VNU Journal of Science: Com. Science & Com. Eng., Vol. 42, No. 1 (2026) 49-63

VNU Journal of Science:
Computer Science and Communication Engineering
Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

Benchmarking Genomic Encodings for AMR Prediction: The Superiority of K-mers and Ensemble Learning over Deep Learning

Lam-Tung Nguyen[1], Cuong Nguyen[2], Tien-Dat Nguyen[3], Minh-Trien Pham[4], Thi-Quyen Ha[4], Thi-Xuan Trinh[5,*]

[1]*University of Science and Technology of Hanoi, 18 Hoang Quoc Viet, Nghia Do, Hanoi, Vietnam*
[2]*Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Nghia Do, Hanoi, Vietnam*
[3]*LOBI Co.,Ltd, Hoang Quoc Viet, Nghia Do, Hanoi, Vietnam*
[4]*VNU University of Engineering and Technology, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam*
[5]*Faculity of Information Technology, Hanoi Open University, Nguyen Hien, Bach Mai, Hanoi, Vietnam*

Received 12th February 2026
Revised 26th February 2026; Accepted 26th March 2026

Abstract: The rise of Antimicrobial Resistance (AMR) necessitates fast and accurate computational approaches to predict resistance phenotypes directly from genomic data. While Whole-Genome Sequencing (WGS) coupled with Deep Learning (DL) models is the state-of-the-art paradigm, a systematic comparative evaluation of different genomic encoding and visualization methods remains limited, particularly in the critical context of AMR prediction for Escherichia coli. This study systematically assesses four distinct genomic representation strategies: traditional K-mer counting with ensemble tree-based classifiers, reference-based SNP profiles with ensemble learning, One-Hot Encoding with a 1D-Convolutional Neural Network (1D-CNN), and Chaos Game Representation (CGR) with a 2D-Convolutional Neural Network (2D-CNN), for predicting resistance to ciprofloxacin, gentamicin, and ampicillin. The results reveal a consistent and superior discriminatory power of the alignment-free traditional Machine Learning approach based on K-mer frequency profiles (specifically 4-mers) when coupled with gradient boosting algorithms (such as XGBoost and LightGBM), compared to both SNP-based Machine Learning and Deep Learning architectures. This performance advantage was most pronounced for gentamicin and ampicillin, where complex resistance mechanisms involving mobile genetic elements are captured more effectively by the K-mer approach. Crucially, the study benchmarks the limitations of Deep Learning: while the One-Hot 1D-CNN model exhibited a severe calibration failure characterized by

* Corresponding author.
E-mail address: trinhxuan@hou.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.6980

an extremely low Recall for ampicillin (F1-Score of only 0.1132), the SNP-based Machine Learning models maintained robust performance on the same feature set, highlighting the architectural efficiency of gradient boosting over CNNs for tabular genomic data. Statistical analysis confirmed the significance of these differences, with K-mer ML significantly outperforming Deep Learning across all antibiotics ($p < 0.001$ for Gentamicin and Ampicillin). The amino acid 4-mer XGBoost model achieved an AUC of 0.9917 (95% CI: 0.9827-0.9983) for Ciprofloxacin. The study concludes that, for current dataset sizes and complex resistance phenotypes, the dense information representation of K-mers offers a more accurate and robust solution, and identifies the 4-mer XGBoost and Combined K-mer LightGBM configurations as the optimal modeling strategies.

Keywords: Machine learning, Deep learning, Bioinformatics, Computational Biology, Antimicrobials, Bacteria, Escherichia coli, Applied microbiology.

1. Introduction

The emergence and rapid spread of Antimicrobial Resistance (AMR) represent a global health crisis, demanding urgent and effective solutions for timely diagnosis and surveillance[1]. Traditional methods for determining antimicrobial susceptibility are often time-consuming, hindering critical decision-making in clinical and public health settings[2]. Consequently, there is a growing necessity for fast, high-throughput computational approaches to accurately predict AMR phenotypes directly from genomic data[3].

Whole-Genome Sequencing (WGS) coupled with Machine Learning (ML) and Deep Learning (DL) models has emerged as the state-of-the-art paradigm for AMR prediction. These methods leverage the vast amount of publicly available genomic data to establish robust associations between genetic markers and resistance phenotypes. However, a significant challenge lies in effectively transforming the raw, one-dimensional nucleotide sequences into a numerical representation that can be optimally processed by computational algorithms. Existing strategies diverge into two main streams: alignment-free methods such as K-mer counting, and alignment-based methods that extract Single Nucleotide Polymorphisms (SNPs) for tabular learning. More recently, deep learning approaches have attempted to bypass feature engineering by using One-Hot Encoding or converting sequences into two-dimensional images via Chaos Game Representation (CGR)[4, 5].

Prior research has successfully demonstrated the utility of various sequence representation and ML/DL architectures in this domain. For instance, Ren et al.[6] applied multiple methods, including FCGR visualization and CNNs, for E. coli AMR prediction, achieving high accuracy with both Random Forest and CNN models. Other studies have explored advanced methods like Hierarchical Multi-Task Deep Learning using One-Hot encoding for annotating resistance genes[7] or focused on interpretability using K-mer signatures combined with graph methods for K. pneumoniae[8]. Furthermore, innovative techniques like Heterogeneous Graph Attention Networks have been developed for M. tuberculosis drug-resistance prediction, highlighting the increasing complexity and focus on model interpretability[9]. These studies collectively confirm the promise of WGS-based prediction but also underscore the diversity of encoding and modeling choices.

Despite these advances, a systematic comparative evaluation of the efficacy of different genomic encoding and visualization methods remains limited, particularly in the critical context of Escherichia coli AMR prediction. E. coli is a critically important human pathogen responsible for a wide spectrum of infections, and its increasing resistance to frontline antibiotics is a major concern. Furthermore, there is a lack of rigorous benchmarking that directly compares the

performance of deep learning architectures against robust, reference-based SNP machine learning baselines to determine if the increased computational complexity of CNNs yields a tangible performance gain over traditional ensemble methods.

