ACIBADEM MEHMET ALİ AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

NEW METHODOLOGIES FOR UNDERSTANDING ANTIBIOTIC RESISTANCE MECHANISMS IN TRANSLATIONAL MEDICINE IN GRAM-POSITIVE BACTERIA

DENİZ ECE KAYA, M.Sc.
DOCTORATE THESIS

DEPARTMENT OF BIOSTATISTICS AND MEDICAL INFORMATICS

SUPERVISOR
Prof. Dr. Özgür Kurt

ISTANBUL – 2023

ACIBADEM MEHMET ALİ AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

NEW METHODOLOGIES FOR UNDERSTANDING ANTIBIOTIC RESISTANCE MECHANISMS IN TRANSLATIONAL MEDICINE IN GRAM-POSITIVE BACTERIA

DENİZ ECE KAYA, M.Sc.

DOCTORATE THESIS

DEPARTMENT OF BIOSTATISTICS AND MEDICAL INFORMATICS

SUPERVISOR
Prof. Dr. Özgür Kurt

ISTANBUL – 2023

Department:	Department of Biostatistics and Medical Informatics
Program:	Biostatistics and Bioinformatics
Thesis Title:	New Methodologies for Understanding Antibiotic Resistance Mechanisms in Translational Medicine in Gram-Positive Bacteria
Student's Name and Surname:	Deniz Ece Kaya
Date of Defense: 01/02/2023	

This is to certify that I have examined this copy of Ph.D. thesis. I have found that she prepared after fulfilling the specified requirements in the associated legislations before the final examining committee, whose signatures are below.

Jury Member (Head of the Defense)	Prof. Osman Uğur Sezerman Acibadem Mehmet Ali Aydinlar University
Jury Member (Thesis Supervisor)	Prof. Özgür Kurt Acibadem Mehmet Ali Aydinlar University
Jury Member	Assoc. Prof. Emel Timuçin Acibadem Mehmet Ali Aydinlar University
Jury Member	Assoc. Prof. Burcu Güngör Abdullah Gul University
Jury Member	Prof. Nurver Ulger Toprak Marmara University

DECLARATION

I declare that this thesis work is my own work, I had no unethical behavior at any stages from the planning to the writing of the thesis, I obtained all the information in this thesis in accordance with academic and ethical rules, I cited all the information and comments that were not obtained with this thesis work, and I provided resources in the list of references. I also declare that there was no violation of any patents and copyrights during the study and writing of this thesis.

Date: 17/01/2023
Deniz Ece Kaya

PREFACE AND ACKNOWLEDGEMENTS

My journey would not have been possible without the support and guidance of many people.

First, I would like to express my warmest appreciation to my advisor, Prof. Özgür Kurt and Prof. Osman Uğur Sezerman, Ph.D., who have continually provided me with wisdom, support, guidance, and expertise. I could not have undertaken this journey without my defense committee, Prof. Sesin Kocagoz, and Assoc. Prof. Emel Timuçin for providing their expertise and knowledge.

It would have been almost impossible to complete this thesis without my colleague, my best friend Ege Ulgen, MD., Ph.D. Without our extensive brainstorming sessions and his endless patience in answering my questions, I could not have finished it. I am deeply thankful to Nogayhan Seymen for all the laughs and excitement we shared. Working with him has been very rewarding, and I am grateful that I had the opportunity to do so, and still, I have.

Thank you to my family, my mum, Leyla, for her endless support, eternal love, and encouragement. Her belief in me has kept my spirits and motivation high during this process. I also thank my father, Kansu, and Fatma, for their never-ending, unconditional support. The accomplishments I have made would not be possible without your endless efforts.

To Nagehan, Alp and Firuze, I am lucky to have you. Thank you for always being there and putting up with all my anxiety. You have made my life so much better and more beautiful because of your friendship.

My final thanks go to Umut, my love and my best friend. I'm not sure how I could do it without you and thank you so much for your patience in dealing with me throughout this process. Thank you.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS AND SYMBOL

AA k-mer	Amino acid k-mer
AMR	Antimicrobial Resistance
AST	Antibiotic Susceptibility Test
CLSI	Clinical and Laboratory Standards Institute
EEA	European Economic Area
EU	European Union
EUCAST	European Committee on Antimicrobial Susceptibility Testing
GBM	Gradient Boosting
MIC	Minimum Inhibitory Concentration
ML	Machine Learning
MRSA	Methicillin-resistant *Staphylococcus aureus*
N k-mer	Nucleotide k-mer
NGS	Next-generation Sequencing
PBPs	Penicillin-binding proteins
RF	Random Forest
SNP	Single Nucleotide Polymorphisms
SVM	Support Vector Machine
WGS	Whole Genome Sequencing
WHO	World Health Organization
XGBoost	Extreme Gradient Boosting

LIST OF FIGURES

LIST OF TABLES

ÖZET

Gram Pozitif Bakterilerde Translasyonel Tıpta Antibiyotik Direnç Mekanizmalarını Anlamak İçin Yeni Metodolojiler

Gram pozitif bakteriler olan Streptococcus pneumoniae, Staphylococcus aureus ve Enterococcus faecium, klinisyenlerin en büyük endişesi ve küresel halk sağlığı sorunlarından biridir. Antimikrobiyal direnç, bu patojenlerden kaynaklanan yüksek morbidite ve mortalite oranları ile ilişkilidir. Makine öğrenimi, azalan genom dizileme maliyetleri nedeniyle gram-pozitif patojen fenotiplerini anlamak, teşhis etmek, tedavi etmek ve tahmin etmek için popüler bir araç haline gelmiştir. Nükleotid k-merler, amino asit k-merler, tek nükleotid polimorfizmler ve bu özelliklerin kombinasyonları zengin genetik bilgiler içermektedir ve bu bilgiler tüm genom dizileme sayesinde ulaşılabilir olmuştur. Bu çalışmamızda, S.pneumoniae, S.aureus ve E.faecium örnekleri farklı makine öğrenimi modelleri kullanılarak fenotip tahmini yapılmış ve sonuçları karşılaştırılmıştır. S.pneumoniae için penisilin, eritromisin ve tetrasiklin, S.aureus için eritromisin, tetrasiklin, gentamisin ve siprofloksasin ve E. faecium için ampisilin ve vankomisin antibiyotikleri kullanılmıştır. Bu modelleri eğitmek için rastgele ormanlar, destek vektör makineleri, stokastik gradyan artırma ve aşırı gradyan artırma gibi çeşitli makine öğrenme yöntemleri kullanılmış ve karşılaştırılmıştır. Modellerin sonuçlarını değerlendirirken, gram-pozitif patojenlerin fenotiplerini yüksek F1 skoru ve doğruluğu ile tahmin edilebildiği görülmüştür ve bu yaklaşımı kullanmak, ileri çalışmalar için AMR tahminini iyileştirmek için kullanılabilir. Bu çalışmada, AMR tahmin modeli kurulumunun temel özelliklerinin ve makine öğrenimi yöntemi seçiminin sonuçları etkilediği görülmüştür.

Anahtar Sözcükler: Antimikrobiyal direnç, K-mer, Makine öğrenimi, Tek nükleotid polimorfizm, Tüm genom dizileme.

ABSTRACT

New Methodologies for Understanding Antibiotic Resistance Mechanisms in Translational Medicine in Gram-Positive Bacteria

Streptococcus pneumoniae, Staphylococcus aureus, and Enterococcus faecium, which are gram-positive bacteria, are the major concerns of clinicians and one of the global public health problems. Antimicrobial resistance is associated with high morbidity and mortality rates from these pathogens. Due to reduced genome sequencing costs, machine learning has become a popular tool for understanding, diagnosing, treating, and predicting gram-positive pathogen phenotypes. As a result of whole-genome sequencing, k-mer nucleotides, amino acid k-mers, single nucleotide polymorphisms, and combinations of these features provide rich genetic information. This study compares different machine learning models for S.pneumoniae, S.aureus, and E.faecium. For three antibiotics: Penicillin, Erythromycin, and Tetracycline, we compared nucleotide k-mers, amino acid k-mers, and SNPs in S.pneumoniae, S.aureus for four antibiotics: Erythromycin, Tetracycline, Gentamicin and Ciprofloxacin, and E.faecium for two antibiotics: Ampicillin and Vancomycin. We used and compared machine learning methods such as random forests, support vector machines, stochastic gradient boosting, and extreme gradient boosting to train these models.

When evaluating the results of models, we can predict phenotypes of gram-positive pathogens with high F1 scores and accuracy. This approach can be used to improve AMR prediction for further studies. In this study, we found that key features of the AMR prediction model setup and the choice of machine learning method affected the results.

Keywords: Antimicrobial resistance, K-mer, Machine learning, Single nucleotide polymorphism, Whole genome sequence.

1 INTRODUCTION AND AIM

Globally, the rate of morbidity and mortality associated with infectious diseases has been increasing due to antimicrobial resistance (AMR). Approximately 10 million people will die from infectious diseases by 2050, according to World Health Organization (WHO), making AMR one of the top 10 global health threats (1, 2). Every year, due to antibiotic-resistant bacteria, more than 33 thousand people are killed in European Union (EU) and European Economic Area (EEA) (62). Increasing antimicrobial resistance in pathogens has led to the ineffectiveness of many commonly used antibiotics (3). New antimicrobial compounds have not been developed as quickly as resistance has spread (4,5). This situation raises global health concerns. The rapid antibiotic susceptibility test (AST) can guide the use of antibiotics and reduce the chances of drug-resistant strains emerging. AST methods that use classical phenotypic culturing methods are currently the gold standard. A pure culture isolate is used to test the effects of antimicrobial drugs by current AST recommendations (63). The breakpoint values used by microbiology laboratories to determine whether microbes are susceptible to or resistant to antibiotics are set by the European Committee on Antimicrobial Susceptibility Testing (EUCAST). EUCAST-calibrated systems are used, including the disk diffusion method. The problem with these methods is that they take a few days to yield results, which delays urgent treatment decisions (6,7). As a result of this delay, drug resistance also spreads (8). The latest advances have inspired new approaches to AST in electronics, biosensor technologies, optics, microfluidics, hybridization, and DNA amplification (63).

Antimicrobial resistance can be prevented through molecular approaches, which have significantly improved recently (9-11). Whole genome sequencing (WGS), or direct metagenomic sequencing of clinical materials, has been proposed as the next-generation genotypic AST due to rapid advances in sequencing technology and decreasing costs (7,12).

Due to reduced genome sequencing costs, AMR phenotypes can be detected directly from sequence data. Recently, Machine Learning (ML) has attracted significant attention in the literature for understanding, diagnosing, treating, and predicting AMR phenotypes (8, 9,22-42).

In this study, we compare different ML models for predicting AMR phenotype for three widespread gram-positive bacteria. These are *Streptococcus pneumoniae*, *Streptococcus aureus,* and *Enterococcus faecium.* We compared six different feature types (Nucleotide k-mer (10-mer), amino acid k-mer (5-mer), SNPs, n k-mer/ aa k-mer, n k-mer/ SNP, aa k-mer/ SNP) to predict AMR. For *S.pneumoniae*, we designed our model for three antibiotics: penicillin, erythromycin, and tetracycline. For S. *a**ur**e**us,* four antibiotics were used. These are erythromycin, tetracycline, gentamicin, and ciprofloxacin. For *E.faecium,* ampicillin and vancomycin were used. Various ML methods (random forest (RF), support vector machine (SVM), stochastic gradient boosting (GBM), and extreme gradient boosting (XGBoost) were compared to train the models. We discuss the strengths and limitations of feature and ML model selection for MIC prediction, and we can say that for each antibiotic, there is a difference in the F1 score and accuracy depending on the feature selection and machine learning model used.

2 BACKGROUND

2.1. Antimicrobial Resistance & Current Technologies

Antibiotic resistance can occur by a variety of mechanisms. Antibiotics can be chemically modified and degraded by phosphorylation, acetylation, monooxygenation, ADP-ribosylation, and glycosylation. A point mutation in the ribosomal protein or changes in the target molecule can make the antibiotic inefficient. There are a variety of antimicrobials and resistance mechanisms that complicate the process of AST. The drug intake can be prevented, efflux increased, and some mechanisms involve reprogramming cell wall synthesis (63-70).

Using phenotypic testing methods such as disk diffusion, broth dilution, agar dilution, gradient, and breakpoint tests, EUCAST and CLSI recommend that reliable antibiotic resistance diagnosis requires phenotypic testing. This method involves continuously exposing a bacterial isolate. In addition to optoelectronics, microfluidics and highly sensitive optical systems will benefit from using advanced optoelectronic systems (63-70). A preliminary step in AST is the identification of the pathogen. Microscopy and gram-staining are the main start points for blood samples. Then samples are cultured to get pure isolate, and minimum inhibitory concentration (MIC) determination is performed. The last part can take several dates unless laboratories have mass spectrometry instruments. Using MALDI-TOF, laser energy evaporates matrix-bound samples, which are analyzed immediately. Vitek AST cards selected based on MALTI-TOF analysis were used for susceptibility testing blood cultures after short-term cultivation on agar plates (63-70).

2.2. Antimicrobial Resistance & Whole Genome Sequencing

Compared to Moore's Law, which is an observation and projection of historic trends, the cost of next-generation sequencing (NGS) has significantly decreased since the Human Genome Project (Figure 1) (71). Technology is enabling NGS to become more affordable as costs continue to decline. Traditional methods such as Sanger sequencing or qPCR may be useful when only a few targets are being analyzed on a

few samples. Still, NGS is more efficient when sequencing large samples or more than 20 target regions (72). Moreover, NGS can detect novel or rare variants with higher power and sensitivity. According to Illumina, targeted gene expression profiling costs $23, and 16S metagenomic sequencing costs $18 per sample (72).

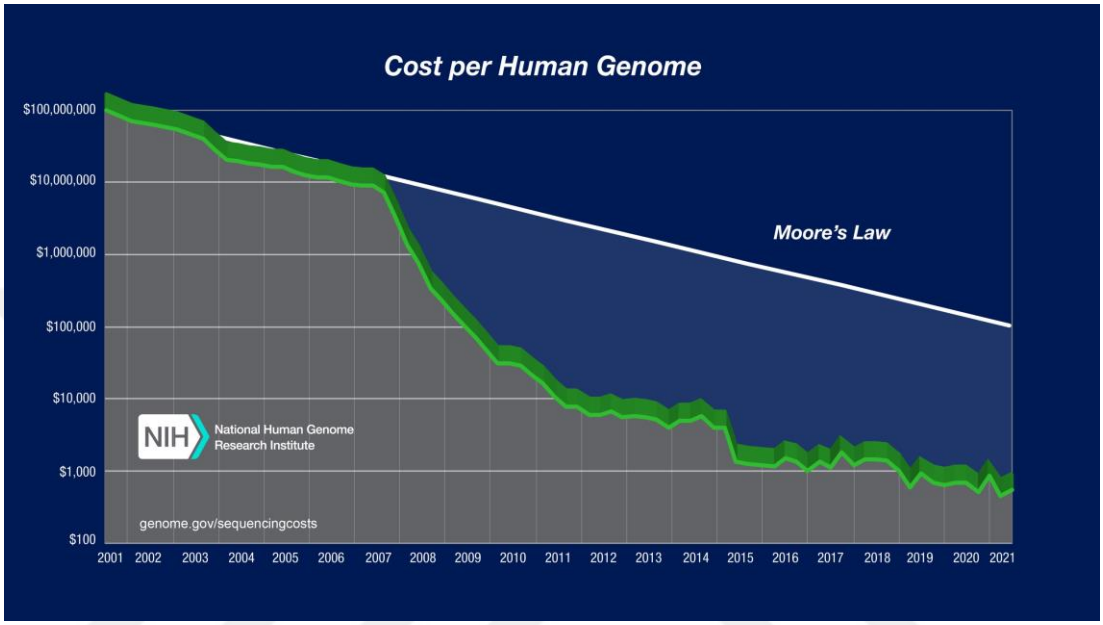

Figure 1: Sequencing costs per human genome between 2001-2021.

WGS can now be performed directly from single bacterial colonies grown under routine diagnostic conditions without requiring subculture to obtain sufficient DNA for sequencing. Recent advances in sample preparation have enabled us to perform WGS directly from bacterial colonies (73,74). Additionally, benchtop sequencers offer sufficient speed to substitute for research and retrospective surveillance as primary diagnostic tools (73,75,76).

In the context of global public health emergencies like tuberculosis (TB), which remains a major threat, WGS has become a vital tool for drug development due to its ability to rapidly identify resistance mechanisms (73,77,78). Additionally, WGS has become an increasingly important tool during clinical trials, in addition to being used to design clinical trials. It is increasingly being used to differentiate exogenous reinfection from the primary infection, which is important for evaluating the efficacy of drug treatment regimens (73,79,80). Surveillance is at the core of infectious disease

control, and molecular resolution provided by WGS has highlighted the importance of antibiotic-resistant organisms spreading internationally (81). The first organism for which this was accomplished was MRSA (82).

2.3. Antimicrobial Resistance & Machine Learning

In several publications (8, 9, 22-42), antimicrobial resistance has been predicted to be accurate for many bacteria based on the genome sequences. Despite the lack of metadata data, such as MIC information, there are fewer publications on AMR phenotype prediction than expected. It is common to share MIC data as classes (R for resistant, I for intermediate, and S for sensitive) but not as concentrations.

AMR in *Streptococcus pneumoniae* (9, 24, 32, 33, 34), *Clostridium difficile* (24), *Pseudomonas aeruginosa* (8, 24), *Mycobacterium tuberculosis* (9, 22-27), *Escherichia coli* (22, 28, 29), *Salmonella enterica* (22), nontyphoidal *Salmonella* (30), *Staphylococcus aureus* (9, 22, 31), *Acinetobacter baumannii* (9), *Actinobacillus pleuropneumoniae* (35), *Elizabethkingia* (36), *Klebsiella pneumoniae* (31, 37), *Campylobacter jejuni* (31), and *Neisseria gonorrhoeae* (31) have been studied in publications. Various k-mer counts have been used in machine learning models (e.g., 8-mers to 11-mers (31), 10-mers (37), 15-mers (30), 31-mers (9, 24), AMR genes (38), SNPs (23, 26, 27, 42), and a combination of these (8, 28, 36). Valizadeh Aslani et al. (31) used amino acid k-mers as features in their recent study and found successful results.

2.4. Streptococcus pneumoniae

It has long been known that *Streptococcus pneumoniae* is one of the bacteria most associated with AMR. Infections such as pneumonia and meningitis are caused mainly by *S.pneumoniae*, a gram-positive human pathogen. Acute otitis media and sinusitis are often caused by this bacterium in childhood (13-15). Geographic location, genetic background, socioeconomic status, and immune status all influence the incidence of invasive pneumococcal disease (14).

In the WHO's opinion, pneumococcal infections play a significant role in global public health, and approximately one million children die yearly because of pneumococcal disease (16).

In 2001, Tettelin et al. published the complete annotated 2.16 Mbps genome of *S.pneumoniae* encoding 2236 predicted proteins (15,83). DNA with a circular structure enclosed by covalent bonds and often accompanied by small cryptic plasmids. Strain TIGR4, one of the original pneumococcal genomes to be sequenced, contains 2236 open reading frames, two-thirds of which have predicted genes assigned to them (83). Approximately 1553 genes form the core set of essential genes for viability, 154 genes constitute the complement of virulence genes, and 176 genes maintain a non-invasive phenotype (14).

Most of the *S.pneumoniae* isolates are resistant to antimicrobial drugs like β-lactams, fluoroquinolones, and macrolides (19). When we checked the main targets of penicillin, we could see the penicillin-binding proteins (PBPs), and for a long-time, penicillin has been the critical choice for treating infections (13,17). β-lactams bind to enzymes essential for bacterial cell wall synthesis and reducing peptidoglycan synthesis. The primary resistance mechanism is the mutation of PBPs to reduce their affinity for antimicrobials (17,18). Pneumococci resistant to beta-lactams are caused by mosaic *PBP* genes that display a reduced affinity for the antimicrobial (20). We can list six PBPs. These are 1a, 1b, 2x, 2a, 2b, and PBP3. The mutation in PBP1a in tandem with PBP2b and PBP2x increases penicillin MIC, while the amino acid changes in PBP2b and PBP2x result in penicillin resistance (20).

During the same time, penicillin resistance spread, and macrolide-resistant pneumococci also increased by inhibiting bacterial protein synthesis and macrolides binding to 23S ribosomal RNA (20). Macrolide resistance in S. pneumoniae has two main mechanisms, efflux of antimicrobials through mef and erm genes or modification of the target site (20). An acquired 23S RNA methylase gene, erm(B), is a major resistance factor in pneumococci, and post-transcriptional modification of 23S occurs after gene expression, reducing the affinity of the macrolide for 23S (20).

A second major mechanism of macrolide resistance is the acquisition of mef genes, which encode an active pump. As a result of the interactions between tet(M) and other tet genes, Streptococcus pneumoniae is highly susceptible to tetracycline inhibition, primarily due to ribosomal protection (21). Streptococci have a less common efflux-mediated mechanism for reducing intracellular tetracycline concentration to subtoxic levels. Tet(K) or Tet(L) encode membrane proteins that participate in this mechanism (21). Tetracyclines inhibit the growth of bacteria via binding to the 30S subunit of the ribosome, and pneumococcal resistance occurs with ribosomal protection (21).

2.5. Staphylococcus aureus

As one of the most common pathogens, Staphylococcus aureus is a major cause of infection in hospitals and communities. S. aureus causes minor skin infections, osteomyelitis, and bacteremia, as well as fatal cases of pneumonia (84,85). In recent years, virtually all antimicrobial drugs used in treatment have developed resistance mechanisms in S. aureus strains. It was reported in 1961 that Jevons isolated an MRSA strain whose resistance was governed by a gene that encoded the PBP2a or PBP2′ embedded in the chromosomal element of methicillin-sensitive S. aureus (84,87). A growing number of places around the globe have reported MRSA as the most common pathogen resistant to antibiotics, including Europe, Africa, and Asia (84,88,89).

There has been an increase in the number of drug-resistant S. aureus infections in the past few years, which makes it more challenging to treat them (84). The endogenous resistance mechanism consists of efflux systems, excessive production of β-lactamase, and outer membrane permeability. Multidrug resistance is related to active drug efflux systems (84,91). NorA, Smr, and QacA are the multidrug pumping proteins on the *S.aerous* cell membrane (93)

Peptidoglycan is an important element of the bacteria cell wall. During the synthesis of peptidoglycan, beta-lactam antibiotics target some enzymes. Bacterial chromosomal genes encode beta-lactamase, an enzyme that catalyzes the hydrolysis of beta-lactam antibiotics (86,94). Specifically, beta-lactams inactivate proteins

known as PBPs, which cause bacterial cell death when inactivated. Currently, research indicates that beta-lactam antibiotics affect bacteria mainly by binding to the penicillin-binding protein, which represses the production of cell wall mucin, damages the cell wall, and leads to bacterial expansion and death. Second, activating the bacteria's autolytic enzyme activity leads to autolysis and death (95).

The bacteria's energy metabolism is affected by lowering the cell membrane's permeability, resulting in reduced drug absorption (90). Several genetic mutations have been shown to lead to drug resistance in Staphylococcus aureus, such as mutations that alter DNA gyrase or reduce outer membrane proteins (96).

A ribosomal RNA methylase modification causes clindamycin and erythromycin resistance (97). There is an overlap in mechanisms of action between macrolides, erythromycin, lincosamides, and streptogramins B. Each target is a four-nucleotide fragment of 23S rRNA in the 50S subunit (86). Typically, resistance to tetracyclines in S. aureus involves active removal of the antibiotic from the bacterial cell and ribosomal protection from the antibiotic. Tetracyclines inhibit protein synthesis by interfering with the 30S subunit of the ribosome. Another antibiotic fusidic acid inhibits protein synthesis by interaction with elongation factor G. A bactericidal antibiotic, aminoglycosides, interferes with the ribosome's 30S subunit to inhibit protein synthesis. There are a variety of mechanisms by which S. aureus may be resistant to aminoglycosides, including the synthesis of transferases that modify the aminoglycoside molecule (98,99) and the lack of enzymes, which are needed for active transport of aminoglycosides (100). As a result of mupirocin resistance, S. aureus produces isoleucyl-tRNA synthetase that is not sensitive to mupirocin and inhibits protein synthesis (86). The mupA gene encodes Isoleucyl-tRNA synthetase. Fluoroquinolones inhibit the activity of topoisomerase II and topoisomerase IV enzymes. This enzyme is responsible for DNA respiralization. Fluoroquinolone resistance in *S. aureus* is caused by mutational changes in the *gyrA, gyrB,* and *parC, which are* genes (86).

2.6. Enterococcus faecium

Various organisms, including humans, are hosts to enterococci, gram-positive facultative anaerobes (102, 103). In terms of gram-positive nosocomial infections, enterococci are the leading cause, followed by staphylococci (102). Although enterococci were once considered a genus with minimal clinical impact, they are now regarded as important organisms due to the emergence of multidrug-resistant strains of Enterococcus faecium and Enterococcus faecalis, which cause about 12% of all nosocomial infections in hospitals (104,105,106). In addition, these bacteria can acquire antibiotic resistance by transferring plasmids and transposons, chromosomal exchanges, or mutations, making therapeutic measures challenging (104).

According to population genetics and genomics, E. faecium has two distinct subpopulations (107). Commensals of the gastrointestinal tract constitute the first subpopulation, which does not generally infect patients. A second subpopulation of E. faecium lineages in hospitalized patients is responsible for nosocomial outbreaks and opportunistic infections (107). Amplified fragment length polymorphism, a fingerprint-based typing method, was first used two decades ago to recognize these distinct subpopulations (108). Using sequence-based methods, such as multi-locus sequence typing and whole genome sequencing, these different subpopulations of E. faecium were further identified and characterized (109,110). Clade A and B are these two populations.

The most effective beta-lactams against enterococci are ampicillin and penicillin, which prevent the production of peptidoglycan, the building block of the bacterial cell wall and a vital element required for bacterial survival (102). In general, penicillin-binding proteins can be divided into two groups: class A, which are bifunctional enzymes with both D, D-transpeptidase, and transglycosylase activity, class B, which only have the transpeptidase domain. Six putative PBP genes were discovered by subsequent genomic research of E. faecalis and E. faecium, three of which are class A, ponA, pbpF, pbpZ, and class B are pbp5, pbpA, pbpB (111). Ampicillin resistance can occur in another way, like the antibiotic is inactivated by beta-lactamase through the beta-lactam ring (112).

