + Genes	R-Blend	TBD	Addresses all gaps

CARD) provide the foundation for resistance gene-based prediction, while WGS-based approaches offer comprehensive but computationally intensive alternatives.

Critical analysis of existing work identified a pervasive accuracy-sensitivity tradeoff: models achieving high overall accuracy often exhibit unacceptably low sensitivity for the clinically critical resistant class. Noman et al.'s BioWeka approach, despite 98

Five methodological gaps were identified: (1) no feature engineering for cumulative gene burden, (2) no hybrid feature selection strategies, (3) limited evaluation of advanced resampling for sparse gene matrices, (4) no weighted ensemble optimization, and (5) insufficient direct comparison between gene-based and WGS approaches. The following chapter presents a comprehensive methodology addressing these gaps through the R-Score feature, hybrid ANOVA-XGBoost feature selection, SMOTE-Tomek resampling, and R-Blend weighted ensemble—demonstrating that optimized resistance gene-based approaches can achieve balanced performance superior to existing methods while maintaining interpretability and computational efficiency.

Chapter 4

Methodology

4.1 Overview

This chapter presents a comprehensive methodology for predicting Antimicrobial Resistance (AMR) using machine learning techniques applied to resistance gene profiles. The proposed approach demonstrates that resistance gene-based prediction can achieve performance comparable to, or exceeding, whole-genome sequence (WGS)-based methods while maintaining computational efficiency and biological interpretability.

The methodology includes six major stages:

1. Dataset Collection and Preprocessing
2. Feature Engineering
3. Feature Selection
4. Class Imbalance Handling
5. Model Training and Ensemble Construction
6. Model Interpretability and Explainability

The pipeline addresses challenges inherent in AMR genomic data such as high dimensionality, extreme sparsity, and class imbalance. Figure 4.1 illustrates the overall workflow.

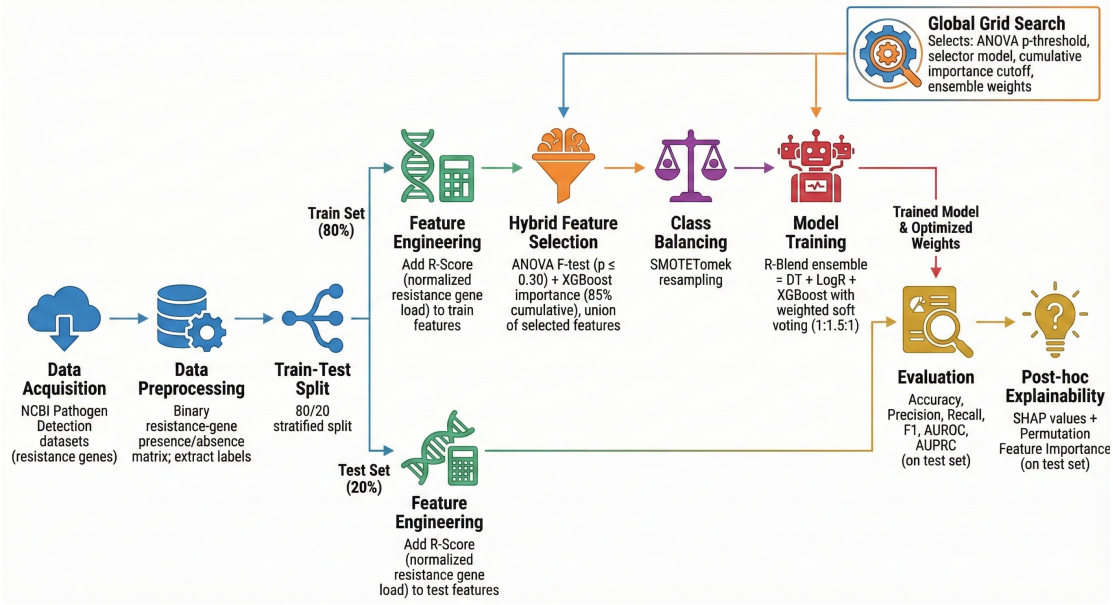

Figure 4.1: End-to-end AMR prediction pipeline of the proposed method

4.2 Dataset Collection and Preprocessing

4.2.1 Dataset Description

Unlike prior works that focus on a single antibiotic–pathogen pair, our study analyzes 12 AMR datasets from Sunuwar et al. [13], each representing a unique antibiotic–pathogen combination. To benchmark against WGS-based methods, we also used the dataset from Noman et al. [14], which includes whole-genome sequences of *Pseudomonas aeruginosa* isolates.