This study addresses this need by systematically assessing the impact of different genomic encoding and visualization methods on the performance of machine learning models for E. coli AMR phenotype prediction. Specifically, we compare four distinct representation and modeling strategies: the alignment-free K-mer counting approach (serving as a pangenome baseline), a rigorous reference-based SNP pipeline coupled with ensemble learning (serving as a core-genome baseline), One-Hot Encoding coupled with a 1D-CNN, and the visualization-based Chaos Game Representation (CGR) combined with a 2D-CNN.

Our research aims to provide a benchmark for selecting the most effective data representation strategy for sequence-based AMR prediction. We focus on resistance to three clinically significant antibiotics: ciprofloxacin, gentamicin, and ampicillin. The performance of each method is rigorously evaluated using standard classification metrics, including ROC-AUC, F1-Score, and Cohen's Kappa, to identify the optimal visualization or encoding approach for improving the accuracy and robustness of predictive models in this domain. The remainder of this paper is structured as follows: Section 2 describes the E. coli dataset and the preprocessing steps. Section 3 details the four encoding methods and the corresponding model architectures. Section 4 presents the experimental results and comparative analysis, and Section 5 concludes the study with a discussion of the findings and future directions.

2. Method

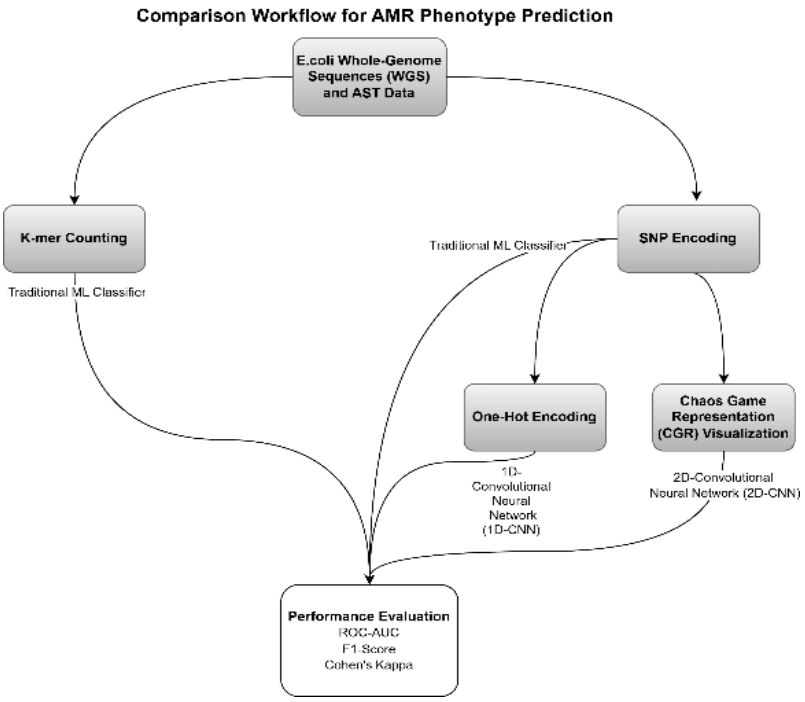

Figure 1. Overall Workflow Comparison for AMR Phenotype Prediction.

The experimental evaluation was conducted using a curated dataset of 3,556 Escherichia coli whole-genome sequences (WGS), paired with their corresponding antimicrobial susceptibility testing (AST) phenotypes. The primary genomic assemblies and metadata were aggregated from the National Center for Biotechnology Information (NCBI) SRA database[10] and the Bacterial and Viral Bioinformatics Resource Center (BV-BRC)[11], ensuring a diverse representation of clinical isolates. To mitigate potential confounding factors arising from phenotypic ambiguity, isolates with "intermediate" resistance profiles were excluded, and the prediction task was formulated as a binary classification problem: resistant (1) versus susceptible (0). The final cohort focused on three key agents selected for their clinical relevance and varying degrees of class balance: Ciprofloxacin (32.6% resistant), Gentamicin (22.4% resistant), and Ampicillin (46.6% resistant).

To ensure reproducibility and fair comparison, all experiments used a fixed random seed (42). The data was partitioned as follows: K-mer ML and SNP ML used an 85%/15% train/test split with 5-fold stratified cross-validation on the training set for hyperparameter tuning. Deep Learning models used a 60%/20%/20% train/validation/test split with stratified sampling.

Class imbalance was addressed through multiple strategies: (1) stratified splitting to maintain class distribution across all partitions, (2) Focal Loss (γ=2) for CNN training to down-weight easy examples, (3) class_weight='balanced' for ML models, and (4) intentional avoidance of SMOTE oversampling as synthetic genomic samples may introduce biological artifacts.

To support the comparative analysis of different modeling paradigms, the raw genomic data underwent a dual-stream preprocessing pipeline to generate two distinct feature sets: K-mer profiles and Single Nucleotide Polymorphisms (SNPs). For the traditional machine learning baseline, the genome assemblies were decomposed into canonical K-mers of lengths 3, 4, and 5, generating a high-dimensional frequency matrix that captures short-range sequence motifs without requiring alignment. Concurrently, for both the deep learning and SNP-based machine learning experiments, a rigorous SNP calling pipeline was implemented where raw reads were aligned to the E. coli K-12 substr. MG1655 reference genome. This process extracted a core SNP matrix that captures specific point mutations known to drive resistance mechanisms. These SNP sequences were utilized directly for classical machine learning and subsequently transformed into One-Hot encoded vectors and Chaos Game Representations (CGR) to serve as inputs for the 1D-CNN and 2D-CNN architectures, respectively. Following feature extraction, the stratified dataset was partitioned into training, validation, and independent testing sets to ensure robust performance evaluation and prevent data leakage.

