# ANTIBIOGRAM PREDICTION THROUGH GENOME ANALYSIS

***Project Reference No.:*** *49S_BE_0112*

*College : Dayananda Sagar College of Engineering, Bengaluru*
*Branch : Department Of Artificial Intelligence & Machine Learning*
*Guide(S) : Prof. Ensteih Silvia*
*Student(S) : Mr. Karthik S*
*Mr. Pradhyuman Singh Shekhawat*

**Keywords:**

Antimicrobial Resistance (AMR), Whole-Genome Sequencing (WGS), k-mer Feature Engineering, Gradient Boosting (XGBoost), DNABERT Transformer

**Introduction:**

Antimicrobial resistance (AMR) caused 1.27 million deaths directly in 2019, with 4.95 million more where it was a contributing factor; cumulative projections to 2050 reach 39 million deaths. The World Bank estimates unchecked AMR could reduce global GDP by 1.1–3.8% by 2050, adding over USD 1 trillion annually to healthcare costs.

The core diagnostic problem is time. Culture-based susceptibility testing has not changed since the 1960s — it requires 48–72 hours that septic patients do not have, forcing clinicians to prescribe broad-spectrum antibiotics empirically and inadvertently accelerating resistance evolution.

Whole-genome sequencing can decode a bacterium's complete resistome within hours, but rule-based tools like ResFinder silently miss any variant not yet catalogued — novel beta-lactamase mutations, silenced efflux operons, porin truncations conferring carbapenem resistance without any acquired gene. Machine learning eliminates this blind spot by learning genotype-to-phenotype mappings empirically. This project implements a dual-pathway ML framework — XGBoost on alignment-free k-mer vectors for rapid clinical diagnostics, and fine-tuned DNABERT for novel pathogen generalisation — trained on 3,204 curated bacterial genomes from BV-BRC, with Qdrant vector similarity search enabling resistance inference for organisms with no prior testing history.

**Objectives:**

1. Design an alignment-free genomic feature engineering pipeline using canonical k=10 k-mer frequency vectors, enabling resistance prediction without the reference-alignment bottleneck.
2. Develop and comparatively evaluate XGBoost (78 antibiotics) and DNABERT (110 antibiotics) as complementary prediction pathways for clinical diagnostics versus research/discovery use cases.
3. Implement Qdrant-based HNSW vector similarity search to infer antibiogram profiles for novel or poorly-characterised pathogens with no prior resistance history.

4. Integrate Explainable AI (SHAP for XGBoost; attention rollup for DNABERT) to provide clinically auditable genomic attribution for each resistance prediction.
5. Deploy a full-stack production system — FastAPI backend with a React+TypeScript frontend — enabling real-time antibiogram prediction within clinical decision timelines.

**Methodology:**

Bacterial genome sequences are collected from publicly available repositories and paired with their laboratory-verified antimicrobial resistance phenotypes. The raw FASTA sequences are passed through a quality control pipeline that removes records with incomplete resistance labels or metadata, producing a curated dataset ready for feature extraction. Each genome is converted into an alignment-free k-mer frequency vector. Every overlapping substring of length k=10 is extracted, mapped to its canonical strand form, and counted. These counts are normalised by the total number of k-mers in the genome to remove length bias. K-mers that appear in very few genomes across the dataset are discarded as noise. The result is a high-dimensional sparse vector that numerically represents a genome without requiring alignment to any reference sequence. This vector representation feeds into two separate prediction pathways. In the first pathway, an XGBoost gradient boosting classifier is trained independently for each antibiotic, learning which k-mer patterns are associated with resistance or susceptibility. After training, SHAP analysis is applied to each classifier to identify which specific k-mer features drive each prediction, making the model output interpretable and traceable to known resistance regions of the genome.

In the second pathway, the DNABERT transformer model is fine-tuned on the same genomes. Sequences are tokenised into overlapping 6-mers and processed through BERT encoder layers, producing a 768-dimensional embedding vector that captures long-range sequence context across the genome. A classification head trained on top of this embedding predicts resistance for each antibiotic. These embeddings also serve a second purpose: they are stored in a Qdrant vector database indexed with HNSW, so that when a completely unknown genome is submitted, the system can find its closest genomic neighbours in the database and infer a probable resistance profile from their phenotypes. The full pipeline is wrapped in a production deployment: a FastAPI backend manages genome uploads, runs the prediction jobs asynchronously, and serves results through REST endpoints. A React frontend provides the user interface, allowing clinicians or researchers to upload a genome file, select a prediction model, and view the resulting antibiogram alongside SHAP-based explanations of which genomic features influenced the prediction.

