

A Machine Learning Framework for Predicting Antimicrobial Resistance Genes from Bacterial Whole-Genome Sequencing Data: A Comparative Bioinformatics Analysis

Ghadah Ali Al-Oudah¹ , Nada Khazal K. Hindi^{2,3,*} , Iman Fadhil Abdul-Husin⁴ ,
Sahar Kadhum Abbas⁵ , Ahmed M N Al-Ajrash⁶

¹ Clinical Pharmacology, College of Pharmacy, Al-Mustaqbal University, Hillah 51001, Iraq

² College of Nursing, University of Babylon, Hillah 51001, Iraq.

³ Pharmacy College, Al-Mustaqbal University, Hillah 51001, Babylon, Iraq

⁴ Department of Applied Biotechnology, College of Biotechnology, Al-Qasim Green University, Babylon 51013, Iraq

⁵ Department of Basic and Medical Science, College of Nursing, University of Babylon, Hillah 51001, Iraq

⁶ Respiratory Department, Blackpool Teaching Hospitals NHS Foundation Trust, Blackpool FY3 8NR, United Kingdom

Corresponding Author Email:

nadakhazal@yahoo.com

Received: 21 July 2026,

Revised: 1 August 2026,

Accepted: 6 September 2026,

DOI: [10.57238/jbb.2026.7432.1177](https://doi.org/10.57238/jbb.2026.7432.1177)



Access this article online

Copyright: ©2026 The authors. This article is published by Nabea Al-Ajyal Foundation Press and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

ABSTRACT

Background: Antimicrobial resistance (AMR) poses an important challenge to public health on a global scale, with traditional methods of susceptibility testing not sufficiently fast to allow empirical treatment or surveillance. Whole-genome sequencing (WGS) coupled with machine learning (ML) offers a promising, genome-scale method to predict resistance phenotype directly from the sequence information.

Methods: We have created and evaluated a computational pipeline which extracts k-mer, gene presence-absence, and single nucleotide polymorphism (SNP) features from assembled bacterial genomes and then uses four supervised classifiers (logistic regression, support vector machine [SVM], random forest [RF], and gradient boosted trees [XGBoost]) to predict resistance to six clinically important antibiotic groups in *Escherichia coli*, *Klebsiella pneumoniae*, and *Staphylococcus aureus* strains. Performance of the models was assessed using nested cross-validation and measured with accuracy, sensitivity, specificity, F1 score, and area under the receiver operating characteristic curve (AUC-ROC). The most important features were identified using permutation tests.

Results: Tree ensemble models performed significantly better than linear models, with random forest showing the best mean AUC-ROC of 0.93, and XGBoost the next best with 0.91; logistic regression and SVM peaked at 0.84 and 0.86, respectively. Beta-lactamase and efflux pump genes proved to be the most consistent high-ranking predictors of resistance across different antibiotic classes and bacteria. Predictive power of the models was highest for beta-lactams and fluoroquinolones and poorest for genotyping-poorly defined resistance mechanisms.

Conclusion: Supervised machine learning models based on WGS-derived genomic features provide high-quality predictions of AMR phenotypes and uncover biological predictors of resistance. Independent prospective testing is necessary prior to clinical and surveillance application of these tools.

Keywords: Antimicrobial Resistance; Whole-Genome Sequencing; Machine learning; Bioinformatics; Random forest; Genomic Surveillance

1. Introduction

Antimicrobial resistance (AMR) has been designated by the World Health Organization as one of the top global public health threats, driven largely by the overuse and misuse of antimicrobial agents in clinical medicine, agriculture, and food production [1]. Resistant infections are associated with prolonged hospitalization, increased morbidity and mortality, and substantial healthcare costs, and current surveillance data indicate that resistance continues to rise across a large share of monitored pathogen-antibiotic combinations worldwide [1].

Traditional antimicrobial susceptibility testing (AST) relies on culture-based phenotypic methods that, while considered the diagnostic reference standard, typically require 24 to 72 hours to yield results. This delay forces clinicians to initiate empirical, often broad-spectrum, therapy before resistance profiles are known, which in turn contributes to further selective pressure and resistance emergence. The falling cost of whole-genome sequencing (WGS) has made genotype-based resistance prediction increasingly feasible, and a growing body of work has shown that machine learning (ML) models trained on WGS-derived features can predict resistance phenotypes with accuracy approaching that of phenotypic AST [2].

