Acta Scientific Microbiology (ISSN: 2581-3226)

Volume 9 Issue 4 April 2026

Review Article

Machine Learning Approaches in Microbial Genomics for Pathogen Identification and Antimicrobial Resistance Prediction

Emmanuel Nkansah[1], Micheal Abimbola Oladosu[2]*, Moses Adondua Abah[3], Abimbola Mary Oluwajembola[2], Fwangmun Ezekiel Gushit[4], Olaide Ayokunmi Oladosu[5], Adesola Esther Adeneye[6] and Bukola Oluwaseyi Olufosoye[7]

[1]Department of Accounting, Economics and Finance, School of Business, La Sierra University, Riverside, CA, USA

[2]Department of Chemical Sciences, Faculty of Science, Anchor University, Ayobo, Ipaja, Lagos, Nigeria

[3]Department of Biochemistry, Faculty of Pure and Applied Sciences, Federal University of Wukari, Wukari, Taraba State, Nigeria

[4]Department of Public Health, Faculty of Health Science, Ahmadu Bello University, Zaria, Kaduna State, Nigeria

[5]Department of Computer Science, Faculty of Science and Technology, Babcock University, Ilishan, Nigeria

[6]Department of Biological Sciences, Faculty of Science, Anchor University, Ayobo, Ipaja, Lagos, Nigeria

[7]Department of Medical Microbiology, Faculty of Medical Laboratory Sciences, Ambrose Alli University, Ekpoma, Edo State, Nigeria

***Corresponding Author:** Micheal Abimbola Oladosu, Department of Chemical Sciences, Faculty of Science, Anchor University, Ayobo, Ipaja, Lagos, Nigeria.

DOI: 10.31080/ASMI.2026.09.1592

Received: November 30, 2020

Published: March 31, 2026

Abstract

The emergence and rapid spread of antimicrobial resistance (AMR) pose a critical threat to global public health, necessitating innovative approaches for pathogen identification and resistance prediction. Machine learning (ML) has revolutionised microbial genomics by enabling rapid, accurate analysis of vast genomic datasets to predict AMR phenotypes and identify pathogens. This review examines recent advances in ML applications for microbial genomics, focusing on supervised and unsupervised learning algorithms, deep learning architectures, and their integration with whole-genome sequencing (WGS) data. We discuss the performance of various ML models, including random forests, support vector machines, convolutional neural networks, and ensemble methods in predicting antimicrobial resistance across different bacterial species. The review highlights challenges in model interpretability, data quality, and clinical implementation while exploring emerging trends in federated learning and transfer learning approaches. Understanding these computational methodologies is essential for developing rapid diagnostic tools and informing antimicrobial stewardship programs in clinical and pharmaceutical settings.

Keywords: Machine Learning; Antimicrobial Resistance; Pathogen Identification; Genomics; Whole-Genome Sequencing; Deep Learning

Introduction

Antimicrobial resistance represents one of the most pressing challenges in modern medicine and pharmaceutical sciences. The World Health Organisation estimates that bacterial AMR was directly responsible for 1.27 million deaths globally in 2019, with projections suggesting this number could rise dramatically without intervention [1]. Traditional culture-based methods for pathogen identification and antimicrobial susceptibility testing (AST) require 24-72 hours, delaying appropriate treatment and contributing to AMR development through empirical antibiotic use [2].

The advent of next-generation sequencing technologies has generated unprecedented volumes of microbial genomic data, creating both opportunities and challenges for pathogen surveillance and resistance prediction. Whole-genome sequencing can identify resistance genes, virulence factors, and phylogenetic relationships, but manual analysis of these complex datasets is time-consuming and requires specialised expertise [3]. Machine learning has emerged as a transformative approach to extract actionable insights from genomic data, offering the potential for rapid, accurate prediction of antimicrobial resistance phenotypes directly from genotypic information [4].

Recent years have witnessed remarkable progress in applying ML algorithms to microbial genomics, with models achieving accuracy rates exceeding 95% for specific pathogen-antibiotic combinations [5]. These computational approaches leverage diverse genomic features, including single-nucleotide polymorphisms (SNPs), k-mers, gene presence/absence patterns, and pangenome variations to construct predictive models [6]. The integration of ML with genomic epidemiology has enhanced outbreak detection, transmission tracking, and prediction of resistance emergence, providing valuable tools for public health surveillance and pharmaceutical development [7].

This review synthesises current knowledge on ML applications in microbial genomics, examining algorithmic approaches, performance metrics, and clinical translation challenges. We analyse the comparative effectiveness of different ML architectures, discuss data preprocessing and feature engineering strategies, and explore future directions, including explainable AI and integration with other omics data. Understanding these technologies is crucial for microbiologists, pharmaceutical scientists, and clinicians working to combat antimicrobial resistance.

Machine learning fundamentals in genomic analysis

Supervised learning approaches

Supervised learning algorithms form the foundation of most AMR prediction models, utilizing labelled training data where genotypes are paired with known resistance phenotypes. Random forest (RF) classifiers have demonstrated particular success in genomic applications due to their ability to handle high-dimensional data, capture non-linear relationships, and provide feature importance metrics [8]. RF models aggregate predictions from multiple decision trees, each trained on bootstrap samples of the data, reducing overfitting while maintaining predictive accuracy [9].

Support vector machines (SVMs) have been extensively applied in microbial genomics, particularly for binary classification tasks, such as distinguishing between susceptible and resistant phenotypes. SVMs identify optimal hyperplanes that maximise the margin between classes in high-dimensional feature spaces, often employing kernel functions to handle non-linearly separable data [10]. Studies have shown that SVM models with radial basis function kernels achieve competitive performance for AMR prediction in *Mycobacterium tuberculosis* and *Staphylococcus aureus* [11].

Gradient boosting methods, including XGBoost and LightGBM, have gained prominence in recent genomic studies due to their superior handling of imbalanced datasets and ability to capture complex interaction effects between genomic features [12]. These ensemble methods sequentially build decision trees, with each iteration focusing on correcting errors from previous models, resulting in highly accurate predictions for diverse bacterial species and antibiotic classes [13]. Table 1 shows the comparison of machine learning algorithms for antimicrobial resistance prediction.