4.2.2 Resistance Gene-Based Datasets (Sunuwar et al.)

The first 12 datasets were collected from the NCBI Pathogen Detection portal and curated by Sunuwar et al. Each dataset contains resistance gene presence/absence information and phenotypic AMR labels across several pathogens and antibiotics:

- **K. pneumoniae (KN)**: Doripenem, Ertapenem, Imipenem, Meropenem
- **E. coli/Shigella (ECS)**: Doripenem, Ertapenem, Imipenem, Meropenem
- **P. aeruginosa (PA)**: Doripenem
- **S. enterica (SE)**: Streptomycin, Kanamycin
- **C. jejuni (CJ)**: Clindamycin

These datasets vary widely in sample size (26–1042 isolates), feature count (11–326 genes), and class distribution.

Table 4.1: Summary of the 12 resistance gene-based datasets

Dataset	Isolates	Genes	Resistant	Susceptible
Doripenem (KN)	316	325	241	75
Ertapenem (KN)	181	324	90	91
Meropenem (ECS)	91	236	45	46
Ertapenem (ECS)	129	236	61	68
Doripenem (ECS)	49	236	25	24
Imipenem (ECS)	64	236	37	27
Doripenem (PA)	44	164	22	22
Streptomycin (SE)	1042	179	542	500
Imipenem (KN)	200	324	113	87
Clindamycin (CJ)	26	43	8	18
Meropenem (KN)	238	324	106	132
Kanamycin (SE)	991	179	493	498

Figure 4.2 shows the class imbalance characteristics.

4.2.3 WGS-Based Dataset (Noman et al.)

To compare resistance gene-based and WGS-based AMR prediction, we used the dataset from Noman et al. [14], consisting of 1437 *P. aeruginosa* isolates annotated with phenotypic resistance against 12 antibiotics.

Table 4.2: Summary of Noman et al. WGS-based dataset

Drug	Isolates	Resistant	Susceptible
Ampicillin	1437	1428	9
Amoxicillin	1437	1423	14
Meropenem	1437	814	623
Cefepime	1437	1415	22
Fosfomycin	1437	1425	12
Ceftazidime	1437	1428	9
Chloramphenicol	1437	1405	32
Erythromycin	1437	36	1401
Tetracycline	1437	205	1232
Gentamycin	1437	608	829
Butirosin	1437	30	1407
Ciprofloxacin	1437	1020	417

