



MACHINE LEARNING-BASED PREDICTION OF ANTIBIOTIC RESISTANCE USING PHENOTYPIC SUSCEPTIBILITY PATTERNS AND RESISTANCE GENE PROFILES

Chinedu E. Okafor^{1*}, Amina Yusuf², Ibrahim M. Bello³

^{1*}Department of Microbiology, University of Nigeria, Nsukka, Nigeria, okafor@yahoo.com

²Department of Medical Laboratory Science, Ahmadu Bello University, Zaria, Nigeria

³Department of Biological Sciences, University of Abuja, Abuja, Nigeria

***Corresponding Author:** Chinedu E. Okafor, okafor@yahoo.com

Received: 20/01/2026, Revised: 24/03/2026, Accepted: 13/05/2026, Published: 27/05/2026

ABSTRACT

Antibiotic resistance is one of the current challenges for successful antimicrobial therapy and calls for analytical methods that can combine phenotypic and molecular data. The aim of this study was to create machine-learning models that predict antibiotic-specific resistance that can be derived from phenotypic susceptibility patterns, resistance gene profiles, and contextual variables. Susceptibility results for amoxicillin, ciprofloxacin, meropenem, vancomycin, and colistin were analyzed besides KPC, NDM-1, OXA-48, and VIM resistance genes. Data preprocessing comprised categorical encoding, evaluation of missing data, partitioning of data in groups by patient identifier, and excluding each target antibiotic from its respective predictor set to avoid information leakage. Accuracy, precision, sensitivity, specificity, F1-score, ROC-AUC, and PR-AUC were used to compare Logistic Regression, Decision Tree, Random Forest, Support Vector Machine and Gradient Boosting models. Susceptibility categories and resistance gene frequencies were fairly even, and genotype-phenotype relationships were not strong. The best mean ROC-AUCs were not far off from chance-level discrimination by the machine-learning algorithms. The inclusion of resistance gene and contextual variables did not significantly increase the predictive value over phenotypic variables. These results will suggest that a more accurate prediction of resistance will need to be developed with more genomic data, more populations, and validation outside of the clinical setting before it can be used in the clinic. However, the framework offers a clear foundation to assess integrated antimicrobial resistance predictors

Keyword: Antimicrobial resistance; Machine learning; Phenotypic susceptibility; Resistance genes; Antibiotic resistance prediction

Introduction

AMR is now one of the greatest global health threats due to its impacts on the effectiveness of treatment, treatment failure, hospitalization, and cost of care. Detection and therapeutic decision-making strategies against resistance are needed as the problem of resistant bacterial infections is growing. Inappropriate use of antimicrobials, transmission of resistant organisms, and acquisition or emergence of resistance determinants are responsible for the emergence and spread of AMR; public health and clinical microbiology surveillance and predictive analytics are crucial to the control of AMR (Fernandez, 2022). Machine learning has been explored as a learning tool to classify resistance using a more complex approach than traditional statistical methods, with the capability to recognize patterns within microbiological data (Anahtar et al., 2021).

However, traditional antimicrobial susceptibility testing is still very important to discern whether a bacterial isolate is sensitive, intermediate, or resistant to a particular antibiotic as well. Phenotypic methods can be time-consuming in the laboratory, however and may not fully explain the molecular mechanisms of resistance. Genotypic information can be used in addition to the phenotypic information to detect the genes associated with resistance and genetic determinants. The genotype/phenotype concept is complex since the presence of a resistance gene is not always 100% correlated with the phenotypic expression. Due to the differences in gene regulation, mutations, plasmid content, and bacterial background, there is the possibility of genotype-phenotype discordance (Petrin et al., 2023). For this reason, there is value in linked phenotypic and genetic data for research and predictive modelling of AMR (Chindelevitch et al., 2023).

Resistance genes related to carbapenemase production are significant since they have the potential of undermining effective antibiotic therapy. Ndm-mediated metallo-beta-lactamases are spread globally and pose a serious threat due to their wide hydrolytic spectrum and horizontal gene transfer (Farhat & Khan, 2020). The Comprehensive Antibiotic Resistance Database (CARDdb) has been used to aid in the identification and annotation of resistance determinants in an internationally consistent manner and has enabled computational and machine-learning applications to support resistome analysis (Alcock et al., 2023). The resources provide enhanced support for biological underpinning of predictive models based on genomic and resistance-gene attributes.

Machine learning offers a model to incorporate phenotypic, genotypic, demographic and clinical factors. Algorithms such as logistic regression, decision trees, random forests, support vector machines, and gradient boosting can capture nonlinear relationships and interactions between variables that are not captured by traditional algorithms. It has been shown that antimicrobial resistance can be predicted from genome data (Van Camp et al., 2020) and resistance mechanisms and new genetic factors can be discovered by machine-learning approaches (Sunuwar & Azad, 2021). Whole-genome sequencing and

machine learning has now further increased the ability to model resistance in bacterial populations (Ren et al., 2022).

However, although these developments have been made, predictive performance is dependent on data quality, representation of the features, bacterial species, antibiotic class, structure of the population and validation strategy. The review of the AMR data-mining methods has pointed out methodological heterogeneity across prediction studies (Li et al., 2021). Some external validation is still not consistent, some data is hard to interpret, there is a risk of data leakage, and there are differences between the development dataset and the population that is used in real-life settings (Kim et al., 2022). New evidence shows that machine-learning models can be usefully discriminatory in some cases, but models have a wide range of performance depending on organisms and resistance outcomes (Lv & Wang, 2024).

There is growing research that incorporates genomic data and machine learning to enhance susceptibility prediction. The underlying idea of whole-genome sequencing (WGS) is that when applied alongside the correct analytical tools, it can yield informative predictors for clinically relevant pathogens, as suggested by systematic evidence. Whole genome sequencing (WGS) is a method that, when combined with the appropriate analytical tools, has the potential to predict clinically relevant pathogens and inform the underlying idea of the WGS method. AI is also finding applications in the diagnosis of infectious diseases, antimicrobial discovery, monitoring of resistance, and support in therapeutic decisions (Cesaro et al., 2025). The use of machine learning alongside antimicrobial discovery in integrative strategies represents the growing impact of computational methods in combating AMR (Bano et al., 2025). Recent viewpoints also highlight the potential of AI resistance prediction for precision therapy, provided the models are strong, explainable and tested (Singh et al., 2026).

