Research Article

Bioinformatics and Biology Insights
Volume 20: 1–8

DOI: 10.1177/11779322261453756
journals.sagepub.com/home/bbi
Sage

A Novel Bioinformatics Pipeline and a Machine-Learning Approach for Antimicrobial Resistance Phenotypic Prediction

Owen Visser[1], Victor Agboli[1] and Somnath Datta[1]

Abstract

Overuse of antimicrobial drugs is known to cause an increase in bacterial resistance among surviving pathogens, reducing the effectiveness of future treatments. Publicly available sequencing collections, such as the National Center for Biotechnology Information Sequence Read Archive (SRA), allow for global investigation of antimicrobial resistance across pathogens. In this study, we developed a pipeline to process 10 803 globally sourced unique bacterial isolates from publicly available SRA datasets (9 pathogens, 4 antibiotics), representing sequencing data generated by multiple laboratories using diverse sequencing platforms and protocols. The pipeline extracted SRA metadata to determine read layout and length, applied quality control, trimming, and decontamination. Preprocessed isolates were mapped reads to an antimicrobial resistance gene class and to a strain-level genome reference library constructed from complete genomes on the SRA submitted between 2009 and 2020. Three classifiers—L1-penalized logistic regression, random forest, and extreme gradient boosting—were trained on the resulting feature matrices, and their outputs were combined in a majority-vote ensemble. Internal training resulted in 83.8% balanced accuracy on average, and external testing yielded 80.2%. Variable importance analyses identified known resistance gene classes and strain markers such as *Acinetobacter baumannii* LAC-4 and *Campylobacter jejuni* 81-176 and NCTC13255, confirming biological relevance. This work demonstrates a scalable approach for antimicrobial resistance prediction using heterogeneous sequencing data.

Keywords

antimicrobial resistance prediction, phenotypic prediction, bioinformatics pipeline

Received: 21 October 2025; accepted: 10 May 2026

Introduction

Antimicrobial resistance (AMR) is a persistent difficulty in modern medicine. Due to global crisis like the COVID-19 pandemic, the indiscriminate use of antibiotics intensified and placed pressure on microbial populations.[1] Notable resistance has emerged across geographical regions and in hospitals.[2,3] It is known that bacterial populations evolve rapidly under selective pressure, the genomic context associated with AMR can change over relatively short time scales.[4] Consequently, predictive models are most reliable when training and evaluation samples are drawn from similar temporal periods, reducing potential bias introduced by evolutionary shifts in the underlying populations.[5] Moreover, the discovery of new resistant strains has been seen in the ecology of individual wounds.[6] These observations have created urgency for the development of methods able to detect and predict resistance to specific antibiotics as to prevent further AMR adaptation.

Growing attention to this matter prompted the Annual International Conference on Critical Assessment of Massive Data Analysis (CAMDA), organized under the International Society for Computational Biology (ISCB), to present AMR prediction as its 2025 challenge. The task was to predict the resistance phenotypes of 9 bacterial pathogens against 4 antibiotics drawn from the World Health Organization (WHO) Priority Pathogen list[7] using globally sourced raw sequencing reads. Two of the challenges addressed emulate conditions encountered in applied work—the heterogeneity of bacterial data and the difficulty of prediction due to the former. Inconsistency imposes a substantial burden on pipeline design, requiring standardized tools and adaptive design to ensure comparable feature matrices across thousands of samples. In this setting, developing models directly from raw sequencing reads,

[1]Department of Biostatistics, University of Florida, Gainesville, USA

Corresponding author:
Somnath Datta, Department of Biostatistics, University of Florida, Gainesville, FL 32603, USA.
Email: somnath.datta@ufl.edu

rather than relying on assembled genomes, avoids variability introduced by differing genome assembly pipelines and enables a consistent feature-construction strategy across heterogeneous sequencing datasets. Ultimately, the challenge reflects the heterogeneous conditions encountered in applied AMR surveillance.

Bioinformatics pipelines that handle bacterial data for the purposes of AMR prediction can accommodate a wide array of sequencing technologies and experimental designs.[8] Public repositories, including the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA), contain samples from multiple sequencing platforms, differing by read length, error profile, and throughput.[9] Each configuration demands careful preprocessing to remove adapters, trim low-quality bases, and filter contaminants before any downstream analysis. Tools like Trimmomatic and Kneaddata are routinely employed for quality trimming and host-sequence removal.[10,11] Contemporary AMR prediction pipelines often incorporate reference-based annotation to derive biologically meaningful features for modeling.[12] Dedicated AMR gene databases, notably ResFinder, are employed to detect and quantify resistance gene classes, while taxonomic genome collections, such as RefSeq or curated species-specific assemblies, enable strain-level identification and abundance estimation.[13,14]

Alignment tools like Bowtie2 or K-mer Alignment (KMA) are used for alignment against these reference libraries to generate gene or genome-level abundance tables suitable for modeling and serve as the principal input variables for downstream predictive analyses.[8,10,15] These components form most contemporary AMR workflows, yet their suitability is data-dependent. Consequently, AMR prediction workflows often require substantial manual adaptation for each dataset, and pipelines developed under 1 sequencing configuration may not transfer reliably to others. Developing a unified preprocessing and feature-construction pipeline that can accommodate heterogeneous sequencing inputs therefore remains an important challenge in AMR prediction.