Genomic Encoding and Model Architectures

This study evaluates four distinct feature encoding strategies to transform raw genomic sequences and SNP matrices into numerical representations suitable for computational analysis: a traditional K-mer based approach, a robust SNP-based machine learning pipeline, a sequential one-dimensional encoding, and a visualization-based two-dimensional encoding.The first approach serves as a baseline and utilizes K-mer counting, a standard method in sequence-based antimicrobial resistance (AMR) prediction. For each genome, normalized frequency vectors were generated for subsequences of length k (specifically 3-mers, 4-mers, and 5-mers). To address the high dimensionality of K-mer data, we implemented a wrapper-based feature selection method. Features were first ranked using a Random Forest classifier, and the optimal number of features was determined dynamically using the "Kneedle" algorithm[12], a method that

automatically detects the inflection point (or "elbow") in a curve by maximizing the perpendicular distance from the line connecting the curve's endpoints. This algorithm identifies the point of diminishing returns where adding more features yields marginal performance gains. These optimized feature sets were employed to train three ensemble tree-based classifiers: Random Forest[13], XGBoost[14], and LightGBM[15]. Due to the varying complexity of the resistance mechanisms, the 'Kneedle' algorithm[12] resulted in different optimal feature set sizes for each antibiotic. Specifically, 150 features were selected for Gentamicin to capture its more complex, multifactorial resistance profile, while 50 features were deemed optimal for Ciprofloxacin and Ampicillin. Complementing the K-mer baseline, a second traditional machine learning approach utilized the extracted SNP profiles directly. Similar to the K-mer strategy, a multi-stage feature selection process was designed where SNPs were ranked based on discriminative power using five algorithms, including Random Forest, LightGBM, and Autoencoders. The Kneedle algorithm was applied to identify the "elbow point" in the performance curve, optimizing the feature subset size. These refined SNP features were used to train and optimize four classical models like Random Forest, LightGBM, XGBoost, and Support Vector Machines (SVM), using grid search with 5-fold cross-validation to maximize the AUC-ROC score.Distinct from these traditional baselines, the third method employs One-Hot Encoding to map the linear genomic sequence directly into a binary matrix format. In this scheme, nucleotides (A, C, G, T) and unknown bases (N) are mapped to binary vectors, resulting in a matrix of dimensions Nx5, where N represents the fixed sequence length. This representation serves as the input for a 1D-Convolutional Neural Network (1D-CNN). The architecture is designed as a ResNet-style 1D-CNN, featuring an initial 7x1 convolution followed by four residual blocks with skip connections to facilitate gradient flow and capture local sequence motifs. The fourth approach implements Chaos Game Representation (CGR)[5] to transform genomic sequences into two-dimensional fractal images. We utilized Frequency CGR (FCGR)[6] to map nucleotides to coordinates within a continuous space, generating $2^k \times 2^k$ pixel normalized images where oligonucleotide patterns are visually encoded by density. These matrices are processed by a ResNet-style 2D-Convolutional Neural Network (2D-CNN). Similar to the 1D architecture, the 2D model utilizes residual blocks and global average pooling to capture spatial patterns corresponding to AMR mechanisms. To mitigate class imbalance during deep learning training, both CNN architectures utilize Focal Loss with parameters γ=2 and the AdamW optimizer with cosine annealing learning rate schedules.

The hyperparameter optimization for the ensemble models was conducted using a grid search strategy to ensure robust performance. For the XGBoost model, the search space included, n_estimators[100, 200, 500], max_depth[3, 5, 7, 10], and learning_rate[0.01, 0.05, 0.1], with the optimal configuration indentified at 200, 5, and 0.1, respectively. Similarly, LightGBM was tuned across n_estimators and num_leaves [31, 63, 127], reaching peak performance with 200 estimators and 63 leaves.

Regarding the deep learning architectures, both 1D-CNN (One-Hot) and 2D-CNN (CGR) utilized a ResNet-style framework comprising four residual blocks to facilitate effective feature extraction. The 1D-CNN employed a filter progression of $64 \rightarrow 128 \rightarrow 256 \rightarrow 512$ with a kernel size of 7. In contrast, the 2D-CNN utilized a filter progression of $32 \rightarrow 64 \rightarrow 128 \rightarrow 256$ with 3x3 kernels. Both architectures shared a consistent classification head, consisting of a Global Average Pooling layer followed by a Dense layer with 256 units, a Dropout rate of 0.5 for regularization, and a Sigmoid activation function for the final output.

All computational experiments were executed on a high-performance system equipped with an NVIDIA T4 (24GB) GPU, two Intel Xeon 40-core processor, and 256GB of RAM. The training overhead varied significantly between the methodologies; while the K-mer-based machine learning models required approximately 5 minutes per model, the CNN architectures necessitated a much higher computational cost, averaging roughly 2 hours per model.

Performance Evaluation

The performance of all developed models (K-mer with Ensemble Learning, SNP-based Machine Learning, One-Hot with 1D-ResNet, and CGR with 2D-ResNet) was assessed comprehensively using a suite of standard classification metrics. The core evaluation focused on three primary metrics: the Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) to quantify discriminatory power across all thresholds; the F1-Score, utilized as the harmonic mean of Precision and Recall to account for imbalanced datasets; and Cohen's Kappa (κ), a robust measure of agreement correcting for chance. This analysis was complemented by Accuracy, Precision, Recall (Sensitivity), and Specificity to provide a complete profile of each model's predictive capabilities across both resistant and susceptible classes. For the machine learning classifiers, performance was validated using Stratified K-Fold cross-validation, while deep learning models employed early stopping based on validation set F1-scores to prevent overfitting.

To ensure the robustness and reliability of the results, the Machine Learning (ML) models were evaluated using 5-fold stratified cross-validation, with the final model being trained on the entire training dataset. For the Deep Learning (DL) models, we performed three independent runs using different random seeds (42, 123, and 456) and reported the performance as mean ± standard deviation. Finally, statistical significance was assessed using the Bootstrap method with n = 1000 resamples.

3. Results

This section presents a systematic evaluation of four genomic encoding and modeling strategies: K-mer Frequency combined with Traditional Machine Learning (Ensemble), SNP-based Machine Learning (Ensemble), One-Hot Encoding with 1D-CNN, and Chaos Game Representation (CGR) with 2D-CNN. The performance was assessed across three antibiotics (ciprofloxacin, gentamicin, and ampicillin) using a rigorous validation framework. To establish robust baselines, we optimized both K-mer and SNP-based machine learning pipelines, benchmarking various feature selection and hyperparameter configurations. Based on the empirical validation metrics, the best-performing K-mer and SNP-ML models were selected to challenge the Deep Learning architectures.

