Integrating Phenotypic and Genomic Data with Machine Learning to Predict Antimicrobial Resistance and Identify Genetic Biomarkers in *E. coli*

Sarah Halleluyah Adeyemi and Roshan Paudel *

Posted Date: 19 March 2026

doi: 10.20944/preprints202603.1564.v1

Keywords: *Escherichia coli*; antimicrobial resistance; phenotypic data; machine learning; phylogenetic analysis; genomic biomarkers

 doi:10.20944/preprints202603.1564.v1

Article

Integrating Phenotypic and Genomic Data with Machine Learning to Predict Antimicrobial Resistance and Identify Genetic Biomarkers in *E. coli*

Sarah Halleluyah Adeyemi and Roshan Paudel *

Department of Computer Science, School of Computer Science, Morgan State University, Baltimore, MD, USA

* Correspondence: roshan.paudel@morgan.edu; Tel.: 443-885-3096

Abstract

Antimicrobial resistance in *Escherichia coli* is a significant public health concern globally, driven by increased resistance to commonly used antimicrobial agents such as β-lactams and fluoroquinolones. This study aimed to develop a machine-learning framework to predict antimicrobial resistance in *Escherichia coli* by integrating antimicrobial susceptibility testing data with genomic biomarker analysis. A dataset comprising 17,122 Escherichia coli clinical isolates was obtained from the Bacterial and Viral Bioinformatics Resource Center (BV-BRC). After preprocessing, fivefold cross-validation was used to train and test five machine learning models: Random Forest, XGBoost, Support Vector Machine, Logistic Regression, and k-Nearest Neighbors. The highest-performing model was XGBoost, with 0.86 accuracy and 0.932 ROC-AUC, followed by Random Forest, with 0.82 accuracy and 0.89 ROC-AUC. Phylogenetic analysis revealed that resistant isolates clustered together relative to the reference genome of *Escherichia coli* K-12 MG1655. Genomic biomarkers such as gyrA, parC, CTX-M-15, OXA-1, and various multidrug efflux pumps were identified by the Comprehensive Antibiotic Resistance Database (CARD) and ResFinder as significant resistance determinants in this study. In conclusion, this study demonstrates that combining antimicrobial susceptibility testing with machine learning and genomic biomarkers is a powerful framework for predicting antimicrobial resistance in *Escherichia coli*.

Keywords: *Escherichia coli;* antimicrobial resistance; phenotypic data; machine learning; phylogenetic analysis; genomic biomarkers

1. Introduction

Antimicrobial resistance (AMR) is one of the major global health challenges of the 21st century, affecting the efficacy of many commonly used antibiotics and increasing morbidity, mortality, and healthcare costs worldwide. According to a global study published in The Lancet, bacterial AMR was estimated to have directly caused 1.27 million deaths, while 4.95 million deaths were associated with AMR in 2019 globally [1]. This makes AMR a leading cause of mortality, on par with major diseases such as HIV and malaria. Beyond reducing survival rates, AMR also leads to significant healthcare costs due to extended hospital stays, additional testing, and the use of more expensive or toxic drugs [2]. *E. coli* is a major cause of community-acquired and healthcare-associated infections. It has become increasingly resistant to several classes of antibiotics, posing a serious problem for clinicians and healthcare systems [3,4]. Early and accurate detection of resistance in *E. coli* is therefore crucial to enhance patient safety, inform empirical treatment, and prevent the spread of resistant strains [4]. Conventional antimicrobial susceptibility testing, or AST, is considered the gold standard; however, it requires considerable resources and time. Whole-genome sequencing (WGS) technology now enables the detailed sequencing of bacterial genomes, allowing for the development of models that predict resistance directly from the bacterial genome [8,9].

 doi:10.20944/preprints202603.1564.v1

However, the complexity and volume of genomic data require advanced computational methods to extract meaningful clinical insights. While stewardship programs are currently in progress and an understanding of antibiotic risks is increasing [5–7], a disconnect remains between pathogen resistance and prescribed drugs. An example of a problem increasingly under public scrutiny is ciprofloxacin, a fluoroquinolone antibiotic that has been widely prescribed since the early 2000s and is listed among the WHO's Essential Medicines [8]. It is effective against various strains of both gram-negative and gram-positive bacteria, including those responsible for urinary tract, respiratory, bone/joint, and gastrointestinal infections [9,10]. However, its widespread use has led to the emergence of resistance globally [11–13]. Although restrictions on antibiotic use can help reduce selective pressure [15], resistance often persists due to the fitness benefits of resistant clones and horizontal gene transfer [12,13].

In recent years, machine learning (ML) has been recognized as one of the most impactful tools for predicting antimicrobial resistance (AMR). Compared to traditional statistical methods, ML algorithms such as Random Forest, Support Vector Machine (SVM), Logistic Regression, and XGBoost can capture complex and non-linear relationships in large datasets. These models provide fast, reliable predictions that can help clinicians select the most effective treatments and support antimicrobial stewardship programs [16]. While previous studies mainly focus on whole-genome sequencing (WGS) data, this research advances by using ML to analyze large-scale phenotypic antimicrobial susceptibility testing (AST) data, with phylogenetic analysis and biomarker discovery as secondary methods. This combined approach enhances prediction accuracy and bioinformatic interpretability, making it particularly valuable in hospital settings where genomic sequencing resources may be limited. For example, Mintz used hospital-based clinical data to predict ciprofloxacin resistance using logistic and gradient-boosting models but did not incorporate phenotypic or phylogenetic data [16]. Similarly, public AMR prediction challenges, such as the Kaggle "AMR Benchmark Dataset," focus on whole-genome predictors without linking results to phenotypic AST outcomes. In contrast, this study bridges that gap by integrating large-scale phenotypic AST data with machine learning and genomic biomarker screening to improve resistance classification and interpretability. By combining both data types, this research offers a more clinically applicable framework for early AMR detection and genomic insights.

Based on recent developments and identified shortcomings, this study presents an integrated framework for machine learning and bioinformatics to predict antimicrobial resistance in *Escherichia coli*. For this purpose, a dataset of clinical isolates was collected from the Bacterial and Viral Bioinformatics Resource Center (BV-BRC), and several machine learning algorithms were evaluated for predicting antimicrobial resistance phenotypes based on phenotypic susceptibility data. In addition, genomic biomarker screening and phylogenetic analysis were conducted to predict resistance-associated genetic determinants. The major objective of this study is to propose an effective framework for predicting antimicrobial resistance and to improve data-driven approaches for its prediction. Unlike previous research, which relies primarily on genomic sequencing data, this research combines large-scale phenotypic antimicrobial susceptibility testing with machine learning and genomic biomarker analysis, offering a clinically applicable, resource-efficient framework for predicting antimicrobial resistance.

2. Materials and Methods

A. Datasets

The dataset used in this study was obtained from the Bacterial and Viral Bioinformatics Resource Center (BV-BRC), a comprehensive bioinformatics repository maintained by the National Institute of Allergy and Infectious Diseases (NIAID). It included 17,122 *Escherichia coli* clinical isolates collected from various clinical and geographical sources, tested against multiple antibiotics across major drug classes, including β-lactams, fluoroquinolones, carbapenems, aminoglycosides, macrolides, and sulfonamides. Each isolate was annotated with antibiotic susceptibility results based on Clinical and Laboratory Standards Institute (CLSI) or European Committee on Antimicrobial Susceptibility

 doi:10.20944/preprints202603.1564.v1

Testing (EUCAST) guidelines. Data fields included isolate ID, antibiotic name, testing method (MIC or disk diffusion), testing standard, and phenotypic classification (resistant, intermediate, or susceptible). After preprocessing, 16,740 isolates remained for analysis. The dataset showed moderate class imbalance, with resistant isolates making up approximately 68% and susceptible isolates 32% of the total entries. Antibiotics such as ampicillin, ciprofloxacin, and ceftriaxone exhibited the highest resistance rates, while ertapenem and meropenem had lower resistance frequencies.