Deep learning architectures

Deep learning represents a paradigm shift in ML-based genomic analysis, utilising artificial neural networks with multiple hidden layers to automatically learn hierarchical feature representations from raw data. Convolutional neural networks (CNNs) have been adapted from image recognition tasks to analyse genomic sequences, treating DNA sequences as one-dimensional signals where convolution operations identify motifs and patterns

Algorithm	Advantages	Disadvantages	Typical Accuracy Range	Computational Cost	Dataset Size (n)	Dataset Source
Random Forest	Handles high-dimensional data; provides feature importance; resistant to overfitting	Requires substantial memory; less interpretable with many trees	85-95%	Moderate	n = 3,000–15,000	PATRIC, Clinical isolates
Support Vector Machine	Effective in high-dimensional spaces; versatile kernel functions	Computationally intensive for large datasets; sensitive to parameter tuning	83-93%	High	n = 1,000–8,000	NCBI, CARD
Gradient Boosting (XGBoost)	Superior performance on imbalanced data; handles missing values	Prone to overfitting with improper tuning; longer training time	87-96%	High	n = 5,000–20,000	Multi-centre surveillance
Convolutional Neural Network	Automatic feature learning; captures sequence motifs	Requires large training datasets; computationally expensive; limited interpretability	84-94%	Very High	n = 10,000–50,000	Public databases (WGS)
Logistic Regression	Highly interpretable; fast training	Assumes linear relationships; limited performance on complex patterns	78-88%	Low	n = 500–5,000	Single-centre studies

Table 1: Comparison of Machine Learning Algorithms for Antimicrobial Resistance Prediction.

Notes: Accuracy ranges represent performance across multiple studies with varying bacterial species, antibiotics, and validation approaches. Dataset sizes indicate typical training set sizes used in published studies. Cross-validation methods varied: most studies employed 10-fold or 5-fold cross-validation, while some used leave-one-out cross-validation or external test set validation. Performance metrics are for binary classification (resistant/susceptible); multi-class prediction typically shows 5–15% lower accuracy. External validation on independent datasets generally revealed 3–10% performance degradation compared to cross-validation estimates, highlighting the importance of rigorous evaluation protocols.

Sources: Data compiled from references [8–19].

associated with resistance [14]. CNN architectures can process raw nucleotide sequences without extensive feature engineering, learning relevant genomic signatures directly from training data [15].

Recurrent neural networks (RNNs), particularly long short-term memory (LSTM) networks, excel at capturing sequential dependencies in genomic data, making them suitable for analysing gene order effects and regulatory relationships [16]. Recent studies have employed bidirectional LSTM architectures to predict AMR phenotypes from assembled genome sequences, achieving accuracies comparable to or exceeding traditional ML methods while providing insights into positional genomic features [17].

Attention mechanisms and transformer architectures, originally developed for natural language processing, have recently been applied to microbial genomics with promising results [18]. These models can identify relevant genomic regions without relying on predefined features, potentially uncovering novel resistance mechanisms that are not captured by existing

databases. Deep learning models require substantial training data and computational resources but offer advantages in handling complex, multi-drug resistance patterns and identifying epistatic interactions between resistance genes [19]. Figure 1 presents the machine learning workflow for antimicrobial resistance prediction from genomic data.

Transformer-based architectures have shown particular promise in recent genomic applications due to their self-attention mechanisms that enable modelling of long-range dependencies without recurrent connections. DNABERT, a BERT-based model pre-trained on genomic sequences, has demonstrated superior performance in tasks such as promoter prediction, splice site identification, and transcription factor binding site recognition when fine-tuned for AMR gene prediction [20]. Evolutionary Scale Modeling (ESM) and similar protein language models have been adapted for analysing resistance-conferring protein sequences, achieving state-of-the-art results in predicting functional effects of mutations in beta-lactamases and other resistance enzymes [21].

BioGPT and genomic foundation models represent an emerging paradigm where large-scale pre-training on diverse genomic datasets enables transfer learning to specific AMR prediction tasks with limited labelled data [22]. These models can process both nucleotide and amino acid sequences, potentially identifying novel resistance mechanisms through unsupervised pattern recognition. Recent studies have applied transformer architectures to metagenomic AMR gene detection, where the ability to capture contextual information across long DNA fragments improves identification of resistance genes in complex microbial communities [18].

Despite their advantages, transformer models present significant computational challenges, typically requiring GPU clusters and substantial training time. The data requirements for effective transformer training often exceed what is available for many pathogen-antibiotic combinations. However, the rapid development of more efficient architectures and the availability of pre-trained genomic models suggest that transformers will play an increasingly important role in AMR prediction as computational resources become more accessible and genomic databases continue to expand [20,21].

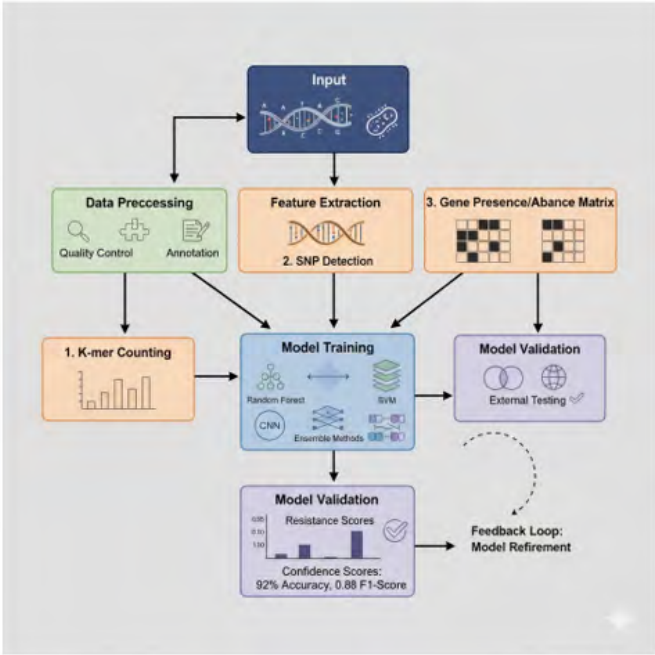

Figure 1: Machine Learning Workflow for Antimicrobial Resistance Prediction from Genomic Data. This workflow illustrates the complete machine learning pipeline for AMR prediction from raw sequencing data. The process begins with DNA extraction and whole-genome sequencing, followed by quality control steps including adapter trimming and quality score filtering. Genome assembly uses de Bruijn graph-based assemblers for short reads or overlap-layout-consensus approaches for long reads. Annotation identifies genes, resistance determinants, and mobile genetic elements using tools like Prokka and AMRFinderPlus. Feature engineering extracts informative genomic signatures including SNPs, k-mers, gene presence/absence patterns, and pangenome variations. The machine learning model training phase employs various algorithms including random forests, support vector machines, gradient boosting, and deep learning architectures, with hyperparameter optimisation through cross-validation. Model validation assesses performance using accuracy, sensitivity, specificity, and F1-scores on independent test sets. Finally, clinical implementation integrates predictions into laboratory information systems for treatment guidance, with continuous model updating as new resistance data becomes available.

