Application of Machine Learning in the Detection of Antimicrobial Resistance

Dissertation

zur Erlangung des Grades eines
Doktor der Naturwissenschaften
(Dr. rer. nat.)

dem Fachbereich Mathematik und Informatik
der Philipps-Universität Marburg von
Yunxiao Ren
aus Luo Yang, China

Marburg, 2023

Gutachter: Prof. Dr. Dominik Heider
Prof. Dr. Ho Ryun Chung

Als Dissertation dem
Fachbereich Mathematik und Informatik
der Philipps-Universität Marburg eingereicht.

Tag der Einreichung:
Tag der Disputation:

Erscheinungort: Marburg
Erscheinungsjahr: 2023
Hochschulkennziffer: 1180

Gutachter: Prof. Dr. Dominik Heider
Prof. Dr. Ho Ryun Chung

Urheberschaftserklärung

Hiermit versichere ich, dass ich die vorgelegte Dissertation selbständig und ohne fremde Hilfe verfasst, nicht andere als die in ihr angegebenen Quellen oder Hilfsmittel benutzt, alle vollständig oder sinngemäß übernommenen Zitate als solche gekennzeichnet, sowie die Dissertation in der vorliegenden oder einer ähnlichen Form noch bei keiner anderen in- oder ausländischen Hochschule anlässlich eines Promotionsgesuchs oder zu anderen Prüfungszwecken eingereicht habe.

Marburg, der _____

Yunxiao Ren

Zusammenfassung

Die Antibiotikaresistenz (AMR) ist zu einer der größten globalen Bedrohungen für die Gesundheit von Mensch und Tier geworden, was den Bedarf an schnellen und präzisen AMR-Diagnoseverfahren erhöht. Traditionelle antimikrobielle Empfindlichkeitstests (AST) sind zeitaufwändig, haben einen geringen Durchsatz und sind auf kultivierbare Bakterien beschränkt. Maschinelles Lernen bietet einen vielversprechenden Weg für die automatische AMR Vorhersage. Die meisten bestehenden Modelle legen jedoch den Schwerpunkt auf Merkmale, die sich nur auf bekannte Resistenzgene und -varianten beziehen, und stützen sich stark auf AMR-Referenzdatenbanken, wodurch neue AMR-bezogene Merkmale übersehen werden können. Um die oben genannten Herausforderungen zu bewältigen, werden in unserer ersten Studie genomweite maschinelle Lernmodelle zur effizienten Erkennung von AMR ohne Abhängigkeit von vorherigem AMR-Wissen eingeführt. Konkret haben wir verschiedene Modelle, darunter logistische Regression (LR), Support Vector Machine (SVM), Random Forest (RF) und Convolutional Neural Network (CNN), zur Vorhersage von Resistenzen gegen vier Antibiotika untersucht. Unsere Ergebnisse zeigen, dass diese Modelle AMR mit Label-Codierung, One-Hot-Codierung und 'Frequency Matrix Chaos Game Representation' (FCGR) auf ganze Genom-Sequenzierungsdaten effektiv vorhersagen können. Im Allgemeinen übertrafen RF und CNN die LR und SVM Modelle. Wichtig ist, dass wir für jedes Antibiotikum spezifische Mutationen identifiziert haben, die mit AMR in Verbindung stehen.

Darüber hinaus konzentrieren sich aktuelle AMR-Studien auf die Vorhersage der Resistenz gegen ein einzelnes Medikament und ignorieren die kumulative Natur der antimikrobiellen Resistenz im Laufe der Zeit, was die schnelle Identifizierung von Multiresistenzen (MDR) zu einer Herausforderung macht. Um diese Einschränkungen zu überwinden, haben wir in unserer zweiten Studie fünf Multi-Label-Klassifikationsmodelle (MLC) für MDR-Probleme entwickelt. Unsere Ergebnisse zeigten, dass das ECC-Modell (Ensemble Classifier Chains) die anderen MLC-Methoden übertraf und eine deutliche Wirksamkeit bei der Vorhersage von MDR zeigte.

Darüber hinaus stellen begrenzte Trainingsstichproben und unausgewogene Daten erhebliche Hindernisse für die Generalisierung und Genauigkeit von AMR-Modellen dar. Um diese Herausforderungen zu überwinden, haben wir in unserer dritten Studie ein Deep-Transfer-Learning-Modell auf der Grundlage einer CNN-Architektur vorgeschlagen. Zunächst trai-

nieren wir das Modell auf vier Datensätzen, dann wird das beste Modell als Ausgangsmodell für das 'Transfer Learning' verwendet, und das Modell wird auf kleinen Datensätzen neu trainiert, indem die Architektur und Gewichte vom Ausgangsmodell übertragen werden. Unsere Ergebnisse zeigen, dass unser Deep-Transfer-Learning-Modell die Modellleistung für AMR-Vorhersagen auf kleinen, unausgewogenen Datensätzen verbessert.

In einer Zeit, in der Datensicherheit und Datenschutz von entscheidender Bedeutung sind, bieten 'Federated Learning' (FL) und 'Swarm Learning' (SL) Lösungen, indem sie Daten während des Trainings lokal halten. Dieser Ansatz reduziert die Notwendigkeit, sensible Informationen an einen zentralen Server zu übertragen und verbessert die Effizienz durch die Verteilung der Rechenlast. Darüber hinaus wird beim Schwarmlernen eine Dezentralisierung erreicht, da im Vergleich zum föderierten Lernen kein zentraler Server zur Verwaltung der Parameter erforderlich ist, was die Sicherheit der Daten weiter verbessert. In unserer vierten Studie befassen wir uns daher mit der Anwendung des Schwarmlernens speziell im Zusammenhang mit AMR.

Abstract

Antimicrobial resistance (AMR) has become one of the significant global threats to both human and animal health, intensifying the need for rapid and precise AMR diagnostic methods. Traditional antimicrobial susceptibility testing (AST) is time-consuming, low throughput, and limited to cultivable bacteria. Machine learning offers a promising avenue for automated AMR prediction. However, most existing models emphasize features related only to known resistance genes and variants, relying heavily on AMR reference databases, and thus may overlook new AMR-related features. To address the above challenges, my first study introduces genome-wide machine learning models to detect AMR without dependence on prior AMR knowledge efficiently. Specifically, I assessed various models, including logistic regression (LR), support vector machine (SVM), random forest (RF), and convolutional neural network (CNN), for predicting resistance against four antibiotics. The findings illustrated that these models can effectively predict AMR with label encoding, one-hot encoding, and frequency matrix chaos game representation (FCGR) encoding on whole-genome sequencing data. Generally, RF and CNN outperformed LR and SVM. Importantly, I identified specific mutations associated with AMR for each antibiotic.

Moreover, current AMR studies focus on single-drug resistance prediction, ignoring the cumulative nature of antimicrobial resistance over time, which makes rapid identification of multi-drug resistance (MDR) a challenge. Therefore, in my second study, in order to overcome these limitations, I constructed five multi-label classification (MLC) models for MDR problems. The findings revealed that the ECC (Ensemble Classifier Chains) model surpassed the other MLC methods, demonstrating marked effectiveness in predicting MDR.

Furthermore, the constraints of limited training samples and data imbalances present significant barriers to the generalization and accuracy of AMR models. To overcome these challenges, in my third study, I have proposed a deep transfer learning model based on a CNN architecture. First, I pre-train the model on four datasets, then the best-performing model is used as the source model for transfer learning, and the model is retrained on small datasets by transferring the architecture and weights from the source model. The results showed that the deep transfer learning model improves model performance for AMR prediction on small and imbalanced datasets.

In an era where data security and privacy are crucial, federated learning (FL) and swarm learn-

ing (SL) present solutions by maintaining data locally during training, which reduces the necessity to transfer sensitive information to a centralized server and improves efficiency by distributing computational load. Moreover, swarm learning achieves decentralization by not requiring a central server to manage the parameters compared to federated learning, which further improves the security of the data. Thus, in my fourth study, I delve into the application of swarm learning specifically within the context of AMR.

Contents

1 Introduction 1
- 1.1 Antimicrobials and Antimicrobial Resistance . . . 1
  - 1.1.1 A brief history of antibiotic discovery . . . 1
  - 1.1.2 The emergence and challenges of antimicrobial resistance . . . 2
  - 1.1.3 Mechanisms of antimicrobial resistance . . . 3
- 1.2 Conventional Detection Methods for Antimicrobial Resistance . . . 5
  - 1.2.1 Antimicrobial susceptibility testing . . . 5
  - 1.2.2 Sequencing-based resistance discovery . . . 6
- 1.3 Machine Learning for Detection of Antimicrobial Resistance . . . 7
  - 1.3.1 Fundamentals of machine learning . . . 7
  - 1.3.2 DNA sequence encoding . . . 9
  - 1.3.3 Machine learning algorithms . . . 11
  - 1.3.4 Machine learning classification tasks . . . 12
  - 1.3.5 Training strategies . . . 14
  - 1.3.6 Evaluation metrics . . . 18
  - 1.3.7 Application of ML to AMR . . . 20
- 1.4 Challenges and Motivation . . . 21
- 1.5 Aims . . . 22
- 1.6 List of Publications . . . 23

2 Methods 25
- 2.1 Datasets Overview . . . 25
- 2.2 Whole Genome Sequencing Analysis . . . 26
- 2.3 Sequences Encoding . . . 26
- 2.4 Model Training and Evaluation . . . 27
- 2.5 Gene Annotation . . . 28
- 2.6 Multi-label Classification . . . 28
- 2.7 Basic CNN Model . . . 29
- 2.8 Deep Transfer Learning Architecture . . . 29

2.9 Swarm Learning . . . 30
2.9.1 Swarm learning framework . . . 30
2.9.2 Algorithm in swarm learning . . . 30
2.9.3 Performance comparison . . . 30

3 Results 32
3.1 Publication 1: Prediction of Antimicrobial Resistance based on Whole-genome Sequencing and Machine Learning. . . . 33
3.1.1 Summary . . . 33
3.1.2 Introduction . . . 35
3.1.3 Materials and methods . . . 37
3.1.4 Results . . . 39
3.1.5 Discussion . . . 41
3.1.6 Conclusion . . . 42
3.2 Publication 2: Multi-label Classification for Multi-drug Resistance Prediction of *Escherichia Coli* . . . 45
3.2.1 Summary . . . 45
3.2.2 Introduction . . . 47
3.2.3 Materials and methods . . . 48
3.2.4 Results . . . 49
3.2.5 Discussion . . . 51
3.2.6 Conclusion . . . 52
3.3 Publication 3: Deep Transfer Learning Enables Robust Prediction of Antimicrobial Resistance for Novel Antibiotics . . . 54
3.3.1 Summary . . . 54
3.3.2 Introduction . . . 55
3.3.3 Results . . . 56
3.3.4 Discussion . . . 61
3.3.5 Materials and methods . . . 62
3.4 Study 4: Swarm Learning Predicts AMR (Unpublished) . . . 67
3.4.1 Aim and motivation . . . 67
3.4.2 Results . . . 67
3.4.3 Discussion and conclusion . . . 73

4 Discussion 74
4.1 Experimental Validation . . . 74
4.2 Species Generalization . . . 75
4.3 Feature Input: SNP and Beyond . . . 75

4.4 Software Development 76
4.5 Focus on AMP 76
4.6 Concluding Remark 77

List of Figures **78**

List of Acronyms and Abbreviations **80**

Bibliography **83**

Appendix A Appendix **106**
A.1 Lebenslauf 107

"Die Wissenschaft nötigt uns, den Glauben an einfachen Kausalitäten aufzugeben"
Friedrich Nietzsche

1 Introduction

1.1 Antimicrobials and Antimicrobial Resistance

1.1.1 A brief history of antibiotic discovery

Antimicrobials are broadly defined as agents used to protect against and combat infections triggered by microorganisms, such as bacteria, fungi, viruses, and parasites, in plants, animals, and humans, which include a large group of substances, such as antibiotics, antivirals, and antifungals (Shankarnarayan et al., 2022). In a narrow sense, it usually refers to antibiotics, a specific antimicrobial class that can inhibit or kill bacteria (Boolchandani et al., 2019). Here, we focus on antibiotics.

The first antibiotic, penicillin, was discovered by Alexander Fleming in 1928, setting the stage for the development of effective antimicrobial agents (Gaynes, 2017) (Figure 1.1). Then, in the 1930s to 1940s, with the realization of penicillin purification technology and in-depth study of its properties, it was widely used in World War II, saving a large number of lives (Hutchings et al., 2019). Along with the successful application of penicillin, researchers turned their attention to discovering new antibiotics. In 1943, streptomycin was discovered and was successfully used to treat tuberculosis (a previously incurable disease), which marked the beginning of a golden age of antibiotic discovery (Aminov, 2010). During the 1940s to 1950s, several important antibiotics were discovered. Chlortetracycline (aureomycin), as the first tetracycline, was isolated in 1948, followed by other tetracyclines such as hygromycin

and doxycycline (Nelson and Levy, 2011). In addition, antibiotics such as chloramphenicol, erythromycin, and vancomycin were also discovered during this period (Hutchings et al., 2019). In the 1960s era, as the need for antibiotics grew, scientists embarked on the exploration of synthesizing antibiotics instead of relying solely on natural sources. A significant breakthrough was achieved with the development of synthetic penicillins, including methicillin, which demonstrated efficacy against bacteria resistant to traditional penicillin treatments (Ribeiro da Cunha et al., 2019). From the 1970s to the 2000s, researchers discovered and developed a variety of novel antibiotics. Examples include cephalosporins, fluoroquinolones, macrolides (such as azithromycin), aminoglycosides, and carbapenems (Figure 1.1). These antibiotics have expanded the therapeutic options for treating bacterial infections (Hutchings et al., 2019).

In summary, the discovery of antibiotics has revolutionized medicine, saving countless lives and transforming the treatment of bacterial infections.

1.1.2 The emergence and challenges of antimicrobial resistance

However, with the overuse and misuse of antibiotics, antimicrobial resistance (AMR) has been gradually reported, in which infectious microorganisms became insensitive to antibiotics, leading to poor outcomes and severe illness and death (Shankarnarayan et al., 2022; Zaman et al., 2017; Palumbi, 2001; Clatworthy et al., 2007).

During the 1940s, just a few years after the mass production of penicillin began during World War II, resistance was observed (Figure 1.1). *Staphylococcus aureus* developed resistance through the production of beta-lactamase (Barber and Rozwadowska-Dowzenko, 1948), an enzyme that inactivates penicillin. Then, the first cases of methicillin-resistant *Staphylococcus aureus* (MRSA) were reported in 1961, just two years later in the introduction of methicillin (Jevons, 1961). Resistance to tetracycline also emerged shortly after its introduction, with resistance genes carried on plasmids (Speer et al., 1992). Vancomycin, introduced in the 1950s, saw its first instances of resistance emerge in the 1980s (Cetinkaya et al., 2000). After the introduction of fluoroquinolones in the late 1960s, resistance began to emerge in the 1970s (Yoshida et al., 1988). During the 1980s-2000s, the discovery and development of new antibiotics slowed. Fewer and fewer new classes of antibiotics were introduced to the market while resistance continued to increase. This has led to a growing gap between the emergence of resistance and the availability of effective treatment options. Moreover, multidrug-resistant tuberculosis (MDR-TB) was identified as a serious problem in the late 1980s and early 1990s (Frieden et al., 1996). The late 1990s and early 2000s also saw the rise of resistance to carbapenems, a class of last-resort antibiotics, in organisms like *Klebsiella pneumoniae* and *Escherichia coli*

(*E. coli*). The WHO (World Health Organization) reported extensively drug-resistant TB (XDR-TB) in 2006 (Nordmann et al., 2009).

Today, antimicrobial resistance remains one of the greatest threats to global health, food security, and social development. The evolution and spread of drug-resistant bacteria continue, and it is estimated that if measures are not taken to address AMR by 2050, the annual global death toll will reach 10 million, and the cost will reach $100 trillion.

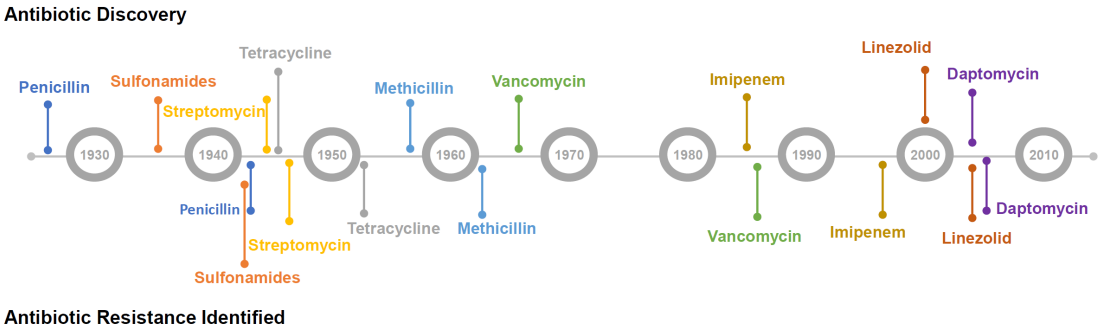

Figure 1.1: History of antibiotic development and observed time of antibiotic resistance. The year each antibiotic was discovered is shown above the timeline, and the year resistance to each antibiotic was identified is indicated below the timeline.

1.1.3 Mechanisms of antimicrobial resistance

AMR can occur through various mechanisms, which can be broadly classified into three categories: genetic mechanisms, biochemical mechanisms, and physical mechanisms (Figure 1.2) (Shankarnarayan et al., 2022; Darby et al., 2023; Munita and Arias, 2016).

Genetic mechanisms are commonly thought to include mutations and horizontal gene transfer (Christaki et al., 2020; Alekshun and Levy, 2007). **a. Mutation**: Bacteria can acquire resistance through random mutations in their genetic material. These mutations can alter the target site of the drug or modify the metabolic pathways, rendering the antimicrobial ineffective. **b. Horizontal gene transfer**: Bacteria can also obtain resistance genes from other bacteria. This can occur through three main mechanisms: **1) Conjugation**: This is a process where one bacterium transfers a copy of a resistance gene to another bacterium through plasmids (small, circular DNA molecules). **2) Transformation**: Bacteria can pick up pieces of DNA from the environment that contain resistance genes and incorporate them into the bacterial genome. **3) Transduction**: This involves the transfer of resistance genes via bac-

teriophages that infect bacteria (Boolchandani et al., 2019).

Biochemical mechanisms include the following directions: **a. Enzymatic inactivation**: Some microorganisms produce enzymes that can chemically modify or degrade antimicrobial agents (Browne et al., 2020). For example, beta-lactamases can break down beta-lactam antibiotics, such as penicillins and cephalosporins. **b. Efflux pumps**: Bacteria can have efflux pumps that actively pump out the antimicrobial agents from inside the cell, preventing their accumulation to effective levels (Browne et al., 2020).

Physical mechanisms include altered permeability and target site modification (Boolchandani et al., 2019). **a. Altered permeability**: Microorganisms can develop mechanisms to modify their outer membrane or cell wall, making it more difficult for drugs to penetrate and reach their targets (Cag et al., 2016; Blair et al., 2015). For example, biofilm formation, which is communities of microorganisms that can attach to surfaces and form a protective matrix. This makes it difficult for antimicrobial drugs to penetrate and reach the bacteria. Additionally, bacteria in biofilms often have slower metabolic rates, making them less susceptible to drugs that target active growth. **b. Target site modification**: Changes in the structure of drug targets, such as receptors or enzymes, can prevent antimicrobial agents from binding effectively, reducing their efficacy (Zaman et al., 2017).

It's important to note that these mechanisms of resistance can act individually or in combination, leading to multi-drug resistance or extensively drug-resistant strains of microorganisms (Shankarnarayan et al., 2022; Munita and Arias, 2016). The misuse and overuse of antimicrobial agents, such as inappropriate prescription or agricultural use, can accelerate the development and spread of antimicrobial resistance. Proper antimicrobial stewardship and infection control measures are crucial to combat the emergence and spread of resistant microorganisms.

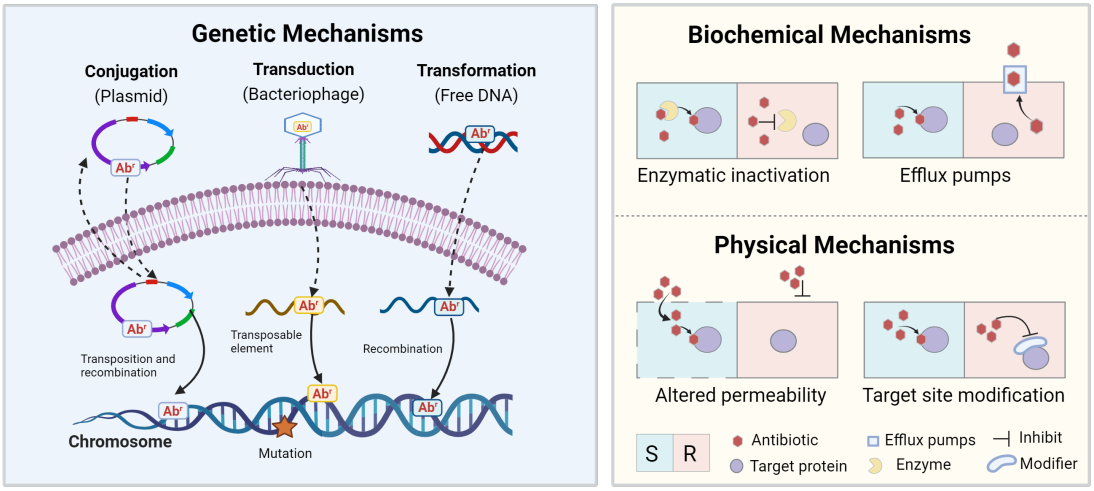

Figure 1.2: Genetic, biochemical, and physical mechanisms of antibiotic resistance. The diagram on the left shows the genetic mechanisms that lead to bacteria acquiring antibiotic resistance (Ab^r), which include both gene mutations and horizontal gene transfer. The latter involves the acquisition of resistance genes through plasmids and conjugative transposons (conjugation), and by bacteriophage (transduction), as well as the integration of foreign free DNA into the bacterial chromosome (transformation). The diagram on the right shows biochemical mechanisms and physical mechanisms, where S represents Susceptible, R represents Resistant. This figure was adapted from Alekshun and Levy (2007) and Boolchandani et al. (2019), which was created with BioRender.

1.2 Conventional Detection Methods for Antimicrobial Resistance

1.2.1 Antimicrobial susceptibility testing

Antimicrobial susceptibility testing (AST) is a laboratory method used to determine the effectiveness of specific antimicrobial agents against bacteria or other microorganisms. It helps guide healthcare professionals in selecting appropriate antibiotics for treating bacterial infections (Boolchandani et al., 2019).

The first step of AST is to isolate and identify the bacterial strains obtained from patient samples, such as blood, urine, or wound cultures. This step is crucial as susceptibility patterns can vary among different bacterial species. Then the isolated bacterium is grown in a laboratory culture medium and exposed to different antibiotics to see how it reacts. Once the testing is done, the results are interpreted based on professional organizations like the Clinical and Laboratory Standards Institute (CLSI) or the European Committee on Antimicrobial Susceptibility Testing (EUCAST) (Boolchandani et al., 2019). The results are reported as either susceptible, intermediate, or resistant, indicating the effectiveness of each antibiotic against the tested bacterium. This information helps guide clinicians in choosing the most

appropriate antibiotic treatment (Boolchandani et al., 2019).

AST is a traditional and standardized method for assaying antimicrobial resistance in bacteria. However, it can sometimes be complex, time-consuming, and low throughput, particularly for organisms that are difficult to grow in a lab or for which standard testing methods are not available (Boolchandani et al., 2019). Thus, a rapid and accurate approach to AMR detection is a critical part of managing infectious diseases, particularly in the era of growing antibiotic resistance.

1.2.2 Sequencing-based resistance discovery

Advances in sequencing technology and decreasing costs have made sequencing-based approaches a viable and effective tool for antimicrobial resistance discovery and surveillance (Boolchandani et al., 2019). These methods leverage high-throughput DNA sequencing technologies, such as whole-genome sequencing (WGS), metagenomic sequencing, and targeted gene sequencing, to analyze microbial genomes and identify specific genetic mutations, resistance genes, and mobile genetic elements contributing to AMR (World Health Organization, 2020). In particular, WGS can provide a comprehensive insight into an isolate's genome, which can promote understanding of AMR mechanisms and distinguish pathogen subtyping with identical AST profiles. This kind of molecular data can also be used for surveillance and development of new diagnostics and therapies for AMR. Moreover, it facilitates identifying the position of AMR determinants on either the bacterial chromosome or plasmids, thereby providing crucial information about the routes of AMR spread (World Health Organization, 2020; Köser et al., 2014).

Reuter et al. (2013) highlights the role of WGS in detecting antibiotic resistance and tracking the spread of multidrug-resistant bacteria. Danko et al. (2021) provided antimicrobial resistance markers in different geospatial contexts by analyzing a global map of 4728 metagenomic samples from 60 urban public transportation systems. Roemer and Boone (2013) reviewed the targeted-sequencing strategy for antimicrobials discovery.

WGS serves as a complementary method to AST, offering comprehensive information on the epidemiology of resistance genes in studying resistance determinants. Moreover, WGS facilitates high-throughput AMR monitoring and the identification of AMR-related markers (Boolchandani et al., 2019; World Health Organization, 2020). Despite these advantages, WGS does exhibit some limitations that need to be known. For example, sequence-based approaches to antimicrobial resistance typically involve identifying resistance determinants by first predicting the protein-coding region and then comparing it to AMR refer-

ence databases, such as Comprehensive Antibiotic Resistance Database (CARD), Antibiotic Resistance Genes Database (ARDB) or the active Antibiotic Resistance Gene Annotation (ARGANNOT). The bias of AMR-related databases thus affects the accuracy of prediction. Most antimicrobial resistance databases lack standardization and effective and sustainable management pipelines, they are usually only maintained for a few short years with a lot of outdated information that is not updated in a timely manner. Another important limitation is that they focus on the identification and characterization of protein-coding resistance genes; they ignore the complexity of the mechanisms of AMR, such as genomic changes or de novo mutations in ribosomal RNA (rRNA) genes and regulatory elements, as well as drug-target mutations (Boolchandani et al., 2019).

1.3 Machine Learning for Detection of Antimicrobial Resistance

Machine learning (ML) techniques have emerged as powerful tools for addressing various challenges related to AMR (Farhat et al., 2023; Kim et al., 2022). Here, we will introduce the basics of ML and its application to AMR.

1.3.1 Fundamentals of machine learning

ML can identify patterns from large amounts of data and make predictions or classifications based on learned patterns (Sarker, 2021; Domingos, 2012). The machine learning process begins with data collection, where understanding the available features and target data is crucial based on specific research questions (Figure 1.3) (Alzubaidi et al., 2021). Subsequently, data undergoes preprocessing, which includes tasks like data cleaning — eliminating missing values, outliers, and duplicates — and feature encoding, which converts the raw data into a format that can be recognized by machine learning (Figure 1.3) (Qu et al., 2019). The data is then split into distinct subsets: a training set for model development, a validation set for fine-tuning, and a test set for final evaluation (Figure 1.3). Following this, the appropriate model, whether for classification, regression, or clustering, is selected. Then multiple models are contrasted to pinpoint the best-performing one (Sarker, 2021). Then comes model training, which involves the actual training of the chosen model, incorporating hyperparameter tuning to optimize its performance (Swanson et al., 2023). The model is first trained on the training data and then validated on the validation set to ensure that it's generalizing well to unseen data. Finally, the model is assessed using an independent test set with suitable evaluation metrics. Interpreting the model's results in the context of the research question is also important (Carvalho et al., 2019; Burkart and Huber, 2021).

Machine learning is usually categorized into supervised, unsupervised as well as reinforce-

ment learning (Swanson et al., 2023; Sarker, 2021). Supervised learning uses labeled training data to learn the relationship between inputs and outputs (Alzubaidi et al., 2021). Common algorithms include linear regression, logistic regression (LR), support vector machines (SVM), and various neural networks. Unsupervised learning recognizes patterns in the data without reference to known labeled results. Common algorithms include clustering methods, such as k-means and hierarchical clustering, and dimensionality reduction methods, such as principal component analysis (PCA) (Alzubaidi et al., 2021; Swanson et al., 2023). Reinforcement learning learns how to behave in their environment by performing certain actions and receiving rewards or penalties. Q-learning and deep Q-networks (DQNs) are examples of reinforcement learning algorithms (Botvinick et al., 2019; Nian et al., 2020; Arulkumaran et al., 2017).

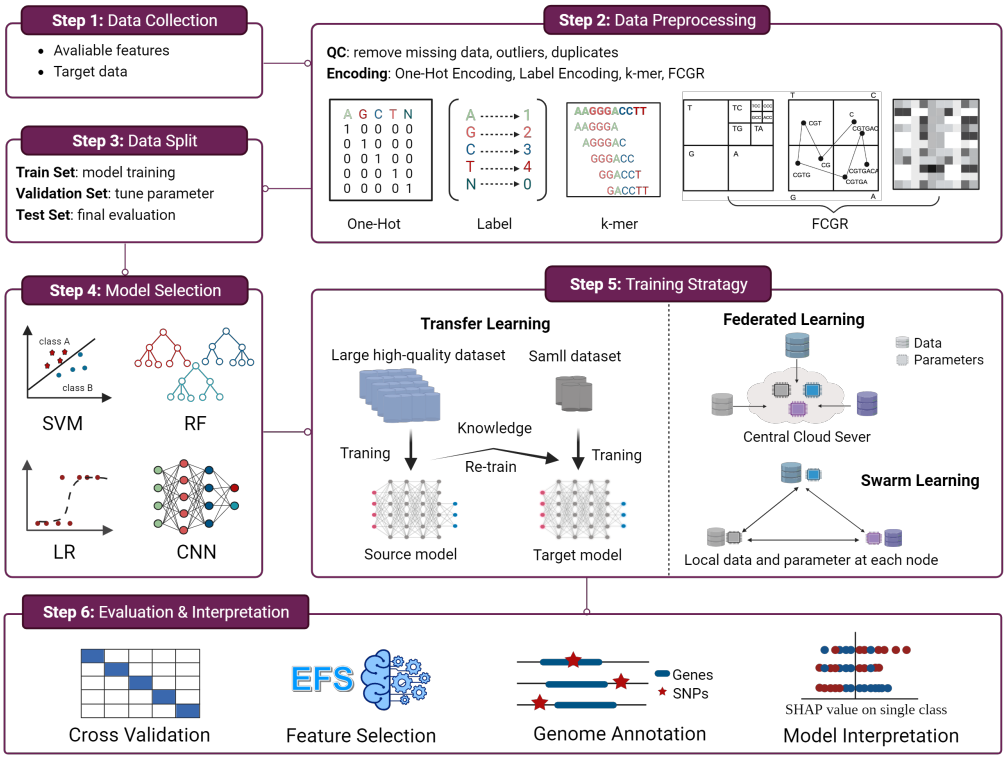

Figure 1.3: Overview of machine learning workflow and project design. This figure was created by BioRender.com.

1.3.2 DNA sequence encoding

DNA sequence encoding is the process of transforming DNA nucleotide sequences, typically represented by the characters A, T, C, and G, into a numerical format that can be recognized by computational algorithms, which is an essential step for ML (Spänig and Heider, 2019; Chen et al., 2020). The common encoding methods for DNA sequences include label encoding, One-Hot encoding, k-mer encoding, and Chaos Game Representation (CGR) encoding (Yu et al., 2018; Spänig and Heider, 2019; Ren et al., 2021).

Label encoding

Label encoding is also named integer encoding. Each nucleotide is mapped to a unique integer value. A common mapping might be A=1, G=2, C=3, T=4 (Yu et al., 2018). This method is straightforward. Gunasekaran et al. (2021) use both label and k-mer encoding techniques to encode DNA sequences. Following this, they employed several neural network models such as convolutional neural networks (CNN), CNN coupled with Long Short-Term Memory (CNN-LSTM), and CNN integrated with Bidirectional LSTM, aiming at sequence classification.

One-Hot encoding

One-hot encoding, also referred to as sparse encoding, encodes the DNA sequence into a binary matrix, which is then vectorized and used as input for the ML models. For example, A=[1, 0, 0, 0], C=[0, 1, 0, 0], G=[0, 0, 1, 0], T=[0, 0, 0, 1]. It's widely applied in genomics, including DNA, RNA, and protein sequence encoding. For example, Zhou et al. (2022) encoded DNA sequences using a One-Hot encoding scheme and then employed deep neural networks to predict the locations of nucleosomes from these DNA sequences. Mittag et al. (2015) implemented coding schemes like label encoding and One-Hot encoding to represent the genotypes of single nucleotide polymorphisms (SNPs), and subsequently examined how these encoding methods influenced the performance of predicting disease risk. Enireddy et al. (2022) employed One-Hot encoding in conjunction with LSTM techniques to predict protein secondary structure. Kuzmin et al. (2020) utilized the widely recognized One-Hot encoding method to transform the sequences into numerical vectors suitable for input into machine learning algorithms and then predicted the host specificity of coronaviruses.

K-mer encoding

K-mer encoding is a method representing genomic sequences by counting the occurrences of all possible substrings of length k (referred to as k-mers) within the sequence (Gunasekaran et al., 2021; Manekar and Sathe, 2018). By offering a fixed-size representation of variable-length sequences, this method is frequently used across various fields, including genomics, metagenomics, and other areas of bioinformatics. Fletez-Brant et al. (2013) developed kmer-

SVM, a web server for identifying predictive regulatory sequence features in genomic data sets based on k-mer encoding. Orozco-Arias et al. (2021) classified long terminal repeat retrotransposons in plant genomes based on k-mer's ML approach. Solis-Reyes et al. (2018) introduced an open-source, supervised, and alignment-free subtyping method called Kameris, which functions by analyzing k-mer frequencies in HIV-1 sequences. Mahé and Tournoud (2018) utilized a k-mer-based genotyping approach and a logistic regression model, combining multiple k-mers into a probabilistic framework for predicting bacterial resistance.

Chaos game representation encoding

Chaos game representation (CGR) encoding is a novel method used to visualize DNA sequences by turning them into a unique pattern or shape, which was first applied CGR algorithm to DNA sequences by Jeffrey (1990). The method is based on a recurrent iterative function system, which can be used to visualize sequences by building fractals from sequences of symbols (Wang et al., 2005; Löchel et al., 2020; Löchel and Heider, 2021). Specifically, this process begins with a square, where each corner in the square represents one of the four DNA bases (A, G, C, T). A dot is initially placed in the center of the square. Then, for each letter in the DNA sequence, the dot is moved halfway to the corner that matches the letter, and a mark is made at the new position. This process is repeated for each subsequent letter in the sequence, with the dot consistently moving halfway to the corner associated with the next letter. Upon completion of the entire sequence, the marks form a unique pattern that visually represents that specific DNA sequence (Almeida et al., 2001; Löchel and Heider, 2021). CGR has a wide range of applications in genomics. Kania and Sarapata (2021) proposed a generalized method for constructing chaos game representations, called serial chaos game representations, which can be used to construct representations that are less sensitive to mutations, thus providing more reliable values for phylogenetic tree construction for free alignment. Hoang et al. (2016) encoded DNA sequences by treating 2D CGR coordinates as complex numbers, and then employed digital signal processing methods to analyze their evolutionary relationship. CGR has also been utilized in the rapid comparison of different strains of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Sengupta et al., 2020). While most existing studies on CGR encoding focused on CGR for DNA, there also exists a smaller number of studies dealing with other alphabets, such as the encoding of protein sequences. For example, Yu et al. (2004) applied CGR algorithm to classify proteins, dividing amino acids into four groups according to their characteristics and then utilizing multifractal and correlation analysis to build a phylogenetic tree for Archaea and Eubacteria. In alternative methods, amino acids were retranslated into DNA for CGR representation (Yang et al., 2009). Sun et al. (2020) employed a three-dimensional CGR technique for protein classification.

Frequency Chaos Game Representation

Frequency Chaos Game Representation (FCGR) is a variant of the standard Chaos Game Representation (CGR) method used for DNA sequence encoding (Rizzo et al., 2016). While CGR provides a unique fractal visualization of a DNA sequence, FCGR takes this a step further by transforming the CGR into a frequency matrix that can be used for quantitative analysis (Löchel et al., 2020; Löchel and Heider, 2021). Lichtblau (2019) used FCGR method to transform sequences into images, followed by dimensionality reduction to create vectors of moderate length. These vectors can then be used for rapidly searching sequences, building phylogenetic trees, and classifying viral genome data. Wang et al. (2005) used FCGR method to compute the image distance between genomes, which was then used to construct phylogenetic trees. And Löchel et al. (2020) utilized FCGR in conjunction with CNN for predicting resistance in HIV-1.

Different encoding methods are suitable for different tasks and models. Simpler methods like label or One-Hot encoding might be used as a starting point, with more complex methods employed as needed based on the requirements of the specific analysis.

1.3.3 Machine learning algorithms

Traditional ML algorithms

Traditional ML algorithms come in various forms. Random forest (RF) is one common algorithm, which is an ensemble learning method that can be used for both classification and regression tasks (Breiman, 2001). RF is composed of multiple decision trees. Each tree is constructed from the training data and is used to make sequential binary decisions about the input features. These decisions ultimately lead to a prediction concerning the label of the data points (Swanson et al., 2023). RF often outperforms models that rely on a single tree as they combine the insights from multiple decision trees. This ensemble approach not only enhances the overall predictive accuracy but also enables random forests to assign an importance value to each feature, reflecting its contribution to the final prediction result. Another popular algorithm is support vector machines (SVM), which is a set of supervised learning methods used for classification, regression, and outliers detection (Chen et al., 2012). The basic concept behind SVM is to find a hyperplane (a line in 2D, a plane in 3D, or a hyperplane in more than three dimensions) that best separates the data into different classes (Boser et al., 1992; Swanson et al., 2023). The optimal hyperplane is the one that maximizes the margin between the closest points (support vectors) of the different classes. With its effectiveness in higher dimensional spaces and robustness to outliers, SVM serves as a powerful tool in vari-

ous analytical applications. Regression models are designed to find a linear combination of input features that can accurately predict continuous outcomes, as seen in linear regression, or binary outcomes, as exemplified by logistic regression (LR) (Swanson et al., 2023; Maulud and Abdulazeez, 2020). In the training process of LR, coefficients are typically estimated using maximum likelihood estimation, optimizing the model's ability for prediction.

Deep learning

Deep learning is a subset of machine learning that involves algorithms inspired by the structure and function of the brain, particularly neural networks (Wainberg et al., 2018). The basic units of a neural network are neurons. They receive input from other neurons, perform a weighted sum of the inputs, pass this through an activation function, and send the output to neurons in the next layer (Swanson et al., 2023). This design allows deep learning models to capture complex patterns and relationships within data. Thus, it can be applied to a wide variety of tasks, including image and speech recognition, natural language processing, and even drug discovery (Alzubaidi et al., 2021).

The common deep learning models include convolutional neural network (CNN), recurrent neural network (RNN), generative adversarial network (GAN), and transformer models (Alzubaidi et al., 2021). CNN is usually used to process grid-structured data like images, utilizing convolutional layers that automatically and adaptively learn spatial hierarchies of features (P and R, 2023). RNN is designed to recognize patterns in sequences of data, such as time series or natural language (Lipton et al., 2015; Sherstinsky, 2020). GAN consists of two networks, a generator, and a discriminator, that are trained together. The generator learns to generate data, and the discriminator learns to distinguish between real and generated data (Aggarwal et al., 2021; Gui et al., 2023). Transformer models are based on attention mechanisms, allowing them to consider other parts of the input when encoding a particular part, which is especially useful in natural language processing (Vaswani et al., 2017; Lin et al., 2022).

1.3.4 Machine learning classification tasks

Classification is one of the main tasks in machine learning and belongs to the category of supervised learning, which involves classifying input information into one of two or more categories. Common classification problems include binary classification, multiple classification, and multi-label classification (MLC).

Binary classification

Binary classification is one of the most common and fundamental tasks in machine learning. It involves categorizing instances into one of two classes, often labeled as 0 or 1, or negative or positive, such as identification of tumor and normal tissue, drug resistance and non-resistance (Kumari and Kr., 2017; Canbek et al., 2022).

Multi-class classification

Multi-class classification, also known as multinomial classification, extends the concept of binary classification to more than two classes. In this task, the goal is to categorize instances into one of three or more classes (Mehra and Gupta, 2013; Grandini et al., 2020; Sharma and Parwekar, 2023). Examples of multi-class classification include classifying handwritten digits into one of the ten classes, determining the sentiment of a text as positive, negative, or neutral, and diagnosing a patient's illness based on symptoms and test results into one of several diseases or conditions. These scenarios illustrate the diverse applications of multi-class classification.