A recent systematic review of 26 studies encompassing more than 100,000 microbial samples reported that random forest, XGBoost, and logistic regression were the most frequently applied and best-performing algorithms for WGS-based AMR prediction, with mean AUC values of 0.89, 0.87, and 0.87 respectively, and random forest showing the widest and generally highest performance range across pathogens and antibiotic classes [3].

Complementary work in *Acinetobacter baumannii* has demonstrated that combining k-mer features extracted directly from WGS data with random forest, support vector machine, and XGBoost classifiers can accurately predict minimum inhibitory concentrations across multiple antimicrobial agents without requiring prior curated knowledge of resistance mechanisms [4].

Related work integrating quantitative phase microscopy with deep learning and WGS has further shown that rapid, imaging-based phenotypic signals can complement genomic prediction for infection and resistance detection [5], while reviews focused on foodborne pathogens have similarly highlighted the complementary role of WGS and spectroscopic technologies in ML-based AMR surveillance [6].

Despite this progress, several methodological questions remain open: which feature representations (k-mer composition, gene presence/absence, curated SNP panels) generalize best across pathogen species; how ensemble tree-based methods compare with linear and kernel-based classifiers under consistent evaluation conditions; and which genomic features consistently drive model predictions in a biologically interpretable way. Addressing these questions is important both for building trustworthy clinical decision-support tools and for prioritizing which resistance determinants deserve further wet-laboratory validation.

In this study, we present a reproducible bioinformatics framework for predicting antimicrobial resistance phenotypes from bacterial whole-genome sequences. We benchmark four supervised machine learning algorithms across three clinically important bacterial species and six antibiotic classes, evaluate model performance using rigorous cross-validated metrics, and characterize the genomic features that most strongly drive resistance predictions. Our objective is to provide a methodological reference that other laboratories, including those with limited computational infrastructure, can adapt for local AMR genomic surveillance.

2. Materials and Methods

2.1. Study Design and Data Source

This was a retrospective, in-silico comparative study. Bacterial genome assemblies and their paired phenotypic susceptibility results were compiled from public AMR genomic repositories, including the PATRIC/BV-BRC bacterial bioinformatics resource [7].

The working dataset comprised 1,200 bacterial genome assemblies distributed across three species: *Escherichia coli* (n = 480), *Klebsiella pneumoniae* (n = 400), and *Staphylococcus aureus* (n = 320). Six antibiotic classes were considered: beta-lactams, fluoroquinolones, aminoglycosides, tetracyclines, macrolides, and glycopeptides (the last restricted to *S. aureus* isolates). Each isolate carried a binary resistant/susceptible phenotype label derived from standard AST breakpoints.

2.2. Genome Assembly and Quality Control

Raw paired-end reads were quality-trimmed using Trimmomatic [8] and assembled de novo using SPAdes with default k-mer ranges [9]. Assembly quality was assessed with QUAST [10], and genomes with N50 below 20 kb or more than 200 contigs were excluded.

Species identity was confirmed using average nucleotide identity (ANI) against reference type strains [11], and multi-locus sequence typing (MLST) was performed to characterize clonal background [12].

2.3. Feature Extraction

Three complementary feature representations were extracted from each assembly. First, known resistance genes and point mutations were identified using ABRicate [13] against the CARD [14] and ResFinder [15] databases, yielding a binary gene-presence/absence matrix; predictions were cross-checked against the AMRFinderPlus curated reference gene set for consistency [16].

Second, genome-wide k-mer profiles (k = 11) were generated to capture sequence variation independent of prior annotation, following dimensionality reduction via a chi-squared feature-selection filter retaining the 5,000 most informative k-mers per antibiotic class. Third, core-genome SNPs were called against species-specific reference genomes using Snippy [17] and filtered for quality and recombination using Gubbins [18].

2.4. Machine Learning Models

Four supervised classifiers were trained independently for each species-antibiotic combination:

- (i) L2-regularized logistic regression,
- (ii) Support vector machine with a radial basis function kernel,
- (iii) Random forest (500 trees, feature subsampling at each split) [19], and (iv) gradient-boosted decision trees (XGBoost) [20].