Sources: Adapted from methodologies described in [4,8,14].

Applications in pathogen identification

Species and strain-level classification

Machine learning algorithms have revolutionised bacterial identification by enabling rapid species and strain-level classification directly from sequencing data. K-mer-based approaches, which decompose genomic sequences into overlapping substrings of length k, provide alignment-free methods for taxonomic classification when combined with ML classifiers [23]. Random forest and neural network models trained on k-mer frequency profiles can accurately identify bacterial species from short-read sequencing data within minutes, significantly faster than traditional BLAST-based approaches [24].

Metagenomic applications present unique challenges due to the complexity of mixed microbial communities in clinical samples. Deep learning models, particularly CNNs, have demonstrated superior performance in taxonomic classification of metagenomic reads compared to conventional bioinformatics tools [25]. These models can simultaneously identify multiple pathogens in polymicrobial infections, a common scenario in respiratory and wound infections where accurate pathogen identification is critical for treatment decisions [26].

Whole-genome sequencing combined with ML has enhanced outbreak investigation by providing high-resolution strain typing and phylogenetic analysis. Core genome multilocus sequence typing (cgMLST) data analysed through clustering algorithms and dimensionality reduction techniques enable rapid identification of clonally related isolates during hospital outbreaks [27]. Machine learning models can integrate genomic data with epidemiological information to predict transmission networks and identify likely sources of infection, informing infection control measures [28]. Table 2 shows the performance metrics of ML models for pathogen Identification across different Bacterial Species.

Benchmark datasets and data resources

The development and evaluation of ML models for AMR prediction critically depends on the availability of high-quality, well-annotated genomic datasets paired with phenotypic resistance information. Several major public databases have emerged as essential resources for training and benchmarking predictive models, each with distinct characteristics, strengths, and limitations.

- **PATRIC and BV-BRC:** The Pathosystems Resource Integration Center (PATRIC), now succeeded by the Bacterial and Viral Bioinformatics Resource Center (BV-BRC), is one of the most comprehensive databases for bacterial genomics and AMR data [30]. BV-BRC contains over 500,000 bacterial genomes with associated AMR phenotype data for multiple antibiotics, derived from both literature curation and direct submissions from surveillance programs. The database provides standardised annotations for resistance genes using the Antibiotic Resistance Ontology (ARO). However, phenotype quality varies significantly across entries, with some isolates having comprehensive AST results for 20+ antibiotics while others have minimal testing data.
- **NCBI Pathogen Detection:** The NCBI Pathogen Detection Database focuses on outbreak surveillance and contains clustered genome assemblies with associated metadata including isolation source, geographic location, and antimicrobial susceptibility profiles where available. This database excels in temporal coverage of foodborne and healthcare-associated pathogens, making it valuable for studying AMR evolution over time. Limitations include inconsistent antimicrobial testing protocols across contributing laboratories and sparse phenotype data for environmental isolates.
- **EMBL-EBI Resources:** The European Nucleotide Archive (ENA) at EMBL-EBI serves as a primary repository for sequence data linked to phenotypic information through the European Genome-phenome Archive (EGA). These resources are particularly strong for isolates from European surveillance programs, including data from the European Antimicrobial Resistance Surveillance Network (EARS-Net). The standardised phenotype reporting following EUCAST guidelines provides high-quality training data, though geographic representation heavily favours European clinical isolates.
- **CARD and ResFinder:** The Comprehensive Antibiotic Resistance Database (CARD) provides curated reference sequences for resistance genes and mutations, along with their associated resistance mechanisms and phenotypes [6]. ResFinder, maintained by the Technical University of Denmark, offers a curated database of acquired resistance genes and a web-based tool for resistance gene identification from WGS data. Associated databases including PointFinder

and PlasmidFinder provide complementary information on chromosomal mutations and mobile genetic elements. These resources are widely used for benchmarking ML models and annotating resistance determinants across diverse genomic studies.

- **Dataset Limitations and Benchmarking Considerations:** Several critical limitations affect all major databases. Geographic sampling bias remains pronounced, with isolates from North America and Europe dramatically overrepresented compared to those from Africa, South Asia, and South America, creating models that may underperform on strains circulating in underrepresented regions. Phenotypic testing methodology also varies across contributing laboratories, introducing noise through differences in AST methods, breakpoint interpretations, and quality control practices. Rare resistance mechanisms are systematically underrepresented, creating imbalanced datasets where ML models struggle with uncommon but clinically significant resistance profiles. For benchmarking, train-test splits should account for phylogenetic relatedness to avoid overoptimistic performance estimates; temporal validation, where models trained on historical data predict resistance in more recent isolates, better reflects real-world deployment scenarios than random cross-validation.

Virulence factor prediction

Beyond species identification, ML approaches facilitate the prediction of virulence factors and pathogenicity potential from genomic sequences. Ensemble learning methods combining multiple classifiers have been developed to identify genes encoding toxins, adhesins, and secretion systems that contribute to bacterial pathogenesis [31]. These predictive models assist pharmaceutical researchers in identifying therapeutic targets and guiding vaccine development by pinpointing conserved virulence factors across pathogenic strains [32]. Figure 2 presents the hierarchical clustering heatmap of virulence factors and antimicrobial resistance genes.