Current methodology for AMR phenotypic prediction often relies on advanced statistical frameworks that accommodate high-dimensional genomic features and complex dependence structures. With well-curated data, methods such as random forests (RFs), support vector machines, penalized regression, and convolution neural networks were shown to predict phenotypes with high accuracy.[16-19] However, these approaches rely on large sets of standardized genomic features and often assume inputs derived from assembled genomes or curated feature matrices, conditions that are rarely satisfied when working with raw sequencing data from heterogeneous public repositories. In recent years, Extreme Gradient Boosting (XGBoost) has been seen to perform well in the prediction of phenotype by efficiently capturing complex non-linear relationships between genomic features and AMR outcomes.[20] This has also been demonstrated in pathogen-specific settings; for example, an XGBoost-based in silico minimum inhibitory concentration panel for *Klebsiella pneumoniae* achieved high predictive accuracy across 20 antibiotics, further supporting the utility of machine-learning approaches for genomic AMR prediction.[21] Ensembles that combine multiple prediction models have also been applied to AMR data to improve stability and accuracy.[22] By aggregating the outputs of several classifiers through a simple rule such as majority voting, an ensemble can reduce the influence of model-specific biases and capture complementary patterns that may be missed by any single method. This approach often yields predictions that are more robust to noise and data imbalance than those produced by individual models, while still allowing the importance of underlying genomic features to be assessed through standard permutation or perturbation analyses.

Despite progress in AMR phenotype prediction, few studies have developed end-to-end pipelines capable of processing heterogeneous raw sequencing reads and producing standardized genomic features suitable for predictive modeling across multiple pathogens. In this study, we address 2 challenges in AMR prediction using isolate data composed of 9 bacteria paired with 4 antibiotics. First, we constructed a bioinformatics pipeline on the University of Florida's HiperGator supercomputer to process raw sequencing reads of differing format and quality.[23] Furthermore, 2 reference resources were prepared: a strain-resolved genome library constructed from NCBI RefSeq assemblies of the 9 pathogens collected between 2009 and 2020 and a curated AMR gene library compiled from ResFinder.[9,13,14] Reads were aligned with KMA to the strain-resolved genome library and to the curated AMR gene library, yielding parallel matrices of strain-level abundances and resistance gene-class counts.[13] Second, we applied an ensemble ofRF, penalized logistic regression, and XGBoost to the matrices and trained and evaluated to predict resistance phenotypes for each bacteria-antibiotic combination.[24-26] We constructed a pipeline capable of handling heterogeneous inputs and using broad genomic references and demonstrated that combined statistical models can provide dependable AMR phenotype predictions despite variation in the data.

Materials and Methods

Materials

The data were sourced from the 2025 CAMDA challenge and consisted of 9 bacterial pathogens (*Acinetobacter baumannii*, *Campylobacter jejuni*, *Escherichia coli*, *K pneumoniae*, *Neisseria gonorrhoeae*, *Pseudomonas aeruginosa*, *Salmonella enterica*, *Staphylococcus aureus*, and *Streptococcus pneumoniae*) paired with 4 antibiotics—Gentamicin (GEN), Erythromycin (ERY), Tetracycline (TET), and Ceftazidime (CAZ)—drawn from the WHO Priority Pathogen List. The raw training data had 6144 samples, which held 5458 unique isolate identifiers. Each record includes the bacterial genus and species, an NCBI SRA accession identifying the raw sequencing reads, the tested antibiotic, and a phenotypic label of Susceptible, Intermediate, or Resistant assigned according to the latest Clinical and Laboratory Standards Institute (CLSI) guidelines. The recorded collection dates for the training isolates ranged from 1999 to 2020, with the majority of samples concentrated within the modern sequencing era.

The training data also included minimum inhibitory concentration (MIC) values with accompanying measurement signs (=, >, <, or ⩽), laboratory typing methods, and platform details are provided when available, although MIC measurement signs are absent for *S aureus* and *A baumannii*. Because these MIC measurement details were incomplete for several species, MIC values were not incorporated into the predictive modeling to maintain consistency across all data. Sequencing data showed considerable variation. Read files included both single-end and paired-end layouts, spanned a wide range of read depths, and originated from multiple sequencing platforms. Some accessions represented technical repeats of the same biological sample, adding further heterogeneity to the collection. File sizes varied greatly, and the levels of quality and contamination differed across submissions. Additional sparse metadata was also included, containing publication identifiers, isolation source, country of origin, and collection date. These accompanying metadata were extremely sparse and often missing and ultimately were not incorporated into the predictive analyses.

The external data consisted of 5345 unique isolates including only the bacterial genus, species, and an NCBI SRA accession code. Similarly, the testing isolates spanned collection years from 1999 to 2020, providing temporal overlap with the training data while extending modestly beyond the primary training window.