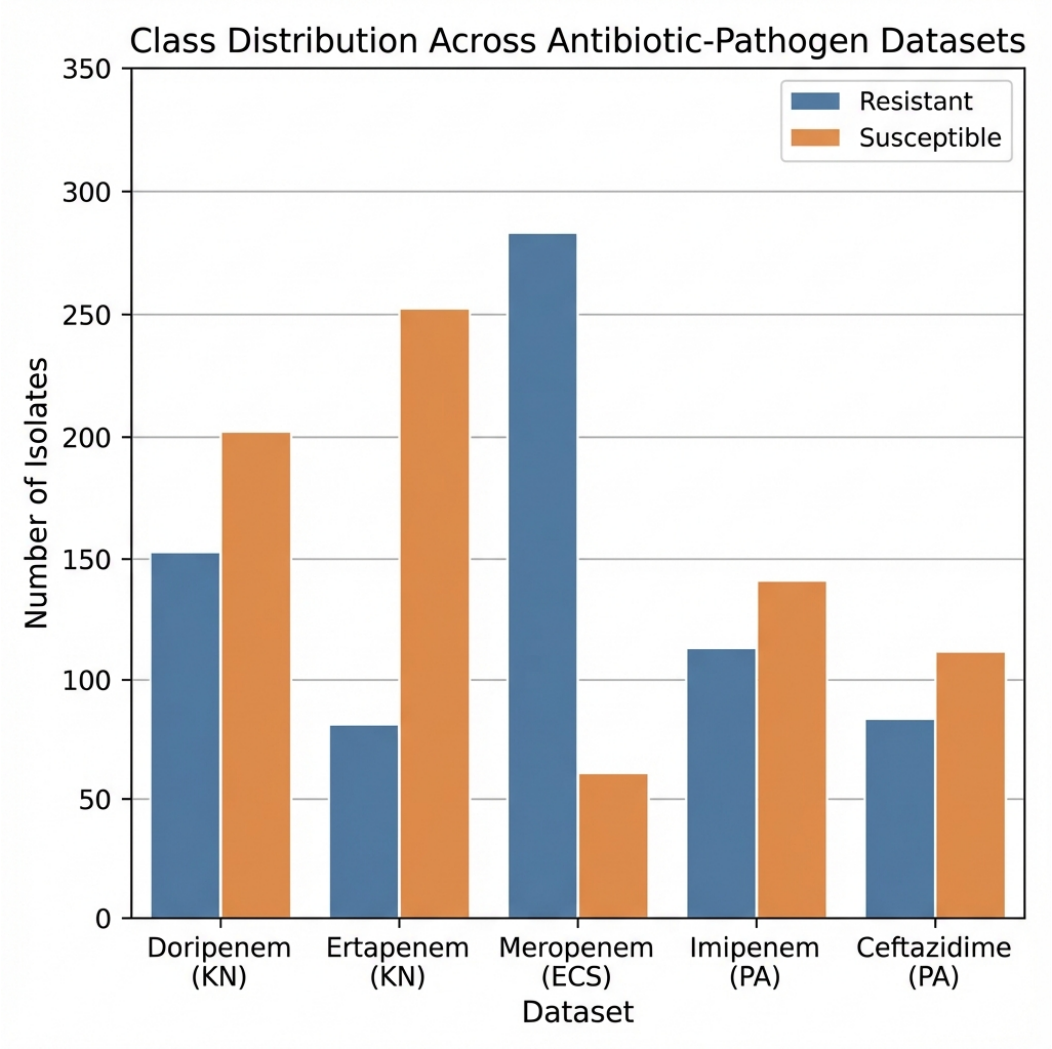

Figure 4.2: Class distribution across antibiotic–pathogen datasets

4.2.4 Binary Gene Matrix Construction

Each dataset was converted into a binary gene presence/absence matrix:

$$x_{i,j} = \begin{cases} 1, & \text{if gene } g_j \text{ is present in isolate } i \\ 0, & \text{otherwise} \end{cases}$$

Gene completeness categories (COMPLETE, PARTIAL, PARTIAL_END_OF_CONTIG) were encoded as additional binary features. The target variable was:

$$y = \begin{cases} 1 & \text{Resistant} \\ 0 & \text{Susceptible} \end{cases}$$

4.3 Feature Engineering

4.3.1 Resistance Gene Load Score (R-Score)

A novel engineered feature capturing cumulative gene burden:

$$\text{R-Score}_i = \sum_{j=1}^{n} x_{i,j}$$

Normalized using Min-Max scaling:

$$\text{R-Score}'_i = \frac{\text{R-Score}_i - \min(\text{R-Score})}{\max(\text{R-Score}) - \min(\text{R-Score})}$$

R-Score consistently improved prediction performance and separated resistant vs. susceptible classes.

4.4 Feature Selection

AMR gene datasets are high-dimensional and sparse. We adopt a hybrid, two-stage feature selection approach.

4.4.1 Stage 1: ANOVA F-test

The ANOVA F-statistic measures linear discriminative power:

$$F(g_j) = \frac{\text{Between-class variance}}{\text{Within-class variance}}$$

Features with p-value ≤ 0.30 were retained. R-Score was always retained.