Multi-label classification

Multi-label classification (MLC) is a type of classification where an instance can be assigned to multiple classes or labels simultaneously (Zhang and Zhou, 2014; Tarekegn et al., 2021). Unlike multi-class classification, where each instance is categorized into one and only one class, MLC allows for a broader and more flexible categorization (Tawiah and Sheng, 2013; Bogatinovski et al., 2022). MLC is well suited to deal with multi-drug resistance issues.

Multi-label problems have traditionally been transformed into single-label problems (Tsoumakas et al., 2009). A common method, known as the binary relevance (BR) approach, simplifies this by treating each label as an independent binary problem (Rokach et al., 2014). However, a significant limitation of the BR approach is its failure to consider dependencies between labels (Read et al., 2021). In contrast to BR, the classifier chain (CC) method explicitly accounts for label correlations by using the predictions from preceding classifiers as additional inputs for subsequent ones (Read et al., 2011). This makes the order of the CC integral to prediction accuracy, leading to the development of the ensemble of classifier chains (ECC). ECC combines several CCs with varied orders to study dependencies between labels (Read et al., 2011, 2021). While CCs and ECCs have been employed for cross-resistance prediction in HIV, specifically focusing on the protein sequences of HIV-1 reverse transcriptase (Heider et al., 2013) and protease (Riemenschneider et al., 2016), these approaches have not been applied to genomic data or multi-drug resistance (MDR) in bacteria.

Additional multi-label techniques include the label powerset (LP) method, which acknowledges label dependencies by treating each label combination as a distinct class (Tsoumakas et al., 2009). Another noteworthy method is the random label space partitioning with label powerset (RD), an effective ensemble technique that leverages label powerset with random subsets of k labels (Read et al., 2011, 2021). These methodologies present varying strategies for addressing the complexity of multi-label classification.

1.3.5 Training strategies

Transfer learning

The limited number and skewed distribution of data hinder the accuracy and generalization of model training (Al-Stouhi and Reddy, 2016). This is often the case with medical diagnoses, such as cancer diagnostics, where datasets are typically imbalanced and may contain a disproportionately low number of cancer samples (Al-Stouhi and Reddy, 2016). Training a machine learning model generally requires a substantial number of samples, but such data may not be readily available, particularly for emerging areas like novel antibiotics. This scarcity and imbalance can pose significant challenges to developing robust and reliable predictive models.

Transfer learning (TL) has emerged as a potent solution to challenges posed by imbalanced and limited datasets, particularly in applications like visual and text classification (Zhuang et al., 2020; Chen, 2021; Yu et al., 2020; Mahbod et al., 2020; Radha et al., 2021; Mallesh et al., 2021; Pan and Yang, 2010). Unlike traditional machine learning methods, where there's usually one domain and one task, transfer learning introduces flexibility by allowing for different but related domains and tasks between training and test data (Farahani et al., 2021; Weiss et al., 2016). In essence, transfer learning leverages knowledge from a source domain, which typically consists of a large collection of high-quality, well-labeled data samples, and applies it to a target domain, where data may be scarcer, or labels may be unbalanced (Ebbehoj et al., 2022; Liu et al., 2020). The goal is to improve model performance in the target domain by utilizing the underlying patterns and insights learned from the source domain. This connectivity between domains, where training and test data can vary yet remain contextually linked, sets transfer learning apart and makes it an appealing strategy for cases where obtaining ample and balanced data is problematic (Plested and Gedeon, 2022; Li et al., 2020; Ling Shao et al., 2015; Schwessinger et al., 2020).

Some researchers have effectively utilized transfer learning to address a variety of challenges across different areas. For example, in computer vision, a common approach involves first

training a CNN on the extensive ImageNet dataset (known as pre-training), and then adapting the learned features to a specific task (known as fine-tuning) to solve various problems (Plested and Gedeon, 2022; Gao and Mosalam, 2018). In the area of text classification, the Word2Vec dataset often serves as a foundational pre-training resource (Mikolov et al., 2013). Specific applications of transfer learning have included the work by Gupta et al. (2021) on enhancing predictive analysis on limited data through a cross-property deep transfer learning model. And the work by Park et al. (2021) to explore data heterogeneity and small sample size issues with single-cell data using meta-transfer learning. Medical fields have also seen the successful deployment of transfer learning, especially in situations dealing with imbalanced labels (Okerinde et al., 2021; Weiss and Khoshgoftaar, 2016; Minvielle et al., 2019; Krawczyk, 2016). For instance, Gao and Cui (2020) implemented deep transfer learning to mitigate healthcare disparities stemming from imbalanced biomedical data. They began by training the model on the data from the majority group and then adapted the learned knowledge to the minority groups to enhance performance. This demonstrates the versatile nature of transfer learning, which can be tailored to various tasks, enhancing efficiency and accuracy in areas ranging from visual recognition to healthcare analytics.

Federated learning

The power of machine learning comes from big data, but the real-world scenarios we face in our daily work and life are often only small. For example, in the medical field, the automatic inspection and diagnosis of computed tomography (CT) chest radiographs require a professional doctor to label the data, but the doctor's time is very precious (Yang et al., 2019). This becomes even more challenging when dealing with rare diseases, where the available case data is minimal. Traditionally, the approach to overcome this limitation is to collect data from multiple partner institutions and then train a machine learning model at a central server, which is called centralized training (Yang et al., 2019). However, this requires each participant to upload their data to the central server, making the data of all participants visible to one another and thereby increasing the risk of data leakage (Yang et al., 2019). As data security and privacy are becoming more and more important, many countries have enacted laws on data privacy that limit the sharing of specific data. Federated learning (FL) has emerged as a solution to this dilemma, allowing collaborative training without compromising the privacy and security of individual data sets (Dasaradharami Reddy and Gadekallu, 2023; Rieke et al., 2020; Banabilah et al., 2022).

Federated learning is a decentralized training methodology that utilizes datasets dispersed across various participants (Liu et al., 2023). By using privacy-preserving techniques, it synthesizes information from these diverse sources to build global models cooperatively, all without

centralizing the data or compromising individual privacy (Yang et al., 2019; Kaissis et al., 2020). This approach is useful for privacy preservation and reducing the need to send large amounts of data to a central location.

Federated learning can be categorized into three distinct types based on the relationships between data feature spaces and sample spaces across different data owners: horizontal federated learning (HFL), vertical federated learning (VFL), and federated transfer learning (FTL) (Yang et al., 2019). Here's an overview of each:

Horizontal federated learning (HFL): This approach is applicable when the data of the federated learning participants have overlapping data features, meaning that they share common characteristics but have different data samples (Yang et al., 2019).
Vertical federated learning (VFL): VFL is suited for scenarios where the participants' training data share common data samples, i.e., the data samples are consistent between participants, but the specific data features vary. Unlike HFL, where feature alignment is key, VFL focuses on aligning samples while allowing for differing features (Yang et al., 2019).
Federated transfer learning (FTL): FTL applies when both the data samples and data features among participants have minimal overlap (Xu et al., 2022). In a typical scenario involving two participants, one acts as the source domain while the other represents the target domain. The model learns the distribution of features in the source domain and transfers this knowledge to the target domain (Saha and Ahmad, 2021; Sun, 2022; Ju et al., 2020; Zhang et al., 2022a). Crucially, this transfer process is conducted in a way that ensures the local data remains within its respective domain and does not leave.

In federated learning systems, commonly utilized privacy-preserving techniques encompass methods based on homomorphic encryption (HE), differential privacy (DP), and secure multiparty computation (MPC) (Yang et al., 2019). These methods form a critical layer of protection, safeguarding the integrity and confidentiality of data during the learning process. The Python open-source package provides a rich set of privacy-preserving implementations, such as the package Pycrypto is commonly used in encryption/decryption algorithms, and the Paillier package provides an implementation that supports partial homomorphic encryption (Yang et al., 2019).

FL is widely used in the medical field. Bai et al. (2021), Dayan et al. (2021), and Dou et al. (2021) applied deep learning models combined with FL training strategies for coronavirus disease (COVID) diagnosis. Several studies have focused on cancer and disease diagnosis using FL and machine learning models (Pati, 2022; Ogier du Terrail et al., 2023). Pati (2022) conducted the most comprehensive FL study to date, encompassing data from 71 locations

across six continents, to develop an automated tumor boundary detection system specifically for glioblastoma, a rare disease. With a dataset comprising 6,314 cases, the largest of its kind reported in the literature, they demonstrated that their model outperformed a publicly trained model. Ogier du Terrail et al. (2023) explored the application of ML, utilizing whole-slide images and clinical data, to predict the histological response to neoadjuvant chemotherapy in early-stage triple-negative breast cancer (TNBC) patients. To circumvent the limitations of small-scale studies and simultaneously maintain data privacy, they carried out a multicentric TNBC study employing federated learning. In this approach, patient information remained securely protected behind the firewalls of individual hospitals. And Wu et al. (2022) introduced a federated graph neural network (GNN) framework known as FedPerGNN. This framework enables collaborative training of GNN models using decentralized graphs inferred from local data, all while employing a privacy-preserving model update method. To enrich the utilization of graph information beyond mere local interactions, they implemented a privacy-preserving graph extension protocol that responsibly integrates higher-order information. Personalized validation was conducted on six distinct datasets across various scenarios. The findings demonstrate that FedPerGNN effectively achieves high performance while also maintaining robust privacy preservation.

Swarm learning

Federated learning alleviates certain concerns by ensuring that data is retained locally, effectively dealing with local confidentiality issues (Warnat-Herresthal et al., 2021). However, the model parameters continue to be managed by central custodians, a factor that centralizes authority (Warnat-Herresthal et al., 2021). Additionally, the adoption of star-shaped architectures in this approach reduces fault tolerance, creating potential weaknesses within the system (Warnat-Herresthal et al., 2021). Warnat-Herresthal et al. (2021) introduced swarm learning (SL), a groundbreaking decentralized machine learning approach that combines edge computing and blockchain-enabled peer-to-peer networking. Unlike traditional federated learning, swarm learning maintains data confidentiality and coordination without a central coordinator, offering an enhanced and more secure method of distributed learning. Warnat-Herresthal et al. (2021) demonstrated the feasibility and efficacy of employing swarm learning to create classifiers for various diseases, including COVID-19, tuberculosis, leukemia, and lung lesions on distributed data. The results of Warnat-Herresthal et al. (2021) also indicated that swarm learning classifiers exhibit superior performance compared to classifiers trained on local data alone.

Bai et al. (2021) developed the unified CT-COVID AI diagnostic initiative, employing a federated learning framework that allows the AI model to be trained distributively and run in-

dependently at each host institution without the necessity of data sharing. Specifically, participants first download and train three-dimensional CNN models using their local cohort data. Once trained, the model parameters are encrypted and sent back to the server. The server then combines the contributions from each participant to create the federated model without having direct access to or explicit knowledge of the individual parameters.

1.3.6 Evaluation metrics

ML model evaluation is an essential part of the development process, as it allows you to understand how well the model is performing. Various metrics can be used, depending on the type of problem you are addressing. Accuracy, precision, and recall are fundamental evaluation metrics for classification models, each serving to quantify different aspects of a model's performance (Vakili et al., 2020).

Accuracy
This metric quantifies the overall correctness of the model by measuring the fraction of both true positive and true negative predictions overall predictions (Vakili et al., 2020). In the context of binary classification, it can be expressed mathematically as:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

Where TP = True Positives, TN = True Negatives, FN = False Negatives, FP = False Positives.

Precision
Precision focuses on the correctness of the positive predictions, representing the ratio of true positive predictions to the total number of positive predictions (true positives plus false positives). It is particularly concerned with minimizing false positive errors (Vakili et al., 2020).

$$Precision = \frac{TP}{TP+FP}$$

Recall
Also known as sensitivity or true positive rate, recall measures the proportion of actual positive samples that are correctly identified (Vakili et al., 2020). It is especially useful when the cost of missing a positive sample (false negative) is high.

$$Recall = \frac{TP}{TP+FN}$$

ROC Curve
The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. It illustrates the tradeoff between correctly identifying positive instances and mistakenly identifying negative instances as positive. A model with a perfect

discriminative ability would result in a curve that hugs the upper left corner, and the area under the ROC curve (AUC-ROC) would be 1.

Precision-Recall curve
The Precision-Recall (PR) curve shows the relationship between precision and recall for different thresholds. Unlike the ROC curve, it focuses solely on the positive class, making it more informative for imbalanced datasets where the positive class is the minority. A higher area under the PR curve (AUC-PR) generally indicates better model performance.

Both of these curves offer insights into a model's performance, but neither is a one-size-fits-all solution. While they provide a comprehensive view of a model's ability to distinguish between classes, they may not always be the most appropriate metrics for heavily imbalanced datasets, particularly when the focus is on the performance related to the minority class.

F1 score
In such cases, metrics like the F1 score, which combines precision and recall into a single value, or custom evaluation metrics tailored to the specific context and requirements, might be more suitable.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

MCC
The Matthews Correlation Coefficient (MCC) is a robust metric used to evaluate the performance of classification models, particularly when dealing with imbalanced datasets. Calculated based on the Pearson correlation coefficient, the MCC ranges from -1 to 1, where "+1" indicates a perfect prediction, "0" represents no better than the random prediction, and "-1" indicates total disagreement between prediction and actual observation (Boughorbel et al., 2017). The strength of the MCC lies in its balanced consideration of true and false positives and negatives, making it a valuable measure when the classes are of different sizes (Boughorbel et al., 2017). In the context of imbalanced datasets, where traditional metrics like accuracy may be misleading, the MCC offers a more nuanced assessment of a model's performance, ensuring that both classes are fairly represented in the evaluation.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}}$$

Given the variability in our datasets, with some being balanced and others extremely imbalanced, relying on a single metric may not adequately capture the overall performance of our model. As a result, we've conducted a comprehensive evaluation using a combination of the metrics mentioned earlier. This multifaceted approach ensures a more nuanced understanding of the model's effectiveness, taking into consideration both the accuracy and the unique

challenges posed by imbalanced data.

Hamming loss and 0/1 loss

The Hamming loss and 0/1 loss are commonly used for the evaluation of MLC models. Hamming loss refers to the proportion of labels that are inaccurately predicted, serving as a gauge for individual label prediction errors across all classes. On the other hand, 0/1 loss examines the correctness of the entire set of predicted labels for a given instance, quantifying the percentage of instances where the full set of predicted labels does not exactly match the true labels. Thus, while Hamming loss provides a finer-grained label-by-label error rate, 0/1 loss offers insight into the overall prediction accuracy of entire label sets.

1.3.7 Application of ML to AMR

Recent studies have underscored the potential of machine learning methods in predicting AMR. By integrating sequencing methodologies with well-established databases and phenotypic information related to AMR, these innovative approaches are laying the groundwork for more precise predictions and actionable insights (Boolchandani et al., 2019; Liu et al., 2020; Lv et al., 2021). For instance, Yang et al. (2018) developed machine learning models using DNA sequencing data from 1839 UK bacterial isolates to classify *Mycobacterium tuberculosis* resistance to eight anti-tuberculosis drugs and to identify multi-drug resistance. However, their models were not based on genome-wide sequence information, they selected 23 known drug-resistance candidate genes and mutations in these 23 candidates and then constructed the models. It has some limitations, such as the prediction of new resistance genes and resistance mechanisms can be restricted. Most studies have employed a similar approach for classifying resistance, determining its presence or absence based on predetermined libraries of variants found in the existing literature (Kouchaki et al., 2019; Moradigaravand et al., 2018; Van Camp et al., 2020; Pesesky et al., 2016).

Deep learning algorithms have also demonstrated substantial potential in predicting new antibiotic drugs, identifying AMR genes, and recognizing AMR peptides (Arango-Argoty et al., 2017; Stokes et al., 2020; Veltri et al., 2018; Popa et al., 2022; Liu et al., 2023; Veltri et al., 2018). Stokes et al. (2020) developed a deep neural network capable of identifying molecules with antimicrobial properties. By applying this network to multiple chemical libraries, they discovered a unique molecule from the Drug Repurposing Hub, namely halicin. Distinct from conventional antibiotics in its structure, halicin demonstrated bactericidal activity against a diverse array of pathogens, including those from the broad phylogenetic spectrum such as *Mycobacterium tuberculosis* and carbapenem-resistant *Enterobacteriaceae* bacteria. Li et al. (2021) proposed a multi-task deep learning framework called HMD-ARG. Initially,

they collected and cleaned resistance gene sequences from seven well-established ARG (Antibiotic Resistance Gene) databases and got the final resulting database, HMD-ARG-DB. This comprehensive collection consists of 17,282 high-quality sequences, coupled with labels of 15 antibiotic classes, 6 underlying resistance mechanisms, and their mobility. Subsequently, HMD-ARG was employed for ARG annotation, encompassing three distinct dimensions: type of antibiotic resistance, underlying mechanism, and gene mobility. Arango-Argoty et al. (2017) developed two deep learning models, DeepARG-SS and DeepARG-LS, based on the metagenome data, for predicting ARGs in short reads and full gene length sequences, respectively. They first also collected ARGs from three major databases: CARD, ARDB, and UNIPROT, and then constructed models based on the presence or absence of resistance genes.

To summarize, machine learning and deep learning have a wide range of applications in AMR detection, new AMR gene prediction, and new antibiotic development.

1.4 Challenges and Motivation

Although these studies applied machine learning to facilitate the prediction of AMR, most of the research models were constructed by focusing only on features related to resistance genes and resistance variants, with a high dependence on previous AMR reference databases, without constructing models from genome-wide features. The predictions may be missing some new features of AMR-related genes and variants. Thus, the development of genome-wide machine learning models to rapidly and accurately detect AMR without prior knowledge of AMR is a significant addition to existing methods.

Another challenge regarding AMR research is that current methods typically focus on single-drug resistance prediction and do not include information on antimicrobial resistance characteristics that accumulate over time. Therefore, rapid identification of multi-drug resistance simultaneously remains a challenge. In our study, we will explore multiple multi-label classification approaches for multidrug resistance modeling of pathogens.

Limited training samples and data imbalance hinder the generalization performance and overall accuracy of the model, which is an important challenge in the AMR detection and development of new antibiotics and a more generalized challenge in the medical field. Therefore, in this study, we will utilize transfer learning to improve this problem.

Data security and privacy have become paramount in machine learning model training. Swarm learning offers a solution by ensuring data remains local during training. This approach not only minimizes the transfer of sensitive data to a centralized server but also enhances train-

ing efficiency by distributing computational tasks. Additionally, it allows models to quickly adapt to emerging data trends, given the continuous updates throughout the network. Therefore, we will explore the application of swarm learning on AMR.

1.5 Aims

The purpose of this dissertation is to apply machine learning to facilitate AMR-related research. Specifically, the first part of the work focuses on the development of fast and accurate detection models for AMR as well as the identification of new AMR genes and mutations. In the second work, we delve into five different MLC approaches dedicated to the problem of multidrug resistance prediction. In the third work, we develop deep transfer learning to facilitate the ability to generalize models with small numbers and label imbalances. Finally, in our fourth work, we employ swarm learning to address the challenges of data privacy and security during AMR model training.

1.6 List of Publications

The publications and contributions during my Ph.D. period are listed below.

Publication 1

Yunxiao Ren, Trinad Chakraborty, Swapnil Doijad, Linda Falgenhauer, Jane Falgenhauer, Alexander Goesmann, Anne-Christin Hauschild, Oliver Schwengers, Dominik Heider. ***Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning***. Bioinformatics, 2022, 38(2), 325-334.

Contributions

D.H. conceived and supervised the study; **Y.R.** analyzed the genome data, developed the machine learning analysis pipeline, and drafted the manuscript; S.D., L.F., and J.F. collected the raw sequencing and antimicrobial resistance (AMR) data. O.S. pre-processed the sequencing data and clinical data. D.H., T.C., and A.G. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Publication 2

Yunxiao Ren, Trinad Chakraborty, Swapnil Doijad, Linda Falgenhauer, Jane Falgenhauer, Alexander Goesmann, Oliver Schwengers, Dominik Heider. ***Multi-label classification for multi-drug resistance prediction of Escherichia coli.*** Computational and Structural Biotechnology Journal, 2022, 20: 1264-1270.

Contributions

D.H. conceived and supervised the study; **Y.R.** analyzed the genome data, developed the multi-label classification pipeline, and drafted the manuscript; S.D., L.F., and J.F. collected the raw sequencing and antimicrobial resistance (AMR) data. O.S. pre-processed the sequencing data and clinical data. D.H., T.C., and A.G. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Publication 3

Yunxiao Ren, Trinad Chakraborty, Swapnil Doijad, Linda Falgenhauer, Jane Falgenhauer, Alexander Goesmann, Oliver Schwengers, Dominik Heider. ***Deep Transfer Learning Enables Robust Prediction of Antimicrobial Resistance for Novel Antibiotics***. Antibiotics, 2022, 11(11): 1611.

Contributions

D.H. conceived and supervised the study; **Y.R.** analyzed the genome data, constructed the transfer learning pipeline, and drafted the manuscript; S.D., L.F., and J.F. collected the raw sequencing and antimicrobial resistance (AMR) data. O.S. pre-processed the sequencing data and clinical data. D.H., T.C., and A.G. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Other contributions not included

Yunxiao Ren, Carmen Li, Dulmini Nanayakkara Sapugahawatte, Chendi Zhu, Sebastian Spänig, Dorota Jamrozy, Julian Rothen, Claudia A Daubenberger, Stephen D Bentley, Margaret Ip, Dominik Heider. ***Predicting hosts and cross-species transmission of Streptococcus agalactiae by machine learning.*** Under Review.

Contributions

Y.R., D.H. and M.I. conceived the study; **Y.R.**, C.L, D.S., C.Z. constructed machine learning methods and related subsequent analysis.**Y.R.**, D.H., C.L., and C.Z. wrote the manuscript. S.S. helped with the DAAD funding application. D.J. and S.D.B partially performed whole genome sequencing on the isolates. J.R. and C.A.D partially provided GBS sequencing dataset. D.H. and M.I. supervised this whole project and revised the manuscript. All authors read and approved the final manuscript.

2 Methods

2.1 Datasets Overview

The species in our work are all based on *E. coli* bacteria. *E. coli* is one of the predominant bacterial agents related to hospital-induced infections and AMR (Shankarnarayan et al., 2022). Serving as an important model organism, it offers valuable insights into severe infections in humans and animals (Poirel et al., 2018). Given the sufficient data available on this species, we have selected it as the basis for developing our model for AMR prediction.

In the first paper, we utilized two datasets: the Giessen data and the public data. The Giessen dataset, specifically collected from our study, contains WGS data along with corresponding phenotypic information related to various antibiotics for a total of 987 *E.coli* strains. These strains were extracted from both animal and human clinical samples. AST was conducted using the VITEK® 2 system (bioMérieux, Nürtingen, Germany) and interpreted in alignment with EUCAST guidelines. The second dataset, referred to as the public dataset, comprises WGS information for 1509 *E.coli* strains, along with corresponding phenotypic data (as documented by Moradigaravand et al. (2018). In the scope of our study, we narrowed our focus to four specific antibiotics: ciprofloxacin (CIP), cefotaxime (CTX), ceftazidime (CTZ), and gentamicin (GEN).

In the second paper, the raw dataset is the same as the Giessen data in the first paper. In

order to do MLC, the isolates need to be filtered for missing antibiotic resistance information. Thus, the final dataset with complete MDR information contains 809 *E. coli* strains.

In the third paper, we utilized two datasets. The first dataset, containing 809 *E. coli* strains, is consistent with the one used in the second paper. The second dataset consists of 1509 *E. coli* strains collected from public sources, which is the same as the data used in the first paper.

In the fourth work, we used three datasets. The first dataset at node 1 was from Giessen, and the second dataset as the test set was from the public source. The third dataset at node 2 was collected from the Chinese University of Hong Kong. See Table 2.1 for more information

2.2 Whole Genome Sequencing Analysis

The raw whole-genome sequencing reads underwent an initial quality assessment and were subsequently filtered low-quality reads using fastp (Chen et al., 2018). The clean reads were then aligned to the *E. coli* reference genome (specifically, the *E. coli* K-12 strain MG1655) using BWA-mem (Li et al., 2009). Variants were then called using Bcftools (Danecek et al., 2021), while the aligned reads were sorted through Samtools (Li and Durbin, 2009). Finally, vcftools (Danecek et al., 2011) was employed to filter the raw variants. All tools were applied using their default parameters.

Firstly, we extracted the reference and variant alleles along with their respective positions. Then we merged all isolates based on the location of the reference alleles. Loci without variation were filtered out (with an "N" designating a locus lacking variation), leading to the construction of the final SNP matrix. In this matrix, rows correspond to individual samples, while columns represent the various variant alleles.

2.3 Sequences Encoding

To prepare the SNPs for machine learning analysis, we employed three encoding techniques: label encoding, one-hot encoding, and FCGR encoding. In label encoding, the nucleotide bases A, G, C, T, and N in the SNP matrix were mapped to numerical values 1, 2, 3, 4, and 0, respectively. With one-hot encoding, the DNA sequence was transformed into a binary matrix and subsequently vectorized using OneHotEncoder from preprocessing class in Scikit-learn python package (Pedregosa et al., 2011). In the case of FCGR encoding, we utilized the R package kaos to convert the sequences into an image-like matrix, setting the resolution at 200 (Löchel et al., 2020).

2.4 Model Training and Evaluation

We constructed four distinct machine learning methods, including LR, SVM, RF, and CNN. We used the Scikit-learn python package to implement LR, SVM, and RF (Pedregosa et al., 2011). LR was configured with default parameters, except for an increase to 1000 iterations. RF was applied using default parameters, including a forest of 200 trees. For SVM, we used a linear kernel algorithm with default parameters.

We implemented CNNs using the Keras (https://keras.io/) library and TensorFlow library (https://tensorflow.org). Our CNN architecture consisted of eleven hidden layers, specifically encompassing four convolutional layers, two batch normalization layers, two pooling layers, one flattening layer, one fully connected layer, and one dropout layer. The CNN structure for both label encoding and one-hot encoding is the same, while differs from FCGR encoding in the convolutional layers and pooling layers. For FCGR, we used the Conv2D and MaxPooling2D functions, whereas the CNN for label encoding and one-hot encoding used the 1D versions instead.

In the CNN architecture, the first two convolutional layers utilized eight filters, each with a kernel size of three, a rectified linear unit (ReLU) activation function, and 'same' padding to maintain the spatial dimensions. The latter two convolutional layers were designed with 16 filters each. All pooling layers in the network employed a pool size of two for spatial down-sampling. The final fully connected layer featured a softmax activation function for class probability estimation. For the training process, we compiled the model using the Adam optimization algorithm, complemented by cross-entropy loss as the objective function.

We fine-tuned the machine learning models through a rigorous optimization process, utilizing five iterations of 5-fold stratified cross-validation. To address class imbalance in the training set, an up-sampling strategy was implemented. For the definitive evaluation conducted on the public data, we assessed performance on both the unmodified public dataset and a balanced version, the latter achieved through a down-sampling strategy.

Model performance was evaluated using several metrics. We plotted the receiver operating characteristic curve (ROC) and computed the AUC to measure the models' ability to distinguish between classes. Additionally, we calculated precision and recall for all models, providing a more comprehensive view of their effectiveness. To conduct statistical comparisons between the models, we applied the DeLong test (Demler et al., 2012), a widely recognized method for evaluating differences in AUC.

2.5 Gene Annotation

To uncover the specific SNPs linked to resistance, we carried out a marker gene identification process utilizing the EFS R package (Neumann et al., 2017). This package integrates eight distinct feature selection methods, all tailored for binary classification tasks (Neumann et al., 2016). We engaged EFS with its default parameters to ensure consistency with established practices. Following the identification of relevant SNPs, we annotated the corresponding genes using the SnpEff software (Cingolani et al., 2012), a specialized tool for variant annotation and effect prediction.

2.6 Multi-label Classification

In this study, we employed various algorithms, including Binary Relevance (BR), Classifier Chain (CC), Ensembled Classifier Chain (ECC), Label Powerset (LP), and Random label space partitioning with label powerset (RD) method for the multi-label classification of MDR in bacteria. BR is often used as a reference model for comparison in multi-label classification scenarios.

To elaborate further, let $L = \{\lambda_1, \ldots, \lambda_m\}$ (where m >1) represent a finite set of class labels corresponding to resistance to specific antibiotics, and let X denote the space of the SNPs, or the instance space. The training set S for MLC can then be defined as $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where these pairs are generated independently and identically according to a probability distribution $P(X)$, over the Cartesian product $X \times Y$. Here, Y represents the set of all possible combinations of labels.

BR tackles a dataset with L labels by dividing it into L separate binary classification problems. In our context, we separated the data into four binary classification challenges, each corresponding to one of the antibiotics (CIP, CTX, CTZ, and GEN).

Contrastingly, the CC method forms a "chain" linking the L binary classifiers. In this scheme, the prediction from one classifier serves as an additional input for all subsequent classifiers in the chain. This design enables the capture of potential dependencies between labels, addressing a limitation in BR. However, CC's performance is highly sensitive to the chain's order. To mitigate this issue, the ECC was introduced, which combines multiple chains with varying orders through majority voting, as proposed by Read et al. (2021).

The LP approach simplifies a multi-label problem into a single-label multi-class issue by training on all unique label combinations found in the training data. Alternatively, the RD method

partitions the label space into groups of size k, training an LP classifier for each group, and then aggregates the predictions from all LP classifiers.

It is worth noting that any conventional binary classification method can be employed in these multi-label strategies. In our study, we specifically evaluated RF, LR, and SVM for the multi-label classification of MDR in bacteria.

2.7 Basic CNN Model

We employed the Keras and Tensorflow Python packages to construct our CNN models. After evaluating various topologies on the training data, a 12-layer structure emerged as the optimal design. This architecture encompasses twelve layers: four convolutional layers each with a kernel size of 3 (implemented using the Conv1D function), two pooling layers (utilizing the MaxPooling1D function), a pair of batch normalization layers, a flattening layer, a fully connected layer containing 128 nodes followed by a dropout layer, and a final output layer employing the "softmax" activation function. The CNN models were compiled using the "categorical_crossentropy" loss function and the "Adam" optimizer, with training carried out over 50 epochs. To enhance computation efficiency, the data was divided into multiple small batches, each containing 8 samples.

2.8 Deep Transfer Learning Architecture

To enhance model performance on small, imbalanced datasets, we implemented deep transfer learning, extending the basic CNN architecture previously detailed. Transfer learning requires specification of both the source and target domains (Ds and Dt, respectively) and tasks (Ts and Tt, respectively) (Cai et al., 2020). In our study, the CIP dataset from our laboratory served as the source domain Ds, while the CTX, CTZ, and GEN datasets constituted the target domain Dt. The tasks Ts and Tt were focused on predicting AMR against various antibiotics.

We executed two distinct transfer learning strategies, namely fine-tuning and freezing. The fine-tuning approach involves a common deep transfer learning method in which the parameters (or weights) from the source domain model (Ds) are transferred to the target domain models (Dt) (Cai et al., 2020). In our implementation, the parameters trained on the CIP dataset were transferred into the models for CTX, CTZ, and GEN. Additionally, to prevent overfitting, we employed the freezing strategy (Mallesh et al., 2021), where two normalization layers and one convolution layer were kept constant, allowing the remaining layers to be retrained in the CNN models.

2.9 Swarm Learning

2.9.1 Swarm learning framework

The principle of Swarm Learning (SL) lies in collaboratively constructing machine learning models across separate computer systems, utilizing private data at each node. This is achieved by sharing parameters across a Swarm network. Unlike the FL, SL operates without the necessity for a centralized server to oversee these parameters. Here, we apply SL to independent data from two distinct nodes. The first dataset, referred to as Node_1, is garnered from Giessen, comprising 809 *E. coli* samples with AMR information against four drugs: CIP, CTX, CTZ, and GEN (Table 2.1). The second dataset, Node_2, originates from Hong Kong, containing *E. coli* samples tested against CIP (n = 979), CTX (n = 977), CTZ (n = 971), and GEN (n = 980) (Table 2.1). After each training session, model weights are synchronized across the nodes. These weights are then averaged during each synchronization event, and subsequent training at each node employs these averaged parameters. The SL framework is efficiently implemented in Python.

2.9.2 Algorithm in swarm learning

We build the CNN as the foundational algorithm within the SL framework, employing the Python packages Keras and TensorFlow. The architecture of our CNN model is comprised of 13 layers. This includes four convolutional layers with a kernel size of three, made possible with the Conv1D function. The model also encompasses two pooling layers, utilizing the MaxPooling1D function, along with two batch normalization layers. Further structure includes a flattening layer, a fully connected layer consisting of 128 nodes, and two dropout layers. The final layer is the output layer, utilizing the "softmax" activation function. We use the "categorical_crossentropy" loss function and the "Adam" optimizer function to compile the CNN models, running it through 50 epochs for optimal performance. To enhance computational speed, the data is partitioned into multiple smaller batches, each containing 16 data points.

2.9.3 Performance comparison

we benchmark the model's performance within the SL framework against both local and centralized training modes. Each dataset is trained independently on each node at the local mode. Conversely, the centralized mode involves training the model on a combined dataset from two nodes. The performance evaluation of the models in these three distinct modes is conducted using independent test data obtained from public sources. This data comprises

E. coli samples tested against CIP (n = 1496), CTX (n = 1428), CTZ (n = 1471), and GEN (n = 1489) (Table 2.1).

Table 2.1: Dataset overview. The local training modes at Node_1 and Node_2 are referred to as Local_1 and Local_2, respectively. The data size for the centralized mode is a combination of Node_1 and Node_2. For the class label, 'R' denotes resistance, while 'S' indicates sensitivity.

Drugs	Nodes	Size	R/S	R/S (%)
CIP	Node_1	809	366/443	45.2/54.8
CIP	Node_2	979	366/613	37.4/62.6
CIP	Test	1496	267/1229	17.8/82.2
CTX	Node_1	809	358/451	44.3/55.7
CTX	Node_2	977	257/720	26.3/73.7
CTX	Test	1428	115/1313	8.1/91.9
CTZ	Node_1	809	276/533	34.1/65.9
CTZ	Node_2	971	62/909	6.4/93.6
CTZ	Test	1471	73/1398	5.0/95.0
GEN	Node_1	809	188/621	23.2/76.8
GEN	Node_2	980	336/644	34.3/65.7
GEN	Test	1489	101/1388	6.8/93.2

3 Results

This section will provide a comprehensive overview of publications related to the dissertation. Each sub-section begins with an extended abstract, followed by the associated manuscript. The first study focuses on the application of machine learning to AMR based on whole-genome sequencing, with the goal of constructing different ML models that do not rely on prior knowledge for accurate AMR prediction and identification of new AMR-associated mutations and genes (Ren et al., 2021). The second work delves into the problem of multi-drug resistance (MDR) problems with the aim of exploring the performance of different multi-label classification (MLC) methods for MDR prediction (Ren et al., 2022a). The third article studies how to address the challenges of data limitation and labeling imbalance that machine learning encounters in training (Ren et al., 2022b). The fourth work ((Unpublished)) focuses on applying swarm learning to cope with data privacy issues in AMR prediction, and since this work has not yet been published, I will briefly describe the motivation and main results of this work.

3.1 Publication 1: Prediction of Antimicrobial Resistance based on Whole-genome Sequencing and Machine Learning.

3.1.1 Summary

Aim and Motivation

The aim of this study (Ren et al., 2021) was to conceive and validate potent machine learning methodologies that can accurately predict antimicrobial resistance (AMR) using whole-genome sequencing data, without relying on any pre-existing knowledge. Additionally, we also strove to discover novel mutations and genes associated with AMR. As the world grapples with the growing problem of AMR, which threatens both human and animal health, the urgency of a rapid and accurate method for AMR detection cannot be overemphasized. Traditional antimicrobial susceptibility testing (AST) strategies have significant drawbacks, including time-consuming, limited throughput, and limitations on culturable bacteria. Machine learning offers a promising solution in this scenario, with its potential to automate AMR prediction using bacterial genomic data. However, the exploration and comparison of various machine learning methodologies to predict AMR, especially while employing different encodings and whole-genome sequencing data without pre-existing knowledge, is a field yet to be extensively explored. Therefore, our study sets out to bridge this gap and contribute to the development of effective solutions for this pressing global issue.

Methods and Results

In our study, we initially collected two whole genome sequencing (WGS) datasets of *E.coli*, the Giessen data consisting of 987 samples, and a public dataset incorporating 1509 samples. Following this, we performed SNP variant calling, focusing on the elimination of only low-quality data rather than filtering data according to known AMR databases. Consequently, we utilized the resulting SNP matrix, where the rows represent the samples and columns are the variant alleles, and corresponding phenotype data relating to four antibiotics, namely ciprofloxacin (CIP), cefotaxime (CTX), ceftazidime (CTZ) and gentamicin (GEN), as input for the subsequent analyses.

Subsequently, we employed three encoding methods including label encoding, One-Hot encoding, and FCGR to transform the sequence into a format that machine learning can use. We then developed four distinct machine learning models, including Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), and Convolutional Neural Network (CNN). We evaluated the performance of these models through cross-validation and testing on independent data. Our findings demonstrated the efficacy of these models in

AMR prediction, with RF and CNN notably outperforming LR and SVM, achieving AUC score of up to 0.96. There was no significant difference between the three coding methods, indicating that all of these methods can be effectively applied to encoding genomic sequences. Lastly, we identified mutations and genes associated with AMR by ensemble feature selection and genome annotation.

Conclusion

This research signifies a critical advancement in the field of AMR prediction. By employing machine learning models and diverse encoding methods on genomic data, we have laid the foundation for a more efficient and informed approach to combat AMR. The knowledge derived from this research could profoundly transform our approach to detecting and managing antimicrobial resistance, potentially playing a vital role in the protection of global health.

Bioinformatics, 38(2), 2022, 325–334
doi: 10.1093/bioinformatics/btab681
Advance Access Publication Date: 6 October 2021
Original Paper

Genome analysis

Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning

Yunxiao Ren[1], Trinad Chakraborty[2,3], Swapnil Doijad[2,3], Linda Falgenhauer[3,4,5], Jane Falgenhauer[2,3], Alexander Goesmann[3,6], Anne-Christin Hauschild[1], Oliver Schwengers[3,6] and Dominik Heider [1,*]

[1]Department of Data Science in Biomedicine, Faculty of Mathematics and Computer Science, Philipps-University of Marburg, Marburg 35032, Germany, [2]Institute of Medical Microbiology, Justus Liebig University Giessen, Giessen 35392, Germany, [3]German Center for Infection Research, Partner site Giessen-Marburg-Langen, Giessen 35392, Germany, [4]Institute of Hygiene and Environmental Medicine, Justus Liebig University Giessen, Giessen 35392, Germany, [5]Hessisches universitäres Kompetenzzentrum Krankenhaushygiene, Giessen 35392, Germany and [6]Department of Bioinformatics and Systems Biology, Justus Liebig University Giessen, Giessen 35392, Germany

*To whom correspondence should be addressed.
Associate Editor: Inanc Birol

Received on June 29, 2021; revised on August 27, 2021; editorial decision on September 17, 2021; accepted on September 24, 2021

Abstract

Motivation: Antimicrobial resistance (AMR) is one of the biggest global problems threatening human and animal health. Rapid and accurate AMR diagnostic methods are thus very urgently needed. However, traditional antimicrobial susceptibility testing (AST) is time-consuming, low throughput and viable only for cultivable bacteria. Machine learning methods may pave the way for automated AMR prediction based on genomic data of the bacteria. However, comparing different machine learning methods for the prediction of AMR based on different encodings and whole-genome sequencing data without previously known knowledge remains to be done.

Results: In this study, we evaluated logistic regression (LR), support vector machine (SVM), random forest (RF) and convolutional neural network (CNN) for the prediction of AMR for the antibiotics ciprofloxacin, cefotaxime, ceftazidime and gentamicin. We could demonstrate that these models can effectively predict AMR with label encoding, one-hot encoding and frequency matrix chaos game representation (FCGR encoding) on whole-genome sequencing data. We trained these models on a large AMR dataset and evaluated them on an independent public dataset. Generally, RFs and CNNs perform better than LR and SVM with AUCs up to 0.96. Furthermore, we were able to identify mutations that are associated with AMR for each antibiotic.

Availability and implementation: Source code in data preparation and model training are provided at GitHub website (https://github.com/YunxiaoRen/ML-iAMR).

