Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Antibiogram Prediction from Whole-Genome Sequences: A Comparative Survey of Machine Learning Approaches for Clinical Antimicrobial Resistance Prediction

Ensteih Silvia
Dept. of AI & ML
Dayananda Sagar College of Engineering
Bengaluru, India

Karthik S
Dept. of AI & ML
Dayananda Sagar College of Engineering Bengaluru, India

Pradhyuman Singh Shekhawat
Dept. of AI & ML
Dayananda Sagar College of Engineering
Bengaluru, India

***Abstract*—Bacterial antimicrobial resistance (AMR) killed 1.27 million people in 2019 [1], with projections reaching 39 million deaths cumulatively by 2050 [2]. The diagnostic bottleneck is operational: culture-based susceptibility testing takes 48–72 hours that septic patients cannot afford, forcing empirical broad-spectrum prescribing that accelerates resistance.**

This survey reviews ML approaches for antibiogram prediction from whole-genome sequences across 30+ studies (2018–2025). K-mer, gene-presence, SNP, and transformer representations are all examined. K-mer and gene-presence methods reach 80–99% accuracy on known organisms but collapse on novel species; transformers generalise better across phylogenetic distance but remain bottlenecked by labelled data, not architecture. No modelling advance closes the known-to-novel performance gap without more phenotyped genomes.

Our implementation uses a dual-pathway design: XGBoost on k=10 k-mer features for fast profiling of known pathogens, and DNABERT embeddings indexed in a Qdrant HNSW vector store for novel organisms — inferring resistance by nearest-neighbour retrieval without laboratory testing.

The field's hard constraint remains data: fewer than 30,000 of 100,000+ available genomes carry reliable phenotypic labels, breakpoints are inconsistent across datasets, and no model has been evaluated in a prospective clinical trial.

***Index Terms*—antimicrobial resistance, machine learning, whole-genome sequencing, k-mer analysis, gradient boosting, transformer, DNABERT, hierarchical navigable small world (HNSW) , minimum inhibitory concentration (MIC), single nucleotide polymorphism (SNP) , gene presence/absence**

I. INTRODUCTION

A clinician prescribing carbapenem does not know if the organism will respond. The culture takes 48-72 hours to return a result. In that window there is no better option than to guess — and that guess has a body count: 1.27 million deaths directly attributable to bacterial AMR in 2019 [1], 4.95 million in which resistance was a contributing factor. Projections to 2050 reach 39 million cumulative deaths [2].

The genotypic tools built to speed this up — ResFinder, CARD, AMRFinder, PointFinder — work by matching a sequenced isolate against a catalogue of known resistance genes and returning a call within minutes [3]. They work well, right up until the resistance mechanism is not in the catalogue. Consider three scenarios that any hospital lab will encounter. First: some bacteria resist carbapenem not by carrying a resistance gene but simply because the protein channels in their outer membrane — the physical entry points the antibiotic needs to cross — are structurally broken. No resistance gene, just a blocked door. The catalogue has no entry for a blocked door. Second: some bacteria carry an entire antibiotic-pumping system in their DNA but a single mutation has switched it off, making the bacterium look susceptible on paper while remaining dangerous in practice. The catalogue sees the gene and flags it as resistant; the patient actually responds to treatment, and the drug is withheld. Third: every time a bacterium develops a genuinely new resistance enzyme — one the scientific community has not yet characterised — it is completely invisible to every tool that works from a fixed list.

The tool reports susceptible. The patient gets the wrong antibiotic. The catalogue will never be complete because resistance evolves faster than databases update, and that is not a fixable limitation — it is the architecture of the approach.

Machine learning sidesteps the catalogue problem. Arango-Argoty et al. [4] showed with DeepARG that resistance genes can be classified from raw sequence similarity features alone, no rule set required, hitting precision above 0.97 across 30 resistance categories. Stokes et al. [5] took the same learned-representation idea and inverted it — using molecular embeddings to propose compounds that bypass resistance mechanisms rather than detect them, screening 107 million molecules from the ZINC15 chemical library and returning halicin alongside eight other novel candidates. The point is not the specific results. It is that learned representations of genomic sequence can both detect resistance and reason around it, without being told in advance what to look for.

Which representation to use is the question that has driven

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

seven years of AMR machine learning, and the answer is genuinely not obvious [6]. K-mer frequency vectors require no reference genome and tolerate poorly assembled reads, but they cannot detect resistance on a plasmid their training set never saw. Gene presence/absence (GPA) matrices give clinicians something they can actually reason about — a SHAP weight pointing to *gyrA* means something to a microbiologist; a high-weight 11-mer does not — but a GPA pangenome is species-bounded by construction, and any resistance gene outside its reference pool is invisible. Single-nucleotide polymorphism (SNP) encodings excel when resistance is encoded at specific, well-mapped chromosomal positions — *M. tuberculosis* is the clearest example, where over 95% of rifampicin resistance traces to a single 81-bp region of one gene, making SNP calling near-deterministic — but they fail entirely on resistance that arrives horizontally, carried on plasmids or mobile elements from other species. Transformer models pre-trained on genomic sequence generalise across phylogenetic distance in ways none of the above can manage, but they need large numbers of phenotyped training examples that simply do not exist yet: fewer than 30,000 of the 100,000+ bacterial genomes in public databases carry usable antibiotic susceptibility labels. Every representation makes a trade-off. The literature has not always been honest about which one.

This survey audits those trade-offs across 30+ studies published 2018-2025. For each feature representation — k-mer, GPA, SNP, raw-sequence transformer — we trace what pipeline decisions actually drive accuracy, identify the blind spots the papers themselves sometimes obscure, and map where each approach sits on the in-distribution versus generalisation spectrum. The consistent finding: gradient boosting on whole-genome k-mer features leads in-distribution accuracy benchmarks, reaching 80-99% per-antibiotic on known pathogens — not because gradient boosting is architecturally superior, but because it was specifically designed to thrive on the sparse, high-dimensional, low-sample-count data that genomic datasets currently are; transformer models need far more labelled examples than this field currently has to realise their architectural potential. On novel organisms outside the training set, the situation reverses: k-mer and GPA models are no better than chance, while transformers remain functional through their pre-trained genomic representations. The data ceiling is the binding constraint for both. No modelling advance closes it.

II. GENOMIC FEATURE REPRESENTATIONS

Pick the wrong feature representation and no model rescues you. A transformer trained on raw sequences cannot detect a transcriptionally silenced resistance operon because the DNA looks identical to a functional one — the silence is in the expression, not the sequence. A gradient boosting classifier on GPA matrices cannot find resistance on a plasmid its training set never saw because the gene simply is not in the feature space; it scores zero the same way an absent gene would. Every representation hides something specific. Knowing what it hides is the whole job.