E. faecium has successfully adapted to the environment in a nosocomial setting and developed resistance against glycopeptides over the past 20 years, making it a widespread nosocomial pathogen (107,113). The van operons on mobile genetic elements contain the resistance genes against glycopeptides. The operon contains regulatory genes that control the expression of ligase genes resistant to glycopeptides, the vanA, and vanB genes being the most common (107). One of the clearest examples of this adaptability is the acquisition of the gene encoding vancomycin resistance. Acquisition of resistance to glycopeptides is an important milestone in the enterococci history into highly resistant organisms. From the first detection of vancomycin-resistant enterococci (VRE) in 1988 to the end of the first decade of this century, the glycopeptides vancomycin and teicoplanin bind to the terminal D-alanine-D-alanine of peptidoglycan, and inhibition of cell wall synthesis (114). There are nine vancomycin resistance clusters described. These are *vanA, vanB, vanC, vanD, vanE, vanG, vanL, vanM* and *vanN* (107, 115). Genes encoding includes three groups: TCS, enzymes for synthesizing peptidoglycan; cell membrane, which regulates a cell; and DAP, which is a lipopeptide antibiotic to targets the cell membrane (107).

Enterococci exhibit inherent resistance to aminoglycosides because of two main factors. Antibiotics are poorly uptaken, require high concentrations to facilitate entry into the intracellular space, and are inactivated by covalent modification of the hydroxyl or amino groups of aminoglycoside molecules by natural enterococcal enzymes. Sterically hindered and binds to its ribosomal target (107).

2.7. CLSI & EUCAST MIC Breakpoints for Streptococcus pneumoniae, Staphylococcus and Enterococcus

The MIC breakpoints of the antibiotics used in our study are shown in Table 1 (116, 117).

Table 1: MIC Breakpoints for *Stretococcus pneumonia*, *Staphylococcus* and *Enterococcus* according to CLSI and EUCAST.

Bacteria	Antibiotic	Standards	S	I	R
Streptococcus pneumoniae	Penicillin	CLSI	≤ 0.06	0.12-1	≥ 2
		EUCAST	≤	-	≥
	Erythromycin	CLSI	≤ 0.25	0.5	≥ 1
		EUCAST	≤	-	≥
	Tetracycline	CLSI	≤ 1	2	≥ 4
		EUCAST	≤	-	≥
Staphylococcus	Erythromycin	CLSI	≤ 0.5	1-4	≥ 8
		EUCAST	≤ 1	-	> 2
	Tetracycline	CLSI	≤ 4	8	≥ 16
		EUCAST	≤ 1	-	>2
	Gentamicin	CLSI	≤ 4	8	≥ 16
		EUCAST	≤ 2	-	> 2
	Ciprofloxacin	CLSI	≤ 1	2	≥ 4
		EUCAST	≤ 0.001	-	> 1
Enterococcus	Ampicillin	CLSI	≤ 8	-	≥ 16
		EUCAST	≤ 4	-	> 8
	Vancomycin	CLSI	≤ 4	8-16	≥ 32
		EUCAST	≤ 4	-	> 4

3 MATERIALS AND METHODS

The overall pipeline of the study is presented in Figure 2. To predict the MIC class for each antibiotic, feature generation, feature selection, and various ML models were trained. The performances were compared using several metrics. The following subsections explain the details of data, feature generation, feature selection, model training, and evaluation. All analyses were performed in R version 4.0.2 (http://www.R-project.org). The R scripts used in this study are available on GitHub at https://github.com/denizecek/AMRprediction.

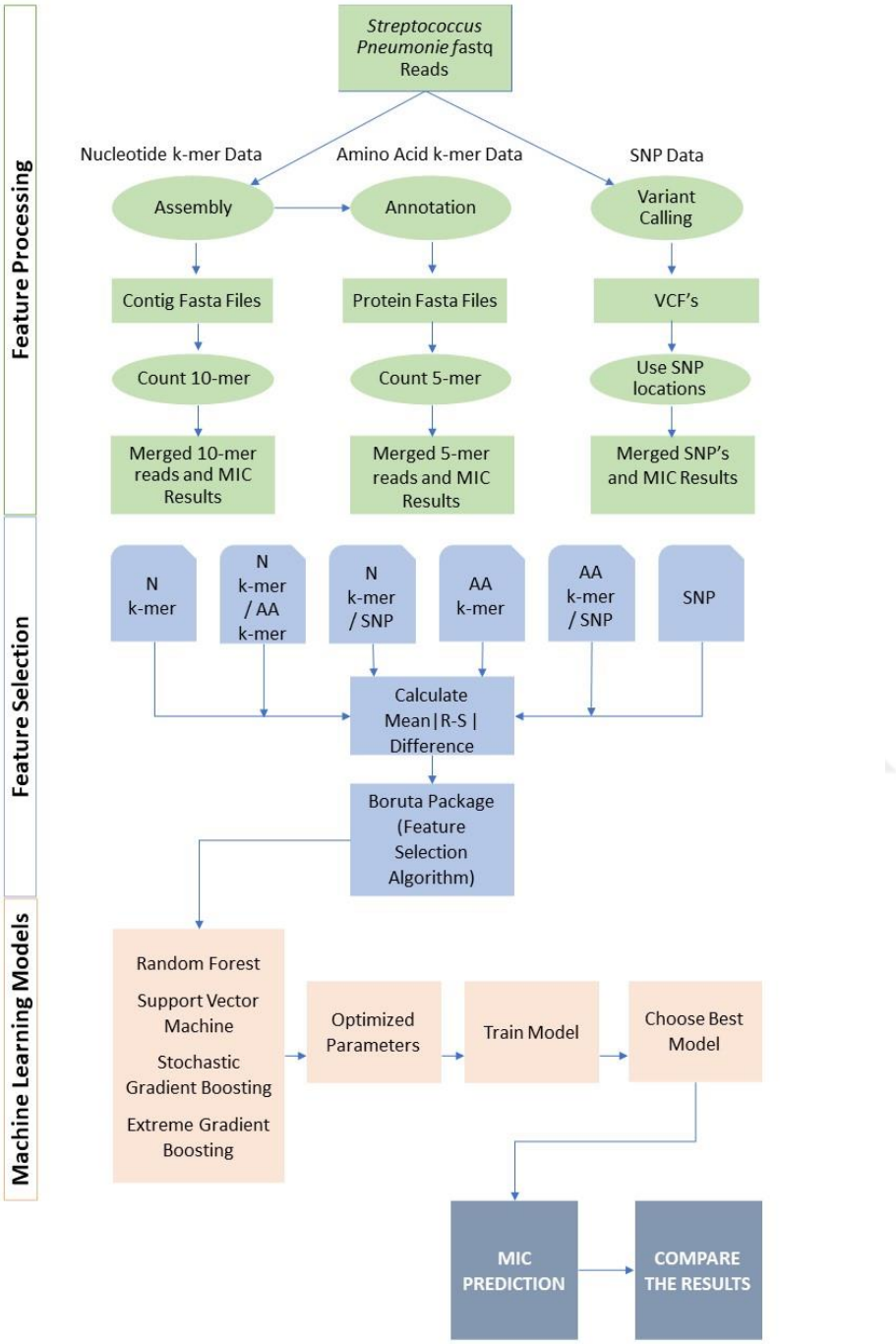

Figure 2: Overall pipeline for all feature extraction and classification approaches for *Streptococcus pneumoniae*

3.1. Datasets and Pre-processing

S. pneumoniae, *S.aerous,* and *E. faecium* metagenomic sequences and the related MIC class information of the antibiotics penicillin, erythromycin, and tetracycline were included in our study. We used three publicly available datasets: 785 S.pneumoniae, 151 S.aerous, and 124 E.faecium strains were downloaded from the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena/) with the project accession code in PRJEB2632 (43), PRJEB2255 (45), PRJEB32286 (60), PRJEB4345 (61). The MIC class information was downloaded from the PATRIC database (46) and PubMLST (118), and the "Resistant" and "Susceptible" classes were matched to the genome data. For each antibiotic, we discarded the "Intermediate" class because these were underrepresented. Table 2, Table 3, and Table 4 present the details of the five datasets, and Appendix 1 contains detailed sample information.

Table 2: *S.pneumoniae* datasets and their sample number of each MIC class.

Antibiotic	**DATASET 1**		**DATASET 2**		**Total**
	S	**R**	**S**	**R**	
Penicillin	453	68	128	49	698
Erythromycin	470	78	17	114	679
Tetracycline	287	36	11	121	455

Table 3: *S.aureus* datasets and their sample number of each MIC class.

	DATASET 3		
Antibiotic	Susceptible	Resistant	Total
Erythromycin	54	97	151
Tetracycline	57	94	151
Gentamicin	102	48	151
Ciprofloxacin	53	98	151

Table 4: *E.faecium* datasets and their sample number of each MIC class.

	DATASET 4		
Antibiotic	Susceptible	Resistant	Total
Ampicillin	25	99	124
Vancomycin	72	52	124

3.2. Feature Generation and Selection

3.2.1. Nucleotide k-mers

PATRIC assembly service (46) and SPAdes (47) were used for genome assembly. Contigs lengths were less than 500 base pairs, and less than 5-fold coverage was discarded. To obtain the frequencies of the 10-mers, the contigs were separated into 10-mers using the R "kmer" package (48). We used antibiotic MIC classes as labels and k-mer counts as features to classify AMR. In this study, we used a 10-mer length instead of a longer k-mer length to reduce matrix size. Memory limitations prevented us from using longer k-mers, and lower initial accuracy prevented us from trying shorter k-mers. This can avoid potential complications such as repeat regions by converting k-mer counts to reflect the presence of "1" or "0" of each k-mer in each genome.

Fitting the machine learning model on a large dataset may be challenging because of high computational costs and long processing times. Large datasets are also less efficient than optimal feature datasets for machine learning. We selected the first feature based on the absolute mean difference between resistant and susceptible samples of each 10-kmer in our dataset since 1,048,577 10-mers were included in our datasets.

Based on the mean difference between the resistant and susceptible samples, features with a difference of 0.4 were selected in *S. pneumoniae*. We aim to reduce the number of features to below 10000 by choosing 0.4. The cutoff value of 0.2 retains

68365 features for penicillin, whereas that of 0.3 includes 28410 features. For penicillin, 9613 features were left when 0.4 was selected before the next feature selection step, 1433 features were left for erythromycin, and 65 were left for tetracycline.

When we worked with the *S.aureus* dataset, the 0.4 cutoff value retained 42 variables for erythromycin. We chose a 0.2 cutoff value and 7425 features left before the final feature selection. 0.2 cutoff value is chosen for gentamicin, and 6133 features are selected. When we tried a 0.2 cutoff for tetracycline and ciprofloxacin, 18360 and 88319 features were left, respectively. Because these feature numbers are greater than our goal (less than 10000). For the final decision, a 0.3 cutoff was selected for tetracycline, 1434 features remained, a 0.4 cutoff was selected for ciprofloxacin, and 8310 features remained.

Finally, for the *E.feacium*, we tried a 0.2 cutoff value for ampicillin and vancomycin. 196,232 and 10,236 features remained. Based on the mean difference between the resistant and susceptible samples, features with a difference of 0.5 were selected in ampicillin. The cutoff value of 0.4 retains 10956 features, whereas that of 0.5 retains 1696 features. For vancomycin, 2916 features were left when 0.3 was selected.

3.2.2. SNP

Single nucleotide polymorphisms (SNPs) are another set of ML model training features. In the PATRIC variant calling service (46), BWA-mem (49) and SAMtools (50) were used. Bcftools (51) was used for filtering with DP > 20 and qual > 50 parameters. Samples were rows and SNP positions compared to the reference genome were columns. The number of samples containing mutations at that site is 1, and the number of samples without mutations is 0.

S.pneumoniae TIGR4 was our reference genome for SNP calling for S.pneumoniae datasets. 151,733 SNPs were used. Comparing the 10-mer features with the SNP features, we found that the number of SNPs was much lower (151723), and

our cutoff value was lower. We calculated the absolute mean difference between resistant and susceptible samples for each SNP and filtered it to 0.05 for all three antibiotics, and we removed all features except this difference. For penicillin, 1278 features remain out of 151733 features. For erythromycin, 6365 features stay out of 145368 features, and for tetracycline, 3915 features remain out of 147815 features.

Staphylococcus aureus subsp. aureus NCTC 8325 (NCBI Reference Sequence: NC_007795.1) was used as a reference in the variant calling for the *S.aureus* dataset. 1023 SNP determined from variant calling. We did not calculate the absolute mean difference for the *S.aureus* dataset because of the low feature number. For the feature selection, only the Boruta package was used.

Enterococcus faecium DO (Taxonomy ID: 333849) was our reference genome for SNP, calling for the *E.faecium* dataset. 17,737 SNPs were found from variant calling. We chose a 0.05 cutoff value and 14,744 features left before the final feature selection in ampicillin and 10,006 features left in vancomycin.

3.2.3 Amino acid k-mers

We built a model using amino acid k-mers to predict the MICs following the methods previously described by Valizadeh Aslani et al. (31). For annotation analysis, we used The Genome Annotation Service in PATRIC (46) and the RAST tool kit (RASTtk) (52). Protein fasta files were downloaded from the PATRIC, and amino acid k-mers were counted using the "kmer" R package (48). Counting longer k-mer sequences was simplified using the Dayhoff-6 alphabet (53) in the package. 5-mers were selected as a feature. The k-mers of shorter amino acids were not chosen because of the lower accuracy, and the k-mers of longer amino acids were not chosen due to memory constraints. In this case, the pre-elimination used in other feature extraction methods was not used because the number of features is less than 10000.

3.2.4. Combinations of features

Another feature extraction method was tested by combining nucleotides of 10mers, amino acids of 5mers, and SNPs in binary combinations. Our selection of features and combination of the best features was based on sections 2.2.1, 2.2.2, and 2.2.3.

3.2.5. Boruta

The final feature selection was performed using the R package Boruta. An algorithm for selecting features using machine learning (54). In this method, Random Forest is encapsulated inside a wrapper method. By creating a shuffled copy of all features, the algorithm adds randomness to the data set. This type of feature is called a Shadow Feature. The algorithm determines the importance of shadow features using the Z-score and decreases accuracy as the shadow features are merged with the original features. By comparing the original feature with the shadow feature, the Boruta algorithm determines whether the original feature is more important. Significant features are retained at every iteration, while unimportant ones are continually removed. This process is repeated until all features have been approved or rejected (54).

3.3. Machine Learning Models for AMR Classification

3.3.1. Model Training

We included all isolates categorized as either "resistant" or "susceptible" to each antibiotic due to the small number of intermediate samples. Single nucleotide polymorphisms (SNPs), nucleotide 10-mers, amino acid 5-mers, and their paired combinations with the presence or absence of information as features were included in the genomic data. For each drug and feature type, we divided our data into a training set of 80% of the resistant and susceptible isolates, respectively. The remaining 20% were equally assigned to a test and validation set. For *S.aerous* and *E.faecium* models,

we did not divide the remaining 20% and used all of them as a test set because the number of samples is less than *S.pneumoniae*. In cross-validation, machine learning model parameters were optimized on the validation set while the test set obtained an independent performance estimate. 10-fold cross-validations used to assess the accuracy and sensitivity of the models used in this study. By using cross-validation, the matrix is partitioned into ten equal parts, each containing the same number of antibiotic-MIC combinations. Eight parts are used for training; one is used for validation, and one is for testing. The validation set is used to avoid overfitting. Accuracy, Kappa, and F1 score are used to compare algorithms. The harmonic mean of the sensitivity and predictive value for a particular class is an F1 Score, and Kappa is Cohen's Kappa statistic averaged across the resampling results. R version 4.0.2 (http://www.R-project.org) is used for analyses.

Four different classifiers were built. Random forest (RF), support vector machine (SVM), stochastic gradient boosting (GBM), and extreme gradient boosting (XGBoost) classifiers were used to detect penicillin, erythromycin, and tetracycline resistance in *S. pneumoniae,* erythromycin, tetracycline, gentamicin, and ciprofloxacin resistance in *S.aureus,* ampicillin and vancomycin resistance in *E.faecium*. For RF, three different models were tested. Model 1 (M1) used the default parameters for each parameter. For model 2 (M2) random search was done. Finally, for model 3 (M3) grid search was done. When we worked with the SVM classifier, Linear model with c parameter, polynomial kernel, and radial kernel functions, our three models, M1, M2, and M3, respectively. For the Gradient Boosting and XGBoost model, the first model, M1, worked with Default Parameters. In the second model, M2, 10-fold cross-validation and two repeats were applied. Using a grid search to tune our important hyperparameters was our M3 model. We compared the performance of different classifiers. The models with the highest F1 score among all created models were compared.

3.3.2. Hyperparameter Tuning

The "randomForest" R package was used for Random Forest models (55), which can deal with many features and contains two parameters, ntree and mtry. Ntree is several trees to grow, and this should not be a small number to be sure that every input row is predicted; the default is 500 trees. Mtry is some variables randomly sampled as candidates at each split, and default values for classification are sqrt(p), p is some variables. Three different models were tested for RF. First, we used the recommended defaults for each parameter (mtry and ntree). We tried a random search for the second model, a random combination of hyperparameters selected and used to train a model. Because we narrowed down the range for each parameter, a grid search was done for the final model to specify every combination of settings to try.

For the SVM classifier, the "caret" R package was used (56). SVM Linear was preferred for the first model. By default, caret used the SVM linear classifier using a parameter c =1 (default caret using c =1 in the SVM linear classifier), also known as cost and determines the possible misclassifications; for our first model, we tuned parameter c to determine the possible misclassifications, as we know the higher the c value, misclassify the point it became less likely by the algorithm. For the third model, we choose the radial kernel function. In RBF kernels, there are two tuning parameters, except the c, there is the sigma parameter (σ), which controls the level of non-linearity introduced in the model. If the sigma is large, the decision boundary will likely be linear (56). For the fourth model, we used a polynomial kernel with three tuning parameters. These are c, scale, which is a positive number for the polynomial scaling factor, and degree, which is a positive number for the polynomial degree.

For the Stochastic Gradient Boosting model, first worked with default parameters. In the second model, 10-fold cross-validation and two repeats were applied. There are four important tunning parameters in the GBM model for a grid search. These are several n.trees, interaction, depth, shrinkage, and n.minobsinnode. The number of trees (n.trees) is the number of gradient-boosting iterations. Increasing n can lower the error, but if you make it higher than the optimal limit, it can cause over-fitting (57). Interaction.depth is several splits it has to perform on a tree, and shrinkage is a learning rate used to reduce the impact of additional trees. When the shrinkage value is close

to 1, this overfit model and too small shrinkage values can cause slow down the learning process. N.minobsinnode is the minimum number of observations in tree terminal nodes. We tried the tuning parameters again at the intervals we determined according to the parameters selected in our previous two models. "gbm" and "caret" R packages were used for the gradient boosting models (56,57).

For XGBoost models, the "xgboost" R package was used (58). There are seven tunning parameters in the XGBoost model. These are shown in table 3. Using default parameter choices, we created our first model. The second model was tested for stability using 10-fold cross-validation and two repeats. We were using a grid search to tune our important hyperparameters from XGBoost.

Table 5: XGBoost parameters and their default values.

	XGBoost Parameters	**Default**
nrounds	The number of decision trees in the model	
eta	After boosting step, we can get the weights of new features, eta shrinks the feature weights to make the boosting process more conservative.	0.3
gamma	Minimum loss reduction is required to make a partition on a leaf node of the tree. When we make gamma larger, the algorithm will be more conservative.	0
max_depth	Maximum depth of a tree. Increasing the value of max depth can make the model more complex.	6
min_child_weight	The minimum sum of instance weight.	1
subsample	Subsample ratio of the training instances.	1

colsample_bytree	Subsample ratio of columns when constructing each tree.	1

4 RESULTS

Two different feature selection methods were used before training our models. The first was the absolute mean difference between resistant and susceptible samples of each dataset; then, the Boruta package was applied to our datasets. According to Boruta's results in this study, we keep important features (Table 6) and use them for our models. If the important feature number is less than 50, we keep the tentative features also. Figure 3 and Figure 4 show the importance, classifier run, and attributes. The green color shows important attributes, the red color is unimportant, and the yellow color shows tentative features.

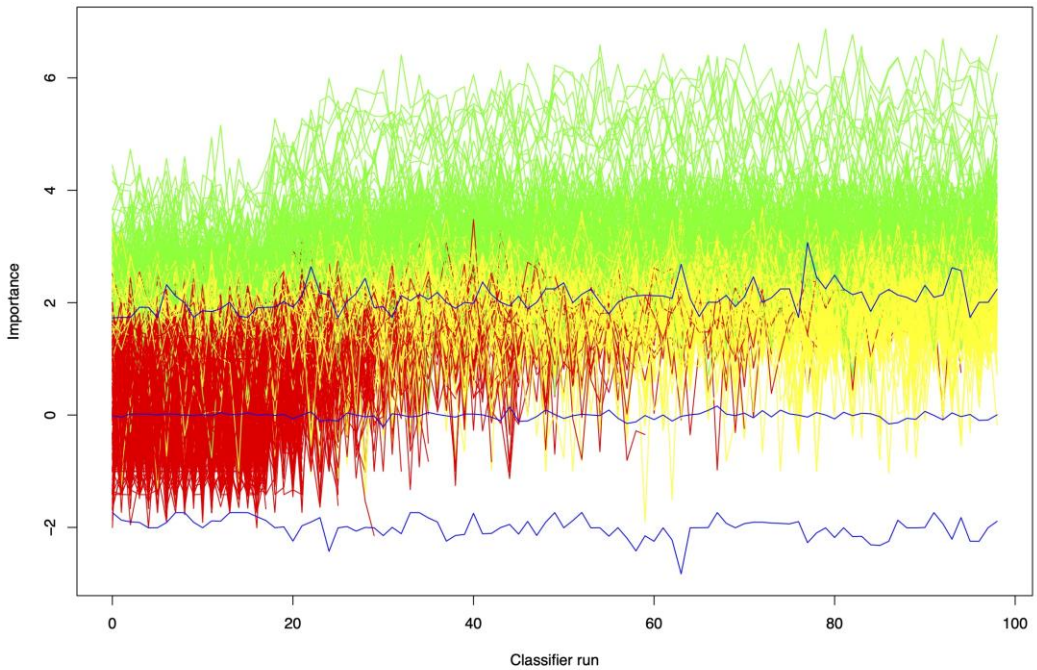

Figure 3: Importance and classifier run results from S.aureus 10-mer (nucleotide kmer) tetracycline feature selection results graph.

Bacteria	Antibiotic	Nucleotide 10-mer Boruta Results		Amino Acid 5-mer Boruta Results		SNP Boruta Results	
		Imp	Ten	Imp	Ten	Imp	Ten
S. pneumoniae	Penicillin	129	220	109	259	58	25
	Erythromycin	256	246	164	102	89	47
	Tetracycline	58	4	130	34	74	56
S. aureus	Erythromycin	79	110	40	126	5	17
	Tetracycline	132	144	61	193	11	16
	Gentamicin	95	73	40	79	8	8
	Ciprofloxacin	47	179	46	144	22	19
E.faecium	Ampicillin	11	14	9	50	19	103
	Vancomycin	8	55	12	44	5	6

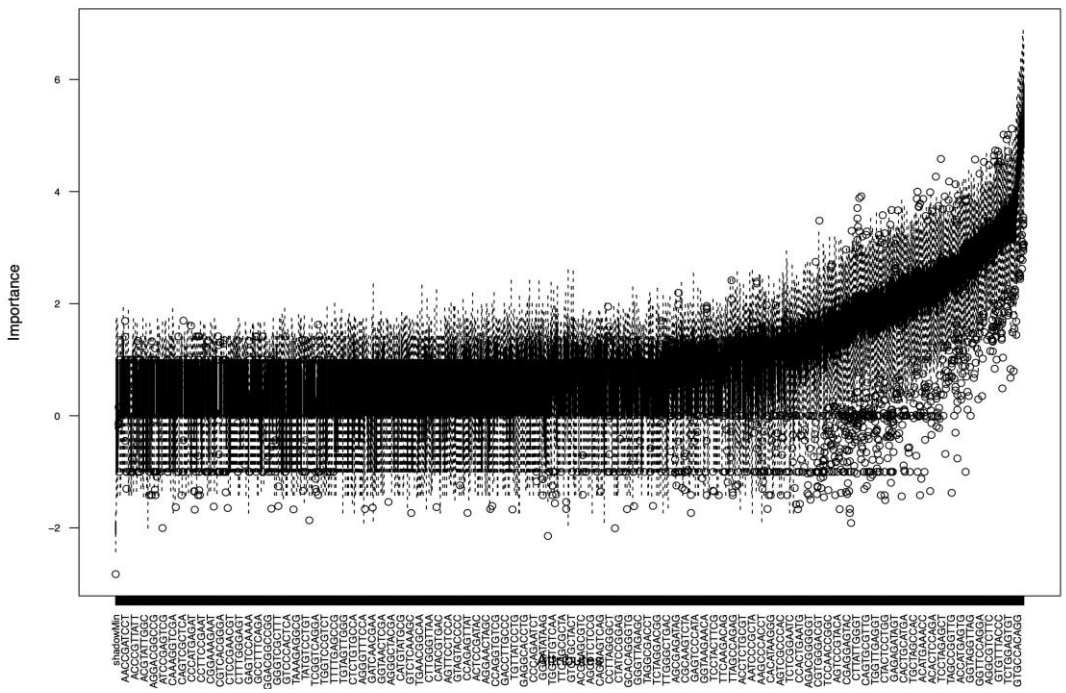

Figure 4: Importance and attributes results from *S.aereus* 10-mer (nucleotide kmer) tetracycline feature selection results graph.

We trained several machine learning classification methods on features individually and together to predict antibiotic susceptibility or resistance. To measure overall performance, we calculated the accuracy, sensitivity, and specificity of resistance (R) and susceptibility (S) assignment and the F1-score based on a classifier trained on a set of specific data types. Figures A1, A2, and A3 in Appendix 2 show the F1 Score, Kappa, and accuracy results for each antibiotic in *S.pneumoniae*. Figures A4, A5, A6, and A7 show results of *S.aureus*, and A8 and A9 are results of *E.faecium*.