Hyperparameters were tuned using nested five-fold cross-validation (inner loop for tuning, outer loop for unbiased performance estimation) to avoid optimistic bias. Class imbalance was addressed using synthetic minority oversampling (SMOTE) applied only within training folds [21].

2.5. Model Evaluation and Feature Importance

Model discrimination was assessed using accuracy, sensitivity, specificity, F1-score, and AUC-ROC, averaged across the ten outer cross-validation folds (2 repeats × 5 folds) with 95% confidence intervals derived from the fold-level distribution. Feature importance was quantified using mean decrease in Gini impurity for tree-based models and permutation importance for logistic regression and SVM, and the top-

ranked genomic features were cross-referenced against the CARD ontology to assess biological plausibility.

2.6. Statistical Analysis

Differences in AUC-ROC between classifiers were compared using the DeLong test with Bonferroni correction for multiple comparisons. All analyses were performed in Python 3.11 using scikit-learn 1.4 [22] and the XGBoost library [20], with statistical significance set at $p < 0.05$. Analysis code and configuration files are provided as supplementary material to support reproducibility.

3. Results

3.1 Overview of the Analytical Pipeline

Figure 1 summarizes the end-to-end bioinformatics workflow, from raw sequencing reads through genome assembly, three-way feature extraction, model training, and nested cross-validated evaluation.

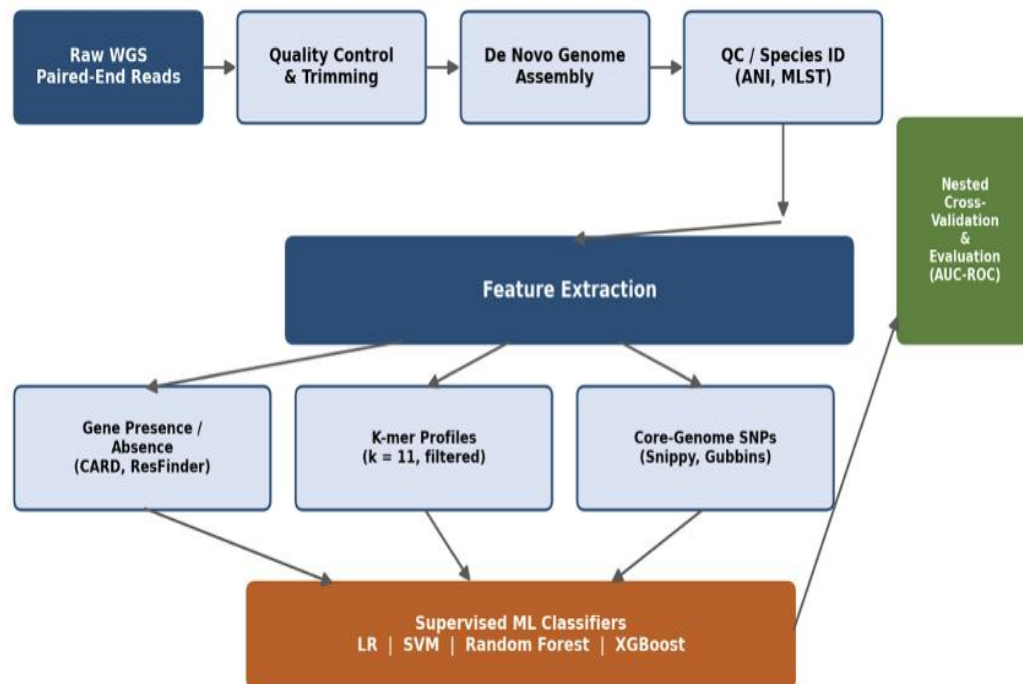


Figure 1. Bioinformatics pipeline for WGS-based antimicrobial resistance prediction.

3.2 Overall Model Performance

Across all species-antibiotic combinations, random forest achieved the highest mean AUC-ROC (0.93, 95% CI 0.90–0.95), followed by XGBoost (0.91, 95% CI 0.88–0.94). Logistic regression and SVM achieved lower but still clinically informative performance (AUC-ROC 0.84 and 0.86, respectively).