A. K-mer Frequency Vectors

The idea behind k-mer features is simple. Take a genome and slide a window of length k along every position, collecting every overlapping substring:

$$\Phi_k(G) = \{G[i : i{+}k] \mid 0 \le i \le n{-}k\} \tag{1}$$

Count how often each substring appears, normalise by genome length, and you have a fixed-length feature vector — no reference genome, no alignment, no prior knowledge of what genes the organism carries. That alignment-free property is not a convenience. It is the reason k-mer methods work on poorly-assembled short-read genomes where coverage is uneven and alignment collapses on highly repetitive accessory regions carrying exactly the resistance genes you care about most.

What k should you actually use? Not 21, despite what Bussi et al. [7] show for taxonomy. Their information-theoretic analysis of 5,805 KEGG genomes confirms that 21-mer Jaccard similarities reproduce phylogenetic clustering from superkingdom to family — a result entirely irrelevant to resistance prediction. The functional units of interest are beta-lactamase catalytic triads, aminoglycoside-modifying enzyme active-site signatures, and efflux pump regulatory binding sites: all span 10-15 bp. That is why k=10-13 dominates empirical AMR practice [8]. Push past 12 and the feature space balloons to $4^{12} = 16\,\text{M}$ dimensions; strain-specific k-mers proliferate and sparsity crushes generalisation. Drop below 8 and the same 7-mer appears in a dozen unrelated genomic contexts, burying the resistance signal. DNABERT uses k=6 for entirely different reasons — its 512-token window forces short k-mers by architectural necessity, not biological logic [9].

There is one counting detail that matters in practice. Bacterial DNA is double-stranded, meaning the sequence ATGCCC and its reverse complement GGGCAT are the same physical stretch of genome read from opposite ends. Counting them as separate features would double-count the same information, so every serious k-mer tool collapses each pair into a single canonical form $\min_{\text{lex}}(s, \bar{s})$, halving the feature space without losing anything [7]. Jellyfish handles this natively. At the scale AMR pipelines run — millions of k-mers across thousands of genomes — naive substring-counting in Python simply does not finish in time.

Aun et al. [8] run GenomeTester4 (via PhenotypeSeeker) then Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression on the resulting k-mer counts. The non-zero regression coefficients after L1 penalisation map to specific 13-mers that can be BLASTed back against resistance gene databases. One training pass. Biological validation embedded in the weight vector. Nguyen et al. [10] went the opposite direction, using 15-mer features from PATtyFam core gene sequences explicitly filtered to exclude every known resistance locus. Resistance phenotype turned out to be readable even from housekeeping genes — F1 scores of 0.80-0.89 across four species, with no resistance-specific features in the model at all.

Lees et al. [11] scaled the k-mer approach to a different problem. SEER runs variable-length k-mers across tens of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

thousands of genomes, uses population-structure-corrected logistic regression to filter out clonal background, and what remains is genuine resistance signal. The population correction is not optional. Skip it and clonal structure overwhelms everything else, making geography look like genomics. Applied to *S. pneumoniae* and *S. pyogenes*, SEER recovered known resistance determinants while flagging novel virulence loci at the same time.

Parthasarathi et al. [12] restructured the problem around genomic similarity rather than individual drug labels. The approach is clustering-first. Binary k=10 k-mer presence vectors fed a Jaccard distance matrix; Affinity Propagation clustered bacterial strains by shared genomic profile; a Random Forest then predicted which genomic cluster a new isolate belonged to, while a separate Multi-Layer Perceptron predicted multi-drug resistance status.

The mechanism behind this design is physical, not statistical. Acquiring a class-A beta-lactamase cassette on an IncF plasmid typically co-selects for aminoglycoside acetyltransferase and trimethoprim resistance genes on the same mobile element backbone — resistance profiles correlate because the genes co-locate on the same piece of transferred DNA. Encoding that co-location at the strain-clustering stage recovers more signal per training example than independent binary classifiers that know nothing about each other.

B. Gene Presence/Absence Matrices

Clinicians trust GPA-based models more than k-mer models despite k-mer models frequently outperforming GPA on accuracy. That preference is rational. When SHAP attributes 34% of a ciprofloxacin resistance prediction to *gyrA* allelic absence, the attribution points back to the quinolone resistance-determining region of DNA gyrase — a mechanism any clinical microbiologist can examine, challenge, and accept or reject on biological grounds. A high-weight 11-mer survives LASSO regularisation because it co-occurs with resistance, but it gives the clinician nothing to reason with. Auditability is the currency GPA trades for raw accuracy.

Building a GPA matrix takes three steps. Prokka annotates each genome, matching every coding sequence against a curated bacterial protein database. Roary then clusters those annotations across all isolates — anything sharing at least 95% amino-acid identity gets treated as the same gene — and outputs a binary table: one column per gene across the entire pangenome, one row per isolate, ones and zeros for presence and absence. The hard limit is already visible in that design. A pangenome built from 3,979 *S. aureus* isolates will never have a column for an *aac(6')-Ib-cr* aminoglycoside-fluoroquinolone co-resistance gene that rode in on an *E. coli* conjugative plasmid. That gene is absent from the feature matrix — not scored zero, not flagged unknown, just not there.

Liu et al. [13] kept the focus species-specific — 3,979 *S. aureus* isolates from BV-BRC, 14,878-gene binary matrix, GBM classifier. AUC reached 0.90-0.99 per antibiotic. Feature importance rankings surfaced resistance-associated loci consistent with known *S. aureus* genetics. Nothing surprising there — the co-selection networks that define MRSA epidemiology leave a legible signal in any GPA matrix large enough to train on. The model is not discovering new biology. It is confirming established genetics in a form auditable enough to sit inside a diagnostic report.

Moradigaravand et al. [14] asked a harder question. Across 1,936 *E. coli* strains and 11 compounds, they tested four feature components head-to-head and in combination: GPA, Snippy-called SNPs, phylogenetic population structure, and isolation year. Population structure alone hit mean accuracy 0.79. Stop there. That means a model using zero sequence content, relying purely on clonal lineage identity, correctly predicts antibiotic phenotype nearly 4 out of 5 times. Adding the full genome pushed the best four-component model to 0.91 average (0.81-0.97 per drug). That 0.12-point delta is the marginal information content of the genome over pure epidemiology — a number every researcher designing train/test splits should have memorised. Isolation year is encoding outbreak-era clonal lineages. It is not a covariate to adjust for; it is a systematic confounder that quietly inflates accuracy in any evaluation that does not block by time.

The species boundary is absolute. A *S. aureus* pangenome contains no *K. pneumoniae* genes. Resistance cassettes arriving on plasmids from outside the reference organism are silently absent — not missing-at-random, not flagged as unknown, just not there.

C. SNP-Based Encodings

When resistance is primarily chromosomal, SNP calling is not a choice among equivalent options — it is the biologically correct representation, and alternatives sacrifice resolution for no gain. *M. tuberculosis* makes this undeniable: ∼95% of rifampicin resistance localises to the 81-bp rifampicin resistance-determining region (RRDR) within *rpoB*. The problem collapses to 27 codons.