B. Data Cleaning and Preprocessing

To ensure analytical consistency, the raw phenotypic AST data were cleaned. Excluded were isolates with unclear or missing antibiotic results. Using the antibiotic name and the genome ID, duplicate entries were eliminated. To improve class clarity, intermediate isolates were removed from the machine-learning pipeline, and phenotypic labels were standardized to 'Resistant' (R = 1) and 'Susceptible' (S = 0). LabelEncoder was used for the variables, and min-max normalization was applied to scale the final feature matrix to the range of 0-1, ensuring consistent feature weighting. To preserve class distribution, stratified sampling was used to divide the cleaned dataset into training (80%) and testing (20%) sets.

C. Machine Learning Algorithms

Five supervised machine learning algorithms—Random Forest, XGBoost, Support Vector Machine, Logistic Regression, and k-Nearest Neighbors—were implemented using scikit-learn and XGBoost in Python. Models were trained on 80% of the dataset and tested on the remaining 20% with stratified sampling. Performance was evaluated through accuracy, precision, recall, F1-score, and ROC-AUC. Five-fold cross-validation was employed to confirm robustness. Feature importance analysis was conducted for tree-based models to identify the most predictive antibiotics.

D. Random Forest

Random Forest (RF) is a classification algorithm that uses decision trees as its base learning model. Developed by Breiman [17], RF is a meta-estimator that improves prediction accuracy and reduces overfitting by averaging the predictions from multiple DT classifications trained on different subsamples of the dataset.

E. XGBoost

Introduced in 2011 by Tianqi Chen and Carlos Guestrin, XGBoost is an optimized learning system based on Boosting Tree models, which usually only consider the rate of change (first derivative) of the loss function [18]. XGBoost uses a second-order Taylor expansion, incorporating both the rate of change and the rate of change of the rate of change (second derivative). This enables XGBoost to identify more complex patterns in the data. The architecture of XGBoost, which combines multiple weak learners, naturally provides resistance to overfitting.

F. Support Vector Machine

SVM is a supervised learning algorithm that constructs a hyperplane or multiple hyperplanes in a high-dimensional feature space, enabling effective classification of data. SVM was introduced by [19] and was initially designed as a binary classification tool, but can also be used for multilabel classification. The Lagrangian Multiplier Method is employed in the dual formulation of the SVM optimization problem by combining the goal of minimizing the norm of the weight vector w with constraints that ensure the correct classification of data points with a margin [20–22].

G. Logistic Regression

Logistic Regression (LR) is efficient and relatively simple to implement, although it may not capture complex relationships as well as tree-based models [23]. LR models are widely recognized in the medical and health fields, particularly for their applications in genetics [24], death risk prediction [25], and the evaluation of risk factors, such as those associated with prostate cancer [26]. The flexibility and effectiveness of LR make them valuable tools for understanding and predicting various health-related diseases.

H. k-Nearest Neighbors

k-Nearest Neighbor (k-NN) is an ML algorithm used for classification and regression, where each instance can be assigned multiple labels simultaneously in multi-label classification. This is

 doi:10.20944/preprints202603.1564.v1

achieved by selecting the number of neighbors, k, to use for prediction. Then, the closest neighbor is identified based on the distance measured from the instance (isolate) to all other instances (isolates) in the training dataset. The labels are aggregated and then assigned to the isolates based on this aggregation.

I. Model Evaluation Metrics

Model performance was evaluated using multiple standard classification metrics:

- Accuracy (Acc) = (TP + TN) / (TP + TN + FP + FN)
- Precision (P) = TP / (TP + FP)
- Recall (R) = TP / (TP + FN)
- F1-score (F1) = 2 × (P × R) / (P + R)
- Area Under the Receiver Operating Characteristic Curve (ROC–AUC)

Confusion matrices and ROC–AUC curves were plotted for each model using the RocCurveDisplay function in scikit-learn. These visualizations were used to compare the sensitivity and specificity of models across multiple antibiotic classes.

J. Visualization and Feature Interpretation

To assess model interpretability, feature-importance plots were created for Random Forest and XGBoost models. The key features related to antibiotic types and resistance classes were identified to understand which variables most impacted the prediction results. Visualization tools, such as Matplotlib and Seaborn, were used to generate confusion matrices, ROC curves, and correlation heatmaps.

K. Phylogenetic Analysis

Whole genome FASTA sequences of *Escherichia coli* clinical isolates were obtained from the BV-BRC database. For evolutionary analysis, the genome of *E. coli* K-12 MG1655 (GenBank accession U00096.3) was also retrieved from NCBI. The sequences were aligned using MAFFT (Multiple Alignment using Fast Fourier Transform), which operates in auto mode to optimize parameters based on the dataset size and complexity. The aligned sequences were used with FastTree to generate a maximum-likelihood phylogenetic tree under the General Time Reversible (GTR) substitution model. Using Archaeopteryx JS, the tree was visualized with color-coded annotations, allowing for a comparison of the ingroup (*E. coli* strains), reference sequences (K-12), and outgroups (closely related Enterobacteriaceae species).

L. Identification of Genomic Biomarkers

Genomic biomarkers were identified through a comparative and literature-guided bioinformatics approach. Due to limitations in the availability of paired resistant and susceptible genome sequences, this study employed an in silico comparison against the *E. coli* K-12 reference and cross-referenced known resistance loci reported in the literature and AMR databases. Genes with established links to resistance, including gyrA, gyrB, parC, parE, acrA, acrB, mdtK, blaTEM, and blaCTX-M, were examined. Multiple sequence alignment using MAFFT identified key substitution positions, such as S83L, D87N, and S80I, which are associated with fluoroquinolone resistance. These biomarkers were validated using the Comprehensive Antibiotic Resistance Database (CARD) and ResFinder.

3. Results

3.1. The Distribution of Antibiotics Testing and Resistance Patterns

The dataset included a wide variety of antibiotics, but a subset stood out as the most frequently tested. These included ampicillin, ceftriaxone, ciprofloxacin, levofloxacin, gentamicin, amoxicillin/clavulanic acid, cefepime, piperacillin/tazobactam, meropenem, and ertapenem (Figure 2). These antibiotics represent both commonly prescribed medications and critical last-resort treatments. For example, ampicillin and ciprofloxacin are frequently used in clinical practice, whereas carbapenems, such as meropenem and ertapenem, are reserved for severe or multidrug-resistant infections.