Phenotypic susceptibility patterns combined with resistance gene profiles can then be used together – a sensible way to explore the potential for improved prediction with complementary resistance information. This integration can help decide if genetic susceptibility variables offer further information over and above the observed susceptibility pattern and what the limitations of the observed susceptibility variables are.

Objectives:

- To describe phenotypic susceptibility patterns and to determine the distribution of the major resistance gene profiles.
- To design and evaluate machine-learning models to predict antibiotic-specific resistance, based on phenotypic, genetic and contextual predictors.
- To assess the predictive performance and model interpretation of resistance gene information and phenotypic susceptibility features.

2. Materials and Methods

2.1 Study Design and Data Source

A retrospective secondary-data analytical design, supervised machine learning, and antibiotic-resistance prediction were used in this study with phenotypic susceptibility patterns and resistance gene profiles. The analysis was based on a data set containing 2200 observations and 12 variables called the Antibiotic Resistance Tracking Dataset. Data collected included demographic data, specimen characteristics, phenotypic susceptibility results for amoxicillin, ciprofloxacin, meropenem, vancomycin, and colistin, laboratory test method, resistance gene data and clinical outcome. The analytical unit was the individual microbiological record, and the main aim was to assess the possibility of using phenotypic and genetic characteristics to predict resistance to antibiotics (Siam & Faiza, 2025).

2.2 Study Variables

The key variables of interest were the susceptibility classes for amoxicillin, ciprofloxacin, meropenem, vancomycin and colistin; each susceptibility class was categorized as Sensitive, Intermediate, or Resistant. Primary binary machine learning analysis: Resistant observations were classified as the positive class, sensitive observations as the negative class, and Intermediate observations were retained for descriptive analysis or evaluated separately if applicable. Susceptibility patterns of the remaining antibiotics, gene category of resistance, age, gender, type of specimen, and the method used for testing were predictor variables. KPC, NDM-1, OXA-48 and VIM were included as resistance gene information.

2.3 Data Preprocessing

Data in the dataset were checked for duplicate records, missing data, misspellings of labels and duplicate patient ID numbers prior to analysis. Categorical variables were coded as binary or one-hot encoding, as needed, and age was left continuous. Resistance gene values that were not found were evaluated independently and were not necessarily viewed as a lack of resistance gene unless indicated by the dataset definition. Patient_ID values were also repeated when partitioning data to ensure that minimal information could be leaked between the training and testing subsets.

2.4 Phenotypic and Resistance Gene Analysis

Descriptive analysis was conducted to obtain the frequency and percentage distribution of Sensitive, Intermediate and Resistant classifications for each antibiotic. Additionally, the prevalence of the resistance genes KPC, NDM-1, OXA-48 and VIM was determined. Associations between resistance gene categories and phenotypic susceptibility patterns were examined by cross-tabulation. Categorical associations were assessed using chi-square tests of independence, and Fisher's exact test was used when the expected frequencies were too low. A p-value of < 0.05 was considered to be statistically significant.

2.5 Feature Engineering and Predictor Configuration

Model-ready features and summary measures of resistance, as appropriate, were created from phenotypic susceptibility variables. For target leakage, the susceptibility value of the antibiotic being predicted was not included in the predictor set. For instance, if meropenem resistance was applied as the target, the meropenem susceptibility was not taken into account as an input variable. Three predictor configurations were tested: phenotypic susceptibility alone, resistance gene plus contextual variables, and a combined model with phenotypic susceptibility, resistance gene and contextual variables.

2.6 Machine-Learning Model Development

Machine-learning models were separately trained to predict each antibiotic resistance outcome, with supervised learning. The algorithms used were: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Gradient Boosting or XGBoost. Logistic regression was used as a baseline model that is interpretable, while tree-based and kernel-based models were used to model more complex relationships between susceptibility patterns, resistance genes, and antibiotic resistance. The hyperparameters were tuned by cross-validation on the training set.

2.7 Model Training and Validation

The dataset was split into training and test subsets with a roughly 80:20 ratio with equal distribution of resistance classes. When repeated patient identifiers were for patients who had been observed multiple times, these were grouped together and the patients were not included in both the training and testing sets. Five-fold stratified cross-validation was used for model development and tuning in the training data set. The independent test dataset was used for the final evaluation of the performance but not during the hyperparameter optimization.

2.8 Model Performance Evaluation

The measures used to assess predictive performance were accuracy, precision, sensitivity, specificity, F1-score, receiver operating characteristic (ROC) area under the curve and precision-recall area under the curve. Confusion matrices were also analyzed to find true positive, true negative, false positive and false negative classifications. Different machine learning algorithms and feature sets were tested to see if adding resistance gene data added value to prediction beyond patterns of phenotypic susceptibility.

2.9 Model Interpretation

Feature importance was calculated to determine the most important phenotypic and genetic features in predicting antibiotic resistance. Where applicable, feature importance, permutation importance or SHAP interpretation was applied to trees. The role of KPC, NDM-1, OXA-48 and of other antibiotic

susceptibility patterns was specifically highlighted to establish the relative importance of phenotype in model-based predictions of the genotype.

2.10 Statistical and Computational Analysis

Descriptive statistics included frequencies and percentages for categorical variables and means and standard deviations or medians and interquartile ranges for continuous variables, as appropriate based on the distribution of data. Statistical tests (inferential) were applied as two-tailed tests with a p-value < 0.05 as the significance criterion. The machine learning, preprocessing, statistical analysis and visualization were performed in Python with suitable analytical libraries.

2.11 Ethical Considerations

No human subjects were involved in the study, and secondary data were used exclusively from the Antibiotic Resistance Tracking Dataset, which does not include any contact or recruitment with human participants or direct intervention. Patient identifiers were only used to evaluate repeated observations and were excluded from predictive modelling to avoid data leakage and repeated observations of patient identifiers. An effort was not made to identify individual patients. The analysis was performed following the guidelines for confidentiality, the responsible use of secondary data, the possibility of reproducing the analysis and adherence to the terms and licences of the source data.