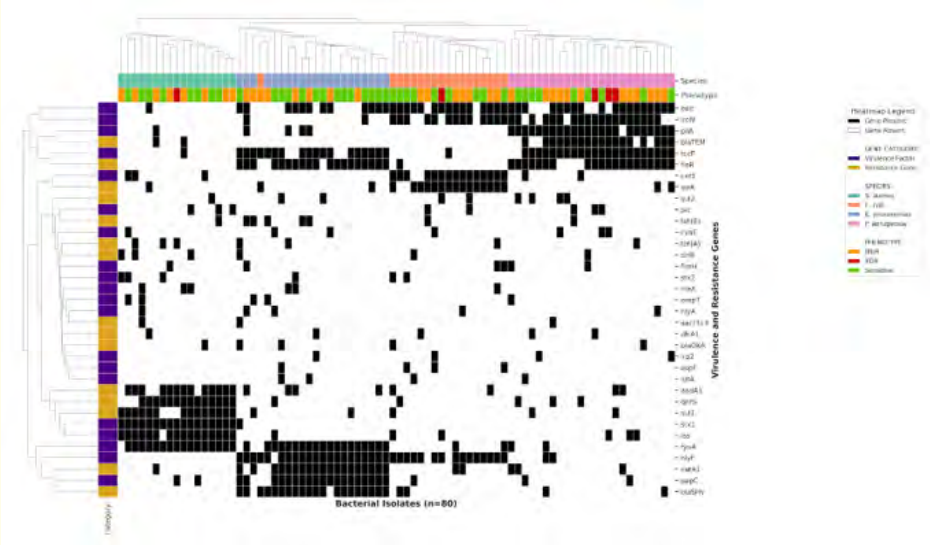

Figure 2: Hierarchical Clustering Heatmap of Virulence Factors and Antimicrobial Resistance Genes. This heatmap illustrates co-occurrence and clustering patterns of virulence factors and antimicrobial resistance genes across bacterial isolates. Rows represent individual genes and columns represent isolates. Colour intensity indicates gene presence/absence or expression level, with the hierarchical dendrogram revealing co-clustering of genes that frequently co-occur, suggesting shared mobile genetic elements or co-selection pressures. Ensemble learning methods combining multiple classifiers were used to identify and score these genetic features. The clustering reveals functionally related resistance and virulence gene clusters, supporting the hypothesis that AMR and virulence co-evolution is driven by shared horizontal gene transfer events, particularly plasmid-mediated dissemination in nosocomial pathogens.

Sources: Visualisation approach based on references [31,32].

Antimicrobial resistance prediction

Genotype-to-phenotype prediction models

The core application of ML in microbial genomics involves predicting antimicrobial resistance phenotypes from genomic data, a task with direct clinical relevance. Early models focused on detecting known resistance genes and mutations from curated databases, but modern ML approaches can identify novel genetic determinants and complex resistance mechanisms [30]. Feature engineering plays a crucial role, with studies comparing gene presence/absence matrices, k-mer frequencies, SNP profiles, and protein sequence embeddings as input features for ML classifiers [33].

Large-scale studies have demonstrated that ensemble ML models can predict resistance to specific antibiotics with area under the receiver operating characteristic curve (AUROC) values exceeding 0.95 for well-studied pathogen-antibiotic combinations such as *Escherichia coli* with fluoroquinolones and *S. aureus* with methicillin [29]. Performance varies considerably across antibiotics and species, with lower accuracy observed for resistance mechanisms involving efflux pumps, permeability changes, or complex regulatory networks not well-represented in training data [34]. Figure 3 shows the performance comparison of ML models across multiple antibiotics

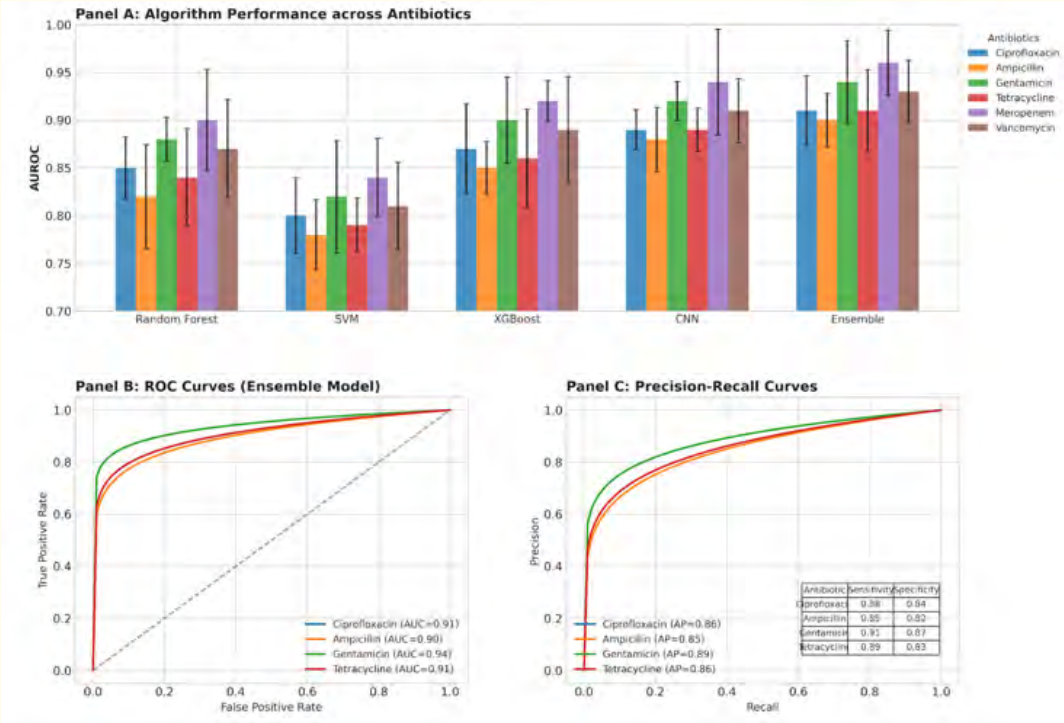

Figure 3: Performance Comparison of ML Models Across Multiple Antibiotics. This figure compares the predictive accuracy of major machine learning algorithms across six antibiotic classes for clinically relevant bacterial species. The x-axis represents AUROC values and accuracy percentages, while each grouped set of bars corresponds to a distinct ML algorithm. Ensemble methods such as gradient boosting and random forest generally outperform single classifiers, particularly for complex multi-gene resistance phenotypes such as carbapenem resistance and multi-drug resistance. Performance is notably higher for resistance mechanisms mediated by single well-characterised genes (e.g., mecA for methicillin resistance) compared to mechanisms involving efflux pumps, permeability changes, or regulatory networks. Simpler algorithms such as logistic regression offer greater interpretability but lower accuracy on complex phenotypes, illustrating the accuracy-interpretability trade-off inherent in clinical ML applications.