For hyperparameter tuning, the best parameters selected for each antibiotic in the final models of *S.pneumoniae* are indicated in Table A6 (Appendix 2). *S.aureus* selected parameters were shown in Table A7, and *E.faecium* in Table A8. The default for each parameter model is M1. Random search is M2, and for the final model, M3, a grid search was done.

4.1. Streptococcus pneumoniae Results

With samples, five repeats of this setup were run to optimize penicillin parameters and average performance estimates. The machine learning classifiers performed almost equally well predicting penicillin susceptibility and resistance when the six input data types were used (k-mer, AA k-mer, SNPs, k-mer/AA k-mer, and SNP/AA k-mer). When feature inputs were compared, it was seen that nucleotide 10-mer F1 scores were higher than other inputs. It was found that amino acid 5-mer and AA k-mer / k-mer combinations were also inputs that contributed to high F1 scores when different feature inputs in penicillin were compared. SNP-included features gave lower accuracy and F1 Score results than other feature inputs. Table 7 and Figure 5 are shown the comparisons of the model results in penicillin. Random Forest had the highest F1 score, with 0.95. This model had an accuracy of 0.98. Regarding F1-Score, SVM was the second-best option, followed by XGBoost, which was close to RF. XGBoost had a significantly lower computational complexity than random forest, however. Penicillin resistance and susceptibility were predicted based on RFs using a 10-mer nucleotide feature with 0.95 sensitivity and 0.99 specificity. SVM linear predicts high predictive sensitivity and specificity values of 0.97 and 0.99, GBM predicts a

sensitivity of 0.97 and a specificity of 0.98, and XGBoost predicts a sensitivity of 0.96 and a specificity of 0.97 (Table 7). All penicillin results, including grid search, are shown in Appendix 3, Table A9.

Table 7: F1 Score, kappa, accuracy, sensitivity and specificity results of *S.pneumoniae* penicillin results.

Algorithm	**Input**	**F1-score**	**Kappa**	**Accur.**	**Sensit.**	**Specific.**
Random Forest	k-mer	0.952	0.942	0.983	0.952	0.989
Support Vector Machine	k-mer	0.951	0.941	0.983	0.967	0.986
Stochastic Gradient Boosting	k-mer	0.943	0.931	0.980	0.966	0.983
Extreme Gradient Boosting	k-mer	0.914	0.896	0.969	0.964	0.973
Random Forest	AA k-mer	0.907	0.889	0.969	0.964	0.970
Support Vector Machine	AA k-mer	0.925	0.910	0.975	0.965	0.976
Stochastic Gradient Boosting	AA k-mer	0.907	0.889	0.969	0.980	0.957
Extreme Gradient Boosting	AA k-mer	0.898	0.878	0.966	0.963	0.967
Random Forest	SNP	0.839	0.810	0.950	0.978	0.945
Support Vector Machine	SNP	0.859	0.833	0.955	0.960	0.954
Stochastic Gradient Boosting	SNP	0.821	0.789	0.944	0.938	0.945
Extreme Gradient Boosting	SNP	0.839	0.810	0.950	0.968	0.902
Random Forest	SNP/AA k-mer	0.912	0.894	0.969	0.919	0.980
Support Vector Machine	SNP/AA k-mer	0.919	0.903	0.972	0.934	0.980
Stochastic Gradient Boosting	SNP/AA k-mer	0.902	0.882	0.967	0.932	0.973
Extreme Gradient Boosting	SNP/AA k-mer	0.918	0.901	0.972	0.949	0.977
Random Forest	SNP/k-mer	0.873	0.846	0.956	0.873	0.973
Support Vector Machine	SNP/k-mer	0.891	0.867	0.961	0.877	0.980
Stochastic Gradient Boosting	SNP/k-mer	0.898	0.876	0.964	0.891	0.903
Extreme Gradient Boosting	SNP/k-mer	0.896	0.874	0.964	0.980	0.977
Random Forest	AA k-mer/k-mer	0.946	0.934	0.981	0.924	0.993
Support Vector Machine	AA k-mer/k-mer	0.945	0.933	0.981	0.938	0.990
Stochastic Gradient Boosting	AA k-mer/k-mer	0.912	0.894	0.969	0.919	0.980
Extreme Gradient Boosting	AA k-mer/k-mer	0.903	0.883	0.967	0.918	0.977

We used 1888 Erythromycin samples for our models, with SNP/AA k-mer with GBM achieving the best results in terms of accuracy and F1 score (F1-Score 0.93, Accuracy 0.96). Apart from the SNP input alone, all classifiers performed equally well for erythromycin AMR prediction. In terms of accuracy and F1-Score, XGBoost and random forest were the second and third-best options. In erythromycin analysis,

combinations of SNP/AA k-mer, SNP/k-mer, and AA k-mer/k-mer F1 scores were higher than other inputs. The highest F1 scores were obtained from amino acid 5-mer and SNP combinations. Table 8 shows the comparisons between the models. When we compare the F1-Score, the second-best option is random forest. Using the SNP/AA k-mer features combination; we could correctly predict erythromycin resistance and susceptibility based on GBM with a specificity of 0.97 and a sensitivity of 0.93. All grid search erythromycin results are shown in Appendix 3, Table A10.

Table 8: F1 Score, kappa, accuracy, sensitivity and specificity results of *S.pneumoniae* erythromycin results.

Algorithm	**input**	**F1-score**	**Kappa**	**Accuracy**	**Sensitivity**	**Specificity**
Random Forest	k-mer	0.906	0.867	0.944	0.910	0.958
Support Vector Machine	k-mer	0.908	0.868	0.944	0.896	0.965
Stochastic Gradient Boosting	k-mer	0.889	0.842	0.933	0.886	0.954
Extreme Gradient Boosting	k-mer	0.880	0.829	0.928	0.883	0.947
Random Forest	AA k-mer	0.842	0.771	0.901	0.811	0.945
Support Vector Machine	AA k-mer	0.844	0.775	0.904	0.823	0.941
Stochastic Gradient Boosting	AA k-mer	0.854	0.789	0.909	0.826	0.949
Extreme Gradient Boosting	AA k-mer	0.843	0.774	0.904	0.829	0.938
Random Forest	SNP	0.811	0.734	0.891	0.846	0.908
Support Vector Machine	SNP	0.782	0.701	0.880	0.861	0.886
Stochastic Gradient Boosting	SNP	0.796	0.712	0.880	0.814	0.907
Extreme Gradient Boosting	SNP	0.750	0.644	0.893	0.756	0.891
Random Forest	SNP/AA k-mer	0.919	0.885	0.952	0.927	0.962
Support Vector Machine	SNP/AA k-mer	0.894	0.852	0.939	0.924	0.944
Stochastic Gradient Boosting	SNP/AA k-mer	0.929	0.898	0.957	0.929	0.969
Extreme Gradient Boosting	SNP/AA k-mer	0.919	0.885	0.952	0.927	0.962
Random Forest	SNP/k-mer	0.896	0.850	0.936	0.873	0.965
Support Vector Machine	SNP/k-mer	0.877	0.821	0.923	0.838	0.964

Table 9: F1 Score, kappa, accuracy, sensitivity and specificity results of *S.pneumoniae* erythromycin results (continue).

Stochastic Gradient Boosting	SNP/k-mer	0.895	0.849	0.936	0.880	0.961
Extreme Gradient Boosting	SNP/k-mer	0.886	0.837	0.931	0.871	0.957
Random Forest	AA k-mer/k-mer	0.909	0.869	0.944	0.889	0.969
Support Vector Machine	AA k-mer/k-mer	0.899	0.855	0.939	0.887	0.961
Stochastic Gradient Boosting	AA k-mer/k-mer	0.904	0.862	0.941	0.888	0.965
Extreme Gradient Boosting	AA k-mer/k-mer	0.904	0.862	0.941	0.888	0.965

Performance estimates were averaged over five repetitions of this setup with 1520 samples for tetracycline using cross-validation. As for predicting tetracycline resistance and susceptibility, machine learning classifiers performed similarly when input data types included k-mer, AA k-mer, SNPs, k-mer/AA k-mer, SNP/k-mer, and SNP/AA k-mer (values > 0.77). It was found that the AA k-mer / k-mer combination F1-score of tetracycline was higher than that of any other input, while SNP input alone also produced high F1 scores. In Table 9, there is a comparison of models. The random forest was the best model, with a 0.84 F1 score and 0.85 accuracy. In terms of F1-Score and accuracy, GBM and XGBoost performed very close to RF, and they are the second and third-best options. A sensitivity of 0.89 and a specificity of 0.97 are predicted correctly with the AA k-mer / k-mer feature tetracycline. All grid search tetracycline results can be found in Appendix 3, Table A11.

Table 10: F1 Score, kappa, accuracy, sensitivity and specifity results of *S.pneumoniae* tetracycline results.

Algorithm	input	F1-score	Kappa	Accuracy	Sensitvity	Specificity
Random Forest	k-mer	0.813	0.660	0.831	0.853	0.815
Support Vector Machine	k-mer	0.765	0.562	0.782	0.777	0.786
Stochastic Gradient Boosting	k-mer	0.780	0.612	0.808	0.851	0.780
Extreme Gradient Boosting	k-mer	0.771	0.593	0.798	0.830	0.776
Random Forest	AA k-mer	0.790	0.608	0.805	0.804	0.806
Support Vector Machine	AA k-mer	0.777	0.576	0.788	0.772	0.803

Table 11: F1 Score, kappa, accuracy, sensitivity and specifity results of *S.pneumoniae* tetracycline results (continue).

Stochastic Gradient Boosting	AA k-mer	0.794	0.615	0.808	0.805	0.811
Extreme Gradient Boosting	AA k-mer	0.792	0.609	0.805	0.795	0.813
Random Forest	SNP	0.835	0.682	0.841	0.818	0.863
Support Vector Machine	SNP	0.813	0.643	0.821	0.802	0.839
Stochastic Gradient Boosting	SNP	0.817	0.649	0.825	0.804	0.845
Extreme Gradient Boosting	SNP	0.828	0.669	0.835	0.812	0.857
Random Forest	SNP/AA k-mer	0.781	0.587	0.794	0.781	0.806
Support Vector Machine	SNP/AA k-mer	0.779	0.569	0.784	0.751	0.818
Stochastic Gradient Boosting	SNP/AA k-mer	0.774	0.574	0.788	0.774	0.8
Extreme Gradient Boosting	SNP/AA k-mer	0.772	0.562	0.781	0.756	0.805
Random Forest	SNP/k-mer	0.786	0.612	0.807	0.823	0.796
Support Vector Machine	SNP/k-mer	0.789	0.619	0.811	0.829	0.797
Stochastic Gradient Boosting	SNP/k-mer	0.775	0.604	0.804	0.843	0.779
Extreme Gradient Boosting	SNP/k-mer	0.804	0.634	0.817	0.812	0.822
Random Forest	AA k-mer/k-mer	0.841	0.695	0.848	0.829	0.865
Support Vector Machine	AA k-mer/k-mer	0.808	0.636	0.818	0.805	0.830
Stochastic Gradient Boosting	AA k-mer/k-mer	0.819	0.655	0.828	0.813	0.841
Extreme Gradient Boosting	AA k-mer/k-mer	0.817	0.649	0.825	0.804	0.845

In Figure 5, six types of feature inputs and four models of penicillin, erythromycin, and tetracycline are compared.

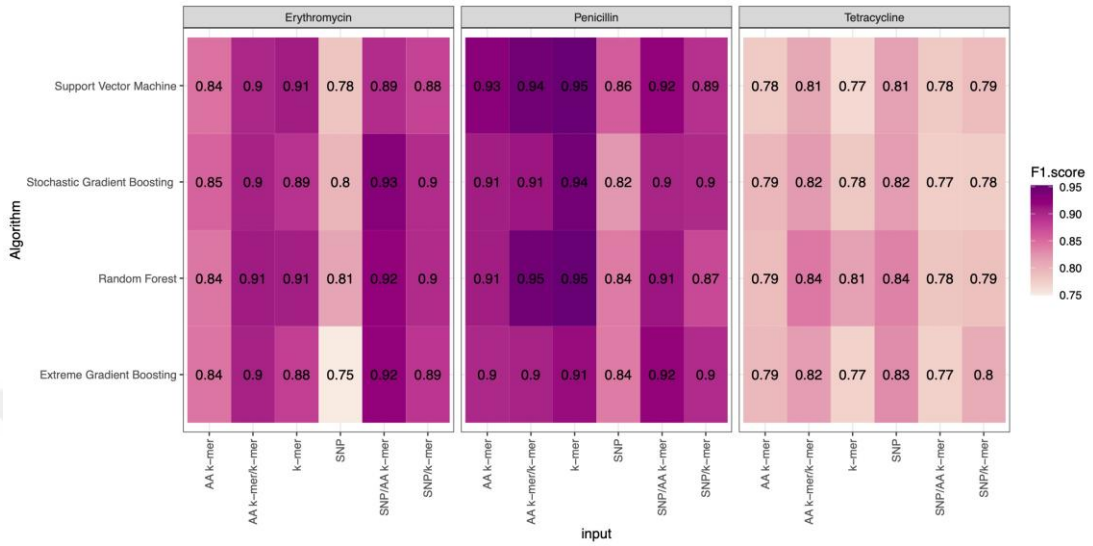

Figure 5: All F1 Score comparisons for penicillin, erythromycin, and tetracycline models are shown in bar graphs.

To simplify the analysis, we also looked for the presence or absence of k-mers rather than k-mer counts. The k-mers with the highest importance values are the most informative for predicting the MICs in a new genome. The models can be used to understand the genomic regions that differentiate MICs by analyzing the feature importance values of each k-mer.

Our next step was to identify k-mers with high-importance values within AMR genes or near AMR genes, to understand the relationship between known AMR genes and the important k-mers selected by each model.

The highest-ranking k-mers from each best model for each antibiotic are shown in Figure 6. In most cases, there was a clear correspondence between the k-mers and known AMR genes. AMR genes related to S. pneumoniae were examined for the top 10 10-mers with the highest importance values, such as penicillin-binding proteins (PBPs), which are involved in cell wall synthesis and are frequently associated with

penicillin resistance (PBP2b, PBP2x, and PBP1a), while macrolide resistance mechanisms are found in Spneumoniae genes such as ermB and mefE, which encode efflux pumps that function. Also, *tetM, tetO,* and *tetK* genes are the most common resistance mechanism to tetracycline in S.pneumoniae.

In S.pneumoniae with penicillin MIC results, the top-ranking kmer is found in the *pbp1a* and *pbp2b* genes. 2nd,3rd,7th,9th, and 10th k-mers are found in *pbp1a* and *pbp2b* genes. The 10th highest ranking k-mer is found in the *pbp2x,* which is the peptidoglycan synthesize gene. For erythromycin, the second top-ranking k-mer is found in the mefE gene. In all 3 cases, the highest-ranking k-mers by random forest results corresponded with known AMR determinants. In erythromycin, k-mers with highest-ranking do not have a known role in AMR. When we checked the tetracycline top 10 features, we could see that the top-ranking k-mers were in *tetK, tetM, tetO,* and *tetS* genes. Figure 6 shows the top 10 important k-mer features for each of the three antibiotics. Table 10 shows the performance rankings of each method based on the average accuracy for all antibiotics.

A) Penicillin B) Erythromycin C) Tetracycline

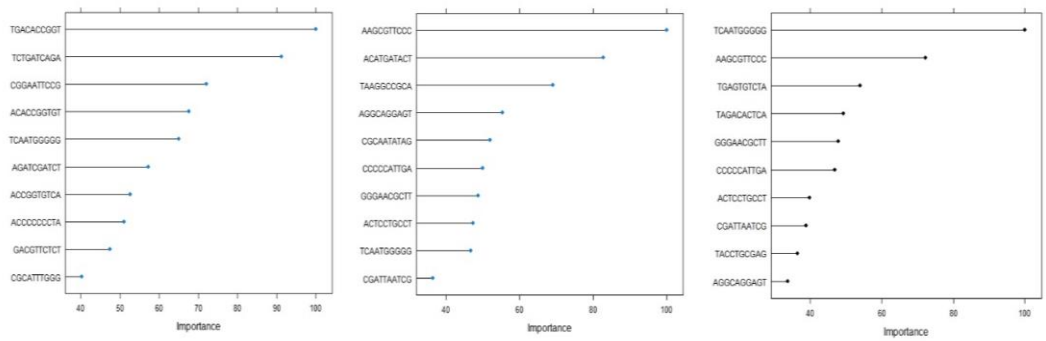

Figure 6: The top 10 important k-mer features for each of the three antibiotics.

Table 12: Performance rankings of each method based on the average accuracy for all antibiotics.

	Penicillin	**Erythromycin**	**Tetracycline**
1	k-mer	SNP/AA k-mer	AA k-mer/k-mer
2	AA k-mer	AA k-mer/k-mer	SNP
3	SNP	k-mer	SNP/k-mer
4	SNP/AA k-mer	SNP/k-mer	k-mer
5	AA k-mer/k-mer	AA k-mer	AA k-mer
6	SNP/k-mer	SNP	SNP/AA k-mer

4.2. Staphylococcus aureus Results

A dataset of 151 samples was used when analyzing the Streptococcus aureus data. A total of four antibiotics were used during training. These are erythromycin, tetracycline, gentamicin, and ciprofloxacin. Although the distribution between the groups was suitable for analysis, a small number of samples caused us difficulty in establishing the model.

This setup was run to optimize the parameters of erythromycin and estimate average performance. When six types of input data were used k-mer, AA k-mer, SNPs, k-mer/AA k-mer, and SNP/AA k-mer, the machine learning classifiers performed almost equally well except for SNP input. Nucleotide 10-mer nucleotide kmer and 5-mer amino acid kmer accuracy were higher than other feature inputs when comparing feature inputs. When different feature inputs in erythromycin were compared, AA k-mer / k-mer combinations contributed to high accuracy. F1 Score results were lower for SNP includes features than for other feature inputs. A comparison of the model results in erythromycin is shown in Table 11, Figure 11, Appendix 4, Table A12, and Figure A4. The highest accuracy was achieved by stochastic gradient boosting with 0.976. Random forest, XGBoost, and SVM were the following best options.

Table 13: Kappa and accuracy results of *S.aureus* erythromycin results.

Algorithm	Input	Accuracy	Kappa
Random Forest	k-mer	0.962	0.916
Support Vector Machine	k-mer	0.974	0.942
Stochastic Gradient Boosting	k-mer	0.976	0.945
Extreme Gradient Boosting	k-mer	0.963	0.919
Random Forest	AA k-mer	0.923	0.839
Support Vector Machine	AA k-mer	0.944	0.881
Stochastic Gradient Boosting	AA k-mer	0.944	0.880
Extreme Gradient Boosting	AA k-mer	0.935	0.863
Random Forest	SNP	0.746	0.388
Support Vector Machine	SNP	0.795	0.481
Stochastic Gradient Boosting	SNP	0.728	0.379
Extreme Gradient Boosting	SNP	0.741	0.404
Random Forest	SNP/AA k-mer	0.936	0.863
Support Vector Machine	SNP/AA k-mer	0.934	0.854
Stochastic Gradient Boosting	SNP/AA k-mer	0.947	0.887
Extreme Gradient Boosting	SNP/AA k-mer	0.943	0.876
Random Forest	SNP/k-mer	0.957	0.910
Support Vector Machine	SNP/k-mer	0.969	0.932
Stochastic Gradient Boosting	SNP/k-mer	0.955	0.906
Extreme Gradient Boosting	SNP/k-mer	0.951	0.895
Random Forest	AA k-mer/k-mer	0.931	0.855
Support Vector Machine	AA k-mer/k-mer	0.949	0.895
Stochastic Gradient Boosting	AA k-mer/k-mer	0.932	0.850
Extreme Gradient Boosting	AA k-mer/k-mer	0.923	0.833

GBM nucleotide kmer has the highest accuracy according erythromycin model and figure 7 shows the importance of top ten ranked kmer features.

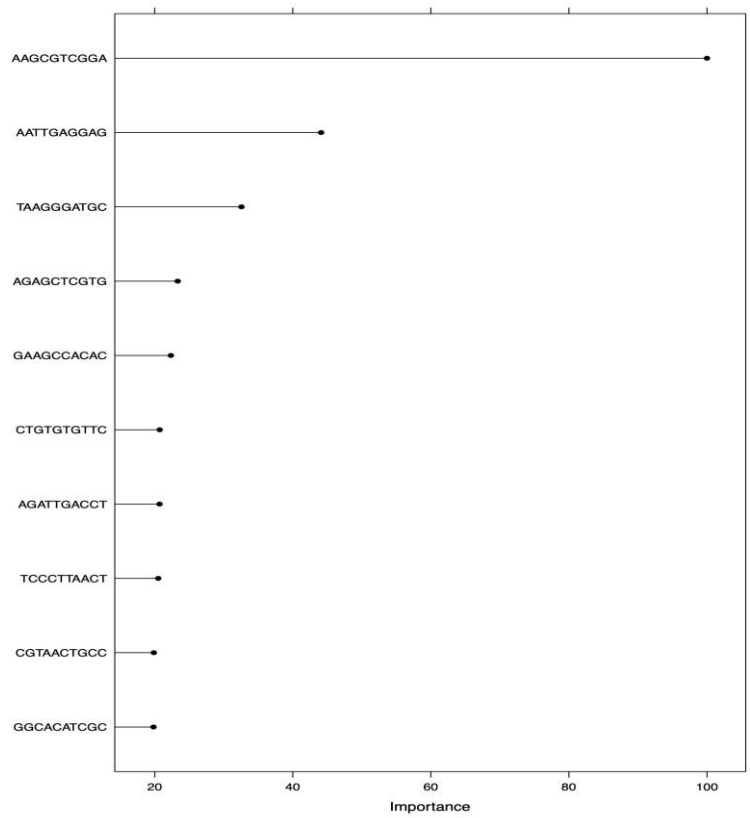

Figure 7: The top 10 important k-mer features for GBM model of erythromycin.

As with erythromycin, tetracycline also was successful in accuracy and kappa results. The highest accuracy was seen in the SNP/kmer xgboost model. Accuracy was 0.996, and kappa 0.991 according to this model. Except for the models in which the SNP feature was used alone, all model results were close to each other and gave results above 0.95. Models using SNP as a feature showed an accuracy between 0.77-0.84. The most accurate nucleotide kmer model SVM and random forest were found. The ten most important features of the XGB SNP/AA kmer model are shown in Figure 8. A comparison of the model results in tetracycline is shown in Table 12, Figure 11, Appendix 4, Table A13, and Figure A5.

Table 14: Kappa and accuracy results of *S.aureus* tetracycline results.

Algorithm	input	Accuracy	Kappa
Random Forest	k-mer	0.989	0.976
Support Vector Machine	k-mer	0.992	0.983
Stochastic Gradient Boosting	k-mer	0.976	0.951
Extreme Gradient Boosting	k-mer	0.971	0.933
Random Forest	AA k-mer	0.978	0.953
Support Vector Machine	AA k-mer	0.969	0.934
Stochastic Gradient Boosting	AA k-mer	0.955	0.901
Extreme Gradient Boosting	AA k-mer	0.968	0.933
Random Forest	SNP	0.836	0.649
Support Vector Machine	SNP	0.845	0.669
Stochastic Gradient Boosting	SNP	0.767	0.514
Extreme Gradient Boosting	SNP	0.770	0.516
Random Forest	SNP/AA k-mer	0.976	0.945
Support Vector Machine	SNP/AA k-mer	0.967	0.931
Extreme Gradient Boosting	SNP/AA k-mer	0.996	0.991
Random Forest	SNP/k-mer	0.992	0.983
Support Vector Machine	SNP/k-mer	0.992	0.983
Stochastic Gradient Boosting	SNP/k-mer	0.979	0.955
Extreme Gradient Boosting	SNP/k-mer	0.971	0.935
Random Forest	AA k-mer/k-mer	0.976	0.948
Support Vector Machine	AA k-mer/k-mer	0.983	0.961
Stochastic Gradient Boosting	AA k-mer/k-mer	0.975	0.946
Extreme Gradient Boosting	AA k-mer/k-mer	0.984	0.965

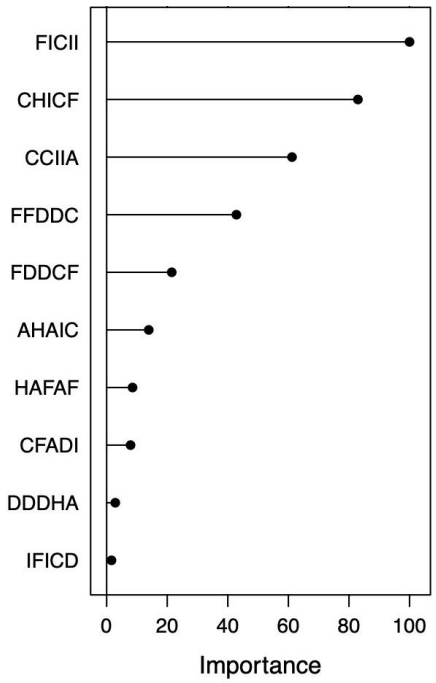

Figure 8: The top 10 important SNP / AA kmer features for GBM model of tetracycline.

When we look at the results of *S.aureus* gentamicin, we see that all models, excluding our models that use SNP locations, as in the other two antibiotics, make highly accurate predictions. The highest accuracy in the nucleotide k-mer model is seen with 0.996 in the GBM model. When we look at the amino acid kmer features, the most successful model was random forest, while the most successful model was SVM when the SNP feature was used. When all models were examined individually, the most successful models were random forests with amino acid kmer. The ten most important features of this model are shown in Figure 9. More detailed results are shown in Table 13, Figure 11, Appendix 4, Table A14, and Figure A6.

Table 15: Kappa and accuracy results of *S.aureus* Gentamicin results.