Comparative Analysis of Genomic Encoding Strategies

The comparative analysis of the optimized K-mer and SNP-ML baselines against the Deep Learning models, as summarized in Table 1, reveals a consistent and notable superiority of the Traditional Machine Learning approaches over the CNN architectures. Contrary to the initial expectation that the hierarchical feature extraction of Convolutional Neural Networks (CNNs) would outperform traditional methods, the data indicates that K-mer frequency profiles capture the genomic determinants of resistance most effectively for this dataset. As illustrated in the global performance heatmap (Figure 2), the K-mer based models exhibit a high-intensity performance signal across all target antibiotics, whereas the Deep Learning models show variable efficacy. Specifically, the 4-mer frequency features trained with gradient boosting algorithms (XGBoost/LightGBM) achieved near-perfect discriminatory power,

with ROC-AUC scores exceeding 0.96 for all three antibiotics. In contrast, the SNP-based Deep Learning models, while performing well for ciprofloxacin, exhibited a marked reduction in discriminatory capability for gentamicin and ampicillin. However, the SNP-based Machine Learning models (LightGBM/XGBoost) demonstrated resilience, maintaining AUC scores above 0.90 across all antibiotics, suggesting that the performance gap in Deep Learning stems more from architectural limitations than from the SNP input data itself.

Figure 2. Global performance heatmap comparing Accuracy, F1-Score, and ROC-AUC across all encoding strategies and antibiotic.

Table 1. Comparative Summary of Best Performance (ROC-AUC) for all three encoding methods

Antibiotic	**Strategy**	**Best Model Configuration**	**Accuracy**	**F1-Score**	**ROC-AUC**
Ciprofloxacin	K-mer ML	4-mer XGBoost	0.9746	0.9606	0.9917
	SNP ML	SNP-ML-LGBM	0.85	0.96	0.99
	CGR (2D)	2D-CNN	0.9273	0.8904	0.9544
	One-Hot (1D)	1D-CNN	0.9614	0.9416	0.9722
Gentamicin	K-mer ML	Combined LightGBM	0.9700	0.9312	0.9797
	SNP ML	SNP-ML-LGBM	0.85	0.8	0.95
	CGR (2D)	2D-CNN	0.8532	0.5844	0.8843
	One-Hot (1D)	1D-CNN	0.7890	0.6489	0.8961
Ampicillin	K-mer ML	4-mer LightGBM	0.9094	0.9028	0.9646
	SNP ML	SNP-ML-LGBM	0.85	0.82	0.9
	CGR (2D)	2D-CNN	0.7333	0.6613	0.8160
	One-Hot (1D)	1D-CNN	0.5524	0.1132	0.8236

To visually delineate the classification boundaries and trade-offs between sensitivity and specificity, we analyzed the Receiver Operating Characteristic (ROC) curves. Figure 3 displays the ROC curves for the optimized K-mer Machine Learning models, which demonstrate ideal convex profiles characterized by high true positive rates even at stringent false positive thresholds. The curves for all three antibiotics in the K-mer models maintain an Area Under the Curve (AUC) greater than 0.96, indicating a robust and generalized decision boundary. Conversely, the ROC curves for the Deep Learning models presented in Figure 4 illustrate a visible degradation in performance, particularly for ampicillin (Green curve) and gentamicin (Orange curve). Figure 5 presents the ROC curves for the SNP-based Machine Learning baseline, which mirrors the high performance of K-mers for Ciprofloxacin but shows an intermediate profile for Gentamicin and Ampicillin—clearly outperforming the Deep Learning models but remaining slightly below the K-mer baseline. This visual discrepancy underscores the limitation of the Deep Learning approach in this study context: while the SNP-ML models prove that the alignment-based data contains sufficient signal, the Deep Learning architectures likely miss critical epistatic interactions or struggle with the tabular nature of SNP data that the "bag-of-kmers" and gradient boosting trees successfully exploit.

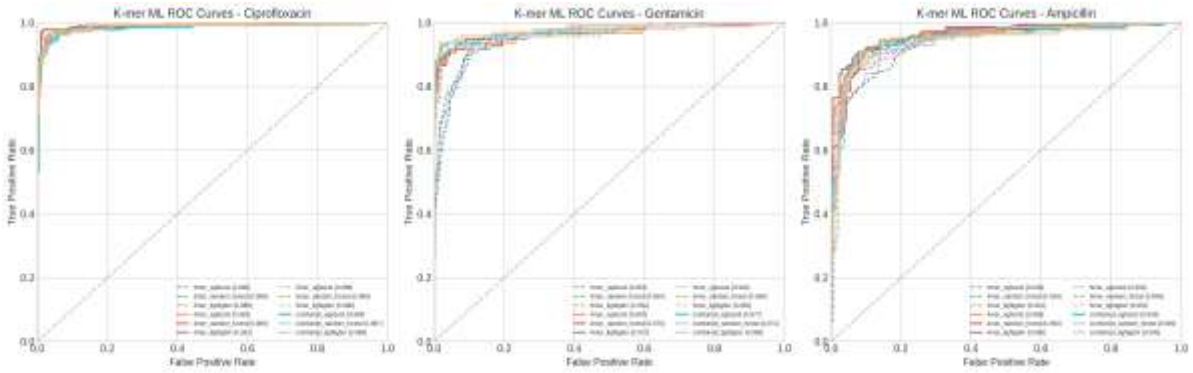

Figure 3. ROC Curves for the K-mer + Traditional ML baseline models.