Sources: Performance data compiled from references [29,34,35].

Multi-drug resistance prediction

Predicting multi-drug resistance (MDR) phenotypes presents additional complexity due to the need for multi-label classification or regression approaches. Multi-output neural networks and multi-task learning frameworks have been developed to simultaneously predict resistance to multiple antibiotics, potentially capturing shared genetic mechanisms and improving prediction efficiency [35]. These models can identify pan-resistant organisms and predict resistance profiles across entire antibiotic classes, valuable information for antimicrobial stewardship programs [36].

Transfer learning approaches, where models pre-trained on well-characterised organisms are fine-tuned for species with limited phenotypic data, have shown promise in addressing data scarcity challenges [37]. This paradigm is particularly relevant for predicting resistance in fastidious organisms or for newly emerging resistance mechanisms where labelled training data is insufficient for traditional supervised learning [38]. Table 3 gives the antimicrobial resistance prediction performance by antibiotic class.

Antibiotic Class	Representative Drugs	Best Performing ML Model	Mean AUROC	Mean Accuracy (%)	Major Resistance Mechanisms	Reference
β-lactams	Penicillin, Cephalosporins	Random Forest	0.94	91.3	β-lactamase production, PBP modifications	[29]
Fluoroquinolones	Ciprofloxacin, Levofloxacin	Gradient Boosting	0.96	93.7	DNA gyrase mutations, efflux pumps	[13]
Aminoglycosides	Gentamicin, Tobramycin	CNN	0.92	89.5	Modifying enzymes, ribosomal mutations	[15]
Glycopeptides	Vancomycin	Ensemble (RF + XGBoost)	0.93	90.8	van gene clusters, cell wall modifications	[35]
Tetracyclines	Tetracycline, Doxycycline	SVM	0.91	88.2	tet genes, efflux systems	[11]
Carbapenems	Meropenem, Imipenem	Deep Neural Network	0.95	92.6	Carbapenemases, porin loss	[19]

Table 3: Antimicrobial Resistance Prediction Performance by Antibiotic Class.

AUROC = Area Under Receiver Operating Characteristic Curve.

Computational challenges and solutions

Data quality and preprocessing

The performance of ML models in microbial genomics critically depends on data quality, including sequencing depth, assembly completeness, and phenotypic testing accuracy. Contamination, incomplete genome assemblies, and heterogeneous phenotypic testing protocols across laboratories introduce noise that can degrade model performance [39]. Robust preprocessing pipelines incorporating quality filtering, genome assembly validation, and phenotypic data curation are essential for developing reliable predictive models [40].

Class imbalance represents a pervasive challenge in AMR prediction, as resistance phenotypes for certain antibiotic-pathogen combinations may be rare in training datasets. Techniques such as synthetic minority over-sampling (SMOTE), cost-sensitive learning, and ensemble methods with balanced bootstrapping have been employed to address imbalance and prevent models from defaulting to majority class predictions [41]. Additionally, geographical and temporal biases in genomic databases can limit model generalizability, necessitating careful validation on external datasets representing diverse populations and resistance epidemiology [42].

Model interpretability and clinical translation

While complex ML models, particularly deep learning architectures, often achieve superior predictive accuracy, their "black box" nature poses challenges for clinical acceptance and regulatory approval. Explainable AI techniques, including SHAP (Shapley Additive exPlanations) values and attention visualisation, provide insights into which genomic features drive predictions, enhancing trust and enabling biological interpretation [43]. Identifying predictive genomic features can also guide hypothesis-driven research to characterise novel resistance mechanisms and inform therapeutic development [44].

Clinical implementation requires not only accurate predictions but also consideration of computational infrastructure, turnaround time, and integration with laboratory workflows. Cloud-based platforms and containerised ML pipelines have been developed to facilitate the deployment of genomic prediction tools in resource-limited settings [45]. Regulatory frameworks for ML-based diagnostic devices are evolving, with agencies requiring prospective validation studies demonstrating clinical utility before approving genomic AMR prediction tools [46]. Figure 4 features the Important Analysis and Model Interpretability.