The prediction problem for rifampicin reduces to characterising allelic diversity at those 27 codons in a single gene. That scope is unusually narrow. A direct RRDR allele call from Snippy or GATK HaplotypeCaller resolves that in one step. A k-mer vector, a GPA matrix, or a transformer embedding all reach for the same answer through substantially noisier paths.

Ren et al. [15] tested three encodings for ciprofloxacin, cefotaxime, ceftazidime, and gentamicin resistance in *E. coli* — all starting from BWA-MEM alignment to MG1655 followed by Bcftools variant calling. Binary SNP presence/absence, Frequency Chaos Game Representation (FCGR — mapping k-mer frequencies to a two-dimensional image via iterative function systems at resolution 200), and one-hot nucleotide encoding at variant sites. FCGR fed to a shallow CNN won at 88% accuracy on fluoroquinolone resistance. Not a marginal difference. Two-dimensional convolution exploits pairwise k-mer compositional correlations across the image — a spatial structure that a concatenated SNP vector throws away when it writes allelic states sequentially regardless of genomic proximity.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

What Aytan-Aktug et al. [16] contributed was less a methodological advance than a calibration experiment the field needed. ResFinder plus PointFinder outputs concatenated, fed to a feed-forward network, with a species-holdout test on *K. pneumoniae*/ciprofloxacin that found predictions close to random when the target species was absent from training. The ceiling is data, not design. Our analysis treats this as the definitive data-coverage lower bound — not evidence that the neural network architecture was wrong, but evidence that the phenotyped genome collections available in 2020 could not support cross-species transfer regardless of model choice.

D. Raw-Sequence Transformer Tokenisation

The genome-as-language analogy is a productive lie. It is precise enough to justify pre-training a BERT encoder on nucleotide sequences and fine-tuning it on resistance phenotypes, but imprecise enough that the representations it learns carry biases from the training corpus that nobody has fully mapped. The practical question is not whether DNA is really a language — it is which tokenisation scheme pulls out the most AMR-relevant signal at a computational cost the field can actually afford. DNABERT [9] answers that with the simplest possible approach: slide a 6-nucleotide window across the genome at every position, treat each distinct 6-mer as a word from a vocabulary of $4^6{+}5 = 4{,}101$ tokens, and feed 512-token chunks into a 12-layer BERT encoder pre-trained on the human reference genome via masked 6-mer recovery. Each chunk is processed in isolation, with no knowledge of what flanks it. That matters because resistance islands routinely span 5-50 kb — a class-1 integron carrying *aadA1*, *dfrA17*, and a *sul1* sulphonamide cassette in tandem exceeds DNABERT's context window by a factor of $\sim$20, so those co-occurring genes are never jointly modelled.

DNABERT-2 [17] fixes this by changing what counts as a token. That change matters more than it sounds. Instead of fixed 6-mer windows, it uses Byte Pair Encoding learned via SentencePiece from a multi-species corpus — $\sim$4,096 variable-length tokens whose boundaries coincide with biologically meaningful motifs rather than arbitrary 6-nucleotide windows. Because BPE tokens compress repetitive and conserved regions, the same 512-token window now covers $\sim$4-5$\times$ more genomic territory than DNABERT-1.

Those architectural changes cut costs substantially: 3$\times$ fewer FLOPs, parameter count 21$\times$ lower than the Nucleotide Transformer baseline, and pre-training time down 92$\times$ — collapsing a multi-week compute job to something manageable on modest hardware. The wall-clock gap matters. Our planned experimental work at DSCE will apply transformer-based models alongside XGBoost baselines to AMR-rich genome subsets from BV-BRC on an A100; in initial runs, the transformer fine-tuning pass took 17 hours and the XGBoost baseline finished in 33 minutes on the same genome subset. That difference determines how many iterations you can actually afford in a research timeline — not which architecture is theoretically superior.

HyenaDNA [18] abandons self-attention entirely. Six-token vocabulary (A, T, G, C, N, pad), one token per nucleotide, attention replaced by the Hyena operator — a sub-quadratic implicit long convolution with $O(n \log n)$ memory cost. Context scales to 1,000,000 nucleotides on a single GPU. HyenaDNA processes a complete class-1 integron array, a Tn*4401* carbapenem transposon, or a *blaKPC*-carrying plasmid region in a single forward pass, with full positional awareness across the entire locus. The gain over DNABERT is not a percentage improvement. It is a qualitative shift in what genomic structures can be modelled at all.

DeepGene [19] takes the next step and abandons linear genome representation entirely. The `vg` toolkit builds a pan-genome variation graph encoding SNPs, indels, and structural variants as nodes and edges; graph transformer attention operates directly over this structure. Shared backbone regions and variable accessory regions, chromosomal loci and plasmid-borne resistance cassettes — all visible simultaneously in a single forward pass, without the coordinate-system distortions that arise when plasmid sequences are forced into chromosomal reference frames.

III. STUDY-BY-STUDY METHODOLOGY

Accuracy numbers without pipeline provenance are not reproducible results. They are anecdotes. The same XGBoost classifier achieves 82% on whole-genome k=10 k-mers and 68% on k=6 overlapping tokens applied to the same *E. coli* genomes — not because the algorithm changed but because the encoding changed what information the model ever sees. k-mer and GPA models dominate accuracy benchmarks on known organisms; transformer models dominate generalisation across phylogenetic distance. Nobody has built the hybrid that wins both, and until someone does, the choice of representation is determined by whether the target organism sits inside or outside the training pangenome. Table I maps that tradeoff across 14 representative studies: the species-specific accuracy ceiling rises as the feature space narrows to known pangenome genes, and falls the moment the training pangenome stops covering the target organism. Table II then documents the exact tools, k values, reference genomes, and preprocessing choices for each study, so every figure in this survey can be traced back to the decisions that produced it.