 doi:10.20944/preprints202603.1564.v1

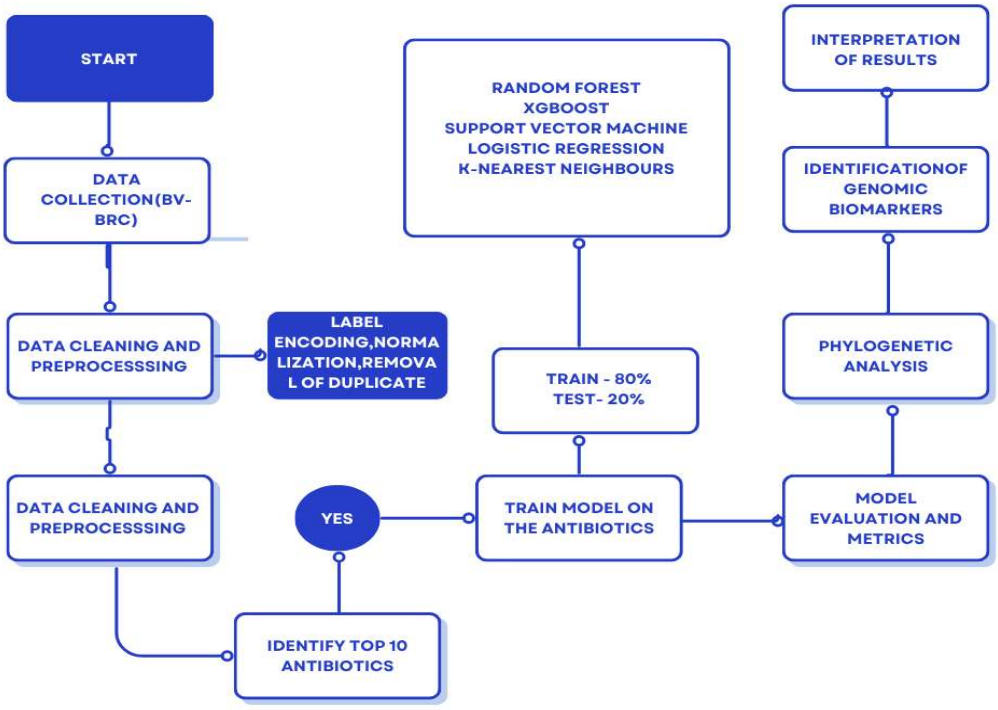


Figure 1. Analytical Workflow for Machine Learning-Based Classification and Biomarker Identification in E. coli. This figure outlines the methodological framework used in this study. The workflow starts with collecting and preprocessing phenotypic and genomic data from the BV-BRC repository, then proceeds to data normalization and feature encoding. The top ten antibiotics are used to train multiple machine learning models (Random Forest, XGBoost, SVM, Logistic Regression, and k-Nearest Neighbors) to predict resistance profiles. The process ends with model evaluation, phylogenetic analysis, genomic biomarker identification, and biological interpretation of the results.

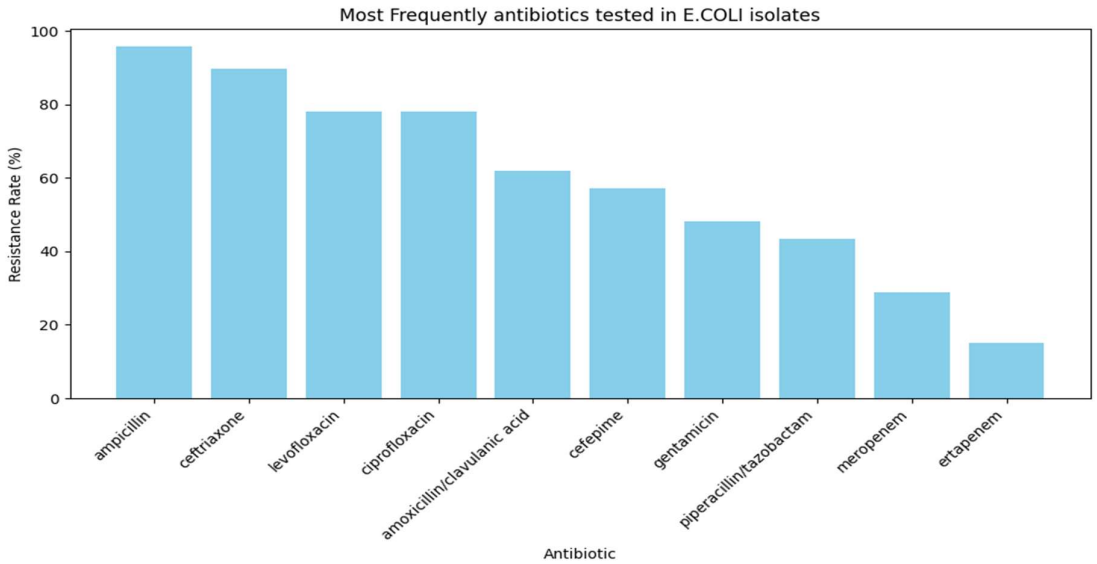


Figure 2. Top 10 most frequently tested antibiotics in E. coli isolates. It displays the proportion of antimicrobial susceptibility tests conducted for each antibiotic in the BV-BRC dataset. Ampicillin, ceftriaxone, levofloxacin, and ciprofloxacin were the most frequently evaluated, highlighting their widespread clinical use. In contrast,

 doi:10.20944/preprints202603.1564.v1

carbapenems such as meropenem and ertapenem were less often tested, reflecting their role as last-resort treatments for multidrug-resistant infections.

3.2. Machine Learning Model Performance

The machine learning models, phylogenetic analysis, and theoretical genomic biomarker identification are presented below. Five ML algorithms (Random Forest, eXtreme Gradient Boosting, Support Vector Machine, K-Nearest Neighbors, and logistic regression) were used to classify antimicrobial resistance in E. coli based on phenotypic antimicrobial susceptibility testing (AST) data. The models were evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to assess their predictive performance.

The performance of all five machine learning models is summarized in Table 1. XGBoost and Random Forest achieved the highest predictive accuracy among the models tested.

Table 1. Performance of Machine Learning Model on AMR classification.

MODEL	ACCURACY	ROC - AUC	Precision	Recall	F1
Random Forest	0.82	0.890	0.56	0.55	0.56
XGBoost	0.86	0.932	0.57	0.55	0.56
SVM	0.67	0.540	0.56	0.34	0.27
Logistic Regression	0.76	0.637	0.40	0.37	0.35
KNN	0.78	0.828	0.52	0.51	0.52

Tree-based ensemble methods outperformed all other approaches. XGBoost achieved the highest performance (accuracy = 0.86, ROC-AUC = 0.932), followed closely by Random Forest (accuracy = 0.82, ROC-AUC = 0.89). Both models achieved strong classification performance for susceptible and resistant isolates but struggled to classify the minority class ("Intermediate"), which comprised fewer than 25 samples. In contrast, SVM and logistic regression performed poorly, with ROC-AUC values of 0.54 and 0.64, respectively. These models exhibited a strong majority-class bias (susceptible) and showed little to no ability to identify resistant or intermediate isolates accurately. The kNN model also demonstrated good predictive performance (accuracy 0.78, ROC-AUC 0.828), but its precision and recall suggest potential overfitting to the majority class. SVM and LR exhibited comparatively lower accuracies of 0.67 and 0.76, suggesting that linear models are less effective at handling the complex, nonlinear feature relationships present in phenotypic AMR data. Ensemble-based approaches (XGBoost and Random Forest) provided the best trade-off between bias and variance, making them robust models for antimicrobial resistance classification. See Figure 3 below for the visualization of the performance accuracy.