4.4.2 Stage 2: XGBoost Embedded Importance

An XGBoost classifier was trained (400 trees, depth 6), and features were ranked by importance. Features contributing to the top 85% cumulative importance were selected:

$$S_{\text{XGB}} = \{g_j : \sum \text{Importance}(g_j) \leq 0.85\}$$

4.4.3 Union-Based Integration

Final feature set:

$$S_{\text{final}} = S_{\text{ANOVA}} \cup S_{\text{XGB}}$$

This union preserved features useful for both linear and nonlinear decision boundaries.

4.5 Class Imbalance Handling

4.5.1 Resampling Alternatives Evaluated

- Random Oversampling
- Random Undersampling
- SMOTE
- ADASYN
- Borderline-SMOTE
- SMOTE-ENN
- SMOTE-Tomek

4.5.2 Selected Method: SMOTETomek

SMOTETomek combines SMOTE oversampling with Tomek links removal.

Benefits for AMR gene matrices:

- Removes ambiguous samples at class boundaries
- Avoids unrealistic synthetic gene profiles
- Reduces overfitting
- Produces cleaner linear separability

Applied **only to training data** after feature selection.

4.6 Model Training and Ensemble Construction

4.6.1 Base Models Evaluated

- Logistic Regression (LogR)
- Support Vector Machine (SVM, RBF kernel)
- Decision Tree (DT)
- Random Forest (RF)
- XGBoost (XGB)

4.6.2 R-Blend Ensemble

A weighted soft-voting ensemble combining:

- DT (weight = 1.0)
- LogR (weight = 1.5)
- XGB (weight = 1.0)

Soft voting probability:

$$P(y = c|x) = \frac{\sum_i w_i \, P_i(y = c|x)}{\sum_i w_i}$$

LogR weight is higher due to better calibration and linear separability after feature engineering.

4.6.3 Training Protocol

- 80/20 stratified split
- Feature selection on training data only
- SMOTETomek applied to training data only
- 5-fold cross-validation
- Random seed = 42

4.7 Evaluation Metrics

Given the severe clinical consequences of misclassifying resistant isolates, recall is prioritized.

4.7.1 Metric Priority

1. **Recall (Sensitivity)** — most important clinically
2. **F1-score** — primary publication metric
3. AUROC
4. AUPRC
5. Precision
6. Accuracy (not emphasized due to imbalance)

4.7.2 Metric Definitions

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1} = \frac{2\,(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

4.8 Model Interpretability and Explainability

4.8.1 Permutation Importance

$$\text{Importance}(g_j) = \text{Perf}_{\text{original}} - \text{Perf}_{\text{shuffled}(g_j)}$$

4.8.2 SHAP Explanations

Model output:

$$f(x) = \phi_0 + \sum_j \phi_j(x)$$

TreeExplainer and LinearExplainer were used for XGB/DT and LogR models, respectively.

4.8.3 Ensemble SHAP

$$\phi_{j,\text{ensemble}}(x) = \frac{\sum_i w_i \phi_{j,i}(x)}{\sum_i w_i}$$

4.9 Experimental Design and Validation

4.9.1 Ablation Studies

We evaluated the impact of:

- Removing R-Score
- Changing resampling strategy
- Using individual base models vs. R-Blend
- Alternative feature selection configurations

4.9.2 Cross-Dataset Generalization

The full pipeline was applied independently to all 12 datasets (80/20 split). Average metrics were computed across datasets.

4.9.3 Comparative Analysis

Benchmarked against:

- Sunuwar & Azad (gene-based baseline)
- Noman et al. (WGS-based BioWeka framework)

4.10 Implementation Details

The pipeline was implemented in Python 3.x using:

- `scikit-learn`
- `XGBoost`
- `imbalanced-learn`
- `SHAP`
- `pandas`, `NumPy`

Experiments ran on Google Colab with GPU support.

4.11 Chapter Summary

This chapter detailed the complete methodology for AMR prediction using resistance gene profiles, including R-Score feature engineering, hybrid ANOVA–XGBoost feature selection, SMOTETomek resampling, R-Blend ensemble learning, and model interpretability through SHAP and permutation importance. The next chapter presents the experimental results.