Although prior work has shown that geographic origin and isolation source can shape bacterial evolution and resistance patterns, the information available in these data was too sparse and inconsistently reported to support statistical modeling.[2,3,6,27] As a result, the predictive framework was restricted to features derived directly from the sequencing reads.

Bioinformatics Pipeline

The bacterial isolates were sourced from the NCBI SRA using the datasets command-line interface.[28] Because the sequencing data originated from diverse laboratories and platforms, a structured sequence of preprocessing steps was implemented to standardize inputs before downstream analysis. All computations were performed on the University of Florida HiPerGator cluster.[23]

Training and testing accession codes were first extracted from the metadata and separated by bacteria-antibiotic combinations to generate accession lists for batch processing. Two reference resources were then prepared. The first was a strain-resolved genome library created by downloading and indexing all annotated RefSeq genomes for the 9 target species meeting our inclusion criteria of being a complete genome and being released between January 1, 1990, and December 31, 2019. The lower bound in the query was specified as a permissive catch-all to avoid inadvertently excluding eligible records; however, after filtering, no complete RefSeq assemblies prior to 2009 were included in the final reference set. A complete list of all assembly reports can be found within the codebase (see the Supplementary Materials section). The second was an AMR gene-class library from the ResFinder repository and indexed for KMA alignment. These 2 libraries comprised, respectively, strain-level genome references and AMR gene sequences grouped by antibiotic class.

To ensure taxonomic consistency, RefSeq genomes were downloaded by taxon using the NCBI Datasets command-line tools (CLI) with atypical assemblies excluded and subsequently indexed for KMA alignment. Cleaned reads were aligned to the corresponding species-specific RefSeq indices as part of preprocessing, providing a computational check that sequencing data were consistent with the assigned species labels.

Raw FASTQ files were retrieved using fasterq-dump in parallel batches. Sequencing layout (single-end vs paired-end) was determined directly from the number of FASTQ files generated per accession, ensuring correct downstream handling without reliance on external metadata.

Downloaded reads were evaluated with FastQC to generate per-sample quality metrics, from which empirical average read length was computed as total bases divided by total reads; this measure guided downstream filtering.

Raw data were then processed using a standardized pipeline consisting of Trimmomatic, kneaddata, and Bowtie2.[10,11,29] Low-quality bases were trimmed using Trimmomatic with a 4-base sliding window and a minimum average Phred quality score of 20. Reads shorter than 60% of the observed average read length (minimum 25 bases) were removed. Following this, paired reads were repaired using the repair.sh utility from bbmap, which restores proper read order and diverts singletons. Quality control was fully automated; samples were not manually inspected or excluded solely on the basis of FastQC module flags, as the objective was to preserve real-world heterogeneity while mitigating low-quality signal through preprocessing and reference alignment checks.

The reads were aligned with KMA against both reference libraries to obtain genome-level counts (strain identification) and AMR gene-class counts (functional profiling).[13,15] The per-sample .mapstat outputs were aggregated into 2 count matrices: 1 for strain-level genomes and 1 for AMR gene classes. Both matrices were filtered to remove low-information features: strain-level genomes were retained only if present in at least 1% of samples with a minimum of 10 mapped reads, and AMR gene classes were similarly filtered to eliminate rare or spurious hits. Samples with no detected strain-level genomes or AMR gene classes were excluded from model training.

Filtered matrices were normalized to relative abundances using the metagenomeSeq package in R, which applies cumulative sum scaling (CSS) to correct for differences in library size and sequencing depth.[30] Normalized values were log-transformed to stabilize variance and expressed per sample. Finally, the normalized strain-level genome and AMR matrices were concatenated by sample to form the combined feature set used in all downstream predictive modeling.

Machine-Learning Models

The combined strain-AMR feature matrix was used as the predictor set for all models, where each row represented a

bacterial isolate and each column represented the normalized relative abundance of a strain-specific marker or an AMR gene class. For each bacteria-antibiotic pair, the binary response variable indicated resistant or susceptible status.

Samples labeled as "Intermediate" were removed from training as intermediate was not an accepted response for testing submission to the CAMDA challenge. Because each modeling method incorporated its own regularization or variable-selection mechanism, no additional dimensionality reduction was applied before fitting.

Random Forest. An RF classifier was trained with the ranger implementation through the caret interface in R.[24,31] Hyperparameters were tuned with 5-fold cross-validation over a grid of candidate values for the number of variables randomly sampled at each split (mtry) centered on $\sqrt{p}$ and the minimum node size ranging from 1 to 5. Each model used 500 trees and the Gini impurity criterion. The final parameter set was chosen to maximize cross-validated accuracy.

L1-Penalized Logistic Regression. An L1-penalized logistic regression (LLR) was fitted using the glmnet package in R.[25] The regularization parameter lambda was selected by 10-fold cross-validation to minimize binomial deviance. This procedure yields sparse coefficient vectors, allowing the model to down-weight noninformative genomic features.