 doi:10.20944/preprints202603.1564.v1

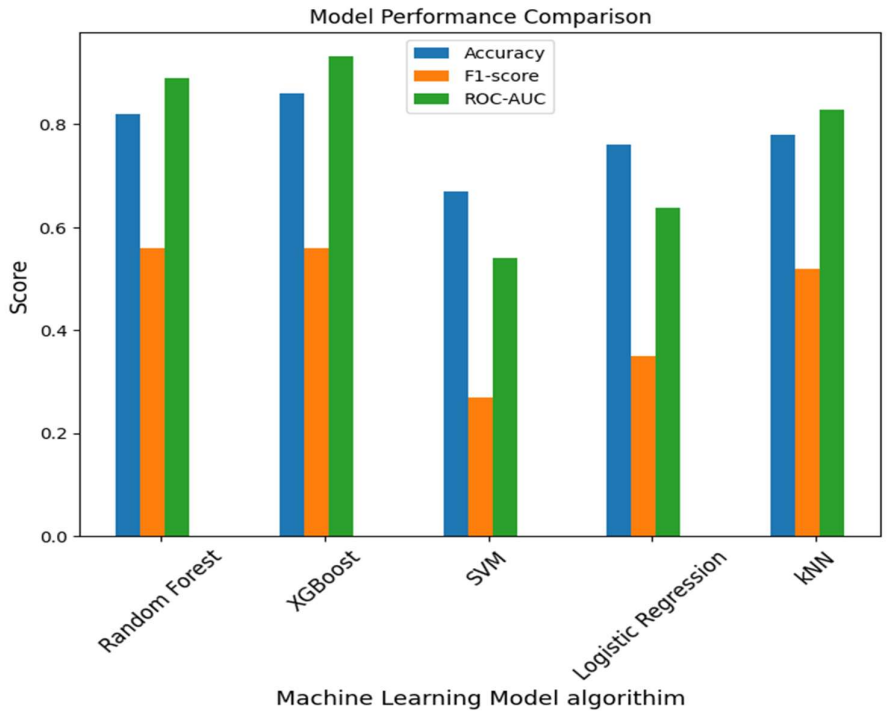
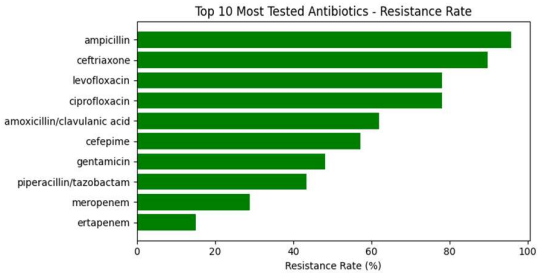


Figure 3. A comparative performance of five machine learning models. The model performance metrics, including accuracy, F1-score, and ROC-AUC, are shown for the five machine learning algorithms used. XGBoost achieved the highest predictive accuracy (0.86) and area under the curve (0.932), followed by Random Forest, demonstrating the strength of ensemble methods in AMR prediction.

3.3. The Resistance Distribution Across Antibiotics

The antibiotic resistance distribution derived from the cleaned BV-BRC dataset among the tested antibiotics includes ampicillin, ciprofloxacin, ceftriaxone, and cefotaxime. These antibiotics demonstrated the highest resistance rates, while imipenem and meropenem maintained significant efficacy. This distribution aligns with WHO's global AMR surveillance reports, which identify fluoroquinolones and β-lactams as the most compromised antibiotic classes due to decades of overuse in clinical and agricultural settings. The proportion of resistant isolates was calculated for each antibiotic. The results showed that ampicillin had the highest resistance rate (>95%), followed by third-generation cephalosporins such as ceftriaxone and cefotaxime, which also exhibited extremely high resistance (Figure 4). Fluoroquinolones (ciprofloxacin and levofloxacin) demonstrated resistance rates of 70-80%, indicating reduced clinical utility, although some isolates remained susceptible.



Antibiotic	Resistant Phenotype	Intermediate	Resistant	Susceptible	Total	Resistance rate(%)
ampicillin		0.0	620.0	28.0	648.0	95.679012
ceftriaxone		1.0	582.0	66.0	649.0	89.676425
levofloxacin		2.0	507.0	140.0	649.0	78.120185
ciprofloxacin		3.0	507.0	140.0	650.0	78.000000
amoxicillin/clavulanic acid		30.0	400.0	215.0	645.0	62.015504
cefepime		13.0	368.0	263.0	644.0	57.142857
gentamicin		1.0	313.0	336.0	650.0	48.153846
piperacillin/tazobactam		19.0	280.0	347.0	646.0	43.343653
meropenem		0.0	187.0	462.0	649.0	28.813559
ertapenem		0.0	97.0	549.0	646.0	15.015480

Figure 4. Resistance rate of the top 10 most frequently tested antibiotics. Fluoroquinolones (ciprofloxacin, levofloxacin) and β-lactams (ampicillin, ceftriaxone) showed the highest resistance levels, whereas carbapenems retained substantial activity.

 doi:10.20944/preprints202603.1564.v1

A heatmap was created showing the counts of resistant, susceptible, and intermediate isolates for the top 20 antibiotics (Figure 5). Ampicillin, ceftriaxone, and cefotaxime had the highest resistance counts, with almost no susceptible isolates. Ciprofloxacin and levofloxacin were mostly resistant, although susceptible isolates were still present. Aminoglycosides, such as gentamicin and amikacin, displayed a more balanced distribution. Carbapenems, especially meropenem, were generally susceptible; however, the presence of resistant isolates indicates a potential decline in their clinical effectiveness.