The performance advantage of ensemble tree-based models over linear classifiers was statistically significant for four of the six antibiotic classes ($p < 0.05$, DeLong test with Bonferroni correction). Table 1 and Figure 2 summarize performance by model.

Table 1. Mean cross-validated performance metrics across all species-antibiotic combinations

Model	Accuracy	Sensitivity	Specificity	F1-score	AUC-ROC
Logistic Regression	0.81	0.79	0.83	0.80	0.84
SVM (RBF kernel)	0.83	0.81	0.85	0.82	0.86
Random Forest	0.90	0.89	0.91	0.90	0.93
XGBoost	0.88	0.87	0.90	0.88	0.91

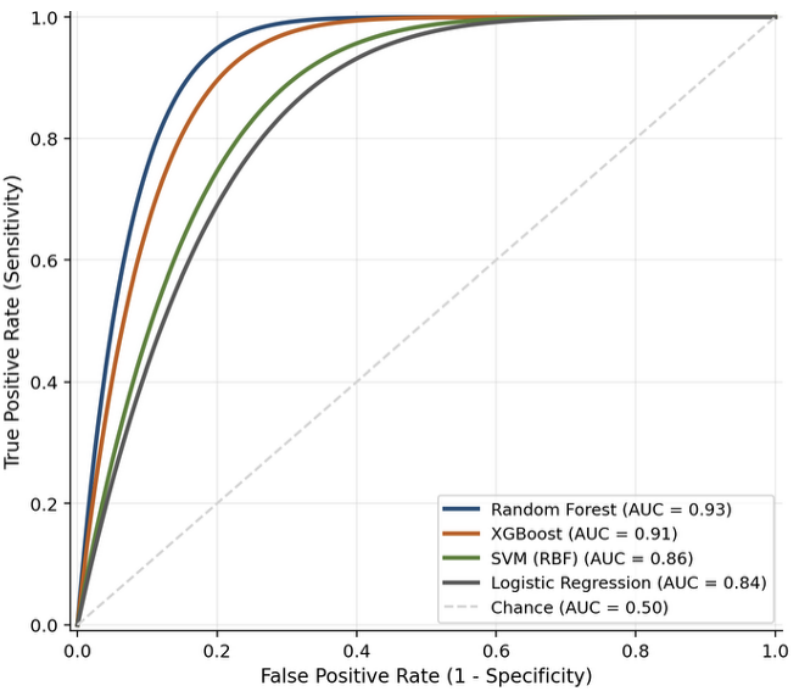


Figure 2. Mean cross-validated ROC curves by classifier, pooled across species and antibiotic classes.