Extreme Gradient Boosting. A gradient boosting model (XGB) was trained using the XGBoost implementation accessed through caret.[26,31] A 5-fold cross-validation scheme optimized the area under the receiver operating characteristic (ROC) curve over a grid including 300 or 500 boosting rounds, tree depths of 3 to 5, and learning rates between 0.03 and 0.07. Other parameters (gamma, colsample_bytree, min_child_weight, subsample) were set to their package defaults.

Majority-Vote Ensemble. Predictions from the 3 base models were combined by majority voting to produce the final ensemble classification. For each sample, if at least 2 of the 3 models predicted resistance, the ensemble output was labeled as "Resistant"; otherwise, it was labeled as "Susceptible."

Feature Importance

To assess the contribution of individual predictors, a leave-one-feature-out (LOFO) procedure was applied in which each feature was removed across samples and the resulting difference in ensemble accuracy was recorded. The change in accuracy for a removed feature served as an importance score to determine the strain and AMR gene-class features most influential to the ensemble decision process.

Model Evaluation

Models were evaluated using repeated train-test splits on the training data. For each bacterial species, the available isolates were randomly partitioned into a 70:30 training-testing split and the models were refit on the training subset. This procedure was repeated for 100 iterations, with fixed seed values ranging from set.seed(1) to set.seed(100), to account for variability arising from the random partitioning and to ensure reproducibility.

Model performance was assessed using balanced accuracy (%), defined as the mean of the true positive rate (TPR) and true negative rate (TNR). For each iteration, class probabilities were converted to predicted labels using a fixed probability threshold of 0.5, and balanced accuracy was computed on the held-out test subset. This produced a distribution of balanced accuracy values across iterations for each bacterial species and model.

Receiver operating characteristic curves were also computed using the predicted probabilities from each model. For every iteration, the TPR and false positive rate were evaluated across a grid of probability thresholds. These ROC curves were then averaged across iterations and across all bacterial species to produce a single-aggregated ROC curve for each modeling approach. This allowed comparison of the overall discriminative performance of the models independent of any specific classification threshold.

Balanced accuracy was selected as the primary threshold-based metric to align with the CAMDA evaluation protocol, allowing direct comparison between the internal cross-validation results and the external challenge evaluation.

Results

Model Evaluation on Training Data

Of the 5458 unique training isolates, 607 intermediate-labeled isolates were removed. In addition, 319 resistant and 115 susceptible isolates lacked strain-level genome hits and were therefore excluded. This resulted in 4417 isolates—2066 resistant and 2351 susceptible—used for model training. The sample sizes of the train-split and test-split can be seen in Table 1. Combinations of bacteria and antibiotics will be referred to solely by the bacteria for brevity.

The balanced accuracy of each model (LLR, RF, XGB) and the majority-vote ensemble (ENS) are reported in Table 2. Performance varied notably across bacteria.

Pseudomonas aeruginosa showed the lowest average balanced accuracy across all models, primarily brought down by the lowest accuracy of the LLR model. The second lowest average balanced accuracy across models was *E coli*, in which the RF had its worst accuracy. The best accuracy (96.6%) was for the prediction of *C jejuni*, where the RF model was the highest, followed in order by ENS, XGB, and LLR, which all exceeded 93%. Other notable bacteria predictions were of *A baumannii* and *S aureus* where the RF, XGB, and ENS models exceeded 90% accuracy.

Random forest achieved the highest single-model accuracy for 4 of the 9 bacteria, while ENS came close with 3 of the highest, and the LLR model did not have the highest score for any bacteria. Notably, the RF model performed worst for 1 bacterium, and the LLR performed worst for the remaining 8. For *C jejuni*, *S enterica*, and *S aureus*, the LLR performed comparably but lagged for more complex resistance patterns such as those in *A baumannii*

Table 1. Counts of Resistant and Susceptible Isolates in the Training Data Separated by the 9 Bacterial Pathogens and 4 Antibiotics (Drugs).

Bacteria	Drug	Resistant	Susceptible
A baumannii	CAZ	234	133
C jejuni	TET	315	201
E coli	GEN	135	327
K pneumoniae	GEN	312	321
N gonorrhoeae	TET	328	267
P aeruginosa	CAZ	205	211
S enterica	GEN	277	310
S aureus	ERY	134	291
S pneumoniae	ERY	126	290
Total		2066	2351

Table 2. Balanced Accuracy (%) of 4 Machine-Learning Models Across 9 Bacterial Pathogens and 4 Antibiotics (Drugs).