Contact: dominik.heider@uni-marburg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The rise of antimicrobial resistance (AMR) is one of the greatest threats to global health, food security and societal development. Estimates indicate that the number of yearly deaths will be at 10 million worldwide with a cost of $100 trillion if no steps to tackle AMR are taken by 2050 (Naylor *et al.*, 2018). Traditional antimicrobial susceptibility testing (AST) is widely used for AMR analysis in clinical practice. However, this approach requires professional facilities and technicians for implementation and is viable only for cultivable bacteria (Boolchandani *et al.*, 2019). Recently, many studies highlight the potential of machine learning methods in predicting AMR combining sequencing methods and well-known databases with phenotypic information for AMR (Boolchandani *et al.*, 2019; Liu *et al.*, 2020; Lv *et al.*, 2021). For instance, Yang *et al.* (2018) and Kouchaki *et al.* (2018) analyzed AMR using different machine learning algorithms [e.g. support vector machine (SVM), logistic regression (LR) and random forest (RF)] trained on whole-genome sequencing and achieved high accuracy on AMR prediction. Deep learning algorithms also showed

© The Author(s) 2021. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article/38/2/325/6382301 by Universitaetsbibliothek Marburg user on 18 May 2022

significant potential for predicting new antibiotic drugs, AMR genes and AMR peptides (Arango-Argoty *et al.*, 2018; Stokes *et al.*, 2020; Veltri *et al.*, 2018). However, these studies focused on genome variants (such as single-nucleotide polymorphisms, SNPs) or other features only related to resistant genes identified in previous studies or resistant databases. The potential of machine learning models for predicting AMR without using known resistance mutation databases or annotated genes remains to be clarified.

To use machine learning methods for the classification of AMR, the input sequences (here: genomic sequences) need to be encoded into numerical values. A practical and informative encoding method for the whole-genome sequence is, thus, crucial for downstream

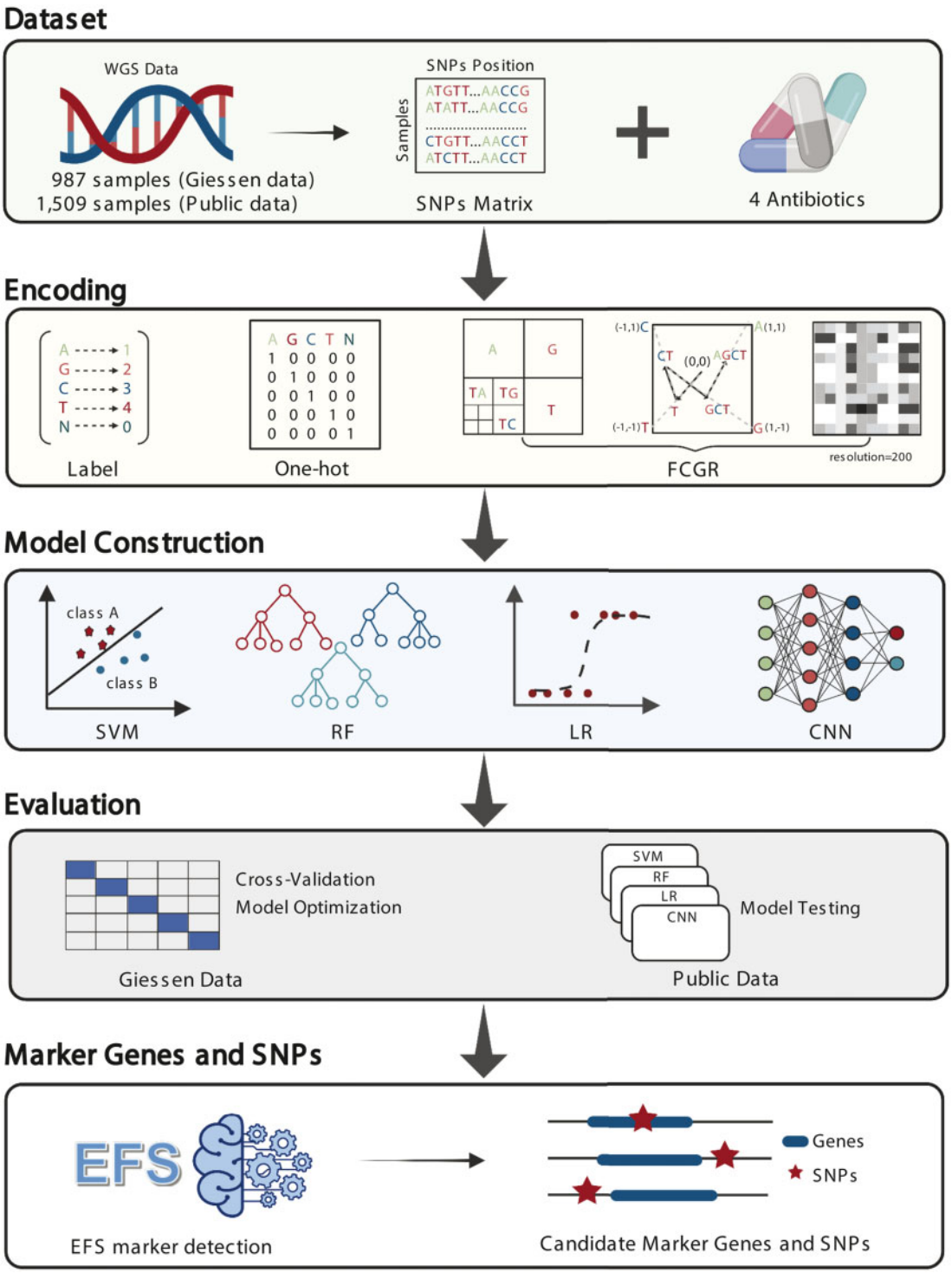

Fig. 1. Workflow of the study. WGS data from Giessen and the public data from Moradigaravand *et al.* (2018) were processed, and single nucleotide polymorphisms (SNPs) were called. The SNP data were encoded by label encoding, one-hot encoding and FCGR encoding for subsequent machine learning. The Giessen dataset was used to train and validate the four machine learning algorithms using cross-validation. The public data were used for the final evaluation of the models. Finally, we analyzed the association of SNPs and SNPs-adjacent genes with AMR using EFS. Created with BioRender.com

Downloaded from https://academic.oup.com/bioinformatics/article/38/2/325/6382301 by Universitaetsbibliothek Marburg user on 18 May 2022

analysis. There are various encoding methods for sequences (Spänig and Heider, 2019), e.g. one-hot encoding or label encoding. One-hot encoding, also referred to as sparse encoding, encodes the DNA sequence into a binary matrix, which is then vectorized and used as input for the machine learning models. Label encoding is another simple and straightforward encoding method, where each label is assigned a unique integer.

Thus, in this study, we use label encoding, one-hot encoding and Chaos Game Representation (CGR) to encode the genomic data. CGR is a recurrent iterative function system, which can be used to visualize sequences by building fractals from sequences of symbols, i.e. from an alphabet $\mathcal{A} = \{S_1, \ldots, S_n\}$. Jeffrey (1990) was the first who applied the CGR algorithm to DNA sequences, i.e. $n = 4$ and $\mathcal{A} = \{A, C, G, T\}$, thus the resulting fractals are constructed from squares. Since the development of the CGR and its application in life science, it has been used for the analysis and alignment-free comparison of whole-genome sequences (Joseph and Sasikumar, 2006; Kania and Sarapata, 2021; Lichtblau, 2019). It has been shown that CGR is an excellent representation for genomes and that CGR-driven phylogeny leads to reliable predictions (Deschavanne *et al.*, 1999). In particular, the comparison between genomes using CGR is straightforward and fast (Hoang *et al.*, 2016). CGR has been used, for instance, for a fast comparison of SARS-CoV2 strains (Sengupta *et al.*, 2020). Extensions of CGR include color grids (Deschavanne *et al.*, 1999) and frequency matrix chaos game representation (FCGR) (Almeida *et al.*, 2001). Wang *et al.* (2005) used FCGR to calculate the image distance between genomes to generate phylogenetic trees. Rizzo *et al.* (2016) showed that deep neural networks (DNNs) trained on genomes encoded with FCGR yielded very accurate predictions. They used a convolutional neural network (CNN) to divide bacteria into three different phyla, order, family and genus, and showed very high accuracy for the method.

While most existing studies on CGR encoding focused on CGR for DNA, there also exist a smaller number of studies dealing with other alphabets, e.g. the encoding of protein sequences. Yu *et al.* (2004) used the CGR algorithm for protein classification by separating the amino acids into four groups based on their properties and used multifractal and correlation analysis to construct a phylogenetic tree of Archaea and Eubacteria. In other approaches, the amino acids were retranslated into DNA for CGR (Yang *et al.*, 2009). Sun *et al.* (2020) used a three-dimensional CGR representation for protein classification, and Löchel *et al.* (2020) used FCGR for resistance prediction in HIV-1 with CNNs.

Thus, in this study, we analyzed the potential of different statistical and machine learning methods, including LR, SVM, RF and CNN with label encoding, one-hot encoding and FCGR encoding for predicting AMR based on whole-genome sequencing of *Escherichia coli* (*E.coli*).

2 Materials and methods

The workflow of the study is shown in Figure 1.

2.1 Data collection and sample phenotype

Escherichia coli is an important model organism that can cause severe infections in humans and animals, it also represents a significant resistance gene pool that may be responsible for treatment failure in humans and veterinary medicine (Poirel *et al.*, 2018).

In our study, we used two datasets, referred to as the Giessen data and the public data. The first dataset (Giessen) was collected as part of our study and contains whole-genome sequencing data (WGS) and corresponding phenotypic information for several antibiotics for, in total, 987 *E.coli* strains. These isolates were obtained from human and animal clinical samples. Antimicrobial susceptibility testing was performed using the VITEK® 2 system (bioMérieux, Nürtingen, Germany) and interpreted following EUCAST guidelines. DNA isolation and whole-genome sequencing were performed, as described by Falgenhauer *et al.* (2020).

The latter dataset (public) consists of WGS of 1509 *E.coli* strains and corresponding phenotypic information (Moradigaravand *et al.*, 2018). In our study, we focused on the four antibiotics ciprofloxacin (CIP), cefotaxime (CTX), ceftazidime (CTZ) and gentamicin (GEN).

CIP belongs to the class of fluoroquinolones and is widely used to treat various infections, including gastroenteritis, respiratory tract infections or urinary tract infections (Heeb *et al.*, 2011). CIP is particularly effective against Gram-negative bacteria, such as *E.coli*. However, due to overuse, resistances evolve rapidly. CTX and CTZ belong to the class of cephalosporins and are also widely used to treat various infections, such as meningitis, pneumonia, urinary tract infections, sepsis and gonorrhea. They are broad-spectrum antibiotics with activity against numerous Gram-positive and Gram-negative bacteria, including *E.coli*. Nevertheless, resistance is also increasing noticeably (Gums *et al.*, 2008; Sharma, 2013).

GEN belongs to the aminoglycoside class and is widely used to treat various infections, including meningitis, pneumonia, urinary tract infections and sepsis. It is active against a wide range of bacterial infections, mostly Gram-negative bacteria including *E.coli*. It binds to the 30S subunit of the bacterial ribosome and negatively affects protein synthesis (Garneau-Tsodikova and Labby, 2016).

We used data of 900 isolates with resistance information for CIP (418 resistant, 482 susceptible), 930 isolates with resistance information for CTX (455 resistant, 475 susceptible), 841 isolated for CTZ (291 resistant, 550 susceptible) and 926 isolates for GEN (216 resistant, 710 susceptible).

While the CIP and CTX data are balanced, the Giessen datasets are imbalanced on the CTZ and GEN data (34% and 23% resistant isolates, respectively). The public dataset is imbalanced for all antibiotics. For CIP, CTX, CTZ and GEN, there are only 267, 115, 73 and 101 resistant samples, representing 18%, 8%, 5% and 7% of all isolates in the public dataset, respectively.

The summary of the datasets is shown in Table 1.

2.2 Variants calling of whole-genome sequencing data

The raw whole-genome sequencing reads were first quality checked and filtered by fastp (Chen *et al.*, 2018). The filtered reads were then aligned to the *E.coli* reference genome (*E.coli* K-12 strain. MG1655) using BWA-mem (Li *et al.*, 2009). Bcftools (Danecek *et al.*, 2021) was used for calling variants. Samtools (Li and Durbin, 2009) was used to sort the aligned reads, and vcftools (Danecek *et al.*, 2011) was used to filter the raw variants. We used default parameters for all tools.

2.3 SNPs pre-processing and encoding

We first extracted reference alleles, variant alleles and their positions, and merged all isolates based on the position of reference alleles. We filtered out the loci without variation (N replaces a locus without variation), and we built the final SNP matrix, where the rows represent the samples and columns are the variant alleles.

To encode the SNPs for subsequent machine learning, we used label encoding, one-hot encoding and FCGR encoding. For the label encoding, the A, G, C, T and N in the SNP matrix were converted to 1, 2, 3, 4 and 0. In one-hot encoding, the DNA sequence is encoded into a binary matrix, which is subsequently vectorized. For the FCGR encoding, we used the R package kaos to transform the sequences into an image-like matrix with a resolution of 200 (Löchel *et al.*, 2020).

2.4 Machine learning and model evaluation

We used four machine learning methods, including LR, SVM, RF and CNN. For LR, RF and SVM, we used the Scikit-learn python

Table 1. Overview of the datasets

Drug	CIP		CTX		CTZ		GEN	
Source	Giessen	Public	Giessen	Public	Giessen	Public	Giessen	Public
Resistant	418	267	455	115	291	73	216	101
Susceptible	482	1229	475	1313	550	1398	710	1398
Total	900	1496	930	1428	841	1471	926	1489

Downloaded from https://academic.oup.com/bioinformatics/article/38/2/325/6382301 by Universitaetsbibliothek Marburg user on 18 May 2022

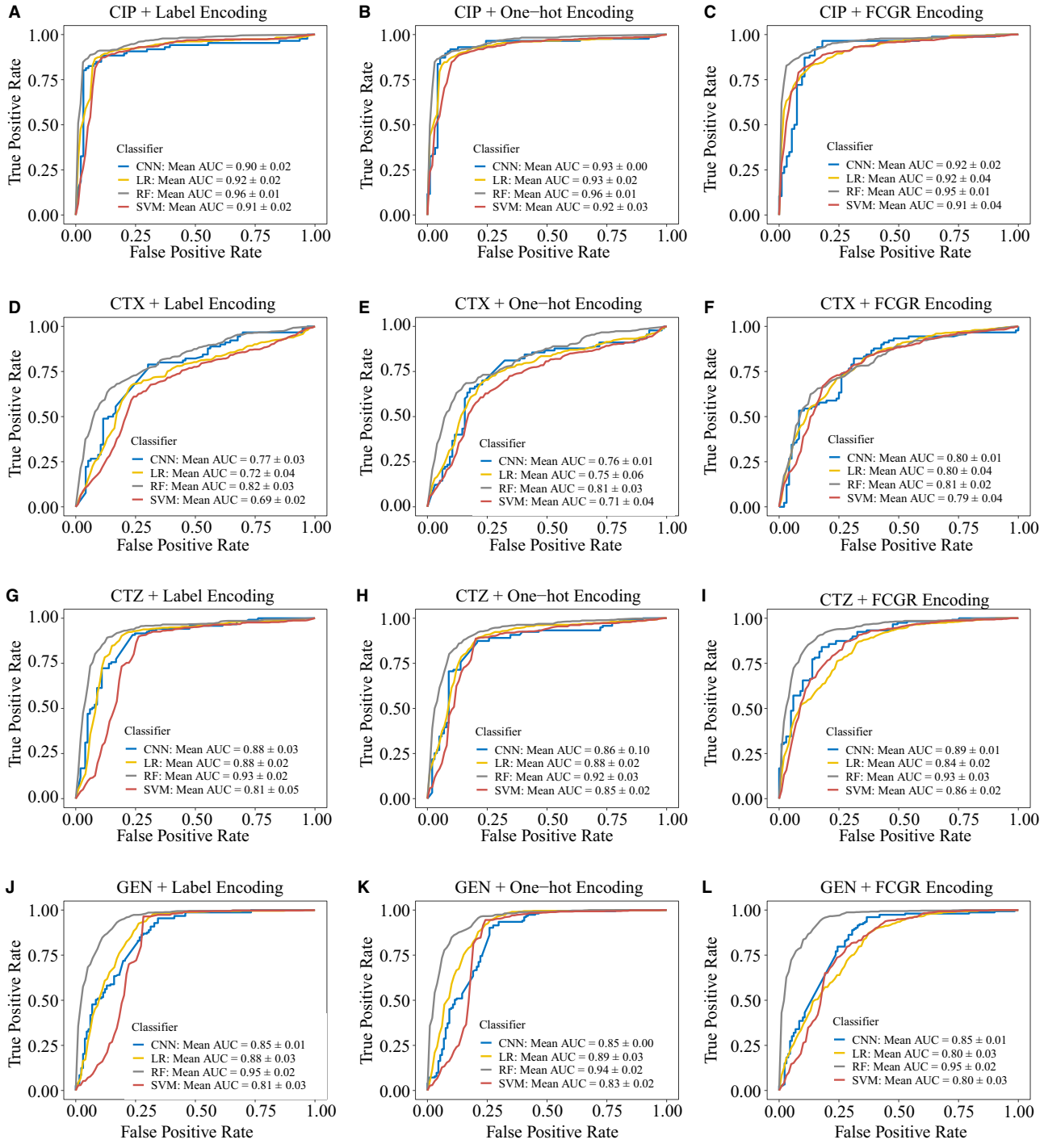

Fig. 2. ROC curves for the models with label encoding, one-hot encoding and FCGR encoding on the Giessen data. First row: ROC curves for CIP with label encoding (**A**), one-hot encoding (**B**) and FCGR encoding (**C**), respectively. Second row: ROC curves for CTX with label encoding (**D**), one-hot encoding (**E**) and FCGR encoding (**F**), respectively. Third row: ROC curves for CTZ with label encoding (**G**), one-hot encoding (**H**) and FCGR encoding (**I**), respectively. Fourth row: ROC curves for GEN with label encoding (**J**), one-hot encoding (**K**) and FCGR encoding (**L**), respectively

Table 2. Results of the four machine learning models with label encoding on the Giessen data

Classifiers/drug	Precision	Precision	Precision	Precision	Recall	Recall	Recall	Recall
	CIP	CTX	CTZ	GEN	CIP	CTX	CTZ	GEN
CNN	0.88 ± 0.04	0.75 ± 0.04	0.81 ± 0.02	0.76 ± 0.03	0.87 ± 0.01	0.65 ± 0.10	0.89 ± 0.03	0.91 ± 0.02
LR	0.88 ± 0.05	0.71 ± 0.04	0.81 ± 0.03	0.77 ± 0.02	0.90 ± 0.03	0.69 ± 0.08	0.92 ± 0.05	0.96 ± 0.03
RF	0.92 ± 0.04	0.75 ± 0.03	0.84 ± 0.03	0.79 ± 0.02	0.89 ± 0.03	0.73 ± 0.07	0.90 ± 0.06	0.97 ± 0.03
SVM	0.85 ± 0.03	0.69 ± 0.02	0.78 ± 0.03	0.75 ± 0.02	0.89 ± 0.04	0.73 ± 0.03	0.89 ± 0.03	0.96 ± 0.03

Downloaded from https://academic.oup.com/bioinformatics/article/38/2/325/6382301 by Universitaetsbibliothek Marburg user on 18 May 2022

Table 3. Results of the four machine learning models with one-hot encoding on the Giessen data

Classifiers/drug	Precision	Precision	Precision	Precision	Recall	Recall	Recall	Recall
	CIP	CTX	CTZ	GEN	CIP	CTX	CTZ	GEN
CNN	0.87 ± 0.05	0.75 ± 0.00	0.84 ± 0.01	0.80 ± 0.00	0.90 ± 0.01	0.71 ± 0.03	0.84 ± 0.03	0.87 ± 0.05
LR	0.89 ± 0.05	0.71 ± 0.04	0.80 ± 0.03	0.78 ± 0.02	0.89 ± 0.03	0.73 ± 0.08	0.89 ± 0.05	0.95 ± 0.02
RF	0.92 ± 0.05	0.75 ± 0.01	0.82 ± 0.02	0.80 ± 0.03	0.90 ± 0.02	0.73 ± 0.07	0.90 ± 0.07	0.97 ± 0.03
SVM	0.86 ± 0.05	0.68 ± 0.03	0.77 ± 0.03	0.76 ± 0.03	0.89 ± 0.03	0.69 ± 0.06	0.89 ± 0.06	0.95 ± 0.04

Table 4. Results of the four machine learning models with FCGR encoding on the Giessen data

Classifiers/drug	Precision	Precision	Precision	Precision	Recall	Recall	Recall	Recall
	CIP	CTX	CTZ	GEN	CIP	CTX	CTZ	GEN
CNN	0.87 ± 0.04	0.74 ± 0.04	0.81 ± 0.03	0.75 ± 0.02	0.91 ± 0.03	0.84 ± 0.04	0.87 ± 0.06	0.96 ± 0.01
LR	0.79 ± 0.08	0.70 ± 0.04	0.73 ± 0.05	0.69 ± 0.04	0.85 ± 0.04	0.79 ± 0.05	0.85 ± 0.04	0.86 ± 0.02
RF	0.91 ± 0.03	0.74 ± 0.01	0.82 ± 0.02	0.80 ± 0.02	0.87 ± 0.03	0.72 ± 0.07	0.90 ± 0.07	0.98 ± 0.01
SVM	0.81 ± 0.03	0.72 ± 0.03	0.73 ± 0.01	0.69 ± 0.02	0.88 ± 0.03	0.81 ± 0.05	0.87 ± 0.03	0.92 ± 0.03

package (Pedregosa *et al.*, 2011). LR was used with default parameters, except that we used 1000 iterations. RF was used with default parameters and 200 trees. For SVM, we used a linear kernel and default parameters.

We implemented CNNs using the Keras (https://keras.io/) package and TensorFlow (https://tensorflow.org). The CNN architecture is based on eleven hidden layers, including four convolutional layers, two batch normalization layers, two pooling layers, one flattening layer, one fully connected layer and one dropout layer. The structure of the networks for label encoding and one-hot encoding are the same, which differ from FCGR encoding-based CNNs only in the convolutional layers and pooling layers (see Supplementary Fig. S1). For FCGR, we used the Conv2D and MaxPooling2D function, while the CNN for the label encoding used the 1D versions instead.

We used eight filters in the first two convolution layers with a kernel size of three, rectified linear unit activation function and same padding. The last two convolution layers used 16 filters instead. The pool size of all pooling layers is two. We used the softmax activation function in the final fully connected layer and compiled the model with Adam optimization and cross-entropy loss.

2.5 Statistical evaluation

We optimized the machine learning models on the Giessen data using five times 5-fold stratified cross-validation. We applied an up-sampling strategy to balance the samples in the training set. For the final evaluation on the public data, we analyzed the performance on the raw public dataset and on a balanced set using a down-sample strategy.

We evaluated the models using the receiver operating characteristics curve (ROC) and the area under the curve (AUC). We also calculated precision and recall for all models. Statistical comparisons were made by the DeLong test (Demler *et al.*, 2012).

2.6 Marker genes identification located around SNPs

To identify the SNPs that are associated with resistance, we performed a marker gene identification using the EFS R package (Neumann *et al.*, 2017). The EFS package aggregates eight feature selection methods for binary classification tasks (Neumann *et al.*, 2016). We used EFS with default parameters. We then annotated the corresponding genes of SNPs using SnpEff software (Cingolani *et al.*, 2012).

3 Results

3.1 Performance of different machine learning methods for predicting AMR on Giessen data

We used the filtered SNPs matrix encoded by label encoding, one-hot encoding and FCGR encoding from the Giessen dataset to train the four machine learning methods LR, RF, SVM and CNN. The performance of the four machine learning models was evaluated using five times 5-fold cross-validation. The ROC curves and AUC values of the different machine learning models range from 0.69 to 0.96, demonstrating that all models can effectively predict AMR compared with random null models (Fig. 2). We observed that the mean AUC of the RFs was higher than for LR, SVM and CNN classifiers for all antibiotics with both encoding methods (Fig. 2). In particular, RFs were significantly better than LR ($P = 0.03$), SVMs ($P = 0.01$) and CNNs ($P = 0.02$) for CIP with label encoding (Supplementary Fig. S2). RFs were also better than the other three classifiers for GEN with label encoding and FCGR encoding ($P < 0.05$). For CTZ, RFs significantly outperformed SVMs with all encoding methods ($P < 0.05$) (Supplementary Fig. S2). For CTX, RFs are significantly better than LR and SVM with label encoding and one-hot encoding ($P < 0.05$), while there are no significant differences if the FCGR encoding is used (Supplementary Fig. S2).

Moreover, all models show high precision and recall using label (Table 2), one-hot (Table 3) and FCGR encoding (Table 4) for CIP. For CTZ and GEN, the models show high recall but lower precision, which may be related to the imbalanced resistant and susceptible isolates. In sum, RF, CNN, LR and SVM can predict AMR for CIP, CTZ, GEN and CTX with three encoding methods in *E.coli*.

3.2 Evaluation of the models on public data

We performed a further evaluation of our models using the public data of *E.coli* of Moradigaravand *et al.* (2018). The public data are highly imbalanced and thus performance metrics are difficult to interpret. Thus, to evaluate the performance of the models, we performed a down-sampling to balance the public data. For completeness, results for the imbalanced set are shown in Supplementary Tables S1–S3.

The resulting ROC curves clearly show that the machine learning models generalize well and can predict AMR (Fig. 3). The AUCs of RFs are higher compared with those from LR, SVM and CNN with three encoding approaches, except for CTZ and GEN with FCGR encoding. Consistent with the results from the Giessen data, all classifiers have high precision and recall for three encoding methods (Tables 5–7).

Downloaded from https://academic.oup.com/bioinformatics/article/38/2/325/6382301 by Universitaetsbibliothek Marburg user on 18 May 2022

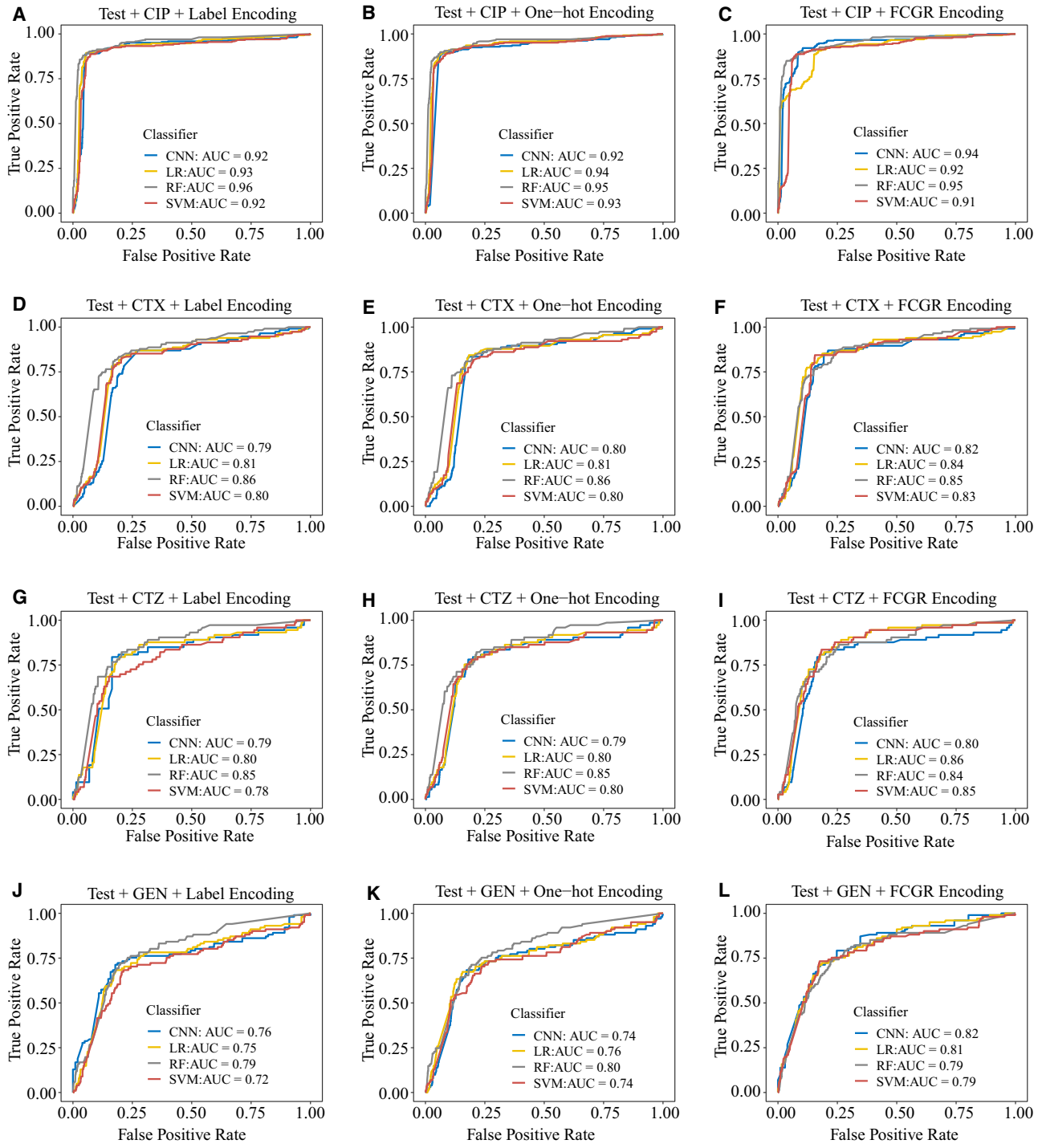

Fig. 3. ROC curves for the models with label, one-hot and FCGR encoding on the public data. First row: ROC curves for CIP with label encoding (**A**), one-hot encoding (**B**) and FCGR encoding (**C**), respectively. Second row: ROC curves for CTX with label encoding (**D**), one-hot encoding (**E**) and FCGR encoding (**F**), respectively. Third row: ROC curves for CTZ with label encoding (**G**), one-hot encoding (**H**) and FCGR encoding (**I**), respectively. Fourth row: ROC curves for GEN with label encoding (**J**), one-hot encoding (**K**) and FCGR encoding (**L**), respectively

3.3 Marker genes associated with antibiotic resistance

We performed an SNP association study on the Giessen and public data using the EFS R package with default parameters. In this analysis, we did not include the known resistance genes. Thus, we aimed at identifying secondary mutations that contribute to the resistance directly or indirectly, e.g. compensatory mutations. This data-driven approach does not need AMR expert knowledge and can also be used and predict resistance even without knowing the resistance genes but by identification of the secondary mutations. EFS provided a ranking of the SNPs for each antibiotic. The ten most important SNPs for each antibiotic are shown in Figure 4. These SNPs are part of 19 different genomic regions. We then annotated and analyzed the corresponding genes of these regions (Table 8).

Downloaded from https://academic.oup.com/bioinformatics/article/38/2/325/6382301 by Universitaetsbibliothek Marburg user on 18 May 2022

Table 5. Evaluation of the machine learning models with label encoding on the public data

Classifiers/drug	Precision	Precision	Precision	Precision	Recall	Recall	Recall	Recall
	CIP	CTX	CTZ	GEN	CIP	CTX	CTZ	GEN
CNN	0.94	0.71	0.79	0.84	0.88	0.88	0.81	0.70
LR	0.93	0.76	0.80	0.82	0.90	0.84	0.75	0.62
RF	0.95	0.75	0.81	0.83	0.90	0.85	0.77	0.61
SVM	0.94	0.71	0.75	0.77	0.87	0.84	0.74	0.60

Note: Precision and recall are calculated based on balanced data using down-sampling.

Table 6. Evaluation of the machine learning models with one-hot encoding on the public data

Classifiers/drug	Precision	Precision	Precision	Precision	Recall	Recall	Recall	Recall
	CIP	CTX	CTZ	GEN	CIP	CTX	CTZ	GEN
CNN	0.95	0.83	0.84	0.80	0.90	0.83	0.78	0.62
LR	0.90	0.80	0.76	0.81	0.90	0.85	0.78	0.63
RF	0.90	0.78	0.73	0.81	0.90	0.86	0.78	0.63
SVM	0.89	0.78	0.75	0.73	0.88	0.83	0.77	0.55

Note: Precision and recall are calculated based on balanced data using down-sampling.

Table 7. Evaluation of the machine learning models with FCGR encoding on the public data

Classifiers/drug	Precision	Precision	Precision	Precision	Recall	Recall	Recall	Recall
	CIP	CTX	CTZ	GEN	CIP	CTX	CTZ	GEN
CNN	0.84	0.71	0.72	0.74	0.93	0.89	0.86	0.71
LR	0.85	0.77	0.79	0.80	0.89	0.87	0.86	0.74
RF	0.92	0.77	0.83	0.83	0.88	0.89	0.78	0.59
SVM	0.88	0.78	0.77	0.75	0.90	0.86	0.86	0.74

Note: Precision and recall are calculated based on balanced data using down-sampling.

Some of these genes are well-known genes conferring antibiotic resistance, such as *marA*. *marA* is a gene related to multiple drug resistance (Abdolmaleki *et al.*, 2019). In comparison, the other genes have not been well studied so far. For instance, the gene *nhaA* (associated with CTX, CTZ and GEN resistance) displays a Na+/H+ antiport activity in *E.coli* that can regulate the permeability, which may further affect drug resistance (Padan *et al.*, 2004). The gene *rlmC* encodes a 23S RNA methyltransferase that methylates the 23S rRNA, of antibiotic binding sites and is related to antibiotic resistance (Pletnev *et al.*, 2020; Stojković *et al.*, 2016). It has been reported that the gene *fliI* encodes a virulence factor, and some studies focused on the correlation between antimicrobial resistance and bacterial virulence (Beceiro *et al.*, 2013; Deng *et al.*, 2019). The gene *pepB* encodes the peptidase B, which is related to the production of bacteriocins, narrow-spectrum antimicrobial peptides produced by bacteria (Suzuki *et al.*, 2001; Telhig *et al.*, 2020). *MurB* is the key biosynthetic enzyme involved in the synthesis of peptidoglycan, the key component of the cell wall (Nasiri *et al.*, 2017; Walsh and Wencewicz, 2014). In sum, the marker genes and SNPs identified by EFS can be used as a reference for further AMR studies.

4 Discussion

This study analyzed four different machine learning methods (RFs, LR, SVMs and CNNs) for predicting four antibiotic resistances in *E.coli* based on whole-genome sequence data with three different encoding schemes, namely, label encoding, one-hot encoding and FCGR encoding. Moreover, our goal was to identify mutations (secondary mutations) contributing to resistance beyond known resistance genes. Thus, we used a reference genome for *E.coli* without known resistance genes. Our study confirmed that label encoding, one-hot encoding and FCGR encoding could encode genomic data for preparing the input data for subsequent machine learning and deep learning methods. Our results show that the four machine learning methods can effectively predict AMR without the need for a database of known resistance genes or SNPs, which is an essential prerequisite for AMR prediction in less well-studied pathogens and drugs. Furthermore, we provide potential genes and SNPs associated with AMR based that can be used as a reference for the subsequent experiments.

Previous studies reported different SNPs in the bacterial genome associated with multiple drug resistance (Brimacombe *et al.*, 2007; Figueroa *et al.*, 2019; Shi *et al.*, 2019; Su *et al.*, 2019; Yang *et al.*, 2018). However, these studies mainly focused on partial SNPs based on available AMR databases (Yang *et al.*, 2018). Machine learning based on the complete set of SNPs from whole-genome sequencing gives further insights and can be used to identify novel biological mechanisms of resistance.

Encoding the genomic features into a readable format for machine learning and deep learning is an essential step. Label encoding, one-hot encoding and CGR encoding can convert SNPs into machine-recognizable formats very efficiently. Our study used the three approaches to encode SNPs and yield excellent predictions for both encoding methods. Many studies indicated that CNNs outperform other machine learning algorithms in image classification, which was the rationale for incorporating FCGR as an encoding scheme.

We compared four machine learning methods, including RFs, LR, SVMs and CNNs. Overall, the four machine learning methods showed good performance in predicting the four antibiotic resistances of *E.coli*. We also demonstrated that our models generalize well on unseen data, as proven by validating the results based on an independent public dataset. We were also able to identify SNPs

Downloaded from https://academic.oup.com/bioinformatics/article/38/2/325/6382301 by Universitaetsbibliothek Marburg user on 18 May 2022

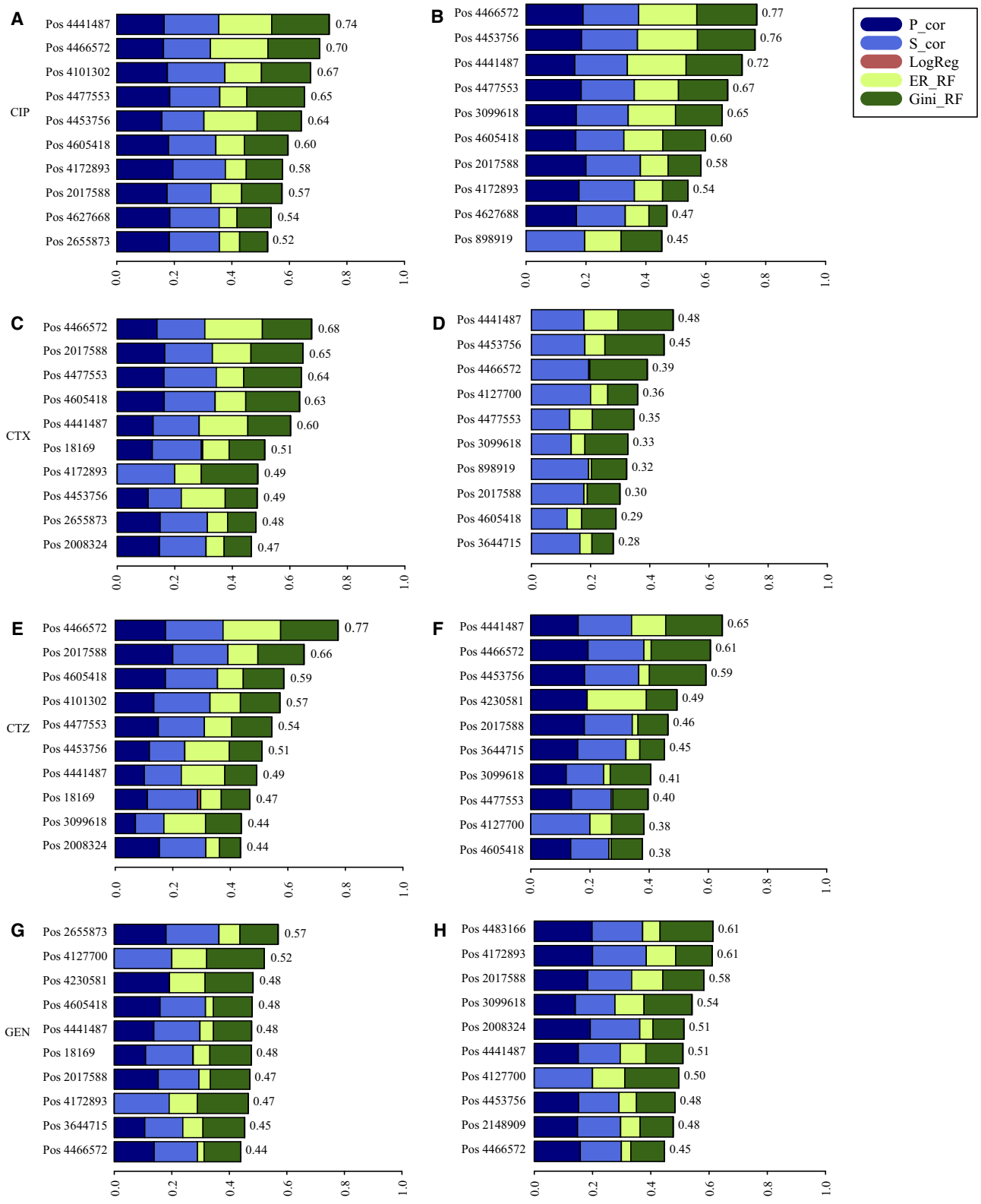

Fig. 4. EFS analysis for each antibiotic for both datasets. The left four figures are the identified ten most important SNPs for CIP (**A**), CTX (**C**), CTZ (**E**) and GEN (**G**) from the Giessen dataset. The right figures are the corresponding SNPs from the public dataset

associated with resistance. However, the marker genes located around the SNPs associated with AMR need experimental validation.