Algorithm	input	Accuracy	Kappa
Random Forest	k-mer	0.984	0.963
Support Vector Machine	k-mer	0.992	0.982
Stochastic Gradient Boosting	k-mer	0.996	0.990
Extreme Gradient Boosting	k-mer	0.984	0.961
Random Forest	AA k-mer	1.000	1.000
Support Vector Machine	AA k-mer	0.992	0.983
Stochastic Gradient Boosting	AA k-mer	0.992	0.980
Extreme Gradient Boosting	AA k-mer	0.991	0.974
Random Forest	SNP	0.854	0.606
Support Vector Machine	SNP	0.860	0.630
Stochastic Gradient Boosting	SNP	0.678	0.172
Extreme Gradient Boosting	SNP	0.732	0.223
Random Forest	SNP/AA k-mer	0.997	0.991
Extreme Gradient Boosting	SNP/AA k-mer	0.992	0.980
Random Forest	SNP/k-mer	0.992	0.980
Support Vector Machine	SNP/k-mer	0.992	0.980
Stochastic Gradient Boosting	SNP/k-mer	0.976	0.941
Extreme Gradient Boosting	SNP/k-mer	0.958	0.892
Random Forest	AA k-mer/k-mer	0.992	0.983
Support Vector Machine	AA k-mer/k-mer	0.992	0.982
Stochastic Gradient Boosting	AA k-mer/k-mer	0.917	0.790
Extreme Gradient Boosting	AA k-mer/k-mer	0.996	0.990

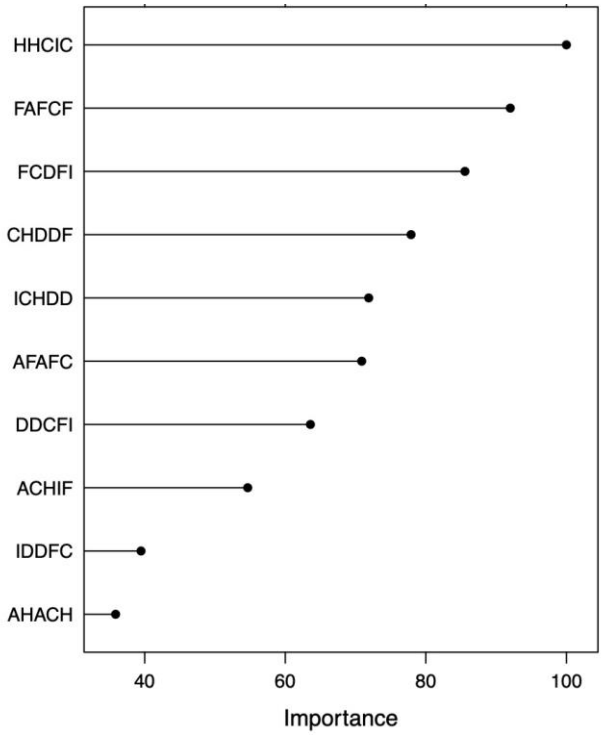

Figure 9: The top 10 important AA kmer features for RF model of gentamicin.

When we examine the models we run with ciprofloxacin, we see that the nucleotide kmer, which we use alone as a feature, gives the highest accuracy. When we compare the nucleotide kmer models within themselves, we see that the best model with 0.97 accuracy and 0.94 kappa results is established with random forest. Figure 10 shows the top ten most important features of this model.

Apart from this, it is seen that SVM, gbm, and xgboost give very close and successful results. After k-mer, we see that the most successful models are the combination features using kmer. These are followed by aakmer and SNP results. More detailed results are shown in Table 14, Figure 11, Appendix 4, Table A15, and Figure A7.

Table 16: Kappa and accuracy results of *S.aureus* Ciprofloxacin results.

Algorithm	Input	Accuracy	Kappa
Random Forest	k-mer	0.9754	0.9440
Support Vector Machine	k-mer	0.9673	0.9290
Stochastic Gradient Boosting	k-mer	0.9638	0.9220
Extreme Gradient Boosting	k-mer	0.9596	0.9111
Random Forest	AA k-mer	0.9158	0.8219
Support Vector Machine	AA k-mer	0.9500	0.8958
Stochastic Gradient Boosting	AA k-mer	0.9349	0.8592
Extreme Gradient Boosting	AA k-mer	0.9220	0.8271
Random Forest	SNP	0.8792	0.7276
Support Vector Machine	SNP	0.8858	0.7662
Stochastic Gradient Boosting	SNP	0.8983	0.7638
Extreme Gradient Boosting	SNP	0.8638	0.7008
Random Forest	SNP/AA k-mer	0.9266	0.8412
Support Vector Machine	SNP/AA k-mer	0.9666	0.9304
Stochastic Gradient Boosting	SNP/AA k-mer	0.9501	0.8928
Extreme Gradient Boosting	SNP/AA k-mer	0.9439	0.8805
Random Forest	SNP/k-mer	0.9608	0.9128
Support Vector Machine	SNP/k-mer	0.9667	0.9257
Stochastic Gradient Boosting	SNP/k-mer	0.9627	0.9224
Extreme Gradient Boosting	SNP/k-mer	0.9544	0.8993
Random Forest	AA k-mer/k-mer	0.9315	0.8568
Support Vector Machine	AA k-mer/k-mer	0.9519	0.9008
Stochastic Gradient Boosting	AA k-mer/k-mer	0.9449	0.8817
Extreme Gradient Boosting	AA k-mer/k-mer	0.9291	0.8407

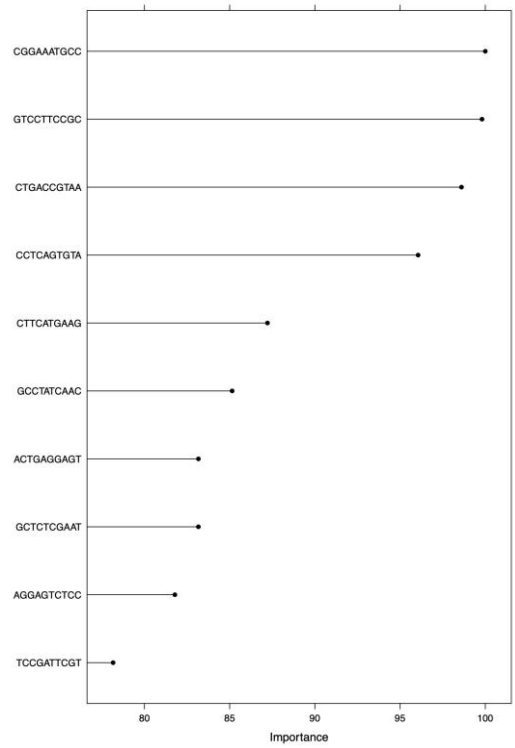

Figure 10: The top 10 important nucleotide kmer features for RF model of ciprofloxacin.

When we look at the results of *S.aureus* gentamicin, we see that all models, excluding our models that use SNP locations, as in the other two antibiotics, make predictions with high accuracy. The highest accuracy in the nucleotide k-mer model is seen with 0.996 in the GBM model. When we look at the amino acid kmer features, the most successful model was random forest, while the most successful model was SVM when the SNP feature was used. When all models were examined individually, the most successful models were random forest with amino acid kmer. The ten most important features of this model are shown in Figure 9. More detailed results are shown in Table 12, Figure 11, Appendix 4, Table A14, and Figure A6.

Table 15 shows the performance rankings of each method based on the average accuracy for all antibiotics in *S.aureus.* Figure 11 shows all accuracy comparisons for erythromycin, tetracycline, gentamicin, and ciprofloxacin models shown in bar graphs.

Table 17: Performance rankings of each antibiotic of *S.aureus*.

	Erythromycin	**Tetracycline**	**Gentamicin**	**Ciprofloxacin**
1	Kmer	SNP / AA kmer	AA kmer	Kmer
2	SNP / kmer	Kmer	SNP / aa kmer	SNP / kmer
3	AA kmer / kmer	SNP / kmer	Kmer	SNP / AA kmer
4	SNP / AA kmer	AA kmer / kmer	AA kmer / kmer	AA kmer / kmer
5	AA kmer	AA kmer	SNP / kmer	AA kmer
6	SNP	SNP	SNP	SNP

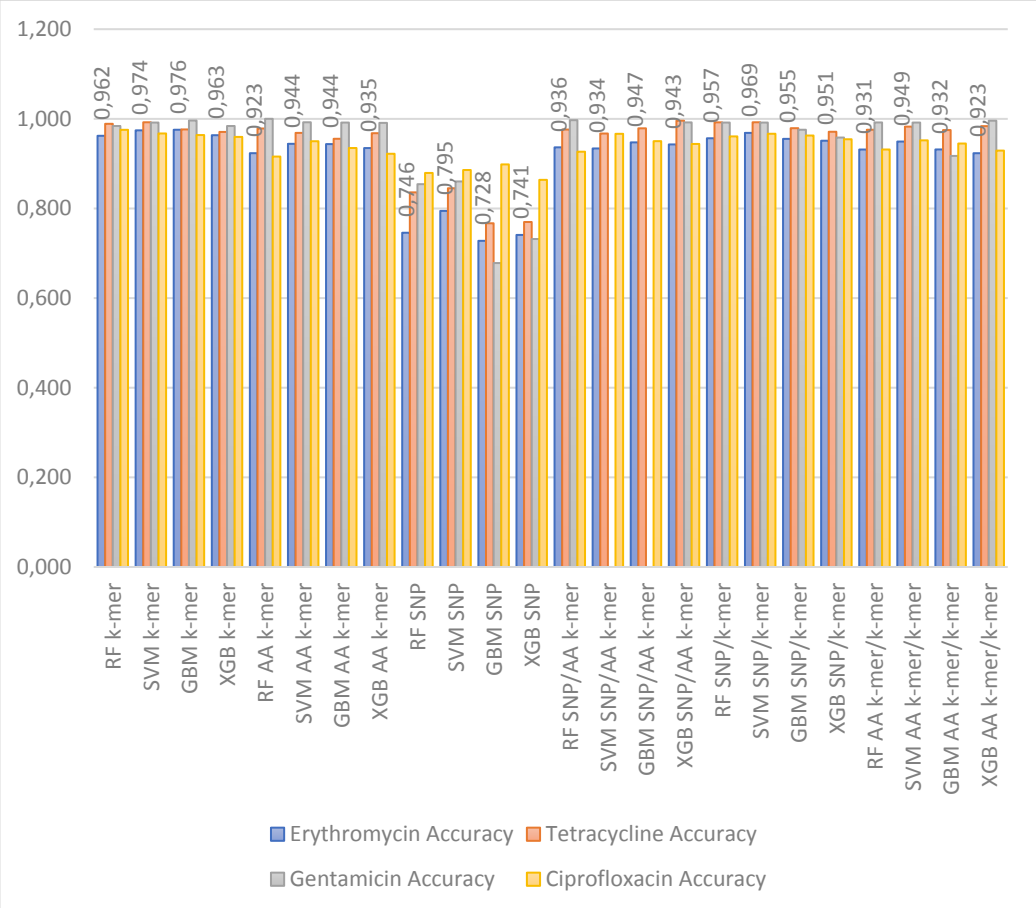

Figure 11: All Accuracy comparisons for erythromycin, tetracycline, gentamicin, and ciprofloxacin models are shown in bar graphs.

4.3. Enterococcus faecium Results

We used 124 samples dataset for E. faecium analysis. During training, ampicillin and vancomycin antibiotics are used. We have 99 resistant and 25 susceptible samples in ampicillin and 52 resistant and 72 susceptible samples in vancomycin.

When different feature inputs in ampicillin were compared, six types of input data were used, and they performed almost equally in terms of accuracy. Accuracy results are between 0.91 to 0.99. However, when we checked the kappa results, the results were between 0.65 to 0.98. SNP and SNP-included feature combinations accuracy was higher than other inputs. Comparisons of Ampicillin models are shown in Table 16, Figure 14, Appendix 5, Table A16, and Figure A8. The support vector machine achieved the highest accuracy with 0.99. Figure 12 shows the top ten most important SNP / AA kmer features for the random forest model of ampicillin.

Table 18: Kappa and accuracy results of *E.faecium* Ampicillin results.

Algorithm	input	Accuracy	Kappa
Random Forest	k-mer	0.953	0.854
Support Vector Machine	k-mer	0.930	0.806
Stochastic Gradient Boosting	k-mer	0.925	0.773
Extreme Gradient Boosting	k-mer	0.930	0.771
Random Forest	AA k-mer	0.963	0.863
Support Vector Machine	AA k-mer	0.940	0.796
Stochastic Gradient Boosting	AA k-mer	0.910	0.705
Extreme Gradient Boosting	AA k-mer	0.950	0.809
Random Forest	SNP	0.923	0.690
Support Vector Machine	SNP	0.990	0.962
Stochastic Gradient Boosting	SNP	0.940	0.746
Extreme Gradient Boosting	SNP	0.920	0.658
Random Forest	SNP/AA k-mer	0.970	0.898
Support Vector Machine	SNP/AA k-mer	0.990	0.974
Stochastic Gradient Boosting	SNP/AA k-mer	0.965	0.878

Table 19: Kappa and accuracy results of *E.faecium* Ampicillin results (continue).

Extreme Gradient Boosting	SNP/AA k-mer	0.965	0.871
Random Forest	SNP/k-mer	0.950	0.800
Support Vector Machine	SNP/k-mer	0.990	0.974
Stochastic Gradient Boosting	SNP/k-mer	0.950	0.802
Extreme Gradient Boosting	SNP/k-mer	0.935	0.738
Random Forest	AA k-mer/k-mer	0.957	0.840
Support Vector Machine	AA k-mer/k-mer	0.940	0.723
Stochastic Gradient Boosting	AA k-mer/k-mer	0.945	0.813
Extreme Gradient Boosting	AA k-mer/k-mer	0.965	0.860

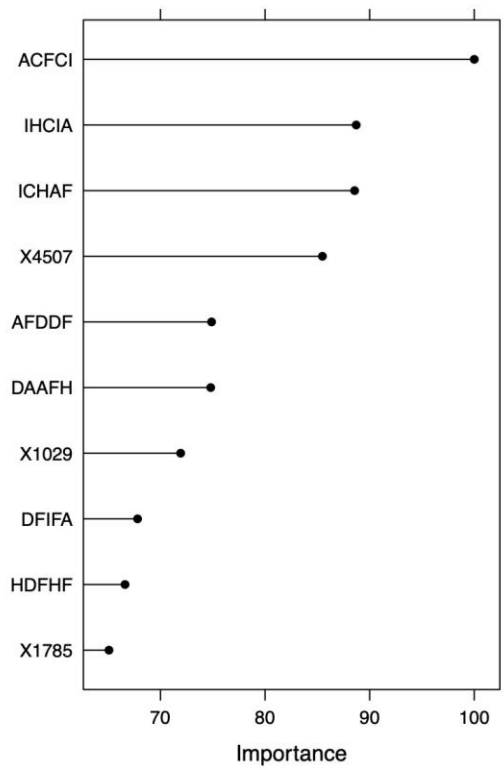

Figure 12: The top 10 important nucleotide SNP / AA kmer features for model of ampicillin.

When we compare the *E.faecium* vancomycin results, we can conclude that all models, except SNP inputs, make predictions with high accuracy. The highest

accuracy in the SNP and amino acid 5-mer model is seen with 0.966 in the GBM model. This result is very similar to the ampicillin result. When we used the amino acid 5-kmer features, the most successful model was XGB. The ten most important features of this model are shown in Figure 13. More detailed results are shown in Table 17, Figure 14, Appendix 5, Table A17, and Figure A9.

Table 20: Kappa and accuracy results of *E.faecium* Vancomycin results.

Algorithm	Input	Accuracy	Kappa
Random Forest	k-mer	0.909	0.798
Support Vector Machine	k-mer	0.912	0.804
Stochastic Gradient Boosting	k-mer	0.901	0.788
Extreme Gradient Boosting	k-mer	0.897	0.779
Random Forest	AA k-mer	0.944	0.886
Support Vector Machine	AA k-mer	0.912	0.808
Stochastic Gradient Boosting	AA k-mer	0.935	0.860
Extreme Gradient Boosting	AA k-mer	0.956	0.909
Random Forest	SNP	0.756	0.514
Support Vector Machine	SNP	0.780	0.580
Stochastic Gradient Boosting	SNP	0.778	0.536
Extreme Gradient Boosting	SNP	0.781	0.548
Random Forest	SNP/AA k-mer	0.960	0.920
Support Vector Machine	SNP/AA k-mer	0.950	0.895
Stochastic Gradient Boosting	SNP/AA k-mer	0.966	0.930
Extreme Gradient Boosting	SNP/AA k-mer	0.954	0.902
Random Forest	SNP/k-mer	0.942	0.886
Support Vector Machine	SNP/k-mer	0.939	0.878
Stochastic Gradient Boosting	SNP/k-mer	0.908	0.814
Extreme Gradient Boosting	SNP/k-mer	0.913	0.820

Table 21: Kappa and accuracy results of *E.faecium* Vancomycin results (continue).

Random Forest	AA k-mer/k-mer	0.948	0.894
Support Vector Machine	AA k-mer/k-mer	0.920	0.840
Stochastic Gradient Boosting	AA k-mer/k-mer	0.959	0.919
Extreme Gradient Boosting	AA k-mer/k-mer	0.944	0.883

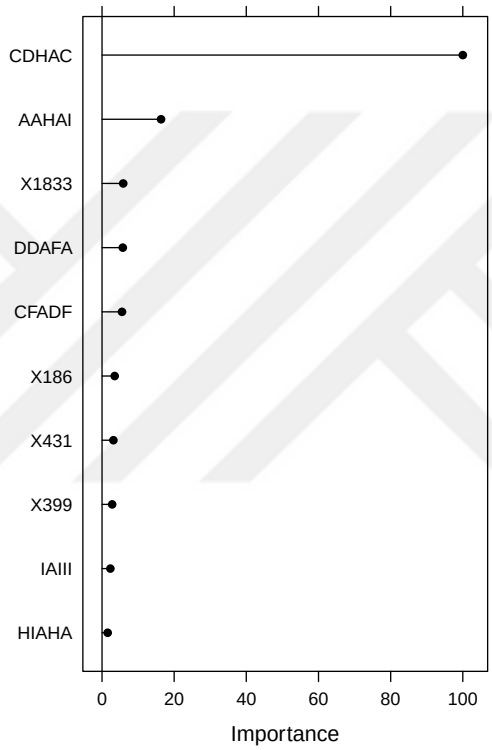

Figure 13: The top 10 important nucleotide SNP/ AA k-mer features for GBM model of vancomycin.

Table 18 is the performance rankings of each method based on the accuracy results of ampicillin and vancomycin in *E.faecium* and Figure 14 is all accuracy comparisons for our models are shown in bar graphs.

Table 22: Performance rankings of each antibiotic of *E.faecium*

	Ampicillin	**Vancomycin**
1	SNP / AA kmer	SNP / AA kmer
2	SNP / kmer	AA kmer / kmer
3	SNP	AA kmer
4	AA kmer / kmer	SNP / kmer
5	AA kmer	Kmer
6	Kmer	SNP

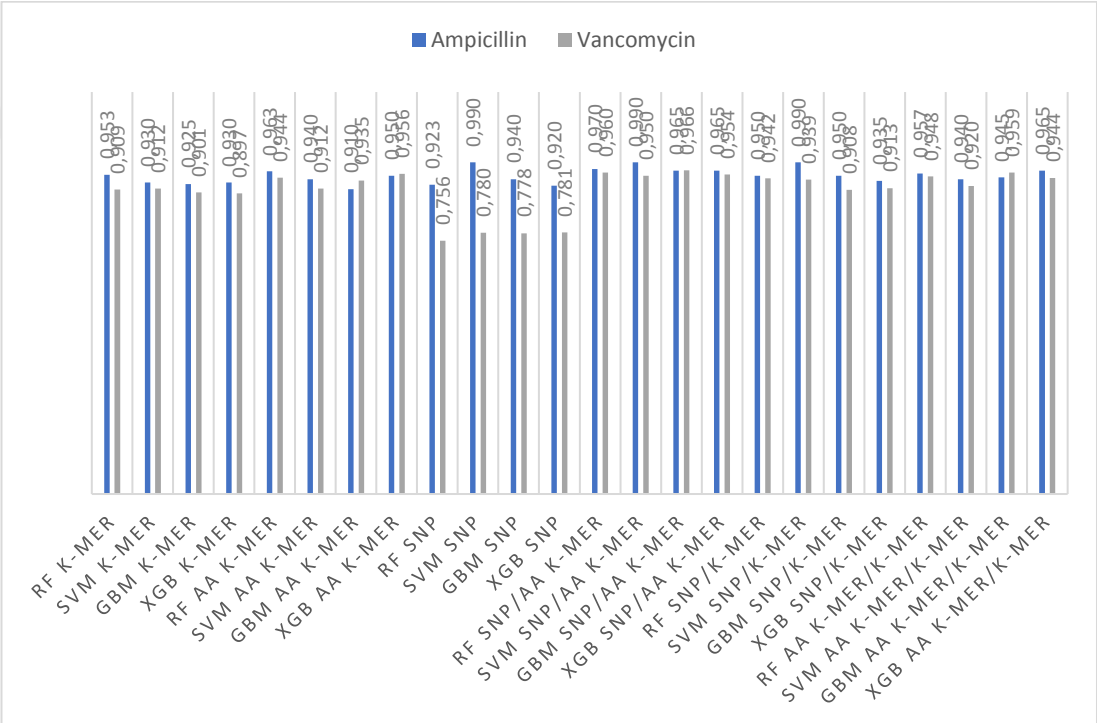

Figure 14: All Accuracy comparisons for ampicillin and vancomycin models are shown in bar graphs.

5 DISCUSSION

There has been a significant increase in morbidity and mortality rates associated with infectious diseases worldwide due to antimicrobial resistance (AMR) (1,2). With the decrease in effective antibiotics over the years, molecular approaches have become increasingly important in the fight against antimicrobial resistance. It is imperative to detect AMR in bacterial genomes as quickly as possible (9-11). To build accurate classifiers, it is also necessary to have a balance between susceptible and resistant genomes. Currently, this is a major limitation. As a result, the number of genomes with AMR data is usually resistant since hospitals and epidemiologists often rely on these genomes for clinical purposes.

Given the current data sets available at PATRIC, we built Random Forest (RF), Support Vector Machine (SVM), Stochastic Gradient Boosting (GBM), and Extreme Gradient Boosting (XGBoost) classifiers for penicillin, erythromycin, and tetracycline resistance *Streptococcus pneumoniae*, erythromycin, tetracycline, gentamicin, and ciprofloxacin resistance *Staphylococcus aureus*, ampicillin, and vancomycin resistance *Enterococcus faecium*.

Three different models were tested for RF. The default for each parameter model is M1, the random search is M2, and for the final model, M3, a grid search was done. Even though we had higher accuracy results than the other machine learning models, RF worked slowly on large datasets, especially when we tried grid search. For the SVM classifier, SVM Linear with tunned parameter c is M1, and the second and third models are M2 and M3; we used polynomial kernel and radial kernel functions.

When we checked the best *S.pneumoniae* parameters tables (Table A6, Appendix 2) in most penicillin models, SVM Linear was given the best accuracy results (M1); however, radial kernel function was given the best accuracy results for all erythromycin and tetracycline models. Our GBM model first worked with default parameters, M1. In the second model, 10-fold cross-validation and two repeats were applied, M2. In all GBM accuracy results, M2 was higher than M1. Two parameters

were used to tune the M3 GBM model for a grid search. These are boosting iterations and maximum tree depth. As shown in Figure 4, in nucleotide k-mer gbm M3 model penicillin samples, we changed boosting iteration parameters between 0 to 1500. We compared the models' accuracy results with three different max tree depth values (1, 2, 3). Increasing boosting iteration can increase accuracy, but it depends on the max tree depth. For XGBoost models, the Minimum sum of instance weight is set to 1 and 7, and a maximum depth of a tree is set to 3 and 5. The subsample ratio of columns was tried between 0.3 to 0.525. Other parameters remained the same as the default. In Figure 16, we can see the accuracy change according to the Penicillin XGBoost nucleotide k-mer model parameter changes.

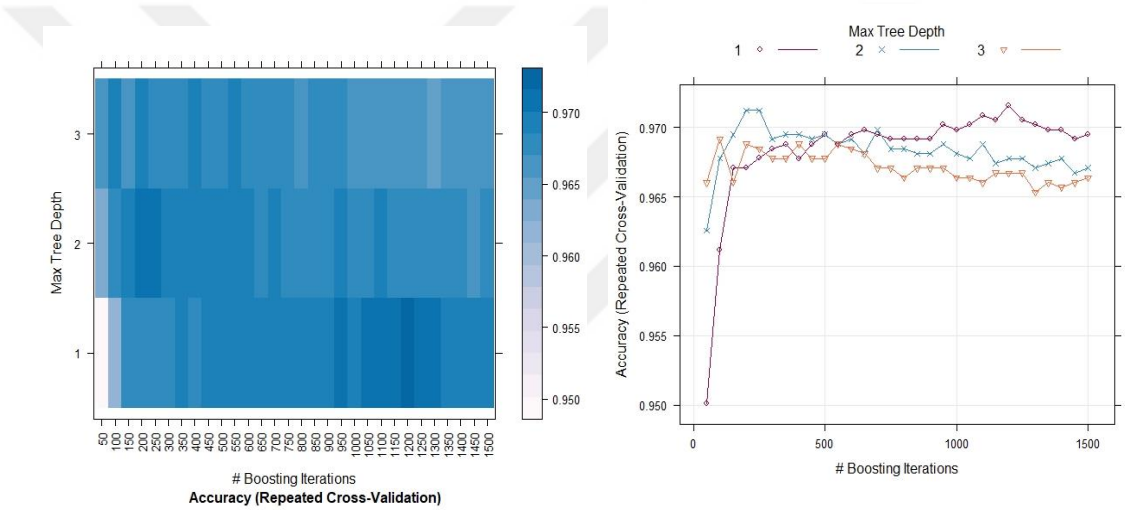

Figure 15: Penicillin nucleotide kmer GBM model accuracy results with different boosting iterations and max tree depth.