Bacteria	Drug	LLR	RF	XGB	ENS
A baumannii	CAZ	(86.9)	**90.6**	90.4	90.2
C jejuni	TET	(93.8)	**96.6**	95.7	96.3
E coli	GEN	69.2	(65.8)	**72.1**	68.7
K pneumoniae	GEN	(77.7)	78.0	**80.7**	80.1
N gonorrhoeae	TET	(75.7)	**83.9**	81.5	82.5
P aeruginosa	CAZ	(66.8)	68.2	67.6	**68.4**
S enterica	GEN	(87.3)	89.2	88.3	**89.3**
S aureus	ERY	(92.9)	94.3	93.3	**94.8**
S pneumoniae	ERY	(75.0)	**85.0**	84.8	84.1
Average		(80.6)	83.5	83.8	**83.8**

The smallest value per pair is surrounded by parentheses, while the largest is bolded.

and *N gonorrhoeae*. The ENS achieved the highest overall performance for *P aeruginosa*, *S enterica*, and *S aureus*. While ENS did not perform as well as RF in an individual comparison, across bacteria, ENS had the highest accuracy on average and never performed worse than the weakest model, demonstrating the stability gained by aggregating classifier predictions.

The ROC analysis, shown in Figure 1, illustrates clearly that the ENS, RF, and XGB outperform the LLR model. As well, the ENS model curve is always outside or in line with the RF and XGB curves, demonstrating its strength.

Variable Importance on Training Data

A LOFO importance analysis was applied on ENS to identify features most predictive of resistance. Table 3 lists the top 5 variable features per species of both strain-specific markers and AMR gene classes. While all strains and gene classes have biological relevance, only a subset of representative examples is highlighted for brevity.

For *A baumannii*, the presence of the aminoglycoside, macrolide, beta.lactam, and sulfonamide resistance genes along with the LAC-4 strain were among the most important predictors in determining AMR. The strain LAC-4 is highly resistant and spreads resistant genes associated with the 3 previously mentioned AMR gene classes.[32]

In *C jejuni*, tetracycline gene-class resistance is well documented, and the NCTC13255 and 81-176 strains have been shown to mediate transfer of TET resistance.[33]

Across bacterial species, predictors of AMR included both established resistance gene classes and specific strain variants. In *A baumannii*, sulfonamide and aminoglycoside resistance genes align with its well-documented multi-drug resistance profile.[34] In *S aureus*, macrolide resistance genes were consistent with widespread resistance mediated by both plasmid-encoded and chromosomal mechanisms.[35]

Model Performance on External Data

Table 4 reports the balanced accuracy of ENS for each bacteria-antibiotic pair on both the internal 70:30 split and the external data.

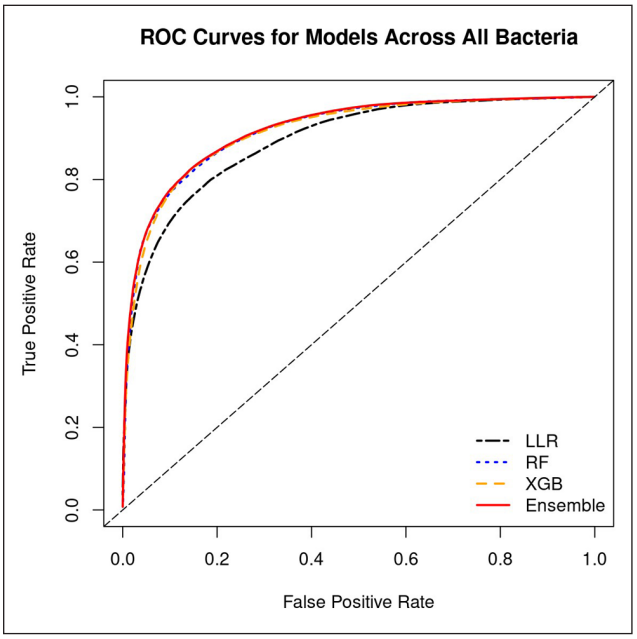

Figure 1. ROC curves comparing predictive performance of the penalized logistic regression (LLR), random forest (RF), extreme gradient boosting (XGB), and ensemble majority-vote models across all bacterial datasets.
Curves represent the average true positive rate vs false positive rate computed over repeated cross-validation iterations and aggregated across species. The diagonal line indicates the performance expected under random classification.

Relative to the internal split, the external data showed substantial decreases for several pathogens, with the largest drops for *E coli* (68.7% vs 49.7%), *N gonorrhoeae* (82.5% vs 69.2%), and *C jejuni* (96.3% vs 81.2%/83.3%). Moderate decreases were observed for *A baumannii* (90.2% vs 88.0%), while higher external scores were seen for *K pneumoniae* (80.1% vs 81.9%), *P aeruginosa* (68.4% vs 75.7%), *S enterica* (89.3% vs 92.4%), *S aureus* (94.8% vs 95.3%), and *S pneumoniae* (84.1% vs 88.5%). The mean balanced accuracy on the external data was 80.4%, indicating that ENS retained strong predictive ability across a diverse set of globally sourced isolates.

Table 3. Top 5 Variable Features Ranked by Model Importance for Prediction of Resistance in Each Bacteria-Antibiotic (Drug) Pair.