Although we only focused on four antibiotics in this study, our method can easily be applied to other antibiotics and can also be extended to other resistance-related SNPs of other pathogens, also

Downloaded from https://academic.oup.com/bioinformatics/article/38/2/325/6382301 by Universitaetsbibliothek Marburg user on 18 May 2022

Table 8. SNPs and corresponding genes associated with AMR

SNP Position	Gene location	SNP annotation	Gene	Gene biotype	Drug
18169	17489 → 18655	Synonymous	*nhaA*	CDS	CTX, CTZ, GEN
898919	898518 → 899645	Synonymous	*rlmC*	CDS	CIP, CTX
2008324	2008277 → 2009482	Synonymous	*yedE*	CDS	CTX, CTZ, GEN
2017588	2016554 → 2017927	synonymous	*fliI*	CDS	CIP, CTX, CTZ, GEN
2148909	2147674 → 2149026	Synonymous	*yegD*	CDS	GEN
2655873	2655075 → 2656358	Synonymous	*pepB*	CDS	CIP, CTX, GEN
3099618	3098558 → 3099565	Upstream gene	*yggM*	CDS	CIP, CTX, CTZ, GEN
3644715	3643140 → 3645182	Synonymous	*prlC*	CDS	CTX, CTZ, GEN
4101302	4100810 → 4101430	Missense	*sodA*	CDS	CIP, CTZ
4127700	4127286 → 4127894	Synonymous	*yiiX*	CDS	CTX, CTZ, GEN
4172893	4172057 → 4173085	Missense	*murB*	CDS	CIP, CTX, GEN
4230581	4230354 → 4231226	Synonymous	*rluF*	CDS	CTZ, GEN
4441487	4439872 → 4441215	Upstream gene	*ytfL*	CDS	CIP, CTX, CTZ, GEN
4453756	4453583 → 4454578	Synonymous	*yjfF*	CDS	CIP, CTX, CTZ, GEN
4466572	4466299 → 4467246	Synonymous	*treR*	CDS	CIP, CTX, CTZ, GEN
4477553	4477307 → 4478311	Missense	*argI*	CDS	CIP, CTX, CTZ
4483166	4480982 → 4483837	Synonymous	*valS*	CDS	GEN
4605418	4604875 → 4605663	Synonymous	*fhuF*	CDS	CIP, CTX, CTZ, GEN
4627668	4627315 → 4628547	Synonymous	*nadR*	CDS	CIP

Note: The first column shows the positions of the identified SNPs for the four antibiotics. The second column and third column show the gene location and SNP annotation. The fourth column and fifth column show the genes annotated from SNPs and gene biotype. The final column is the antibiotics that are associated with the SNPs.

from species other than bacteria. Furthermore, our approach can also be applied to other biomedical areas, e.g. for cancer resistance prediction. More importantly, our method may have huge potential in systems medicine, to improve the diagnosis, targeted therapy and disease prevention.

There are also some limitations in our study. For example, we only used SNP data in our models that have been called based on a single reference genome. This, however, spares many genomic regions that might be important resistance factors. This is especially true for diverse species like *E.coli*. One approach to mitigate this issue would be the selection of more suitable or multiple reference genomes. Another option potentially leading to a more holistic set of potential SNPs would be to use an artificial pseudo-pan-genome incorporating many genomes of a particular species as a reference within the SNP detection workflow. However, other features, e.g. transcriptomics or proteomics data, might be important for AMR as well (Moradigaravand *et al.*, 2018). Moreover, several other important drugs have not been taken into account yet. However, they may be analyzed with the same methodology when enough data are available.

5 Conclusion

We investigated four machine learning methods for predicting AMR to four different drugs in *E.coli* from whole-genome sequence data with label encoding, one-hot encoding and FCGR encoding. Our results demonstrated that all methods perform very well also for unseen data. Overall, our study provides a new machine learning-driven approach for resistance prediction and thus, may improve treatment of patients in the future.

We evaluated the performance based on cross-validation on our own data and tested the model performance on public data. Moreover, we identified potential SNPs and corresponding genes that are associated with AMR.

We could demonstrate that label encoding, one-hot encoding and FCGR encoding can be used for whole-genome sequence analyses. Moreover, we provide a comprehensive evaluation of different machine learning algorithms for AMR prediction in *E.coli*. The results of the study give a rich reference resource for further research on both experimental and computational aspects of AMR.

Acknowledgements

The authors thank Moradigaravand *et al.* for making their data publicly available.

Funding

This work was financially supported by the German Federal Ministry of Education and Research (BMBF) [031L0209B] (Deep-iAMR).

Data availability

The public data is publicly available (see material and methods). The Giessen data is available upon request.

Conflict of Interest: none declared.

References

Abdolmaleki,Z. *et al.* (2019) Phenotypic and genotypic characterization of antibiotic resistance in the methicillin-resistant *Staphylococcus aureus* strains isolated from hospital cockroaches. *Antimicrob. Resist. Infect. Control*, **8**, 54.

Almeida,J.S. *et al.* (2001) Analysis of genomic sequences by chaos game representation. *Bioinformatics*, **17**, 429–437.

Arango-Argoty,G. *et al.* (2018) DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, **6**, 1–15.

Beceiro,A. *et al.* (2013) Antimicrobial resistance and virulence: a successful or deleterious association in the bacterial world? *Clin. Microbiol. Rev.*, **26**, 185–230.

Boolchandani,M. *et al.* (2019) Sequencing-based methods and resources to study antimicrobial resistance. *Nat. Rev. Genet.*, **20**, 356–370.

Brimacombe,M. *et al.* (2007) Antibiotic resistance and single-nucleotide polymorphism cluster grouping type in a multinational sample of resistant mycobacterium tuberculosis isolates. *Antimicrob. Agents Chemother.*, **51**, 4157–4159.

Chen,S. *et al.* (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.

Cingolani,P. *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms. *SnpEff. Fly*, **6**, 80–92.

Danecek,P. *et al.*; 1000 Genomes Project Analysis Group. (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Downloaded from https://academic.oup.com/bioinformatics/article/38/2/325/6382301 by Universitaetsbibliothek Marburg user on 18 May 2022

Danecek,P. *et al.* (2021) Twelve years of SAMtools and BCFtools. *GigaScience*, **10**, 1–4.
Demler,O.V. *et al.* (2012) Misuse of DeLong test to compare AUCs for nested models. *Stat. Med.*, **31**, 2577–2587.
Deng,Y. *et al.* (2019) Horizontal gene transfer contributes to virulence and antibiotic resistance of vibrio harveyi 345 based on complete genome sequence analysis. *BMC Genomics*, **20**, 761.
Deschavanne,P.J. *et al.* (1999) Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol.*, **16**, 1391–1399.
Falgenhauer,L. *et al.* (2020) Cross-border emergence of clonal lineages of ST38 *Escherichia coli* producing the OXA-48-like carbapenemase OXA-244 in Germany and Switzerland. *Int. J. Antimicrob. Agents*, **56**, 106157.
Figueroa,J. *et al.* (2019) Analysis of single nucleotide polymorphisms (SNPs) associated with antibiotic resistance genes in Chilean *Piscirickettsia salmonis* strains. *J. Fish Dis.*, **42**, 1645–1655.
Garneau-Tsodikova,S. and Labby,K.J. (2016) Mechanisms of resistance to aminoglycoside antibiotics: overview and perspectives. *MedChemComm*, **7**, 11–27.
Gums,J.G. *et al.* (2008) Differences between ceftriaxone and cefotaxime: microbiological inconsistencies. *Ann. Pharmacother.*, **42**, 71–79.
Heeb,S. *et al.* (2011) Quinolones: from antibiotics to autoinducers. *FEMS Microbiol. Rev.*, **35**, 247–274.
Hoang,T. *et al.* (2016) Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison. *Genomics*, **108**, 134–142.
Jeffrey,H. (1990) Chaos game representation of gene structure. *Nucleic Acids Res.*, **18**, 2163–2170.
Joseph,J. and Sasikumar,R. (2006) Chaos game representation for comparison of whole genomes. *BMC Bioinformatics*, **7**, 243.
Kania,A. and Sarapata,K. (2021) The robustness of the chaos game representation to mutations and its application in free-alignment methods. *Genomics*, **113**, 1428–1437.
Kouchaki,S. *et al.*; CRyPTIC Consortium. (2019) Application of machine learning techniques to tuberculosis drug resistance analysis. *Bioinformatics*, **35**, 2276–2282.
Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, **25**, 1754–1760.
Li,H. *et al.*; 1000 Genome Project Data Processing Subgroup. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
Lichtblau,D. (2019) Alignment-free genomic sequence comparison using FCGR and signal processing. *BMC Bioinformatics*, **20**, 742.
Liu,Z. *et al.* (2020) Evaluation of machine learning models for predicting antimicrobial resistance of *Actinobacillus pleuropneumoniae* from whole genome sequences. *Front. Microbiol.*, **11, doi: 10.3389/fmicb.2020.00048**.
Löchel,H.F. *et al.* (2020) Deep learning on chaos game representation for proteins. *Bioinformatics*, **36**, 272–279.
Lv,J. *et al.* (2021) A review of artificial intelligence applications for antimicrobial resistance. *Biosafety Health*, **3**, 22–31.
Moradigaravand,D. *et al.* (2018) Prediction of antibiotic resistance in *Escherichia coli* from large-scale pan-genome data. *PLoS Comput. Biol.*, **14**, e1006258.
Nasiri,M.J. *et al.* (2017) New insights in to the intrinsic and acquired drug resistance mechanisms in mycobacteria. *Front. Microbiol.*, **8**, 681.
Naylor,N.R. *et al.* (2018) Estimating the burden of antimicrobial resistance: a systematic literature review. *Antimicrob. Resist. Infect. Control*, **7**, 58.
Neumann,U. *et al.* (2016) Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection approach. *BioData Mining*, **9**, 36.
Neumann,U. *et al.* (2017) EFS: an ensemble feature selection tool implemented as r-package and web-application. *BioData Min.*, **10**, 21.
Padan,E. *et al.* (2004) NhaA of escherichia coli, as a model of a pH-regulated na+h+antiporter. *Biochim. Biophys. Acta (BBA) Bioenerget.*, **1658**, 2–13.
Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
Pletnev,P. *et al.* (2020) Comprehensive functional analysis of *Escherichia coli* ribosomal RNA methyltransferases. *Front. Genet.*, **11**, 97.
Poirel,L. *et al.* (2018) Antimicrobial resistance in escherichia coli. *Microbiol. Spectrum*, **6, doi: 10.1128/microbiolspec.ARBA-0026-2017**.
Rizzo,R. *et al.* (2016) Classification experiments of DNA sequences by using a deep neural network and chaos game representation. In: *Proceedings of the 17th International Conference on Computer Systems and Technologies*. ACM, Palermo, Italy, pp. 222–228.
Sengupta,D.C. *et al.* (2020) Similarity studies of corona viruses through chaos game representation. *Comput. Mol. Biosci.*, **10**, 61–72.
Sharma,M. (2013) Prevalence and antibiogram of extended spectrum beta-lactamase (ESBL) producing gram negative bacilli and further molecular characterization of ESBL producing *Escherichia coli* and *Klebsiella* spp. *J. Clin. Diagn. Res.*, **7**, 2173–7.
Shi,J. *et al.* (2019) Antimicrobial resistance genetic factor identification from whole-genome sequence data using deep feature selection. *BMC Bioinformatics*, **20**, 535.
Spänig,S. and Heider,D. (2019) Encodings and models for antimicrobial peptide classification for multi-resistant pathogens. *BioData Mining*, **12**, 7.
Stojković,V. *et al.* (2016) Antibiotic resistance evolved via inactivation of a ribosomal RNA methylating enzyme. *Nucleic Acids Res.*, **44**, 8897–8907.
Stokes,J.M. *et al.* (2020) A deep learning approach to antibiotic discovery. *Cell*, **180**, 688–702.e13.
Su,M. *et al.* (2019) Genome-based prediction of bacterial antibiotic resistance. *J. Clin. Microbiol.*, **57**, e01405-18.
Sun,Z. *et al.* (2020) A novel numerical representation for proteins: three-dimensional chaos game representation and its extended natural vector. *Comput. Struct. Biotechnol. J.*, **18**, 1904–1913.
Suzuki,H. *et al.* (2001) Purification and characterization of aminopeptidase b from *Escherichia coli* k-12. *Biosci. Biotechnol. Biochem.*, **65**, 1549–1558.
Telhig,S. *et al.* (2020) Bacteriocins to thwart bacterial resistance in gram negative bacteria. *Front. Microbiol.*, **11**, 586433.
Veltri,D. *et al.* (2018) Deep learning improves antimicrobial peptide recognition. *Bioinformatics*, **34**, 2740–2747.
Walsh,C.T. and Wencewicz,T.A. (2014) Prospects for new antibiotics: a molecule-centered perspective. *J. Antibiot.*, **67**, 7–22.
Wang,Y. *et al.* (2005) The spectrum of genomic signatures: from dinucleotides to chaos game representation. *Gene*, **346**, 173–185.
Yang,J.-Y. *et al.* (2009) Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation. *J. Theor. Biol.*, **257**, 618–626.
Yang,Y. *et al.* (2018) Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics*, **34**, 1666–1671.
Yu,Z.-G. *et al.* (2004) Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses. *J. Theor. Biol.*, **226**, 341–348.

Downloaded from https://academic.oup.com/bioinformatics/article/38/2/325/6382301 by Universitaetsbibliothek Marburg user on 18 May 2022

3.2 Publication 2: Multi-label Classification for Multi-drug Resistance Prediction of *Escherichia Coli*

3.2.1 Summary

Aim and Motivation

The aim of this work (Ren et al., 2022a) focuses on addressing the multi-drug resistance (MDR) problem. Its objective is to explore and evaluate the effectiveness of various multi-label classification (MLC) methods for predicting MDR. MDR within pathogenic bacteria poses a significant threat to global health. MDR is typically the consequence of genetic mutations and the aggregation of resistance genes, often leading to treatment failure and increasing public health risks. Although machine learning methods offer a broad spectrum of applications for AMR prediction, they predominantly focus on predicting single drug resistance and overlook the temporal accumulation of AMR traits. This leaves the simultaneous and rapid identification of multi-drug resistance as an unaddressed challenge.

Methods and Results

In this study, we used 809 whole-genome sequencing (WGS) data of *E. coli* strains with resistance information for four antibiotics, namely ciprofloxacin (CIP), cefotaxime (CTX), ceftazidime (CTZ), and gentamicin (GEN). We called for SNP variants and performed preprocessing analysis, following the same procedure as our previous study (Ren et al., 2021). To achieve the multi-label classification of MDR in bacteria, we deployed five different methodologies: Binary Relevance (BR), Classifier Chain (CC), Ensemble Classifier Chains (ECC), Label Powerset (LP), and Random Label Space Partitioning with Label Powerset (RD). Our results demonstrated the potential of MLC methods in accurately modeling multi-drug resistance in pathogens. Importantly, we found the ECC model achieves accurate MDR prediction and outperforms other MLC methods.

Conclusion

Our study broadens the array of tools available for predicting MDR, thus catalyzing advancements in diagnosing patient infections. The multi-label classification methods that we have introduced not only expedite the identification of pathogens and resistance but also enhance its accuracy. Consequently, these methodologies hold the potential to mitigate the public health threats posed by antimicrobial resistance, and in the long term, reduce the number of fatalities associated with such resistance.

Graphical abstract

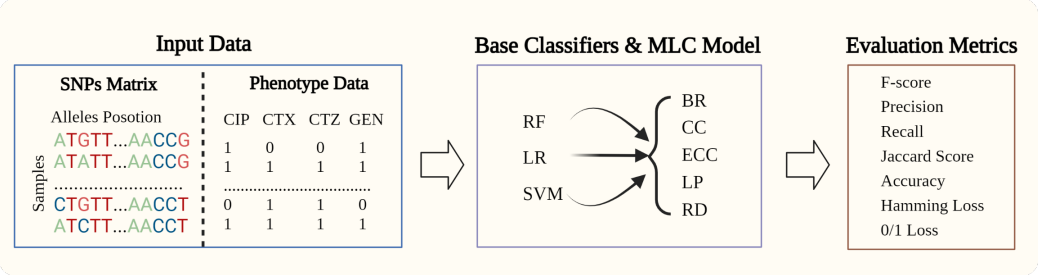

Figure 3.1: Workflow of this study. This figure was created by BioRender.com

Computational and Structural Biotechnology Journal 20 (2022) 1264–1270

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Multi-label classification for multi-drug resistance prediction of *Escherichia coli*

Yunxiao Ren [a], Trinad Chakraborty [b,c], Swapnil Doijad [b,c], Linda Falgenhauer [c,d,e], Jane Falgenhauer [b,c], Alexander Goesmann [c,f], Oliver Schwengers [c,f], Dominik Heider [a,*]

[a] Department of Data Science in Biomedicine, Faculty of Mathematics and Computer Science, Philipps-University of Marburg, Germany
[b] Institute of Medical Microbiology, Justus Liebig University Giessen, Germany
[c] German Center for Infection Research, Partner site Giessen-Marburg-Langen, Germany
[d] Institute of Hygiene and Environmental Medicine, Justus Liebig University Giessen, Germany
[e] Hessisches universitäres Kompetenzzentrum Krankenhaushygiene, Germany
[f] Department of Bioinformatics and Systems Biology, Justus Liebig University Giessen, Germany

ARTICLE INFO

Article history:
Received 31 January 2022
Received in revised form 8 March 2022
Accepted 8 March 2022
Available online 10 March 2022
Dataset link: https://github.com/YunxiaoRen/Multi_Label-Classification

Keywords:
Multi-drug resistance
Machine learning
Multi-label classification

ABSTRACT

Antimicrobial resistance (AMR) is a global health and development threat. In particular, multi-drug resistance (MDR) is increasingly common in pathogenic bacteria. It has become a serious problem to public health, as MDR can lead to the failure of treatment of patients. MDR is typically the result of mutations and the accumulation of multiple resistance genes within a single cell. Machine learning methods have a wide range of applications for AMR prediction. However, these approaches typically focus on single drug resistance prediction and do not incorporate information on accumulating antimicrobial resistance traits over time. Thus, identifying multi-drug resistance simultaneously and rapidly remains an open challenge. In our study, we could demonstrate that multi-label classification (MLC) methods can be used to model multi-drug resistance in pathogens. Importantly, we found the ensemble of classifier chains (ECC) model achieves accurate MDR prediction and outperforms other MLC methods. Thus, our study extends the available tools for MDR prediction and paves the way for improving diagnostics of infections in patients. Furthermore, the MLC methods we introduced here would contribute to reducing the threat of antimicrobial resistance and related deaths in the future by improving the speed and accuracy of the identification of pathogens and resistance.

© 2022 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Antimicrobial resistance (AMR) is rapidly increasing and is, therefore, one of the greatest threats to global health and also causes significant economic problems. According to WHO estimates, without countermeasures, up to 10 million deaths will be caused by AMR in the future, with immense costs to the healthcare system of approximately $100 trillion by 2050 [1]. In particular, infection due to multi-drug resistance (MDR) pathogens has become most threatening to public health, as MDR can lead to failure of treatment of patients [2,3]. For instance, the emergence of MDR in *Escherichia coli (E. coli)* has become one of the global health concerns [4–6]. In general, bacteria are resistant to antibiotics by spontaneous mutations in existing genes or by the acquisition of extraneous genes [6,7]. Many previous studies investigating AMR have focused on well-known resistance genes or mutations in well-known genes, such as mutations in the *gyrA* gene and *parC* gene in *E. coli* [8,9]. However, there is a lack of AMR studies based on overall mutations without previous knowledge.

While antimicrobial susceptibility testing (AST) is widely used for AMR profiles in clinical practice, machine learning models have been shown to produce highly reliable predictions in a shorter turnaround time. Typically, these machine learning models combine sequencing data with antibiotic resistance databases with phenotypic information [10,11]. For instance, Yang *et al.*, [12] and Kouchaki *et al.*, [13] used different machine learning algorithms, namely support vector machine (SVM), logistic regression (LR), and random forest (RF) to predict AMR from whole-genome

Abbreviations: AMR, Antimicrobial Resistance; MDR, Multi-Drug Resistance; MLC, Multi-Label Classification.

* Corresponding author.
E-mail address: dominik.heider@uni-marburg.de (D. Heider).

https://doi.org/10.1016/j.csbj.2022.03.007
2001-0370/© 2022 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

sequencing data and achieved high accuracy prediction. Other approaches also included deep learning to predict new antibiotic drugs, AMR genes, and AMR peptides [14–20]. However, all of these studies are based on single drug resistance information and do not take into account the MDR information of the bacteria.

Multi-label classification (MLC) offers a potential solution for AMR prediction based on MDR information. Traditionally, multi-label problems are transformed into single-label problems [21]. For instance, the widely known binary relevance (BR) approach, is a simple and straightforward method that treats each label as an independent binary problem [22]. One of the limitations of the BR approach is that it does not take into account the dependencies between the labels [23]. Unlike BR, the classifier chain (CC) takes into account the correlation among labels and uses the predicted results from the previous classifiers as an additional input for the following classifier [24]. Obviously, the order of the CC affects the prediction accuracy. Thus, the ensemble of classifier chains (ECC) was proposed, which contains several CCs with different orders and can be applied to study the dependencies between labels [23,24]. CCs and ECCs have been used for cross-resistance prediction in HIV based on protein sequences of the HIV-1 reverse transcriptase [25] and protease [26], however, it has never been used with genomic data and MDR of bacteria. Other multi-label approaches include the label powerset (LP) method, which considers the dependency among labels, and each label combination is considered as a class [21]. Random label space partitioning with label powerset (RD) method is another effective ensemble method, which is based on label powerset with a random subset of *k* labels [23,24].

In our study, we gave the applications of MLC methods on multi-drug resistance prediction. We aimed at identifying secondary mutations that contribute to the resistance directly or indirectly, e.g., compensatory mutations. We did not include the known resistance genes. Our approach does not need any AMR expert knowledge and can also predict resistance even without knowing the resistance genes by identifying secondary mutations. The results demonstrated that the ECC model can significantly improve overall resistance prediction in bacteria compared to the other four MLC methods. MLC models will improve patient care, in particular the treatment of patients, reduce the threat of antimicrobial resistance and related deaths in the future, and improve the speed and accuracy of the identification of pathogens and resistance.

2. Materials and methods

2.1. Dataset

In our analysis, we used 987 whole-genome sequencing (WGS) data of *E. coli* strains with resistance information for four antibiotics, namely ciprofloxacin (CIP), cefotaxime (CTX), ceftazidime (CTZ), and gentamicin (GEN). These data were collected by our partner institution, the University of Giessen. The isolates were obtained from human and animal clinical samples. Antimicrobial susceptibility testing was performed using the VITEK® 2 system (bioMérieux, Nürtingen, Germany) and interpreted following EUCAST guidelines. DNA isolation and whole-genome sequencing was performed as described in Falgenhauer et al. [27].

In order to use MLC, the isolates need to be filtered for missing antibiotic resistance information. The final dataset with complete MDR information contains 809 *E. coli* strains (see Table 1). CIP is a fluoroquinolone and is widely used to treat infections with Gram-negative bacteria, e.g., gastroenteritis, respiratory tract infections, or urinary tract infections [28]. CTX and CTZ are broad-spectrum antibiotics from the class of cephalosporins and are widely used to treat infections of Gram-positive and Gram-negative bacteria, such as meningitis, pneumonia, urinary tract infections, sepsis, and gonorrhea [29,30]. GEN is an aminoglycoside and is widely used to treat various infections of Gram-negative bacteria, including meningitis, pneumonia, urinary tract infections, and sepsis [31].

2.2. Dataset pre-processing and encoding

The pre-processing step of raw WGS data refer to our previous study [20]. Briefly, we filtered bad quality reads by fastp (v0.23.2) software [32] and then mapped the clean reads to *E. coli* reference genome (*E. coli* K-12 strain. MG1655) through BWA-MEM with default parameters [33]. We called single nucleotide polymorphisms (SNPs) variants using bcftools (v1.14) via 'call' function with default parameters [34,35]. We extracted reference alleles, variant alleles and their positions, and merged all isolates based on the position of reference alleles. We retained the alleles existing variant more than half in samples. Finally, we got an SNP matrix, where the rows represent the samples and columns are the variant alleles. We utilized one-hot encoding to transform the SNP matrix into a binary matrix for subsequent machine learning.

2.3. Multi-label classification

In the current study, we used BR, CC, ECC, LP, and RD for the multi-label classification of MDR in bacteria. BR is typically used as a baseline model to compare multi-label classification models. Let $L := \{\lambda_1, ..., \lambda_m\}$ with $m > 1$ be a finite set of class labels (here: resistance for the four antibiotics), and let X be the instance space, i.e., the SNPs. The training set S in MLC is then defined as $S := \{(x_1, y_1), ..., (x_n, y_n)\}$, generated independently and identically according to a probability distribution $P(X,)$ on $X \times Y$. Y is the set of possible label combinations, i.e., the powerset of L (Fig. 1A).

BR divides the dataset with L labels into L binary classification problems (Fig. 1B). Accordingly, we split the data into four binary classification problems, one for each antibiotic (CIP, CTX, CTZ, and GEN). In contrast, the CC approach links the L binary classifiers into a "chain" such that the output prediction of one classifier is used as an additional input for all subsequent classifiers, which overcomes the disadvantage of not considering dependencies between labels and captures possible dependencies between the labels (Fig. 1C). The performance of CC depends heavily on the order of the chain, thus, Read *et al.*, [23] proposed the use of ECC, which aggregates several chains with different orders by majority vote (Fig. 1D). The LP approach transforms a multi-label problem into a single-label multi-class problem, which is trained on all unique label combinations found in the training data [36] (Fig. 1E). The RD method divides the label space into partitions of size k, trains an LP classifier per partition, and predicts the testing data by aggregating the result of all LP classifiers (Fig. 1F). It is important to note that any standard method for binary classification can be used in these multi-label approaches. In the current study, we evaluated RFs, LR, and SVMs for multi-label classification of MDR in bacteria.

2.4. Evaluation metrics

In MLC, the predictions for each instance are a collection of labels, and the performance of classifiers can be calculated through the average score of an evaluation metric or directly by comparing the scores for each class. In this study, we employed seven different metrics that are widely used to evaluate the performance of the classifiers including hamming loss, 0/1 loss, F-score, accuracy, precision, recall, and Jaccard similarity.

Table 1
Overview of the dataset.

Antibiotics	CIP	CTX	CTZ	GEN
Resistant	366	358	276	188
Susceptible	443	451	533	621

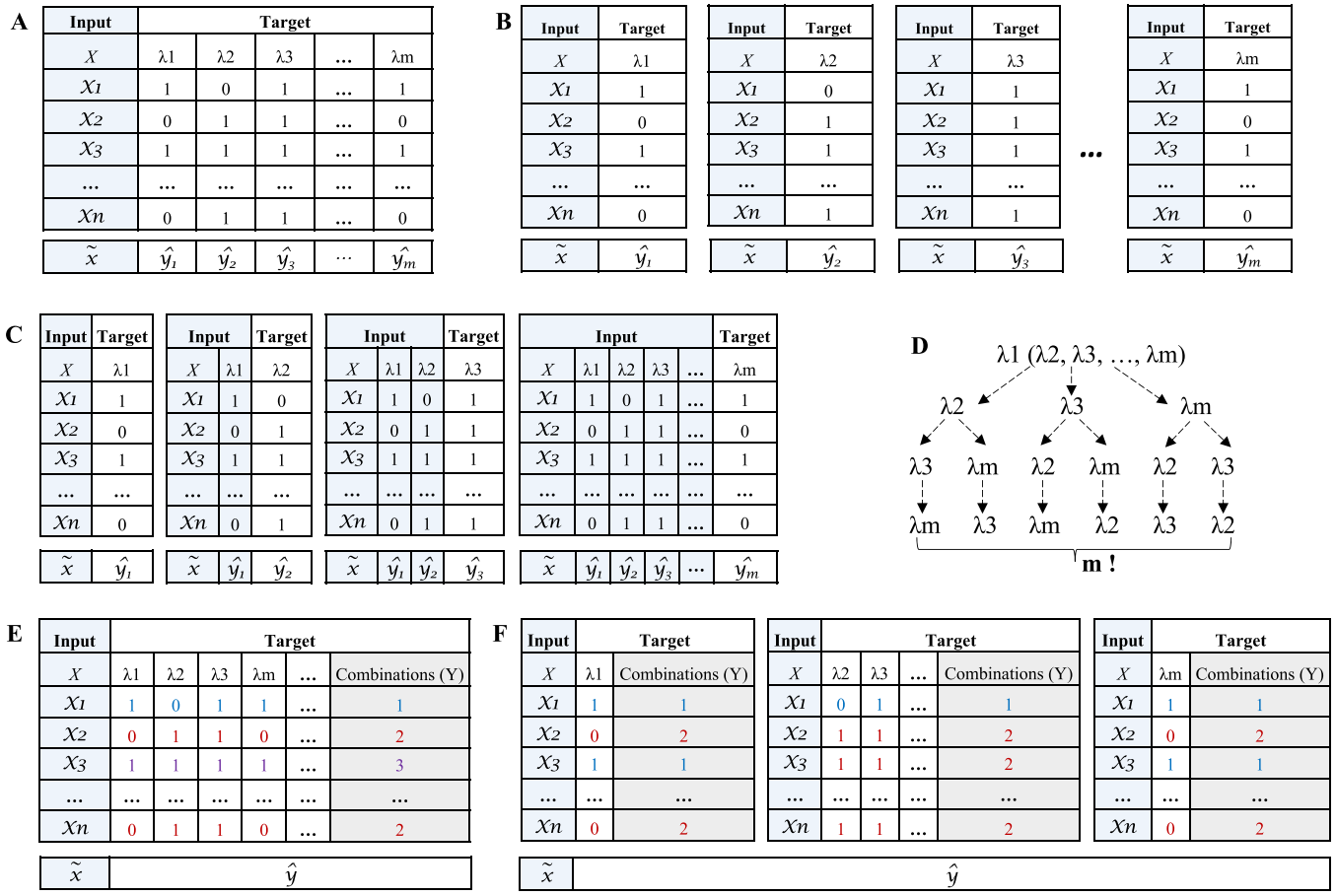

Fig. 1. Transformation methods of multi-label classification problems. (A) One multi-label dataset. $\chi_i \in$ xis a training instance. (B) Binary relevance (BR) transforms the multi-label dataset with m labels into m independent binary datasets. (C) The process of classifier chain (CC) for multi-label data. (D) The possible number of label orders for ensemble classifier chains (ECC). (E) The transformation of the multi-label dataset by label powerset (LP). Labels with different colors represent the different combinations of labels. (F) The transformation of a multi-label dataset by random label space partitioning with label powerset (RD). Labels with different colors represent the different combinations of labels.

The Hamming loss and 0/1 loss are commonly used for the evaluation of MLC models [37]. For Hamming loss, it is defined as the fraction of labels that are incorrectly predicted. The 0/1 loss simply checks whether the complete label subset is predicted correctly or not, represented as the percentage of incorrectly predicted labels.

Accuracy is defined as the proportion of correct predictions, while precision is defined as the number of resistant samples divided by the overall number of samples that are predicted to be resistant. Recall (also called sensitivity) is defined as the number of correctly predicted resistant samples divided by the total number of resistant samples. The F-score can be calculated as the weighted average of precision and recall. Jaccard similarity indicates the overlap between the ground truth and the predictions, focusing on true positives and ignoring true negatives [38]. The classifiers were trained and evaluated based on five-times 5-fold cross-validation, which means the dataset is randomly divided into 5 equal sub-groups, and one of the groups is used as the test set and the rest are used as the training set. The model is trained on the training set and scored on the test set. Then the process is repeated until each unique group has been used as the test set. Statistical significance has been calculated based on the Wilcoxon signed-rank test and T-test.

3. Results

3.1. Performance of different MLC methods on RF base classifier

We firstly constructed five MLC models (BR, CC, ECC, LP, and RD) based on RF base classifier for MDR prediction of four antibiotics (CIP, CTX, CTZ, and GEN). We compared the performance by F-score, Precision and Recall, and Jaccard score. As shown in Fig. 2, the ECC model has the highest F-score, Precision and Recall, and Jaccard score for resistance prediction against four antibiotics. For instance, the ECC model reached a F-score, precision, recall, and Jaccard score on the CIP dataset of 0.93 ± 0.04, 0.94 ± 0.05, 0.98 ± 0.03, and 0.92 ± 0.06, respectively. Especially, the ECC model significantly outperformed the BR, CC, LP, and RD for predicting

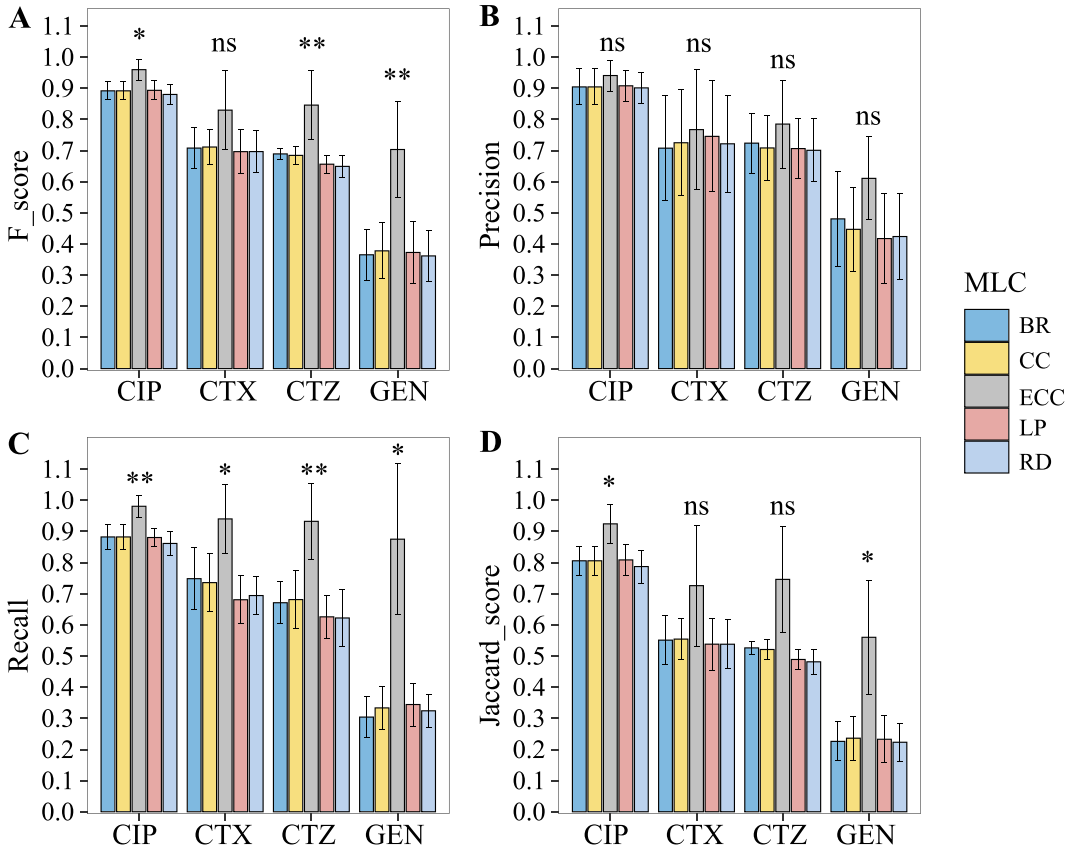

Fig. 2. Performance of different MLC methods with RF base classifiers for resistance prediction for each antibiotic. (A) F-scores, (B) Precision, (C) Recall, and (D) Jaccard score of five MLC methods with RF base classifiers for predicting resistance against each antibiotic. ∗ p < 0.05, ∗∗p < 0.01, ∗∗∗p < 0.001, ns: no significance.

resistance against CIP, CTZ, and GEN based on the F-score metric. Moreover, we observed from the Recall metric that the performance of the ECC model is significantly better than other models, which represents the ECC model has a better sensitivity to detect resistant samples. Besides, the ECC model reached, in general, the highest accuracy, as well as, lowest hamming loss, and 0/1 loss for RF (Table 2). Taken together, our results indicated that the ECC models can significantly improve the prediction performance for MDR prediction in *E. coli*.

3.2. Performance of different MLC methods on LR base classifier

We also compared the performance of the five MLC methods (BR, CC, ECC, LP, and RD) on the LR base classifier. We found the ECC model still got a higher F-score, precision, recall, and Jaccard score (Fig. 3), which showed the consistent performance of the ECC model on LR with RF base classifier. The results on F-score suggested that ECC model is significantly better than other models for CIP, CTZ, and GEN drug, reached 0.94 ± 0.04, 0.80 ± 0.15, and 0.64 ± 0.13 (p-value < 0.05). We also found a similar trend in recall results of the ECC model, and the ECC model achieved a higher sensitivity performance for MDR prediction. Moreover, ECC model significantly outperformed other four MLC methods on CIP and GEN drug based on recall results (0.98 ± 0.03, 0.87 ± 0.23, p-value < 0.05) and Jaccard score (0.89 ± 0.07, 0.48 ± 0.14, p-value < 0.05). As well, the ECC model got the highest accuracy, lowest hamming loss, and 0/1 loss on the LR base classifier (Table 3). These results demonstrated that the ECC model still has robust performance for MDR prediction.

Table 2
Accuracy, hamming loss, and 0/1 loss of five MLC methods with RF base classifier for predicting resistance against four antibiotics. Mean ± standard deviations (significance label of p-value) are shown in table. The statistical significances were compared each group to all (base-mean). ∗p < 0.05, ∗∗p < 0.01, ∗∗∗p < 0.001, ns: no significance.

MLC	Accuracy	Hamming Loss	0/1 Loss
BR	0.51 ± 0.07 (ns)	0.20 ± 0.03 (ns)	0.49 ± 0.07 (ns)
CC	0.52 ± 0.07 (ns)	0.20 ± 0.04 (ns)	0.48 ± 0.06 (ns)
ECC	0.72 ± 0.13 (ns)	0.11 ± 0.05 (*)	0.28 ± 0.13 (ns)
LP	0.53 ± 0.08 (ns)	0.11 ± 0.05 (ns)	0.47 ± 0.08 (ns)
RD	0.51 ± 0.09 (ns)	0.21 ± 0.04 (ns)	0.49 ± 0.09 (ns)

3.3. Performance of different MLC methods on SVM base classifier

For SVM, the F-score of ECC model is significantly better than BR, CC, LP, and RD only for CIP (Fig. 4A) (F-scores of 0.93 ± 0.04, 0.86 ± 0.03, 0.86 ± 0.03, 0.88 ± 0.03, and 0.87 ± 0.04, respectively). There are, however, no significant differences between BR, CC, LP, and RD models. In comparison, CC, LP, and RD did not improve the precision or recall significantly, and in some cases even performed worse compared to the BR (Fig. 4B-C). For the CCs, this might be due to the known problem of error propagation [39]. We found the same conclusion from Jaccard score that the ECC model got better performance than the other four MLC methods, and the Jaccard score of the ECC ranged from 0.42 ± 0.18 for the drug GEN to 0.88 ± 0.07 for the drug CIP (Fig. 4D). Moreover, the ECC model based on the SVM base classifier reached consistent performance with the highest accuracy, lowest hamming loss, and 0/1 loss for RF (Table 4). In summary, the results based on the SVM classifier also demonstrated that the ECC models can significantly improve the prediction performance for MDR prediction in *E. coli*.

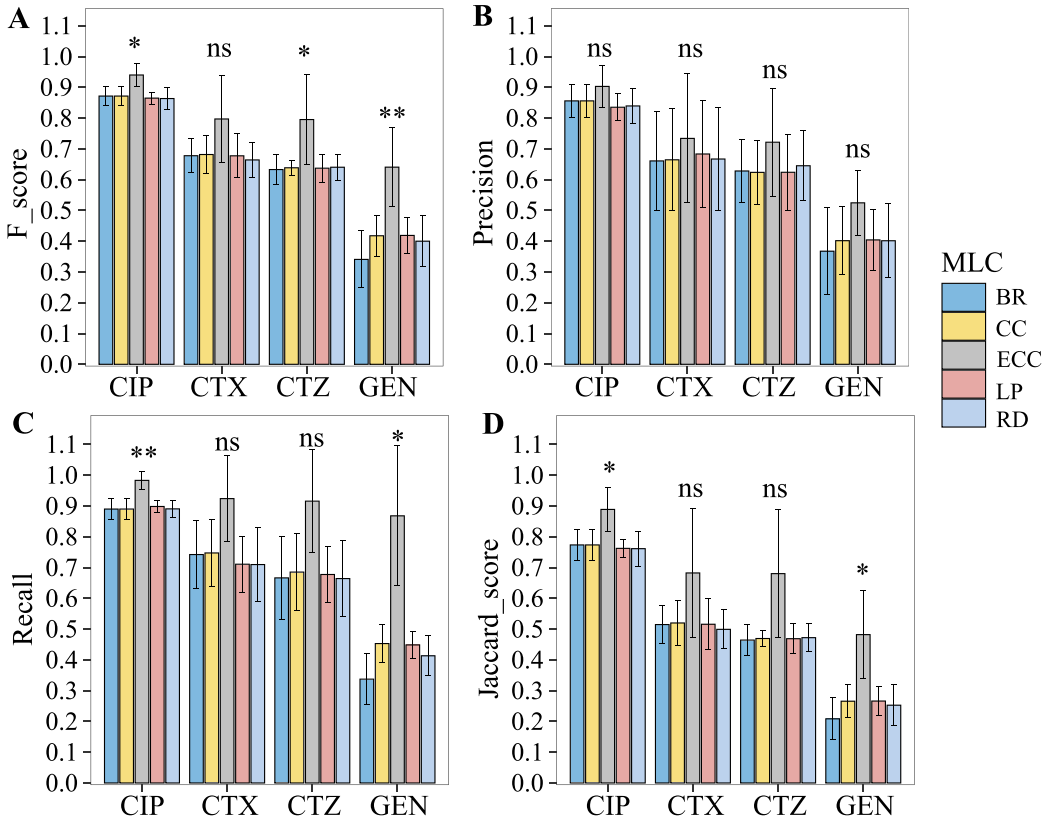

Fig. 3. Performance of different MLC methods with LR base classifiers for resistance prediction for each antibiotic. (A) F-scores, (B) Precision, (C) Recall, and (D) Jaccard score of five MLC methods with RF base classifiers for predicting resistance against each antibiotic. ∗p < 0.05, ∗∗p < 0.01, ∗∗∗p < 0.001, ns: no significance.