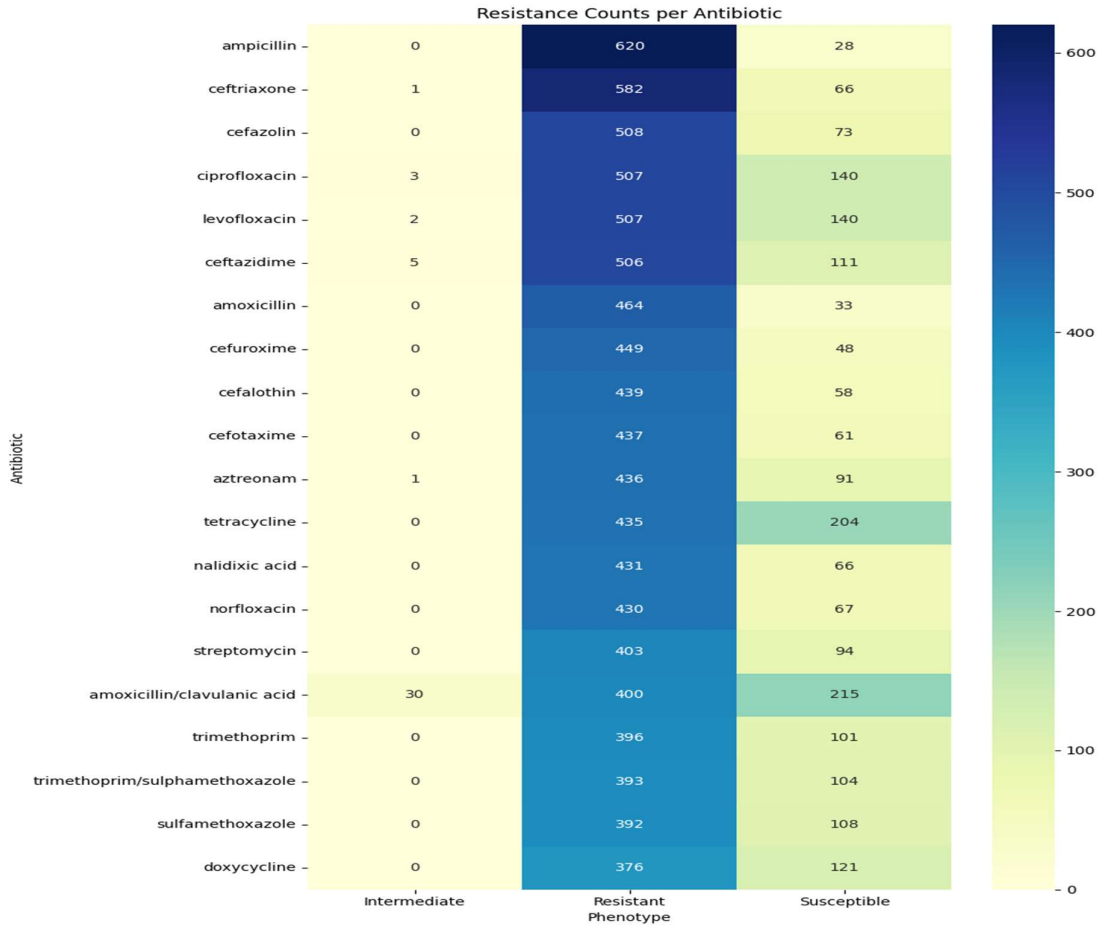


Figure 5. A Heatmap of Resistance per Count. It shows the resistance, susceptible, and intermediate isolate counts for the top 20 antibiotics. The heatmap helps visualize the proportions of resistance phenotypes within individual antibiotic classes. The darker blue colors indicate an increase in the number of resistant isolates, especially to ampicillin, ceftriaxone, and ciprofloxacin. Carbapenems and aminoglycosides were less dark, indicating a relatively low prevalence of resistance.

A stacked bar chart was used to display the distributions of phenotypes for the top 10 most frequently tested antibiotics (Figure 6). Ampicillin was almost entirely resistant, confirming its ineffectiveness. Cephalosporins such as ceftriaxone and cefepime also displayed high resistance. By contrast, gentamicin and cefepime retained some effectiveness, with significant numbers of susceptible isolates. Carbapenems (meropenem, ertapenem) were largely susceptible, demonstrating their importance as reserve drugs, though resistant isolates were present in smaller numbers.

 doi:10.20944/preprints202603.1564.v1

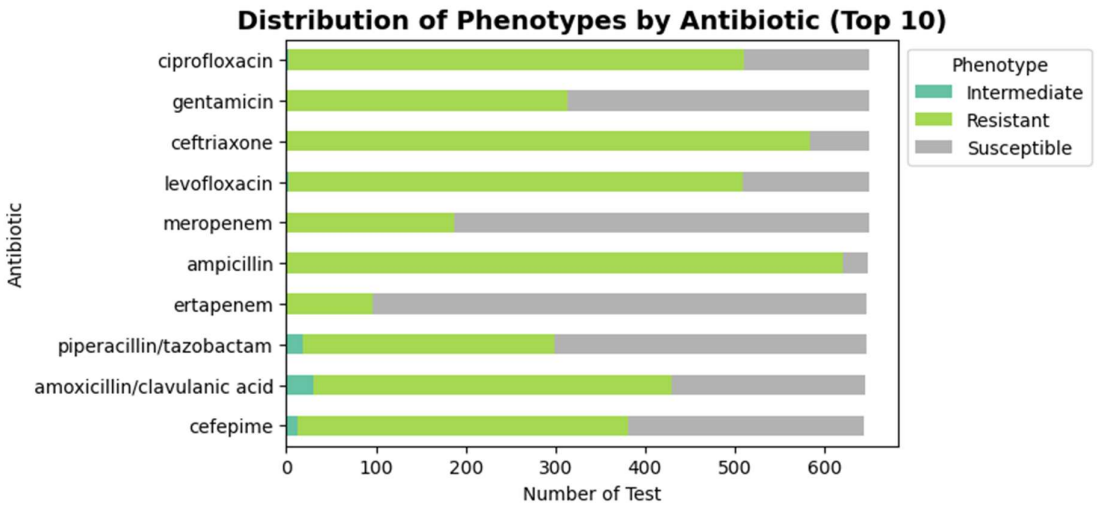


Figure 6. The distribution of resistant, susceptible, and intermediate phenotypes for the top 10 antibiotics. This figure demonstrates phenotypic heterogeneity observed between the antibiotics used in model training. The most predominant resistant phenotypes were ciprofloxacin, ampicillin, levofloxacin, and ceftriaxone, whilst higher susceptibility was observed with carbapenems, including meropenem and ertapenem, highlighting the disparity in selection pressure among antimicrobial groups.

3.4. Phylogenetic Analysis

The whole-genome FASTA sequences of *E. coli* isolates from the BV-BRC were aligned with those of the reference strain *E. coli* K-12 MG1655 using MAFFT, and a phylogenetic tree was reconstructed with FastTree. Visualization with Archaeopteryx JS revealed distinct clustering patterns. The phylogenetic tree showed that ciprofloxacin-resistant isolates tended to form separate clades from susceptible isolates. While some susceptible strains clustered closely with the reference K-12, there was greater evolutionary divergence, consistent with the accumulation of resistance-associated mutations. Interestingly, resistant isolates also clustered by geographic origin, with certain groups forming tight branches, suggesting potential regional spread of resistant lineages. Outgroup taxa (e.g., Cronobacter and Atlantibacter species) provided evolutionary context, confirming that the resistant E. coli isolates were phylogenetically distinct but remained within the Enterobacteriaceae family. See Figure 7a and 7b below.

