Bridging the Gap Between Genotype and Phenotype: A Machine Learning Model for Rapid Antimicrobial Susceptibility Prediction

Jamil As-ad, Safaet Jaman Arman
International Islamic University Chittagong

Abstract

Antimicrobial resistance (AMR), particularly multi-drug resistance (MDR), is a critical global health threat necessitating rapid and accurate diagnostic methods to supplement or replace time-consuming culture-based antimicrobial susceptibility testing (AST). Machine learning (ML) and deep learning (DL) models that predict AMR phenotypes from whole-genome sequencing (WGS) data offer a promising solution. This study evaluates and develops several advanced ML approaches to predict AMR across multiple bacterial species, focusing on improving performance, addressing data limitations, and enhancing model interpretability. We applied a range of ML algorithms, including Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), Convolutional Neural Networks (CNN), and ensemble methods, to predict resistance phenotypes from genomic features such as single nucleotide polymorphisms (SNPs), k-mers, and gene content. A key challenge in AMR prediction is handling multi-drug resistance, where pathogens are resistant to multiple antibiotics simultaneously. Traditional models often predict resistance to single drugs in isolation. To address this, we developed a multi-label classification (MLC) approach using an Ensemble of Classifier Chains (ECC) model, which simultaneously predicts resistance to multiple drugs by accounting for correlations between them. When applied to 809 E. coli isolates, the ECC model significantly outperformed other MLC methods, demonstrating its potential for accurate MDR prediction. Furthermore, ML models often struggle with novel antibiotics or imbalanced datasets where resistant samples are scarce. We demonstrate that deep transfer learning can overcome these limitations by transferring knowledge from a well-trained model (e.g., for an antibiotic with abundant data) to a new task with limited data. This approach significantly improved prediction performance for antibiotics with small, imbalanced datasets, highlighting its potential for rapid diagnostics of emerging resistances. To enhance predictive accuracy, we also developed a novel discriminative position-fused deep learning classifier that integrates an attention mechanism with positional features, enabling the model to focus on core SNPs critical for AMR. This attention-based model significantly outperformed traditional CNNs and other ML models, improving the average AUROC to 0.80 and the F1-score to 0.82. Our findings show that ML models can achieve high predictive accuracy, with some models reaching 95–96In conclusion, advanced ML techniques like multi-label classification, deep transfer learning, and attention-based neural networks offer robust and accurate solutions for predicting AMR from genomic data. These approaches can handle complex MDR patterns, overcome data scarcity for novel antibiotics, and provide insights into resistance mechanisms, thereby paving the way for faster clinical diagnostics and improved patient outcomes

I. Introduction

Antimicrobial resistance (AMR) is a rapidly escalating global health crisis, representing one of the greatest threats to modern medicine. When disease-causing microorganisms like bacteria develop resistance to the drugs designed to treat them, the effectiveness of these treatments is nullified, leading to delayed patient recovery and increased mortality. According to World Health Organization (WHO) estimates, AMR could cause up to 10 million deaths annually by 2050, with an associated cost to healthcare systems of approximately 100 trillion. This threat is exacerbated by the overuse and misuse of antimicrobial drugs in clinical and agricultural settings, which fuels the selection and spread of resistant strains. A particularly serious concern is the rise of multi-drug resistance (MDR), where pathogens become resistant to multiple classes of antibiotics, leading to treatment failures and posing a significant challenge to public health. Escherichia coli, a major bacterial pathogen in clinical settings, is a prime example where the emergence of MDR has become a global health concern. The gold standard for determining AMR in clinical practice is antimicrobial susceptibility testing (AST). However, traditional culture-based AST methods are time-consuming, often taking two days for common bacteria and several weeks for slow-growing organisms like Mycobacterium tuberculosis. This delay often forces clinicians to prescribe broad-spectrum empiric antibiotics, a practice that can contribute to the further spread of resistance. Consequently, there is an urgent need for rapid and accurate diagnostic methods to guide appropriate, targeted therapy. Whole-genome sequencing (WGS) has emerged as a powerful alternative, offering a genetic-based approach to predict AMR phenotypes much faster than culture-based methods. Early genotypic prediction methods relied on rule-based approaches, which use prior knowledge to detect the presence or absence of known resistance genes

or mutations. While effective for well-studied organisms and resistance mechanisms, these methods are difficult to scale, require constant manual curation, and struggle with complex or novel resistance patterns. To overcome these limitations, machine learning (ML) and deep learning (DL) have been increasingly applied to predict AMR from WGS data. These data-driven approaches can learn the underlying patterns that connect genomic features—such as single nucleotide polymorphisms (SNPs), gene content, or k-mers (short DNA subsequences)—to resistance phenotypes without requiring a priori knowledge of the specific mechanisms involved. This allows ML models to not only predict resistance with high accuracy but also to potentially discover novel genetic determinants of AMR. Previous studies have demonstrated the high accuracy of ML models, such as Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF), and Convolutional Neural Networks (CNNs), in predicting AMR for a variety of pathogens, including E. coli, Salmonella, and M. tuberculosis. However, significant challenges remain in applying these models clinically. One major issue is the prediction of MDR, as most models focus on single-drug resistance and fail to account for the correlations and co-occurrence of resistance to multiple drugs. Another challenge is improving model performance for novel antibiotics or when faced with small, imbalanced datasets, which often hinder a model's accuracy and generalization. Furthermore, many complex DL models act as "black boxes," making it difficult to interpret the biological reasoning behind their predictions, a key requirement for clinical trust and application. This paper addresses these gaps by exploring advanced ML methodologies for AMR prediction. We investigate the use of Multi-Label Classification (MLC) models, such as the Ensemble of Classifier Chains (ECC), to simultaneously predict resistance to multiple drugs in E. coli by modeling the dependencies between them. We also propose a discriminative position-fused deep learning classifier that incorporates an attention mechanism to enhance predictive performance by focusing on core SNPs critical to resistance. Finally, we demonstrate that deep transfer learning can significantly improve prediction accuracy for novel antibiotics or on small, imbalanced datasets by leveraging knowledge from well-trained models. By systematically evaluating these approaches, we aim to extend the available tools for AMR prediction, paving the way for improved diagnostics and more effective patient treatment in the fight against this global health threat.