Table 3
Accuracy, hamming loss, and 0/1 loss of five MLC methods with LR base classifier for predicting resistance against four antibiotics. Mean ± standard deviations (significance label of p-value) are shown in table. The statistical significances were compared each group to all (base-mean). ∗p < 0.05, ∗∗p < 0.01, ∗∗∗p < 0.001, ns: no significance.

MLC	Accuracy	Hamming Loss	0/1 Loss
BR	0.45 ± 0.08 (ns)	0.24 ± 0.04 (ns)	0.55 ± 0.08 (ns)
CC	0.47 ± 0.08 (ns)	0.23 ± 0.04 (ns)	0.53 ± 0.08 (ns)
ECC	0.65 ± 0.11 (ns)	0.14 ± 0.05 (*)	0.35 ± 0.11 (ns)
LP	0.50 ± 0.08 (ns)	0.23 ± 0.04 (ns)	0.50 ± 0.08 (ns)
RD	0.47 ± 0.07 (ns)	0.24 ± 0.05 (ns)	0.53 ± 0.07 (ns)

4. Discussion

In our study, we compared five MLC models (BR, CC, ECC, LP, and RD) based on three base classifiers (RF, LR, and SVM) for MDR predictions in *E. coli* and evaluated the performance with seven different metrics. Our results illustrated that the ECC model outperforms the other MLC methods and can effectively predict MDR.

The ECC multi-label classification model has a wide range of applications, e.g., for cancers, chronic diseases, and viruses. For instance, Zhou *et al.*, [40] reported that the ECC performed best in the diagnosis of four diabetic complications. ECCs have also been used for cross-resistance prediction in viral infections, e.g., in HIV-1 [25,26]. Here, we firstly applied ECC models on multi-label drug resistance prediction based on all mutations, which could contribute to improving the MDR prediction in other model organisms or poorly known organisms.

Our results also showed that ECC obtained the highest accuracy in all three base classifiers compared to the other four MLC methods, which indicates that the ECC model has good scalability, and can be combined with multiple base classifiers, such as neural networks. Among them, the ECC model based on RF base classifier performs best compared to LR and SVM, which is consistent with our previous research results [20].

The performance of five MLC methods on each drug is different. In general, all MLC methods performed well on CIP drug, and worse on GEN drug. The comparatively lower performance for GEN may be based on the fact that bacterial resistance to GEN is predominantly mediated by plasmids carrying the resistance genes. We focused here solely on chromosomal sequences of the bacteria and did not take into account the effect of alterations in other genetic components on the MDR, like the plasmids, transposons, and integrons [41,42]. This is one of the limitations of our study. The other limitation in our study is our MLC models are built only on four drugs, and we should integrate more types of antibiotics to further investigate the MDR prediction in the future.

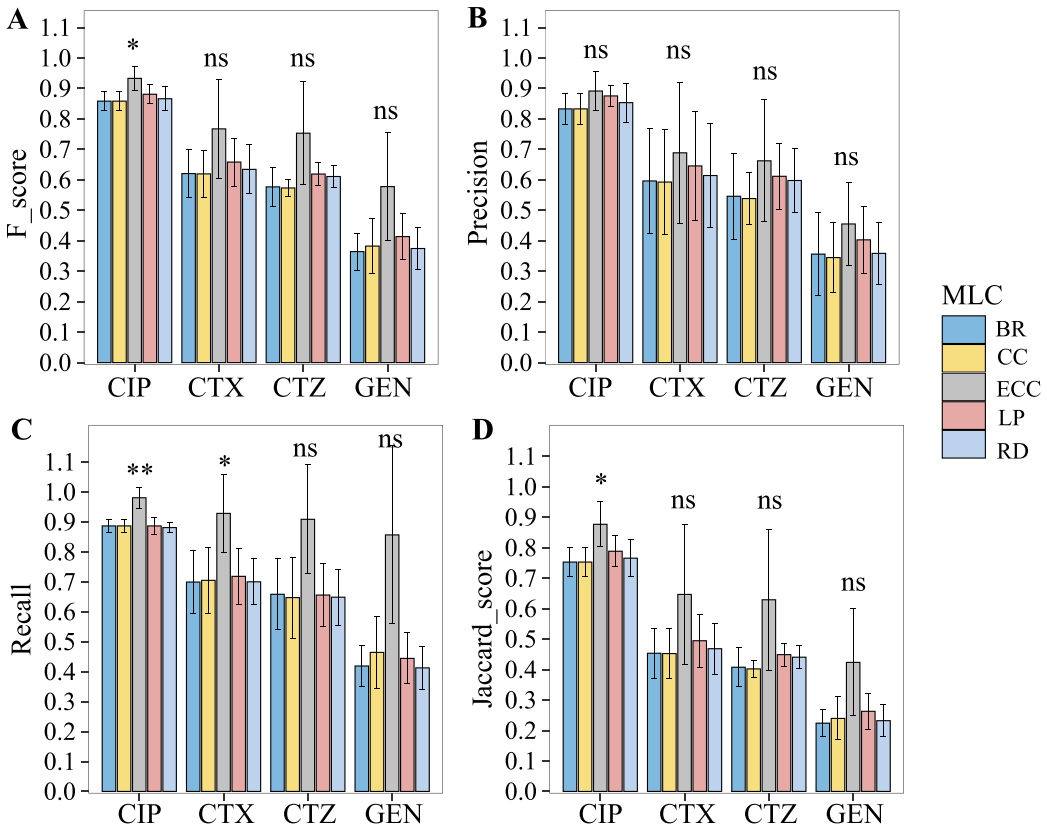

Fig. 4. Performance of different MLC methods with SVM base classifiers for resistance prediction for each antibiotic. (A) F-scores, (B) Precision, (C) Recall, and (D) Jaccard score of five MLC methods with RF base classifiers for predicting resistance against each antibiotic. ∗p < 0.05, ∗∗p < 0.01, ∗∗∗p < 0.001, ns: no significance.

Table 4
Accuracy, hamming loss, and 0/1 loss of five MLC methods with SVM base classifier for predicting resistance against four antibiotics. Mean ± standard deviations (significance label of p-value) are shown in table. The statistical significances were compared each group to all (base-mean). ∗p < 0.05, ∗∗p < 0.01, ∗∗∗p < 0.001, ns: no significance.

MLC	Accuracy	Hamming Loss	0/1 Loss
BR	0.37 ± 0.08 (ns)	0.28 ± 0.05 (ns)	0.63 ± 0.08 (ns)
CC	0.39 ± 0.08 (ns)	0.28 ± 0.05 (ns)	0.61 ± 0.08 (ns)
ECC	0.57 ± 0.12 (ns)	0.18 ± 0.07 (ns)	0.43 ± 0.12 (ns)
LP	0.47 ± 0.07 (ns)	0.24 ± 0.03 (ns)	0.53 ± 0.07 (ns)
RD	0.41 ± 0.09 (ns)	0.26 ± 0.05 (ns)	0.59 ± 0.09 (ns)

5. Conclusions

In summary, our study illustrates five MLC methods based on three base classifiers that achieved accurate MDR prediction. Our results suggest ECC is a promising MLC method for MDR identification, which could be used as a reference approach for clinical staff to improve the diagnostics and patient treatments and thus contribute to reducing the threat of antimicrobial resistance and related deaths in the future.

Data availability

Source codes for data preparation and model training are provided at Github website https://github.com/YunxiaoRen/Multi_Label-Classification.

And the final SNP matrix datasets we used for model training in this paper are also available at https://github.com/YunxiaoRen/Multi_Label-Classification.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank de.NBI - German Network for Bioinformatics for providing cloud computing platform.

Funding

This work is financially supported by the German Federal Ministry of Education and Research (BMBF) under grant number 031L0209B (Deep-iAMR).

Author's contributions

D. H. conceived and supervised the study; Y. R. analyzed the data and drafted the manuscript; S. D., L. F., and J. F. collected the raw sequencing data and the clinical data. O. S. preprocessed the sequencing data and clinical data. D. H., T. C., and A. G. revised the manuscript, and all authors read and approved the final manuscript.

References

[1] Naylor NR, Atun R, Zhu N, et al. Estimating the burden of antimicrobial resistance: a systematic literature review. Antimicrob Resist Infect Control 2018;7:58.

[2] Obolski U, Dellus-Gur E, Stein GY, et al. Antibiotic cross-resistance in the lab and resistance co-occurrence in the clinic: Discrepancies and implications in E. coli. Infect Genet Evol 2016;40:155–61.

[3] Vivas R, Barbosa AAT, Dolabela SS, et al. Multidrug-resistant bacteria and alternative methods to control them: an overview. Microb Drug Resist 2019;25:890–908.

[4] Tanwar J, Das S, Fatima Z, et al. Multidrug resistance: an emerging crisis. Interdisc Perspect Infect Dis 2014;2014:1–7.

[5] Magiorakos A-P, Srinivasan A, Carey RB, et al. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. Clin Microbiol Infect 2012;18:268–81.

[6] Nikaido H. Multidrug resistance in bacteria. Annu. Rev. Biochem. 2009;78:119–46.

[7] Ramadan H, Soliman AM, Hiott LM, et al. Emergence of multidrug-resistant Escherichia coli producing CTX-M, MCR-1, and FosA in retail food from Egypt. Front. Cell. Infect. Microbiol. 2021;11:681588.

[8] Ramírez Castillo FY, Avelar González FJ, Garneau P, et al. Presence of multi-drug resistant pathogenic Escherichia coli in the San Pedro River located in the State of Aguascalientes, Mexico. Front Microbiol 2013;4.

[9] Cag Y, Caskurlu H, Fan Y, et al. Resistance mechanisms. Ann Transl Med 2016; 4:326–326.

[10] Boolchandani M, D'Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. Nat Rev Genet 2019;20:356–70.

[11] Liu Z, Deng D, Lu H, et al. Evaluation of machine learning models for predicting antimicrobial resistance of Actinobacillus pleuropneumoniae from whole genome sequences. Front. Microbiol. 2020;11:48.

[12] Yang Y, Niehaus KE, Walker TM, et al. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. Bioinformatics 2018;34:1666–71.

[13] Kouchaki S, Yang Y, Walker TM, et al. Application of machine learning techniques to tuberculosis drug resistance analysis. Bioinformatics 2019;35:2276–82.

[14] Radha M, Fonseca P, Moreau A, et al. A deep transfer learning approach for wearable sleep stage classification with photoplethysmography. NPJ Digit Med 2021:4:135.

[15] Arango-Argoty GA, Garner E, Pruden A, et al. DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. 2017.

[16] Veltri D, Kamath U, Shehu A. Deep learning improves antimicrobial peptide recognition. Bioinformatics 2018;34:2740–7.

[17] Her H-L, Wu Y-W. A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the Escherichia coli strains. Bioinformatics 2018;34:i89–95.

[18] Kavvas ES, Catoiu E, Mih N, et al. Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. Nat Commun 2018;9:4306.

[19] Khaledi A, Weimann A, Schniederjans M, et al. Predicting antimicrobial resistance in *Pseudomonas aeruginosa* with machine learning-enabled molecular diagnostics. EMBO Mol Med 2020.

[20] Ren Y, Chakraborty T, Doijad S, et al. Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. Bioinformatics 2021:btab681.

[21] Tsoumakas G, Katakis I, Vlahavas I. Mining Multi-label Data. Data Mining and Knowledge Discovery Handbook 2009; 667–685.

[22] Rokach L, Schclar A, Itach E. Ensemble methods for multi-label classification. Expert Syst Appl 2014;41:7507–23.

[23] Read J, Pfahringer B, Holmes G, et al. Classifier chains: A review and perspectives. JAIR 2021; 70:683–718.

[24] Read J, Pfahringer B, Holmes G, et al. Classifier chains for multi-label classification. 2011; 16

[25] Heider D, Senge R, Cheng W, et al. Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction. Bioinformatics 2013;29:1946–52.

[26] Riemenschneider M, Senge R, Neumann U, et al. Exploiting HIV-1 protease and reverse transcriptase cross-resistance information for improved drug resistance prediction by means of multi-label classification. BioData Mining 2016;9:10.

[27] Falgenhauer L, Nordmann P, Imirzalioglu C, et al. Cross-border emergence of clonal lineages of ST38 Escherichia coli producing the OXA-48-like carbapenemase OXA-244 in Germany and Switzerland. Int J Antimicrob Agents 2020;56:106157.

[28] Heeb S, Fletcher MP, Chhabra SR, et al. Quinolones: from antibiotics to autoinducers. FEMS Microbiol Rev 2011;35:247–74.

[29] Sharma M. Prevalence and antibiogram of Extended Spectrum β-Lactamase (ESBL) producing Gram negative bacilli and further molecular characterization of ESBL producing Escherichia coli and Klebsiella spp.. JCDR 2013.

[30] Gums JG, Boatwright DW, Camblin M, et al. Differences between ceftriaxone and cefotaxime: microbiological inconsistencies. Ann Pharmacother 2008;42:71–9.

[31] Garneau-Tsodikova S, Labby KJ. Mechanisms of resistance to aminoglycoside antibiotics: overview and perspectives. Medchemcomm 2016;7:11–27.

[32] Chen S, Zhou Y, Chen Y, et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018;34:i884–90.

[33] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60.

[34] Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. GigaScience 2021;10(giab008).

[35] Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078–9.

[36] Junior JDC, Faria ER, Silva JA, et al. Label powerset for multi-label data streams. Classification with Concept Drift. 2017;9.

[37] Dembczyński K, Waegeman W, Cheng W, et al. Regret analysis for performance metrics in multi-label classification: the case of hamming and subset zero-one loss. Mach Learn Knowl Disc Datab 2010;6321:280–95.

[38] Shikalgar NR. JIBCA: Jaccard index based clustering algorithm for mining online review. Int J Comput Appl 105:6.

[39] Senge R, del Coz JJ, Hüllermeier E. On the problem of error propagation in classifier chains for multi-label classification. Data Anal Mach Learn Knowl Discov 2014; 163–170.

[40] Zhou H, Beltrán JF, Brito IL. Functions predict horizontal gene transfer and the emergence of antibiotic resistance. Sci Adv 2021;7:eabj5056.

[41] Alekshun MN, Levy SB. Molecular mechanisms of antibacterial multidrug resistance. Cell 2007;128:1037–50.

[42] Karczmarczyk M, Abbott Y, Walsh C, et al. Characterization of multidrug-resistant Escherichia coli isolates from animals presenting at a university veterinary hospital. Appl Environ Microbiol 2011;77:7104–12.

3.3 Publication 3: Deep Transfer Learning Enables Robust Prediction of Antimicrobial Resistance for Novel Antibiotics

3.3.1 Summary

Aim and Motivation

This paper (Ren et al., 2022b) aims to explore strategies for overcoming the difficulties posed by data constraints and label imbalances, which are common obstacles of machine learning. Machine learning model training often encounters hurdles due to data size limitations and skewed data distributions, which can negatively impact the accuracy and generalizability of the models. This problem is particularly prominent in many medical diagnostic datasets, such as those used for cancer diagnosis, where the datasets are unbalanced and usually consist of a relatively small number of samples. However, machine learning models often require a large number of data for training. This challenge is not exclusive to the medical field but is also faced in the development of novel antibiotics. Employing transfer learning (TL) holds the potential for effectively addressing these issues.

Methods and Results

Building upon our prior research (Ren et al., 2021), it was observed that our models, particularly the CNN, exhibited impressive performance in AMR prediction based on whole-genome mutations. However, the performance could be enhanced when dealing with imbalanced label distribution. To address this, we initially constructed a fundamental CNN model for each antibiotic included in our dataset, namely CIP, CTX, CTZ, and GEN. We then selected the best-performing CNN, the model for CIP, as our pre-trained model, leveraging its learned knowledge to enhance the prediction for the remaining antibiotics: CTX, CTZ, and GEN.

Our results illustrated that transfer learning can notably improve the prediction performance for other antibiotics. Furthermore, our research demonstrated that the pre-trained model can effectively generalize to unseen, extremely imbalanced public datasets characterized by a small number of samples for the resistance class.

Conclusion

To summarize, we offer a deep transfer learning model capable of achieving accurate and robust AMR prediction on small, imbalanced datasets. By combining secondary mutation profiles with our pre-trained network, we lay the groundwork for future training tasks dealing with AMR in small, imbalanced datasets. This approach can contribute to the development of comprehensive solutions for novel antibiotics and future AMR challenges.

antibiotics

MDPI

Article

Deep Transfer Learning Enables Robust Prediction of Antimicrobial Resistance for Novel Antibiotics

Yunxiao Ren [1,2], Trinad Chakraborty [3,4], Swapnil Doijad [3,4], Linda Falgenhauer [4,5,6], Jane Falgenhauer [3,4], Alexander Goesmann [4,7], Oliver Schwengers [4,7] and Dominik Heider [1,2,*]

1 Department of Data Science in Biomedicine, Faculty of Mathematics and Computer Science, Philipps-University of Marburg, 35032 Marburg, Germany
2 Center for Synthetic Microbiology (SYNMIKRO), Philipps-University of Marburg, 35032 Marburg, Germany
3 Institute of Medical Microbiology, Justus Liebig University Giessen, 35392 Giessen, Germany
4 German Center for Infection Research, Partner Site Giessen-Marburg-Langen, 35392 Giessen, Germany
5 Institute of Hygiene and Environmental Medicine, Justus Liebig University Giessen, 35392 Giessen, Germany
6 Hessisches Universitäres Kompetenzzentrum Krankenhaushygiene, 35392 Giessen, Germany
7 Department of Bioinformatics and Systems Biology, Justus Liebig University Giessen, 35392 Giessen, Germany
* Correspondence: dominik.heider@uni-marburg.de

Abstract: Antimicrobial resistance (AMR) has become one of the serious global health problems, threatening the effective treatment of a growing number of infections. Machine learning and deep learning show great potential in rapid and accurate AMR predictions. However, a large number of samples for the training of these models is essential. In particular, for novel antibiotics, limited training samples and data imbalance hinder the models' generalization performance and overall accuracy. We propose a deep transfer learning model that can improve model performance for AMR prediction on small, imbalanced datasets. As our approach relies on transfer learning and secondary mutations, it is also applicable to novel antibiotics and emerging resistances in the future and enables quick diagnostics and personalized treatments.

Keywords: transfer learning; antimicrobial resistance; small data with imbalanced label

check for updates

Citation: Ren, Y.; Chakraborty, T.; Doijad, S.; Falgenhauer, L.; Falgenhauer, J.; Goesmann, A.; Schwengers, O.; Heider, D. Deep Transfer Learning Enables Robust Prediction of Antimicrobial Resistance for Novel Antibiotics. *Antibiotics* **2022**, *11*, 1611. https://doi.org/10.3390/antibiotics11111611

Academic Editor: Asad Mustafa Karim

Received: 18 October 2022
Accepted: 10 November 2022
Published: 12 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Antimicrobial resistance (AMR) has become one of the serious public health problems worldwide, threatening the effective treatment of a growing number of infections [1]. There were over 700,000 deaths from drug-resistant infections in 2019, and it could rise to 10 million deaths by 2050 according to estimations from the World Health Organization (WHO) [2].

Machine learning and deep learning approaches have played significant roles in antibiotic resistance prediction in recent years [3–6]. A number of deep-learning-based models and tools for predicting AMR genes or peptides have been developed, e.g., DeepARG [7] or Deep-AmPEP [8]. These methods also promoted the discovery of new antibiotics. For example, Stokes et al. trained a deep learning model based on multiple chemical libraries [9]. They found a molecule showing bactericidal activity against a broad phylogenetic spectrum of pathogens, and thus has the potential to be the basis for a new antibiotic [9]. However, skewed distribution of the data in machine learning often obstructs the accuracy and generalization of model training [10]. In fact, many datasets about medical diagnoses, such as cancer diagnostics, are imbalanced datasets and typically have a low number of samples [10]. For training a machine learning model, a large number of samples is necessary. However, these data are typically not available for novel antibiotics.

Transfer learning (TL) has shown promising applications for such challenges in recent years [11–18]. The basic idea of transfer learning is to transfer knowledge from source domains to target domains for improving the model performance [11,15,19]. In contrast to

traditional machine learning (including deep learning), having only one domain and one task, transfer learning extends the notion of domain and task, in which the domains and tasks between the training and test data can be different but related in some ways [20–22]. Generally, the source domain is a set of data with a large number of data samples with high-quality labels. In contrast, data in the target domain may include a limited number of samples with unbalanced labels. Thus, transfer learning is widely used to solve the issue with limited datasets for visual classification and text classification [21,23–27]. For example, many researchers firstly trained a convolutional neural network (CNN) model on the ImageNet dataset (pre-training) and then transferred the information from the pre-trained model into a new task (fine-tuning) to solve a wide range of computer vision problems [23–25]. The Word2Vec dataset is also commonly used as a pre-training dataset for text classification [28]. Gupta et al. enhanced predictive analysis on small data using a cross-property deep transfer learning model [29]. Park et al. used meta-transfer learning to explore the data heterogeneity and extremely small sample size problem based on single cell data [30]. Transfer learning is also widely used in the medical area with an imbalanced label [10,31–34]. For example, Gao et al. used deep transfer learning to reduce healthcare disparities arising from imbalanced biomedical data [35]. They first trained the model on the majority group data, then transferred the knowledge learned to each minority group to improve the model performance. Thus, our study aims to transfer the knowledge from a well-trained model to a small amount of imbalanced label data to explore whether the performance for AMR prediction can be improved.

Based on our previous work [6], our models, especially the CNN, performed well for AMR prediction based on whole genome mutations, while the performance on the data with the imbalanced label can still be improved. Therefore, in our work, we firstly constructed a basic CNN model for each antibiotic in our dataset, including ciprofloxacin (CIP), cefotaxime (CTX), ceftazidime (CTZ), and gentamicin (GEN). We then used the model for CIP, i.e., the best-performing CNN, as the pre-trained model and transferred the knowledge to improve the prediction of the other three antibiotics, i.e., CTX, CTZ, and GEN (see Study design). Our results show that transfer learning can significantly improve the prediction performance on the other antibiotics. Our work also illustrates that the pre-trained model can generalize well on unseen public datasets that are extremely imbalanced, i.e., have a low number of samples for the resistance class. We provide a deep transfer learning model that can achieve accurate and robust AMR prediction on small, imbalanced datasets. By combining secondary mutation profiles and our pre-trained network, we pave the way for other training tasks concerning AMR with small, imbalanced datasets in the future, and thus enable a quick and generic solution for novel antibiotics and AMR in the future.

2. Results

2.1. Datasets

In this work, we used two datasets of *Escherichia coli* (*E. coli)* with whole-genome sequencing (WGS) and resistance information for four antibiotics, namely ciprofloxacin (CIP), cefotaxime (CTX), ceftazidime (CTZ), and gentamicin (GEN). The first dataset contains 809 *E. coli* strains, produced by our laboratory. The isolates were collected from human and animal clinical samples. Antimicrobial susceptibility testing was performed using the VITEK® 2 system (bioMérieux, Nürtingen, Germany) and interpreted following EUCAST guidelines. DNA isolation and whole-genome sequencing were performed as described in Falgenhauer et al. [36]. The percentage of isolates resistant to CIP, CTX, CTZ, and GEN are 45%, 44%, 34%, and 23%, respectively (see Figure 1). This dataset was split into the training dataset and testing dataset (see Section 2.2). The second dataset comprises 1509 *E. coli* strains collected from public datasets [37]. This dataset is highly imbalanced concerning resistant and sensitive isolates. The isolates that are resistant to CIP, CTX, CTZ, and GEN are 18%, 8%, 5%, and 7% of all isolates, respectively (see Figure 1). We used this dataset as

the external validation dataset to demonstrate the application of transfer learning on an imbalanced, small, and unseen dataset.

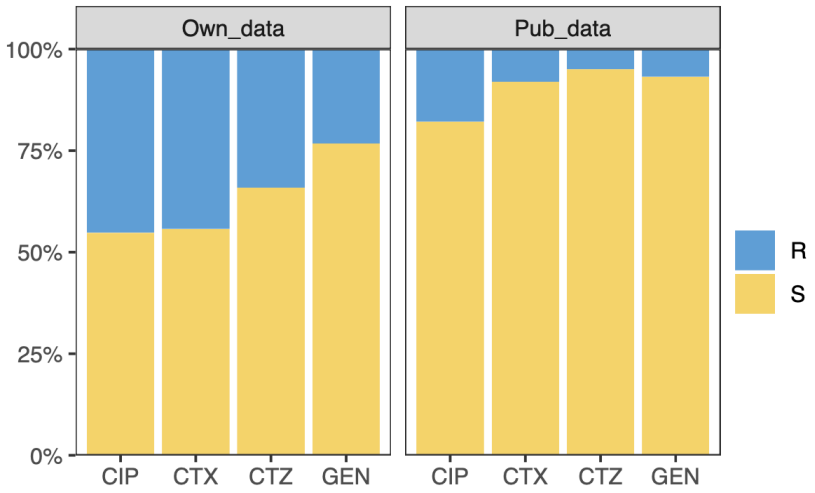

Figure 1. Overview of the samples. The samples are resistant (R) or susceptible (S) to ciprofloxacin (CIP), cefotaxime (CTX), ceftazidime (CTZ), and gentamicin (GEN). The left and right panel show the resistant and susceptible sample information on our and public dataset considered for this study, respectively.

2.2. Study Design

Transfer learning generally uses a known pre-trained model with a large amount of data as the source model [12,14,19,38]. Here, we used the model that performs the best on our AMR dataset as the pre-trained model instead of the public uncorrelated dataset. Thus, we firstly constructed basic CNN architectures for each antibiotic with our data (see Figure 2). The CNN architectures were implemented using the Keras (https://keras.io/, accessed on 15 October 2021) package and TensorFlow (https://tensorflow.org, accessed on 15 October 2021). We evaluated the performance of the CNNs based on accuracy, receiver operating characteristics curve (ROC), and the precision–recall curve (P_R curve), then selected the best-performing model, namely the CIP model, as the source model for transfer learning. The source model based on CIP data not only performed well, but more importantly, the source task was also closely related to the other target tasks, i.e., the prediction of CTX, CTZ, and GEN resistance. We thus transferred the architecture and weights of the source model from the CIP data and retrained the model with CTX, CTZ, and GEN, respectively (see Figure 2). Our dataset was separated into a test set with 20% of the samples, and the remaining data were used for fivefold cross-validation to split the training set and validation set. The public dataset was used as an external validation set to further validate the performance of the models on independent data.

2.3. Performance of the Basic CNN Models

We built basic convolutional neural network (CNN) models for each antibiotic in our dataset [6]. The dataset was randomly split at 20% to create a testing set, and the remaining data was used in fivefold cross-validation, where we trained the models and fine-tuned the hyper-parameters. We observed that the training accuracy and validation accuracy of the CNN model on CIP data reached a plateau around 0.98 and 0.91, respectively, and there is less bias in each cycle training process (see Figure 3). The training and validation accuracies of the other CNNs trained on the other antibiotics were lower, e.g., the CTX model had accuracies of around 0.89 and 0.79 for training and validation (see Figure 3). For the CTZ data, the training and validation accuracies of the model in fivefold cross-validation were around 0.87 and 0.83. For the GEN data, the accuracies were around 0.86 and 0.79 (see Figure 3). These results indicate that the model on CIP data has the highest accuracy

compared with the other models on CTX, CTZ, and GEN data. Thus, we selected the CIP model as the source model for transfer learning.

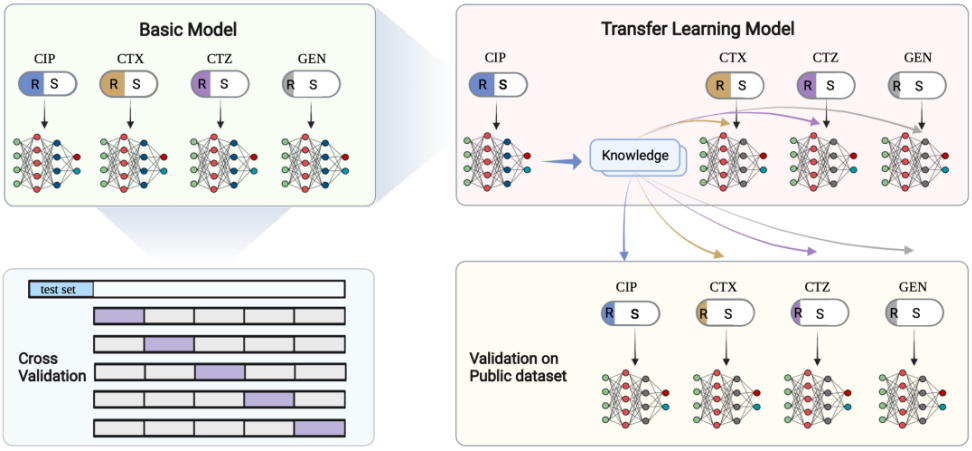

Figure 2. Deep transfer learning schemes. In the top left panel, the basic CNN models are shown. Each model is trained on independent antibiotics and evaluated on a new dataset. The top right panel shows the model trained on CIP that is then used as the pre-trained model to transfer the knowledge to the other three antibiotics. The bottom left panel shows the 5-fold cross-validation scheme. The dataset was firstly split, and 20% was used for testing. The remaining data were used in the cross-validation. The bottom right panel shows our validation scheme for the transfer learning model on an independent public dataset. This figure was created with BioRender.com.

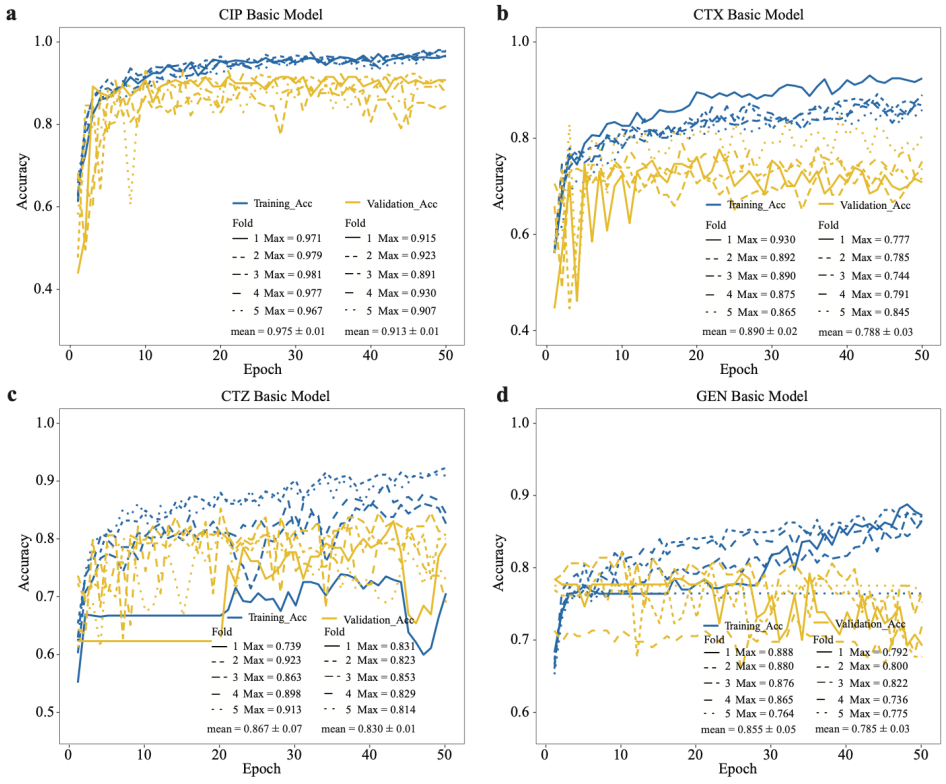

Figure 3. Accuracy of basic CNN models on training and validation datasets based on our dataset. Training accuracy and validation accuracy on (**a**) CIP, (**b**) CTX, (**c**) CTZ, and (**d**) GEN. The legend shows the maximum accuracy in each fold and its mean value.

We also evaluated the model performance on the testing set using the receiver operating characteristics curve (ROC) and the precision–recall curve (P_R curve). We observed the same results based on the area under the ROC (AUROC) and P_R curves (AUPRC) for CIP (0.97 ± 0.01, 0.95 ± 0.01) and CTX (0.78 ± 0.02, 0.75 ± 0.01) testing data (see Figure 4), which show that the CNN model can generalize well. However, the AUROC and AUPRC are much lower for CTZ (0.75 ± 0.07, 0.64 ± 0.01) and GEN (0.81 ± 0.02, 0.55 ± 0.02) in the testing datasets (see Figure 4).

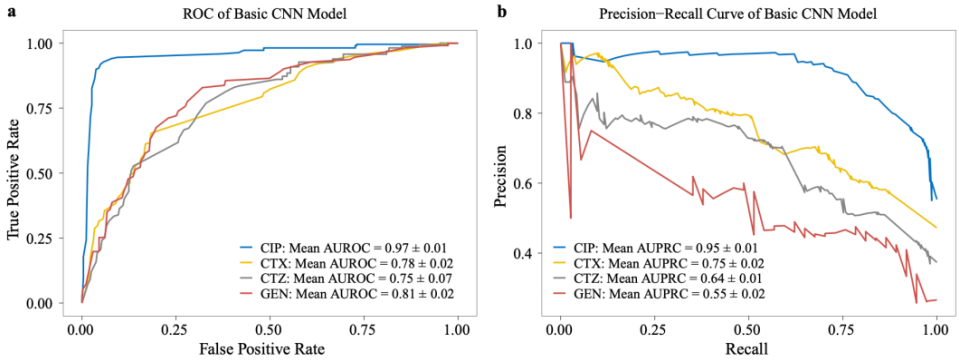

Figure 4. Performance of basic models on the testing dataset of our dataset. (**a**) The ROC curve and (**b**) precision–recall curve (P_R) on CIP, CTX, CTZ, and GEN antibiotics.

2.4. Deep Transfer Learning Improves the Model Performance on the Minority Group

Based on the basic CNN model's performance, we used the model trained on CIP data as the pre-trained model, transferred the learned weights, and retrained the models for CTX, CTZ, and GEN. To evaluate the model performance on the imbalanced datasets, we used the Matthews correlation coefficient (MCC) as one of the evaluation metrics, which is widely used for dealing with binary classification problems on imbalanced data [39–41]. Since we are more interested in the resistance phenotype, we also compared the F1 score regarding resistance (F1-R). Our results show that the transfer learning model significantly improves MCC for CTX (p = 0.009), CTZ (p = 0.023), and GEN (p = 0.001) compared with the basic models (see Figure 5a, Table 1). Moreover, the F1-Rs for CTX (p = 0.007), CTZ (p = 0.014), and GEN ($p = 6.1 \times 10^{-5}$) of the transfer learning models were significantly higher than the basic models (see Figure 5b, Table 1). We also observed that the maximum accuracy of the transfer learning models stabilize over 0.9 in both the training and validation sets for CTX, CTZ, and GEN. Thus, all of them were significantly improved (Figure 6). These results indicate that transfer learning can improve the model performance, especially for the minority groups, and thus is also applicable for small, imbalanced datasets.

Table 1. MCC values and F1-R values (F1 on resistance class) of deep transfer learning models and basic CNN models on the testing set of our dataset.

Drugs	**CTX**		**CTZ**		**GEN**	
Metrics	**MCC**	**F1-R**	**MCC**	**F1-R**	**MCC**	**F1-R**
Basic	0.47 ± 0.03	0.70 ± 0.02	0.46 ± 0.03	0.65 ± 0.02	0.33 ± 0.01	0.41 ± 0.02
TL	0.56 ± 0.03	0.76 ± 0.02	0.55 ± 0.03	0.71 ± 0.02	0.53 ± 0.03	0.63 ± 0.02

2.5. Model Evaluation on Independent Public Data

We further evaluated the deep transfer learning models on an independent public dataset. The public dataset contains data from *E. coli* resistance to the four antibiotics, CIP, CTX, CTZ, and GEN. There is an extreme imbalance between resistant and susceptible phenotypes in this dataset, with a very low number of resistant strains (see Figure 1). We firstly evaluated the model performance based on the MCC metric, which shows that

the transfer learning models are significantly better than the original models for CTX ($p = 4.6 \times 10^{-3}$), CTZ ($p = 5.6 \times 10^{-4}$), and GEN ($p = 6.9 \times 10^{-3}$) (see Figure 7a, Table 2). Again, we also observed that the F1-Rs of the transfer learning models were significantly higher than for the basic models for CTX, CTZ, and GEN data (see Figure 7b, Table 2). The MCC and F1-R of the transfer learning model for CIP data were also better than for the basic model. Moreover, we compared the transfer learning models and basic models based on AUROC and AUPRC metrics. The AUROC results suggest that transfer learning significantly improved drug resistance prediction for CTX ($p = 2.4 \times 10^{-4}$) and CTZ ($p = 0.012$) (see Figure 7c, Table 2). Moreover, the results of AUPRC show that the transfer learning models significantly improved for CTX ($p = 7.1 \times 10^{-3}$), CTZ ($p = 4.1 \times 10^{-4}$), and GEN ($p = 8.1 \times 10^{-3}$) (see Figure 7d, Table 2). Taken together, the results on the public dataset also clearly show that the deep transfer learning models can compensate for class imbalance and thus improve AMR prediction also for small, imbalanced datasets, and thus is also a very promising approach for novel antibiotics in the future where available data on resistance are limited.

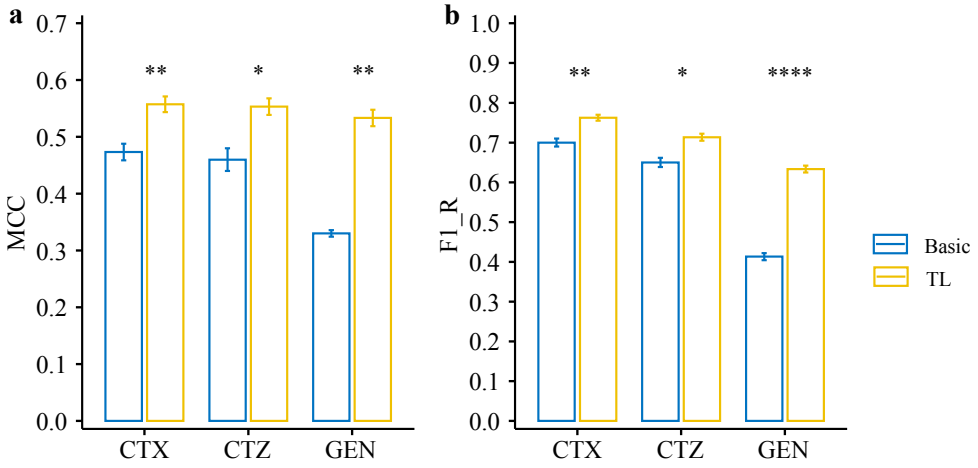

Figure 5. Performance comparison between deep transfer learning models and basic CNN models on the testing set of our dataset. (**a**) MCC of the deep transfer learning models and basic CNN models on each dataset. (**b**) F1_R (F1 resistance) of the deep transfer learning models and basic CNN models on each dataset. Statistical comparisons were performed using the Student's *t*-test. * $p < 0.05$; ** $p < 0.01$; **** $p < 0.0001$.