3. Results

3.1 Characteristics of the Study Dataset

There were 2,200 microbiological records in the Antibiotic Resistance Tracking Dataset. The mean age was 45.65 ± 26.27 years, with a median age of 45 years and a range of 1–90 years. Females accounted for 1,110 records (50.5%), while males represented 1,090 records (49.5%). The distribution of specimen types was fairly even, with blood accounting for 20.4 per cent of the records, sputum and wound swabs for 20.4 percent, urine for 19.7 percent and stool for 19.2 percent. Susceptibility testing methods recorded the most frequently used were Disc diffusion (34.3%), MIC testing (32.9%) and automated systems (32.8%). Table 1 summarizes the main aspects of the study dataset.

Table 1. Demographic, specimen, laboratory, and outcome characteristics of the study dataset

Characteristic	Category/Statistic	n	%
Total records	—	2,200	100.0
Age, years	Mean $\pm$ SD	45.65 ± 26.27	—
Age, years	Median	45	—

Age, years	Range	1–90	—
Gender	Female	1,110	50.5
	Male	1,090	49.5
Specimen type	Blood	449	20.4
	Sputum	448	20.4
	Wound swab	448	20.4
	Urine	433	19.7
	Stool	422	19.2
Test method	Disc diffusion	755	34.3
	MIC	724	32.9
	Automated system	721	32.8
Clinical outcome	ICU	775	35.2
	Recovered	739	33.6
	Deceased	686	31.2

The fairly uniform distribution of specimen types and laboratory methods minimized the risk that a particular specimen or laboratory type would swamp subsequent analysis.

3.2 Phenotypic Antibiotic Susceptibility Patterns

All five antibiotics showed widespread distributions of the three classifications: Sensitive, Intermediate and Resistant. The highest number and percentage of resistant observations were seen in the case of meropenem, where 755 observations were recorded corresponding to 34.3% of the observations. Amoxicillin resistance was found in 730 records (33.2%) followed by ciprofloxacin and vancomycin at 32.6 % each. The resistance proportion of colistin was the lowest at 31.9%. Phenotypic susceptibility distributions are presented completely in Table 2.

Table 2. Distribution of phenotypic susceptibility classifications across antibiotics

Antibiotic	Sensitive, n (%)	Intermediate, n (%)	Resistant, n (%)
Amoxicillin	726 (33.0)	744 (33.8)	730 (33.2)
Ciprofloxacin	752 (34.2)	730 (33.2)	718 (32.6)
Meropenem	708 (32.2)	737 (33.5)	755 (34.3)
Vancomycin	757 (34.4)	726 (33.0)	717 (32.6)
Colistin	761 (34.6)	738 (33.5)	701 (31.9)

Figure 1 illustrates the distribution of the different susceptibility classes for each antibiotic and shows no antibiotic was clearly predominant. This was a relatively balanced distribution of phenotypes, with no major class imbalance, which was suitable for resistance classification.

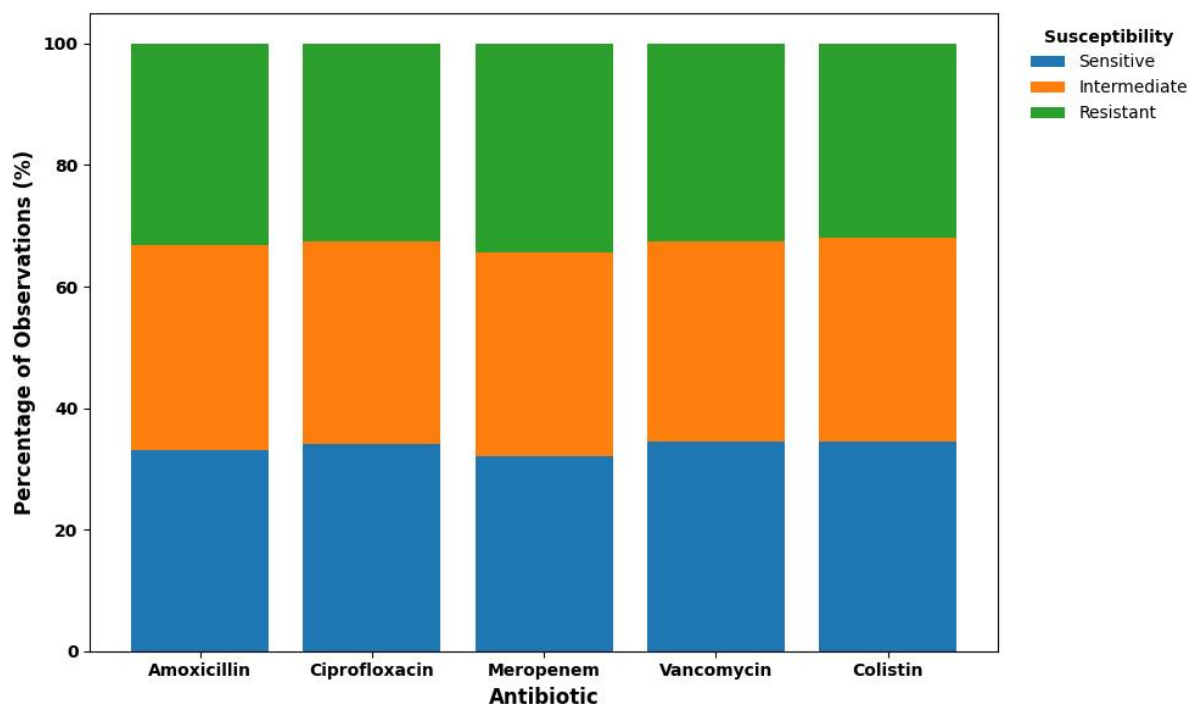


Figure 1. Distribution of Sensitive, Intermediate, and Resistant phenotypes across the five antibiotics

3.3 Distribution of Resistance Gene Profiles

Of the 2,200 records, VIM was the most common recorded resistance gene at 463 (21.0%) of the records. 450 records contained OXA-48 (20.5%) and 425 records contained KPC (19.3%) and NDM-1 (19.3%). In 437 observations (19.9%) no resistance gene information was recorded. Table 3 summarizes the distribution of the resistance genes, and Figure 2 shows the gene frequencies graphically.