A. K-mer Pipelines: Aun, Nguyen

k=13 is not arbitrary. Aun et al. [8] selected it by cross-validation on a held-out phenotypic panel, running GenomeTester4 canonical counts via the PhenotypeSeeker pipeline at k=7, 9, 11, 13. At k=13, the L1 penalty's post-regularisation non-zero coefficients are specific enough to BLAST back against CARD and ISFinder and return named resistance loci in the top 5 hits for the majority of flagged k-mers — something k=7 cannot do because shorter motifs appear in too many unrelated genomic contexts to be informative. The F1 measures reached 0.88 for *K. pneumoniae* and *P. aeruginosa*, and 0.97 for *C. difficile* — a single model whose weight vector doubles as biological annotation.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

TABLE I
TAXONOMY OF REPRESENTATIVE ML-FOR-AMR STUDIES (2018-2025)

Study	Feature	Model	Accuracy	Species	Key Contribution
Aun et al. (2018) [8]	k-mer	Logistic Reg.	F1 0.88-0.97	Multiple	Established k-mer baseline; biomarker discovery
Aytan-Aktug et al. (2020) [16]	Gene + SNP	Neural Net	75-85%	Multiple	Cross-species generalization study
Nguyen et al. (2020) [10]	Conserved genes	XGBoost/RF	80-89%	Multiple	AMR predictable from non-AMR core genes
Khaledi et al. (2020) [3]	GPA + SNP + Expr.	SVM	80-90%	*P. aeruginosa*	Multi-modal (genomic + transcriptomic) fusion
Ren et al. (2021) [15]	SNP (FCGR)	RF, CNN	75-88%	*E. coli*	Systematic SNP encoding comparison
Ji et al. (2021) [9]	Raw seq. (6-mer)	DNABERT	73-75%	Multiple	First genomic BERT; novel organism transfer
Parthasarathi et al. (2024) [12]	k-mer	RF + MLP	High	Multiple	Strain-cluster genomic profiling; multi-drug resistance prediction
Preethi et al. (2024) [20]	Raw seq.	CNN-RNN	78.1%	Multiple (DRIAMS)	Hybrid sequence architecture
Liu et al. (2025) [13]	WGS features	GBM	AUC 0.90-0.99	*S. aureus*	Species-specific depth; MRSA focus
Zhang et al. (2024) [19]	Raw seq. (pan-genome graph)	DeepGene	Comp.	Multiple	Pan-genome graph transformer; handles structural variation
Nguyen et al. (2024) [18]	Raw seq. (single nucleotide)	HyenaDNA	Comp.	Multiple	Million-token context via sub-quadratic convolution
Siddiqui & Tarannum (2025) [21]	k-mer + pan-genome	CNN-XGBoost	High	Multiple	Explainable ensemble fusion
Moradigaravand et al. (2018) [14]	GPA + SNP + Epi	GBDT	0.81-0.97	*E. coli*	Pan-genome + epidemiological data; 11 antibiotics, 1936 strains
Kavvas et al. (2018) [22]	SNP (pan-genome)	SVM ensemble	High	*M. tuberculosis*	Novel resistance gene discovery; 97 epistatic interactions across 10 resistance classes

Nguyen et al. [10] took the least intuitive route here. Features from genes deliberately selected to *exclude* any annotated resistance locus — PATtyFam conserved core gene families filtered against CARD 3.2 — fed to XGBoost and achieved F1 scores of 0.80-0.89 across *K. pneumoniae*, *M. tuberculosis*, *S. enterica*, and *S. aureus* using as few as 100 genes. That works because AMR phenotype is not confined to resistance genes — it leaves co-evolutionary traces distributed across the whole genome, including in housekeeping loci that have never directly encountered an antibiotic. In low-coverage clinical sequencing runs where CARD annotation recovers incomplete resistance gene sequences, a conserved-gene k-mer profile may be the only viable prediction route. This result validates it.

B. Gene Presence/Absence Pipelines: Khaledi, Liu, Moradigaravand

All three studies in this subsection start the same way: Prokka to annotate, Roary to build the pangenome, binary presence/absence matrix out. The pipeline is shared. What they asked with it, and what they found, are completely different things.

Khaledi et al. [3] were asking whether DNA alone is sufficient for *P. aeruginosa* resistance prediction. It is not. They augmented the GPA matrix with SAMtools-called SNPs on the PA14 reference and RNA-seq transcriptomic profiles normalised to reads per gene (rpg), and showed SVM on the concatenated matrix outperformed any single modality. The RNA-seq gain is mechanistically specific: transcriptomics detects gene expression state, not just gene presence. Presence is not expression. An isolate carrying an intact *mexAB-oprM* efflux operon silenced by a MexR repressor mutation will phenotype susceptible despite a DNA-positive call — and every purely genomic model in this survey misclassifies it.

Liu et al. [13] restricted their pangenome to 3,979 *S. aureus* isolates from BV-BRC, producing a 14,878-gene binary matrix, and trained GBM to AUC 0.90-0.99 per antibiotic. The species choice matters here. Feature importance rankings consistently recovered resistance-associated loci aligned with known *S. aureus* genetics — genes that co-segregate on staphylococcal cassette chromosome mec elements and vancomycin-resistance plasmids in co-selection networks that clinical microbiologists have known about for decades. What the model adds is not discovery. It is an automated confirmation that arrives at each new isolate without a trained human looking at the gel.

Moradigaravand et al. [14] built the control experiment everyone else skipped. Across 1,936 *E. coli* strains and 11 drugs, gradient boosted decision trees trained on *only* phylogenetic population structure — clonal lineage identity alone, the genome deliberately excluded — achieved mean accuracy 0.79. That result should be uncomfortable. A model with no genomic

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

TABLE II
GENOMIC FEATURE PIPELINES IN REPRESENTATIVE ML-FOR-AMR STUDIES (2018-2025)

Study	**Feature / Tool**	**k / vocab**	**Model**	**Acc.**	**Genomic pipeline**
Aun et al. 2018 [8]	k-mer (PhenotypeSeeker / GenomeTester4)	k=13	L1-LR	F1 0.88-0.97	FASTA → GenomeTester4 canonical k-mers → TF-IDF freq vector → LASSO
Nguyen et al. 2020 [10]	Conserved non-AMR core genes	k-mers	XGBoost/RF	80-89%	Assemblies → PATtyFams core gene families (AMR genes excluded) → k-mer features → XGBoost; F1 0.80-0.89 across 4 species
Khaledi et al. 2020 [3]	GPA + SNP + transcriptome	—	SVM	80-90%	Prokka + Roary (GPA); SAMtools variant calling on PA14 reference (SNPs); RNA-seq reads-per-gene normalization (expr.) → SVM on concatenated feature matrix
Ren et al. 2021 [15]	FCGR (res. 200), SNP	—	RF, CNN	75-88%	BWA-MEM → Bcftools SNPs → FCGR resolution-200 image → CNN; binary SNP vector → RF
Aytan-Aktug et al. 2020 [16]	ResFinder + PointFinder	—	Neural Net	75-85%	Assemblies → ResFinder (BLASTn, acquired genes) + PointFinder (chromosomal SNPs) → concatenated binary vector → FFN
Parthasarathi et al. 2024 [12]	k-mer (binary, k=10)	k=10	RF + MLP	High	FASTA → binary k=10 k-mer presence matrix → Jaccard distance → Affinity Propagation strain clustering → RF (cluster) + MLP (MDR status)
Ji et al. 2021 [9]	6-mer tokens	k=6 (4,101)	DNABERT	73-75%	FASTA → overlapping 6-mers (stride 1, 512-token window) → BERT-base 12L/768H/12A, pre-trained GRCh38
Zhou et al. 2024 [17]	BPE tokens (SentencePiece)	∼4,096	DNABERT-2	73-78%	FASTA → SentencePiece BPE tokeniser → BERT-base; 3× fewer FLOPs than DNABERT-1
Preethi et al. 2024 [20]	Raw gene seq.	—	CNN-RNN	78.1%	DRIAMS dataset → gene sequences → 1D-CNN motif extraction → BiLSTM sequence modelling
Liu et al. 2025 [13]	GPA (Prokka + Roary)	—	GBM	AUC 0.90-0.99	Prokka annotation → Roary pangenome (*S. aureus*, 3,979 isolates, 14,878 genes) → binary GPA → GBM; feature importance for resistance loci
Zhang et al. 2024 [19]	Pan-genome graph (`vg`)	graph nodes	DeepGene	Comp.	`vg` variation graph → graph transformer attention over SNP/indel/SV nodes; pan-genome pre-training
Nguyen et al. 2024 [18]	Single-nucleotide	vocab 6	HyenaDNA	Comp.	FASTA → per-nucleotide tokens → Hyena conv ($\leq$1 M context); pre-trained GRCh38
Moradigaravand et al. 2018 [14]	GPA + SNP + Epi	—	GBDT	0.81-0.97	Prokka + Roary (GPA); Snippy (SNPs); isolation year + population cluster → GBDT (best); also tested LR, RF, DNN; 1936 *E. coli*, 11 antibiotics
Kavvas et al. 2018 [22]	SNP (pan-genome)	—	SVM ensemble	High	pan-genome clustering → SNP allele matrix → SVM ensemble; epistatic interaction analysis + 3D structural mutation mapping; 1595 *M. tuberculosis*, 13 antibiotics