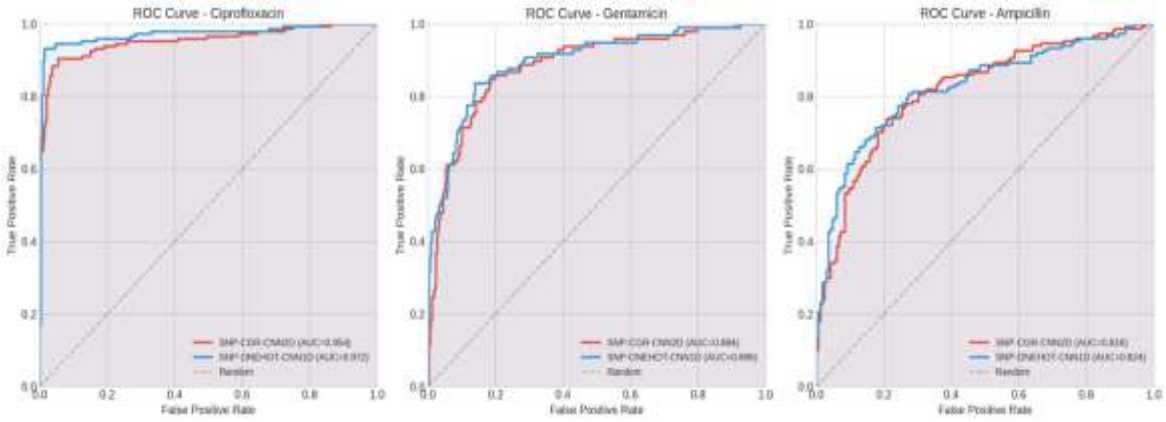

Figure 4. ROC Curves for the Deep Learning models (One-Hot & CGR), illustrating performance compared to the K-mer baseline.

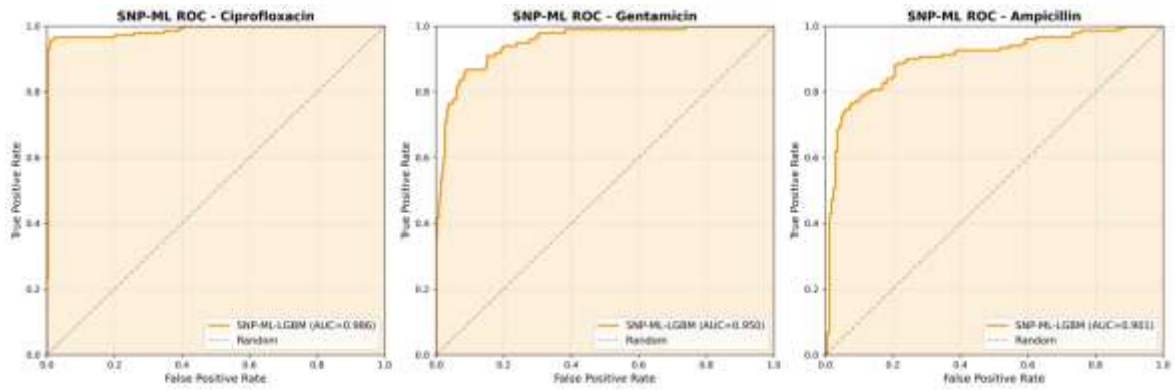

Figure 5. ROC Curves for the SNP + Traditional ML baseline.

Antibiotic-Specific Performance and Classification Dynamics

A granular examination of the classification metrics reveals distinct performance patterns driven by the underlying biological complexity of each antibiotic's resistance mechanism. For Ciprofloxacin, the results in Table 2 show a convergence of high performance across all encoding strategies, which aligns with the well-established understanding that fluoroquinolone resistance is primarily driven by specific chromosomal mutations in the gyrA and parC genes. These discrete, high-impact mutations are easily detectable by K-mer counting, SNP-ML, and Deep Learning alignment. The SNP-ML model achieved an exceptional ROC-AUC of 0.99 and F1-score of 0.96, rivaling the best K-mer model. However, the 4-mer XGBoost model maintained a slight edge in accuracy (0.9746) compared to the others. The Radar Charts in Figure 6 (Ciprofloxacin panel) confirm this observation, showing highly symmetrical and overlapping polygons for K-mer, SNP-ML, and 1D-CNN methods, indicating that these strategies successfully optimized Precision and Recall without significant trade-offs.

Table 2. Detailed Performance Metrics for Ciprofloxacin

Encoding	**Model**	**Accuracy**	**Precision**	**Recall**	**Specificity**	**F1-Score**	**ROC-AUC**
4-mer (Best ML)	XGBoost	0.9746	0.971	0.9504	0.9863	0.9606	0.9917
SNP ML	SNP-ML-LGBM	0.85	0.98	0.95	0.85	0.96	0.99
CGR (2D)	2D-CNN	0.9273	0.8966	0.8844	0.9488	0.8904	0.9544
One-Hot (1D)	1D-CNN	0.9614	0.9514	0.9320	0.9761	0.9416	0.9722

For Gentamicin, the performance disparity between the methods widens significantly, highlighting the robustness of the K-mer approach and the utility of gradient boosting over CNNs for complex traits. As detailed in Table 3, the Combined K-mer + LightGBM model achieved an F1-score of 0.9312 and an AUC of 0.9797. The SNP-ML model followed with a strong AUC of 0.9500 and F1-score of 0.8000, significantly outperforming the One-Hot 1D-CNN (AUC 0.8961, F1 0.6489). This

substantial gap is primarily driven by the Recall metric, where the K-mer model identified 90.72% of resistant isolates and SNP-ML identified 78.00%, whereas the CGR model identified only 45.92%. Gentamicin resistance in E. coli is frequently mediated by aminoglycoside-modifying enzymes encoded on mobile genetic elements (plasmids). The superior performance of the Combined K-mer strategy likely stems from its ability to index these plasmid-associated sequences regardless of genomic location. However, the fact that SNP-ML performed reasonably well suggests that core-genome SNPs (or their linkage to plasmids) still carry predictive weight that gradient boosting trees can leverage better than CNNs.