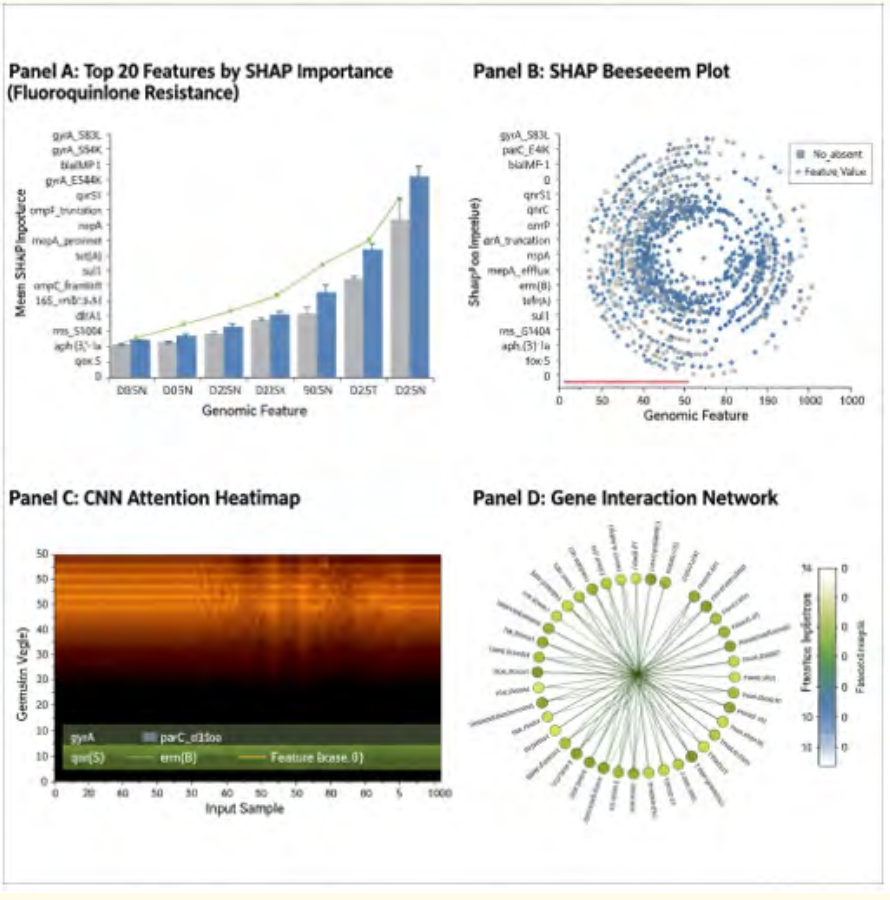

Figure 4: Feature Importance Analysis and Model Interpretability. This figure displays SHAP (Shapley Additive exPlanations) values and feature importance scores for the top genomic predictors of antimicrobial resistance across multiple ML models. Each bar represents the mean absolute SHAP value of a genomic feature, indicating its average contribution to model predictions. Features are ranked from most to least influential, revealing that known resistance genes (e.g., bla genes for beta-lactam resistance, gyrA mutations for fluoroquinolone resistance) dominate the top positions, validating the biological relevance of ML predictions. Colour coding distinguishes positive (resistance-promoting) from negative (susceptibility-associated) feature effects. This explainability analysis bridges the gap between black-box ML predictions and clinically actionable biological interpretation, enabling microbiologists to identify novel candidate resistance determinants and supporting regulatory acceptance of ML-based diagnostic tools.

Sources: Surveillance applications described in references [7,18].

Critical limitations in current approaches

While machine learning has demonstrated remarkable success in AMR prediction, several fundamental limitations constrain current methodologies and must be addressed for robust clinical implementation.

- **Overfitting in High-Dimensional Genomic Spaces:** The curse of dimensionality presents a pervasive challenge in genomic ML applications, where feature numbers often vastly exceed sample sizes [23]. A typical bacterial genome contains 3-5 million base pairs and 3,000-5,000 genes, yet training datasets may contain only hundreds or thousands of phenotypically characterised isolates for specific pathogen-antibiotic combinations [9]. This high-dimensional, low-sample-size regime creates substantial overfitting risk, where models learn training set idiosyncrasies rather than generalisable resistance patterns. K-mer-based approaches exacerbate this problem, as k = 10 generates approximately one million potential features, many representing spurious phylogenetic associations rather than causal resistance determinants [33]. Standard k-fold cross-validation provides limited overfitting protection when closely related strains appear in multiple folds, inflating performance estimates; phylogeny-aware splits are therefore essential for reliable evaluation [38]. Regularisation techniques including L1/L2 penalties, dropout in neural networks, and early stopping help control overfitting but require careful tuning [44]. External validation on completely independent datasets from different geographic regions or time periods provides the only reliable overfitting assessment, yet such validation is rarely performed and frequently reveals substantial performance degradation compared to cross-validation estimates [9,38].
- **Geographic and Demographic Biases:** Current AMR prediction models suffer from systematic geographic and demographic biases that limit their generalisability [42]. Training datasets predominantly comprise isolates from high-income countries, creating models optimised for resistance patterns that may not reflect strains circulating in low- and middle-income countries [39]. For example, carbapenem resistance in Klebsiella pneumoniae in North America is primarily mediated by KPC enzymes, while OXA-48-like carbapenemases dominate in many European and Middle Eastern countries, and NDM enzymes are most prevalent in South Asia [40]. Models trained on KPC-dominated datasets may underperform in regions where OXA-48 or NDM predominate. Socioeconomic factors further compound these biases, as isolates from high-income settings typically have comprehensive antimicrobial susceptibility testing across multiple drug classes while those from resource-limited settings may have minimal testing data [39]. Addressing these biases requires intentional efforts to expand genomic surveillance in underrepresented regions, develop regionally adapted models, and rigorously evaluate cross-geographic performance before clinical deployment [47].
- **Limitations of Genomic-Only Prediction Approaches:** A fundamental constraint of current genomic ML approaches is their inability to capture resistance phenotypes influenced by gene expression regulation, epigenetic modifications, and environmental factors [40]. While genomic sequence determines resistance potential, actual resistance expression depends on complex regulatory networks, growth conditions, and inter-bacterial interactions that sequence data alone cannot predict [8]. Efflux pump overexpression mediated through promoter mutations or regulatory gene disruption can confer clinically significant multi-drug resistance, yet identical mutations may produce variable phenotypes depending on strain background and growth phase [8]. Two-component regulatory systems, as seen in Pseudomonas aeruginosa MexAB-OprM regulation, can modulate resistance gene expression in response to environmental signals without any underlying sequence change [8]. Phase variation, epigenetic modifications, persister cell formation, and inoculum-dependent effects further limit the predictive power of purely sequence-based models [40]. These constraints suggest genomic ML models should be viewed as resistance potential predictors rather than definitive phenotype classifiers, and hybrid approaches combining rapid genomic prediction for initial treatment guidance with confirmatory phenotypic testing for complex cases represent the most clinically pragmatic implementation pathway [48].

Emerging trends and future directions

Integration of multi-omics data

Future advances in ML-based microbial genomics will likely involve integration of complementary data modalities, including transcriptomics, proteomics, and metabolomics. Multi-omics integration can capture dynamic aspects of resistance expression, including regulatory responses to antibiotic exposure that are not apparent from genomic data alone [48]. Graph neural networks and

other advanced architectures capable of modelling heterogeneous data types and their relationships represent promising avenues for multi-omics integration [49].

Federated learning and privacy-preserving approaches

As genomic data generation accelerates globally, federated learning frameworks that enable model training across distributed datasets without centralising sensitive patient data are gaining attention [50]. These privacy-preserving approaches allow institutions to collaboratively develop ML models while maintaining data sovereignty and compliance with data protection regulations. Federated learning could dramatically expand training data diversity, improving model generalizability across geographic regions and resistance epidemiology patterns [47].

sReal-Time surveillance and outbreak prediction

Integration of ML-based genomic analysis with real-time surveillance systems represents a frontier in public health microbiology. Streaming ML algorithms that continuously update predictions as new genomic and phenotypic data become available could provide early warning of emerging resistance mechanisms or outbreak clusters [18]. Combining genomic predictions with epidemiological data, antibiotic consumption patterns, and environmental factors through integrated ML frameworks may enable proactive interventions before resistance becomes widespread [7]. Figure 5 highlights the global distribution and temporal trends of ML-Predicted AMR.