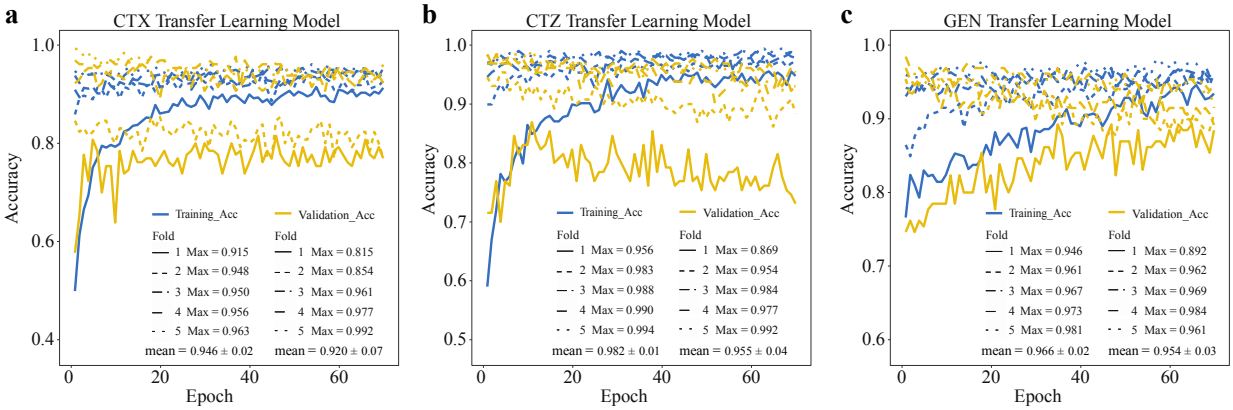

Figure 6. Accuracy of deep transfer learning models on training and validation datasets on our data. Training accuracy and validation accuracy of deep transfer learning models on (**a**) CTX, (**b**) CTZ, and (**c**) GEN. The legends show the maximum accuracy in each fold and its mean value.

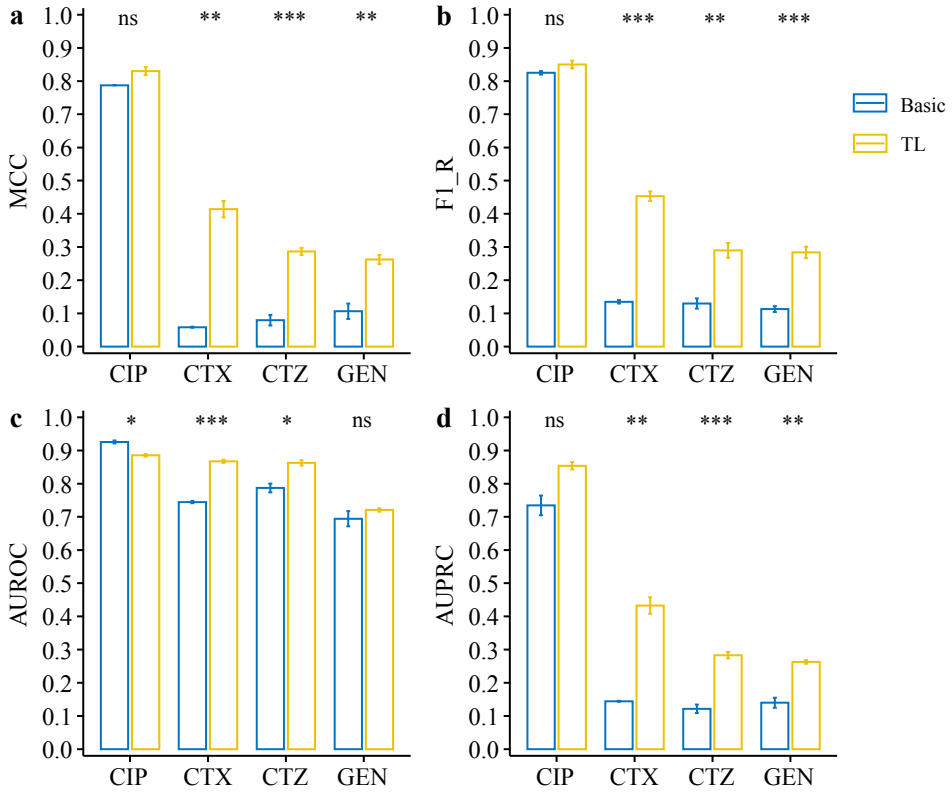

Figure 7. Performance comparison between deep transfer learning models and basic CNN models on the testing dataset of the public dataset. (**a**) MCC of the deep transfer learning models and basic CNN models on each dataset. (**b**) F1_R (F1 resistance) of the deep transfer learning models and basic CNN models on each dataset. (**c**,**d**) AUC of ROC curve (**c**) and precision–recall curve (**d**) of the deep transfer learning models and basic CNN models on each dataset. Statistical comparisons were performed using the Student's *t*-test. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; ns: not significant.

Table 2. MCC values, F1-R values (F1 on resistance class), AUROC, and AUPRC of deep transfer learning models and basic CNN models on the testing set of public dataset.

Drugs	**CIP**		**CTX**		**CTZ**		**GEN**	
Model	**Basic**	**TL**	**Basic**	**TL**	**Basic**	**TL**	**Basic**	**TL**
MCC	0.79 ± 0.00	0.83 ± 0.02	0.06 ± 0.00	0.41 ± 0.04	0.08 ± 0.03	0.29 ± 0.02	0.11 ± 0.04	0.26 + 0.03
F1-R	0.83 ± 0.01	0.85 ± 0.02	0.14 ± 0.01	0.45 ± 0.03	0.13 ± 0.03	0.29 ± 0.05	0.11 ± 0.02	0.28 + 0.04
AUROC	0.93 ± 0.01	0.89 ± 0.01	0.74 ± 0.00	0.87 ± 0.01	0.79 ± 0.02	0.86 ± 0.02	0.69 ± 0.04	0.72 + 0.01
AUPRC	0.73 ± 0.04	0.85 ± 0.02	0.14 ± 0.00	0.43 ± 0.04	0.12 ± 0.02	0.28 ± 0.02	0.14 ± 0.03	0.26 + 0.01

3. Discussion

In this work, we propose a deep transfer learning model that performs well on small, imbalanced data for AMR prediction. Transfer learning typically pre-trains a model on a larger well-known dataset [30,38]. Here, we used a CNN model on a balanced dataset (CIP dataset) with high accuracy as the pre-trained model. The knowledge obtained from the pre-trained model was then transferred to other datasets with resistance to CTX, CTZ, and GEN. We found that our deep transfer learning model can significantly improve the prediction performance compared with the basic CNN models, ranging from 0.06–0.22 based on different evaluation metrics (see Figure 5, Table 1). Especially, the results indicate that our deep transfer learning model can facilitate the resistance prediction on small, imbalanced

datasets. These findings are also supported and validated by an independent evaluation with an unseen, public dataset. The performance was significantly improved, ranging from 0.02–0.35 based on different evaluation metrics (see Figure 7, Table 2). Moreover, we can extend our approach to other species and various antibiotic drugs using our pre-trained model in the future, which will improve the accuracy of resistance prediction and save treatment time, especially for small data sizes with imbalanced labels.

Another interesting result is that we found the performance for CIP data on the public dataset is better than for CTX, CTZ, and GEN public datasets. This result indicates that the closer the correlation between the source task and target task is, the better the performance of the final models. Thus, it is more important to focus on the relevance between the source task and the target tasks when we choose the source domain. The evaluation metrics of the models should be carefully chosen when we are faced with extreme class imbalance. In this article, we provide the commonly used evaluation metrics such as the F1 score, ROC curve, and P_R curve, as well as the evaluation metrics applicable to imbalanced data such as the MCC.

Transfer learning has gained more attention in recent years. For example, Al-Stouhi et al. previously proposed that transfer learning can be used to solve class imbalance problems with inadequate data and provided theoretical and empirical validation on healthcare and text classification applications [10]. Minvielle et al. explored the impact of class imbalance using transfer learning on decision trees [33]. However, only a few studies have been carried out on AMR so far. The proportion of the susceptible and resistant isolates in AMR datasets varies depending on the antibiotic/bacterial species combinations. For the majority of the antibiotics, the AMR data are imbalanced, and the resistant classes of interest are in the minority group. This is particularly true for novel antibiotics in the future, where data of resistant strains are limited. Therefore, our proposed deep transfer learning model paves the way to improve AMR prediction accuracy, as well as for small datasets of novel antibiotics in the future. Moreover, in this analysis, we aimed at identifying secondary mutations that contribute to the resistance directly or indirectly, e.g., compensatory mutations. Thus, we did not include the known resistance genes. Our pre-trained model may not be as effective in predicting resistance due to the transfer of resistance genes compared with resistance due to mutations. Our approach does not need any AMR expert knowledge and can also predict resistance even without knowing the resistance genes by identifying secondary mutations. By combining this data-driven approach with transfer learning, AMR predictions can be significantly improved. It can also be used when only small data are available and information on resistance mechanisms is missing or when the resistance mechanisms are not fully understood yet, e.g., for novel antibiotics.

4. Materials and Methods

4.1. Data Pre-processing

We performed quality checking and filtering on the raw whole-genome sequencing reads using fastp (v0.23.2) software [42]. The filtered reads were then aligned to the *E. coli* reference genome (*E. coli* K-12 strain. MG1655) using BWA-mem with default parameters [43]. We then called variants from the sequencing data using Bcftools software (v1.14) via the "call" function with default parameters [44]. We extracted SNPs variants, reference alleles, and their positions and merged all isolates based on the positions of reference alleles. We filtered out the loci without variation (N replaces a locus without variation) and retained the existing allele variants of more than half in samples. The final SNP matrix, where each column represents the variant allele, and each row is a sample, was encoded into numerical values by one-hot encoding that can be used for subsequent machine learning. The pre-processing process was carried out according to Ren et al. [6].

4.2. Basic CNN Model

We used the Keras (https://keras.io/, accessed on 15 October 2021) and Tensorflow (https://tensorflow.org, accessed on 15 October 2021) Python packages to build the CNN models. We evaluated different topologies in the training data and found that a model with 12 layers performed the best. Thus, the architecture of the CNN models (see Figure 8a) contains twelve layers, including four convolutional layers with a kernel size of 3, implemented by the Conv1D function, two pooling layers using the MaxPooling1D function, two batch normalization layers, one flattening layer, one fully connected layer with 128 nodes followed by a dropout layer, and one output layer with the "softmax" activation function. We used the "categorical_crossentropy" loss function and the "Adam" optimizer function to compile the CNN models with 50 epochs. In order to improve the computation speed, we split the data into multiple small batches, with a batch size of 8.

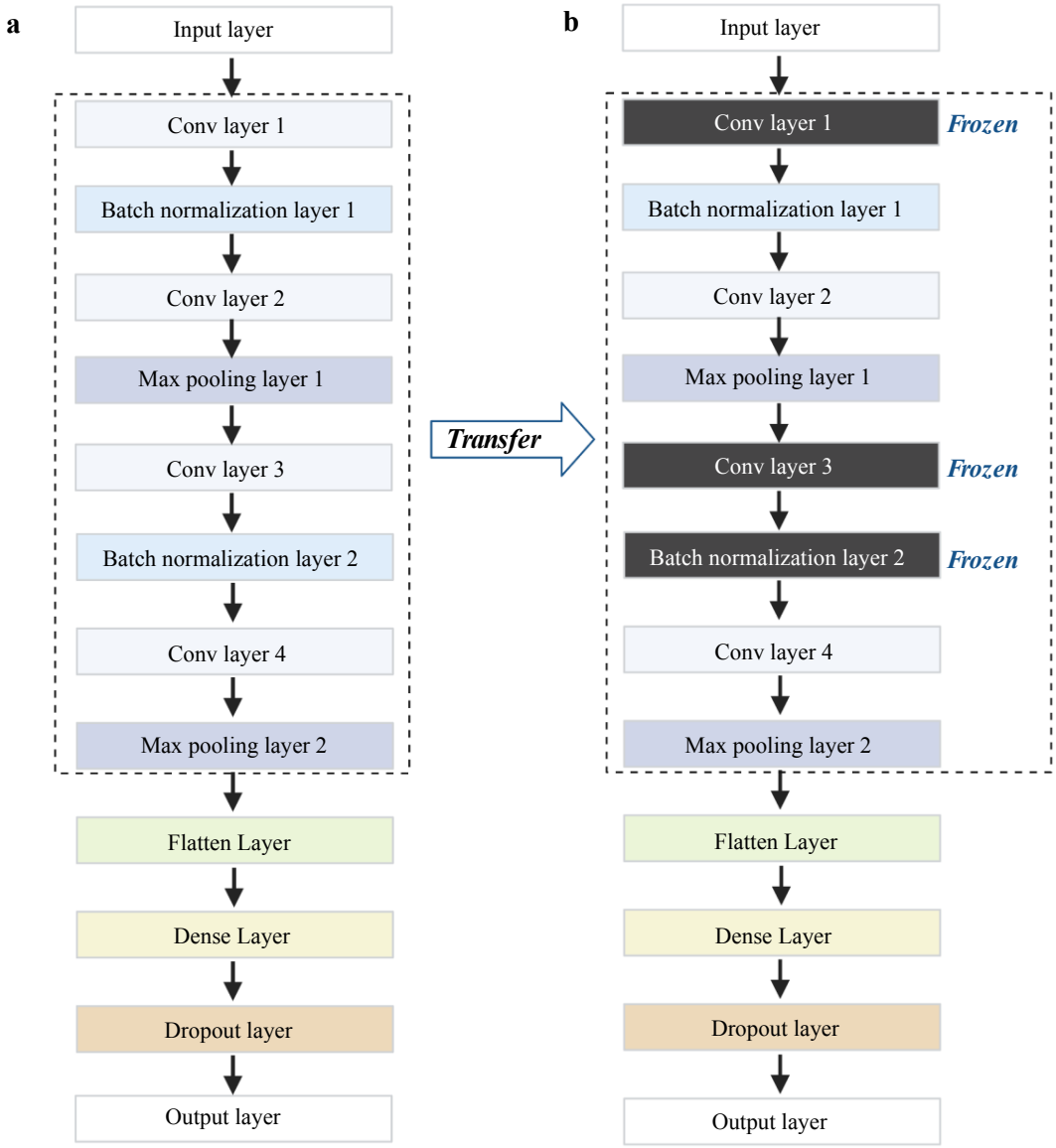

Figure 8. Our framework of basic CNN models and transfer learning models. (**a**) The architecture of the basic CNN models. (**b**) The architecture of the transfer learning models. Conv layer represents convolution layers. This figure was created with BioRender.com.

4.3. Deep Transfer Learning Architecture

In order to facilitate the model performance on small, imbalanced data, we employed deep transfer learning. The deep learning architecture is built based on the basic CNN models as previously described (see Figure 8b). In transfer learning, we have to specify the source domain Ds and the target domain Dt and the source task Ts and the target task Tt [38]. Here, we used the CIP dataset from our lab as the source domain Ds; CTX, CTZ, and GEN datasets were used as the target domain Dt. The tasks of Ts and Tt are predicting AMR against different antibiotics. We incorporated two transfer learning strategies, namely fine-tuning and freezing in our work. The fine-tuning strategy is a common deep transfer learning approach based on transferring parameters (weights) from the Ds model to the Dt models [38]. Therefore, we transferred the parameters (weights) of the model trained on CIP into the CTX, CTZ, GEN models, respectively. Furthermore, we froze two normalization layers and one convolution layer and retrained the CNN models on other layers to avoid overfitting [17].

4.4. Model Evaluation Metrics

Accuracy, precision, and recall are the basic evaluation metrics for classification models in our study. Accuracy measures the fraction of correct predictions, including positive and negative samples [45]. For binary classification, it can be calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

where *TP* = True Positives (the predicted positive value matches the actual positive value), *TN* = True Negatives (the predicted negative value matches the actual negative value), *FN* = False Negatives (the actual positive value was predicted as negative value), and *FP* = False Positives (the actual negative value was classified as positive value). Precision represents the ratio of true positives to the total predicted positives [45]:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall refers to how many of the actual positives are captured [45]. It is calculated as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

F1 score combines precision and recall into one metric [45]:

$$\text{F1} = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The ROC curve (receiver operating characteristic curve) is a chart showing the trade-off between the true positive rate (TPR) and the false-positive rate (FPR). The PR curve (precision–recall curve) is a graph that combines precision and recall in a single visualization. The higher the area under the curve score, the better the performance of a model. However, accuracy, F1 score, ROC curve, and PR curve are not the best metrics for heavily imbalanced datasets, especially when you are more interested in the minority group. The MCC (Matthews correlation coefficient) is another alternative metric, which is calculated based on the Pearson correlation coefficient between actual and predicted values ranging from $[-1, 1]$ [41]. It is the method of choice for imbalanced datasets [41]:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{5}$$

Since some of our datasets are balanced and some are extremely imbalanced, a single metric may not reflect the model performance well. Therefore, we comprehensively evaluated our results based on the above metrics.

Author Contributions: D.H. conceived and supervised the study; Y.R. analyzed the data and drafted the manuscript; S.D., L.F. and J.F. collected the raw sequencing and antimicrobial resistance (AMR) data. O.S. pre-processed the sequencing data and clinical data. D.H., T.C. and A.G. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work is financially supported by the German Federal Ministry of Education and Research (BMBF) under grant number 031L0209B (Deep-iAMR).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets from our laboratory used in the current study are publicly available at https://github.com/YunxiaoRen/deep_transfer_learning_AMR (accessed on 15 October 2021). The public dataset information is publicly available at https://doi.org/10.1371/journal.pcbi.1006258.s010 (accessed on 15 October 2021).

Acknowledgments: We would like to thank Moradigaravand et al. for making their data publicly available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Prestinaci, F.; Pezzotti, P.; Pantosti, A. Antimicrobial Resistance: A Global Multifaceted Phenomenon. *Pathog. Glob. Health* **2015**, *109*, 309–318. [CrossRef] [PubMed]
2. WHO-Antimicrobial_Resistance_Whitepaper. Available online: https://www.who.int/docs/default-source/documents/no-time-to-wait-securing-the-future-from-drug-resistant-infections-en.pdf (accessed on 15 October 2021).
3. Boolchandani, M.; D'Souza, A.W.; Dantas, G. Sequencing-Based Methods and Resources to Study Antimicrobial Resistance. *Nat. Rev. Genet.* **2019**, *20*, 356–370. [CrossRef]
4. Macesic, N.; Polubriaginof, F.; Tatonetti, N.P. Machine Learning: Novel Bioinformatics Approaches for Combating Antimicrobial Resistance. *Curr. Opin. Infect. Dis.* **2017**, *30*, 511–517. [CrossRef] [PubMed]
5. Yang, J.H.; Wright, S.N.; Hamblin, M.; McCloskey, D.; Alcantar, M.A.; Schrübbers, L.; Lopatkin, A.J.; Satish, S.; Nili, A.; Palsson, B.O.; et al. A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action. *Cell* **2019**, *177*, 1649–1661.e9. [CrossRef] [PubMed]
6. Ren, Y.; Chakraborty, T.; Doijad, S.; Falgenhauer, L.; Falgenhauer, J.; Goesmann, A.; Hauschild, A.-C.; Schwengers, O.; Heider, D. Prediction of Antimicrobial Resistance Based on Whole-Genome Sequencing and Machine Learning. *Bioinformatics* **2021**, *38*, 325–334. [CrossRef]
7. Arango-Argoty, G.A.; Garner, E.; Pruden, A.; Heath, L.S.; Vikesland, P.; Zhang, L. DeepARG: A Deep Learning Approach for Predicting Antibiotic Resistance Genes from Metagenomic Data. *Microbiome* **2018**, *6*, 23. [CrossRef] [PubMed]
8. Yan, J.; Bhadra, P.; Li, A.; Sethiya, P.; Qin, L.; Tai, H.K.; Wong, K.H.; Siu, S.W.I. Deep-AmPEP30: Improve Short Antimicrobial Peptides Prediction with Deep Learning. *Mol. Ther. Nucleic Acids* **2020**, *20*, 882–894. [CrossRef]
9. Stokes, J.M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N.M.; MacNair, C.R.; French, S.; Carfrae, L.A.; Bloom-Ackerman, Z.; et al. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180*, 688–702.e13. [CrossRef]
10. Al-Stouhi, S.; Reddy, C.K. Transfer Learning for Class Imbalance Problems with Inadequate Data. *Knowl. Inf. Syst.* **2016**, *48*, 201–228. [CrossRef]
11. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. *Proc. IEEE* **2020**, *109*, 43–76. [CrossRef]
12. Chen, Y. A Transfer Learning Model with Multi-Source Domains for Biomedical Event Trigger Extraction. *BMC Genom.* **2021**, *22*, 31. [CrossRef]
13. Yu, J.; Deng, Y.; Liu, T.; Zhou, J.; Jia, X.; Xiao, T.; Zhou, S.; Li, J.; Guo, Y.; Wang, Y.; et al. Lymph Node Metastasis Prediction of Papillary Thyroid Carcinoma Based on Transfer Learning Radiomics. *Nat. Commun.* **2020**, *11*, 4807. [CrossRef]
14. Mahbod, A.; Schaefer, G.; Wang, C.; Dorffner, G.; Ecker, R.; Ellinger, I. Transfer Learning Using a Multi-Scale and Multi-Network Ensemble for Skin Lesion Classification. *Comput. Methods Programs Biomed.* **2020**, *193*, 105475. [CrossRef]
15. Farahani, A.; Pourshojae, B.; Rasheed, K.; Arabnia, H.R. A Concise Review of Transfer Learning. In Proceedings of the 2020 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 16–18 December 2020.
16. Radha, M.; Fonseca, P.; Moreau, A.; Ross, M.; Cerny, A.; Anderer, P.; Long, X.; Aarts, R.M. A Deep Transfer Learning Approach for Wearable Sleep Stage Classification with Photoplethysmography. *NPJ Digit. Med.* **2021**, *4*, 135. [CrossRef]
17. Mallesh, N.; Zhao, M.; Meintker, L.; Höllein, A.; Elsner, F.; Lüling, H.; Haferlach, T.; Kern, W.; Westermann, J.; Brossart, P.; et al. Knowledge Transfer to Enhance the Performance of Deep Learning Models for Automated Classification of B Cell Neoplasms. *Patterns* **2021**, *2*, 100351. [CrossRef] [PubMed]
18. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
19. Weiss, K.; Khoshgoftaar, T.M.; Wang, D. A Survey of Transfer Learning. *J. Big Data* **2016**, *3*, 9. [CrossRef]

20. Kopas, L.; Kusalik, A.; Schneider, D. Antimicrobial Resistance Prediction from Whole-Genome Sequence Data Using Transfer Learning. *F1000Research* **2019**, *8*, 1333. [CrossRef]
21. Ebbehoj, A.; Thunbo, M.Ø.; Andersen, O.E.; Glindtvad, M.V.; Hulman, A. Transfer Learning for Non-Image Data in Clinical Research: A Scoping Review. *PLoS Digit. Health* **2022**, *1*, e0000014. [CrossRef]
22. Liu, Z.; Jiang, M.; Luo, T. Leverage Electron Properties to Predict Phonon Properties via Transfer Learning for Semiconductors. *Sci. Adv.* **2020**, *6*, eabd1356. [CrossRef]
23. Plested, J.; Gedeon, T. Deep Transfer Learning for Image Classification: A Survey. *arXiv* **2022**, arXiv:2205.09904.
24. Li, X.; Grandvalet, Y.; Davoine, F.; Cheng, J.; Cui, Y.; Zhang, H.; Belongie, S.; Tsai, Y.-H.; Yang, M.-H. Transfer Learning in Computer Vision Tasks: Remember Where You Come From. *Image Vis. Comput.* **2020**, *93*, 103853. [CrossRef]
25. Gao, Y.; Mosalam, K.M. Deep Transfer Learning for Image-Based Structural Damage Recognition: Deep Transfer Learning for Image-Based Structural Damage Recognition. *Comput.-Aided Civ. Infrastruct. Eng.* **2018**, *33*, 748–768. [CrossRef]
26. Shao, L.; Zhu, F.; Li, X. Transfer Learning for Visual Categorization: A Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2015**, *26*, 1019–1034. [CrossRef] [PubMed]
27. Schwessinger, R.; Gosden, M.; Downes, D.; Brown, R.C.; Oudelaar, A.M.; Telenius, J.; Teh, Y.W.; Lunter, G.; Hughes, J.R. DeepC: Predicting 3D Genome Folding Using Megabase-Scale Transfer Learning. *Nat. Methods* **2020**, *17*, 1118–1124. [CrossRef]
28. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
29. Gupta, V.; Choudhary, K.; Tavazza, F.; Campbell, C.; Liao, W.; Choudhary, A.; Agrawal, A. Cross-Property Deep Transfer Learning Framework for Enhanced Predictive Analytics on Small Materials Data. *Nat. Commun.* **2021**, *12*, 6595. [CrossRef]
30. Park, Y.; Hauschild, A.-C.; Heider, D. Transfer Learning Compensates Limited Data, Batch-Effects, And Technical Heterogeneity In Single-Cell Sequencing. *bioRxiv* **2021**. [CrossRef]
31. Okerinde, A.; Shamir, L.; Hsu, W.; Theis, T.; Nafi, N. EGAN: Unsupervised Approach to Class Imbalance Using Transfer Learning. *arXiv* **2021**, arXiv:2104.04162.
32. Weiss, K.R.; Khoshgoftaar, T.M. Investigating Transfer Learners for Robustness to Domain Class Imbalance. In Proceedings of the 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 18–20 December 2016; pp. 207–213.
33. Minvielle, L.; Atiq, M.; Peignier, S.; Mougeot, M. Transfer Learning on Decision Tree with Class Imbalance. In Proceedings of the 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, 4–6 November 2019; pp. 1003–1010.
34. Krawczyk, B. Learning from Imbalanced Data: Open Challenges and Future Directions. *Prog. Artif. Intell.* **2016**, *5*, 221–232. [CrossRef]
35. Gao, Y.; Cui, Y. Deep Transfer Learning for Reducing Health Care Disparities Arising from Biomedical Data Inequality. *Nat. Commun.* **2020**, *11*, 5131. [CrossRef] [PubMed]
36. Falgenhauer, L.; Nordmann, P.; Imirzalioglu, C.; Yao, Y.; Falgenhauer, J.; Hauri, A.M.; Heinmüller, P.; Chakraborty, T. Cross-Border Emergence of Clonal Lineages of ST38 Escherichia Coli Producing the OXA-48-like Carbapenemase OXA-244 in Germany and Switzerland. *Int. J. Antimicrob. Agents* **2020**, *56*, 106157. [CrossRef] [PubMed]
37. Moradigaravand, D.; Palm, M.; Farewell, A.; Mustonen, V.; Warringer, J.; Parts, L. Prediction of Antibiotic Resistance in Escherichia Coli from Large-Scale Pan-Genome Data. *PLoS Comput. Biol.* **2018**, *14*, e1006258. [CrossRef] [PubMed]
38. Cai, C.; Wang, S.; Xu, Y.; Zhang, W.; Tang, K.; Ouyang, Q.; Lai, L.; Pei, J. Transfer Learning for Drug Discovery. *J. Med. Chem.* **2020**, *63*, 8683–8694. [CrossRef]
39. Chicco, D.; Jurman, G. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef]
40. Chicco, D.; Starovoitov, V.; Jurman, G. The Benefits of the Matthews Correlation Coefficient (MCC) Over the Diagnostic Odds Ratio (DOR) in Binary Classification Assessment. *IEEE Access* **2021**, *9*, 47112–47124. [CrossRef]
41. Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal Classifier for Imbalanced Data Using Matthews Correlation Coefficient Metric. *PLoS ONE* **2017**, *12*, e0177678. [CrossRef]
42. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor. *Bioinformatics* **2018**, *34*, i884–i890. [CrossRef]
43. Li, H.; Durbin, R. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef]
44. Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M.O.; Whitwham, A.; Keane, T.; McCarthy, S.A.; Davies, R.M.; et al. Twelve Years of SAMtools and BCFtools. *GigaScience* **2021**, *10*, giab008. [CrossRef]
45. Vakili, M.; Ghamsari, M.; Rezaei, M. Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification. *arXiv* **2020**, arXiv:2001.09636.

3.4 Study 4: Swarm Learning Predicts AMR (Unpublished)

3.4.1 Aim and motivation

With rising concerns over data security and privacy, numerous countries and institutions have implemented data privacy laws that restrict specific data sharing, especially in the medical field. However, this has also hindered, to some extent, the models' training with data size limitations. The emergence of federated learning (FL) overcomes this challenge by allowing collaborative training without compromising the privacy and security of individual datasets. Yet, model parameters are still managed by a central server, indicating a centralization of power. Thus, this research delves into the swarm learning (SL) approach, a groundbreaking decentralized machine learning method that combines edge computing and blockchain-enabled peer-to-peer networks. We employ SL on data from two distinct nodes to predict AMR for four drugs and evaluate its performance against both locally and centrally trained modes.

3.4.2 Results

SL for AMR identification against CIP

Initially, we compared various training modes' efficacy in predicting resistance against CIP on training and test datasets. The model's performance of MCC scores was charted within 50 epochs in local, centralized, and swarm modes (Figure 3.2A). A focused performance comparison on test data was also made after 10 epochs (Figure 3.2 B). The results showed that the swarm mode achieved effective prediction of resistance to CIP, as the median value of the MCC score exceeded 0.85. More importantly, the swarm mode significantly surpassed both local and centralized approaches, particularly enhancing local mode performance at node 2, with the p-value of 2.36e-6 compared to local_1, 1.03e-57 compared to local_2, and 1.87e-5 compared to centralized mode, respectively (Figure 3.2 B).

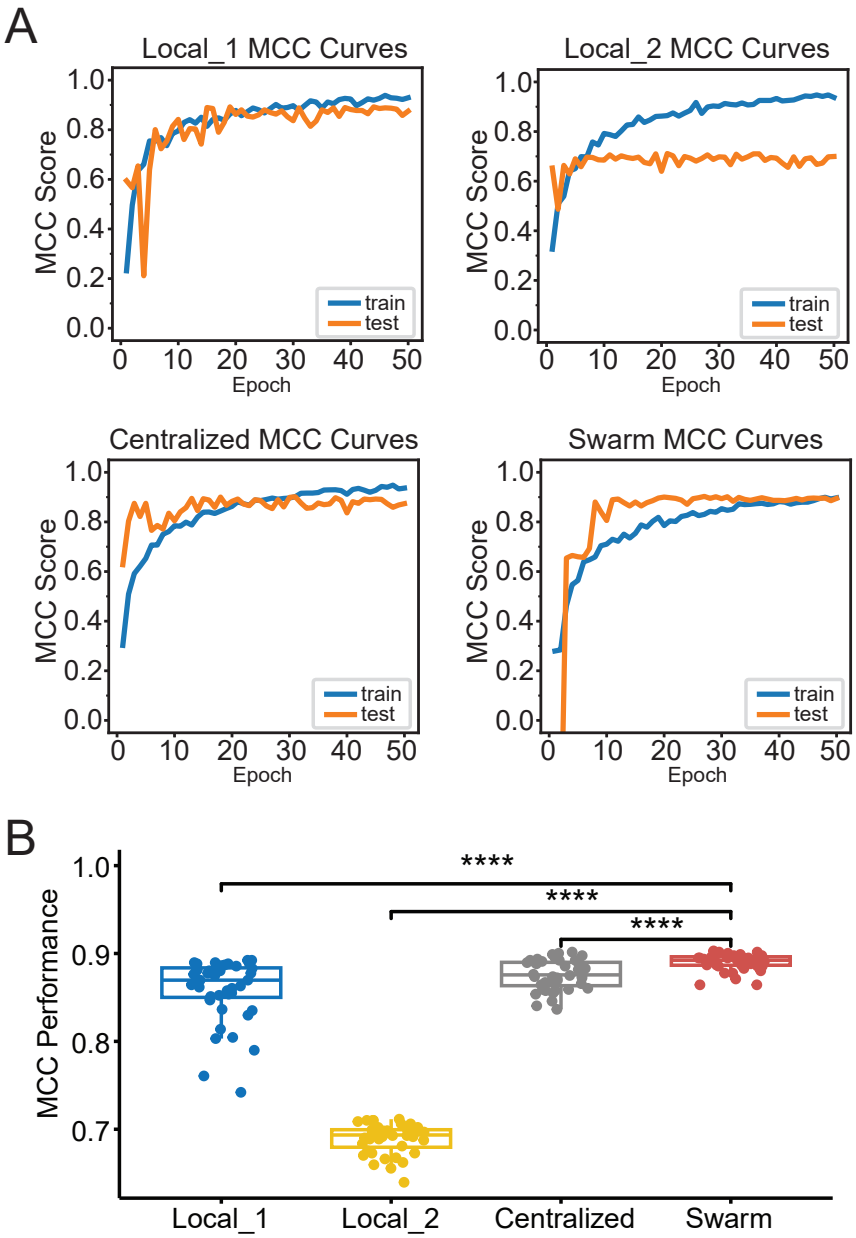

Figure 3.2: Performance for AMR identification against CIP. A: Performance curve plot about MCC scores within 50 epochs on training and test datasets. B: Boxplot about MCC scores after 10 epochs on the test dataset. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001, ns: no significance. The p-values were calculated by T-test.

SL for AMR identification against CTX

Subsequently, we evaluated the modes for CTX resistance predictions. Performance curves for MCC were detailed within 50 epochs for all modes (Figure 3.3 A), and boxplots post-10 epochs were highlighted in Figure 3.3 B. The analysis indicated that the centralized mode by integrating data at node 1 and node 2 didn't notably enhance model performance compared to the local mode at node 2. This implies that one of the local datasets might be of subpar quality, and simply amassing more data didn't improve the model performance. The swarm mode was superior to the centralized mode, with a p-value of 2.10e-9. However, when compared to the local model, the swarm mode enhanced the performance for node 1 but weakened it for node 2, highlighting the influence of data quality on both centralized and swarm mode outcomes.

SL for AMR identification against CTZ

For the CTZ resistance predictions, the performance curves for the MCC score are shown across 50 epochs for each mode (Figure 3.4 A), and the boxplots showed the performance after 10 epochs (Figure 3.4 B). We observed that the swarm mode significantly outperformed the centralized mode and local mode at node 1, with p-values of 7.26e-13 and 5.10e-20, respectively. However, at node 2, the swarm mode trailed behind the local mode, a trend consistent with the CTX results.

SL for AMR identification against GEN

In the final assessment, we focused on GEN resistance predictions. Performance across 50 epochs for each mode is shown in Figure 3.5 A, and the boxplots for each mode after 10 epochs are shown in Figure 3.5 B. Swarm mode emerged as superior to centralized and the local mode at node 2 concerning median values. However, the performance after 10 epochs in swarm mode showed fluctuating results.

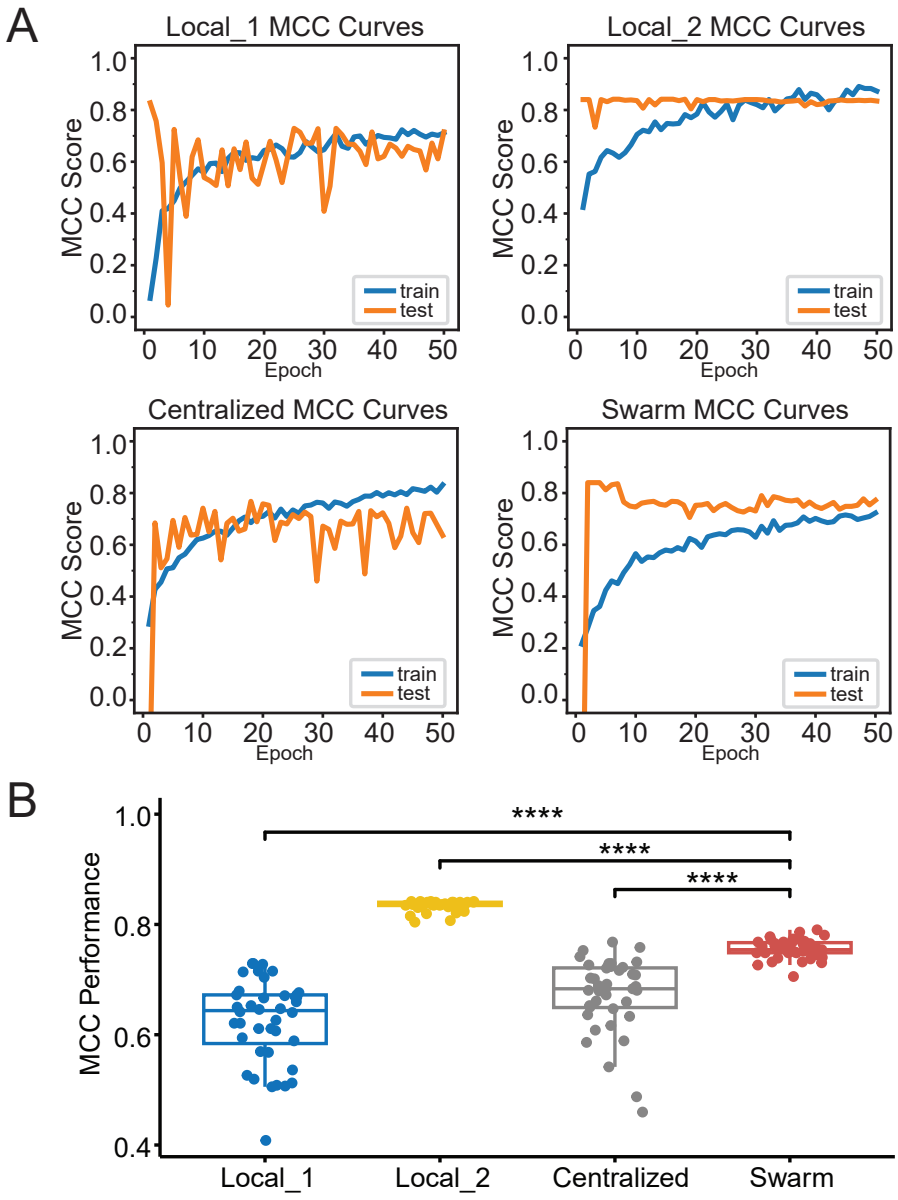

Figure 3.3: Performance for AMR identification against CTX. A: Performance curve plot about MCC scores within 50 epochs on training and test datasets. B: Boxplot about MCC scores after 10 epochs on the test dataset. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001, ns: no significance. The p-values were calculated by T-test.

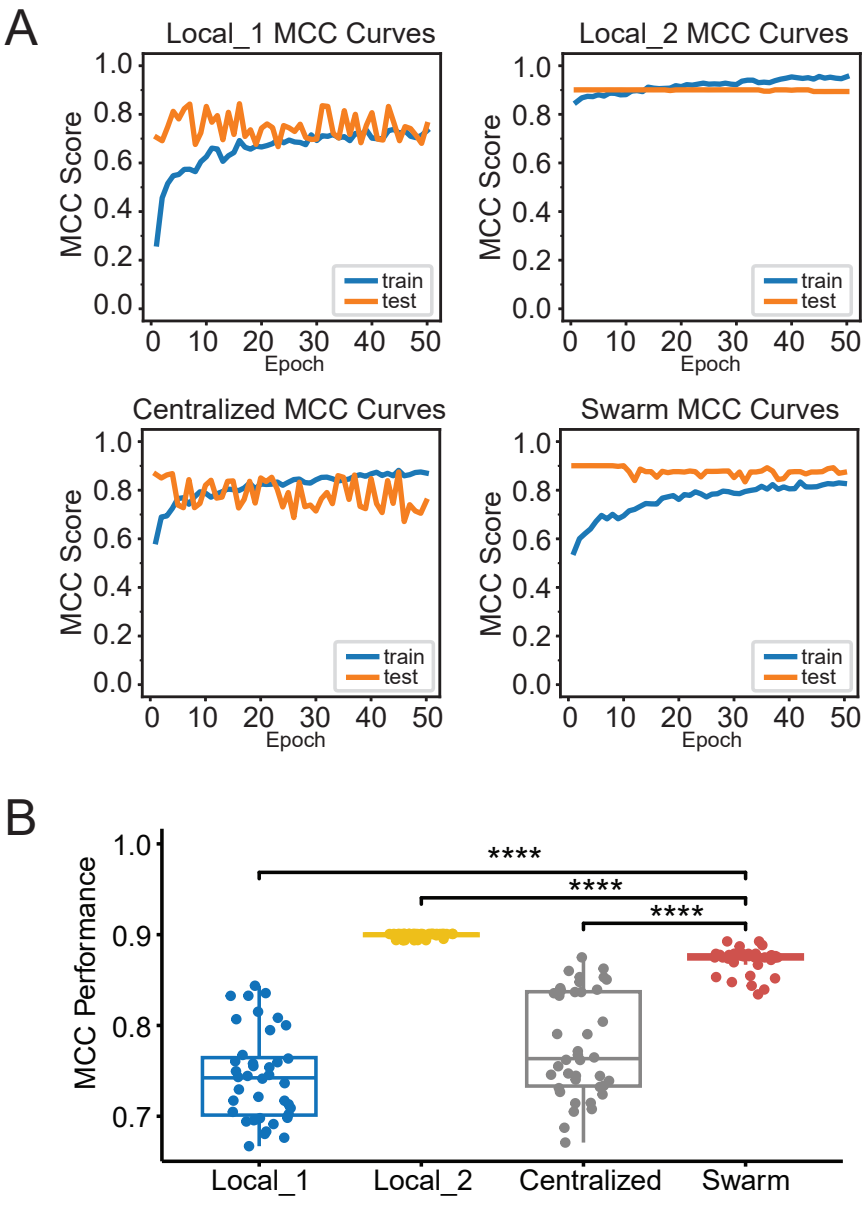

Figure 3.4: Performance for AMR identification against CTZ. A: Performance curve plot about MCC scores within 50 epochs on training and test datasets. B: Boxplot about MCC scores after 10 epochs on the test dataset. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001, ns: no significance. The p-values were calculated by T-test.