Table 3. Distribution of recorded resistance gene profiles

Resistance gene category	n	%
VIM	463	21.0
OXA-48	450	20.5
KPC	425	19.3
NDM-1	425	19.3
Not recorded	437	19.9
Total	2,200	100.0

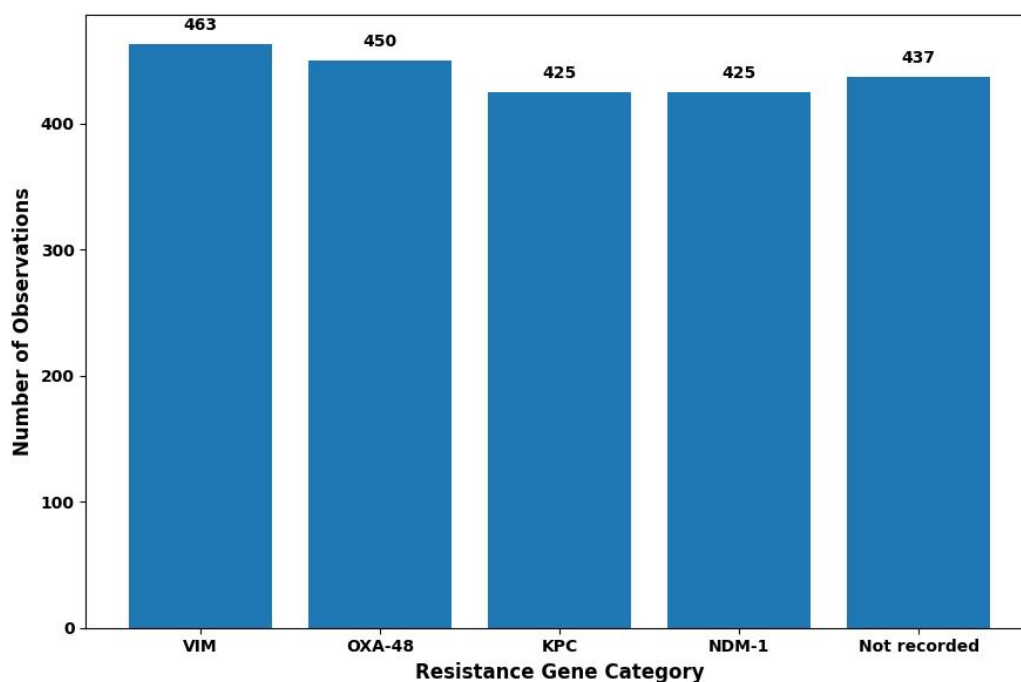


Figure 2. Frequency distribution of recorded resistance gene categories in the study dataset

The relative proportions of resistances within each gene category also were fairly consistent for the five antibiotics. For instance, the resistance rates for meropenem are 31.8% for NDM-1 and 35.6% for OXA-48 in OXA-48 records, respectively. And the resistance rates to amoxiillin are 28.9% and 36.5% for NDM-1 and OXA-48, respectively, in the OXA-48 records. These genotype–phenotype patterns are visualized in Figure 3.

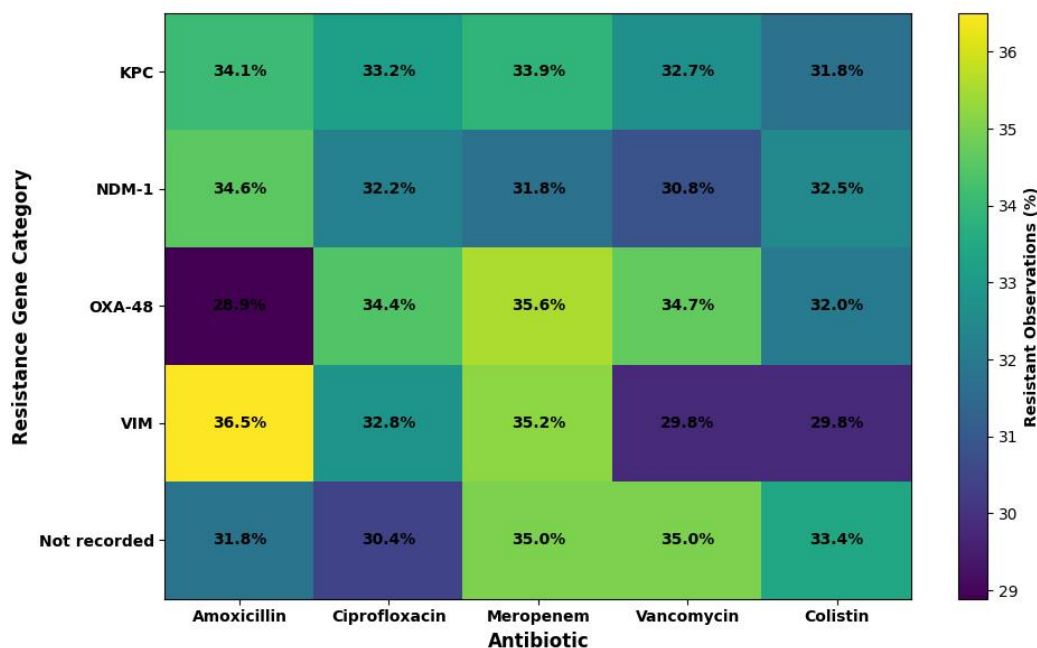


Figure 3. Heatmap showing the percentage of phenotypically resistant observations for each antibiotic across resistance gene categories

3.4 Association Between Resistance Gene Profiles and Phenotypic Resistance

Statistically significant associations with susceptibility classification and the recorded resistance gene category were not found for any of the five types of antibiotics for the Chi-square analyses. The association was weakest with meropenem ($\chi^2 = 2.77$, $p = 0.948$) and the greatest chi-square resulted for amoxicillin ($\chi^2 = 8.28$, $p = 0.407$) as seen in Table 4. The associations among the available gene categories and phenotypic susceptibility classifications were all very low, with all Cramer's V values less than 0.05.