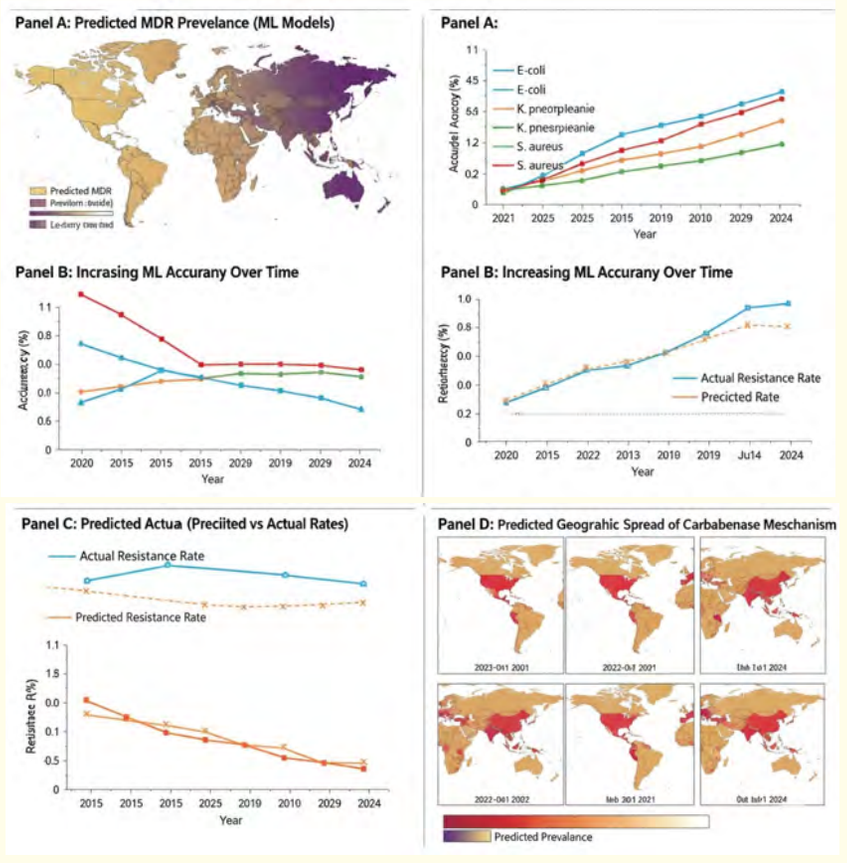

Figure 5: Global Distribution and Temporal Trends of ML-Predicted AMR. This figure depicts world maps and time-series plots illustrating ML-predicted antimicrobial resistance prevalence across geographic regions and over time. Heat-mapped global panels show country-level resistance predictions for key pathogen-antibiotic combinations, revealing pronounced geographic heterogeneity that reflects differences in antibiotic usage, infection control practices, and sampling intensity. Time-series plots demonstrate temporal trends in predicted resistance rates, with upward trajectories visible for carbapenem-resistant Enterobacteriaceae and extensively drug-resistant tuberculosis in multiple regions. Emerging trends such as federated learning for privacy-preserving multi-site model training, multi-omics data integration, and real-time streaming surveillance algorithms are highlighted as key future directions that will enhance the sensitivity and timeliness of ML-based AMR surveillance. The figure underscores the critical need for geographically representative training data and global data-sharing frameworks to enable equitable AMR prediction across all regions.

Sources: Peng C, Chen L, Wang Y., et al. [18]; Sundermann AJ., *et al.* [7]. Surveillance applications described in references [7,18].

Conclusion

Machine learning has fundamentally transformed microbial genomics, providing powerful tools for pathogen identification and antimicrobial resistance prediction. Supervised learning algorithms, including random forests, support vector machines, and gradient boosting methods have demonstrated robust performance across diverse bacterial species and antibiotics, while deep learning architectures offer potential for discovering novel resistance mechanisms and handling complex multi-drug resistance patterns. Despite impressive predictive accuracy in research settings, challenges remain in data quality, model interpretability, and clinical translation.

The pharmaceutical and clinical microbiology communities must address these challenges through collaborative efforts to generate high-quality, diverse training datasets, develop explainable AI approaches, and conduct prospective validation studies. Emerging technologies, including multi-omics integration, federated learning, and real-time surveillance systems, promise to further enhance the utility of ML in combating antimicrobial resistance. As these computational approaches mature, they will become indispensable tools in antimicrobial stewardship, pharmaceutical development, and public health surveillance, contributing to global efforts to preserve antibiotic effectiveness for future generations.

Acknowledgments

The authors acknowledge the contributions of researchers worldwide advancing AI applications in drug discovery and the open-source software community enabling these developments.

Conflict of Interest

The authors declare no conflicts of interest.

Funding

This review paper does not have any funding from external or internal organisations.

Bibliography

1. Antimicrobial Resistance Collaborators. "Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis". *Lancet* 399.10325 (2022): 629-655.
2. Schlaberg R., *et al*. "Validation of metagenomic next-generation sequencing tests for universal pathogen detection". *Archives of Pathology and Laboratory Medicine* 144.12 (2020): 1423-1432.
3. Hendriksen RS., *et al*. "Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage". *Nature Communication* 10.1 (2019): 1124.
4. Arango-Argoty G., *et al*. "DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data". *Microbiome* 8.1 (2020): 108.
5. Wheeler NE., *et al*. "Machine learning identifies signatures of host adaptation in the bacterial pathogen Salmonella enterica". *PLoS Genetics* 16.5 (2020): e1008850.
6. Macesic N., *et al*. "Machine learning: novel bioinformatics approaches for combating antimicrobial resistance". *Current Opinion in Infectious Diseases* 33.5 (2020): 382-388.
7. Sundermann AJ., *et al*. "Whole-genome sequencing surveillance and machine learning of the electronic health record for enhanced healthcare outbreak detection". *Clinical Infectious Disease* 75.3 (2022): 476-482.
8. Khaledi A., *et al*. "Predicting antimicrobial resistance in Pseudomonas aeruginosa with machine learning-enabled molecular diagnostics". *EMBO Molecular Medicine* 12.3 (2020): e10264.
9. Su M., *et al*. "Genome-based prediction of bacterial antibiotic resistance". *Journal of Clinical Microbiology* 57.3 (2019): e01405-18.
10. Yang Y., *et al*. "Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data". *Bioinformatics* 34.10 (2018): 1666-1671.
11. Avershina E and Ahmad R. "Machine learning approach for phenotype prediction from metagenomic composition". *Artificial Intelligence in Medicine* 111 (2021): 101998.
12. Kuang X., *et al*. "Accurate and rapid prediction of tuberculosis drug resistance from genome sequence data using traditional machine learning algorithms and CNN". *Scientific Report* 12.1 (2022): 2427.
13. Peiffer-Smadja N., *et al*. "Machine learning in the clinical microbiology laboratory: has the time come for routine practice?" *Clinical Microbiology and Infection* 26.10 (2020): 1300-1309.