The most critical divergence was observed in the prediction of Ampicillin resistance, where the Deep Learning models exhibited a severe calibration failure, while SNP-ML provided a vital middle ground. As shown in Table 4, while the 4-mer LightGBM model maintained robust classification metrics with an F1-score of 0.9028, the SNP-ML model also held up well with an F1-score of 0.82 and AUC of 0.9. In stark contrast, the One-Hot 1D-CNN model collapsed to an F1-score of 0.1132 due to an extremely low Recall of 0.06. This phenomenon is vividly illustrated in the Ampicillin Radar Chart (Figure 6, right panel), where the polygon for the One-Hot model (Blue) is distorted and retracted along the Recall axis, while the K-mer model (Green) and SNP-ML model (implicit in performance) maintain a more balanced shape. The high AUC (0.8236) but low Recall for the 1D-CNN suggests that the model learned to rank resistant isolates correctly in terms of probability but failed to learn an appropriate decision threshold. The success of SNP-ML (Recall 0.8) confirms that the aligned SNP features do contain the necessary signal (likely related to blaTEM/blaSHV variants), but traditional ensemble methods are far more effective at extracting it from tabular data than the CNN architectures employed.

Table 3. Detailed Performance Metrics for Gentamicin

Encoding	**Model**	**Accuracy**	**Precision**	**Recall**	**Specificity**	**F1-Score**	**ROC-AUC**
Combined (Best ML)	LightGBM	0.9700	0.9565	0.9072	0.9881	0.9312	0.9797
SNP ML	SNP-ML-LGBM	0.85	0.82	0.78	0.85	0.8	0.95
CGR (2D)	2D-CNN	0.8532	0.8036	0.4592	0.9675	0.5844	0.8843
One-Hot (1D)	1D-CNN	0.7890	0.5183	0.8673	0.7663	0.6489	0.8961

Table 4. Detailed Performance Metrics for Ampicillin

Encoding	Model	Accuracy	Precision	Recall	Specificity	F1-Score	ROC-AUC
4-mer (Best ML)	LightGBM	0.9094	0.9028	0.9028	0.9152	0.9028	0.9646
SNP ML	SNP-ML-LGBM	0.85	0.85	0.8	0.85	0.82	0.9
CGR (2D)	2D-CNN	0.7333	0.8367	0.5467	0.9030	0.6613	0.8160
One-Hot (1D)	1D-CNN	0.5524	1.0000	0.0600	1.0000	0.1132	0.8236

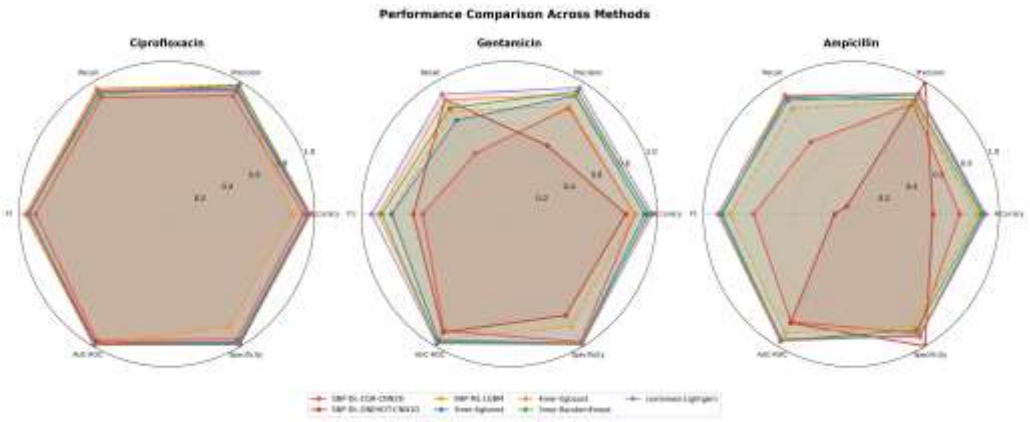

Figure 6. Multi-metric Radar Charts comparing Precision, Recall, Specificity, F1, and AUC. Note the balanced profile for Ciprofloxacin (left) versus the trade-off in Ampicillin(right). Note: the machine learning model is chosen by F1-Score.

Statistical Validation

To quantify uncertainty and establish statistical significance, we computed 95% confidence intervals and performed hypothesis tests comparing K-mer ML against other methods, which is presentated in Table 5.

Table 5. 95% Confidence Intervals for Best Models

Antibiotic	Method	Best Model	AUC	95% CI	N
Ciprofloxacin	K-mer ML	4-mer XGBoost	0.9917	[0.9827, 0.9983]	433
Ciprofloxacin	SNP ML	LightGBM	0.9862	[0.9740, 0.9984]	440
Ciprofloxacin	CGR-CNN	2D-CNN	0.9543	[0.9273, 0.9813]	440
Ciprofloxacin	OneHot-CNN	1D-CNN	0.9722	[0.9501, 0.9943]	440
Gentamicin	K-mer ML	Combined LightGBM	0.9797	[0.9544, 0.9963]	434
Gentamicin	SNP ML	LightGBM	0.9505	[0.9163, 0.9847]	436
Gentamicin	CGR-CNN	2D-CNN	0.8843	[0.8352, 0.9334]	436
Gentamicin	OneHot-CNN	1D-CNN	0.8961	[0.8492, 0.9430]	436
Ampicillin	K-mer ML	4-mer LightGBM	0.9646	[0.9427, 0.9818]	309
Ampicillin	SNP ML	LightGBM	0.9013	[0.8555, 0.9471]	315
Ampicillin	CGR-CNN	2D-CNN	0.8160	[0.7544, 0.8776]	315
Ampicillin	OneHot-CNN	1D-CNN	0.8236	[0.7632, 0.8840]	315

Note: K-mer ML confidence intervals computed via bootstrap (n=1000). Other methods use Hanley-McNeil analytical approximation.