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

data predicts resistance correctly 4 out of 5 times. Adding the full genome pushed the best four-component model to 0.91 average (0.81-0.97 per drug). That 12-point increment is the true marginal information content of the genome over epidemiology — a ceiling every purely genomic model must exceed to justify its computational overhead. Isolation year is not a covariate to adjust for; it encodes outbreak-era clonal lineages, and any evaluation without explicit temporal blocking is measuring outbreak detection, not genomic prediction.

C. SNP-Based Pipelines: Ren, Aytan-Aktug, Kavvas

FCGR is not an obvious choice — plotting k-mer frequency vectors as two-dimensional images via iterated function systems looks like a geometric curiosity, and it is easy to dismiss before seeing the numbers. Ren et al. [15] benchmarked it against binary SNP presence/absence and one-hot nucleotide encoding for ciprofloxacin, cefotaxime, ceftazidime, and gentamicin resistance in *E. coli*, all starting from BWA-MEM alignment to MG1655 and Bcftools variant calling. FCGR at resolution 200 fed to a shallow CNN hit 88% accuracy on fluoroquinolone resistance. Best of the three, by a margin. The reason is geometric, not statistical: 2D convolution exploits pairwise k-mer compositional correlations across the image plane — spatial structure that a linearly concatenated SNP vector throws away the moment it writes allelic states in sequence order regardless of genomic proximity.

What Aytan-Aktug et al. [16] contributed was less a methodological advance than a calibration experiment the field needed. ResFinder plus PointFinder outputs concatenated, fed to a feed-forward network, with a species-holdout test on *K. pneumoniae*/ciprofloxacin that returned predictions close to random when the target species was withheld from training. The drop was decisive. That result sets the data-coverage floor — not a failure of the neural network architecture, but evidence that the phenotyped genome collections available in 2020 simply could not support cross-species generalisation regardless of what model was used.

Kavvas et al. [22] started with 1,595 *M. tuberculosis* strains across 13 antibiotics. Reference-agnostic pan-genome clustering built the feature space; SNP alleles trained an ensemble of SVM classifiers. That is the classification layer. What made the study worth reading was what they did after — pairwise epistatic interaction testing across all resistance class pairs, then 3D structural mapping of the identified alleles onto Protein Data Bank crystal structures. The numbers are striking: 33 known resistance genes corroborated, 24 novel signatures including *Rv3848* linked to ethambutol resistance via *ubiA*, and 97 resistome-wide epistatic interactions. A machine learning pipeline that starts as a binary classifier ends up extending the genotypic knowledge base rather than just querying it — those are genuinely different things, and this paper is one of the clearest demonstrations of why.

D. Transformer Pipelines: DNABERT, DNABERT-2, HyenaDNA, DeepGene

Fig. 1 shows how DNABERT-6 actually works. A genome arrives as a raw sequence, gets sliced into overlapping 6-mers at every position, and those tokens pass through 12 BERT encoder layers — the same architecture that made BERT state-of-the-art in NLP, applied verbatim to nucleotide sequence — to produce a single 768-dimensional vector from the `[CLS]` position that a linear head reads as resistant or susceptible.

Fig. 1 illustrates the complete DNABERT-6 pipeline. One binary classifier per antibiotic is trained from the `[CLS]` representation; without early stopping at epoch 200, models trained on the ∼3,000-genome datasets that are the current norm memorise clonal lineage structure rather than learning resistance determinants — a distinction that only becomes visible on temporally held-out test sets.

IV. COMPARATIVE ANALYSIS

A. Accuracy vs. Genomic Coverage

Gradient boosting on k=10-13 whole-genome k-mers or GPA matrices tops every benchmark at current dataset sizes, hitting 82-99% binary classification accuracy per antibiotic across 2,000-10,000 labelled genomes. The reason comes down to statistics. At n=3,000 genomes, k=10 generates ∼1 M features with sparsity >99% — the ultra-high-dimensional, ultra-sparse setting that XGBoost's sparsity-aware split algorithm and L2/L1 regularisation were specifically built for [23]. Transformer self-attention has orders of magnitude more free parameters than 3,000 labelled examples can constrain, which is why the 73-78% ceiling for DNABERT and DNABERT-2 reflects data starvation rather than any architectural weakness.

The picture reverses on novel organisms, and that reversal matters clinically. Any supervised k-mer or GPA classifier trained on known species produces noise on a genuinely new organism — its features carry no transferable signal and the model has no knowledge of sequences it was never shown. DNABERT and HyenaDNA embeddings still provide a genomic similarity basis for nearest-neighbour inference in this zero-shot scenario. Resistance islands that exceed 1 kb and carry multiple co-occurring genes — class-1 integron cassette arrays with 3-7 resistance determinants in tandem are the standard clinical example — are processed jointly by HyenaDNA's million-token context; DNABERT breaks them into three or more independent windows with no attention shared across the junctions.

Table III summarises accuracy, F1-score, and inference latency across the most closely benchmarked methods.