14. Kim J., *et al*. "VAMPr: VAriant Mapping and Prediction of antibiotic resistance via explainable features and machine learning". *PLOS Computational Biology* 16.1 (2020): e1007511.

15. Nguyen M., *et al*. "Predicting antimicrobial resistance using conserved genes". *PLOS Computational Biology* 16.11 (2020): e1008319.

16. Eyre DW., *et al*. "WGS to predict antibiotic MICs for Neisseria gonorrhoeae". J*ournal of Antimicrobial Chemotherapy* 72.7 (2017): 1937-1947.

17. Yang Y., *et al*. "Prediction of antimicrobial resistance in Mycobacterium tuberculosis using a graph convolutional network". *Computational and Structural Biotechnology Journal* 19 (2021): 4096-4105.

18. Peng C., *et al*. "A transformer-based model for antimicrobial resistance prediction from whole genome sequencing data". *Brief Bioinformation* 24.1 (2023): bbac543.

19. Tian Y., *et al*. "Deep learning for antimicrobial resistance prediction: current applications and challenges". *Brief Bioinformation* 24.2 (2023): bbad024.

20. Ji Y., *et al*. "DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome". *Bioinformatics* 37.15 (2021): 2112-2120.

21. Rives A., *et al*. "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences". *PNAS* 118.15 (2021): e2016239118.

22. Devlin J., *et al*. "BERT: Pre-training of deep bidirectional transformers for language understanding". arXiv preprint arXiv:1810.04805 (2018).

23. Nguyen LH and Holmes S. "Ten quick tips for effective dimensionality reduction". *PLOS Computational Biology* 15.6 (2019): e1006907.

24. Pearman WS., *et al*. "The advantages and disadvantages of short- and long-read metagenomics to infer bacterial and eukaryotic community composition". *Annals of the New York Academy of Sciences* 1476.1 (2020): 42-50.

25. Liang Q., *et al*. "DeepMicrobes: taxonomic classification for metagenomics with deep learning". *NAR Genome Bioinformation* 2.1 (2020): lqaa009.

26. Wolters M., *et al*. "Rapid molecular diagnostics of respiratory tract infections caused by multidrug-resistant bacteria". *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 63.5 (2020): 601-608.

27. Moura A., *et al*. "Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes". *Nature Microbiology* 2 (2017): 16185.

28. Snitkin ES., *et al*. "Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing". *Science Translational Medicine* 4.148 (2012): 148ra116.

29. Nguyen M., *et al*. "Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal Salmonella". *Journal of Clinical Microbiology* 57.2 (2019): e01260-18.

30. Davis JJ., *et al*. "Antimicrobial resistance prediction in PATRIC and RAST". *Scientific Report* 6 (2016): 27930.

31. Hyun JC., *et al*. "Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity". *BMC Genomics* 23.1 (2020): 7.

32. Sheppard AE., *et al*. "Nested Russian doll-like genetic mobility drives rapid dissemination of the carbapenem resistance gene blaKPC". *Antimicrobe Agents Chemotherapy* 60.6 (2016): 3767-3778.

33. Drouin A., *et al*. "Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons". *BMC Genomics* 17.1 (2016): 754.

34. Kouchaki S., *et al*. "Application of machine learning techniques to tuberculosis drug resistance analysis". *Bioinformatics* 35.13 (2019): 2276-2282.

35. Aytan-Aktug D., *et al*. "Prediction of acquired antimicrobial resistance for multiple bacterial species using neural networks". *mSystems* 5.1 (2020): e00774-19.

36. Mahé P and Tournoud M. "Predicting bacterial resistance from whole-genome sequences using k-mers and stability selection". *BMC Bioinformatics* 19.1 (2018): 383.

37. Anahtar MN., *et al*. "Applications of machine learning to the problem of antimicrobial resistance: an emerging model for translational research". *Journal of Clinical Microbiology* 59.7 (2021): e0126020.

38. Moradigaravand D., *et al*. "Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data". *PLOS Computational Biology* 14.12 (2018): e1006258.

39. Ellington MJ., *et al*. "The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee". *Clinical Microbiology Infectious* 23.1 (2017): 2-22.

40. Moran RA and van Schaik W. "The dynamics of antimicrobial resistance: sources, sinks, and the global context". *Nature Reviews Microbiology* 21.1 (2023): 16-29.

41. Chawla NV., *et al*. "SMOTE: synthetic minority over-sampling technique". *Journal of Artificial Intelligence Research* 16 (2002): 321-357.

42. van Boeckel TP., *et al*. "Global antibiotic consumption 2000 to 2010: an analysis of national pharmaceutical sales data". *Lancet Infectious Disease* 14.8 (2014): 742-750.

43. Lundberg SM., *et al*. "A unified approach to interpreting model predictions". Adv Neural Inf Process Syst. 30 (2017): 4765-4774.

44. Chen T and Guestrin C. "XGBoost: a scalable tree boosting system". Proc 22nd ACM SIGKDD Int Conf Knowledge Discovery Data Min. (2016): 785-794.

45. Wattam AR., *et al*. "Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center". *Nucleic Acids Research* 45.D1 (2017): D535-D542.

46. US Food and Drug Administration. Proposed regulatory framework for modifications to artificial intelligence/machine learning-based software as a medical device (2019).

47. Dayan I., *et al*. "Federated learning for predicting clinical outcomes in patients with COVID-19". *Nature Medicine* 27.10 (2021): 1735-1743.

48. Subramanian I., *et al*. "Multi-omics data integration, interpretation, and its application". *Bioinformatics and Biology Insights* 14 (2020): 1177932219899051.

49. Ma T and Zhang A. "Integrate multi-omic data using affinity network fusion (ANF) for cancer patient clustering". Proc IEEE Int Conf Bioinform Biomed. (2017): 398-403.

50. Rieke N., *et al*. "The future of digital health with federated learning". *NPJ Digital Medicine* 3.1 (2020): 119.