Bacteria	Drug	Top 5 variable strains/AMR classes
A baumannii	CAZ	sulfonamide, (LAC-4), aminoglycoside, macrolide, beta.lactam
C jejuni	TET	tetracycline, (81-176), (NCTC 13255), beta.lactam, (M129)
E coli	GEN	beta.lactam, phenicol, tetracycline, aminoglycoside, colistin
K pneumoniae	GEN	aminoglycoside, (AUSMDU00008119), (SCKP020049),(015625), (SCKP020046)
N gonorrhoeae	TET	(FQ02-chrom), (TFG-A2), (34769), (FQ02-plasmid), (WHO U)
P aeruginosa	CAZ	quinolone, tetracycline, (124-4(59)), (AR442), (IMP68)
S enterica	GEN	rifampicin, (SL254), (CFSAN001921), (0307213), (CFSA231)
S aureus	ERY	macrolide, beta.lactam, fusidic acid, quinolone, trimethoprim
S pneumoniae	ERY	(GPSC13), macrolide, (JJA), (KK0981), (Xen35)

Strain identifiers are shown in parentheses; "-chrom" and "-plasmid" indicate chromosomal or plasmid location.

Table 4. Balanced Accuracy (%) of ENS on the Internal 70:30 Train-Test Split and the Results Returned From the External Data.

Bacteria	Drug	Train split	External test
A baumannii	CAZ	**90.2**	88.0
C jejuni	TET	**96.3**	(81.2) 83.3
E coli	GEN	**68.7**	49.7
K pneumoniae	GEN	80.1	**81.9**
N gonorrhoeae	TET	**82.5**	69.2
P aeruginosa	CAZ	68.4	**75.7**
S enterica	GEN	89.3	**92.4**
S aureus	ERY	94.8	**95.3**
S pneumoniae	ERY	84.1	**88.5**
Average		**83.8**	(80.2) 80.4

Bold text denotes the highest value per pair. Two test results for *C jejuni* were reported by CAMDA without explanation (81.2% and 83.3%); the lower of the 2 is encompassed by parentheses.

Discussion

Public SRA submissions varied widely in layout, read length, and overall quality. The pipeline resolved these mismatches by detecting layout and length directly from SRA records and then applying standardized trimming, filtering, decontamination, and alignment, allowing isolates to be compared on the same footing. Sequencing reads were aligned to both a strain-resolved genome reference and a curated AMR gene-class library, producing 2 complementary feature sets that captured lineage-specific and gene-based variation. These references were assembled and indexed in-house for use with KMA, ensuring uniform mapping across the dataset.

On the processed data, the majority-vote ensemble consistently matched or exceeded the best individual classifier on the internal evaluation of balanced accuracy and ROC. On the external evaluation, it reached a mean balanced accuracy of 80.4%. Performance was uneven across species: *E coli*, *N gonorrhoeae*, and *C jejuni* showed reduced external accuracy while *S enterica*, *S aureus*, *S pneumoniae*, and *P aeruginosa* maintained or exceeded their internal results. These differences are presented descriptively; potential causes include dataset shift in lineage composition, variation in read quality and layout, or incomplete labeling, but the available data do not allow attribution. In addition, incomplete SRA metadata limited model stratification: many runs lacked information on collection site, time, or source, preventing adjustment for ecological or epidemiological context. Such omissions likely contributed to the modest internal-external performance gap by obscuring structure within the training data.

Feature importance analysis identified beta-lactam, aminoglycoside, macrolide, and tetracycline classes as the strongest predictors across pathogens, with strain-level features adding additional discriminatory power. This correspondence with known resistance mechanisms supports the biological plausibility of the learned associations.

Furthermore, strain-level features showed strength by identifying known strains used in resistance research.[32,33] The model may capture broader genomic variation beyond annotated AMR genes, potentially reflecting genetic signals not currently represented in curated resistance databases such as ResFinder; however, identifying and validating the specific genomic determinants underlying these associations lie beyond the scope of the present study.

Recent machine-learning approaches to AMR prediction often rely on high-dimensional genomic representations such as single nucleotide polymorphisms (SNPs) or k-mer frequencies. Frameworks such as AMRLearn use genome-wide SNP matrices to link sequence variation to resistance phenotypes, allowing models to capture fine-scale mutational signals associated with AMR.[36] Similarly, k-mer-based models have been applied to pathogens such as *P aeruginosa* to construct highly expressive feature spaces derived directly from sequence composition.[37] These representations can capture fine-scale genomic variation associated with resistance, but they typically generate extremely large feature spaces and may incorporate many variants that are difficult to interpret biologically. In contrast, the feature representation used here summarizes genomic information through strain-level abundance and resistance gene-class counts derived from curated reference libraries. These features capture both a proxy of lineage structure through strain abundance as well as the presence of known resistance mechanisms while maintaining a substantially

lower-dimensional and biologically interpretable feature space. In practice, this representation allows the models to incorporate broader genomic context, including resistance patterns associated with genomic backgrounds rather than individual mutations alone. Conceptually, this approach prioritizes interpretable genomic signals and broader genomic context, whereas SNP- and k-mer-based models aim to capture fine-scale sequence variation across the entire genome.