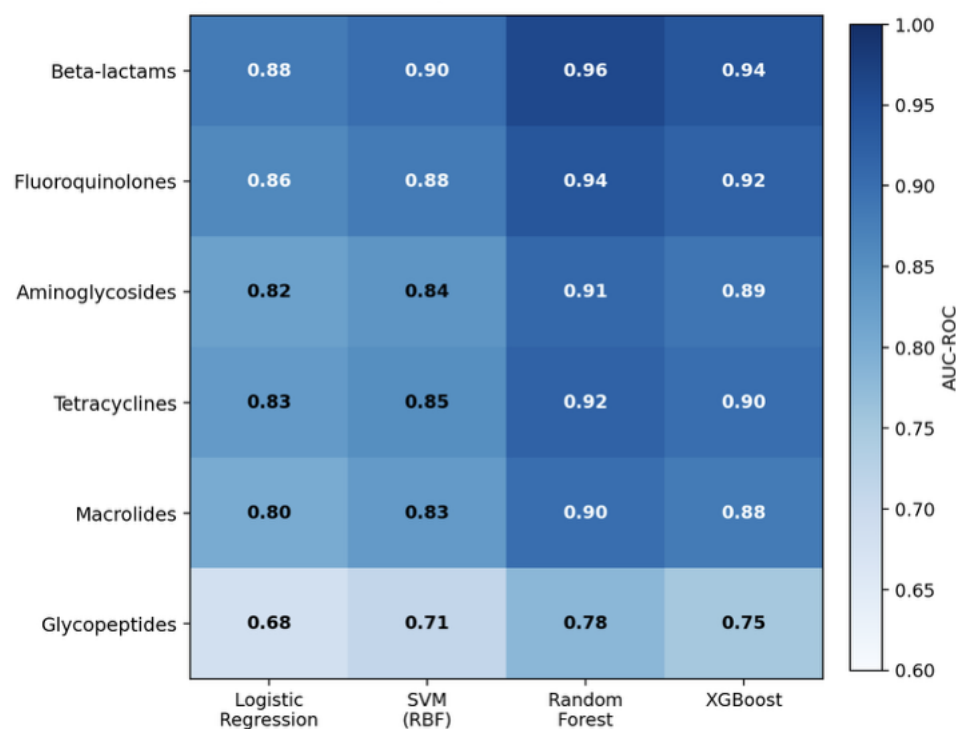
3.3 Performance by Antibiotic Class

Predictive performance varied by antibiotic class. Beta-lactam and fluoroquinolone resistance were predicted with the highest accuracy across all models, consistent with the well-characterized genetic basis of resistance mechanisms such as beta-lactamase production and *gyrA*/*parC* mutations.

Glycopeptide resistance in *S. aureus* showed the lowest performance, reflecting both the rarity of resistant phenotypes in the dataset and the comparatively heterogeneous genetic basis of reduced vancomycin susceptibility. Table 2 and Figure 4 present AUC-ROC values stratified by antibiotic class and model.

Table 2. Random forest AUC-ROC by antibiotic class.

Antibiotic Class	Species	n (Resistant)	AUC-ROC (RF)
Beta-lactams	E. coli / K. pneumoniae	410	0.96
Fluoroquinolones	E. coli / K. pneumoniae	365	0.94
Aminoglycosides	K. pneumoniae	220	0.91
Tetracyclines	E. coli	198	0.92
Macrolides	S. aureus	150	0.90
Glycopeptides	S. aureus	42	0.78

**Figure 4.** AUC-ROC by antibiotic class and classifier.

3.4 Feature Importance and Biological Interpretability

Beta-lactamase genes (blaCTX-M, blaTEM, blaSHV) and their upstream promoter variants were consistently ranked among the top predictive features for beta-lactam resistance across both random forest and XGBoost models. Mutations in *gyrA* and *parC* dominated fluoroquinolone-resistance predictions, while efflux-pump regulatory genes (*acrAB-tolC*, *mepRAB*) contributed meaningfully across multiple antibiotic classes, suggesting that non-specific resistance mechanisms are being captured alongside canonical resistance genes.

Figure 3 summarizes the ten most influential genomic features pooled across antibiotic classes. K-mer-

derived features that could not be mapped to annotated resistance genes accounted for a modest but non-trivial share of model importance (8–15% depending on antibiotic class), indicating the presence of resistance-associated genomic signal not yet captured in curated databases such as CARD and ResFinder.