Table 4. Association between resistance gene category and antibiotic susceptibility phenotype

Antibiotic	χ^2	df	p-value	Cramer's V
Amoxicillin	8.276	8	0.407	0.043
Ciprofloxacin	4.590	8	0.800	0.032
Meropenem	2.770	8	0.948	0.025
Vancomycin	6.480	8	0.594	0.038
Colistin	3.579	8	0.893	0.029

Results of this analysis suggest that there was no strong discrimination between the Resistant, Intermediate and Sensitive phenotypes in this dataset using only the available category of resistance gene.

3.5 Machine-Learning Prediction of Antibiotic Resistance

Intermediate cases were not included and Resistant observations were paired against Sensitive observations for primary binary machine-learning analysis. The resulting sample sizes for the two groups were between 1,456 and 1,474 observations, depending on the type of antibiotic used. To avoid target leakage, each antibiotic was removed from the predictor matrix. Overall predictive discrimination was moderate for the five antibiotics. Significant results included Support Vector Machine with the highest mean ROC-AUC (0.516), followed by Gradient Boosting and Random Forest, both of which had a mean ROC-AUC close to chance level. Gradient Boosting achieved the best performance in terms of mean accuracy (0.506) and mean F1 score (0.470). The performance of the models is presented in a comparison in Table 5.

Table 5. Mean predictive performance of machine-learning algorithms across the five antibiotic targets

Model	Accuracy	Precision	Sensitivity	Specificity	F1-score	ROC-AUC	PR-AUC
Logistic Regression	0.485	0.495	0.410	0.566	0.441	0.485	0.507
Decision Tree	0.494	0.499	0.381	0.612	0.411	0.488	0.508
Random Forest	0.492	0.500	0.443	0.546	0.462	0.492	0.522
Support Vector Machine	0.487	0.494	0.437	0.539	0.458	0.516	0.536
Gradient Boosting	0.506	0.516	0.446	0.573	0.470	0.493	0.513

For individual antibiotics, Gradient Boosting had the best ROC-AUC of 0.533 for amoxicillin, Support Vector Machine the best ROC-AUC of 0.541 for both ciprofloxacin and vancomycin. Random Forest and Decision Tree had the highest ROC-AUC for colistin (0.516 and 0.519, respectively), and for meropenem, Random Forest and logistic regression had the highest values (0.516 and 0.514, respectively). Even so, none of the models showed good discrimination with all best antibiotic-specific ROC-AUC values being around 0.52 to 0.54.

The ROC curves presented in Figure 4 also highlight the lack of separation between resistant and sensitive observations. The curves were close to the diagonal reference line, indicating that the information available for the phenotype, gene, demographic, specimen and testing variables was limited in their ability to predict the binary resistance outcomes.

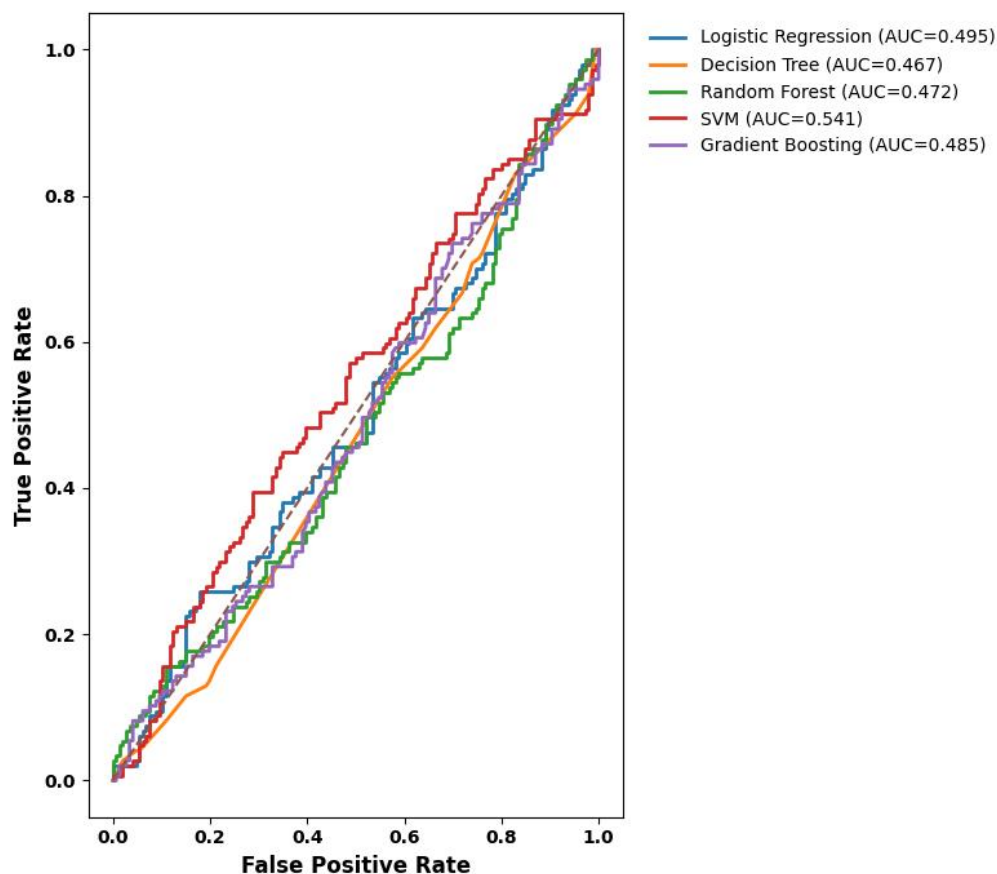


Figure 4. Receiver operating characteristic curves comparing machine-learning classifiers for ciprofloxacin resistance prediction

3.6 Comparative Contribution of Phenotypic and Resistance Gene Features

To assess the effect of using gene and phenotypic data to enhance the prediction of resistance, the Random Forest models were also tested with phenotypic-only, gene-plus-contextual, and gene-plus-phenotype models. The data for each of the five antibiotic targets are reported as means in Table 6.