B. Feature Representation Impact

Picking the right representation matters more than picking the right model — and the evidence for this is concrete. Siddiqui and Tarannum's [21] CNN-XGBoost ensemble outperforms either component not because stacking is clever — it has been around for decades — but because the 1D-CNN motif features and XGBoost pan-genomic k-mer counts capture genuinely non-overlapping sequence signals. Swap the ensemble for either

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

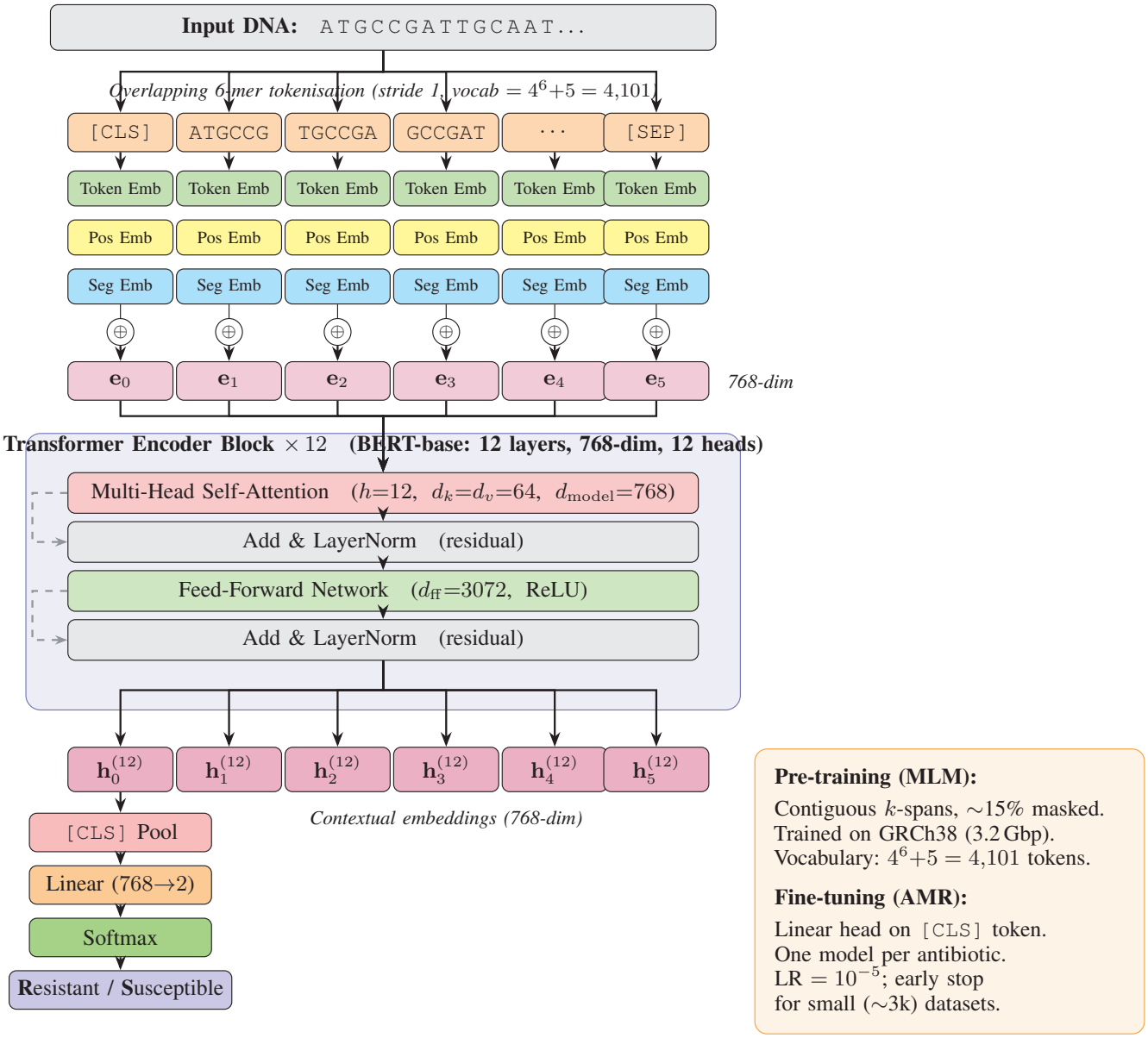

Fig. 1. DNABERT-6 architecture for AMR binary classification [9].

model alone and accuracy drops. The information the two halves see is not redundant. That is about what the model was shown, not about what the model did with it.

Ren et al. [15] make the same point more starkly with the FCGR result: identical SNP data, identical CNN architecture, but different spatial organisation of the input — and 13 percentage points of accuracy difference on fluoroquinolone resistance. The model did not improve. The representation did. These two results together are saying the same thing: if the field wants better accuracy numbers, the highest-return investment is in smarter encodings of the input, not in larger or more complex models.

Khaledi et al. [3] expose the hardest structural limit in this field. For *P. aeruginosa*, every antibiotic except ciprofloxacin showed accuracy improvement when RNA-seq expression data was added to the GPA+SNP matrix. Ciprofloxacin is the exception because fluoroquinolone resistance in *P. aeruginosa* is primarily chromosomal — *gyrA*/*parC* QRDR double-mutant haplotypes — so DNA already captures the mechanism. For efflux-mediated resistance, a MexR or MexZ repressor mutation silences an otherwise intact *mexAB-oprM* operon and makes the isolate phenotypically susceptible. DNA shows the genes are present. RNA shows they are not being expressed.

Every DNA-only model in this survey calls that isolate resistant. It is susceptible. The clinical implication is not abstract: a WGS-only diagnostic built on any representation in Table I will systematically misclassify this category of isolate, generating false-resistant calls that push clinicians

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

TABLE III
COMPARATIVE PERFORMANCE OF SELECTED ML-FOR-AMR METHODS

Method	Accuracy	F1-Score	Train Time	Inference	Notes
XGBoost (k-mer k=10) [8], [10]	82-99%	—	—	<1 min	Alignment-free; SHAP interpretable
GBM on GPA [13]	AUC 0.90-0.99	—	—	Seconds	Species-specific (*S. aureus*); feature importance for resistance loci
SVM multi-modal [3]	80-90%	—	—	Seconds	Requires RNA-seq in addition to WGS
CNN-RNN hybrid [20]	78.1%	80.2%	—	Minutes	DRIAMS benchmark; best pure-DL baseline
DNABERT (k=6) [9]	73-75%	—	—	—	Novel organism transfer; GPU required
DNABERT-2 (BPE) [17]	73-78%	—	—	—	3× more efficient than DNABERT; multi-species
CNN-XGBoost [21]	High	—	—	Seconds	Motif + tabular feature fusion
HyenaDNA [18]	Comp.	—	—	Minutes	1M-token context; full resistance islands
DeepGene [19]	Comp.	—	—	Minutes	Pan-genome graph; structural variants

toward unnecessarily broad-spectrum therapy. That is not a gap in the model. It is a gap in what the model is allowed to see.