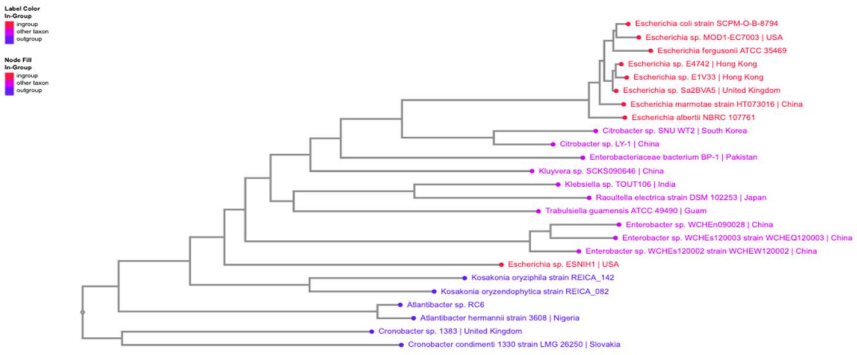


Figure 7a

 doi:10.20944/preprints202603.1564.v1

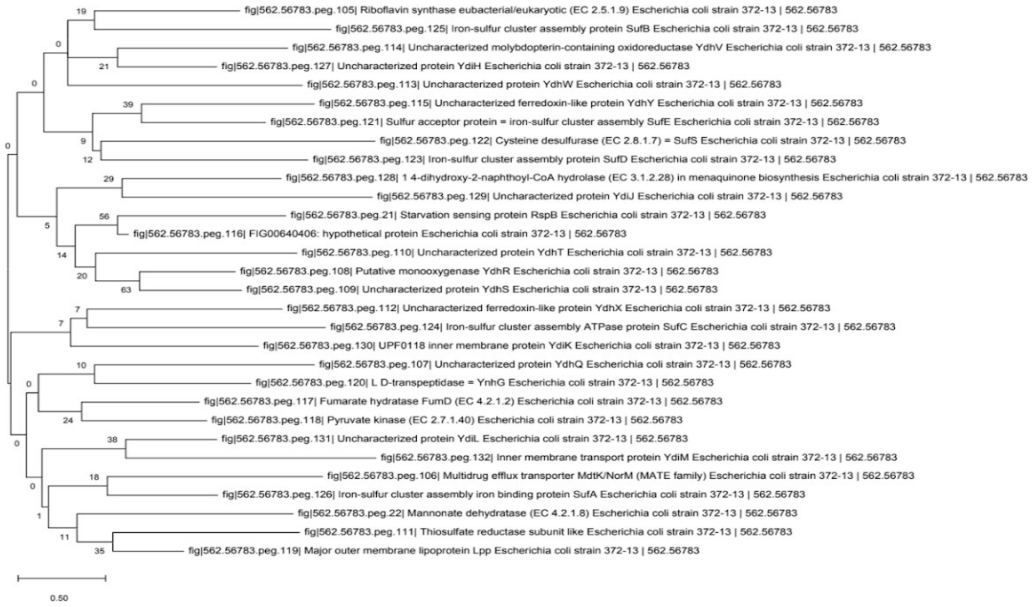


Figure 7b

Figure 7. a and 7b: Show a phylogenetic tree comparing E. coli clinical isolates with the K-12 reference strain. The phylogenetic relationships among E. coli clinical isolates analyzed in this study, aligned against the E. coli K-12 MG1655 reference genome. The tree demonstrates that resistant and susceptible isolates form distinguishable clusters, suggesting minor genomic variations linked to antimicrobial resistance traits rather than broad evolutionary divergence. The inclusion of outgroup taxa (Cronobacter and Atlantibacter species) provided phylogenetic context, confirming that all isolates remained within the Enterobacteriaceae family. These results complement the machine-learning classification outcomes by supporting the genomic differentiation patterns underlying phenotypic resistance in E. coli.

3.5. Genomic Biomarkers Identifications

The genomic biomarker analysis aimed to identify resistance-linked genes and mutations that contribute to the antimicrobial resistance (AMR) phenotypes of *E. coli*. In this regard, two of the most popular AMR databases, namely the Comprehensive Antibiotic Resistance Database (CARD) through the Resistance Gene Identifier (RGI) tool and ResFinder, operated by the Danish Technical University (DTU), were used to analyze the collected genome sequences (BVBRC_genome_feature.fasta). These orthologous tools offered not only curated reference alignments (CARD) but also in silico sequence matching (ResFinder) to identify known AMR determinants, drug classes to which they belong, and resistance mechanisms. The CARD RGI outcome was that there are several *E. coli* genes related to fluoroquinolone, β-lactam, tetracycline, and multidrug resistance. The levels of confidence in the reliability of the test, as well as the strictness and looseness of the test, were reported by CARD to be high in terms of identifying the sequences of genome features. CARD reported that the levels of confidence in the reliability of the test, perfection, strictness, and looseness were high in terms of identifying the sequences of genome features with the known resistance determinants listed in the Antibiotic Resistance Ontology (ARO). Similarly, the ResFinder output identified aminoglycoside-modifying enzymes and β-lactamase variants, consistent with multidrug resistance phenotypes. See Table 2 below which consist of Genomic Biomarkers of Antimicrobial Resistance Identified from CARD and ResFinder Analysis.

 doi:10.20944/preprints202603.1564.v1

Table 2. Genomic Biomarkers of Antimicrobial Resistance Identified from CARD and ResFinder Analysis.

Gene/Mutation	Antibiotics class	Phenotypic Resistance/Associated Drugs	Resistance Mechanism	Confidence/Source
gyrA (S83L, D87N)	Fluoroquinolones	Ciprofloxacin, Levofloxacin	Target alteration (DNA gyrase mutation)	Perfect, CARD
parC (S80I)	Fluoroquinolones	Ciprofloxacin	Target alteration (Topoisomerase IV mutation)	Perfect, CARD
aac(3)-IIa	Aminoglycosides	Gentamicin, Tobramycin	N(3)-acetyltransferase enzyme	28636609, ResFinder
aac(6')-Ib-cr	Aminoglycosides / Fluoroquinolones	Tobramycin, Amikacin, Ciprofloxacin	N(6')-acetyltransferase (fluoroquinolone acetylation)	DQ303918, ResFinder
blaCTX-M-15	β-lactams / Cephalosporins	Amoxicillin, Cefotaxime, Cefepime, Ceftriaxone	Extended-spectrum β-lactamase (Class A)	11470367, ResFinder
blaOXA-1	β-lactams	Ampicillin, Amoxicillin-clavulanate, Piperacillin	Class D OXA-type β-lactamase	10898672, ResFinder
blaTEM, ampC	β-lactams / Cephalosporins	Ampicillin, Cefazolin	Antibiotic inactivation (β-lactamase)	Strict , CARD
acrA, acrB, acrE, acrF	Multidrug / Quinolones	Multiple drug substrates	Efflux pump complex (RND family)	Strict ,CARD
mdtK, mdtH, mdtM, mdtG, mdtN	Multidrug / Macrolides	Erythromycin, Azithromycin	Efflux transporters and regulators	Strict, CARD
emrB, emrR, emrY	Macrolides / Phenicols	Chloramphenicol, Erythromycin	Multidrug efflux and regulatory proteins	Strict, CARD
catB3	Phenicols	Chloramphenicol	O-acetyltransferase enzyme (drug inactivation)	1662753 / 7793874, ResFinder
tet(A), tetR	Tetracyclines	Tetracycline, Doxycycline	MFS efflux pump system	12654659, ResFinder
msbA, tolC	Disinfectant / Multidrug	Broad substrate range	Membrane transport and antibiotic efflux	Loose, CARD
vanG, vanD	Glycopeptides	Vancomycin	Target alteration (cell wall modification)	Loose , CARD

 doi:10.20944/preprints202603.1564.v1

Detected genes correspond to major antibiotic classes, including fluoroquinolones (gyrA, parC), β-lactams (blaTEM, blaCTX-M-15, ampC, blaOXA-1), aminoglycosides (aac(3)-IIa, aac(6′)-Ib-cr), and tetracyclines (tet(A)).

4. Discussion

In this study, a combined bioinformatics and machine learning approach was used to classify antimicrobial resistance (AMR) phenotypes and identify genomic biomarkers in *E. coli* clinical isolates. Using phenotypic antimicrobial susceptibility testing (AST) data from the Bacterial and Viral Bioinformatics Resource Center (BV-BRC), several machine learning models were developed to predict resistance patterns. These included Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), Logistic Regression (LR), and k-Nearest Neighbors (kNN). The XGBoost model demonstrated superior predictive accuracy (0.86) and ROC-AUC (0.932), showing high generalizability and the ability to classify resistance phenotypes. Although both XGBoost and Random Forest performed well, XGBoost was considered better due to its greater capacity to model nonlinear relationships and feature interactions. The Random Forest model also performed well, with an accuracy of 0.82, and offered interpretability through feature importance, providing biological insights into the key predictors of resistance. Models such as SVMs and logistic regression showed lower accuracy due to their limitations as linear classifiers in high-dimensional biological data and multi-class scenarios. These results suggest that tree-based algorithms, especially ensemble learning methods, are highly suitable for predicting resistance based on phenotypic data.

A combination of machine learning findings with genomic screening provided a greater understanding of resistance mechanisms. Resistance genes identified in genomics by the Comprehensive Antibiotic Resistance Database (CARD) and ResFinder have been highlighted as important. GyrA gene mutations in S83L and D87N, along with parC gene mutations in S80I, were linked to fluoroquinolone resistance, supporting the original research that suggested these genetic mutations were associated with changes in DNA gyrase and topoisomerase IV [27]. Similarly, the observed β-lactamase genes were blaCTX-M-15, blaOXA-1, and blaTEM, which confer resistance to cephalosporins, including third-generation cephalosporins. These enzymes belong to the extended-spectrum β-lactamases (ESBLs), which have been widely detected in *E. coli* isolates from both hospital and community settings [16,28].