Future improvements will depend on both methodological and infrastructural developments. On the methodological side, richer feature representation and inclusion of the intermediate susceptibility category could refine decision boundaries and better reflect laboratory practice.[38] In addition, class-weighting and resampling strategies, which may have modestly improved performance, were not applied because class imbalance was moderate and balanced accuracy was used for evaluation. For datasets exhibiting more severe imbalance, however, their inclusion would be recommended. On the infrastructural side, more consistent and complete metadata are required for models to learn relationships between genotype and the settings in which resistance evolves. Broader and better-annotated training data, drawn from geographically and temporally diverse sources, will be essential to capture rare or emerging resistance patterns.

Conclusion

This study introduced a bioinformatics pipeline purpose-built to process public SRA sequencing data of heterogeneous format, quality, and origin. The system automated metadata extraction, web-based retrieval of layout and read length, batch downloading, multi-stage quality control, and contaminant filtering on the University of Florida HiPerGator cluster.[23] Two reference resources were created to serve as alignment targets: a strain-resolved genome library constructed from RefSeq assemblies of 9 target species (2009-2020) and a curated AMR gene-class library derived from ResFinder.[9,13,14] The pipeline integrated these components into a standardized workflow that generated normalized strain-level and AMR feature matrices suitable for downstream learning.

On these prepared data, an ensemble of LLR, RF, and gradient boosting achieved a mean balanced accuracy of 80.4% on external evaluation. Feature importance analyses highlighted beta-lactam, aminoglycoside, macrolide, and tetracycline classes, together with strain-level markers, consistent with established resistance mechanisms.

Future extensions should emphasize richer and more consistent metadata describing time, location, and source of isolation, allowing ecological and epidemiological context to inform model learning. Incorporating a 3-category outcome that includes the intermediate susceptibility class would improve interpretability and reduce type I error.[38] Finally, prospective multi-site validation with prespecified evaluation criteria will be required to confirm generalizability and establish this pipeline as a reproducible framework for large-scale AMR prediction.

Acknowledgements

The authors thank the Critical Assessment of Massive Data Analysis (CAMDA) 2025 Antimicrobial Resistance Prediction Challenge, an initiative of the International Society for Computational Biology (ISCB), for providing the sequencing data and metadata used in this study. The authors also thank the 2 anonymous reviewers for their constructive comments and suggestions, which helped improve the clarity and rigor of this work. In addition, the authors acknowledge Shoumi Sarkar for providing example code whose general structure served as a reference in the early development of the present analysis.

ORCID iDs

Owen Visser https://orcid.org/0009-0001-4877-0560
Victor Agboli https://orcid.org/0009-0006-5597-8315

Author Contributions

Owen Visser: Methodology; Software; Investigation; Formal analysis; Visualization; Writing—original draft; Conceptualization; Data curation.
Victor Agboli: Validation; Writing—review & editing.
Somnath Datta: Project administration; Supervision; Writing—review & editing; Resources.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statement

The raw sequencing data analyzed in this study are publicly available through the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA) at https://www.ncbi.nlm.nih.gov/sra via accession lookup. The GitHub repository for this study (https://github.com/shupasneaky/AMR-Prediction-Code) provides all SRA accession codes used in the analysis, including those for the training dataset, test dataset, and the reference genomes used to construct the genomic libraries. The true biological labels for the external test dataset are held by CAMDA as part of the blinded challenge evaluation process and were not accessible to the authors.

Supplemental Material

Supplemental material for this article is available online at https://github.com/shupasneaky/AMR-Prediction-Code.

References

1. Afshinnekoo E, Bhattacharya C, Burguete-García A, et al. COVID-19 drug practices risk antimicrobial resistance evolution. *Lancet Microbe*. 2021;2(4):e135-e136.
2. Zhang R, Ellis D, Walker AR, Datta S. Unraveling city-specific microbial signatures and identifying sample origins for the data from CAMDA 2020 metagenomic geolocation challenge. *Front Genet*. 2021;12:659650.
3. Chng KR, Li C, Bertrand D, et al. Cartography of opportunistic pathogens and antibiotic resistance genes in a tertiary hospital environment. *Nat Med*. 2020;26(6):941-951.