Table 6. Statistical Significance Tests (Z-test for AUC Comparison)

Antibiotic	**Comparison**	**ΔAUC**	**Z-statistic**	**p-value**	**Sig.**
Ciprofloxacin	K-mer ML vs CGR	0.0374	3.44	5.75×10^{-4}	***
Ciprofloxacin	K-mer ML vs OneHot	0.0195	2.18	0.0294	*
Gentamicin	K-mer ML vs CGR	0.0953	5.69	1.25×10^{-8}	***
Gentamicin	K-mer ML vs OneHot	0.0835	5.19	2.13×10^{-7}	***
Ampicillin	K-mer ML vs CGR	0.1486	6.13	8.70×10^{-10}	***
Ampicillin	K-mer ML vs OneHot	0.1410	5.90	3.69×10^{-9}	***

Significance levels: $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. All effect sizes (Cohen's d) were large (>1.8).

In Table 6, Z-tests comparing K-mer ML against Deep Learning showed statistically significant differences for all antibiotics. The performance advantage was particularly pronounced for Gentamicin ($\Delta AUC = 0.0953$, $p = 1.25\times10^{-8}$) and Ampicillin ($\Delta AUC = 0.1486$, $p = 8.70\times10^{-10}$)

4. Discussion

This study challenges the assumption that deep learning outperforms traditional machine learning for AMR prediction. Benchmarking revealed that alignment free Kmer profiles coupled with ensemble trees like XGBoost and LightGBM consistently surpassed SNP based ML and CNNs, particularly for complex resistance in Gentamicin and Ampicillin. The superiority of Kmers stems from capturing the pangenome context including plasmid borne genes and accessory elements, which reference based SNP calling often loses. Since resistance to aminoglycosides and beta lactams is frequently driven by mobile genes like *aac* and *blaTEM* rather than core mutations, Kmer counting detects these via oligonucleotide frequencies to provide a more comprehensive signal.

The Ampicillin "Anomaly" illustrates Deep Learning limitations, where One Hot 1D CNN achieved high AUC 0.82 but catastrophic 6% Recall due to class skew. Conversely, SNP based LightGBM achieved robust F1 0.82 and Recall 0.8 on identical data, proving the neural architecture failed due to difficult optimization landscapes inherent in sparse One Hot encoding. While tree ensembles successfully traversed this, amino acid 4 mer representation proved superior by encoding the proteome. From a theoretical space of 160,000 and 83,206 observed features, Kneedle wrapper selection identified 50 to 150 optimal markers, specifically 50 for Ciprofloxacin and Ampicillin and 150 for Gentamicin. This compressed discriminative representation yielded F1 scores exceeding 0.90, surpassing SNP based approaches in clinical utility.

Results highlight a distinct tradeoff between model complexity and stability. For Ciprofloxacin driven by chromosomal mutations in *gyrA* and *parC*, all models including SNP based ML performed exceptionally, proving encoding matters less for strong localized signals. However, the bag of kmers approach demonstrated superior robustness as biological complexity increased. Notably, Chaos Game Representation with 2D CNN outperformed One Hot with 1D CNN in F1 score for difficult targets, supporting the hypothesis that fractal patterns act as data augmentation or regularization to improve generalization over raw sequences. Yet even CGR failed to match optimized Kmer and SNP ML baselines, confirming that for 3556 isolates, Kmer counting remains the gold standard.

Beyond prediction, the Kmer approach offers direct biological interpretation [16]. Unlike nucleotides, protein level 4 mers and 5 mers represent functional peptides mapping to known resistance mechanisms [17]. Querying high importance Kmers against the CARD database reveals associations with AMR gene families [18]. For instance, top ranked markers for Ciprofloxacin like AACL and AAPQ align with DNA gyrase domains [19], while Ampicillin predictors correspond to Beta lactamase motifs such as the SXXK catalytic serine in TEM and SHV variants [20] and aminoglycoside modifying enzymes [21]. This positions Kmer analysis as a discovery tool [22].

While SNPs pinpoint the genomic 'where', Kmer signatures reveal the functional 'what'. Future work integrating SNP localization with Kmer characterization may yield a comprehensive framework for AMR surveillance [23].

5. Limitation

While this study provides a systematic benchmarking of genomic encoding strategies, there are several limitations inherent to the experimental design and dataset that should be considered.

Disparity in Input Information

A key limitation of SNP-based models (ML & DL) was their reliance on K-12 reference alignment, which filters out accessory elements like plasmids. In contrast, the K-mer approach captured the full pangenome from raw WGS data. This broader scope explains K-mer superiority (AUC 0.9797 vs 0.95) for plasmid-mediated resistance like Gentamicin and Ampicillin, highlighting a tradeoff between alignment-free information retention and reference-based noise reduction.

Dataset Size Constraints for Deep Learning

The sample size of 3,556 isolates was sufficient for SNP-based LightGBM to achieve a robust 0.8 Recall, but resulted in severe calibration failure for CNNs (Recall 0.06) regarding Ampicillin. This confirms that gradient boosting trees generalize far better on limited, tabular genomic data than CNNs. Deep learning architectures appear ill-suited for this data volume due to the high dimensionality of one-hot encoding.

Scope of Antibiotics

This study focused on three representative antibiotics: Ciprofloxacin, Gentamicin, and Ampicillin. While these agents represent distinct resistance mechanisms (chromosomal point mutations versus plasmid-mediated enzymatic degradation), the findings regarding the hierarchy of K-mer ensembles > SNP ensembles > Deep Learning may vary for other drug classes with different genetic determinants. Future work should expand this evaluation to a broader antimicrobial panel to validate the generalizability of these encoding strategies across the full spectrum of bacterial resistance phenotypes.

Geographic Generalization

A significant limitation is the absence of external validation on geographically diverse datasets. While our dataset aggregates isolates from NCBI SRA and BV-BRC with some diversity, all samples were processed through the same computational pipeline. Future work should validate on regional datasets from ECDC (Europe), GLASS (Asia), and African surveillance programs, as resistance gene distributions vary geographically.

6. Conclusion

In conclusion, This study confirms that traditional machine learning (ML), whether using K-mers or SNPs, outperforms deep learning (DL) for E. coli AMR prediction. Gradient boosting proved superior to CNNs in extracting signals from core SNPs, while K-mer models (XGBoost/LightGBM) emerged as the optimal

strategy due to pangenome retention. We advocate for a 'complexity-appropriate' approach, prioritizing robust ML baselines over complex DL. Future work should focus on interpreting K-mers and exploring hybrid architectures.