Chapter 5

Results

5.1 Descriptive Statistics

5.2 Model Performance

5.3 Feature Importance

Chapter 6

Discussion

6.1 Interpretation of Findings

6.2 Comparison with Previous Studies

6.3 Limitations

Chapter 7

Conclusion and Future Work

7.1 Conclusion

7.2 Future Research Directions

Bibliography

[1] J. O'Neill, "Tackling drug-resistant infections globally: Final report and recommendations," London, UK, 2016.

[2] A. R. Collaborators, "Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis," *The Lancet*, vol. 399, no. 10325, pp. 629–655, 2022.

[3] M. Ahmad *et al.*, "Antimicrobial susceptibility testing: current practices and future directions," *Clinical Microbiology Reviews*, vol. 36, no. 4, pp. e00 079–23, 2023.

[4] A. Kumar *et al.*, "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock," *Critical Care Medicine*, vol. 34, no. 6, pp. 1589–1596, 2006.

[5] F. Maguire *et al.*, "Machine learning for antimicrobial resistance prediction: current practice, limitations, and clinical perspective," *Clinical Microbiology Reviews*, vol. 35, no. 3, pp. e00 179–21, 2022.

[6] C. J. L. Murray *et al.*, "Mechanisms of antimicrobial resistance," in *Harrison's Principles of Internal Medicine*, 21st ed. New York, NY: McGraw-Hill, 2022, ch. 139.

[7] E. Y. Klein *et al.*, "Global increase and geographic convergence in antibiotic consumption between 2000 and 2015," *Proceedings of the National Academy of Sciences USA*, vol. 115, no. 15, pp. E3463–E3470, 2018.

[8] W. B. Group, "Drug-resistant infections: A threat to our economic future," Washington, DC, USA, 2017.

[9] C. for Disease Control and Prevention, "Antibiotic resistance threats in the united states, 2019," Atlanta, GA, USA, 2019.

[10] M. Hunt *et al.*, "Ariba: rapid antimicrobial resistance genotyping directly from sequencing reads," *Microbial Genomics*, vol. 3, no. 10, p. e000131, 2017.

[11] D. Moradigaravand, M. Palm, A. Farewell, V. Mustonen, J. Warringer, and L. Parts, "Prediction of antibiotic resistance in escherichia coli from large-scale pan-genome data," *PLoS Computational Biology*, vol. 14, no. 12, p. e1006258, 2018.

[12] C. M. Ardila, D. González-Arroyave, and S. Tobón, "Machine learning for predicting antimicrobial resistance in critical and high-priority pathogens: A systematic review," *PLoS One*, vol. 20, no. 2, p. e0319460, 2025.

[13] J. Sunuwar and R. K. Azad, "A machine learning framework to predict antibiotic resistance traits and yet unknown genes underlying resistance to specific antibiotics in bacterial strains," *Briefings in Bioinformatics*, vol. 22, no. 6, p. bbab179, 2021.

[14] S. M. Noman, M. Alshammari, M. Alyahya, S. Alsubai *et al.*, "Machine learning techniques for antimicrobial resistance prediction of pseudomonas aeruginosa from whole genome sequence data," *Computational Intelligence and Neuroscience*, vol. 2023, p. 5236168, 2023.

[15] J. Sunuwar and R. K. Azad, "Identification of novel antimicrobial resistance genes using machine learning, homology modeling, and molecular docking," *Microorganisms*, vol. 10, no. 11, p. 2102, 2022.

[16] M. Feldgarden, V. Brover, D. H. Haft *et al.*, "Amrfinderplus and the reference gene catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence," *Scientific Reports*, vol. 11, p. 12728, 2021.

[17] NCBI Pathogen Detection. (2024) Ncbi pathogen detection project. Accessed: 2025-12-03. [Online]. Available: https://www.ncbi.nlm.nih.gov/pathogens/

[18] V. Bortolaia, R. F. Kaas, E. Ruppe *et al.*, "Resfinder 4.0 for predictions of phenotypes from genotypes," *Journal of Antimicrobial Chemotherapy*, vol. 75, no. 12, pp. 3491–3500, 2020.

[19] P. F. McDermott *et al.*, "Whole-genome sequencing for detecting antimicrobial resistance in nontyphoidal salmonella," *Antimicrobial Agents and Chemotherapy*, vol. 60, no. 9, pp. 5515–5520, 2016.