Table 6. Mean Random Forest performance according to predictor configuration

Predictor configuration	Accuracy	F1-score	ROC-AUC	PR-AUC	Sensitivity	Specificity
Phenotypic features only	0.501	0.471	0.509	0.518	0.443	0.557
Gene + contextual features	0.472	0.462	0.474	0.507	0.457	0.491
Combined features	0.492	0.462	0.492	0.522	0.443	0.546

The models with phenotypic data only had the highest mean ROC-AUC score of 0.509 and the highest mean accuracy score of 0.501. The overall ROC-AUC was not improved by adding resistance gene and

contextual information, but the combination of resistance gene and contextual information resulted in a slightly higher PR-AUC of 0.522. The comparison shown in Figure 5 suggests that the three configurations were comparable and achieved near chance discrimination.

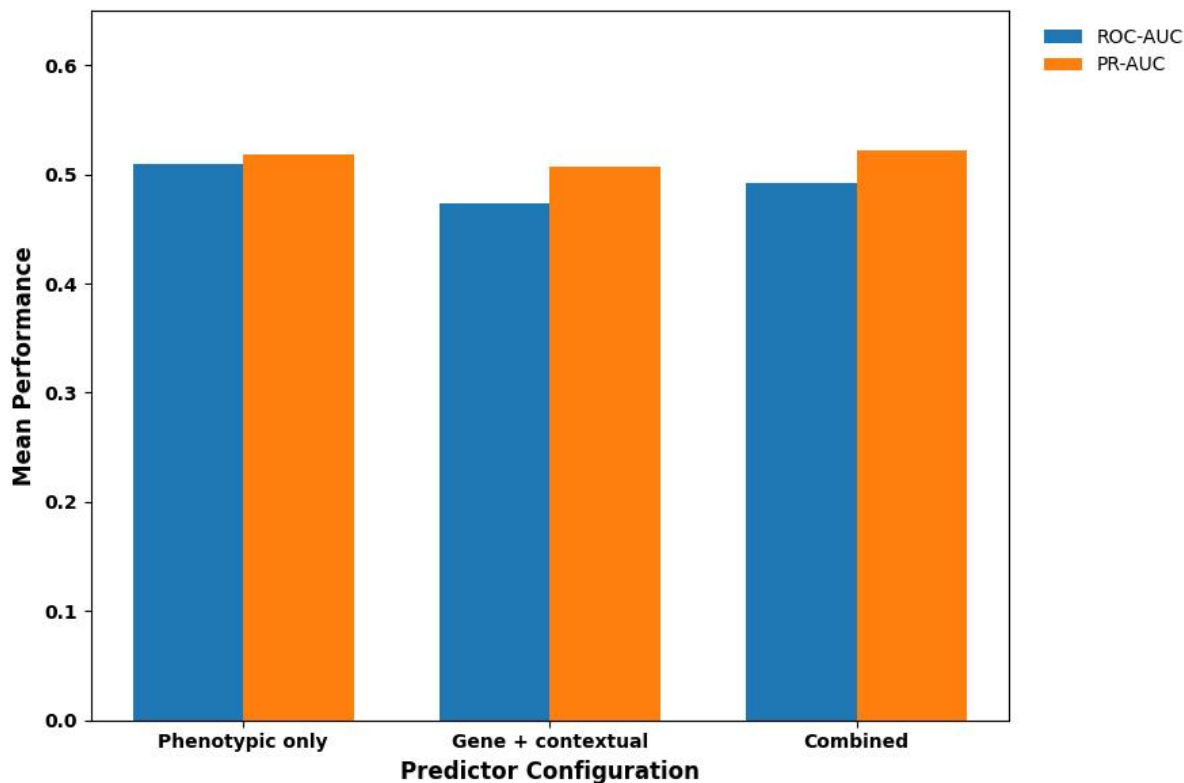


Figure 5. Comparison of mean ROC-AUC and PR-AUC across phenotypic-only, gene-plus-contextual, and combined Random Forest predictor configurations

Overall, balanced susceptibility was found, as well as a relatively even distribution of the resistance gene categories, but the resistance genes exhibited weak univariate associations with phenotypic susceptibility and machine-learning models had limited predictive discrimination. The results thus indicate that the variables included in this dataset account for resistance status descriptively, but are not enough to reliably predict resistance when the susceptibility result for the antibiotic of interest is excluded from the predictor matrix.

4. Discussion

In the present work, the feasibility of predicting antibiotic-specific resistance was assessed by the combination of phenotypic susceptibility, resistance gene profile, and contextual variables by supervised machine-learning algorithms. The susceptibility categories were fairly equally distributed for the five antibiotics and the VIM, OXA-48, KPC and NDM-1 were observed at comparable rates. But the weak associations between recorded resistance genes and phenotypic susceptibility and the modest discrimination in the models indicate that other genes or gene combinations must significantly influence the individuals that are susceptible to *Armillaria*. The results suggest that the use of predictive models and

the choice of algorithm is only as successful as the biological richness, granularity and independence of the available features.

The relative performance of Logistic Regression, Decision Tree, Random Forest, Support Vector Machine and Gradient Boosting across the datasets and resistance targets is consistent with the evidence of the large variance in machine-learning performance across different datasets and resistance targets. There have been significant variations in model discrimination across different reviews due to variations in organisms, antibiotics, sample size, feature construction and validation methods (Lv & Wang, 2024). These issues also exist in the prediction of AMR in genomic and clinical applications, where success in a small set of data may fail to extend to larger populations (Kim et al., 2022). As such, the limited results shown here should be interpreted as indicative of the lack of predictive power within the variables available rather than as a failure of machine learning. This presentation is significant as the class distributions were nicely balanced so that the concern that poor discrimination was due to severe outcome imbalance was diminished. The results instead suggest that the lack of feature specificity and low molecular resolution are the most likely limitations to predictive performance.