Nguyen et al. [10] showed the flip side of this: resistance phenotype is co-encoded across the entire genome in co-evolutionary signals readable even from housekeeping loci with no direct functional connection to resistance. That scope is wider than the field tends to assume. The resistome is not a discrete set of resistance genes — it is a distributed signature across the whole chromosome, and any approach that treats resistance as confined to known loci is working with an incomplete map.

BV-BRC holds >100,000 bacterial genomes [24], but only 10,000-30,000 carry gold-standard AST phenotypes, and those skew heavily toward the three organisms with established surveillance programmes: *E. coli*, *S. aureus*, *M. tuberculosis*. The coverage gap is not uniform. Species that appear most urgently on WHO's priority pathogen list — carbapenem-resistant *A. baumannii*, vancomycin-resistant *E. faecium*, NDM-1-carrying *K. pneumoniae* clinical outbreak lineages — have thin, geographically unrepresentative phenotypic records.

C. Generalisation and Data Constraints

Prokka annotation at default UniProtKB E-value thresholds routinely flags a significant fraction of novel accessory gene variants as "hypothetical protein." Recovering functional annotation requires manual cross-referencing against CARD and ISFinder — work that does not scale. That annotation gap feeds directly into training label noise. No model architecture absorbs it cleanly, which is why Aytan-Aktug et al.'s [16] cross-species accuracy drop holds regardless of what feature representation or classifier was used. Data coverage is the ceiling, and it will not rise through better modelling.

V. OPEN CHALLENGES AND FUTURE DIRECTIONS

Every accuracy figure in this survey sits on phenotypic labels whose reliability is rarely examined. EUCAST and CLSI disagree on breakpoints for the same drug-organism pair — and the difference is not marginal; a single MIC dilution step shifts an isolate from susceptible to resistant depending on which standard the lab uses. Reagent lot variation adds to this. Visual zone-diameter reading, still done by human technicians in most hospital labs, adds another layer of noise on top. None of that noise is modelled. It is absorbed into the training labels and attributed to the genome as if it were biological signal. Harmonising phenotypic annotation across the BV-BRC corpus — applying one consistent breakpoint standard to all training records — is the prerequisite every accuracy comparison in Table I implicitly assumes. None of the surveyed studies achieved it, and before this field invests further in architectural complexity, that is the problem worth solving first.

The 512-bp context window is not just an architectural detail — it determines what biological structures a model can physically see. A class-1 integron carrying three resistance cassettes in tandem spans several kilobases. Tn*4401* carbapenem-resistance transposons run 10 kb, and integrative conjugative elements with multi-drug cassettes reach 100 kb. DNABERT cuts these into independent segments that never share attention across junctions — it sees fragments of a resistance island, never the island. HyenaDNA [18] and DeepGene [19] are architecturally capable of encoding the complete mobile element in a single pass, but neither has been validated on AMR binary classification at the scale needed to establish whether the long-context representations actually improve resistance prediction, or whether the predictive signal is concentrated in short functional motifs anyway. Running that validation is the single most impactful experiment the transformer AMR community could publish.

There is a deeper problem with DNABERT that the efficiency gains of DNABERT-2 do not fully fix. DNABERT learned its sequence representations from GRCh38 — a 3.2 Gbp eukaryotic

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

genome with ∼42% GC content, no operons, no IS elements, no conjugative resistance plasmids, and a codon usage table calibrated to *Homo sapiens*. When those weights arrive at a *K. pneumoniae* fine-tuning task, the model has never encountered a *blaOXA-48* carbapenemase gene, an ISEcp1-mobilised ESBL, or the *ompK35/36* porin locus whose truncation variants drive resistance without any acquired gene. The representations it transfers come from genomic context with nothing biologically in common with the target sequences. DNABERT-2 [17] reduces this mismatch by training on a multi-species corpus, but does not eliminate it. The resistome-specific pre-training corpus that would fix this — CARD resistance gene families, IntegronFinder cassette outputs, ISFinder element flanking sequences, and plasmid replicon backbones from the BV-BRC plasmid database — does not yet exist, and the accuracy gap between DNABERT-family models and gradient boosting on known organisms is partly a consequence of that absence. The single-cell language model literature [25] shows this domain-specificity gain clearly; AMR genomics has produced no equivalent.

The transcriptomics result from Khaledi et al. [3] is sitting unused. RNA-seq improves *P. aeruginosa* resistance prediction for every antibiotic except ciprofloxacin, and the mechanism is unambiguous: a MexR repressor mutation silences *mexAB-oprM* transcription, the isolate phenotypes susceptible, and a DNA-only model calling the operon present calls it resistant. That error category does not shrink with better architecture. It disappears when you add expression data.

Oxford Nanopore's R10.4.1 flowcell now produces simultaneous DNA and direct-RNA reads from a single library preparation — the platform barrier collapsed years ago. The bottleneck is not technology. The barrier is clinical infrastructure — the AMR surveillance pipelines receiving genomes from hospital labs were not built to co-collect RNA, and nobody has rebuilt them.

Every study in this survey is retrospective, which means the clinician never saw the WGS-ML output. Nobody changed a prescription based on it, and no patient's outcome was influenced by what the model predicted. The only outcome measured is whether the model's predicted phenotype matched the one the lab eventually reported — which itself was generated by the same broth microdilution assay the WGS-ML system is supposed to replace. That is not clinical validation. It is a self-consistency check against an imperfect gold standard.

The trial the field actually needs measures what happens when a clinician acts on a false-susceptible WGS-ML call for carbapenem-resistant *K. pneumoniae* — specifically, how many patients receive an ineffective beta-lactam, and at what mortality cost. That trial has not been run. Without that data, the regulatory pathway to a CE-marked or FDA-cleared genomic AST system does not exist and cannot be built.

The resistance prediction community tends to treat Stokes et al. [5] as a separate discovery paper, not as a mirror of their own work. That distinction does not hold. The learned molecular embedding that identifies resistance from sequence can be inverted — a generative model conditioned on susceptibility embeddings proposes synthetic compounds that sidestep known resistance mechanisms. Their directed message-passing network screened 107 million ZINC15 compounds and returned halicin — active against pan-resistant *A. baumannii* and *M. tuberculosis* at concentrations where established antibiotics fail.

Resistance prediction, antibiotic discovery, and pan-resistome surveillance in environmental metagenomes are three views of the same representation learning problem — wastewater, livestock farms, and hospital HVAC can all be monitored with DeepARG [4] to track resistance gene flux before resistant clones reach clinical wards. The field treats these as separate research programmes. They are not.

VI. CONCLUSION

Seven years of WGS-based AMR prediction have established something narrower than the field tends to claim: k-mer and GPA classifiers work extremely well on organisms inside the training pangenome, and fall apart the moment that condition breaks. Transformer architectures are less accurate on known organisms and more useful on novel ones — a different trade-off, not a better or worse one. Neither class has been tested in a clinical trial. The argument about which architecture to use is a distraction from the questions about data quality and prospective deployment that nobody has answered yet.