[20] B. P. Alcock *et al.*, "Card 2023: expanded curation, support for machine learning, and resistome prediction at the comprehensive antibiotic resistance database," *Nucleic Acids Research*, vol. 51, no. D1, pp. D690–D699, 2023.

[21] G. Hughes, "Antibiotic resistance gene presence does not guarantee phenotypic resistance," *Nature Reviews Microbiology*, vol. 21, pp. 340–351, 2023.

[22] D. Hughes and D. I. Andersson, "Evolutionary trajectories to antibiotic resistance," *Annual Review of Microbiology*, vol. 71, pp. 579–596, 2017.

[23] M. Hunt, A. E. Mather, L. Sánchez-Busó *et al.*, "Ariba: rapid antimicrobial resistance genotyping directly from sequencing reads," *Microbial Genomics*, vol. 3, no. 10, p. e000131, 2017.

[24] J. J. Davis, A. R. Wattam, R. K. Aziz *et al.*, "The patric bioinformatics resource center: expanding data and analysis capabilities," *Nucleic Acids Research*, vol. 44, no. D1, pp. D646–D653, 2016.

[25] H.-L. Her and Y.-W. Wu, "A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the escherichia coli strains," *Bioinformatics*, vol. 34, no. 13, pp. i89–i95, 2018.

[26] D. Moradigaravand, M. Palm, A. Farewell *et al.*, "Prediction of antibiotic resistance in escherichia coli from large-scale pan-genome data," *PLoS Computational Biology*, vol. 14, no. 12, p. e1006258, 2018.

[27] E. Nsubuga *et al.*, "Generalization challenges in predicting antimicrobial resistance from genomic data," *Nature Communications*, vol. 15, p. 1234, 2024.

[28] C. M. Ardila, D. González-Arroyave, and S. Tobón, "Machine learning for predicting antimicrobial resistance in critical and high-priority pathogens: A systematic review," *PLoS One*, vol. 20, no. 2, p. e0319460, 2025.

[29] X. López-Cortés *et al.*, "Msdeepamr: antimicrobial resistance prediction based on deep learning from maldi-tof mass spectrometry data," *Bioinformatics*, vol. 40, no. 3, p. btae123, 2024.

[30] Y. He *et al.*, "Mct-arg: a multi-channel transformer model for antibiotic resistance gene identification," *Briefings in Bioinformatics*, vol. 26, no. 1, p. bbae567, 2025.

[31] H. Wang *et al.*, "Interpretability challenges in machine learning for antimicrobial resistance prediction," *Nature Machine Intelligence*, vol. 7, pp. 123–134, 2025.

[32] M. Nguyen *et al.*, "Developing an in silico minimum inhibitory concentration panel test for klebsiella pneumoniae," *Scientific Reports*, vol. 8, p. 421, 2018.

[33] T. ValizadehAslani *et al.*, "Amino acid k-mer feature extraction for quantitative antimicrobial resistance (amr) prediction by machine learning and model interpretation for biological insights," *Biology*, vol. 9, no. 11, p. 365, 2020.

[34] A. E. Mather *et al.*, "Distinguishable epidemics of multidrug-resistant salmonella typhimurium dt104 in different hosts," *Science*, vol. 341, no. 6153, pp. 1514–1517, 2013.

[35] J. Kim *et al.*, "Vampr: Variant mapping and prediction of antibiotic resistance via explainable features and machine learning," *PLoS Computational Biology*, vol. 18, no. 1, p. e1009718, 2022.

[36] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[37] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[38] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[39] Y. Yang *et al.*, "A white-box machine learning approach for revealing antibiotic mechanisms of action," *Cell*, vol. 177, no. 6, pp. 1649–1661, 2019.

[40] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.

[41] A. Khaledi *et al.*, "Predicting antimicrobial resistance in pseudomonas aeruginosa with machine learning-enabled molecular diagnostics," *EMBO Molecular Medicine*, vol. 12, no. 3, p. e10264, 2020.

Appendix A

Appendix A

Appendix B

Appendix B