Previous studies have demonstrated that richer genomic representations can lead to better resistance prediction. In several bacterial species, whole genome sequencing and engineered genomic features have helped predict (Van Camp et al., 2020; Ren et al., 2022). The use of feature extraction has helped to improve AMR modelling in *Acinetobacter baumannii* by identifying informative determinants (Gao et al., 2024). It has been predicted that the whole genome sequence of *E. coli* can be used to predict the susceptibility to cefepime and other resistance patterns (Humphries et al., 2023; Wan et al., 2026). The copy-number information can also generalize better than single nucleotide polymorphisms (SNPs) in some cases (Fistarol et al., 2025). These results indicate that single categorical resistance-gene variable is less informative than high-dimensional genomic profiles.

Biologically plausible weak genotype–phenotype relationships are also observed. Phenotypic resistance is not always directly correlated with the presence of resistance genes because expression levels, regulatory mechanisms of resistance genes, plasmid context, compensatory mutations and interactions between resistance genes can alter susceptibility. There is considerable genotype–phenotype variation among bacterial isolates (Petrin et al., 2023). AMR databases also demonstrate that resistance is not necessarily attributed to a single gene, allele, mutation or regulatory change, but to combinations of these (Alcock et al., 2023). In this regard, the use of wider sets of genomic features and multivariate learning methods is increasingly emerging in studies looking for unknown resistance determinants (Sunuwar & Azad, 2021).

The predictor-configuration analysis revealed that the inclusion of resistance-gene and contextual variables did not noticeably further enhance ROC-AUC over the phenotypic variables. This is important for susceptibility data from surveillance, as the predictive value might rely strongly on previous phenotypic trends and the local epidemiology (Valavarasu et al., 2025). In *E. coli*, similar susceptibility-

prediction frameworks indicate that the usefulness of models depends on the structure and representativeness of the input data (Talamantes-Becerra et al., 2024). Comparative studies on other organisms also show that the performance of the algorithm varies depending on the resistance mechanism and the availability of the features (Liu et al., 2020; Noman et al., 2023).

The findings also further confirm the need to be methodologically transparent. The need for careful preprocessing, leakage prevention, appropriate partitioning and biologically meaningful feature selection are highlighted by genomic machine-learning guidance (Orcales et al., 2024). This is important because the inclusion of the susceptibility result for the specific antibiotic would inflate the performance. Exclusion of each individual target antibiotic from its own predictor set resulted in more conservative and methodologically sound estimates.

Richer genomic profiles, multicenter samples, external validation and explainable AIs should be included in future studies. The integration of detailed genomic information with comprehensive validation frameworks is supported by marker-discovery research in *E. coli* and systematic reviews (Moat et al., 2025; Ardila et al., 2024). When informative data are available, broader AI approaches can be used to enhance resistance surveillance, therapeutic optimization and precision antimicrobial selection (de la Lastra et al., 2024; Bano et al., 2025). However, before models can be used in routine decision making, they must be validated, interpretable, and generalizable (Cesaro et al., 2025; Singh et al., 2026). In general, the integration of phenotypic susceptibility with only a few resistance-gene categories is possible, but more molecular data seems to be required to achieve more accurate and useful resistance prediction.

5. Conclusion

This study showed how machine learning can predict antibiotic resistance, using a combination of phenotypic susceptibility patterns, resistance gene profiles and contextual characteristics. The distribution of phenotypic susceptibility to amoxicillin, ciprofloxacin, meropenem, vancomycin and colistin was relatively balanced, as were the representation of KPC, NDM-1, OXA-48 and VIM resistance genes. The associations of resistance gene categories with phenotypic susceptibility, however, were weak, reflecting the fact that the resistance patterns observed could not be explained by individual gene categories. Modest predictive discrimination was further shown through comparative evaluations of Logistic Regression, Decision Tree, Random Forest, Support Vector Machine and Gradient Boosting and was comparable to chance for antibiotic-specific outcomes. The addition of resistance gene and contextual data did not significantly enhance the overall discrimination added to that of the phenotypic data. The results highlight the biological richness, specificity and quality of the predictor variables, not just the algorithm complexity, as being important to machine-learning performance. The modelling strategy was also leakage-safe, which meant it gave conservative performance estimates as well as methodologically sound. Future studies need to include detailed genomic information, more multicenter populations, longitudinal surveillance data, external validation, and explainable artificial intelligence (XAI) techniques. These

improvements can help with genotype–phenotype characterization and with the construction of accurate, meaningful, and clinically applicable antibiotic resistance prediction systems.