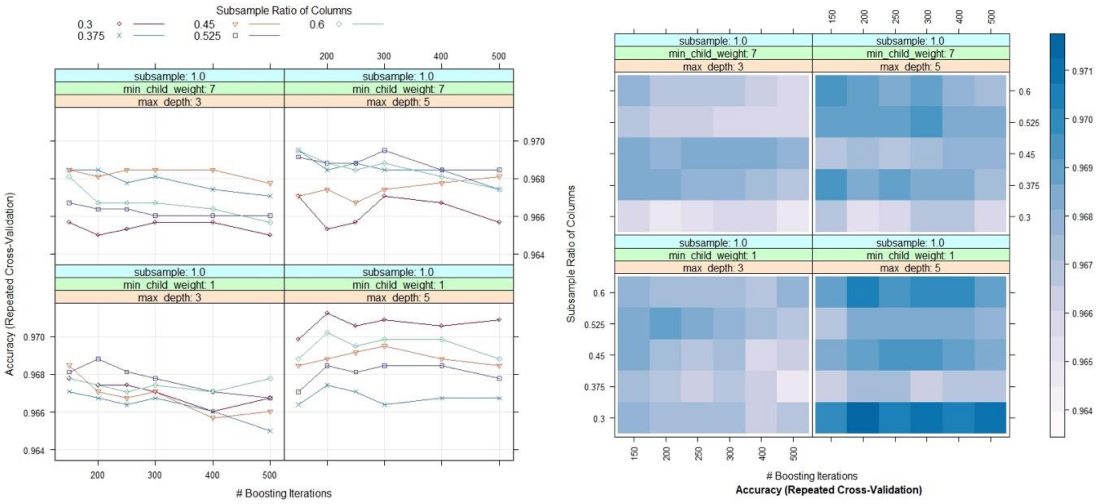

Figure 16: Penicillin nucleotide kmer XGBoost model accuracy results with different subsample ratios of columns, min weight, and maximum depth.

Our results demonstrate that different considerations support using alternative methods when considering the feature extraction method for AMR prediction. Different pre-processing steps are involved in the feature extraction approaches presented here. SNP features must be aligned to the reference genome since every SNP is unique compared to the reference genome. Since databases such as NCBI provide the reference genomes, alignment to the reference genomes is usually not an issue. On the other hand, an amino acid sequence is an input to the feature extraction method for amino acid k-mers. The top priority is predicting AMR as quickly and as cheaply as possible. This method is more interpretable and has a smaller feature size, making amino acid k-mers the better choice.

Compared to *S.pneumonie* results, we found that different feature inputs produced optimal results for other antibiotics. The best performance for penicillin was obtained with nucleotide 10-mers. Alternatively, amino acid 5-mers combined with SNPs produced the best results for erythromycin, and nucleotide 10-mers combined with amino acid 5-mers produced the best results for tetracycline.

In comparison to *S.aerous* results, like *S.pneumoniae,* we found that different feature inputs produced the best results for other antibiotics; however, nucleotide kmer

is the best feature for both erythromycin and ciprofloxacin and second for tetracycline. The best performance for tetracycline was obtained with SNP / AA kmer. Alternatively, amino acid k-mers produced the best results for gentamicin. Rather than other antibiotic results, SNP and amino acid kmer combination inputs have higher accuracy in both ampicillin and vancomycin *E. facium* results.

Unlike other studies, we used kmer, aakmer, SNP, and their binary combinations as inputs. When we evaluated the results, we saw that the SNP / AA kmer combination in *E.faecium* samples gave higher accuracy than all other inputs in ampicillin and vancomycin. When we look at the results of *S.aureus* with tetracycline and *S.pneumoniae* with erythromycin, we see that the most successful model was established with the combination of SNP / AA kmer. When we evaluate the top 10 features of these models, we see that 6/10 of the model get aa inputs and 4/10 from SNP inputs (Figure 13).

Excessive features in machine learning can result in overfitting and increased memory requirements. However, after increasing the k-mer length, the dataset becomes too large to handle by the machine learning algorithm. Using long k-mers is complicated because the number of features increases. However, we have shown that for amino acid k-mers, this increase in feature size is less severe than for nucleotide k-mers. In contrast to nucleotide k-mers, amino acid k-mers represent biological information more compactly: each codon has three nucleotides corresponding to one amino acid.

Furthermore, for example, amino acid k-mers and their combinations deliver better accuracy than other inputs in E.faecium models. As in previous Genome-Wide Association Studies, we also found that RF discovered k-mers associated with penicillin resistance in *S. pneumoniae* that corresponded with the pbp2x gene (9, 59). RF analysis of pbp1a and pbp2a penicillin-binding proteins showed significant variations related to resistance in Chewapreecha and colleagues' study (59).

6 CONCLUSION

This study compares different ML models for predicting AMR phenotype for *Streptococcus pneumoniae*, *Staphylococcus aureus*, and *Enterococcus faecium*.

We compared nucleotide k-mers, amino acid k-mers, and SNPs and their combinations to predict AMR for three antibiotics: Penicillin, Erythromycin, and Tetracycline for *S. pneumoniae*, for four antibiotics: Erythromycin, Tetracycline, Gentamicin, and Ciprofloxacin for *S. aureus*, for two antibiotics Ampicillin and Vancomycin for *E.faecium*. We used the Boruta package as a feature elimination. We compared various machine learning methods to train classification models, including random forests, support vector machines, stochastic gradient boosting, and extreme gradient boosting. For MIC prediction, we discussed the strengths and limitations of feature and ML model selection. From our work, we have observed that the model setup features, and the ML method choice affect the results of each antibiotic. In particular, the combination of features giving high accuracy and F1 score for some antibiotics suggests that these feature inputs should be evaluated in the future. The approach undertaken by this study can be used in further studies to improve AMR prediction performance and accuracy.

7 REFERENCES

1. World Health Organization. Antimicrobial resistance. World Health Organization from https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance.
2. World Health Organization. A new report calls for urgent action to avert the antimicrobial resistance crisis. World Health Organization. from https://www.who.int/news/item/29-04-2019-new-report-calls-for-urgent-action-to-avert-antimicrobial-resistance-crisis.
3. Blair J, Webber M, Baylay A. Molecular mechanisms of antibiotic resistance. Nat Rev Microbiology. 2015;13:42–51.
4. Michael CA, Dominey-Howes D, Labbate M. The antimicrobial resistance crisis: Causes, consequences, and management. Frontiers in Public Health. 2014;2.
5. VanOeffelen M, Nguyen M, Aytan-Aktug D, Brettin T, Dietrich EM, Kenyon RW, Machi D, Mao C, Olson R, Pusch GD, Shukla M, Stevens R, Vonstein V, Warren AS, Wattam AR, Yoo H, Davis JJ. A genomic data resource for predicting antimicrobial resistance from laboratory-derived antimicrobial susceptibility phenotypes. Briefings in Bioinformatics, 2021;22(6).
6. Review on Antimicrobial Resistance (AMR Review). Rapid Diagnostics: Stopping Unnecessary Use of Antibiotics. The Review on Antimicrobial Resistance, chaired by Jim O'Neill. https://amr-review.org/Publications.html. 2015.
7. Zhang C, Ju Y, Tang N, Li Y, Zhang G, Song Y, Fang H, Yang L, Feng J. Systematic analysis of supervised machine learning as an effective approach to predicate β-lactam resistance phenotype in Streptococcus pneumoniae. Briefings in Bioinformatics. 2019;21(4), 1347–1355.
8. Khaledi A, Weimann A, Schniederjans M, Asgari E, Kuo TH, Oliver A, Cabot G, Kola A, Gastmeier P, Hogardt M, Jonas D, Mofrad MRK, Bremges A, McHardy AC, Häussler S. Predicting antimicrobial resistance in pseudomonas aeruginosa with machine learning- enabled molecular diagnostics. EMBO Molecular Medicine. 2020;12(3).
9. Davis JJ, Boisvert S, Brettin T, Kenyon RW, Mao C, Olson R, Overbeek R, Santerre J, Shukla M, Wattam AR, Will R, Xia F, Stevens R. Antimicrobial Resistance Prediction in Patric and rast. Scientific Reports.2016;6(1).
10. Leski TA, Vora GJ, Barrows BR, Pimentel G, House BL, Nicklasson M, Wasfy M, Abdel-Maksoud M, Taitt CR. Molecular characterization of multidrug-resistant hospital isolates using the antimicrobial resistance determinant microarray. PLoS ONE. 2013;8(7).
11. Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ, Tomita T, Zobel J, Holt KE. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Medicine.2014; 6(11).
12. Dunne Jr WM, Jaillard M, Rochas O, Van Belkum A. Microbial Genomics and antimicrobial susceptibility testing. Expert Review of Molecular Diagnostics. 2017;17(3), 257–269.
13. Jindal HM, Ramanathan B, Le CF, Gudimella R, Razali R, Manikam R, Sekarausing machinen S D. Comparative genomic analysis of ten clinical streptococcus pneumoniae collected from a

Malaysian hospital reveal 31 new unique drug-resistant SNPs using whole genome sequencing. Journal of Biomedical Science.2018;25(1).

14. van der Poll T, Opal SM. Pathogenesis, treatment, and prevention of pneumococcal pneumonia. The Lancet. 2009;374(9700), 1543–1556.

15. Henriques-Normark B, Tuomanen E.I. The pneumococcus: Epidemiology, microbiology, and pathogenesis. Cold Spring Harbor Perspectives in Medicine.2013;3(7).

16. World Health Organization. Pneumococcal disease. World Health Organization. from https://www.who.int/teams/health-product-policy-and-standards/standards-and-specifications/vaccine-standardization/pneumococcal-disease.

17. Zapun A, Contreras-Martel C, Vernet T. Penicillin-binding proteins and β-lactam resistance. FEMS Microbiology Reviews, 2008;32(2), 361–385.

18. Poole K. Resistance to B-lactam antibiotics. Cellular and Molecular Life Sciences. 2004;61(17).

19. Sader HS, Mendes RE, Le J, Denys G, Flamm, RK, Jones, RN. Antimicrobial susceptibility of streptococcus pneumoniae from North America, Europe, Latin America, and the Asia-Pacific region: Results from 20 years of the Sentry Antimicrobial Surveillance Program (1997–2016). Open Forum Infectious Diseases,2019.

20. Cornick JE,Bentley SD. Streptococcus pneumoniae: The evolution of antimicrobial resistance to beta-lactams, fluoroquinolones, and macrolides. Microbes and Infection.2012;14(7-8), 573–583.

21. Montanari MP, Cochetti I, Mingoia M, Varaldo PE. Phenotypic and molecular characterization of tetracycline- and erythromycin-resistant strains of streptococcus pneumoniae. Antimicrobial Agents and Chemotherapy. 2003; 47(7), 2236–2241.

22. Aytan-Aktug D, Clausen PT, Bortolaia V, Aarestrup FM, Lund O. Prediction of acquired antimicrobial resistance for multiple bacterial species using neural networks. MSystems. 2020; 5(1).

23. Deelder W, Christakoudi S, Phelan J, Benavente ED, Campino S, McNerney R, Palla L, Clark TG. Machine learning predicts accurately mycobacterium tuberculosis drug resistance from whole genome sequencing data. Frontiers in Genetics.2019;10.

24. Drouin A, Giguère S, Déraspe M, Marchand M, Tyers M, Loo VG, Bourgault AM, Laviolette F, Corbeil J. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. BMC Genomics.2016;17(1).

25. Wang L, Yang J, Chen L, Wang W, Yu F, Xiong H. Whole-genome sequencing of mycobacterium tuberculosis for prediction of drug resistance. Epidemiology and Infection.2022; 150.

26. Yang Y, Walker TM, Walker AS, Wilson DJ, Peto TEA, Crook DW, Shamout F, CRyPTIC Consortium, Zhu T, Clifton DA, DeepAMR for predicting co-occurrent resistance of Mycobacterium tuberculosis, Bioinformatics. 2019;18(35), 3240–3249.

27. Yang Y, Niehaus KE, Walker TM, Iqbal Z, Walker AS, Wilson DJ, Peto TE, Crook DW, Smith E G, Zhu T, Clifton DA. Machine learning for classifying tuberculosis drug resistance from DNA sequencing data. Bioinformatics. 2017;34(10), 1666–1671.

28. 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 Computational Biology. 2018;14(12).

29. Pataki BÁ, Matamoros S, van der Putten BCL, Remondini D, Giampieri E, Aytan-Aktug D, Hendriksen RS, Lund O, Csabai I, Schultsz C. Understanding and predicting ciprofloxacin minimum inhibitory concentration in escherichia coli with machine learning.2019.

30. Nguyen M, Long SW, McDermott PF, Olsen RJ, Olson R, Stevens RL, Tyson GH, Zhao S, Davis JJ. Using machine learning to predict antimicrobial mics and associated genomic features for nontyphoidal salmonella. Journal of Clinical Microbiology.2019; 57(2).

31. ValizadehAslani T, Zhao Z, Sokhansanj BA, Rosen GL. Amino acid K-Mer feature extraction for quantitative antimicrobial resistance (AMR) prediction by machine learning and model interpretation for Biological Insights. Biology.2020;9(11), 365.

32. Li Y, Metcalf BJ, Chochua S, Li Z, Gertz RE, Walker H, Hawkins PA, Tran T, Whitney CG, McGee L, Beall BW. Penicillin-binding protein transpeptidase signatures for tracking and predicting β-lactam resistance levels in streptococcus pneumoniae. MBio.2016;7(3).

33. Li Y, Metcalf BJ, Chochua S, Li Z, Gertz RE, Walker H, Hawkins PA, Tran T, McGee L, Beall BW. Validation of β-lactam minimum inhibitory concentration predictions for pneumococcal isolates with newly encountered penicillin-binding protein (PBP) sequences. BMC Genomics.2017;18(1).

34. Zhang C, Ju Y, Tang N, Li Y, Zhang G, Song Y, Fang H, Yang L, Feng J. Systematic analysis of supervised machine learning as an effective approach to predicate β-lactam resistance phenotype in Streptococcus pneumoniae. Briefings in Bioinformatics.2019; 21(4), 1347–1355.

35. Liu Z, Deng D, Lu H, Sun J, Lv L, Li S, Peng G, Ma X, Li J, Li Z, Rong T, Wang G. Evaluation of machine learning models for predicting antimicrobial resistance of actinobacillus pleuropneumoniae from whole genome sequences. Frontiers in Microbiology.2020; 11.

36. Naidenov B, Lim A, Willyerd K, Torres NJ, Johnson, WL, Hwang HJ, Hoyt P, Gustafson JE, Chen C. Pan-genomic and polymorphic driven prediction of antibiotic resistance in Elizabethkingia. Frontiers in Microbiology.2019;10.

37. Nguyen M, Brettin T, Long SW, Musser JM, Olsen RJ, Olson R, Shukla M, Stevens RL, Xia F, Yoo H, Davis JJ. Developing an in silico minimum inhibitory concentration panel test for Klebsiella pneumoniae. Scientific Reports.2018;8(1).

38. Her HL, Wu YWA pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the escherichia coli strains. Bioinformatics.2018;34(13), i89–i95.

39. Hyun JC, Kavvas ES, Monk JM, Palsson BO. Machine learning with random subspace ensembles identifies antimicrobial resistance determinants from pan-genomes of three pathogens. PLOS Computational Biology.2020;16(3).

40. Lees JA, Mai TT, Galardini M, Wheeler NE, Horsfield ST, Parkhill J, Corander J. Improved prediction of bacterial genotype-phenotype associations using interpretable pangenome-spanning regressions. MBio.2020;11(4).

41. Eyre DW, De Silva D, Cole K, Peters J, Cole MJ, Grad YH, Demczuk W, Martin I, Mulvey MR, Crook DW, Walker AS, Peto TE, Paul J. WGS to predict antibiotic mics for Neisseria gonorrhoeae. Journal of Antimicrobial Chemotherapy.2017;72(7), 1937–1947.

42. Shi J, Yan Y, Links MG, Li L, Dillon JAR, Horsch M, Kusalik A. Antimicrobial resistance genetic factor identification from whole-genome sequence data using deep feature selection. BMC Bioinformatics.2019;20(S15).

43. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, Bentley SD, Hanage WP, Lipsitch M. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nature Genetics.2013;45(6), 656–663.

44. van Tonder AJ, Bray JE, Roalfe L, White R, Zancolli M, Quirk SJ, Haraldsson G, Jolley KA, Maiden MC, Bentley SD, Haraldsson Á, Erlendsdóttir H, Kristinsson KG, Goldblatt D, Brueggemann AB. Genomics reveals the worldwide distribution of multidrug-resistant serotype 6e pneumococci. Journal of Clinical Microbiology.2015;53(7), 2271–2285.

45. Croucher NJ, Hanage WP, Harris SR, McGee L, van der Linden M, de Lencastre H, Sá-Leão R Song JH, Ko KS, Beall B, Klugman KP, Parkhill J, Tomasz A, Kristinsson KG, Bentley SD. Variable recombination dynamics during the emergence, transmission, and 'disarming' of a multidrug-resistant pneumococcal clone. BMC Biology.2014; 12(1).

46. Davis JJ, Wattam AR, Aziz RK, Brettin T, Butler R, Butler RM, Chlenski P, Conrad N, Dickerman A, Dietrich EM, Gabbard JL, Gerdes S, Guard A, Kenyon RW, Machi D, Mao C, Murphy-Olson D, Nguyen M, Nordberg EK, Olsen GJ, Olson RD, Overbeek JC, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomas C, VanOeffelen M, Vonstein V, Warren AS, Xia F, Xie D, Yoo H, Stevens R. The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. Nucleic Acids Res. 2020; 8;48

47. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. Spades: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology.2012;19(5), 455–477.

48. Wilkinson SP. kmer an R package for fast alignment-free clustering of biological sequences. R package version 1.0.0.2018.

49. Li H. Aligning sequence reads, clone sequences, and assembly contigs with BWA-MEM. ArXiv.

50. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. and 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics.2009;25, 2078-9.

51. Li H. A statistical framework for SNP calling, mutation discovery, association mapping, and population genetic parameter estimation from sequencing data. Bioinformatics. 2011;1;27(21):2987-93.

52. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomason JA 3rd, Stevens R, Vonstein V, Wattam AR, Xia F. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Scientific reports.2015;5: 8365

53. Edgar RC. Local homology recognition and distance measures in linear time using compressed amino acid alphabets. Nucleic Acids Research.2004;32, 380-385.

54. Kursa, M. B., & Rudnicki, W. R. Feature selection with the**boruta**package. Journal of Statistical Software.2010;36(11).

55. Liaw A, Wiener M. Classification and Regression by randomForest. R News2002;2(3), 18-22.

56. Kuhn M. caret: Classification and Regression Training. R package version 6.0-91.2022.

57. Greenwell B, Boehmke B, Cunningham J, and GBM Developers. gbm: Generalized Boosted Regression Models. R package version 2.1.8.2020.

58. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T, Li M, Xie J, Lin M, Geng Y, Li Y , Yuan J. xgboost: Extreme Gradient Boosting. R package version 1.5.2.1.2022.

59. Chewapreecha C. et al. Comprehensive identification of single nucleotide polymorphism associated with beta-lactam resistance within pneumococcal mosaic genes.2014.

60. Kumar N, Raven KE, Blane B, Leek D, Brown NM, Bragin E, Rhodes PA, Parkhill J, Peacock SJ. Evaluation of a fully automated bioinformatics tool to predict antibiotic resistance from MRSA genomes. *Journal of Antimicrobial Chemotherapy*. 2020;*75*(5), 1117–1122.

61. Gouliouris T, Raven KE, Ludden C, Blane B, Corander J, Horner CS, Hernandez-Garcia J, Wood P, Hadjirin NF, Radakovic M, Holmes MA, de Goffau M, Brown NM, Parkhill J, Peacock SJ. Genomic surveillance of enterococcus faecium reveals limited sharing of strains and resistance genes between livestock and humans in the United Kingdom. *MBio*. 2018;*9*(6).

62. Cassini A, Högberg LD, Plachouras D, Quattrocchi A, Hoxha A, Simonsen GS, et al. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European economic area in 2015: a population-level modelling analysis. Lancet. Infect. Dis. 2019;19, 56–66. doi: 10.1016/S1473-3099(18)30 605-4.

63. Vasala A, Hytönen VP, Laitinen OH. Modern tools for rapid diagnostics of antimicrobial resistance. *Frontiers in Cellular and Infection Microbiology*.2020;*10*.

64. Pulido MR, García-Quintanilla M, Martín-Peña R, Cisneros JM, McConnell MJ. Progress on the development of rapid methods for antimicrobial susceptibility testing. J. Antimicrob. Chemother. 2013;68, 2710–2717.

65. Plüddemann A, Onakpoya I, Harrison S, Shinkins B, Tompson A, Davis R, et al. “Position paper on anti-microbial resistance diagnostics,” in Centre for Evidence-Based Medicine. 2015;1–142.

66. Li Y, Yang X, and Zhao. Emerging microtechnologies and automated systems for rapid bacterial identification and antibiotic susceptibility testing. SLAS Technol. 2017;22, 585–608.

67. Maurer FP, Christner M, Hentschke M, Rohde H. Advances in rapid identification and susceptibility testing of bacteria in the clinical microbiology laboratory: implications for patient care and antimicrobial stewardship programs. Infect. Dis. Rep. 2017; 9:6839.

68. Syal K, Mo M, Yu H, Iriya R, Jing W, Guodong S, et al. Current and emerging techniques for antibiotic susceptibility tests. Theranostics 2017;7, 1795–1805.

69. Maugeri G, Lychko I, Sobral R, Roque ACA. (2019). Identification and antibiotic-susceptibility profiling of infectious bacterial agents: a review of current and future trends. Biotechnol. J. 2019; 14:e1700750.

70. Idelevich EA, Schüle I, Grünastel B, Wüllenweber J, Peters G, Becker K. Rapid identification of microorganisms from positive blood cultures by MALDI-TOF mass spectrometry subsequent to very shortterm incubation on solid medium. Clin. Microbiol. Infect. 2014; 20, 1001–1006.

71. *The cost of sequencing a human genome*. Genome.gov. (n.d.). Retrieved October 18, 2022, from https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

72. *Cost of NGS: Comparisons and budget guidance*. Cost of NGS | Comparisons and budget guidance. (n.d.). Retrieved October 18, 2022, from https://www.illumina.com/science/technology/next-generation-sequencing/beginners/ngs-cost.html

73. Koser CU, Ellington MJ, Peacock SJ. Whole-genome sequencing to control antimicrobial resistance. *Trends in Genetics*. 2014;*30*(9), 401–407.

74. Koser CU. et al. Rapid single-colony whole-genome sequencing of bacterial pathogens. J. Antimicrob. Chemother. 2014;69,1275–1281.

75. Koser CU et al. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog. 2012;8, e1002824.

76. Didelot X. et al. Transforming clinical microbiology with bacterial genome sequencing. Nat. Rev. Genet. 2012;13, 601–612.

77. Roemer T and Boone C. Systems-level antimicrobial drug and drug synergy discovery. Nat. Chem. Biol. 2013;9, 222–231.

78. Zumla A. et al. (2013) Advances in the development of new tuberculosis drugs and treatment regimens. Nat. Rev. Drug Discov. 2013;12, 388–404.

79. Niemann S. et al. Genomic diversity among drug sensitive and multidrug resistant isolates of Mycobacterium tuberculosis with identical DNA fingerprints. PLoS ONE. 2009;4, e7407.

80. Eyre DW. et al. Whole-genome sequencing demonstrates that fidaxomicin is superior to vancomycin for preventing reinfection and relapse of Infection with Clostridium difficile. J. Infect. Dis. 2014;209,1446–1451.

81. Grad YH et al. Genomic epidemiology of Neisseria gonorrhoeae with reduced susceptibility to cefixime in the USA: a retrospective observational study. Lancet Infect. Dis. 2014; 14, 220–226.

82. Harris SR. et al.Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. Lancet Infect. Dis. 2013;13, 130–136.

83.Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, Heidelberg J, DeBoy RT, Haft DH, Dodson RJ, et al. 2001. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science 293: 498–506.

84. Guo Y, Song G, Sun M, Wang J and Wang. Prevalence and Therapies of Antibiotic-Resistance in Staphylococcus aureus. Front. Cell. Infect. Microbiol. 2020;10:107. doi: 10.3389/fcimb.2020.00107

85. Lowy, F. D. Staphylococcus aureus infections. N. Engl. J. Med. 1998;339, 520–532. doi: 10.1056/NEJM199808203390806

86. Mlynarczyk-Bonikowska B, Kowalewski, C., Krolak-Ulinska, A., Marusza, W. Molecular Mechanisms of Drug Resistance in *Staphylococcus aureus*. *Int. J. Mol. Sci.* **2022**,*23*,8088. https://doi.org/ 10.3390/ijms23158088

87. Schulte RH., and Munson, E. *Staphylococcus aureus* resistance patterns in wisconsin: 2018 surveillance of wisconsin organisms for trends in antimicrobial resistance and epidemiology (SWOTARE) program report. *Clin. Med. Res*. 2019;17, 72–81. doi: 10.3121/cmr.2019.1503

88. Mediavilla JR., Chen, L., Mathema, B., and Kreiswirth, B. N. Global epidemiology of community-associated methicillin resistant Staphylococcus aureus (CA-MRSA). Curr. Opin. Microbiol. 2012;15, 588–595. doi: 10.1016/j.mib.2012.08.003

89. Lakhundi S., and Zhang, K. Methicillin-Resistant Staphylococcus aureus: molecular characterization, evolution, and epidemiology. Clin. Microbiol. Rev. 2018;31:e00020–18. doi: 10.1128/CMR.00020-18

90. Anuj SA., Gajera, H. P., Hirpara, D. G., and Golakiya, B. A. (2019). Interruption in membrane permeability of drug-resistant *Staphylococcus aureus* with cationic particles of nanosilver. *Eur. J. Pharm. Sci.* 127, 208–216. doi: 10.1016/j.ejps.2018.11.005

91. Costa SS., Sobkowiak, B., Parreira, R., Edgeworth, J. D., Viveiros, M., Clark, T. G., et al. (2018). Genetic diversity of norA, coding for a main efflux pump of *Staphylococcus aureus*. *Front. Genet.* 9:710. doi: 10.3389/fgene.2018.00710

92. Foster TJ. (2017). Antibiotic resistance in *Staphylococcus aureus*. Current status and future prospects. *FEMS Microb. Rev.* 41, 430–449. doi: 10.1093/femsre/fux007

93. Jang S. (2016). Multidrug efflux pumps in Staphylococcus aureus and their clinical implications. J. Microbiol. 54, 1–8. doi: 10.1007/s12275-016-5159-z

94. Lee, Y. D., and Park, J. H. (2016). Phage conversion for beta-lactam antibiotic resistance of Staphylococcus aureus from foods. J. Microbiol. Biotechnol. 26, 263–269. doi: 10.4014/jmb.1508.08042

95. Matono, T., Nagashima, M., Mezaki, K., Motohashi, A., Kutsuna, S., Hayakawa, K., et al. (2018). Molecular epidemiology of beta-lactamase production in penicillin-susceptible Staphylococcus aureus under high-susceptibility conditions. J. Infect. Chemother. 24, 153–155. doi: 10.1016/j.jiac.2017.10.014

96. Yang, J. J., Cheng, A., Tai, H. M., Chang, L. W., Hsu, M. C., and Sheng, W. H. (2019). Selected mutations by nemonoxacin and fluoroquinolone exposure among relevant gram-positive bacterial strains in Taiwan. Microb. Drug Resist. 26, 110–117. doi: 10.1089/mdr.2019.0048

97. Martinez, N., Luque, R., Milani, C., Ventura, M., Banuelos, O., and Margolles, A. (2018). A gene homologous to rRNA methylase genes confers erythromycin and clindamycin resistance in bifidobacterium breve. Appl. Environ. Microbiol. 84, 166–176. doi: 10.1128/AEM.02888-17

98. Chandrakanth, R.; Raju, S.; Patil, S. A. Aminoglycoside-resistance mechanisms in multidrug-resistant Staphylococcus aureus clinical isolates. Curr. Microbiol. 2008, 56, 558–562.