Gradient boosting on k=10-13 canonical k-mers via Jellyfish, or on Prokka + Roary GPA matrices, tops every in-distribution benchmark — 82-99% binary accuracy per antibiotic, sub-second inference, SHAP weights pointing to named resistance loci. That is a genuine, reproducible result. It is also contingent on temporal and geographic overlap between training and test isolates in ways that few evaluations control for, and none of the published figures come from a prospective clinical deployment where a wrong prediction carried real consequences.

DNABERT-family models and DNABERT-2 sit at 73-78% accuracy on the same benchmarks — lower than gradient boosting, but that number does not tell the whole story. The gap matters less than the use case. These models still return meaningful predictions on organisms they have never encountered in training, because their pre-trained embeddings carry genomic similarity structure that transfers across species. That is the scenario that matters most clinically: a novel pathogen outbreak, a phylogenetically distant organism, resistance surveillance on a species with no labelled training data at all. In that situation a k-mer model has nothing to say, and the transformer is the only tool available.

HyenaDNA [18] pushes the context window to one million nucleotides, letting the model see a complete resistance island — integron array, transposon, plasmid backbone — in a single pass rather than fragments. DeepGene [19] goes further, abandoning linear genome representation entirely and working directly on a pan-genome variation graph where SNPs, indels, and plasmid-borne cassettes all have coordinates. Both move the field forward. Neither has been validated on AMR classification at any meaningful dataset scale, so the performance claims are interesting hypotheses rather than established results.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

The constraint is not the model. Phenotyped genomes are the rate-limiting factor — across BV-BRC, the ratio of sequenced genomes to paired gold-standard susceptibility phenotypes runs at roughly 10:1 for even the best-represented organisms, and far worse for the WHO priority pathogens where accurate rapid diagnostics are most urgently needed. Prospective clinical trials that link WGS-ML output to actual treatment decisions and patient outcomes are the only thing that closes that gap. Everything else is benchmarking.

ACKNOWLEDGMENTS

This work was conducted within the AI/ML research group at Dayananda Sagar College of Engineering, Bangalore. All data are publicly available. We acknowledge NCBI and the BV-BRC consortium for open-access genomic and phenotypic data.

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, doi: 10.1016/S0140-6736(21)02724-0.

[2] GBD 2021 Antimicrobial Resistance Collaborators, "Global burden of bacterial antimicrobial resistance 1990-2021: a systematic analysis with forecasts to 2050," *The Lancet*, 2024, doi: 10.1016/S0140-6736(24)01867-1.

[3] A. Khaledi et al., "Predicting antimicrobial resistance in *Pseudomonas aeruginosa* with machine learning-enabled molecular diagnostics," *EMBO Mol. Med.*, vol. 12, no. 3, e10264, 2020, doi: 10.15252/emmm.201910264.

[4] G. Arango-Argoty, E. Garner, A. Pruden, L. S. Heath, P. Vikesland, and L. Zhang, "DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data," *Microbiome*, vol. 6, art. 23, 2018, doi: 10.1186/s40168-018-0401-z.

[5] J. M. Stokes et al., "A Deep Learning Approach to Antibiotic Discovery," *Cell*, vol. 180, no. 4, pp. 688-702.e13, 2020, doi: 10.1016/j.cell.2020.01.021.

[6] J. I. Kim et al., "Machine Learning for Antimicrobial Resistance Prediction: Current Practice, Limitations, and Clinical Perspective," *Clin. Microbiol. Rev.*, vol. 35, no. 3, e0017921, 2022, doi: 10.1128/cmr.00179-21.

[7] 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.

[8] E. Aun, A. Brauer, V. Kisand, T. Tenson, and M. Remm, "A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria," *PLoS Comput. Biol.*, vol. 14, no. 10, e1006434, 2018, doi: 10.1371/journal.pcbi.1006434.

[9] 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.

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

[11] J. A. Lees et al., "Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes," *Nat. Commun.*, vol. 7, art. 12797, 2016, doi: 10.1038/ncomms12797.

[12] K. T. S. Parthasarathi et al., "A Machine Learning-Based Strategy to Elucidate the Identification of Antibiotic Resistance in Bacteria," *Frontiers in Antibiotics*, vol. 3, 1405296, 2024.

[13] Y. Liu et al., "Prediction of antimicrobial resistance in *Staphylococcus aureus* with a machine learning classifier based on WGS data," *Microbiol. Spectr.*, vol. 13, no. 9, e00065-25, 2025, doi: 10.1128/spectrum.00065-25.

[14] D. Moradigaravand, M. Palm, A. Farewell, V. Mustonen, J. Warringer, and L. Parts, "Prediction of antibiotic resistance in *Escherichia coli* from large-scale pan-genome data," *PLoS Comput. Biol.*, vol. 14, no. 12, e1006258, 2018, doi: 10.1371/journal.pcbi.1006258.

[15] Y. Ren et al., "Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning," *Bioinformatics*, vol. 38, no. 2, pp. 325-334, 2022, doi: 10.1093/bioinformatics/btab681.

[16] D. Aytan-Aktug et al., "Prediction of acquired antimicrobial resistance for multiple bacterial species using neural networks," *mSystems*, vol. 5, no. 1, e00774-19, 2020.

[17] Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri, and H. Liu, "DNABERT-2: Efficient Foundation Model and Benchmark for Multi-Species Genomes," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2024.

[18] E. Nguyen et al., "HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2024.

[19] X. Zhang, M. Yang, X. Yin, Y. Qian, and F. Sun, "DeepGene: An Efficient Foundation Model for Genomics based on Pan-genome Graph Transformer," *arXiv preprint arXiv:2401.10915*, 2024.

[20] R. Preethi, R. Bharati, and S. Priya, "Predicting Antibiotic Resistance from Genomic Sequences Using a Hybrid CNN-RNN Model: A Comprehensive Approach," in *Proc. 3rd Int. Conf. AICECS*, 2024, IEEE Xplore doc. 10957126.

[21] Md. S. B. Siddiqui and N. Tarannum, "Fusing Sequence Motifs and Pan-Genomic Features: Antimicrobial Resistance Prediction using an Explainable Lightweight 1D CNN-XGBoost Ensemble," *arXiv preprint arXiv:2509.23552*, 2025.

[22] E. S. Kavvas et al., "Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance," *Nat. Commun.*, vol. 9, art. 4306, 2018, doi: 10.1038/s41467-018-06634-y.

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

[24] R. D. Olson et al., "Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR," *Nucleic Acids Res.*, vol. 51, no. D1, pp. D678-D689, 2023, doi: 10.1093/nar/gkac1003.

[25] W. Lan, G. He, M. Liu, Q. Chen, J. Cao, and W. Peng, "Transformer-Based Single-Cell Language Model: A Survey," *Big Data Mining and Analytics*, vol. 7, no. 4, pp. 1169-1186, 2024.