**Result and Conclusion:**

XGBoost achieved 82.0% accuracy and 67.3% F1 across 78 antibiotics, outperforming the Preethi et al. CNN-RNN baseline (78.1%) by 3.9 points in 33 minutes of training; DNABERT achieved 73.5% accuracy and 50.0% F1 across 110 antibiotics in 17.2 hours. Inference runs at 14 seconds per genome for XGBoost and 8–12 minutes for DNABERT on a GTX 1650 Ti.

| Model | Accuracy | F1 | Jaccard | Train Time | Antibiotics |
|---|---|---|---|---|---|
| XGBoost (k=10) | 82.0% | 67.3% | 59.3% | 33 min | 78 |
| DNABERT | 73.5% | 50.0% | 40.6% | 17.2 h | 110 |
| CNN-RNN [Preethi et al.] | 78.1% | 80.2% | — | N/A | — |
| | | | | | |

XGBoost hit 100% accuracy on 10 antibiotics including vancomycin and cefazolin, with strong results on methicillin (97.4%), isoniazid (93.2%), and kanamycin (92.4%). DNABERT outperformed XGBoost only on colistin (94.8% vs. 84.6%), where long-range mcr-1 regulatory context exceeds k-mer resolution. SHAP confirmed XGBoost learns class-specific signatures across aminoglycosides, beta-lactams, and fluoroquinolones traceable to known resistance loci.

In conclusion, k-mer features with gradient boosting deliver optimal accuracy, speed, and interpretability at ~3K-genome scale. DNABERT's value is generalisation to phylogenetically novel organisms — a capability XGBoost cannot replicate. Together they form a complete diagnostic pipeline.

**References:**


[1] Antimicrobial Resistance Collaborators, "Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis," *The Lancet*, vol. 399, no. 10325, pp. 629–655, 2022.

[2] J. I. Kim et al., "Machine Learning for Antimicrobial Resistance Prediction: Current Practice, Limitations, and Clinical Perspective," *Clinical Microbiology Reviews*, vol. 35, no. 3, 2022.

[3] B. P. Alcock et al., "BV-BRC: Bacterial and Viral Bioinformatics Resource Center," *Nucleic Acids Res.*, 2023.

[4] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD*, 2016, pp. 785–794.

[5] Y. Ji et al., "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome," *Bioinformatics*, vol. 37, no. 15, pp. 2112–2120, 2021.

[6] Y. Bussi, R. Kapon, and Z. Reich, "Large-scale k-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy," *PLoS One*, vol. 16, no. 10, e0258693, 2021.

[7] M. Nguyen et al., "Predicting antimicrobial resistance using conserved genes," *PLoS Comput. Biol.*, vol. 16, no. 10, e1008319, 2020.

[8] R. Preethi, R. Bharati, and S. Priya, "Predicting Antibiotic Resistance from Genomic Sequences Using a Hybrid CNN-RNN Model," in *Proc. 3rd Int. Conf. AICECS*, 2024.

[9] R. Aytan-Aktug et al., "Prediction of acquired antimicrobial resistance for multiple bacterial species using neural networks," *mSystems*, vol. 8, no. 3, e00774-22, 2023.

**Future Scope:**

The future scope of this project includes:

1. Dataset scaling: The current 3,204-genome dataset constrains DNABERT to the data-starvation regime. Expanding to 10,000+ phenotyped genomes from BV-BRC and DRIAMS — particularly for WHO priority pathogens (carbapenem-resistant A. baumannii, NDM-1-carrying K. pneumoniae) currently underrepresented in AMR surveillance collections — is the single highest-impact improvement.
2. Resistome-specific pre-training: DNABERT's pre-training on GRCh38 (a eukaryotic human genome) means its transferred representations have zero overlap with bacterial resistome sequences. Pre-training a BERT encoder on a curated corpus of CARD resistance gene families, IntegronFinder cassette outputs, and plasmid replicon backbones from BV-BRC would substantially improve fine-tuning convergence and generalisation.
3. Long-context architectures: HyenaDNA's million-token context window would allow the model to encode a complete class-1 integron or Tn4401 carbapenem-resistance transposon in a single forward pass — a qualitative improvement over DNABERT's 512-token window that fragments multi-gene resistance cassettes across independent segments with no shared attention.