99. Emaneini, M.; Taherikalani, M.; Eslampour, M.A.; Sedaghat, H.; Aligholi, M.; Jabalameli, F.; Shahsavan, S.; Sotoudeh, N. Phenotypic and genotypic evaluation of aminoglycoside resistance in clinical isolates of staphylococci in Tehran, Iran. Microb. Drug Resist. 2009, 15, 129–132.

100. Melter, O.; Radojevicˇ, B. Small colony variants of Staphylococcus aureus Review.FoliaMicrobiol.2010,55,548–558.

101. Lowy, D. Antimicrobial resistance: The example of Staphylococcus aureus. J. Clin. Investig. 2003, 111, 1265–1273.

102. Miller WR, Munita JM, Arias CA. Mechanisms of antibiotic resistance in enterococci. Expert Rev Anti Infect Ther. 2014 Oct;12(10):1221-36. doi: 10.1586/14787210.2014.956092. PMID: 25199988; PMCID: PMC4433168.

103. Murray BE. The life and times of the Enterococcus. Clin Microbiol Rev. 1990; 3:46–65. [PubMed: 2404568]

104. Johnston, L. M., Jaykus, L.-A. (2004). Antimicrobial resistance of enterococcus species isolated from produce. Applied and Environmental Microbiology, 70(5), 3133–3137. https://doi.org/10.1128/aem.70.5.3133-3137.2004

105. Kuhn, I., A. Iversen, L. G. Burman, B. Olsson-Liljequist, A. Franklin, M. Finn, F. Aarestrup, A. M. Seyfarth, A. R. Blanch, H. Taylor, J. Caplin, M. A. Moreno, L. Dominguez, and R. Mollby 2000. Epidemiology and ecology of enterococci, with special reference to antibiotic resistance strains in animals, humans, and the environment. Example of an ongoing project within the European research programme. Int. J. Antimicrob. Agents 14:337–342.

106. Linden, P. K., and C. B. Miller. 1999. Vancomycin-resistant enterococci: the clinical effect of a common nosocomial pathogen. Diagn. Microbiol. Infect. Dis. 33:113–120.

107. Zhou, X., Willems, R. J., Friedrich, A. W., Rossen, J. W., & Bathoorn, E. (2020). Enterococcus faecium: From microbiological insights to practical recommendations for Infection Control and Diagnostics. Antimicrobial Resistance & Infection Control, 9(1). https://doi.org/10.1186/s13756-020-00770-1

108. Willems RJ, Top J, van Den Braak N, van Belkum A, Endtz H, Mevius D, et al. Host specificity of vancomycin-resistant Enterococcus faecium. J Infect Dis. 2000;182(3):816–23.

109. Galloway-Pena J, Roh JH, Latorre M, Qin X, Murray BE. Genomic and SNP Analyses Demonstrate a Distant Separation of the Hospital and Community- Associated Clades of Enterococcus faecium. PLoS One. 2012;7(1):e30187.

110. Leavis HL, Willems RJ, van Wamel WJ, Schuren FH, Caspers MP, Bonten MJ. Insertion sequence-driven diversification creates a globally dispersed emerging multiresistant subspecies of E. faecium. PLoS Pathog. 2007;3(1):e7.

111. Duez C, Hallut S, Rhazi N, et al. The ponA gene of Enterococcus faecalis JH2-2 codes for a low- affinity class A penicillin-binding protein. J Bacteriol. 2004; 186:4412–16. [PubMed: 15205448]

112. Murray BE. Beta-lactamase-producing enterococci. Antimicrob Agents Chemother. 1992; 36:2355–9. [PubMed: 1489177]

113. Bonten MJ, Willems R, Weinstein RA. Vancomycin-resistant enterococci: why are they here, and where do they come from? Lancet Infect Dis. 2001;1(5): 314–25.

114. Arias CA, Murray BE. The rise of the Enterococcus: beyond vancomycin resistance. Nat Rev Microbiol. 2012; 10:266–78. [PubMed: 22421879]

115. Boyd DA, Willey BM, Fawcett D, et al. Molecular characterization of Enterococcus faecalis N06-0364 with low-level vancomycin resistance harboring a novel D-Ala-D-Ser gene cluster, vanL. Antimicrob Agents Chemother. 2008; 52:2667–72. [PubMed: 18458129]

116. 30th ed. CLSI supplement M100. Wayne, PA: Clinical and Laboratory Standards Institute; 2020.

117. "The European Committee on Antimicrobial Susceptibility Testing. Breakpoint tables for interpretation of MICs and zone diameters. Version 12.0, 2022. http://www.eucast.org."

118. Jolley, K. A., Bray, J.E., Maiden M.C.J. (2018) Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res. September 24; 3:124. Doi: 10.12688/wellcomeopenres.14826.1. PMID: 30345391; PMCID: PMC6192448.

8 APPENDICES

Appendix 1

Table A1: Streptococcus Pneumoniae Dataset 1

ENA_accession	Penicillin (R-S-I or MIC (µg/mL))	Erythromycin (R-S-I or MIC (µg/mL))	Tetracycline (R-S-I or MIC (µg/mL))
ERR065287	<=0.03	<=0.06	<=0.50
ERR065288	<=0.03	<=0.06	<=0.50
ERR065289	<=0.03	<=0.06	<=0.50
ERR065290	<=0.03	<=0.06	<=0.50
ERR065291	0.06	<=0.06	<=0.50
ERR065292	<=0.03	<=0.06	<=0.50
ERR065293	0.25	<=0.06	<=0.50
ERR065294	2	>0.50	<=0.50
ERR065295	<=0.03	<=0.06	<=0.50
ERR065296	0.12	<=0.06	<=0.50
ERR065297	<=0.03	>0.50	<=0.50
ERR065298	<=0.03	<=0.06	<=0.50
ERR065299	<=0.03	<=0.06	<=0.50
ERR065300	1	<=0.06	<=0.50
ERR065301	1	<=0.06	<=0.50
ERR065302	<=0.03	<=0.06	<=0.50
ERR065303	0.12	<=0.06	<=0.50
ERR065304	<=0.03	<=0.06	<=0.50
ERR065305	<=0.03	<=0.06	<=0.50
ERR065306	<=0.03	<=0.06	<=0.50
ERR065307	<=0.03	<=0.06	<=0.50
ERR065308	<=0.03	<=0.06	<=0.50
ERR065309	<=0.03	>0.50	<=0.50
ERR065310	<=0.03	<=0.06	<=0.50
ERR065311	<=0.03	<=0.06	<=0.50
ERR065312	<=0.03	<=0.06	<=0.50
ERR065313	<=0.03	<=0.06	<=0.50
ERR065314	2	>0.50	>4.00
ERR065315	<=0.03	<=0.06	<=0.50
ERR065316	<=0.03	<=0.06	<=0.50
ERR065317	<=0.03	<=0.06	<=0.50
ERR065318	0.06	<=0.06	<=0.50
ERR065319	<=0.03	<=0.06	<=0.50
ERR065320	0.12	>0.50	>4.00

Table A1: Streptococcus Pneumoniae Dataset 1 (continue).

ERR065321	0.25	<=0.06	<=0.50
ERR065322	1	<=0.06	<=0.50
ERR065323	0.12	<=0.06	<=0.50
ERR065324	0.06	<=0.06	<=0.50
ERR065325	2	<=0.06	<=0.50
ERR065326	<=0.03	<=0.06	<=0.50
ERR065327	<=0.03	<=0.06	<=0.50
ERR065328	0.06	<=0.06	<=0.50
ERR065329	0.06	<=0.06	<=0.50
ERR065330	<=0.03	<=0.06	<=0.50
ERR065331	<=0.03	<=0.06	<=0.50
ERR065332	0.12	>0.50	>4.00
ERR065337	0.06	>0.50	<=0.50
ERR065338	0.06	<=0.06	<=0.50
ERR065339	1	<=0.06	<=0.50
ERR065340	0.06	>0.50	<=0.50
ERR065341	<=0.03	<=0.06	<=0.50
ERR065342	2	<=0.06	<=0.50
ERR065343	0.06	<=0.06	<=0.50
ERR065344	0.06	>0.50	<=0.50
ERR065345	<=0.03	<=0.06	<=0.50
ERR065346	<=0.03	<=0.06	<=0.50
ERR065347	<=0.03	<=0.06	<=0.50
ERR065348	<=0.03	<=0.06	<=0.50
ERR065349	<=0.03	<=0.06	<=0.50
ERR065350	<=0.03	<=0.06	<=0.50
ERR065351	1	<=0.06	<=0.50
ERR065352	1	>0.50	>4.00
ERR065353	0.12	>0.50	>4.00
ERR065354	<=0.03	<=0.06	<=0.50
ERR065355	<=0.03	<=0.06	<=0.50
ERR065953	0.25	>1	0.25
ERR065955	>4	NA	0.12
ERR065956	<0.03	<0.03	0.25
ERR065957	1	<0.03	0.12
ERR065958	<0.03	<0.03	0.25
ERR065959	<0.03	<0.03	0.25
ERR065960	<0.03	<0.03	0.25
ERR065962	0.06	<0.03	0.25
ERR065963	2	>1	0.25

Table A1: Streptococcus Pneumoniae Dataset 1 (continue).

ERR065964	<0.03	0.06	0.25
ERR065965	<0.03	>1	0.25
ERR065966	0.12	<0.03	0.12
ERR065967	0.06	<0.03	0.25
ERR065968	<0.03	<0.03	0.12
ERR065969	<0.03	<0.03	0.25
ERR065970	<0.03	<0.03	0.25
ERR065971	<0.03	<0.03	0.25
ERR065972	<0.03	<0.03	0.12
ERR065973	>4	>1	>8
ERR065974	<0.03	<0.03	0.25
ERR065975	0.06	<0.03	0.25
ERR067960	<=0.03	<=0.06	<=0.50
ERR067961	<=0.03	<=0.06	<=0.50
ERR067962	2	<=0.06	<=0.50
ERR067963	2	>0.50	<=0.50
ERR067964	<=0.03	<=0.06	<=0.50
ERR067965	0.06	<=0.06	<=0.50
ERR067966	2	>0.50	<=0.50
ERR067967	<=0.03	<=0.06	<=0.50
ERR067968	<=0.03	<=0.06	<=0.50
ERR067969	1	>0.50	<=0.50
ERR067970	2	>0.50	>4.00
ERR067972	<=0.03	<=0.06	<=0.50
ERR067973	<=0.03	<=0.06	<=0.50
ERR067974	<=0.03	<=0.06	<=0.50
ERR067975	<=0.03	<=0.06	<=0.50
ERR067976	<=0.03	<=0.06	<=0.50
ERR067977	<=0.03	<=0.06	<=0.50
ERR067978	0.25	>0.50	>4.00
ERR067979	0.12	<=0.06	<=0.50
ERR067980	<=0.03	<=0.06	<=0.50
ERR067981	<=0.03	<=0.06	<=0.50
ERR067982	2	>0.50	>4.00
ERR067983	<=0.03	<=0.06	<=0.50
ERR067984	0.25	<=0.06	<=0.50
ERR067985	<=0.03	<=0.06	<=0.50
ERR067986	<=0.03	<=0.06	<=0.50
ERR067987	0.06	<=0.06	<=0.50
ERR067988	2	<=0.06	<=0.50

Table A1: Streptococcus Pneumoniae Dataset 1 (continue).

ERR067989	<=0.03	<=0.06	<=0.50
ERR067990	4	>0.50	>4.00
ERR067991	<=0.03	<=0.06	<=0.50
ERR067992	1	<=0.06	<=0.50
ERR067994	<=0.03	<=0.06	<=0.50
ERR067995	0.06	<=0.06	<=0.50
ERR067996	<=0.03	<=0.06	<=0.50
ERR067997	<=0.03	<=0.06	<=0.50
ERR067998	0.25	>0.50	<=0.50
ERR067999	2	>0.50	>4.00
ERR068000	<=0.03	<=0.06	<=0.50
ERR068001	0.25	>0.50	<=0.50
ERR068002	2	<=0.06	<=0.50
ERR068003	1	>0.50	>4.00
ERR068004	0.25	>0.50	<=0.50
ERR068005	0.12	<=0.06	<=0.50
ERR068006	<=0.03	<=0.06	<=0.50
ERR068007	0.12	>0.50	<=0.50
ERR068008	<=0.03	<=0.06	<=0.50
ERR068009	2	<=0.06	<=0.50
ERR068011	<=0.03	<=0.06	<=0.50
ERR068012	0.12	<=0.06	<=0.50
ERR068013	<=0.03	<=0.06	<=0.50
ERR068014	<=0.03	<=0.06	<=0.50
ERR068015	<=0.03	<=0.06	<=0.50
ERR068016	<=0.03	<=0.06	<=0.50
ERR068017	<=0.03	<=0.06	<=0.50
ERR068018	<=0.03	<=0.06	<=0.50
ERR068019	1	<=0.06	<=0.50
ERR068020	<=0.03	<=0.06	<=0.50
ERR068021	<=0.03	<=0.06	<=0.50
ERR068022	<=0.03	<=0.06	<=0.50
ERR068023	0.12	<=0.06	<=0.50
ERR068024	0.06	<=0.06	<=0.50
ERR068025	<=0.03	<=0.06	<=0.50
ERR068026	0.25	>0.50	>4.00
ERR068027	<=0.03	<=0.06	<=0.50
ERR068028	0.25	>0.50	>4.00
ERR068029	2	<=0.06	<=0.50
ERR068030	0.12	>0.50	>4.00

Table A1: Streptococcus Pneumoniae Dataset 1 (continue).

ERR068031	>4.00	<=0.06	<=0.50
ERR068032	0.25	>0.50	>4.00
ERR068033	2	<=0.06	<=0.50
ERR068034	0.06	<=0.06	<=0.50
ERR068035	4	>0.50	>4.00
ERR068036	4	>0.50	>4.00
ERR068037	0.12	<=0.06	<=0.50
ERR068038	<=0.03	>0.50	<=0.50
ERR068039	<=0.03	<=0.06	<=0.50
ERR068041	<=0.03	<=0.06	<=0.50
ERR068042	0.25	<=0.06	<=0.50
ERR068043	0.25	>0.50	<=0.50
ERR068044	<=0.03	<=0.06	<=0.50
ERR068045	<=0.03	<=0.06	<=0.50
ERR068046	<=0.03	<=0.06	<=0.50
ERR068047	<=0.03	<=0.06	<=0.50
ERR068048	<=0.03	<=0.06	<=0.50
ERR068049	0.25	>0.50	>4.00
ERR068050	<=0.03	0.12	<=0.50
ERR069683	<=0.03	<=0.06	<=0.50
ERR069684	0.12	<=0.06	<=0.50
ERR069685	<=0.03	<=0.06	<=0.50
ERR069686	<=0.03	<=0.06	<=0.50
ERR069687	0.06	<=0.06	<=0.50
ERR069688	<=0.03	<=0.06	<=0.50
ERR069689	<=0.03	<=0.06	<=0.50
ERR069690	<=0.03	<=0.06	<=0.50
ERR069691	0.06	<=0.06	<=0.50
ERR069692	<=0.03	<=0.06	<=0.50
ERR069693	<=0.03	<=0.06	<=0.50
ERR069694	<=0.03	>0.50	<=0.50
ERR069695	<=0.03	<=0.06	<=0.50
ERR069696	0.12	<=0.06	<=0.50
ERR069697	<=0.03	<=0.06	<=0.50
ERR069698	0.25	>0.50	>4.00
ERR069699	2	>0.50	>4.00
ERR069700	4	>0.50	<=0.50
ERR069701	<=0.03	<=0.06	<=0.50
ERR069702	<=0.03	<=0.06	<=0.50
ERR069703	2	<=0.06	<=0.50

Table A1: Streptococcus Pneumoniae Dataset 1 (continue).

ERR069704	<=0.03	<=0.06	<=0.50
ERR069705	0.06	<=0.06	<=0.50
ERR069706	2	>0.50	>4.00
ERR069707	<=0.03	<=0.06	<=0.50
ERR069708	0.12	<=0.06	<=0.50
ERR069709	0.5	>0.50	<=0.50
ERR069710	0.25	>0.50	<=0.50
ERR069711	0.06	<=0.06	<=0.50
ERR069712	0.06	<=0.06	<=0.50
ERR069713	<=0.03	<=0.06	<=0.50
ERR069714	0.25	>0.50	<=0.50
ERR069715	0.06	<=0.06	<=0.50
ERR069716	<=0.03	<=0.06	<=0.50
ERR069717	0.12	<=0.06	<=0.50
ERR069718	0.06	<=0.06	<=0.50
ERR069719	0.12	<=0.06	<=0.50
ERR069720	>4.00	>0.50	>4.00
ERR069721	0.12	<=0.06	<=0.50
ERR069722	<=0.03	<=0.06	<=0.50
ERR069723	0.06	<=0.06	<=0.50
ERR069724	0.25	>0.50	>4.00
ERR069725	0.25	>0.50	>4.00
ERR069726	2	<=0.06	<=0.50
ERR069727	0.06	<=0.06	<=0.50
ERR069728	2	<=0.06	<=0.50
ERR069729	0.06	<=0.06	<=0.50
ERR069730	0.06	<=0.06	<=0.50
ERR069731	<0.03	<0.03	0.25
ERR069732	<0.03	<0.03	0.25
ERR069733	2	>1	0.25
ERR069734	0.5	>1	0.25
ERR069735	4	>1	0.5
ERR069736	<0.03	NA	NA
ERR069737	0.06	<0.03	0.12
ERR069738	<0.03	<0.03	0.25
ERR069739	<0.03	<0.03	0.25
ERR069740	0.06	>1	0.12
ERR069741	0.12	<0.03	0.5
ERR069742	2	>1	>8
ERR069743	0.06	<0.03	0.25
ERR069744	4	>1	0.25

Table A1: Streptococcus Pneumoniae Dataset 1 (continue).

ERR069745	0.12	NA	NA
ERR069746	0.06	NA	0.25
ERR069747	4	>1	0.25
ERR069748	<0.03	<0.03	0.12
ERR069749	2	<0.03	0.25
ERR069750	<0.03	<0.03	0.25
ERR069751	<0.03	<0.03	0.12
ERR069752	<0.03	NA	NA
ERR069753	0.12	<0.03	0.25
ERR069755	0.06	<0.03	0.25
ERR069757	<0.03	<0.03	0.25
ERR069758	<0.03	<0.03	0.12
ERR069759	0.06	<0.03	0.25
ERR069760	0.06	<0.03	0.25
ERR069761	0.06	0.5	>8
ERR069762	2	<0.03	0.12
ERR069763	2	>1	0.25
ERR069764	<0.03	<0.03	0.25
ERR069765	0.06	<0.03	0.25
ERR069766	<0.03	<0.03	0.25
ERR069767	<0.03	<0.03	0.25
ERR069768	0.06	<0.03	2
ERR069769	<0.03	<0.03	0.25
ERR069770	0.06	<0.03	0.25
ERR069771	0.12	<0.03	0.25
ERR069772	<0.03	<0.03	0.25
ERR069773	<0.03	<0.03	0.25
ERR069774	<0.03	<0.03	0.25
ERR069775	0.06	<0.03	0.25
ERR069776	4	0.06	0.25
ERR069777	<0.03	>1	0.25
ERR069778	<0.03	<0.03	0.12
ERR069779	0.06	<0.03	0.25
ERR069780	<0.03	<0.03	0.25
ERR069781	0.12	<0.03	0.25
ERR069782	2	0.06	1
ERR069783	<0.03	<0.03	0.12
ERR069784	0.5	>1	0.5
ERR069785	4	1	>8
ERR069786	<0.03	<0.03	0.25
ERR069787	2	>1	0.25

Table A1: Streptococcus Pneumoniae Dataset 1 (continue).

ERR069788	0.5	>1	0.5
ERR069789	<0.03	<0.03	0.25
ERR069790	<0.03	<0.03	0.12
ERR069791	2	0.25	>8
ERR069792	<0.03	<0.03	0.25
ERR069793	2	>1	>8
ERR069794	2	<0.03	0.12
ERR069795	<0.03	<0.03	0.25
ERR069796	0.12	<0.03	0.12
ERR069797	>4	>1	>8
ERR069799	4	<0.03	0.5
ERR069800	<0.03	<0.03	<0.06
ERR069801	<0.03	<0.03	<0.06
ERR069802	<0.03	<0.03	0.25
ERR069804	<0.03	<0.03	0.25
ERR069805	0.12	<0.03	0.12
ERR069806	<0.03	0.06	0.25
ERR069807	<0.03	<0.03	<0.06
ERR069808	0.06	<0.03	>8
ERR069809	<0.03	<0.03	NA
ERR069810	0.06	<0.03	0.25
ERR069811	<0.03	<0.03	0.25
ERR069812	0.06	<0.03	0.12
ERR069813	<0.03	>1	0.5
ERR069814	<0.03	<0.03	0.12
ERR069815	<0.03	<0.03	0.25
ERR069816	<0.03	<0.03	0.12
ERR069817	0.12	>1	0.12
ERR069818	4	<0.03	0.25
ERR069819	<0.03	<0.03	0.5
ERR069820	<0.03	<0.03	0.25
ERR069821	>4	>1	>8
ERR069822	<0.03	<0.03	0.25
ERR069823	<0.03	<0.03	0.25
ERR069824	<0.03	<0.03	0.25
ERR069825	1	<0.03	0.12
ERR069826	<0.03	0.06	0.25
ERR069827	0.5	>1	0.25
ERR069829	<0.03	<0.03	0.12
ERR069830	>4	<0.03	0.25

Table A1: Streptococcus Pneumoniae Dataset 1 (continue).

ERR069831	0.25	<0.03	NA
ERR069832	<0.03	0.06	0.5
ERR069833	<0.03	<0.03	0.12
ERR069834	4	NA	0.5
ERR069835	<0.03	<0.03	0.12
ERR069836	<0.03	NA	NA
ERR069837	<0.03	<0.03	0.25
ERR069838	0.25	>1	0.25
ERR069839	0.06	<0.03	0.25
ERR069840	0.25	>1	>8
ERR069841	<0.03	1	0.12
ERR069842	4	>1	>8
ERR069843	4	>1	>8
ERR124217	0.023	0.064	NA
ERR124218	0.023	0.125	NA
ERR124219	0.023	0.064	NA
ERR124220	0.023	0.032	NA
ERR124221	0.19	>256	NA
ERR124222	6	>256	NA
ERR124223	0.016	0.064	NA
ERR124224	0.023	0.094	NA
ERR124225	0.032	0.094	NA
ERR124226	0.023	0.094	NA
ERR124227	0.125	0.064	NA
ERR124228	0.023	0.064	NA
ERR124229	0.032	0.125	NA
ERR124230	0.023	0.094	NA
ERR124231	0.064	0.094	NA
ERR124232	0.023	0.064	NA
ERR124233	4	>256	NA
ERR124234	0.094	0.064	NA
ERR124235	0.38	>256	NA
ERR124236	2	0.047	NA
ERR124237	0.016	0.064	NA
ERR124238	0.023	0.094	NA
ERR124239	0.38	>256	NA
ERR124240	0.023	0.064	NA
ERR124241	1.5	0.032	NA
ERR124242	0.023	0.064	NA
ERR124243	0.023	0.094	NA

Table A1: Streptococcus Pneumoniae Dataset 1 (continue).

ERR124244	0.38	0.064	NA
ERR124245	0.023	0.047	NA
ERR124246	0.023	0.064	NA
ERR124247	0.032	0.094	NA
ERR124248	0.023	0.047	NA
ERR124249	0.19	>256	NA
ERR124250	1	6	NA
ERR124251	0.023	0.094	NA
ERR124252	0.125	0.094	NA
ERR124253	0.023	0.094	NA
ERR124254	0.023	0.094	NA
ERR124255	4	>256	NA
ERR124256	0.023	0.094	NA
ERR124257	0.023	4	NA
ERR124258	0.023	0.064	NA
ERR124260	0.032	0.064	NA
ERR124263	0.023	3	NA
ERR124264	0.016	0.032	NA
ERR124265	0.016	0.032	NA
ERR124266	0.5	1.5	NA
ERR124267	0.023	6	NA
ERR124268	0.023	0.064	NA
ERR124269	0.094	0.047	NA
ERR124270	0.094	8	NA
ERR124271	<.016	0.064	NA
ERR124272	0.19	>256	NA
ERR124273	1.5	0.047	NA
ERR124274	0.25	>256	NA
ERR124275	1.5	0.047	NA
ERR124276	0.016	0.064	NA
ERR124277	1.5	0.094	NA
ERR124278	<0.016	0.094	NA
ERR124279	1.5	0.094	NA
ERR124280	0.023	0.094	NA
ERR124281	0.032	0.094	NA
ERR124282	0.023	0.094	NA
ERR124283	0.25	>256	NA
ERR124284	0.125	0.094	NA
ERR124285	0.023	0.094	NA

Table A1: Streptococcus Pneumoniae Dataset 1 (continue).