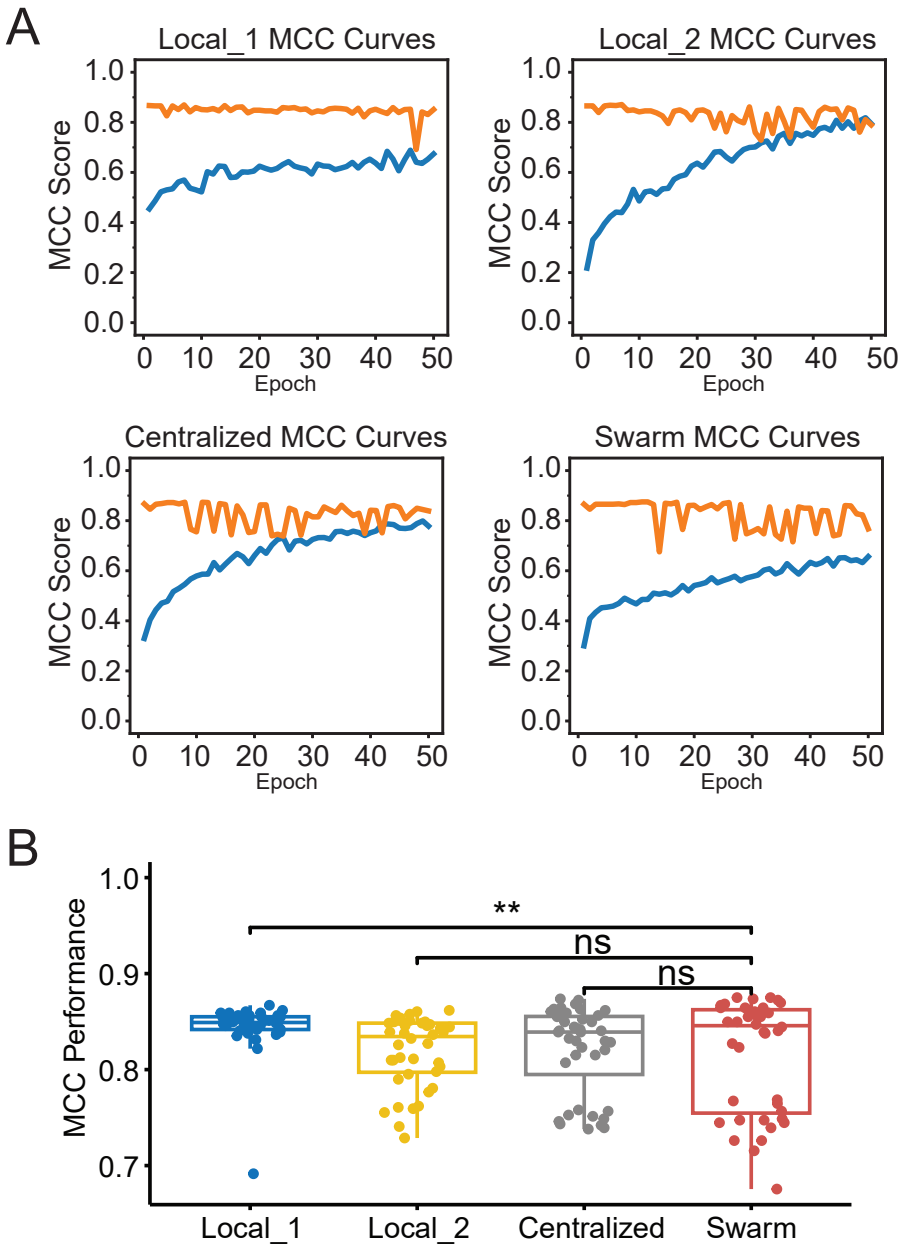

Figure 3.5: Performance for AMR identification against GEN. A: Performance curve plot about MCC scores within 50 epochs on training and test datasets. B: Boxplot about MCC scores after 10 epochs on test dataset. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001, ns: no significance. The p-values were calculated by T-test.

3.4.3 Discussion and conclusion

In this study, we employed Swarm Learning (SL) to predict resistance to four different drugs across two independent nodes. We conducted an extensive evaluation of model performance under three distinct modes: SL, centralized, and local, using independent test datasets. Our findings reveal that the swarm mode consistently outperforms the centralized mode. However, its relative performance compared to the local mode varies among different nodes.

The reason behind SL and centralized modes not achieving the same level of training performance as the specific local mode model is likely multifaceted. Firstly, it may be attributed to variations in data quality across nodes. Some nodes may possess high-quality data in sufficient quantities for robust model training, while others may have lower-quality data. The integration of data in the centralized mode, although it enhances the training performance of one node, might adversely impact the performance of another node.

Furthermore, an imbalance in label distribution within our test and training datasets could be contributing to these disparities in model performance. For instance, in the case of the CTZ data, the ratios of resistance (R) and sensitivity (S) labels in nodes 1, 2, and the test data are 34.1/65.9, 6.4/93.6, and 5.0/95.0, respectively. It's possible that the local training model in node 2 has a bias towards identifying sensitivity samples. Interestingly, the similarity between the label distribution in the test and training datasets may also influence the final local model performance, with potential implications from both the centralized and SL modes.

In light of these findings, for a more robust assessment of the performance of different training modes, we advocate the balancing of training data and the inclusion of a more diverse set of test data with varying label distribution ratios. This approach will enable a more comprehensive evaluation of the model's generalization capabilities across different operational modes.

In summary, our study highlights the complex interplay of data quality, quantity, label distribution, and the chosen mode of operation in predicting drug resistance, shedding light on the intricate factors affecting model performance across different nodes.

4 Discussion

In these studies, we have successfully developed efficient and precise models for predicting both AMR and MDR. Our innovative approach includes a deep transfer learning model that enhances prediction accuracy in the context of small and label-imbalanced samples. Notably, we have identified critical AMR-associated mutations and genes, setting a foundation for further exploration. However, there are some aspects that can continue to be improved in future studies.

4.1 Experimental Validation

Firstly, there is a need for more comprehensive experimental validation to substantiate the predictive results of our model. In the first work, we identified some genes associated with antibiotic resistance. Some of these genes have been well-studied, such as *marA*, which is related to multiple drug resistance (Abdolmaleki et al., 2019). While others remain less explored. For example, gene *nhaA*, associated with CTX, CTZ, and GEN resistance, displays Na^+/H^+ antiport activity in *E.coli* that may influence drug resistance by regulating permeability (Padan et al., 2004). Gene *rlmC* encodes a 23S RNA methyltransferase that methylates the 23S rRNA at antibiotic binding sites and thus may be related to antibiotic resistance (Pletnev et al., 2020; Stojković et al., 2016). Gene *fliI* is known to encode a virulence factor, with studies highlighting the correlation between antimicrobial resistance and bacterial virulence (Beceiro et al., 2013; Deng et al., 2019). *pepB* encodes peptidase B, linked to the production of

bacteriocins, which are narrow-spectrum antimicrobial peptides (Suzuki et al., 2001; Telhig et al., 2020). *MurB* is a key enzyme in the synthesis of peptidoglycan, a crucial component of the bacterial cell wall (Nasiri et al., 2017; Walsh and Wencewicz, 2014).

Although these findings contribute to a more comprehensive understanding of antibiotic resistance, additional in-depth experiments will strengthen the reliability of our findings.

4.2 Species Generalization

Our present model is specifically designed to analyze the resistance of *E. coli* to four targeted drugs. *E. coli*, a prominent bacterial pathogen, is frequently linked with hospital-acquired infections and AMR. It is part of the ESKAPE group of pathogens, an acronym representing six critical multidrug-resistant bacterial species including *Enterococcus faecalis*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacteriaceae* (Shankarnarayan et al., 2022). The focus on *E. coli* in our model reflects its significant role within this group, but future research could expand the model's scope to encompass other ESKAPE pathogens and more drugs, which can enhance the model's generalization capabilities and attain a more universally applicable model that could serve a wider array of needs in the medical field.

4.3 Feature Input: SNP and Beyond

Our constructed models focus on genome-wide variant information to identify secondary mutations that contribute to the resistance directly or indirectly, e.g. compensatory mutations. Mutation represents an inherent mechanism leading to AMR, yet the pathways to AMR are multifaceted and also encompass horizontal gene transfer (Lerminiaux and Cameron, 2019; Evans et al., 2020; Sun et al., 2019; Zhang et al., 2022b). Numerous studies have integrated metagenomic analyses of resistance gene distribution across various environments with expression abundance assessments to comprehend the health risks associated with ARGs (Antibiotic Resistance Genes) and their capability for HGT. For instance, Danko et al. (2021) created the first urban metagenome map, utilizing 4728 metagenomic samples from 60 urban public transportation systems. Analyzing the distribution and transmission of ARGs across global habitats is essential from a worldwide health standpoint, especially considering the transition from environmental compartments to humans.

In another study, Zhang et al. (2022b) conducted an extensive study across six unique habitats, analyzing 4572 samples at the metagenomic level. They identified 2561 ARG that jointly confer resistance to 24 antibiotic classes. The research further explored the prevalence, po-

tential for transmission, and expression characteristics of these ARGs within the pathogen, shedding light on the complex interplay of factors influencing antimicrobial resistance.

Therefore, future research should incorporate additional features to capture the complexity of AMR mechanisms, providing a richer and more accurate predictive model.

4.4 Software Development

We shared all the source code and data for our four topics, facilitating the possibility for interested researchers to repeat our process or apply it to their own data. But for clinicians and individuals without a computational background, there are still challenges to using our methods. So we envision developing a toolkit or web-based tool. This user-friendly interface would extend the reach of our methods and make it a valuable asset in the fight against drug resistance.

4.5 Focus on AMP

Conventional antibiotics are facing a growing challenge as drug-resistant strains continue to emerge, leading to a global health crisis. In response to this pressing need to combat AMR, researchers are increasingly focusing on antimicrobial peptides (AMPs) to develop innovative antibiotics. AMPs are small proteins found in a wide range of organisms, from bacteria to humans, that play a crucial role in the innate immune response, targeting pathogenic microorganisms including bacteria, fungi, viruses, and parasites (Huan et al., 2020; Lei et al., 2019; Brogden, 2005). AMPs exhibit unique structural attributes, allowing them to disrupt microbial cell membranes and perform multifaceted roles in host defense mechanisms. Their broad-spectrum activity and unconventional modes of action make AMPs particularly promising candidates in the discovery of novel antibiotics, including antiviral and antibacterial drugs (Mba and Nweze, 2022; Spohn et al., 2019).

For example, Ma et al. (2022) combined several natural language processing neural network models, including LSTM, Attention, and BERT, to identify candidate AMPs from human gut microbiome data, ultimately identifying 181 that showed antimicrobial activity.

Furthermore, the discovery of new AMPs is being revolutionized through the application of generative models (Das et al., 2021; Szymczak et al., 2023). Szymczak et al. (2023) introduced HydrAMP, a conditional variance autoencoder that skillfully learns a low-dimensional continuous representation of a peptide while simultaneously capturing its antimicrobial properties. The model separates the learned representation of a peptide from its antimicrobial

conditions and leverages the ingenuity of parameter control. Complemented by wet-lab validation, their approach yielded nine highly active peptides generated as analogs of clinically relevant prototypes, along with six analogs of an inactive peptide. HydrAMP's capability to spawn a diverse array of potent peptides represents a forward leap in the ongoing battle against the antimicrobial resistance crisis.

In summary, AMPs, with their distinctive characteristics and varied mechanisms of action, emerge as promising alternatives to traditional antibiotics. Their exploration and development through modern computational techniques herald a new era in the fight against AMR, offering hope for more effective treatments and interventions.

4.6 Concluding Remark

Overall, we have developed accurate AMR prediction models that serve as valuable tools for both AMR monitoring and clinical treatment. Our models have enabled us to identify crucial mutations and genes associated with AMR, providing a rich reference resource for further experimental and computational studies of AMR. Furthermore, we compared different multi-label classification methods, providing a novel approach for simultaneously identifying multiple drug resistance. In addition, our innovative approach includes a deep transfer learning model that enhances prediction accuracy with a limited number of samples and label imbalances. Moreover, we have also developed federated transfer learning, a strategy allowing different data owners to train models locally at their data stores. This method not only achieves precise prediction but also ensures the utmost data security and privacy.

In conclusion, our comprehensive approach to combating the challenge of AMR incorporates diverse machine learning algorithms. These address the specific needs and constraints of AMR prediction, including considerations for multiple drug resistance classification, constraints imposed by small sample sizes and label imbalances, and the imperatives of data privacy and security.

List of Figures

1.1 **History of antibiotic development and observed time of antibiotic resistance.** The year each antibiotic was discovered is shown above the timeline, and the year resistance to each antibiotic was identified is indicated below the timeline. 3

1.2 **Genetic, biochemical, and physical mechanisms of antibiotic resistance.** The diagram on the left shows the genetic mechanisms that lead to bacteria acquiring antibiotic resistance (Ab^r), which include both gene mutations and horizontal gene transfer. The latter involves the acquisition of resistance genes through plasmids and conjugative transposons (conjugation), and by bacteriophage (transduction), as well as the integration of foreign free DNA into the bacterial chromosome (transformation). The diagram on the right shows biochemical mechanisms and physical mechanisms, where S represents Susceptible, R represents Resistant. This figure was adapted from Alekshun and Levy (2007) and Boolchandani et al. (2019), which was created with BioRender. 5

1.3 **Overview of machine learning workflow and project design.** This figure was created by BioRender.com. 8

3.1 **Workflow of this study.** This figure was created by BioRender.com 46

3.2 **Performance for AMR identification against CIP.** A: Performance curve plot about MCC scores within 50 epochs on training and test datasets. B: Boxplot about MCC scores after 10 epochs on the test dataset. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001, ns: no significance. The p-values were calculated by T-test. 68

3.3 **Performance for AMR identification against CTX.** A: Performance curve plot about MCC scores within 50 epochs on training and test datasets. B: Boxplot about MCC scores after 10 epochs on the test dataset. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001, ns: no significance. The p-values were calculated by T-test. 70

3.4 **Performance for AMR identification against CTZ.** A: Performance curve plot about MCC scores within 50 epochs on training and test datasets. B: Boxplot about MCC scores after 10 epochs on the test dataset. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001, ns: no significance. The p-values were calculated by T-test. 71
3.5 **Performance for AMR identification against GEN.** A: Performance curve plot about MCC scores within 50 epochs on training and test datasets. B: Boxplot about MCC scores after 10 epochs on test dataset. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001, ns: no significance. The p-values were calculated by T-test. 72

List of Acronyms and Abbreviations

AI	Artificial Intelligence
AMR	Antimicrobial Resistance
ARDB	Antibiotic Resistance Genes Database
ARG	Antibiotic Resistance Gene
ARGANNOT	Active Antibiotic Resistance Gene Annotation
AST	Antimicrobial Susceptibility Testing
AUC	Area Under Curve
AMP	Antimicrobial Peptides
BR	Binary Relevance
CARD	Comprehensive Antibiotic Resistance Database
CC	Classifier chain
CGR	Chaos Game Representation
CIP	Ciprofloxacin
CNN	Convolutional Neural Network
CT	Computed Tomography
CTX	Cefotaxime
CTZ	Ceftazidime
CLSI	Clinical and Laboratory Standards Institute
COVID	Corona Virus Disease
DP	Differential Privacy
ECC	Ensemble Classifier Chains
E. coli	*Escherichia coli*
EUCAST	European Committee on Antimicrobial Susceptibility Testing

FCGR	Frequency Matrix Chaos Game Representation
FL	Federated Learning
FTL	ederated Transfer Learning
FP	False Positives
FPR	False Positives Rate
FN	False Negatives
GAN	Generative Adversarial Network
GEN	Gentamicin
GNN	Graph Neural Network
HE	Homomorphic Encryption
HFL	Horizontal Federated Learning
LP	Label Powerset
LR	Logistic Regression
LSTM	Long Short-Term Memory
MCC	Matthews Correlation Coefficient
MDR	Multi-Drug Resistance
MDR-TB	Multi-Drug Resistance Tuberculosis
ML	Machine Learning
MLC	Multi-Label Classification
MPC	Multi-Party Computation
MRSA	methicillin-resistant *Staphylococcus aureus*
PCA	Principal Component Analysis
RD	Random Label Space partitioning with Label Powerset
RF	Random Forest
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SL	Swarm Learning
SNP	Single Nucleotide Polymorphisms
SVM	Support Vector Machine
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2

TL	Transfer Learning
TP	True Positives
TN	True Negatives
TPR	True Positive Rate
VFL	Vertical Federated Learning
WGS	Whole-Genome Sequencing
WHO	World Health Organization
XDR-TB	Extensively Drug-Resistant Tuberculosis

Bibliography

Zohreh Abdolmaleki, Zohreh Mashak, and Farhad Safarpoor Dehkordi. Phenotypic and genotypic characterization of antibiotic resistance in the methicillin-resistant Staphylococcus aureus strains isolated from hospital cockroaches. *Antimicrobial Resistance and Infection Control*, 8(1):54, 2019. ISSN 2047-2994. doi: 10.1186/s13756-019-0505-7.

Alankrita Aggarwal, Mamta Mittal, and Gopi Battineni. Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*, 1(1):100004, 2021. ISSN 2667-0968. doi: 10.1016/j.jjimei.2020.100004.

Samir Al-Stouhi and Chandan K. Reddy. Transfer learning for class imbalance problems with inadequate data. *Knowl Inf Syst*, 48(1):201–228, 2016. ISSN 0219-1377, 0219-3116. doi: 10.1007/s10115-015-0870-3.

Michael N. Alekshun and Stuart B. Levy. Molecular Mechanisms of Antibacterial Multidrug Resistance. *Cell*, 128(6):1037–1050, 2007. ISSN 00928674. doi: 10.1016/j.cell.2007.03.004.

J. S. Almeida, J. A. Carrico, A. Maretzek, P. A. Noble, and M. Fletcher. Analysis of genomic sequences by Chaos Game Representation. *Bioinformatics*, 17(5):429–437, 2001. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/17.5.429.

Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J. Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1):53, 2021. ISSN 2196-1115. doi: 10.1186/s40537-021-00444-8.

Rustam I. Aminov. A Brief History of the Antibiotic Era: Lessons Learned and Challenges for the Future. *Front Microbiol*, 1:134, 2010. ISSN 1664-302X. doi: 10.3389/fmicb.2010.00134.

G. A. 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. Preprint, Bioinformatics, 2017.

Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. A Brief Survey of Deep Reinforcement Learning. *IEEE Signal Process. Mag.*, 34(6):26–38, 2017. ISSN 1053-5888. doi: 10.1109/MSP.2017.2743240.

Xiang Bai, Hanchen Wang, Liya Ma, Yongchao Xu, Jiefeng Gan, Ziwei Fan, Fan Yang, Ke Ma, Jiehua Yang, Song Bai, Chang Shu, Xinyu Zou, Renhao Huang, Changzheng Zhang, Xiaowu Liu, Dandan Tu, Chuou Xu, Wenqing Zhang, Xi Wang, Anguo Chen, Yu Zeng, Dehua Yang, Ming-Wei Wang, Nagaraj Holalkere, Neil J. Halin, Ihab R. Kamel, Jia Wu, Xuehua Peng, Xiang Wang, Jianbo Shao, Pattanasak Mongkolwat, Jianjun Zhang, Weiyang Liu, Michael Roberts, Zhongzhao Teng, Lucian Beer, Lorena E. Sanchez, Evis Sala, Daniel L. Rubin, Adrian Weller, Joan Lasenby, Chuansheng Zheng, Jianming Wang, Zhen Li, Carola Schönlieb, and Tian Xia. Advancing COVID-19 diagnosis with privacy-preserving collaboration in artificial intelligence. *Nat Mach Intell*, 3(12):1081–1089, 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00421-z.

Syreen Banabilah, Moayad Aloqaily, Eitaa Alsayed, Nida Malik, and Yaser Jararweh. Federated learning review: Fundamentals, enabling technologies, and future applications. *Information Processing and Management*, 59(6):103061, 2022. ISSN 0306-4573. doi: 10.1016/j.ipm.2022.103061.

Mary Barber and Mary Rozwadowska-Dowzenko. INFECTION BY PENICILLIN-RESISTANT STAPHYLOCOCCI. *The Lancet*, 252(6530):641–644, 1948. ISSN 0140-6736. doi: 10.1016/S0140-6736(48)92166-7.

Alejandro Beceiro, María Tomás, and Germán Bou. Antimicrobial Resistance and Virulence: A Successful or Deleterious Association in the Bacterial World? *Clin Microbiol Rev*, 26(2):185–230, 2013. ISSN 0893-8512. doi: 10.1128/CMR.00059-12.

Jessica M. A. Blair, Mark A. Webber, Alison J. Baylay, David O. Ogbolu, and Laura J. V. Piddock. Molecular mechanisms of antibiotic resistance. *Nat Rev Microbiol*, 13(1):42–51, 2015. ISSN 1740-1534. doi: 10.1038/nrmicro3380.

Jasmin Bogatinovski, Ljupčo Todorovski, Sašo Džeroski, and Dragi Kocev. Comprehensive comparative study of multi-label classification methods. *Expert Systems with Applications*, 203:117215, 2022. ISSN 0957-4174. doi: 10.1016/j.eswa.2022.117215.

Manish Boolchandani, Alaric W. D'Souza, and Gautam Dantas. Sequencing-based methods and resources to study antimicrobial resistance. *Nat Rev Genet*, 20(6):356–370, 2019. ISSN 1471-0056, 1471-0064. doi: 10.1038/s41576-019-0108-4.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh Pennsylvania USA, 1992. ISBN 978-0-89791-497-0. doi: 10.1145/130385.130401.

Matthew Botvinick, Sam Ritter, Jane X. Wang, Zeb Kurth-Nelson, Charles Blundell, and Demis Hassabis. Reinforcement Learning, Fast and Slow. *Trends in Cognitive Sciences*, 23 (5):408–422, 2019. ISSN 1364-6613. doi: 10.1016/j.tics.2019.02.006.

Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS ONE*, 12(6):e0177678, 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0177678.

Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324.

Kim A. Brogden. Antimicrobial peptides: Pore formers or metabolic inhibitors in bacteria? *Nat Rev Microbiol*, 3(3):238–250, 2005. ISSN 1740-1534. doi: 10.1038/nrmicro1098.

Katrina Browne, Sudip Chakraborty, Renxun Chen, Mark DP Willcox, David StClair Black, William R. Walsh, and Naresh Kumar. A New Era of Antibiotics: The Clinical Potential of Antimicrobial Peptides. *International Journal of Molecular Sciences*, 21(19): 7047, 2020. ISSN 1422-0067. doi: 10.3390/ijms21197047.

Nadia Burkart and Marco F. Huber. A Survey on the Explainability of Supervised Machine Learning. *jair*, 70:245–317, 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12228.

Yasemin Cag, Hulya Caskurlu, Yanyan Fan, Bin Cao, and Haluk Vahaboglu. Resistance mechanisms. *Ann. Transl. Med.*, 4(17):326–326, 2016. ISSN 23055839, 23055847. doi: 10.21037/atm.2016.09.14.

Chenjing Cai, Shiwei Wang, Youjun Xu, Weilin Zhang, Ke Tang, Qi Ouyang, Luhua Lai, and Jianfeng Pei. Transfer Learning for Drug Discovery. *J. Med. Chem.*, 63(16):8683–8694, 2020. ISSN 0022-2623, 1520-4804. doi: 10.1021/acs.jmedchem.9b02147.

Gürol Canbek, Tugba Taskaya Temizel, and Seref Sagiroglu. PToPI: A Comprehensive Review, Analysis, and Knowledge Representation of Binary Classification Performance Measures/Metrics. *SN COMPUT. SCI.*, 4(1):13, 2022. ISSN 2661-8907. doi: 10.1007/s42979-022-01409-1.

Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8):832, 2019. ISSN 2079-9292. doi: 10.3390/electronics8080832.

Yesim Cetinkaya, Pamela Falk, and C. Glen Mayhall. Vancomycin-Resistant Enterococci. *Clin Microbiol Rev*, 13(4):686–707, 2000. ISSN 0893-8512.

Hui-Ling Chen, Bo Yang, Gang Wang, Su-Jing Wang, Jie Liu, and Da-You Liu. Support vector machine based diagnostic system for breast cancer using swarm intelligence. *J Med Syst*, 36(4):2505–2519, 2012. ISSN 0148-5598. doi: 10.1007/s10916-011-9723-0.

Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890, 2018. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bty560.

Yifei Chen. A transfer learning model with multi-source domains for biomedical event trigger extraction. *BMC Genomics*, 22(1):31, 2021. ISSN 1471-2164. doi: 10.1186/s12864-020-07315-1.

Zhen Chen, Pei Zhao, Fuyi Li, Tatiana T Marquez-Lago, André Leier, Jerico Revote, Yan Zhu, David R Powell, Tatsuya Akutsu, Geoffrey I Webb, Kuo-Chen Chou, A Ian Smith, Roger J Daly, Jian Li, and Jiangning Song. iLearn: An integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Briefings in Bioinformatics*, 21(3):1047–1057, 2020. ISSN 1477-4054. doi: 10.1093/bib/bbz041.

Eirini Christaki, Markella Marcou, and Andreas Tofarides. Antimicrobial Resistance in Bacteria: Mechanisms, Evolution, and Persistence. *J Mol Evol*, 88(1):26–40, 2020. ISSN 1432-1432. doi: 10.1007/s00239-019-09914-3.

Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Xiangyi Lu, and Douglas M Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, 6(2):80–92, 2012. ISSN 1933-6942. doi: 10.4161/fly.19695.

Anne E. Clatworthy, Emily Pierson, and Deborah T. Hung. Targeting virulence: A new paradigm for antimicrobial therapy. *Nat Chem Biol*, 3(9):541–548, 2007. ISSN 1552-4469. doi: 10.1038/nchembio.2007.24.

P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*, 27(15): 2156–2158, 2011. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btr330.

Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, and Heng Li. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2):giab008, 2021. ISSN 2047-217X. doi: 10.1093/gigascience/giab008.

David Danko, Daniela Bezdan, Evan E. Afshin, Sofia Ahsanuddin, Chandrima Bhattacharya, Daniel J. Butler, Kern Rei Chng, Daisy Donnellan, Jochen Hecht, Katelyn Jackson, Katerina Kuchin, Mikhail Karasikov, Abigail Lyons, Lauren Mak, Dmitry Meleshko, Harun Mustafa, Beth Mutai, Russell Y. Neches, Amanda Ng, Olga Nikolayeva, Tatyana Nikolayeva, Eileen Png, Krista A. Ryon, Jorge L. Sanchez, Heba Shaaban, Maria A. Sierra, Dominique Thomas, Ben Young, Omar O. Abudayyeh, Josue Alicea, Malay Bhattacharyya, Ran Blekhman, Eduardo Castro-Nallar, Ana M. Cañas, Aspassia D. Chatziefthimiou, Robert W. Crawford, Francesca De Filippis, Youping Deng, Christelle Desnues, Emmanuel Dias-Neto, Marius Dybwad, Eran Elhaik, Danilo Ercolini, Alina Frolova, Dennis Gankin, Jonathan S. Gootenberg, Alexandra B. Graf, David C. Green, Iman Hajirasouliha, Jaden J. A. Hastings, Mark Hernandez, Gregorio Iraola, Soojin Jang, Andre Kahles, Frank J. Kelly, Kaymisha Knights, Nikos C. Kyrpides, Paweł P. Łabaj, Patrick K. H. Lee, Marcus H. Y. Leung, Per O. Ljungdahl, Gabriella Mason-Buck, Ken McGrath, Cem Meydan, Emmanuel F. Mongodin, Milton Ozorio Moraes, Niranjan Nagarajan, Marina Nieto-Caballero, Houtan Noushmehr, Manuela Oliveira, Stephan Ossowski, Olayinka O. Osuolale, Orhan Özcan, David Paez-Espino, Nicolás Rascovan, Hugues Richard, Gunnar Rätsch, Lynn M. Schriml, Torsten Semmler, Osman U. Sezerman, Leming Shi, Tieliu Shi, Rania Siam, Le Huu Song, Haruo Suzuki, Denise Syndercombe Court, Scott W. Tighe, Xinzhao Tong, Klas I. Udekwu, Juan A. Ugalde, Brandon Valentine, Dimitar I. Vassilev, Elena M. Vayndorf, Thirumalaisamy P. Velavan, Jun Wu, María M. Zambrano, Jifeng Zhu, Sibo Zhu, Christopher E. Mason, and International MetaSUB Consortium. A global metagenomic map of urban microbiomes and antimicrobial resistance. *Cell*, 184(13):3376–3393.e17, 2021. ISSN 1097-4172. doi: 10.1016/j.cell.2021.05.002.

Elizabeth M. Darby, Eleftheria Trampari, Pauline Siasat, Maria Solsona Gaya, Ilyas Alav, Mark A. Webber, and Jessica M. A. Blair. Molecular mechanisms of antibiotic resistance revisited. *Nat Rev Microbiol*, 21(5):280–295, 2023. ISSN 1740-1534. doi: 10.1038/s41579-022-00820-y.

Payel Das, Tom Sercu, Kahini Wadhawan, Inkit Padhi, Sebastian Gehrmann, Flaviu Cipcigan, Vijil Chenthamarakshan, Hendrik Strobelt, Cicero dos Santos, Pin-Yu Chen, Yi Yan Yang, Jeremy P. K. Tan, James Hedrick, Jason Crain, and Aleksandra Mojsilovic. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat Biomed Eng*, 5(6):613–623, 2021. ISSN 2157-846X. doi: 10.1038/s41551-021-00689-x.

K. Dasaradharami Reddy and Thippa Reddy Gadekallu. A Comprehensive Survey on Federated Learning Techniques for Healthcare Informatics. *Comput Intell Neurosci*, 2023: 8393990, 2023. ISSN 1687-5265. doi: 10.1155/2023/8393990.

Ittai Dayan, Holger R. Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Z. Abidin, Andrew Liu, Anthony Beardsworth Costa, Bradford J. Wood, Chien-Sung Tsai, Chih-Hung Wang, Chun-Nan Hsu, C. K. Lee, Peiying Ruan, Daguang Xu, Dufan Wu, Eddie Huang, Felipe Campos Kitamura, Griffin Lacey, Gustavo César de Antônio Corradi, Gustavo Nino, Hao-Hsin Shin, Hirofumi Obinata, Hui Ren, Jason C. Crane, Jesse Tetreault, Jiahui Guan, John W. Garrett, Joshua D. Kaggie, Jung Gil Park, Keith Dreyer, Krishna Juluru, Kristopher Kersten, Marcio Aloisio Bezerra Cavalcanti Rockenbach, Marius George Linguraru, Masoom A. Haider, Meena AbdelMaseeh, Nicola Rieke, Pablo F. Damasceno, Pedro Mario Cruz e Silva, Pochuan Wang, Sheng Xu, Shuichi Kawano, Sira Sriswasdi, Soo Young Park, Thomas M. Grist, Varun Buch, Watsamon Jantarabenjakul, Weichung Wang, Won Young Tak, Xiang Li, Xihong Lin, Young Joon Kwon, Abood Quraini, Andrew Feng, Andrew N. Priest, Baris Turkbey, Benjamin Glicksberg, Bernardo Bizzo, Byung Seok Kim, Carlos Tor-Díez, Chia-Cheng Lee, Chia-Jung Hsu, Chin Lin, Chiu-Ling Lai, Christopher P. Hess, Colin Compas, Deepeksha Bhatia, Eric K. Oermann, Evan Leibovitz, Hisashi Sasaki, Hitoshi Mori, Isaac Yang, Jae Ho Sohn, Krishna Nand Keshava Murthy, Li-Chen Fu, Matheus Ribeiro Furtado de Mendonça, Mike Fralick, Min Kyu Kang, Mohammad Adil, Natalie Gangai, Peerapon Vateekul, Pierre Elnajjar, Sarah Hickman, Sharmila Majumdar, Shelley L. McLeod, Sheridan Reed, Stefan Gräf, Stephanie Harmon, Tatsuya Kodama, Thanyawee Puthanakit, Tony Mazzulli, Vitor Lima de Lavor, Yothin Rakvongthai, Yu Rim Lee, Yuhong Wen, Fiona J. Gilbert, Mona G. Flores, and Quanzheng Li. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat Med*, 27(10):1735–1743, 2021. ISSN 1546-170X. doi: 10.1038/s41591-021-01506-3.

Olga V. Demler, Michael J. Pencina, and Ralph B. D'Agostino. Misuse of DeLong test to compare AUCs for nested models. *Statist. Med.*, 31(23):2577–2587, 2012. ISSN 02776715. doi: 10.1002/sim.5328.

Yiqin Deng, Haidong Xu, Youlu Su, Songlin Liu, Liwen Xu, Zhixun Guo, Jinjun Wu, Changhong Cheng, and Juan Feng. Horizontal gene transfer contributes to virulence and antibiotic resistance of Vibrio harveyi 345 based on complete genome sequence analysis. *BMC Genomics*, 20(1):761, 2019. ISSN 1471-2164. doi: 10.1186/s12864-019-6137-8.

Pedro Domingos. A few useful things to know about machine learning. *Commun. ACM*, 55(10):78–87, 2012. ISSN 0001-0782, 1557-7317. doi: 10.1145/2347736.2347755.

Qi Dou, Tiffany Y. So, Meirui Jiang, Quande Liu, Varut Vardhanabhuti, Georgios Kaissis, Zeju Li, Weixin Si, Heather H. C. Lee, Kevin Yu, Zuxin Feng, Li Dong, Egon Burian, Friederike Jungmann, Rickmer Braren, Marcus Makowski, Bernhard Kainz, Daniel Rueckert, Ben Glocker, Simon C. H. Yu, and Pheng Ann Heng. Federated deep learning for detecting COVID-19 lung abnormalities in CT: A privacy-preserving multinational validation study. *npj Digit. Med.*, 4(1):60, 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00431-6.

Andreas Ebbehoj, Mette Østergaard Thunbo, Ole Emil Andersen, Michala Vilstrup Glindtvad, and Adam Hulman. Transfer learning for non-image data in clinical research: A scoping review. *PLOS Digit Health*, 1(2):e0000014, 2022. ISSN 2767-3170. doi: 10.1371/journal.pdig.0000014.

Vamsidhar Enireddy, C. Karthikeyan, and D. Vijendra Babu. OneHotEncoding and LSTM-based deep learning models for protein secondary structure prediction. *Soft Comput*, 26(8): 3825–3836, 2022. ISSN 1433-7479. doi: 10.1007/s00500-022-06783-9.

Daniel R Evans, Marissa P Griffith, Alexander J Sundermann, Kathleen A Shutt, Melissa I Saul, Mustapha M Mustapha, Jane W Marsh, Vaughn S Cooper, Lee H Harrison, and Daria Van Tyne. Systematic detection of horizontal gene transfer across genera among multidrug-resistant bacteria in a single hospital. *eLife*, 9:e53886, 2020. ISSN 2050-084X. doi: 10.7554/eLife.53886.

Abolfazl Farahani, Behrouz Pourshojae, Khaled Rasheed, and Hamid R. Arabnia. A Concise Review of Transfer Learning, 2021.

Faiza Farhat, Md Tanwir Athar, Sultan Ahmad, Dag Øivind Madsen, and Shahab Saquib Sohail. Antimicrobial resistance and machine learning: Past, present, and future. *Frontiers in Microbiology*, 14, 2023. ISSN 1664-302X. doi: 10.3389/fmicb.2023.1179312.

Christopher Fletez-Brant, Dongwon Lee, Andrew S. McCallion, and Michael A. Beer. Kmer-SVM: A web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic Acids Research*, 41(W1):W544–W556, 2013. ISSN 0305-1048. doi: 10.1093/nar/gkt519.

T. R. Frieden, L. F. Sherman, K. L. Maw, P. I. Fujiwara, J. T. Crawford, B. Nivin, V. Sharp, D. Hewlett, K. Brudney, D. Alland, and B. N. Kreisworth. A multi-institutional outbreak of highly drug-resistant tuberculosis: Epidemiology and clinical outcomes. *JAMA*, 276 (15):1229–1235, 1996. ISSN 0098-7484.

Yan Gao and Yan Cui. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. *Nat Commun*, 11(1):5131, 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-18918-3.

Yuqing Gao and Khalid M. Mosalam. Deep Transfer Learning for Image-Based Structural Damage Recognition: Deep transfer learning for image-based structural damage recognition. *Computer-Aided Civil and Infrastructure Engineering*, 33(9):748–768, 2018. ISSN 10939687. doi: 10.1111/mice.12363.

Robert Gaynes. The Discovery of Penicillin—New Insights After More Than 75 Years of Clinical Use. *Emerg Infect Dis*, 23(5):849–853, 2017. ISSN 1080-6040. doi: 10.3201/eid2305.161556.

Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for Multi-Class Classification: An Overview, 2020.

Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications. *IEEE Transactions on Knowledge and Data Engineering*, 35(04):3313–3332, 2023. ISSN 1041-4347. doi: 10.1109/TKDE.2021.3130191.

Hemalatha Gunasekaran, K. Ramalakshmi, A. Rex Macedo Arokiaraj, S. Deepa Kanmani, Chandran Venkatesan, and C. Suresh Gnana Dhas. Analysis of DNA Sequence Classification Using CNN and Hybrid Models. *Computational and Mathematical Methods in Medicine*, 2021:e1835056, 2021. ISSN 1748-670X. doi: 10.1155/2021/1835056.

Vishu Gupta, Kamal Choudhary, Francesca Tavazza, Carelyn Campbell, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data. *Nat Commun*, 12(1):6595, 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-26921-5.

Dominik Heider, Robin Senge, Weiwei Cheng, and Eyke Hüllermeier. Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction. *Bioinformatics*, 29(16):1946–1952, 2013. ISSN 1460-2059, 1367-4803. doi: 10.1093/bioinformatics/btt331.

Tung Hoang, Changchuan Yin, and Stephen S.-T. Yau. Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison. *Genomics*, 108(3):134–142, 2016. ISSN 0888-7543. doi: 10.1016/j.ygeno.2016.08.002.

Yuchen Huan, Qing Kong, Haijin Mou, and Huaxi Yi. Antimicrobial Peptides: Classification, Design, Application and Research Progress in Multiple Fields. *Frontiers in Microbiology*, 11, 2020. ISSN 1664-302X. doi: 10.3389/fmicb.2020.582779.

Matthew I Hutchings, Andrew W Truman, and Barrie Wilkinson. Antibiotics: Past, present and future. *Current Opinion in Microbiology*, 51:72–80, 2019. ISSN 13695274. doi: 10.1016/j.mib.2019.10.008.

H.Joel Jeffrey. Chaos game representation of gene structure. *NAR*, 18:8, 1990. doi: 10.1093/nar/18.8.2163.

M. Patricia Jevons. "Celbenin" - resistant Staphylococci. *Br Med J*, 1(5219):124–125, 1961. ISSN 0007-1447. doi: 10.1136/bmj.1.5219.124-a.

Ce Ju, Dashan Gao, Ravikiran Mane, Ben Tan, Yang Liu, and Cuntai Guan. Federated Transfer Learning for EEG Signal Classification. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3040–3045, 2020. doi: 10.1109/EMBC44109.2020.9175344.

Georgios A. Kaissis, Marcus R. Makowski, Daniel Rückert, and Rickmer F. Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat Mach Intell*, 2 (6):305–311, 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-0186-1.