References

1. Alcock, B. P., Huynh, W., Chalil, R., Smith, K. W., Raphenya, A. R., Wlodarski, M. A., ... & McArthur, A. G. (2023). CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic acids research*, 51(D1), D690-D699.
2. Anahtar, M. N., Yang, J. H., & Kanjilal, S. (2021). Applications of machine learning to the problem of antimicrobial resistance: an emerging model for translational research. *Journal of clinical microbiology*, 59(7), 10-1128.
3. Ardila, C. M., Yadalam, P. K., & González-Arroyave, D. (2024). Integrating whole genome sequencing and machine learning for predicting antimicrobial resistance in critical pathogens: a systematic review of antimicrobial susceptibility tests. *PeerJ*, 12, e18213.
4. Bano, N., Mohammed, S. A., & Raza, K. (2025). Integrating machine learning and multitargeted drug design to combat antimicrobial resistance: a systematic review. *Journal of Drug Targeting*, 33(3), 384-396.
5. Cesaro, A., Hoffman, S. C., Das, P., & de la Fuente-Nunez, C. (2025). Challenges and applications of artificial intelligence in infectious diseases and antimicrobial resistance. *npj Antimicrobials and Resistance*, 3(1), 2.
6. Chindelevitch, L., van Dongen, M., Graz, H., Pedrotta, A., Suresh, A., Uplekar, S., ... & Wheeler, N. (2023). Ten simple rules for the sharing of bacterial genotype—Phenotype data on antimicrobial resistance. *PLoS Computational Biology*, 19(6), e1011129.
7. de la Lastra, J. M. P., Wardell, S. J., Pal, T., de la Fuente-Nunez, C., & Pletzer, D. (2024). From data to decisions: leveraging artificial intelligence and machine learning in combating antimicrobial resistance—a comprehensive review. *Journal of medical systems*, 48(1), 71.
8. Farhat, N., & Khan, A. U. (2020). Evolving trends of New Delhi Metallo-beta-lactamase (NDM) variants: A threat to antimicrobial resistance. *Infection, Genetics and Evolution*, 86, 104588.
9. Fernandez, G. (2022). Turning the juggernaut. *Lancet Planetary Health*, 6(2), E75-E75.
10. Fistarol, B. F., Gervasio, J. D., & Szöllösi, G. J. (2025). Gene copy-number features generalize better than SNPs for antimicrobial resistance prediction in *Staphylococcus aureus*. *npj Antimicrobials and Resistance*, 3(1), 100.
11. Gao, Y., Li, H., Zhao, C., Li, S., Yin, G., & Wang, H. (2024). Machine learning and feature extraction for rapid antimicrobial resistance prediction of *Acinetobacter baumannii* from whole-genome sequencing data. *Frontiers in microbiology*, 14, 1320312.
12. Humphries, R. M., Bragin, E., Parkhill, J., Morales, G., Schmitz, J. E., & Rhodes, P. A. (2023). Machine-learning model for prediction of cefepime susceptibility in *Escherichia coli* from whole-genome sequencing data. *Journal of clinical microbiology*, 61(3), e01431-22.
13. Kim, J. I., Maguire, F., Tsang, K. K., Gouliouris, T., Peacock, S. J., McAllister, T. A., ... & Beiko, R. G. (2022). Machine learning for antimicrobial resistance prediction: current practice, limitations, and clinical perspective. *Clinical microbiology reviews*, 35(3), e00179-21.
14. Li, X., Zhang, Z., Liang, B., Ye, F., & Gong, W. (2021). A review: antimicrobial resistance data mining models and prediction methods study for pathogenic bacteria. *The Journal of Antibiotics*, 74(12), 838-849.
15. Liu, Z., Deng, D., Lu, H., Sun, J., Lv, L., Li, S., ... & Wang, G. (2020). Evaluation of machine learning models for predicting antimicrobial resistance of *Actinobacillus pleuropneumoniae* from whole genome sequences. *Frontiers in microbiology*, 11, 48.
16. Lv, G., & Wang, Y. (2024). Machine learning-based antibiotic resistance prediction models: An updated systematic review and meta-analysis. *Technology and Health Care*, 32(5), 2865-2882.
17. Moat, J., Zovoilis, A., Steinkey, R., Zaheer, R., McAllister, T., & Laing, C. (2025). Machine learning methods to identify markers and predict antimicrobial resistance in *Escherichia coli*. *Canadian Journal of Microbiology*, 71, 1-15.
18. Noman, S. M., Zeeshan, M., Arshad, J., Deressa Amentie, M., Shafiq, M., Yuan, Y., ... & Jiao, X.

- (2023). Machine learning techniques for antimicrobial resistance prediction of *Pseudomonas aeruginosa* from whole genome sequence data. *Computational intelligence and neuroscience*, 2023(1), 5236168.
19. Orcales, F., Moctezuma Tan, L., Johnson-Hagler, M., Suntay, J. M., Ali, J., Recto, K., ... & Pennings, P. (2024). Using genomic data and machine learning to predict antibiotic resistance: A tutorial paper. *PLoS computational biology*, 20(12), e1012579.
 20. Petrin, S., Orsini, M., Massaro, A., Olsen, J. E., Barco, L., & Losasso, C. (2023). Phenotypic and genotypic antimicrobial resistance correlation and plasmid characterization in *Salmonella* spp. isolates from Italy reveal high heterogeneity among serovars. *Frontiers in public health*, 11, 1221351.
 21. Ren, Y., Chakraborty, T., Doijad, S., Falgenhauer, L., Falgenhauer, J., Goesmann, A., ... & Heider, D. (2022). Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. *Bioinformatics*, 38(2), 325-334.
 22. Siam, M. S., & Faiza, A. T. (2025). Antibiotic resistance tracking dataset (Version 2) [Data set]. Mendeley Data. <https://doi.org/10.17632/h4byb28gcv.2>
 23. Singh, R., Lin, Y. W., Zhao, J., & Li, J. (2026). Artificial intelligence and machine learning in antimicrobial discovery, resistance prediction, and precision therapy. *International Journal of Antimicrobial Agents*, 107715.
 24. Sunuwar, J., & Azad, R. K. (2021). A machine learning framework to predict antibiotic resistance traits and yet unknown genes underlying resistance to specific antibiotics in bacterial strains. *Briefings in Bioinformatics*, 22(6), bbab179.
 25. Talamantes-Becerra, B., Wickramarachchi, A., Bauer, D. C., Sng, L., Nguyen, A., Twine, N., & Leroux, H. (2024, September). Prediction of antibiotic susceptibility in *E. coli* Isolates using machine learning. In *Health. Innovation. Community: It Starts With Us: Papers from the 28th Australian Digital Health and Health Informatics Conference (HIC 2024)–Brisbane, Australia, 5–7 August 2024* (pp. 150-155). 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP: SAGE Publications Pvt. Ltd.
 26. Valavarasu, S., Sangu, Y., & Mahapatra, T. (2025). Prediction of antibiotic resistance from antibiotic susceptibility testing results from surveillance data using machine learning. *Scientific reports*, 15(1), 30509.
 27. Van Camp, P. J., Haslam, D. B., & Porollo, A. (2020). Prediction of antimicrobial resistance in gram-negative bacteria from whole-genome sequencing data. *Frontiers in microbiology*, 11, 1013.
 28. Wan, F., Tong, W., Wu, W., Liu, J., Huang, H., Zhang, P., & Jiang, W. (2026). Utilizing whole genome sequencing data for machine learning driven prediction of antibiotic resistance in *Escherichia coli*. *Frontiers in Microbiology*, 17, 1842717.

This article is licensed under a [Creative Commons Attribution 4.0 International \(CC BY NC ND\)](https://creativecommons.org/licenses/by-nc-nd/4.0/) License; copyright remains with the author(s).