ERR124286	0.25	0.5	NA
ERR124287	0.023	0.094	NA
ERR124288	0.25	0.25	NA
ERR124289	0.023	0.064	NA
ERR124290	1.5	0.064	NA
ERR124291	0.032	0.032	NA
ERR124292	0.023	0.094	NA
ERR124293	1.5	0.064	NA
ERR124294	0.25	0.38	NA
ERR124295	0.125	0.064	NA
ERR124296	0.023	0.047	NA
ERR124297	0.38	0.25	NA
ERR124298	0.016	0.094	NA
ERR124299	1.5	0.064	NA
ERR124300	0.25	32	NA
ERR124301	1	0.094	NA
ERR124302	0.023	0.047	NA
ERR124303	1	3	NA
ERR124304	0.023	0.064	NA
ERR124305	0.5	1.5	NA
ERR124306	<.016	0.047	NA
ERR124307	0.016	0.047	NA
ERR124308	0.094	0.047	NA
ERR124309	0.016	0.032	NA
ERR124310	0.5	1.5	NA
ERR124311	0.023	0.047	NA
ERR124312	0.016	0.047	NA
ERR124313	0.023	3	NA
ERR124314	2	3	NA
ERR124315	0.094	0.094	NA
ERR124316	1.5	0.064	NA
ERR124317	0.023	0.094	NA
ERR124318	0.023	0.094	NA
ERR124319	0.38	0.5	NA
ERR124320	4	<0.03	0.25
ERR129025	0.094	0.094	NA
ERR129026	0.19	>256	NA
ERR129027	0.047	0.064	NA
ERR129028	0.094	0.094	NA
ERR129029	0.047	0.064	NA

Table A1: Streptococcus Pneumoniae Dataset 1 (continue).

ERR129030	0.094	0.064	NA
ERR129031	3	>256	NA
ERR129032	0.023	0.094	NA
ERR129033	0.023	0.064	NA
ERR129034	0.023	0.047	NA
ERR129035	1.5	0.064	NA
ERR129036	3	>256	NA
ERR129037	0.023	0.047	NA
ERR129038	0.023	0.064	NA
ERR129039	0.023	0.064	NA
ERR129040	0.016	0.016	NA
ERR129041	0.016	0.047	NA
ERR129042	0.064	0.047	NA
ERR129043	0.125	0.032	NA
ERR129045	0.032	0.047	NA
ERR129046	0.023	0.064	NA
ERR129047	0.023	0.032	NA
ERR129048	0.016	0.064	NA
ERR129049	3	>256	NA
ERR129050	0.023	0.094	NA
ERR129051	0.016	0.064	NA
ERR129052	0.023	0.094	NA
ERR129053	0.38	0.094	NA
ERR129054	0.064	0.094	NA
ERR129055	0.023	0.094	NA
ERR129056	0.016	0.064	NA
ERR129057	0.023	0.064	NA
ERR129058	0.016	0.047	NA
ERR129059	0.125	0.064	NA
ERR129060	0.064	>256	NA
ERR129061	0.125	>256	NA
ERR129062	1	>256	NA
ERR129063	0.094	0.094	NA
ERR129064	0.023	0.047	NA
ERR129065	0.023	0.047	NA
ERR129066	0.016	0.064	NA
ERR129068	0.016	0.047	NA
ERR129071	0.023	0.094	NA
ERR129072	0.125	0.094	NA
ERR129073	0.023	0.094	NA
ERR129074	1.5	0.064	NA

Table A1: Streptococcus Pneumoniae Dataset 1 (continue).

ERR129075	6	4	NA
ERR129076	0.064	0.064	NA
ERR129077	0.023	0.094	NA
ERR129078	3	4	NA
ERR129079	0.023	0.032	NA
ERR129080	0.023	0.032	NA
ERR129081	0.032	0.094	NA
ERR129082	0.023	0.094	NA
ERR129083	0.064	0.064	NA
ERR129084	0.016	0.047	NA
ERR129086	1.5	0.047	NA
ERR129087	1.5	4	NA
ERR129088	0.023	0.064	NA
ERR129089	1.5	0.094	NA
ERR129090	0.016	0.094	NA
ERR129091	0.023	0.094	NA
ERR129092	0.38	3	NA
ERR129093	0.023	0.064	NA
ERR129094	0.016	0.064	NA
ERR129095	0.19	>256	NA
ERR129096	0.016	0.047	NA
ERR129098	0.023	0.094	NA
ERR129099	1.5	0.064	NA
ERR129100	0.023	0.064	NA
ERR129102	0.047	1.5	NA
ERR129103	0.094	0.094	NA
ERR129104	1.5	0.064	NA
ERR129105	0.125	8	NA
ERR129106	0.19	12	NA
ERR129107	0.023	0.064	NA
ERR129108	0.5	6	NA
ERR129110	0.023	0.047	NA
ERR129111	0.016	0.064	NA
ERR129112	0.094	0.064	NA
ERR129113	0.016	0.064	NA
ERR129115	0.032	0.064	NA
ERR129116	0.023	0.094	NA
ERR129117	0.125	0.064	NA
ERR129118	2	0.064	NA
ERR129119	1.5	0.064	NA
ERR129120	0.094	0.064	NA

Table A1: Streptococcus Pneumoniae Dataset 1 (continue).

ERR129121	0.023	0.047	NA
ERR129122	0.016	0.064	NA
ERR129123	0.094	0.064	NA
ERR129124	0.016	0.064	NA
ERR129125	0.016	0.064	NA
ERR129126	0.023	0.094	NA
ERR129127	0.023	0.094	NA
ERR129128	3	12	NA
ERR129129	0.023	3	NA
ERR129130	<0.016	0.064	NA
ERR129131	0.023	0.064	NA
ERR129132	0.023	0.094	NA
ERR129133	0.023	0.094	NA
ERR129134	2	0.064	NA
ERR129135	0.032	0.094	NA
ERR129136	0.094	0.094	NA
ERR129137	0.032	6	NA
ERR129138	0.032	0.094	NA
ERR129139	0.032	0.094	NA
ERR129140	0.125	0.094	NA
ERR129141	0.023	0.094	NA
ERR129142	0.094	0.094	NA
ERR129143	0.75	0.064	NA
ERR129144	0.023	0.064	NA
ERR129145	0.5	2	NA
ERR129146	0.023	0.064	NA
ERR129147	0.023	0.047	NA
ERR129148	1	0.047	NA
ERR129149	0.023	0.094	NA
ERR129151	0.023	0.032	NA
ERR129152	0.023	0.064	NA
ERR129153	0.094	0.064	NA
ERR129154	0.023	0.064	NA
ERR129155	0.023	0.064	NA
ERR129156	0.016	0.064	NA
ERR129157	0.032	0.064	NA
ERR129158	0.023	0.047	NA
ERR129159	0.023	0.032	NA
ERR129160	0.25	4	NA
ERR129161	0.016	0.032	NA
ERR129162	0.023	0.064	NA

Table A1: Streptococcus Pneumoniae Dataset 1 (continue).

ERR129163	1.5	0.064	NA
ERR129164	0.032	0.064	NA
ERR129167	0.094	0.064	NA
ERR129168	< 0.01	0.064	NA
ERR129169	0.032	0.064	NA
ERR129170	0.023	0.094	NA
ERR129171	0.047	0.094	NA
ERR129172	0.023	0.047	NA
ERR129173	1.5	16	NA
ERR129174	0.023	0.047	NA
ERR129175	0.032	0.047	NA
ERR129176	0.094	0.032	NA
ERR129177	0.023	0.032	NA
ERR129178	0.016	0.032	NA
ERR129179	2	0.094	NA
ERR129180	0.016	0.064	NA
ERR129181	0.016	0.064	NA
ERR129182	0.023	0.064	NA
ERR129183	0.023	0.064	NA
ERR129184	0.023	0.094	NA
ERR129185	0.023	2	NA
ERR129186	0.023	8	NA
ERR129187	<0.016	0.094	NA
ERR129188	0.023	0.094	NA
ERR129190	0.023	0.094	NA
ERR129192	0.023	0.094	NA
ERR129193	4	>256	NA
ERR129195	0.19	12	NA
ERR129196	0.023	0.064	NA
ERR129197	0.094	0.064	NA
ERR129198	0.38	>256	NA
ERR129199	0.023	0.094	NA
ERR129200	1.5	>256	NA
ERR129201	0.016	0.094	NA
ERR129202	0.094	0.094	NA
ERR129203	0.023	0.094	NA
ERR129204	0.032	0.64	NA
ERR129205	0.032	0.064	NA
ERR129206	0.023	0.064	NA
ERR129207	0.023	0.047	NA
ERR129208	0.023	0.064	NA

Table A1: Streptococcus Pneumoniae Dataset 1 (continue).

ERR129209	0.125	0.094	NA
ERR129210	0.023	0.064	NA
ERR129211	0.023	0.064	NA
ERR129212	0.016	0.064	NA
ERR129213	0.125	0.047	NA
ERR129214	0.023	0.023	NA
ERR129215	0.016	0.047	NA
ERR129216	0.094	0.094	NA

Table A2: Streptococcus Pneumoniae Dataset 2

ENA_accession	Penicillin (R-S-I)	Erythromycin (R-S-I)	Tetracycline (R-S-I)
ERR028590	R	NA	NA
ERR028591	R	R	NA
ERR028592	R	NA	NA
ERR028594	R	NA	R
ERR028595	R	R	R
ERR028596	R	R	R
ERR028597	R	R	R
ERR028598	R	R	NA
ERR028599	R	NA	R
ERR028600	R	R	S
ERR028601	R	R	R
ERR028734	S	R	R
ERR028735	R	R	R
ERR028736	R	S	R
ERR028737	R	R	R
ERR028738	S	NA	R
ERR028739	R	R	R
ERR028740	R	R	I
ERR028741	R	R	S
ERR028742	R	R	NA
ERR028743	R	NA	R
ERR028744	R	S	S
ERR028745	R	NA	R
ERR028747	R	R	R
ERR028748	R	R	NA
ERR028749	R	R	R
ERR028750	R	R	R
ERR028751	R	NA	R
ERR028752	R	R	R
ERR028753	R	R	NA
ERR028754	R	R	R
ERR028755	R	R	R
ERR028756	R	R	R
ERR028757	R	I	R
ERR028758	R	R	R
ERR028760	R	R	R
ERR028764	R	R	R
ERR028766	R	R	R
ERR028767	R	R	R
ERR028768	R	R	S
ERR028769	R	R	R
ERR028770	R	R	NA
ERR028771	R	R	R
ERR051799	I	R	R
ERR051829	S	R	NA
ERR051863	R	NA	R
ERR051883	S	R	R
ERR051888	R	I	NA
ERR051889	R	R	R

Table A3: Streptococcus Pneumoniae Dataset 3 (continue).

ERR051898	S	S	S
ERR051900	S	R	R
ERR051921	S	R	R
ERR051924	S	S	NA
ERR051930	S	R	R
ERR051938	S	R	R
ERR051948	S	R	R
ERR051950	R	R	R
ERR051951	R	NA	R
ERR051955	S	R	NA
ERR051960	S	NA	NA
ERR051963	S	R	R
ERR051965	S	R	NA
ERR051966	R	R	R
ERR051968	S	R	R
ERR052017	S	NA	R
ERR052023	S	R	R
ERR052038	S	I	NA
ERR052043	S	R	R
ERR052325	S	R	R
ERR052331	S	R	R
ERR052390	S	R	R
ERR052391	S	R	R
ERR052392	S	R	R
ERR052393	S	R	R
ERR052394	S	R	R
ERR052395	S	R	NA
ERR052396	S	R	R
ERR052397	S	R	R
ERR052398	S	R	R
ERR052399	S	R	R
ERR052400	S	NA	R
ERR052401	S	S	R
ERR052402	S	R	R
ERR052438	S	R	NA
ERR052439	S	R	S
ERR052440	S	R	R
ERR052441	S	NA	R
ERR052452	S	S	R
ERR052456	S	R	R
ERR052464	S	S	NA
ERR052465	S	R	R
ERR052466	S	R	R
ERR052467	S	S	S
ERR052505	S	R	R
ERR052506	S	NA	R
ERR052511	S	R	R
ERR055616	I	NA	R
ERR055617	S	R	R
ERR107482	S	R	NA
ERR107483	S	S	R
ERR107484	S	NA	R
ERR107485	S	NA	R

Table A3: Streptococcus Pneumoniae Dataset 3 (continue).

ERR107486	S	R	NA
ERR107487	S	R	NA
ERR107489	S	R	R
ERR107490	S	R	R
ERR107491	S	R	R
ERR107492	S	R	R
ERR107493	S	R	R
ERR107494	S	R	NA
ERR107495	S	R	R
ERR107496	S	R	R
ERR107497	S	R	R
ERR107498	S	R	R
ERR107499	S	S	R
ERR107500	S	R	R
ERR107501	S	S	R
ERR107502	S	R	R
ERR107503	S	R	R
ERR107504	S	NA	R
ERR107505	S	R	R
ERR107506	S	NA	R
ERR107507	S	R	R
ERR107508	S	NA	NA
ERR107509	S	R	R
ERR107510	S	R	R
ERR107511	S	NA	R
ERR107512	S	S	NA
ERR107513	S	R	R
ERR107514	S	NA	R
ERR107515	S	NA	NA
ERR107516	S	NA	NA
ERR107517	S	NA	NA
ERR107518	S	NA	NA
ERR107519	S	NA	NA
ERR107520	S	NA	NA
ERR107521	S	R	R
ERR107523	S	NA	NA
ERR107524	S	NA	NA
ERR107525	S	NA	NA
ERR107528	S	NA	NA
ERR107529	S	R	R
ERR107530	S	R	R
ERR107531	S	R	R
ERR107532	S	R	R
ERR107533	S	R	R
ERR107534	S	R	R
ERR107535	S	R	R
ERR107536	S	R	R
ERR107537	S	R	R
ERR107538	S	R	R
ERR107539	S	R	R
ERR107540	S	R	S
ERR107541	S	NA	NA
ERR107542	S	NA	NA

Table A3: Streptococcus Pneumoniae Dataset 3 (continue).

ERR107543	S	I	R
ERR107544	S	R	R
ERR107545	S	R	R
ERR107546	S	R	R
ERR107547	S	R	R
ERR107548	S	R	R
ERR107549	S	S	S
ERR107550	S	S	S
ERR107561	S	S	S
ERR107562	S	R	R
ERR107563	S	S	R
ERR107564	S	S	R
ERR107565	S	R	R
ERR107566	S	R	R
ERR107567	S	R	NA
ERR107568	S	NA	NA
ERR107569	S	NA	NA
ERR107570	S	NA	NA
ERR107571	S	NA	NA
ERR107572	S	NA	NA
ERR271773	R	NA	NA
ERR271785	R	NA	NA
ERR271792	S	NA	NA
ERR271793	S	NA	NA

Table A4: Stapylococcus Aerous Dataset

AccessionIDs	Erythromycin	Tetracycline	Gentamicin	Ciprofloxacin
ERS3476315	R	R	S	R
ERS3476310	S	S	S	R
ERS3476327	R	R	S	R
ERS3476329	R	R	S	S
ERS3476333	R	R	S	S
ERS3476334	R	R	S	S
ERS3476337	R	R	S	S
ERS3476347	R	R	S	R
ERS3476354	S	R	S	R
ERS3476361	R	R	S	R
ERS3476363	R	R	S	R
ERS3476364	R	R	S	R
ERS3476365	R	R	S	R
ERS3476372	R	R	S	R
ERS3476373	R	R	S	R
ERS3476378	S	R	S	R
ERS3476380	R	S	R	R
ERS3476392	S	R	S	R
ERS3476394	R	R	S	R
ERS3476395	R	R	S	R
ERS3476402	S	S	R	R
ERS3476407	R	R	S	R
ERS3476411	R	R	R	R
ERS3476415	R	S	S	R
ERS3476427	S	R	R	S

Table A4: Stapylococcus Aerous Dataset (continue)

ERS3476433	R	R	S	S
ERS3476437	S	R	S	R
ERS3476438	R	R	S	R
ERS3476439	R	R	S	S
ERS3476440	R	R	S	R
ERS3476450	R	R	S	S
ERS3476457	S	S	R	S
ERS3476471	R	R	S	S
ERS3476479	S	R	S	S
ERS3476486	R	S	S	R
ERS3476495	S	S	R	S
ERS3476500	R	R	S	R
ERS3476501	S	S	R	R
ERS3476503	R	S	R	R
ERS3476512	R	S	S	R
ERS3476515	S	S	S	R
ERS3476520	R	R	S	S
ERS3476522	R	R	S	R
ERS3476523	R	R	S	R
ERS3476528	R	S	S	R
ERS3476529	R	R	S	R
ERS3476532	R	R	S	R
ERS3476538	R	R	S	R
ERS3476539	R	S	R	R
ERS3476541	R	R	S	R
ERS3476542	R	S	R	R
ERS3476545	R	S	R	R
ERS3476547	S	S	R	R
ERS3476585	R	S	S	R
ERS3476567	R	R	S	R
ERS3476571	R	R	S	S
ERS3476588	R	R	S	R
ERS3476595	R	R	S	R
ERS3476596	R	R	S	S
ERS3476598	R	S	R	R
ERS3476601	S	R	S	S
ERS3476619	S	R	S	S
ERS3476626	R	S	R	R
ERS3476630	R	R	S	R
ERS3476635	S	R	S	S
ERS3476640	S	R	S	R
ERS3476649	R	S	R	R
ERS3476651	R	R	S	R
ERS3476656	S	R	R	S
ERS3476657	R	S	R	R
ERS3476664	R	S	R	R
ERS3476670	R	R	S	S
ERS3476672	S	R	S	R
ERS3476673	R	S	R	R
ERS3476674	S	R	S	R
ERS3476681	R	S	R	R
ERS3476684	S	S	S	R
ERS3476697	R	S	S	R

Table A4: Stapylococcus Aerous Dataset (continue)

ERS3476698	S	R	R	S
ERS3476700	R	R	S	S
ERS3476703	S	R	R	R
ERS3476704	R	R	R	S
ERS3476712	S	R	S	S
ERS3476724	R	S	R	R
ERS3476730	S	R	S	S
ERS3476740	S	S	R	R
ERS3476743	S	R	R	S
ERS3476745	R	S	R	R
ERS3476746	R	S	R	R
ERS3476754	R	R	S	S
ERS3476764	R	S	R	R
ERS3476777	R	R	S	S
ERS3476781	R	S	R	R
ERS3476789	S	R	S	S
ERS3476795	R	R	S	S
ERS3476797	R	S	S	R
ERS3476799	R	S	S	R
ERS3476801	R	R	S	S
ERS3476809	R	R	S	R
ERS3476810	R	S	R	R
ERS3476834	S	R	S	S
ERS3476835	R	S	S	R
ERS3476836	S	R	S	S
ERS3476848	R	S	S	R
ERS3476852	R	S	R	R
ERS3476870	S	S	S	R
ERS3476872	S	R	S	S
ERS3476887	R	R	S	R
ERS3476897	R	R	R	S
ERS3476903	R	R	S	R
ERS3476905	R	S	S	R
ERS3476908	S	R	S	S
ERS3476916	S	S	R	R
ERS3476918	R	S	S	R
ERS3476919	S	R	S	S
ERS3476924	S	R	S	S
ERS3476926	R	R	S	S
ERS3476943	S	S	S	R
ERS3476948	S	R	R	S
ERS3476969	S	R	R	R
ERS3476971	S	S	R	R
ERS3476976	R	R	S	S
ERS3476982	R	S	R	R
ERS3476989	R	S	R	S
ERS3476996	R	S	R	R
ERS3477000	S	S	R	R
ERS3477003	S	S	R	S
ERS3477004	S	S	R	R
ERS3477005	S	R	S	S
ERS3477007	S	R	S	R
ERS3477008	R	R	R	S

Table A4: Stapylococcus Aerous Dataset (continue)

ERS3477009	R	R	S	S
ERS3477014	S	R	S	R
ERS3477031	S	S	S	R
ERS3477034	R	R	S	R
ERS3477035	R	R	S	S
ERS3477037	R	R	S	R
ERS3477043	S	R	R	S
ERS3477047	R	S	R	R
ERS3477048	S	R	S	S
ERS3477056	S	R	R	S
ERS3477057	R	R	R	R
ERS3477058	R	S	S	S
ERS3477060	S	S	S	R
ERS3477067	S	S	S	R
ERS3477068	S	S	S	R
ERS3477073	R	R	S	S
ERS3477076	R	R	S	S
ERS3477090	S	R	S	R
ERS3477083	R	R	S	R
ERS3477088	R	S	S	R

Table A5: Enterococcus Faecium Dataset

Accession number	**Ampicillin resistance**	**Vancomycin resistance**
ERR369948	R	R
ERR369951	R	R
ERR369954	R	R
ERR369956	R	R
ERR369958	R	R
ERR369962	R	R
ERR369965	R	R
ERR369967	R	R
ERR369969	R	R
ERR369971	R	R
ERR369972	R	R
ERR369974	R	R
ERR369980	R	R
ERR369981	R	R
ERR369982	R	R
ERR369983	R	R
ERR369985	R	R
ERR369991	R	R
ERR369992	R	R
ERR369993	R	R
ERR369995	R	R
ERR369996	R	R
ERR369999	R	R
ERR370000	R	R
ERR370001	R	R
ERR370003	R	R
ERR369944	R	R
ERR370007	R	R

Table A5: Enterococcus Faecium Dataset (continue)

ERR370011	R	R
ERR370013	R	R
ERR370018	R	R
ERR370021	R	R
ERR370024	R	R
ERR370025	R	R
ERR369946	R	R
ERR370029	R	R
ERR370031	R	R
ERR370033	R	R
ERR374660	R	R
ERR374662	R	R
ERR374663	R	R
ERR374664	R	R
ERR374668	R	R
ERR374670	R	R
ERR374671	R	R
ERR374673	R	R
ERR374674	R	R
ERR374682	R	R
ERR374689	R	R
ERR369939	R	S
ERR369964	R	S
ERR369966	R	S
ERR369940	R	S
ERR369976	R	S
ERR369978	R	S
ERR369990	R	S
ERR369997	R	S
ERR369998	R	S
ERR370005	R	S
ERR370012	R	S
ERR370015	R	S
ERR369945	R	S
ERR370019	R	S
ERR370023	R	S
ERR370030	R	S
ERR370032	R	S
ERR374669	R	S
ERR374675	R	S
ERR374676	R	S
ERR374679	R	S
ERR374680	R	S
ERR374681	R	S
ERR374683	R	S
ERR374685	R	S
ERR374686	R	S
ERR374688	R	S
ERR374692	R	S
ERR374694	R	S
ERR374696	R	S
ERR374703	R	S
ERR374705	R	S

Table A5: Enterococcus Faecium Dataset (continue)

ERR374714	R	S
ERR375304	R	S
ERR375333	R	S
ERR375335	R	S
ERR375339	R	S
ERR375343	R	S
ERR375347	R	S
ERR375361	R	S
ERR375363	R	S
ERR375365	R	S
ERR375371	R	S
ERR375377	R	S
ERR375381	R	S
ERR375382	R	S
ERR375385	R	S
ERR375389	R	S
ERR374726	R	S
ERR374728	R	S
ERR369955	S	S
ERR369957	S	S
ERR374678	S	S
ERR375323	S	S
ERR375376	S	S
ERR374734	S	S
ERR374720	S	S
ERR374766	S	S
ERR374778	S	S
ERR374785	S	S
ERR374902	S	S
ERR374951	S	S
ERR374956	S	S
ERR375047	S	S
ERR375069	S	S
ERR375115	S	S
ERR375150	S	S
ERR375176	S	S
ERR375189	S	S
ERR375443	S	S
ERR375541	S	S
ERR377516	S	S
ERR374713	S	R
ERR374733	S	R

Appendix 2

Table A6: The best parameters selected for each antibiotic in the final models in *S.pneumoniae*

Algorithm	Antibiotic	input	Final Parameters
Random Forest	Penicillin	k-mer	mtry = 26
Support Vector Machine	Penicillin	k-mer	C = 0.88
Stochastic Gradient Boosting	Penicillin	k-mer	n.trees = 4808, interaction.depth = 6, shrinkage = 0.03587614 and n.minobsinnode = 19
Extreme Gradient Boosting	Penicillin	k-mer	nrounds = 150, max_depth = 5, eta = 0.3, gamma = 0, colsample_bytree = 0.5, min_child_weight = 1 subsample = 1
Random Forest	Penicillin	AA k-mer	mtry = 15
Support Vector Machine	Penicillin	AA k-mer	C = 13
Stochastic Gradient Boosting	Penicillin	AA k-mer	n.trees = 1822, interaction.depth = 5, shrinkage = 0.08477812 and n.minobsinnode = 10
Extreme Gradient Boosting	Penicillin	AA k-mer	nrounds = 50, max_depth = 2, eta = 0.3, gamma = 0, colsample_bytree = 0.6, min_child_weight = 1 and subsample = 0.75.
Random Forest	Penicillin	SNP	mtry = 11
Support Vector Machine	Penicillin	SNP	sigma = 0.02192786 and C = 1.324154
Stochastic Gradient Boosting	Penicillin	SNP	n.trees = 150, interaction.depth = 2, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Penicillin	SNP	nrounds = 50, max_depth = 3, eta = 0.4, gamma = 0, colsample_bytree = 0.45, min_child_weight = 1 subsample = 0.5
Random Forest	Penicillin	SNP/AA k-mer	mtry = 30
Support Vector Machine	Penicillin	SNP/AA k-mer	C = 0.10
Stochastic Gradient Boosting	Penicillin	SNP/AA k-mer	n.trees = 500, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10.
Extreme Gradient Boosting	Penicillin	SNP/AA k-mer	nrounds = 50, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.3, min_child_weight = 2 and subsample = 1
Random Forest	Penicillin	SNP/k-mer	mtry = 10
Support Vector Machine	Penicillin	SNP/k-mer	C = 0.10
Stochastic Gradient Boosting	Penicillin	SNP/k-mer	n.trees = 100, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10

Table A6: The best parameters selected for each antibiotic in the final models in *S.pneumoniae* (continue)