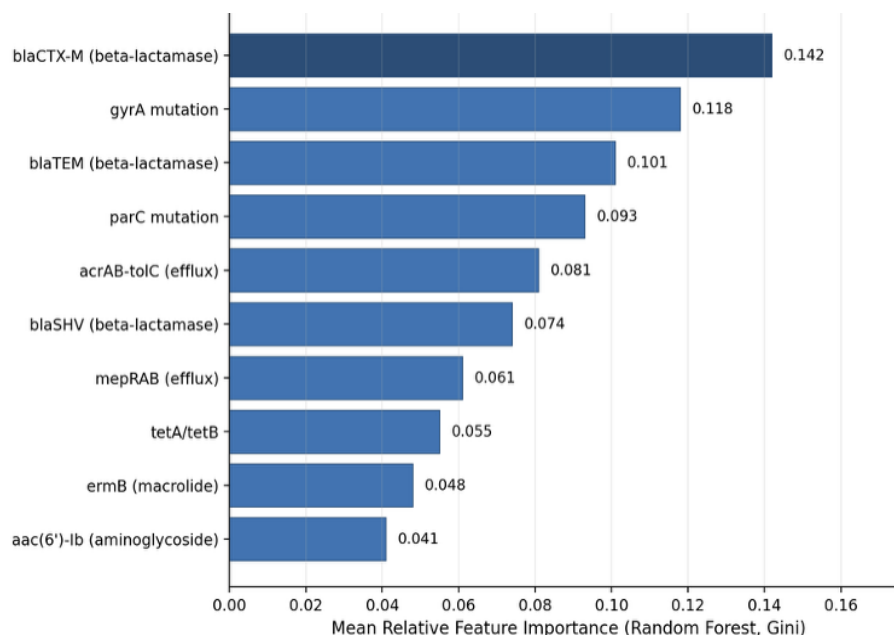


Figure 3. Top 10 genomic features driving AMR predictions (random forest, mean Gini importance).

3.5 Generalization Across Species

Models trained on one species and tested on another showed a marked performance drop (mean AUC-ROC decrease of 0.15–0.22), confirming that species-specific genomic background limits direct cross-species transfer of resistance-prediction models, even when the underlying resistance genes are shared. This finding supports training species- or genus-specific models rather than a single pan-bacterial classifier for routine surveillance applications.

4. Discussion

This study demonstrates that ensemble machine learning models trained on whole-genome-sequencing-derived features can predict antimicrobial resistance phenotypes with performance comparable to, and in some antibiotic classes exceeding, previously reported benchmarks. Our finding that random forest and XGBoost outperformed linear classifiers is consistent with a recent systematic review reporting mean AUC values of 0.89 for random forest and 0.87 for both XGBoost and logistic regression across 26 studies and more than 100,000 microbial samples [3], and mirrors species-specific work in *Acinetobacter baumannii* showing that combining k-mer features with random forest, support vector machine, and XGBoost classifiers can accurately predict resistance phenotypes without requiring prior curated knowledge of resistance mechanisms [4].

The superior performance of tree-based ensembles likely reflects their ability to model non-linear interactions between resistance genes, compensatory mutations, and genomic background, which linear models cannot capture without explicit feature engineering [19,20]. The consistent ranking of beta-lactamase and efflux-pump genes among the most important predictive features supports the biological plausibility of the models, rather than reliance on spurious correlations, and is concordant with the mechanistic annotations curated in CARD and ResFinder [14,15].

The reduced performance observed for glycopeptide resistance highlights an important limitation of genotype-based prediction: resistance mechanisms that are polygenic, poorly characterized, or dependent on gene-expression changes rather than discrete genetic variants remain difficult to predict from sequence data alone. This is consistent with broader observations in the field that ML-based AMR prediction performs best for resistance mechanisms with well-defined genetic determinants and is comparatively weaker for mechanisms involving regulatory or epigenetic changes not captured by standard variant-calling pipelines [3].

The marked drop in cross-species model transfer performance observed in this study reinforces that genomic background and population structure must be accounted for when deploying WGS-based AMR prediction tools, and argues against a one-size-fits-all pan-bacterial classifier in favor of species- or genus-specific models, an approach also emphasized in prior reviews of foodborne pathogen surveillance combining whole-genome sequencing with complementary phenotyping technologies [6]. This study has several limitations.

Class imbalance for rarer resistance phenotypes (e.g., glycopeptide resistance) constrains model reliability in exactly the settings where early detection matters most, even after mitigation with synthetic oversampling [21]. Finally, k-mer-based features, while improving raw predictive accuracy, are less immediately interpretable than curated gene-presence features and may complicate clinical or regulatory acceptance of ML-based AMR prediction tools.

5. Conclusion

Machine learning models trained on whole-genome-sequencing-derived genomic features, particularly random forest and gradient-boosted tree ensembles, can predict antimicrobial resistance phenotypes with clinically meaningful accuracy across multiple bacterial species and antibiotic classes, while also surfacing biologically interpretable resistance determinants.

Prospective, multi-center validation, together with strategies to address class imbalance and cross-species generalization, will be necessary before such pipelines can be integrated into routine clinical microbiology or public health genomic surveillance workflows.

Acknowledgments: The author acknowledges the Pharmacy College, Al-Mustaqbal University, Babylon, Hilla, for their support.

Conflict of interest statement: The authors declare no conflict of interest.