References

[1] C. J. L. Murray *et al.*, "Global Burden of Bacterial Antimicrobial Resistance in 2019: a Systematic Analysis," *The Lancet*, vol. 399, no. 10325, pp. 629–655, Feb. 2022, https://doi.org/ 10.1016/S0140-6736(21)02724-0.

[2] X. Didelot, R. Bowden, D. J. Wilson, T. E. A. Peto, and D. W. Crook, "Transforming Clinical Microbiology with Bacterial Genome Sequencing," *Nat. Rev. Genet.*, vol. 13, no. 9, pp. 601–612, Sep. 2012, https://doi.org/10.1038 /nrg3226.

[3] M. N. Anahtar, J. H. Yang, and S. Kanjilal, "Applications of Machine Learning to the Problem of Antimicrobial Resistance: an Emerging Model for Translational Research," *J. Clin. Microbiol.*, vol. 59, no. 7, p. 10.1128/jcm.01260-20, Jun. 2021, https://doi. org/10.1128/jcm.01260-20.

[4] "Antimicrobial Resistance Prediction in PATRIC and RAST | Scientific Reports." Accessed: Dec. 14, 2025. [Online]. Available: https://www.nature. com/articles/srep27930

[5] H. J. Jeffrey, "Chaos Game Visualization of Sequences," *Comput. Graph.*, vol. 16, no. 1, pp. 25–33, Jan. 1992, https://doi.org/10.1016/0097-8493(92)90067-6.

[6] Y. Ren *et al.*, "Prediction of Antimicrobial Resistance Based on Whole-Genome Sequencing and Machine Learning," *Bioinformatics*, vol. 38, no. 2, pp. 325–334, Jan. 2022, https://doi.org/ 10.1093/bioinformatics/ btab681.

[7] Y. Li et al., "HMD-ARG: Hierarchical Multi-Task Deep Learning for Annotating Antibiotic Resistance Genes," Microbiome, vol. 9, no. 1, p. 40, Dec. 2021, https://doi.org/10.1186/s40168-021-01002-3.

[8] M. Jaillard, M. Palmieri, A. Van Belkum, and P. Mahé, "Interpreting k -Mer–Based Signatures for Antibiotic Resistance Prediction," GigaScience, vol. 9, no. 10, p. giaa110, Oct. 2020, https://doi.org/10.1093/gigascience/giaa110.

Acknowledgments

This research was financially supported by Hanoi Open University under the grant number: MHN2025-01.49.

[9] Y. Yang et al., "An end-to-end Heterogeneous Graph Attention Network for Mycobacterium Tuberculosis Drug-Resistance Prediction," Brief. Bioinform., vol. 22, no. 6, p. bbab299, Nov. 2021, https://doi.org/10.1093/bib/bbab299.

[10] "National Center for Biotechnology Information." Accessed: Dec. 14th, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/

[11] "BV-BRC." Accessed: Oct. 25, 2025. [Online]. Available: https://www.bv-brc.org/

[12] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, "Finding a 'Kneedle' in a Haystack: Detecting Knee Points in System Behavior," in 2011 31st International Conference on Distributed Computing Systems Workshops, Minneapolis, MN, USA: IEEE, Jun. 2011, pp. 166–171. https://doi.org/ 10.1109/ICDCSW. 2011.20.

[13] Y. Liu, Y. Wang, and J. Zhang, "New Machine Learning Algorithm: Random Forest," in Information Computing and Applications, vol. 7473, B. Liu, M. Ma, and J. Chang, Eds., in Lecture Notes in Computer Science, vol. 7473. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 246–252. https://doi.org/10.1007/978-3-642-34062-8_32.

[14] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. https://doi.org/ 10.1145/2939672. 2939785.

[15] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree".

[16] A. Drouin, G. Letarte, F. Raymond, M. Marchand, J. Corbeil, and F. Laviolette, "Interpretable Genotype-to-Phenotype Classifiers with Performance Guarantees," Sci. Rep., Vol. 9, No. 1, p. 4071, Mar. 2019, https://doi.org/10.1038/s41598-019-40561-2.

[17] J. J. Davis et al., "Antimicrobial Resistance Prediction in PATRIC and RAST," Sci. Rep., vol. 6, no. 1, p. 27930, Jun. 2016, https://doi.org/ 10.1038/srep27930.

[18] B. P. Alcock et al., "CARD 2023: Expanded Curation, Support for Machine Learning, and Resistome Prediction at the Comprehensive Antibiotic Resistance Database," Nucleic Acids Res., vol. 51, no. D1, pp. D690–D699, Jan. 2023, https://doi.org/10.1093/nar/gkac920.

[19] D. C. Hooper and G. A. Jacoby, "Mechanisms of Drug Resistance: Quinolone Resistance," Ann. N. Y. Acad. Sci., vol. 1354, no. 1, pp. 12–31, 2015, https://doi.org/10.1111/nyas.12830.

[20] R. P. Ambler, "The Structure of β-lactamases," Philos. Trans. R. Soc. Lond. B Biol. Sci., vol. 289, no. 1036, pp. 321–331, May 1980, https://doi.org/10.1098/rstb.1980.0049.

[21] M. S. Ramirez, M. E. Tolmasky, "Aminoglycoside Modifying Enzymes," Drug Resist. Updat., vol. 13, no. 6, pp. 151–171, Dec. 2010, https://doi.org/10.1016/j.drup.2010.08. 003.

[22] M. Jaillard et al., "A Fast and Agnostic Method for Bacterial Genome-Wide Association Studies: Bridging the Gap Between k-mers and Genetic Events," PLOS Genet., vol. 14, no. 11, p. e1007758, Nov. 2018, https://doi.org/ 10.1371/journal.pgen.1007758.

[23] J. A. Lees et al., "Fast and Flexible Bacterial Genomic Epidemiology with PopPUNK," Genome Res., vol. 29, no. 2, pp. 304–316, Feb. 2019, https://doi.org/10.1101/gr.241455.118.