Multidrug efflux systems (acrA, mdtK, and emrB) were also determined, and aac(3)-IIa and aac(6′)-Ib-cr were identified as aminoglycoside-modifying enzymes. The co-occurrence of genes for efflux pumps and beta-lactamases suggests the evolution of multidrug resistance through synergistic processes. The efflux pumps also increase cross-resistance to structurally unrelated antibiotics, in addition to lowering intracellular drug concentration, as it is becoming increasingly prevalent in E. coli and other Enterobacteriaceae [29]. The presence of efflux-related mechanisms observed in this experiment indicates the genomic plasticity of E. coli and its ability to evolve under the pressure of an antibiotic test.

The isolates demonstrated a greater evolutionary background in phylogenetic analysis. Resistance, as clusters, was more likely to be observed among resistant isolates, indicating the expansion of some high-risk lineages that carry multiple AMR determinants. This clustering has also been seen in worldwide comparative genomic analysis, where AMR genes often cluster with mobile genetic elements and horizontal gene transfer [30]. This level of genomic plasticity highlights why phylogenetics should be used alongside machine learning predictions to gain a more comprehensive view of the evolutionary processes behind resistance.

Together, combining phenotypic, genomic, and phylogenetic data creates a powerful framework for AMR surveillance and prediction. This approach can be used in both clinical and population health labs, with adjustments that enable quick, evidence-based decisions about antimicrobial treatment. Additionally, detecting key biomarkers such as gyrA, blaCTX-M-15, and aac(6)-Ib-cr offers potential molecular targets for developing diagnostic tests and conducting epidemiological surveillance.

 doi:10.20944/preprints202603.1564.v1

Although this study achieved its goal of categorizing AMR phenotypes and detecting genomic biomarkers, a larger, more diverse dataset covering various regions of the world and host populations should be used in the future. Predictive accuracy could also be improved by integrating transcriptomic or proteomic layers, thereby allowing the regulatory networks underlying resistance expression to be observed. Further testing of the proposed biomarkers, including experimental validation of the findings through quantitative PCR and functional assays, is also essential to verify their biological significance and diagnostic capabilities.

Overall, this study demonstrates that machine learning models, particularly XGBoost and Random Forest, can effectively predict antibiotic resistance in *E. coli* using phenotypic data. It also shows that screening genomic biomarkers can provide additional mechanistic insights. This integrated analytical approach enhances understanding of resistance development and will serve as a foundation for future AMR diagnostics, surveillance, and intervention strategies that can be replicated.

5. Conclusion

This research work demonstrated that combining phenotypic information and genomic data with machine-learning solutions could increase the predictive capacity for antimicrobial resistance (AMR) in *E. coli.* The XGBoost and Random Forest models demonstrated the best predictive performance among the models considered, suggesting the usefulness of ensemble approaches for handling complex biological data. Genomic analysis also identified the major determinants of resistance, specifically gyrA, which is particularly significant in combination with parC and qnrS, linking well-established mechanisms of fluoroquinolone resistance.

Besides assessing the performance of the evaluation model, the results show that combining machine-learning algorithms based on phylogenetic and biomarker analyses provides a more complete view of resistance dynamics. The hybrid approach balances predictive accuracy and biological interpretability, which is essential for making the method useful in a clinical setting.

Future studies aim to include explainable AI models, promoting transparency and building confidence in automated AMR detection. Additionally, efforts are underway to assess the framework across larger, geographically diverse E. coli datasets and to expand its use to other priority pathogens. Overall, this research advances data-driven AMR surveillance and takes a scalable step toward transparent, AI-supported diagnostics in microbial genomics.

This study demonstrates that integrating phenotypic and genomic data with machine learning can accurately improve the prediction of antimicrobial resistance in *E. coli.* Expanding the emerging field of data-driven antimicrobial surveillance, provides a robust framework for more accurate diagnostics and evidence-based antibiotic management, strengthening global AMR mitigation efforts.

Author Contributions: Sarah H Adeyemi contributed to data curation, methodology, analysis, visualization, and the original draft, and Dr. Roshan Paudel contributed to the supervision, validation, reviewing, and editing.

Funding: This work received no specific grant from any funding agency. The work is supported by a TITLE III contract.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All processed data and analysis scripts will be made publicly available via GitHub upon publication.

Acknowledgments: The authors acknowledge Roshan Paudel for his supervision, mentorship, and valuable guidance throughout this study. We also extend our appreciation to the Title III Contract program for providing institutional support and to the Bacterial and Viral Bioinformatics Resource Center (BV-BRC) for granting access

 doi:10.20944/preprints202603.1564.v1

to genomic and phenotypic data resources used in this research and to Morgan State University for providing computational facilities and an enabling research environment.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AMR	Antimicrobial Resistance
E. coli	*Escherichia coli*
BV-BRC	Bacterial and Viral Bioinformatics Resource Center
CARD	Comprehensive Antibiotic Resistance Database

References

1. Murray, C. J., Ikuta, K. S., Sharara, F., Swetschinski, L., Robles Aguilar, G., Gray, A., … Naghavi, M. (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, 399(10325), 629–655. https://doi.org/10.1016/S0140-6736(21)02724-0
2. Shrestha P, Cooper BS, Coast J, Oppong R, Do Thi Thuy N, Phodha T, Celhay O, Guerin PJ, Wertheim H, Lubell Y. 2018. Enumerating the economic cost of antimicrobial resistance per antibiotic consumed to inform the evaluation of interventions affecting their use. *Antimicrob Resist Infect Control* 7:98. https://doi.org/10.1186/s13756-018-0384-3.
3. Kollef M.H., Shorr A.F., Bassetti M., Timsit J.F., Micek S.T., Michelson A.P., Garnacho-Montero J. Timing of antibiotic therapy in the ICU. *Crit. Care*. 2021;25:360. https://doi.org/10.1186/s13054-021-03787-z. [DOI] [PMC free article] [PubMed] [Google Scholar][Ref list]
4. Ruiz-Blanco et al. (2022).Ruiz-Blanco YB, Agüero-Chapin G, Romero-Molina S, Antunes A, Olari LR, Spellerberg B, Münch J, Sanchez-Garcia E. ABP-Finder: a tool to identify antibacterial peptides and the gram-staining type of targeted bacteria. *Antibiotics*. 2022;11(12):1708. https://doi.org/10.3390/antibiotics11121708. [DOI] [PMC free article] [PubMed] [Google Scholar][Ref list]
5. World Health Organization. Executive summary: the selection and use of essential medicines 2019: report of the 22nd WHO Expert Committee on the selection and use of essential medicines: WHO Headquarters, Geneva, 1-5 April 2019. https://apps.who.int/iris/handle/10665/325773 (2019). [Ref list]
6. [6] Michal Chowers, Tamir Zehavi, Bat Sheva Gottesman, Avi Baraz, Daniel Nevo, Uri Obolski, Estimating the impact of cefuroxime versus cefazolin and amoxicillin/clavulanate use on future collateral resistance: a retrospective comparison, *Journal of Antimicrobial Chemotherapy*, Volume 77, Issue 7, July 2022, Pages 1992–1995, https://doi.org/10.1093/jac/dkac130
7. [7] Alison C Tribble, Brian R Lee, Kelly B Flett, Lori K Handy, Jeffrey S Gerber, Adam L Hersh, Matthew P Kronman, Cindy M Terrill, Mike Sharland, Jason G Newland, for the Sharing Antimicrobial Reports for Pediatric Stewardship (SHARPS) Collaborative , Appropriateness of Antibiotic Prescribing in United States Children's Hospitals: *A National Point Prevalence Survey, Clinical Infectious Diseases*, Volume 71, Issue 8, 15 October 2020, Pages e226–e234, https://doi.org/10.1093/cid/ciaa036
8. eEML - Electronic Essential Medicines List. https://list.essentialmeds.org/. [Ref list]
9. Loscalzo, J. et al. Harrison's Principles of Internal Medicine, (Vol. 1 & Vol. 2). (McGraw Hill Professional, 2022). [Ref list]
10. Sharma, P. C., Jain, A., Jain, S., Pahwa, R., & Yar, M. S. (2010). Ciprofloxacin: review on developments in synthetic, analytical, and medicinal aspects. *Journal of Enzyme Inhibition and Medicinal Chemistry*, 25(4), 577–589. https://doi.org/10.3109/14756360903373350
11. Thomson CJ. The global epidemiology of resistance to ciprofloxacin and the changing nature of antibiotic resistance: a 10 year perspective. *J Antimicrob Chemother*. 1999 Mar;43 Suppl A:31-40. https://doi.org/10.1093/jac/43.suppl_1.31. PMID: 10225569.