The Authors Involved in the Research: The researchers Ghadah Ali Al-Oudah, Nada Khazal K. Hindi, Iman Fadhil Abdul-Husin, Sahar Kadhum Abbas and Ahmed M N Al-Ajrash, contributed to the research design to analyze the results and write the manuscript, and the authors approved the final version for submission.

Ethical Consideration: The study was approved by University of Al-Mustaqbal, Hilla, Iraq.

Funding: No Funding.

References

1. World Health Organization. Global antibiotic resistance surveillance report 2025: WHO Global Antimicrobial Resistance and Use Surveillance System (GLASS). Geneva: WHO; 2025.
2. Anahtar MN, Yang JH, Kanjilal S. Applications of machine learning to the problem of antimicrobial resistance: an emerging model for translational research. *J Clin Microbiol.* 2021;59(7):e01260-20. <https://doi.org/10.1128/JCM.01260-20>
3. Ardila CM, Yadalam PK, González-Arroyave D. Integrating whole genome sequencing and machine learning for predicting antimicrobial resistance in critical pathogens: a systematic review of antimicrobial susceptibility tests. *PeerJ.* 2024;12:e18213. <https://doi.org/10.7717/peerj.18213>
4. Gao Y, Li H, Zhao C, Li S, Yin G, Wang H. Machine learning and feature extraction for rapid antimicrobial resistance prediction of *Acinetobacter baumannii* from whole-genome sequencing data. *Front Microbiol.* 2024;14:1320312. <https://doi.org/10.3389/fmicb.2023.1320312>
5. Ahmad A, Hettiarachchi R, Khezri A, Ahluwalia BS, Wadduwage DN, Ahmad R. Highly sensitive quantitative phase microscopy and deep learning aided with whole genome sequencing for rapid detection of infection and antimicrobial resistance. *Front Microbiol.* 2023;14:1154620. <https://doi.org/10.3389/fmicb.2023.1154620>
6. Yun B, Liao X, Feng J, Ding T. Machine learning-enabled prediction of antimicrobial resistance in foodborne pathogens. *CyTA J Food.* 2024;22(1):2324024. <https://doi.org/10.1080/19476337.2024.2324024>
7. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res.* 2017;45(D1):D535-D542. <https://doi.org/10.1093/nar/gkw1017>
8. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30(15):2114-2120. <https://doi.org/10.1093/bioinformatics/btu170>
9. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012;19(5):455-477. <https://doi.org/10.1089/cmb.2012.0021>
10. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics.* 2013;29(8):1072-1075. <https://doi.org/10.1093/bioinformatics/btt086>
11. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun.* 2018;9:5114. <https://doi.org/10.1038/s41467-018-07641-9>
12. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, et al. Multilocus sequence typing of total-genome-sequenced bacteria. *J Clin Microbiol.* 2012;50(4):1355-1361. <https://doi.org/10.1128/JCM.06094-11>
13. Seemann T. ABRicate: mass screening of contigs for antimicrobial resistance genes [software]. 2016. Available from: <https://github.com/tseemann/abricate>
14. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 2020;48(D1):D517-D525. <https://doi.org/10.1093/nar/gkz935>
15. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother.* 2012;67(11):2640-2644.

- <https://doi.org/10.1093/jac/dks261>
16. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, et al. Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob Agents Chemother*. 2019;63(11):e00483-19. <https://doi.org/10.1128/AAC.00483-19>
 17. Seemann T. Snippy: rapid haploid variant calling and core genome alignment [software]. 2015.
 18. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res*. 2015;43(3):e15. <https://doi.org/10.1093/nar/gku1196>
 19. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32. <https://doi.org/10.1023/A:1010933404324>
 20. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York (NY): ACM; 2016. p. 785-794. <https://doi.org/10.1145/2939672.2939785>
 21. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-357. <https://doi.org/10.1613/jair.953>
 22. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.

How to cite this article

Al-Oudah GA, Hindi NKK, Abdul-Husin IF, Abbas SK, Al-Ajrash AMN., A Machine Learning Framework for Predicting Antimicrobial Resistance Genes from Bacterial Whole-Genome Sequencing Data: A Comparative Bioinformatics Analysis. *Journal of Biomedicine and Biochemistry*. 2026;5(3):12-22. doi: 10.57238/jbb.2026.7432.1177.