Adrian Kania and Krzysztof Sarapata. The robustness of the chaos game representation to mutations and its application in free-alignment methods. *Genomics*, 113(3):1428–1437, 2021. ISSN 0888-7543. doi: 10.1016/j.ygeno.2021.03.015.

Jee In Kim, Finlay Maguire, Kara K. Tsang, Theodore Gouliouris, Sharon J. Peacock, Tim A. McAllister, Andrew G. McArthur, and Robert G. Beiko. Machine Learning for Antimicrobial Resistance Prediction: Current Practice, Limitations, and Clinical Perspective. *Clin Microbiol Rev*, 35(3):e0017921, 2022. ISSN 1098-6618. doi: 10.1128/cmr.00179-21.

Claudio U. Köser, Matthew J. Ellington, and Sharon J. Peacock. Whole-genome sequencing to control antimicrobial resistance. *Trends Genet*, 30(9):401–407, 2014. ISSN 0168-9525. doi: 10.1016/j.tig.2014.07.003.

Samaneh Kouchaki, Yang Yang, Timothy M Walker, A Sarah Walker, Daniel J Wilson, Timothy E A Peto, Derrick W Crook, CRyPTIC Consortium, and David A Clifton. Application of machine learning techniques to tuberculosis drug resistance analysis. *Bioinformatics*, 35(13):2276–2282, 2019. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bty949.

Bartosz Krawczyk. Learning from imbalanced data: Open challenges and future directions. *Prog Artif Intell*, 5(4):221–232, 2016. ISSN 2192-6352, 2192-6360. doi: 10.1007/s13748-016-0094-0.

Roshan Kumari and Saurabh Kr. Machine Learning: A Review on Binary Classification. *IJCA*, 160(7):11–15, 2017. ISSN 09758887. doi: 10.5120/ijca2017913083.

Kiril Kuzmin, Ayotomiwa Ezekiel Adeniyi, Arthur Kevin DaSouza, Deuk Lim, Huyen Nguyen, Nuria Ramirez Molina, Lanqiao Xiong, Irene T. Weber, and Robert W. Harrison. Machine learning methods accurately predict host specificity of coronaviruses based on spike sequences alone. *Biochem Biophys Res Commun*, 533(3):553–558, 2020. ISSN 0006-291X. doi: 10.1016/j.bbrc.2020.09.010.

Jun Lei, Lichun Sun, Siyu Huang, Chenhong Zhu, Ping Li, Jun He, Vienna Mackey, David H Coy, and Quanyong He. The antimicrobial peptides and their potential clinical applications. *Am J Transl Res*, 11(7):3919–3931, 2019. ISSN 1943-8141.

Nicole A. Lerminiaux and Andrew D. S. Cameron. Horizontal transfer of antibiotic resistance genes in clinical environments. *Can J Microbiol*, 65(1):34–44, 2019. ISSN 1480-3275. doi: 10.1139/cjm-2018-0275.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btp324.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btp352.

Xuhong Li, Yves Grandvalet, Franck Davoine, Jingchun Cheng, Yin Cui, Hang Zhang, Serge Belongie, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Transfer learning in computer vision tasks: Remember where you come from. *Image and Vision Computing*, 93:103853, 2020. ISSN 02628856. doi: 10.1016/j.imavis.2019.103853.

Yu Li, Zeling Xu, Wenkai Han, Huiluo Cao, Ramzan Umarov, Aixin Yan, Ming Fan, Huan Chen, Carlos M. Duarte, Lihua Li, Pak-Leung Ho, and Xin Gao. HMD-ARG: Hierarchical multi-task deep learning for annotating antibiotic resistance genes. *Microbiome*, 9(1):40, 2021. ISSN 2049-2618. doi: 10.1186/s40168-021-01002-3.

Daniel Lichtblau. Alignment-free genomic sequence comparison using FCGR and signal processing. *BMC Bioinformatics*, 20(1):742, 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-3330-3.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI Open*, 3:111–132, 2022. ISSN 2666-6510. doi: 10.1016/j.aiopen.2022.10.001.

Ling Shao, Fan Zhu, and Xuelong Li. Transfer Learning for Visual Categorization: A Survey. *IEEE Trans. Neural Netw. Learning Syst.*, 26(5):1019–1034, 2015. ISSN 2162-237X, 2162-2388. doi: 10.1109/TNNLS.2014.2330900.

Zachary C. Lipton, John Berkowitz, and Charles Elkan. A Critical Review of Recurrent Neural Networks for Sequence Learning, 2015.

Bingyan Liu, Nuoyan Lv, Yuanchun Guo, and Yawen Li. Recent Advances on Federated Learning: A Systematic Survey, 2023.

Zhichang Liu, Dun Deng, Huijie Lu, Jian Sun, Luchao Lv, Shuhong Li, Guanghui Peng, Xianyong Ma, Jiazhou Li, Zhenming Li, Ting Rong, and Gang Wang. Evaluation of Machine Learning Models for Predicting Antimicrobial Resistance of Actinobacillus pleuropneumoniae From Whole Genome Sequences. *Front. Microbiol.*, 11:48, 2020. ISSN 1664-302X. doi: 10.3389/fmicb.2020.00048.

Hannah F Löchel, Dominic Eger, Theodor Sperlea, and Dominik Heider. Deep learning on chaos game representation for proteins. *Bioinformatics*, 36(1):272–279, 2020. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btz493.

Hannah Franziska Löchel and Dominik Heider. Chaos game representation and its applications in bioinformatics. *Comput Struct Biotechnol J*, 19:6263–6271, 2021. ISSN 2001-0370. doi: 10.1016/j.csbj.2021.11.008.

Ji Lv, Senyi Deng, and Le Zhang. A review of artificial intelligence applications for antimicrobial resistance. *Biosafety and Health*, 03(01):22–31, 2021. doi: 10.1016/j.bsheal.2020.08.003.

Yue Ma, Zhengyan Guo, Binbin Xia, Yuwei Zhang, Xiaolin Liu, Ying Yu, Na Tang, Xiaomei Tong, Min Wang, Xin Ye, Jie Feng, Yihua Chen, and Jun Wang. Identification of antimicrobial peptides from the human gut microbiome using deep learning. *Nat Biotechnol*, 40(6):921–931, 2022. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-022-01226-0.

Amirreza Mahbod, Gerald Schaefer, Chunliang Wang, Georg Dorffner, Rupert Ecker, and Isabella Ellinger. Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification. *Computer Methods and Programs in Biomedicine*, 193:105475, 2020. ISSN 01692607. doi: 10.1016/j.cmpb.2020.105475.

Pierre Mahé and Maud Tournoud. Predicting bacterial resistance from whole-genome sequences using k-mers and stability selection. *BMC Bioinformatics*, 19(1):383, 2018. ISSN 1471-2105. doi: 10.1186/s12859-018-2403-z.

Nanditha Mallesh, Max Zhao, Lisa Meintker, Alexander Höllein, Franz Elsner, Hannes Lüling, Torsten Haferlach, Wolfgang Kern, Jörg Westermann, Peter Brossart, Stefan W. Krause, and Peter M. Krawitz. Knowledge transfer to enhance the performance of deep learning models for automated classification of B cell neoplasms. *Patterns*, 2(10):100351, 2021. ISSN 26663899. doi: 10.1016/j.patter.2021.100351.

Swati C Manekar and Shailesh R Sathe. A benchmark study of k-mer counting methods for high-throughput sequencing. *GigaScience*, 7(12):giy125, 2018. ISSN 2047-217X. doi: 10.1093/gigascience/giy125.

Dastan Maulud and Adnan M. Abdulazeez. A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, 1(4):140–147, 2020. ISSN 2708-0757. doi: 10.38094/jastt1457.

Ifeanyi Elibe Mba and Emeka Innocent Nweze. Antimicrobial Peptides Therapy: An Emerging Alternative for Treating Drug-Resistant Bacteria. *Yale J Biol Med*, 95(4):445–463, 2022. ISSN 0044-0086.

Neha Mehra and Surendra Gupta. Survey on Multiclass Classification Methods. 4, 2013.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space, 2013.

Ludovic Minvielle, Mounir Atiq, Sergio Peignier, and Mathilde Mougeot. Transfer Learning on Decision Tree with Class Imbalance. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1003–1010, Portland, OR, USA, 2019. ISBN 978-1-72813-798-8. doi: 10.1109/ictai.2019.00141.

Florian Mittag, Michael Römer, and Andreas Zell. Influence of Feature Encoding and Choice of Classifier on Disease Risk Prediction in Genome-Wide Association Studies. *PLoS One*, 10(8):e0135832, 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0135832.

Danesh Moradigaravand, Martin Palm, Anne Farewell, Ville Mustonen, Jonas Warringer, and Leopold Parts. Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data. *PLoS Comput Biol*, 14(12):e1006258, 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006258.

Jose M. Munita and Cesar A. Arias. Mechanisms of Antibiotic Resistance. *Microbiol Spectr*, 4(2):10.1128/microbiolspec.VMBF–0016–2015, 2016. ISSN 2165-0497. doi: 10.1128/microbiolspec.VMBF-0016-2015.

Mohammad J. Nasiri, Mehri Haeili, Mona Ghazi, Hossein Goudarzi, Ali Pormohammad, Abbas A. Imani Fooladi, and Mohammad M. Feizabadi. New Insights in to the Intrinsic and Acquired Drug Resistance Mechanisms in Mycobacteria. *Front Microbiol*, 8:681, 2017. ISSN 1664-302X. doi: 10.3389/fmicb.2017.00681.

Mark L. Nelson and Stuart B. Levy. The history of the tetracyclines. *Ann N Y Acad Sci*, 1241:17–32, 2011. ISSN 1749-6632. doi: 10.1111/j.1749-6632.2011.06354.x.

Ursula Neumann, Mona Riemenschneider, Jan-Peter Sowa, Theodor Baars, Julia Kälsch, Ali Canbay, and Dominik Heider. Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection approach. *BioData Mining*, 9(1):36, 2016. ISSN 1756-0381. doi: 10.1186/s13040-016-0114-4.

Ursula Neumann, Nikita Genze, and Dominik Heider. EFS: An ensemble feature selection tool implemented as R-package and web-application. *BioData Mining*, 10(1):21, 2017. ISSN 1756-0381. doi: 10.1186/s13040-017-0142-8.

Rui Nian, Jinfeng Liu, and Biao Huang. A review On reinforcement learning: Introduction and applications in industrial process control. *Computers and Chemical Engineering*, 139:106886, 2020. ISSN 0098-1354. doi: 10.1016/j.compchemeng.2020.106886.

Patrice Nordmann, Gaelle Cuzon, and Thierry Naas. The real threat of Klebsiella pneumoniae carbapenemase-producing bacteria. *The Lancet Infectious Diseases*, 9(4):228–236, 2009. ISSN 1473-3099. doi: 10.1016/S1473-3099(09)70054-4.

Jean Ogier du Terrail, Armand Leopold, Clément Joly, Constance Béguier, Mathieu Andreux, Charles Maussion, Benoît Schmauch, Eric W. Tramel, Etienne Bendjebbar, Mikhail

Zaslavskiy, Gilles Wainrib, Maud Milder, Julie Gervasoni, Julien Guerin, Thierry Durand, Alain Livartowski, Kelvin Moutet, Clément Gautier, Inal Djafar, Anne-Laure Moisson, Camille Marini, Mathieu Galtier, Félix Balazard, Rémy Dubois, Jeverson Moreira, Antoine Simon, Damien Drubay, Magali Lacroix-Triki, Camille Franchet, Guillaume Bataillon, and Pierre-Etienne Heudel. Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. *Nat Med*, 29(1):135–146, 2023. ISSN 1546-170X. doi: 10.1038/s41591-022-02155-w.

Ademola Okerinde, Lior Shamir, William Hsu, Tom Theis, and Nasik Nafi. eGAN: Unsupervised approach to class imbalance using transfer learning. *arXiv:2104.04162 [cs]*, 2021.

Simon Orozco-Arias, Mariana S. Candamil-Cortés, Paula A. Jaimes, Johan S. Piña, Reinel Tabares-Soto, Romain Guyot, and Gustavo Isaza. K-mer-based machine learning method to classify LTR-retrotransposons in plant genomes. *PeerJ*, 9:e11456, 2021. ISSN 2167-8359. doi: 10.7717/peerj.11456.

Shruti P and Rekha R. A Review of Convolutional Neural Networks, its Variants and Applications. In *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, pages 31–36, 2023. doi: 10.1109/ICISCoIS56541.2023.10100412.

E Padan, T Tzubery, K Herz, L Kozachkov, A Rimon, and L Galili. NhaA of Escherichia coli, as a model of a pH-regulated Na+/H+antiporter. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1658(1-2):2–13, 2004. ISSN 00052728. doi: 10.1016/j.bbabio.2004.04.018.

Stephen R. Palumbi. Humans as the World's Greatest Evolutionary Force. *Science*, 293 (5536):1786–1790, 2001. doi: 10.1126/science.293.5536.1786.

Sinno Jialin Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359, 2010. ISSN 1041-4347. doi: 10.1109/tkde.2009.191.

Youngjun Park, Anne-Christin Hauschild, and Dominik Heider. Transfer learning compensates limited data, batch effects and technological heterogeneity in single-cell sequencing. *NAR Genomics and Bioinformatics*, 3(4):lqab104, 2021. ISSN 2631-9268. doi: 10.1093/nargab/lqab104.

Sarthak Pati. Federated learning enables big data for rare cancer boundary detection. *Nature Communications*, page 17, 2022. doi: 10.1038/s41467-022-33407-5.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg,

Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

Mitchell W. Pesesky, Tahir Hussain, Meghan Wallace, Sanket Patel, Saadia Andleeb, Carey-Ann D. Burnham, and Gautam Dantas. Evaluation of Machine Learning and Rules-Based Approaches for Predicting Antimicrobial Resistance Profiles in Gram-negative Bacilli from Whole Genome Sequence Data. *Front. Microbiol.*, 7, 2016. ISSN 1664-302X. doi: 10.3389/fmicb.2016.01887.

Jo Plested and Tom Gedeon. Deep transfer learning for image classification: A survey, 2022.

Philipp Pletnev, Ekaterina Guseva, Anna Zanina, Sergey Evfratov, Margarita Dzama, Vsevolod Treshin, Alexandra Pogorel'skaya, Ilya Osterman, Anna Golovina, Maria Rubtsova, Marina Serebryakova, Olga V. Pobeguts, Vadim M. Govorun, Alexey A. Bogdanov, Olga A. Dontsova, and Petr V. Sergiev. Comprehensive Functional Analysis of Escherichia coli Ribosomal RNA Methyltransferases. *Front Genet*, 11:97, 2020. ISSN 1664-8021. doi: 10.3389/fgene.2020.00097.

Laurent Poirel, Jean-Yves Madec, Agnese Lupo, Anne-Kathrin Schink, Nicolas Kieffer, Patrice Nordmann, and Stefan Schwarz. Antimicrobial Resistance in Escherichia coli. *Microbiology Spectrum*, 6(4), 2018. ISSN 2165-0497. doi: 10.1128/microbiolspec.arba-0026-2017.

Stefan Lucian Popa, Cristina Pop, Miruna Oana Dita, Vlad Dumitru Brata, Roxana Bolchis, Zoltan Czako, Mohamed Mehdi Saadani, Abdulrahman Ismaiel, Dinu Iuliu Dumitrascu, Simona Grad, Liliana David, Gabriel Cismaru, and Alexandru Marius Padureanu. Deep Learning and Antibiotic Resistance. *Antibiotics (Basel)*, 11(11):1674, 2022. ISSN 2079-6382. doi: 10.3390/antibiotics11111674.

Kaiyang Qu, Fei Guo, Xiangrong Liu, Yuan Lin, and Quan Zou. Application of Machine Learning in Microbiology. *Front. Microbiol.*, 10:827, 2019. ISSN 1664-302X. doi: 10.3389/fmicb.2019.00827.

Mustafa Radha, Pedro Fonseca, Arnaud Moreau, Marco Ross, Andreas Cerny, Peter Anderer, Xi Long, and Ronald M. Aarts. A deep transfer learning approach for wearable sleep stage classification with photoplethysmography. *npj Digit. Med.*, 4(1):135, 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00510-8.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier Chains for Multi-label Classification. page 16, 2011. doi: 10.1007/s10994-011-5256-5.

Jesse Read, Bernhard Pfahringer, Geoffrey Holmes, and Eibe Frank. Classifier Chains: A Review and Perspectives. *jair*, 70:683–718, 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12376.

Yunxiao Ren, Trinad Chakraborty, Swapnil Doijad, Linda Falgenhauer, Jane Falgenhauer, Alexander Goesmann, Anne-Christin Hauschild, Oliver Schwengers, and Dominik Heider. Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. *Bioinformatics*, page btab681, 2021. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btab681.

Yunxiao Ren, Trinad Chakraborty, Swapnil Doijad, Linda Falgenhauer, Jane Falgenhauer, Alexander Goesmann, Oliver Schwengers, and Dominik Heider. Multi-label classification for multi-drug resistance prediction of Escherichia coli. *Comput Struct Biotechnol J*, 20:1264–1270, 2022a. ISSN 2001-0370. doi: 10.1016/j.csbj.2022.03.007.

Yunxiao Ren, Trinad Chakraborty, Swapnil Doijad, Linda Falgenhauer, Jane Falgenhauer, Alexander Goesmann, Oliver Schwengers, and Dominik Heider. Deep Transfer Learning Enables Robust Prediction of Antimicrobial Resistance for Novel Antibiotics. *Antibiotics*, 11(11):1611, 2022b. ISSN 2079-6382. doi: 10.3390/antibiotics11111611.

Sandra Reuter, Matthew J. Ellington, Edward J. P. Cartwright, Claudio U. Köser, M. Estée Török, Theodore Gouliouris, Simon R. Harris, Nicholas M. Brown, Matthew T. G. Holden, Mike Quail, Julian Parkhill, Geoffrey P. Smith, Stephen D. Bentley, and Sharon J. Peacock. Rapid Bacterial Whole-Genome Sequencing to Enhance Diagnostic and Public Health Microbiology. *JAMA Intern Med*, 173(15):1397–1404, 2013. ISSN 2168-6106. doi: 10.1001/jamainternmed.2013.7734.

Bernardo Ribeiro da Cunha, Luís P. Fonseca, and Cecília R. C. Calado. Antibiotic Discovery: Where Have We Come from, Where Do We Go? *Antibiotics*, 8(2):45, 2019. ISSN 2079-6382. doi: 10.3390/antibiotics8020045.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, Sébastien Ourselin, Micah Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. The future of digital health with federated learning. *npj Digit. Med.*, 3(1):1–7, 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-00323-1.

Mona Riemenschneider, Robin Senge, Ursula Neumann, Eyke Hüllermeier, and Dominik Heider. Exploiting HIV-1 protease and reverse transcriptase cross-resistance information for improved drug resistance prediction by means of multi-label classification. *BioData Mining*, 9(1):10, 2016. ISSN 1756-0381. doi: 10.1186/s13040-016-0089-1.

Riccardo Rizzo, Antonino Fiannaca, Massimo La Rosa, and Alfonso Urso. Classification Experiments of DNA Sequences by Using a Deep Neural Network and Chaos Game Representation. In *Proceedings of the 17th International Conference on Computer Systems and Technologies 2016*, pages 222–228, Palermo Italy, 2016. ISBN 978-1-4503-4182-0. doi: 10.1145/2983468.2983489.

Terry Roemer and Charles Boone. Systems-level antimicrobial drug and drug synergy discovery. *Nat Chem Biol*, 9(4):222–231, 2013. ISSN 1552-4469. doi: 10.1038/nchembio.1205.

Lior Rokach, Alon Schclar, and Ehud Itach. Ensemble methods for multi-label classification. *Expert Systems with Applications*, 41(16):7507–7523, 2014. ISSN 09574174. doi: 10.1016/j.eswa.2014.06.015.

Sudipan Saha and Tahir Ahmad. Federated Transfer Learning: Concept and applications, 2021.

Iqbal H. Sarker. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN COMPUT. SCI.*, 2(3):160, 2021. ISSN 2661-8907. doi: 10.1007/s42979-021-00592-x.

Ron Schwessinger, Matthew Gosden, Damien Downes, Richard C. Brown, A. Marieke Oudelaar, Jelena Telenius, Yee Whye Teh, Gerton Lunter, and Jim R. Hughes. DeepC: Predicting 3D genome folding using megabase-scale transfer learning. *Nat Methods*, 17 (11):1118–1124, 2020. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-020-0960-3.

Dipendra C Sengupta, Matthew D Hill, Kevin R Benton, and Hirendra N Banerjee. Similarity Studies of Corona Viruses through Chaos Game Representation. *Comput Mol Biosci*, 10(3):61–72, 2020. ISSN 2165-3445. doi: 10.4236/cmb.2020.103004.

Shamanth A. Shankarnarayan, Joshua D. Guthrie, Daniel A. Charlebois, Shamanth A. Shankarnarayan, Joshua D. Guthrie, and Daniel A. Charlebois. Machine Learning for Antimicrobial Resistance Research and Drug Development. In *The Global Antimicrobial Resistance Epidemic - Innovative Approaches and Cutting-Edge Solutions*. 2022. ISBN 978-1-80356-042-7. doi: 10.5772/intechopen.104841.

Priyanka Sharma and Pritee Parwekar. Multiclass Classification of Online Reviews Using NLP and Machine Learning for Non-english Language. In Hakimjon Zaynidinov, Madhusudan Singh, Uma Shanker Tiwary, and Dhananjay Singh, editors, *Intelligent Human Computer Interaction*, Lecture Notes in Computer Science, pages 85–94, Cham, 2023. ISBN 978-3-031-27199-1. doi: 10.1007/978-3-031-27199-1_9.

Alex Sherstinsky. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020. ISSN 0167-2789. doi: 10.1016/j.physd.2019.132306.

Stephen Solis-Reyes, Mariano Avino, Art Poon, and Lila Kari. An open-source k-mer based machine learning tool for fast and accurate subtyping of HIV-1 genomes. *PLOS ONE*, 13 (11):e0206409, 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0206409.

Sebastian Spänig and Dominik Heider. Encodings and models for antimicrobial peptide classification for multi-resistant pathogens. *BioData Mining*, 12(1):7, 2019. ISSN 1756-0381. doi: 10.1186/s13040-019-0196-x.

B S Speer, N B Shoemaker, and A A Salyers. Bacterial resistance to tetracycline: Mechanisms, transfer, and clinical significance. *Clin Microbiol Rev*, 5(4):387–399, 1992. ISSN 0893-8512.

Réka Spohn, Lejla Daruka, Viktória Lázár, Ana Martins, Fanni Vidovics, Gábor Grézal, Orsolya Méhi, Bálint Kintses, Mónika Számel, Pramod K. Jangir, Bálint Csörgő, Ádám Györkei, Zoltán Bódi, Anikó Faragó, László Bodai, Imre Földesi, Diána Kata, Gergely Maróti, Bernadett Pap, Roland Wirth, Balázs Papp, and Csaba Pál. Integrated evolutionary analysis reveals antimicrobial peptides with limited resistance. *Nat Commun*, 10(1):4538, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-12364-6.

Vanja Stojković, Lianet Noda-Garcia, Dan S. Tawfik, and Danica Galonić Fujimori. Antibiotic resistance evolved via inactivation of a ribosomal RNA methylating enzyme. *Nucleic Acids Res*, 44(18):8897–8907, 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw699.

Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackerman, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. A Deep Learning Approach to Antibiotic Discovery. *Cell*, 180(4): 688–702.e13, 2020. ISSN 00928674. doi: 10.1016/j.cell.2020.01.021.

Dongchang Sun, Katy Jeannot, Yonghong Xiao, and Charles W. Knapp. Editorial: Horizontal Gene Transfer Mediated Bacterial Antibiotic Resistance. *Frontiers in Microbiology*, 10, 2019. ISSN 1664-302X. doi: 10.3389/fmicb.2019.01933.

Yulian Sun. Federated Transfer Learning with Multimodal Data, 2022.

Zeju Sun, Shaojun Pei, Rong Lucy He, and Stephen S.-T. Yau. A novel numerical representation for proteins: Three-dimensional Chaos Game Representation and its Extended Natural Vector. *Computational and Structural Biotechnology Journal*, 18:1904–1913, 2020. ISSN 2001-0370. doi: 10.1016/j.csbj.2020.07.004.

H. Suzuki, S. Kamatani, and H. Kumagai. Purification and characterization of aminopeptidase B from Escherichia coli K-12. *Biosci Biotechnol Biochem*, 65(7):1549–1558, 2001. ISSN 0916-8451. doi: 10.1271/bbb.65.1549.

Kyle Swanson, Eric Wu, Angela Zhang, Ash A. Alizadeh, and James Zou. From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell*, 186(8):1772–1791, 2023. ISSN 00928674. doi: 10.1016/j.cell.2023.01.035.

Paulina Szymczak, Marcin Możejko, Tomasz Grzegorzek, Radosław Jurczak, Marta Bauer, Damian Neubauer, Karol Sikora, Michał Michalski, Jacek Sroka, Piotr Setny, Wojciech Kamysz, and Ewa Szczurek. Discovering highly potent antimicrobial peptides with deep generative model HydrAMP. *Nat Commun*, 14(1):1453, 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-36994-z.

Adane Nega Tarekegn, Mario Giacobini, and Krzysztof Michalak. A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:107965, 2021. ISSN 0031-3203. doi: 10.1016/j.patcog.2021.107965.

Clifford A. Tawiah and Victor S. Sheng. A Study on Multi-label Classification. In Petra Perner, editor, *Advances in Data Mining. Applications and Theoretical Aspects*, Lecture Notes in Computer Science, pages 137–150, Berlin, Heidelberg, 2013. ISBN 978-3-642-39736-3. doi: 10.1007/978-3-642-39736-3_11.

Soufiane Telhig, Laila Ben Said, Séverine Zirah, Ismail Fliss, and Sylvie Rebuffat. Bacteriocins to Thwart Bacterial Resistance in Gram Negative Bacteria. *Front Microbiol*, 11: 586433, 2020. ISSN 1664-302X. doi: 10.3389/fmicb.2020.586433.

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining Multi-label Data. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Boston, MA, 2009. ISBN 978-0-387-09822-7 978-0-387-09823-4. doi: 10.1007/978-0-387-09823-4_34.

Meysam Vakili, Mohammad Ghamsari, and Masoumeh Rezaei. Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification. *arXiv*, page 13, 2020.

Pieter-Jan Van Camp, David B. Haslam, and Aleksey Porollo. Prediction of Antimicrobial Resistance in Gram-Negative Bacteria From Whole-Genome Sequencing Data. *Front. Microbiol.*, 11:1013, 2020. ISSN 1664-302X. doi: 10.3389/fmicb.2020.01013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, 2017.

Daniel Veltri, Uday Kamath, and Amarda Shehu. Deep learning improves antimicrobial peptide recognition. *Bioinformatics*, 34(16):2740–2747, 2018. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bty179.

Michael Wainberg, Daniele Merico, Andrew Delong, and Brendan J Frey. Deep learning in biomedicine. *Nat Biotechnol*, 36(9):829–838, 2018. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.4233.

Christopher T. Walsh and Timothy A. Wencewicz. Prospects for new antibiotics: A molecule-centered perspective. *J Antibiot*, 67(1):7–22, 2014. ISSN 1881-1469. doi: 10.1038/ja.2013.49.

Yingwei Wang, Kathleen Hill, Shiva Singh, and Lila Kari. The spectrum of genomic signatures: From dinucleotides to chaos game representation. *Gene*, 346:173–185, 2005. ISSN 0378-1119. doi: 10.1016/j.gene.2004.10.021.

Stefanie Warnat-Herresthal, Hartmut Schultze, Krishnaprasad Lingadahalli Shastry, Sathyanarayanan Manamohan, Saikat Mukherjee, Vishesh Garg, Ravi Sarveswara, Kristian Händler, Peter Pickkers, N. Ahmad Aziz, Sofia Ktena, Florian Tran, Michael Bitzer, Stephan Ossowski, Nicolas Casadei, Christian Herr, Daniel Petersheim, Uta Behrends, Fabian Kern, Tobias Fehlmann, Philipp Schommers, Clara Lehmann, Max Augustin, Jan Rybniker, Janine Altmüller, Neha Mishra, Joana P. Bernardes, Benjamin Krämer, Lorenzo Bonaguro, Jonas Schulte-Schrepping, Elena De Domenico, Christian Siever, Michael Kraut, Milind Desai, Bruno Monnet, Maria Saridaki, Charles Martin Siegel, Anna Drews, Melanie Nuesch-Germano, Heidi Theis, Jan Heyckendorf, Stefan Schreiber, Sarah Kim-Hellmuth, COVID-19 Aachen Study (COVAS), Paul Balfanz, Thomas Eggermann, Peter Boor, Ralf Hausmann, Hannah Kuhn, Susanne Isfort, Julia Carolin Stingl, Günther Schmalzing, Christiane K. Kuhl, Rainer Röhrig, Gernot Marx, Stefan Uhlig, Edgar Dahl, Dirk Müller-Wieland, Michael Dreher, Nikolaus Marx, Jacob Nattermann, Dirk Skowasch, Ingo Kurth, Andreas Keller, Robert Bals, Peter Nürnberg, Olaf Rieß, Philip Rosenstiel, Mihai G. Netea, Fabian Theis, Sach Mukherjee, Michael Backes, Anna C. Aschenbrenner, Thomas Ulas, Deutsche COVID-19 Omics Initiative (DeCOI), Angel Angelov, Alexander Bartholomäus, Anke Becker, Daniela Bezdan, Conny Blumert, Ezio Bonifacio, Peer Bork, Bunk Boyke,

Helmut Blum, Thomas Clavel, Maria Colome-Tatche, Markus Cornberg, Inti Alberto De La Rosa Velázquez, Andreas Diefenbach, Alexander Dilthey, Nicole Fischer, Konrad Förstner, Sören Franzenburg, Julia-Stefanie Frick, Gisela Gabernet, Julien Gagneur, Tina Ganzenmueller, Marie Gauder, Janina Geißert, Alexander Goesmann, Siri Göpel, Adam Grundhoff, Hajo Grundmann, Torsten Hain, Frank Hanses, Ute Hehr, André Heimbach, Marius Hoeper, Friedemann Horn, Daniel Hübschmann, Michael Hummel, Thomas Iftner, Angelika Iftner, Thomas Illig, Stefan Janssen, Jörn Kalinowski, René Kallies, Birte Kehr, Oliver T. Keppler, Christoph Klein, Michael Knop, Oliver Kohlbacher, Karl Köhrer, Jan Korbel, Peter G. Kremsner, Denise Kühnert, Markus Landthaler, Yang Li, Kerstin U. Ludwig, Oliwia Makarewicz, Manja Marz, Alice C. McHardy, Christian Mertes, Maximilian Münchhoff, Sven Nahnsen, Markus Nöthen, Francine Ntoumi, Jörg Overmann, Silke Peter, Klaus Pfeffer, Isabell Pink, Anna R. Poetsch, Ulrike Protzer, Alfred Pühler, Nikolaus Rajewsky, Markus Ralser, Kristin Reiche, Stephan Ripke, Ulisses Nunes da Rocha, Antoine-Emmanuel Saliba, Leif Erik Sander, Birgit Sawitzki, Simone Scheithauer, Philipp Schiffer, Jonathan Schmid-Burgk, Wulf Schneider, Eva-Christina Schulte, Alexander Sczyrba, Mariam L. Sharaf, Yogesh Singh, Michael Sonnabend, Oliver Stegle, Jens Stoye, Janne Vehreschild, Thirumalaisamy P. Velavan, Jörg Vogel, Sonja Volland, Max von Kleist, Andreas Walker, Jörn Walter, Dagmar Wieczorek, Sylke Winkler, John Ziebuhr, Monique M. B. Breteler, Evangelos J. Giamarellos-Bourboulis, Matthijs Kox, Matthias Becker, Sorin Cheran, Michael S. Woodacre, Eng Lim Goh, and Joachim L. Schultze. Swarm Learning for decentralized and confidential clinical machine learning. *Nature*, 594(7862):265–270, 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03583-3.

Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *J Big Data*, 3(1):9, 2016. ISSN 2196-1115. doi: 10.1186/s40537-016-0043-6.

Karl R. Weiss and Taghi M. Khoshgoftaar. Investigating Transfer Learners for Robustness to Domain Class Imbalance. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 207–213, Anaheim, CA, USA, 2016. ISBN 978-1-5090-6167-9. doi: 10.1109/icmla.2016.0042.

WHO World Health Organization. GLASS: Whole-genome sequencing for surveillance of antimicrobial resistance, 2020.

Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Tao Qi, Yongfeng Huang, and Xing Xie. A federated graph neural network framework for privacy-preserving personalization. *Nat Commun*, 13(1):3091, 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-30714-9.

Bin Xu, Sheng Yan, Shuai Li, and Yidi Du. A Federated Transfer Learning Framework Based on Heterogeneous Domain Adaptation for Students' Grades Classification. *Applied Sciences*, 12(21):10711, 2022. ISSN 2076-3417. doi: 10.3390/app122110711.

Jian-Yi Yang, Zhen-Ling Peng, Zu-Guo Yu, Rui-Jie Zhang, Vo Anh, and Desheng Wang. Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation. *Journal of Theoretical Biology*, 257(4):618–626, 2009. ISSN 0022-5193. doi: 10.1016/j.jtbi.2008.12.027.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated Machine Learning: Concept and Applications, 2019.

Yang Yang, Katherine E Niehaus, Timothy M Walker, Zamin Iqbal, A Sarah Walker, Daniel J Wilson, Tim E A Peto, Derrick W Crook, E Grace Smith, Tingting Zhu, and David A Clifton. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics*, 34(10):1666–1671, 2018. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btx801.

Hiroaki Yoshida, Tsuyoshi Kojima, Jun-ichi Yamagishi, and Shinichi Nakamura. Quinolone-resistant mutations of the gyrA gene of Escherichia coli. *Mol Gen Genet*, 211(1):1–7, 1988. ISSN 1432-1874. doi: 10.1007/BF00338386.

Jinhua Yu, Yinhui Deng, Tongtong Liu, Jin Zhou, Xiaohong Jia, Tianlei Xiao, Shichong Zhou, Jiawei Li, Yi Guo, Yuanyuan Wang, Jianqiao Zhou, and Cai Chang. Lymph node metastasis prediction of papillary thyroid carcinoma based on transfer learning radiomics. *Nat Commun*, 11(1):4807, 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-18497-3.

Ning Yu, Zhihua Li, and Zeng Yu. Survey on encoding schemes for genomic data representation and feature learning—from signal processing to machine learning. *Big Data Mining and Analytics*, 1(3):191–210, 2018. ISSN 2096-0654. doi: 10.26599/BDMA.2018.9020018.

Zu-Guo Yu, Vo Anh, and Ka-Sing Lau. Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses. *Journal of Theoretical Biology*, 226(3):341–348, 2004. ISSN 0022-5193. doi: 10.1016/j.jtbi.2003.09.009.

Sojib Bin Zaman, Muhammed Awlad Hussain, Rachel Nye, Varshil Mehta, Kazi Taib Mamun, and Naznin Hossain. A Review on Antibiotic Resistance: Alarm Bells are Ringing. *Cureus*, 2017. ISSN 2168-8184. doi: 10.7759/cureus.1403.

Min-Ling Zhang and Zhi-Hua Zhou. A Review on Multi-Label Learning Algorithms. *IEEE Trans. Knowl. Data Eng.*, 26(8):1819–1837, 2014. ISSN 1041-4347. doi: 10.1109/TKDE.2013.39.

Zehui Zhang, Ningxin He, Dongyu Li, Hang Gao, Tiegang Gao, and Chuan Zhou. Federated transfer learning for disaster classification in social computing networks. *Journal of Safety Science and Resilience*, 3(1):15–23, 2022a. ISSN 2666-4496. doi: 10.1016/j.jnlssr.2021.10.007.

Zhenyan Zhang, Qi Zhang, Tingzhang Wang, Nuohan Xu, Tao Lu, Wenjie Hong, Josep Penuelas, Michael Gillings, Meixia Wang, Wenwen Gao, and Haifeng Qian. Assessment of global health risk of antibiotic resistance genes. *Nat Commun*, 13(1):1553, 2022b. ISSN 2041-1723. doi: 10.1038/s41467-022-29283-8.

Yiting Zhou, Tingfang Wu, Yelu Jiang, Yan Li, Kailong Li, Lijun Quan, and Qiang Lyu. DeepNup: Prediction of Nucleosome Positioning from DNA Sequences Using Deep Neural Network. *Genes*, 13(11):1983, 2022. ISSN 2073-4425. doi: 10.3390/genes13111983.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A Comprehensive Survey on Transfer Learning, 2020.

A
Appendix

Danksagung

Time flies, and in the blink of an eye, my doctoral career has come to an end. I still remember clearly that on the second day of the Chinese New Year in 2020, the sudden outbreak of the pandemic pressed the pause button on the transportation of all the cities. I hurriedly packed my bags and said goodbye to my family and came to Germany. Just a month later, schools here began shutting down. I never anticipated that the pandemic would span three years, which took up most of my doctoral career. Although the pandemic has brought a lot of inconvenience to my life and work, I still feel very lucky to have joined Prof. Dr. Dominik Heider's lab in the past three years, and met a group of warm and kind colleagues and supervisor, which have made my Ph.D. career very fulfilling and happy, and I am very thankful for all of them.

First of all, I would like to give special thanks to my advisor, Prof. Dr. Dominik Heider. I am grateful for his guidance on my project and his help in my life. Even though he was busy, he arranged regular one-on-one meetings once every two weeks so that we could communicate the progress of the project and problems encountered in time. I remember during the first year of my PhD, I encountered a lot of difficulties in my project and made less progress. He did not criticize me but encouraged me all the time and gave me a lot of attentive guidance and innovative suggestions. He also gave me a lot of support in writing my research articles. I also appreciate that he always replies to emails very promptly and is always happy to help, no matter what difficulties I encounter in life or work.

I am also grateful to Prof. Dr. Anne-Christin Hauschild. I remember on my first day in Germany, she gave up her Saturday to pick me up at the train station, helped me with my suitcase, and took me to my dormitory. I am also grateful to her for her guidance and help in my project as the leader of our small team in the following year.

I am also very grateful to Dr. Georges Hattab, who was always very warm and helped me with some of the difficulties I encountered in my life, such as applying for a bank card and registering with the city government when I first came to Germany.

I am also very grateful to Dr. Sebastian Spänig. We worked in the same office and exchanged ideas frequently, I learned a lot from our discussion. I would also like to thank his efforts in applying for the DAAD funding, but in the end, I am sorry that he could not go to Hong Kong with me due to various delays.

I am also thankful to my colleague Mohammad Tajabadi, we collaborated on advancing a project and he spent a lot of time helping me with my queries.

I am also grateful to Marius Welzel, who helped me analyze a batch of data using a server with GPUs at the beginning of my project.

Thanks to Dr. Roman Martin, Leon Fehse, and Jan Ruhland for organizing the activities in the lab, which allowed me to have more communication with everyone, learn more about the food in Marburg, watch many sci-fi movies I hadn't seen before, and participate in a dragon boat race in Germany. Thanks to Jan Henric Klau and his wife, his wife was so kind to teach me swimming.

Thanks to Chisom Ezekannagha, we shared our happiness and stress with each other, and urged each other to write our thesis. Thanks to Aleksandar Anžel, I learned a lot from his talk, and thank him for sharing in the graduation process.

Thanks also to Dr. Hannah Franziska Löchel, Dr. Oluwafemi A. Sarumi, Dr. Adèle Ribeiro, and Sandra Clemens for their comments and suggestions on my project at the group meeting.

Thanks also to Prof. Margaret Ip and the members. Thanks for the successful collaboration with them and for their hospitality during the exchange. I would also like to thank Prof. Alexander Goesmann and Prof. Trinad Chakraborty and their group members, especially Dr. Oliver Schwengers and Dr. Swapnil Doijad, for providing me with the data used in my Ph.D. project and processing the raw data I have learned a lot from working with them.

Thanks to Prof. Dr. Ho Ryun Chung, Dr. Roman Martin, Mohammad Tajabadi, and two of my friends for reading my thesis and for their suggestions.

Thanks to Dr. Ines Block and Katrin Bopper for managing our daily laboratory affairs.

Thanks to my friends, knowing them made me feel less lonely in a foreign country and in the pandemic. We encouraged each other, helped each other, talked about our stress, shared

our joys, traveled together, cooked together, learned, and improved together.

Finally, I would like to thank my family for their constant support, concern, and encouragement. But I feel very sorry for them, I have not been back home for nearly four years due to the pandemic, even when my grandparents passed away. I didn't also give my parents much companionship and care when they were sick.