4. Duchêne S, Holt KE, Weill FX, et al. Genome-scale rates of evolutionary change in bacteria. *Microb Genom*. 2016;2(11):e000094.
5. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*. 2020;21(2):345-352.
6. Kirkup Benjamin C Jr. Bacterial strain diversity within wounds. *Advances in Wound Care*. 2015;4:12-23.
7. World Health Organization. WHO publishes list of bacteria for which new antibiotics are urgently needed. 2017. Accessed May 16, 2026. https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed
8. Angers-Loustau A, Petrillo M, Bengtsson-Palme J, et al. The challenges of designing a benchmark strategy for bioinformatics pipelines in the identification of antimicrobial resistance determinants using next generation sequencing technologies. *F1000res*. 2018;7:ISCB Comm J-459.
9. National Center for Biotechnology Information. Sequence read archive (SRA). 2025. Accessed May 12, 2025. https://www.ncbi.nlm.nih.gov/sra
10. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170
11. Lab TH. Kneaddata: quality control tool for metagenomic data. 2018. Accessed May 12, 2025. https://huttenhower.sph.harvard.edu/kneaddata/
12. Van Oeffelen M, Nguyen M, Aytan-Aktug D, et al. A genomic data resource for predicting antimicrobial resistance from laboratory-derived antimicrobial susceptibility phenotypes. *Brief Bioinform*. 2021;22(6):bbab313.
13. Florensa AF, Kaas RS, Clausen PTLC, Aytan-Aktug D, Aarestrup FM. Resfinder—an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. *Microb Genom*. 2022;8(1):000748.
14. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*. 2007;35(Database issue):D61-65.
15. Clausen PTC, Aarestrup FM, Lund O. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics*. 2018;19(1):307. doi:10.1186/s12859-018
16. Ren Y, Chakraborty T, Doijad S, et al. Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. *Bioinformatics*. 2022;38(2):325-334.
17. Kim JI, Maguire F, Tsang KK, et al. Machine learning for antimicrobial resistance prediction: current practice, limitations, and clinical perspective. *Clin Microbiol Rev*. 2022;35(3):e00179-00121.
18. Tang R, Luo R, Tang S, Song H, Chen X. Machine learning in predicting antimicrobial resistance: a systematic review and meta-analysis. *Int J Antimicrob Agents*. 2022;60(5-6):106684.
19. Mishra S, Han TA, Lopes BS, et al. Predicting antimicrobial resistance (AMR) in campylobacter, a foodborne pathogen, and cost burden analysis using machine learning. *arXiv Preprint*. 2025; arXiv:250903551.
20. Woh PY, Soengkono F, Chen Y, et al. Genomic insights into nontyphoidal salmonella: prediction of antimicrobial resistance with whole genome-based machine learning. *Int J Antimicrob Agents*. 2025;66(5):107575.
21. Nguyen M, Brettin T, Long S, et al. Developing an in silico minimum inhibitory concentration panel test for Klebsiella pneumoniae. *Sci Rep*. 2018;8(1):421.
22. Lüftinger L, Májek P, Beisken S, Rattei T, Posch AE. Learning from limited data: towards best practice techniques for antimicrobial resistance prediction from whole genome sequencing data. *Front Cell Infect Microbiol*. 2021;11:610348.
23. UFIT Research Computing. *HiPerGator Supercomputing Cluster*. University of Florida. 2025. Accessed May 16, 2026. https://www.rc.ufl.edu/about/hipergator/
24. Wright MN, Ziegler A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Softw*. 2017;77(1):1-17. doi:10.18637/jss.v077.i01
25. Tay JK, Narasimhan B, Hastie T. Elastic net regularization paths for all generalized linear models. *J Stat Softw*. 2023;106. doi:10.18637/jss.v106.i01
26. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016:785-794; San Francisco, CA. ACM. doi:10.1145/2939672.2939785
27. Greenlon A, Chang PL, Damtew ZM, et al. Global level population genomics reveals differential effects of geography and phylogeny on horizontal gene transfer in soil bacteria. *Proc Nat Acad Sci*. 2019;116(30):15200-15209. doi:10.1073/pnas.1900056116
28. O'Leary N, Cox E, Holmes J, et al. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI datasets. *Sci Data*. 2024;11(1):732. doi:10.1038/s41597-024
29. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nature Methods*. 2012;9(4):357-359. doi:10.1038/nmeth.1923
30. Paulson JN, Pop M, Bravo HC. Metagenomeseq: statistical analysis for sparse high-throughput sequencing. *Bioconductor Package*. 2013;1:191.
31. Kuhn M. Building predictive models in r using the caret package. *J Stat Softw*. 2008;28(5):1-26. doi:10.18637/jss.v028.i05
32. Ou HY, Kuang SN, He X, et al. Complete genome sequence of hypervirulent and outbreak-associated *Acinetobacter baumannii* strain LAC-4: epidemiology, resistance genetic determinants and potential virulence factors. *Sci Rep*. 2015;5(1):8643.
33. Pratt A, Korolik V. Tetracycline resistance of Australian Campylobacter jejuni and campylobacter coli isolates. *J Antimicrob Chemother*. 2005;55(4):452-460.
34. Peleg AY, Seifert H, Paterson DL. Acinetobacter baumannii: emergence of a successful pathogen. *Clin Microbiol Rev*. 2008;21(3):538-582. doi:10.1128/CMR.00058-07
35. Lowy FD. Staphylococcus aureus infections. *N Engl J Med*. 1998;339(8):520-532. doi:10.1056/NEJM199808203390806
36. Zhang X, Hu Y, Cheng Z, et al. AMRLearn: protocol for a machine learning pipeline for characterization of antimicrobial resistance determinants in microbial genomic data. *STAR Protocols*. 2025;6(2):103733.
37. Arango-Argoty G, Garner E, Pruden A, et al. Deeparg: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*. 2018;6(1):23.
38. Hombach M, Böttger EC, Roos M. The critical influence of the intermediate category on interpretation errors in revised EUCAST and CLSI antimicrobial susceptibility testing guidelines. *Clin Microbiol Infect*. 2013;19(2):E59-71.