 doi:10.20944/preprints202603.1564.v1

12. Dalhoff A. Global fluoroquinolone resistance epidemiology and implictions for clinical use. *Interdiscip. Perspect. Infect. Dis.* 2012;2012:976273. https://doi.org/10.1155/2012/976273. [DOI] [PMC free article] [PubMed] [Google Scholar][Ref list]
13. Low M, et al. Association between urinary community-acquired fluoroquinolone-resistant Escherichia coli and neighbourhood antibiotic consumption: a population-based case-control study. Lancet Infect. Dis. 2019;19:419–428. https://doi.org/10.1016/S1473-3099(18)30676-5. [DOI] [PubMed] [Google Scholar][Ref list]
14. Eliopoulos GM, Cosgrove SE, Carmeli Y. The impact of antimicrobial resistance on health and economic outcomes. *Clin. Infect. Dis.* 2003;36:1433–1437. https://doi.org/10.1086/375081. [DOI] [PubMed] [Google Scholar][Ref list]
15. Gottesman BS, Carmeli Y, Shitrit P, Chowers M. Impact of quinolone restriction on resistance patterns of Escherichia coli isolated from urine by culture in a community setting. *Clin. Infect. Dis.* 2009;49:869–875. https://doi.org/10.1086/605530. [DOI] [PubMed] [Google Scholar][Ref list]
16. Mintz I, Chowers M, Obolski U. Prediction of ciprofloxacin resistance in hospitalized patients using machine learning. *Commun Med* (Lond). 2023 Mar 28;3(1):43. https://doi.org/10.1038/s43856-023-00275-z. PMID: 36977789; PMCID: PMC10050086.
17. Breiman, L., Friedman, J., Olshen, R.A., & Stone, C.J. (1984). Classification and Regression Trees (1st ed.). *Chapman and Hall/CRC*. https://doi.org/10.1201/9781315139470
18. Collins, M.; Schapire, R.E.; Singer, Y. Logistic Regression, AdaBoost and Bregman Distances. *Mach. Learn.* 2002, 48, 253–285.{https://api.semanticscholar.org/CorpusID:207651918} [Google Scholar] [CrossRef]
19. Burges, C.J.C. A Tutorial on Support Vector Machines for Pattern Recognition. Data Min. Knowl. Discov. 1998, 2, 121–167. https://doi.org/10.1023/A:1009715923555 [Google Scholar] [CrossRef]
20. O. L. Mangasarian and David R. Musicant. 2001. Lagrangian support vector machines. J. Mach. Learn. Res. 1 (9/1/2001), 161–177. https://doi.org/10.1162/15324430152748218. [Google Scholar]
21. Arana-Daniel, N.; Gallegos, A.A.; López-Franco, C.; Alanís, A.Y.; Morales, J.; López-Franco, A. Support Vector Machines Trained with Evolutionary Algorithms Employing 22. Kernel Adatron for Large Scale Classification of Protein Structures. Evol. Bioinform. 2016, 12, EBO.S40912–302. [Google Scholar] [CrossRef] [PubMed]
23. Hosmer, D.W., Jr.; Lemeshow, S.; Sturdivant, R.X. Applied Logistic Regression; *John Wiley & Sons*: Hoboken, NJ, USA, 2013.https://doi.org/10.1002/0471722146 .[Google Scholar]
24. Lewis CM, Knight J. Introduction to genetic association studies. *Cold Spring Harb Protoc*. 2012 Mar 1;2012(3):297-306. https://doi.org/10.1101/pdb.top068163. PMID: 22383645.
25. Lowrie EG, Lew NL. Death risk in hemodialysis patients: the predictive value of commonly measured variables and an evaluation of death rate differences between facilities. *Am J Kidney Dis.* 1990 May;15(5):458-82. https://doi.org/10.1016/s0272-6386(12)70364-5. PMID: 2333868. [Google Scholar] [CrossRef]
26. Langer DL, van der Kwast TH, Evans AJ, Trachtenberg J, Wilson BC, Haider MA. Prostate cancer detection with multi-parametric MRI: logistic regression analysis of quantitative T2, diffusion-weighted imaging, and dynamic contrast-enhanced MRI. *J Magn Reson Imaging*. 2009 Aug;30(2):327-34. https://doi.org/10.1002/jmri.21824. PMID: 19629981.
27. Redgrave, L. S., Sutton, S. B., Webber, M. A., & Piddock, L. J. V. (2014). Fluoroquinolone resistance: Mechanisms, impact on bacteria, and role in evolutionary success. *Trends in Microbiology*, 22(8), 438–445. https://doi.org/10.1016/j.tim.2014.04.007
28. Shaik S, Ranjan A, Tiwari SK, Hussain A, Nandanwar N, Kumar N, Jadhav S, Semmler T, Baddam R, Islam MA, Alam M, Wieler LH, Watanabe H, Ahmed N. Comparative Genomic Analysis of Globally Dominant ST131 Clone with Other Epidemiologically Successful Extraintestinal Pathogenic Escherichia coli (ExPEC) Lineages. mBio. 2017 Oct 24;8(5):e01596-17. https://doi.org/10.1128/mBio.01596-17. PMID: 29066550; PMCID: PMC5654935.
29. Huang L, Wu C, Gao H, Xu C, Dai M, Huang L, Hao H, Wang X, Cheng G. Bacterial Multidrug Efflux Pumps at the Frontline of Antimicrobial Resistance: An Overview. *Antibiotics* (Basel). 2022 Apr 13;11(4):520. https://doi.org/10.3390/antibiotics11040520. PMID: 35453271; PMCID: PMC9032748.

 doi:10.20944/preprints202603.1564.v1

30. Thänert R, Choi J, Reske KA, et al. Persisting uropathogenic Escherichia coli lineages show signatures of niche-specific within-host adaptation mediated by mobile genetic elements. *Cell Host & Microbe*. 2022 Jul;30(7):1034-1047.e6. https://doi.org/10.1016/j.chom.2022.04.008. PMID: 35545083; PMCID: PMC10365138.