Extreme Gradient Boosting	Penicillin	SNP/k-mer	nrounds = 100, max_depth = 2, eta = 0.3, gamma = 0, colsample_bytree = 0.3, min_child_weight = 1 and subsample = 1
Random Forest	Penicillin	AA k-mer/k-mer	mtry = 27
Support Vector Machine	Penicillin	AA k-mer/k-mer	sigma = 0.0008980129 and C = 336.0581
Stochastic Gradient Boosting	Penicillin	AA k-mer/k-mer	n.trees = 1564, interaction.depth = 9, shrinkage = 0.4044957 and n.minobsinnode = 13
Extreme Gradient Boosting	Penicillin	AA k-mer/k-mer	nrounds = 150, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.4, min_child_weight = 2 and subsample = 0.75
Random Forest	Erythromycin	k-mer	mtry = 20
Support Vector Machine	Erythromycin	k-mer	sigma = 0.00904821 and C = 3.07926
Stochastic Gradient Boosting	Erythromycin	k-mer	n.trees = 250, interaction.depth = 4, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Erythromycin	k-mer	nrounds = 100, max_depth = 5, eta = 0.3, gamma = 0, colsample_bytree = 0.5, min_child_weight = 1
Random Forest	Erythromycin	AA k-mer	mtry = 12.80
Support Vector Machine	Erythromycin	AA k-mer	C = 0.6315789
Stochastic Gradient Boosting	Erythromycin	AA k-mer	n.trees = 500, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Erythromycin	AA k-mer	nrounds = 100, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.8, min_child_weight = 1# and subsample = 0.5.
Random Forest	Erythromycin	SNP	mtry = 12
Support Vector Machine	Erythromycin	SNP	C = 0.73
Stochastic Gradient Boosting	Erythromycin	SNP	n.trees = 500, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Erythromycin	SNP	nrounds = 150, max_depth = 3, eta = 0.4, gamma = 0, colsample_bytree = 0.675, min_child_weight = 1 subsample = 0.5
Random Forest	Erythromycin	SNP/AA k-mer	mtry = 30
Support Vector Machine	Erythromycin	SNP/AA k-mer	sigma = 0.0003544345 and C = 211.443
Stochastic Gradient Boosting	Erythromycin	SNP/AA k-mer	n.trees = 350, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10

Table A6: The best parameters selected for each antibiotic in the final models in *S.pneumoniae* (continue)

Extreme Gradient Boosting	Erythromycin	SNP/AA k-mer	nrounds = 200, max_depth = 2, eta = 0.3, gamma = 0, colsample_bytree = 0.4, min_child_weight = 2 and subsample = 1
Random Forest	Erythromycin	SNP/k-mer	mtry = 25
Support Vector Machine	Erythromycin	SNP/k-mer	sigma = 0.01002054 and C = 10.98855
Stochastic Gradient Boosting	Erythromycin	SNP/k-mer	n.trees = 150, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Erythromycin	SNP/k-mer	nrounds = 100, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.3, min_child_weight = 1and subsample = 1
Random Forest	Erythromycin	AA k-mer/k-mer	mtry = 45
Support Vector Machine	Erythromycin	AA k-mer/k-mer	degree = 2, scale = 0.009069009 and C = 1.241622
Stochastic Gradient Boosting	Erythromycin	AA k-mer/k-mer	n.trees = 3650, interaction.depth = 9, shrinkage = 0.01502375 and n.minobsinnode = 5
Extreme Gradient Boosting	Erythromycin	AA k-mer/k-mer	nrounds = 200, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.3, min_child_weight = 2 and subsample = 1
Random Forest	Tetracycline	k-mer	mtry = 7.61
Support Vector Machine	Tetracycline	k-mer	sigma = 0.06456923 and C = 0.08250606
Stochastic Gradient Boosting	Tetracycline	k-mer	n.trees = 150, interaction.depth = 2, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Tetracycline	k-mer	nrounds = 150, max_depth = 1, eta = 0.4, gamma = 0, colsample_bytree = 0.8, min_child_weight = 1
Random Forest	Tetracycline	AA k-mer	mtry = 11.40
Support Vector Machine	Tetracycline	AA k-mer	sigma = 0.1153326 and C = 28.17864
Stochastic Gradient Boosting	Tetracycline	AA k-mer	n.trees = 500, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Tetracycline	AA k-mer	nrounds = 50, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.4, min_child_weight = 3 and subsample = 1
Random Forest	Tetracycline	SNP	mtry = 14
Support Vector Machine	Tetracycline	SNP	sigma = 0.002651096 and C = 210.9548
Stochastic Gradient Boosting	Tetracycline	SNP	n.trees = 300, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10

Table A6: The best parameters selected for each antibiotic in the final models in *S.pneumoniae* (continue)

Extreme Gradient Boosting	Tetracycline	SNP	nrounds = 150, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.6, min_child_weight = 1 subsample = 0.75.
Random Forest	Tetracycline	SNP/AA k-mer	mtry = 47
Support Vector Machine	Tetracycline	SNP/AA k-mer	sigma = 0.0005349413 and C = 522.9208
Stochastic Gradient Boosting	Tetracycline	SNP/AA k-mer	n.trees = 696, interaction.depth = 2, shrinkage = 0.04072671 and n.minobsinnode = 21
Extreme Gradient Boosting	Tetracycline	SNP/AA k-mer	nrounds = 200, max_depth = 2, eta = 0.3, gamma = 0, colsample_bytree = 0.3, min_child_weight = 1 and subsample = 1
Random Forest	Tetracycline	SNP/k-mer	mtry = 31
Support Vector Machine	Tetracycline	SNP/k-mer	C = 0.1052632
Stochastic Gradient Boosting	Tetracycline	SNP/k-mer	n.trees = 100, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Tetracycline	SNP/k-mer	nrounds = 150, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.5, min_child_weight = 1 subsample = 0.5
Random Forest	Tetracycline	AA k-mer/k-mer	mtry = 13.71
Support Vector Machine	Tetracycline	AA k-mer/k-mer	sigma = 0.0008980129 and C = 336.0581
Stochastic Gradient Boosting	Tetracycline	AA k-mer/k-mer	n.trees = 150, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Tetracycline	AA k-mer/k-mer	nrounds = 150, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.4, min_child_weight = 2 subsample = 0.75

Table A7: The best parameters selected for each antibiotic in the final models in *S.aureus*

Algorithm	Antibiotic	Input	Parameters
Random Forest	Erythromycin	k-mer	mtry = 32
Support Vector Machine	Erythromycin	k-mer	degree = 3, scale = 2.441974e-05 and C = 635.1844
Stochastic Gradient Boosting	Erythromycin	k-mer	n.trees = 1422, interaction.depth = 9, shrinkage = 0.06716136 and n.minobsinnode = 7.
Extreme Gradient Boosting	Erythromycin	k-mer	nrounds = 100, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.625, min_child_weight = 1 and subsample = 1
Random Forest	Erythromycin	AA k-mer	12.8841
Support Vector Machine	Erythromycin	AA k-mer	C = 0.1052632
Stochastic Gradient Boosting	Erythromycin	AA k-mer	degree = 3, scale = 1.227223e-05 and C = 601.7368
Extreme Gradient Boosting	Erythromycin	AA k-mer	n.trees = 1450, interaction.depth = 4, shrinkage = 0.1 and n.minobsinnode = 10.
Random Forest	Erythromycin	SNP	4.690416
Support Vector Machine	Erythromycin	SNP	C = 0.1052632
Stochastic Gradient Boosting	Erythromycin	SNP	n.trees = 900, interaction.depth = 2, shrinkage = 0.1 and # n.minobsinnode = 10
Extreme Gradient Boosting	Erythromycin	SNP	nrounds = 50, max_depth = 2, eta = 0.4, gamma = 0, colsample_bytree # = 0.8, min_child_weight = 1 and subsample = 0.75
Random Forest	Erythromycin	SNP/AA k-mer	mtry = 22
Support Vector Machine	Erythromycin	SNP/AA k-mer	sigma = 0.0004912753 and C = 18.98775
Stochastic Gradient Boosting	Erythromycin	SNP/AA k-mer	n.trees = 100, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Erythromycin	SNP/AA k-mer	nrounds = 308, max_depth = 1, eta = 0.102727, gamma = 0.3009597, colsample_bytree = 0.6346631, min_child_weight = 2 and subsample = 0.6606891
Random Forest	Erythromycin	SNP/k-mer	mtry = 22
Support Vector Machine	Erythromycin	SNP/k-mer	degree = 3, scale = 0.0002247632 and C = 26.66623
Stochastic Gradient Boosting	Erythromycin	SNP/k-mer	n.trees = 150, interaction.depth = 1, shrinkage = 0.1 and n.minobsinnode = 10.
Extreme Gradient Boosting	Erythromycin	SNP/k-mer	nrounds = 100, max_depth = 5, eta = 0.3, gamma = 0, colsample_

Table A7: The best parameters selected for each antibiotic in the final models in *S.aureus* (continue)

Random Forest	Erythromycin	AA k-mer/k-mer	mtry = 5
Support Vector Machine	Erythromycin	AA k-mer/k-mer	sigma = 0.001590236 and C = 1.032432
Stochastic Gradient Boosting	Erythromycin	AA k-mer/k-mer	n.trees = 150, interaction.depth = 2, shrinkage = 0.1 and n.minobsinnode = 10.
Extreme Gradient Boosting	Erythromycin	AA k-mer/k-mer	nrounds = 100, max_depth = 1, eta = 0.3, gamma = 0, colsample_bytree = 0.875, min_child_weight = 1 and subsample = 1
Random Forest	Tetracycline	k-mer	mtry = 10
Support Vector Machine	Tetracycline	k-mer	degree = 3, scale = 2.441974e-05 and C = 635.1844
Stochastic Gradient Boosting	Tetracycline	k-mer	n.trees = 1422, interaction.depth = 9, shrinkage = 0.06716136 and n.minobsinnode = 7
Extreme Gradient Boosting	Tetracycline	k-mer	nrounds = 100, max_depth = 1, eta = 0.3, gamma = 0, colsample_bytree = 1, min_child_weight = 1 and subsample = 1
Random Forest	Tetracycline	AA k-mer	mtry = 168
Support Vector Machine	Tetracycline	AA k-mer	sigma = 0.004656746 and C = 16.55807
Stochastic Gradient Boosting	Tetracycline	AA k-mer	n.trees = 150, interaction.depth = 2, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Tetracycline	AA k-mer	nrounds = 50, max_depth = 1, eta = 0.4, gamma = 0, colsample_bytree = 0.8, min_child_weight = 1 and subsample = 1
Random Forest	Tetracycline	SNP	mtry = 10
Support Vector Machine	Tetracycline	SNP	sigma = 0.003462953 and C = 107.7598
Stochastic Gradient Boosting	Tetracycline	SNP	n.trees = 550, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Tetracycline	SNP	nrounds = 100, max_depth = 1, eta = 0.3, gamma = 0, colsample_bytree = # 0.625, min_child_weight = 1 and subsample = 1
Random Forest	Tetracycline	SNP/AA k-mer	mtry = 194
Support Vector Machine	Tetracycline	SNP/AA k-mer	C = 0.07063114
Extreme Gradient Boosting	Tetracycline	SNP/AA k-mer	nrounds = 100, max_depth = 5, eta = 0.3, gamma = 0, colsample_bytree = 0.75, min_child_weight = 1 and subsample = 1.

Table A7: The best parameters selected for each antibiotic in the final models in *S.aureus* (continue)

Random Forest	Tetracycline	SNP/k-mer	mtry = 5
Support Vector Machine	Tetracycline	SNP/k-mer	C = 0.1189676
Stochastic Gradient Boosting	Tetracycline	SNP/k-mer	n.trees = 100, interaction.depth = 2, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Tetracycline	SNP/k-mer	nrounds = 50, max_depth = 2, eta = 0.4, gamma = 0, colsample_bytree = 0.6, min_child_weight = 1 and subsample = 1
Random Forest	Tetracycline	AA k-mer/k-mer	mtry = 5
Support Vector Machine	Tetracycline	AA k-mer/k-mer	sigma = 0.000815399 and C = 0.2998059
Stochastic Gradient Boosting	Tetracycline	AA k-mer/k-mer	n.trees = 100, interaction.depth = 2, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Tetracycline	AA k-mer/k-mer	nrounds = 100, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.75, min_child_weight = 1 and subsample = 1
Random Forest	Gentamicin	k-mer	mtry = 5
Support Vector Machine	Gentamicin	k-mer	sigma = 0.01174385 and C = 0.153824
Stochastic Gradient Boosting	Gentamicin	k-mer	n.trees = 750, interaction.depth = 2, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Gentamicin	k-mer	nrounds = 100, max_depth = 1, eta = 0.3, gamma = 0, colsample_bytree = 0.5, min_child_weight = 1 and subsample = 1.
Random Forest	Gentamicin	AA k-mer	mtry = 5
Support Vector Machine	Gentamicin	AA k-mer	degree = 3, scale = 2.441974e-05 and C = 635.1844
Stochastic Gradient Boosting	Gentamicin	AA k-mer	n.trees = 50, interaction.depth = 2, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Gentamicin	AA k-mer	nrounds = 382, max_depth = 1, eta = 0.3509363, gamma = 2.022498, colsample_bytree = 0.6198912, min_child_weight = 2 and subsample = 0.4006813
Random Forest	Gentamicin	SNP	mtry = 8
Support Vector Machine	Gentamicin	SNP	sigma = 0.01043077 and C = 370.9932
Stochastic Gradient Boosting	Gentamicin	SNP	n.trees = 3721, interaction.depth = 7, shrinkage = 0.4781242

Table A7: The best parameters selected for each antibiotic in the final models in *S.aureus* (continue)

Extreme Gradient Boosting	Gentamicin	SNP	nrounds = 100, max_depth = 5, eta = 0.3, gamma = 0, colsample_bytree = # 0.625, min_child_weight = 1 and subsample = 1
Random Forest	Gentamicin	SNP/AA k-mer	mtry = 5
Extreme Gradient Boosting	Gentamicin	SNP/AA k-mer	nrounds = 100, max_depth = 5, eta = 0.3, gamma = 0, colsample_bytree = 0.5, min_child_weight = 1 and subsample = 1
Random Forest	Gentamicin	SNP/k-mer	mtry = 5
Support Vector Machine	Gentamicin	SNP/k-mer	C = 0.08017363
Stochastic Gradient Boosting	Gentamicin	SNP/k-mer	n.trees = 4361, interaction.depth = 9, shrinkage = 0.2147973 and n.minobsinnode = 15
Extreme Gradient Boosting	Gentamicin	SNP/k-mer	nrounds = 226, max_depth = 9, eta = 0.3112471, gamma = 9.275322, colsample_bytree = 0.4206683, min_child_weight = 2 and subsample = 0.4767639
Random Forest	Gentamicin	AA k-mer/k-mer	mtry = 13
Support Vector Machine	Gentamicin	AA k-mer/k-mer	degree = 3, scale = 0.001765149 and C = 513.4441
Stochastic Gradient Boosting	Gentamicin	AA k-mer/k-mer	n.trees = 3134, interaction.depth = 2, shrinkage = 0.2299327 and n.minobsinnode = 12
Extreme Gradient Boosting	Gentamicin	AA k-mer/k-mer	nrounds = 50, max_depth = 1, eta = 0.3, gamma = 0, colsample_bytree = 0.6, min_child_weight = 1 and subsample = 0.5
Random Forest	Ciprofloxacin	k-mer	mtry = 2
Support Vector Machine	Ciprofloxacin	k-mer	C = 0.1052632
Stochastic Gradient Boosting	Ciprofloxacin	k-mer	n.trees = 1500, interaction.depth = 3, shrinkage = 0.1 # and n.minobsinnode = 10
Extreme Gradient Boosting	Ciprofloxacin	k-mer	nrounds = 100, max_depth = 3, eta = 0.4, gamma = 0, colsample_bytree = 0.6, min_child_weight = 1 and subsample = 0.75.
Random Forest	Ciprofloxacin	AA k-mer	mtry = 12
Support Vector Machine	Ciprofloxacin	AA k-mer	sigma = 0.00218975 and C = 33.99218
Stochastic Gradient Boosting	Ciprofloxacin	AA k-mer	n.trees = 200, interaction.depth = 2, shrinkage = 0.1 and n.minobsinnode = 10

Table A7: The best parameters selected for each antibiotic in the final models in *S.aureus* (continue)

Extreme Gradient Boosting	Ciprofloxacin	AA k-mer	nrounds = 100, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.8, min_child_weight = 1 and subsample = 0.5
Random Forest	Ciprofloxacin	SNP	mtry = 13
Support Vector Machine	Ciprofloxacin	SNP	sigma = 0.08006517 and C = 16.55807
Stochastic Gradient Boosting	Ciprofloxacin	SNP	n.trees = 783, interaction.depth = 2, shrinkage = 0.1694268 and n.minobsinnode = 5
Extreme Gradient Boosting	Ciprofloxacin	SNP	nrounds = 100, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.875, min_child_weight = 1 and subsample = 1
Random Forest	Ciprofloxacin	SNP/AA k-mer	mtry = 44
Support Vector Machine	Ciprofloxacin	SNP/AA k-mer	sigma = 0.001735752 and C = 80.06231
Stochastic Gradient Boosting	Ciprofloxacin	SNP/AA k-mer	n.trees = 150, interaction.depth = 1, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Ciprofloxacin	SNP/AA k-mer	nrounds = 50, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.6, min_child_weight = 1 and subsample = 0.75
Random Forest	Ciprofloxacin	SNP/k-mer	mtry = 6
Support Vector Machine	Ciprofloxacin	SNP/k-mer	C = 0.5586492
Stochastic Gradient Boosting	Ciprofloxacin	SNP/k-mer	n.trees = 450, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Ciprofloxacin	SNP/k-mer	nrounds = 100, max_depth = 5, eta = 0.3, gamma = 0, colsample_bytree = 0.625, min_child_weight = 1 and subsample = 1
Random Forest	Ciprofloxacin	AA k-mer/k-mer	mtry = 6
Support Vector Machine	Ciprofloxacin	AA k-mer/k-mer	sigma = 0.002450468 and C = 0.6039416
Stochastic Gradient Boosting	Ciprofloxacin	AA k-mer/k-mer	n.trees = 463, interaction.depth = 2, shrinkage = 0.2616397 and n.minobsinnode = 15
Extreme Gradient Boosting	Ciprofloxacin	AA k-mer/k-mer	nrounds = 100, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.625, min_child_weight = 1 and subsample = 1

Table A8: The best parameters selected for each antibiotic in the final models in *E. faecium*

Algorithm	Antibiotic	Input	Parameters
Random Forest	Ampicillin	k-mer	mtry = 5
Support Vector Machine	Ampicillin	k-mer	degree = 3, scale = 1.073719 and C = 0.08843755
Stochastic Gradient Boosting	Ampicillin	k-mer	n.trees = 350, interaction.depth = 4, shrinkage = 0.1 # and n.minobsinnode = 10
Extreme Gradient Boosting	Ampicillin	k-mer	nrounds = 100, max_depth = 1, eta = 0.3, gamma = 0, colsample_bytree = # 0.5, min_child_weight = 1 and subsample = 1
Random Forest	Ampicillin	AA k-mer	mtry = 15
Support Vector Machine	Ampicillin	AA k-mer	sigma = 0.008839918 and C = 305.7902
Stochastic Gradient Boosting	Ampicillin	AA k-mer	n.trees = 1450, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Ampicillin	AA k-mer	nrounds = 100, max_depth = 1, eta = 0.3, gamma = 0, colsample_bytree = 0.625, min_child_weight = 1 and subsample = 1
Random Forest	Ampicillin	SNP	mtry = 5
Support Vector Machine	Ampicillin	SNP	sigma = 0.02323982 and C = 0.1361988
Stochastic Gradient Boosting	Ampicillin	SNP	n.trees = 783, interaction.depth = 2, shrinkage = 0.1694268 and n.minobsinnode = 5.
Extreme Gradient Boosting	Ampicillin	SNP	nrounds = 50, max_depth = 2, eta = 0.4, gamma = 0, colsample_bytree = 0.6, min_child_weight = 1 and subsample = 0.5
Random Forest	Ampicillin	SNP/AA k-mer	mtry = 5
Support Vector Machine	Ampicillin	SNP/AA k-mer	sigma = 0.0484458 and C = 957.5366
Stochastic Gradient Boosting	Ampicillin	SNP/AA k-mer	n.trees = 150, interaction.depth = 4, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Ampicillin	SNP/AA k-mer	nrounds = 50, max_depth = 2, eta = 0.4, gamma = 0, colsample_bytree = 0.6, min_child_weight = 1 and subsample = 0.75
Random Forest	Ampicillin	SNP/k-mer	mtry = 5
Support Vector Machine	Ampicillin	SNP/k-mer	sigma = 0.02904212 and C = 0.6248541
Stochastic Gradient Boosting	Ampicillin	SNP/k-mer	n.trees = 200, interaction.depth = 4, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Ampicillin	SNP/k-mer	nrounds = 100, max_depth = 2, eta = 0.3, gamma = 0, colsample_bytree = 0.8

Table A8: The best parameters selected for each antibiotic in the final models in *E. faecium* (continue)

Random Forest	Ampicillin	AA k-mer/k-mer	mtry = 35
Support Vector Machine	Ampicillin	AA k-mer/k-mer	sigma = 0.005578232 and C = 0.1921897
Stochastic Gradient Boosting	Ampicillin	AA k-mer/k-mer	n.trees = 550, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Ampicillin	AA k-mer/k-mer	nrounds = 150, max_depth = 1, eta = 0.3, gamma = 0, colsample_bytree = 0.8, min_child_weight = 1 and subsample = 0.5
Random Forest	Vancomycin	k-mer	mtry = 2
Support Vector Machine	Vancomycin	k-mer	C = 0.2544103
Stochastic Gradient Boosting	Vancomycin	k-mer	n.trees = 350, interaction.depth = 2, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Vancomycin	k-mer	nrounds = 100, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.625, min_child_weight = 1 and subsample = 1
Random Forest	Vancomycin	AA k-mer	mtry = 8
Support Vector Machine	Vancomycin	AA k-mer	C = 0.1052632
Stochastic Gradient Boosting	Vancomycin	AA k-mer	n.trees = 1150, interaction.depth = 2, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Vancomycin	AA k-mer	nrounds = 100, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.6, min_child_weight = 1 and subsample = 0.75
Random Forest	Vancomycin	SNP	mtry = 2
Support Vector Machine	Vancomycin	SNP	sigma = 0.01909461 and C = 26.803
Stochastic Gradient Boosting	Vancomycin	SNP	n.trees = 4073, interaction.depth = 4, shrinkage = 0.08245028 and n.minobsinnode = 17
Extreme Gradient Boosting	Vancomycin	SNP	nrounds = 150, max_depth = 3, eta = 0.4, gamma = 0, colsample_bytree = 0.8, min_child_weight = 1 and subsample = 0.5
Random Forest	Vancomycin	SNP/AA k-mer	mtry = 24
Support Vector Machine	Vancomycin	SNP/AA k-mer	sigma = 0.2245857 and C = 20.43015
Stochastic Gradient Boosting	Vancomycin	SNP/AA k-mer	n.trees = 100, interaction.depth = 1, shrinkage = 0.1 and n.minobsinnode = 10

Table A8: The best parameters selected for each antibiotic in the final models in *E. faecium* (continue)

Extreme Gradient Boosting	Vancomycin	SNP/AA k-mer	nrounds = 50, max_depth = 1, eta = 0.4, gamma = 0, colsample_bytree = 0.6, min_child_weight = 1 and subsample = 0.75
Random Forest	Vancomycin	SNP/k-mer	mtry = 17
Support Vector Machine	Vancomycin	SNP/k-mer	sigma = 0.0132135 and C = 160.0053
Stochastic Gradient Boosting	Vancomycin	SNP/k-mer	n.trees = 450, interaction.depth = 2, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Vancomycin	SNP/k-mer	nrounds = 50, max_depth = 3, eta = 0.4, gamma = 0, colsample_bytree = 0.6, min_child_weight = 1 and subsample = 0.75
Random Forest	Vancomycin	AA k-mer/k-mer	mtry = 11
Support Vector Machine	Vancomycin	AA k-mer/k-mer	sigma = 0.1796718 and C = 219.1946
Stochastic Gradient Boosting	Vancomycin	AA k-mer/k-mer	n.trees = 1300, interaction.depth = 2, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting	Vancomycin	AA k-mer/k-mer	nrounds = 100, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.5, min_child_weight = 1 and subsample = 1

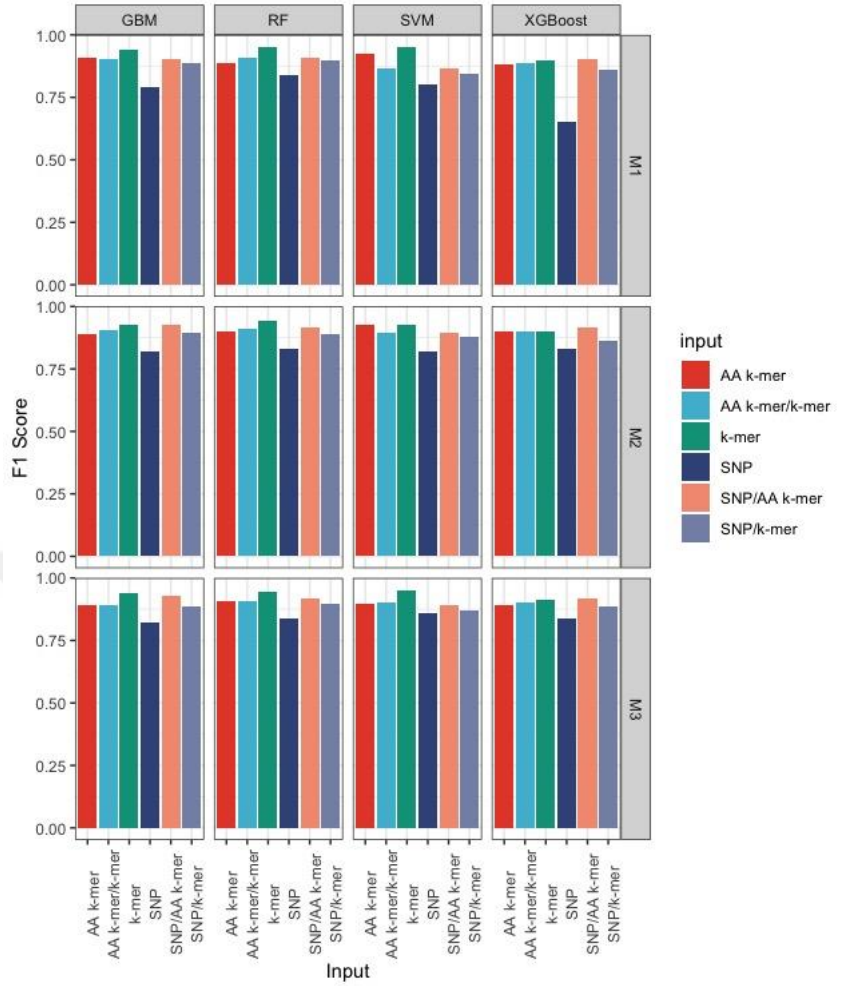

Figure A1. Comparison of F1 Scores of all penicillin S. pneumoniae models.

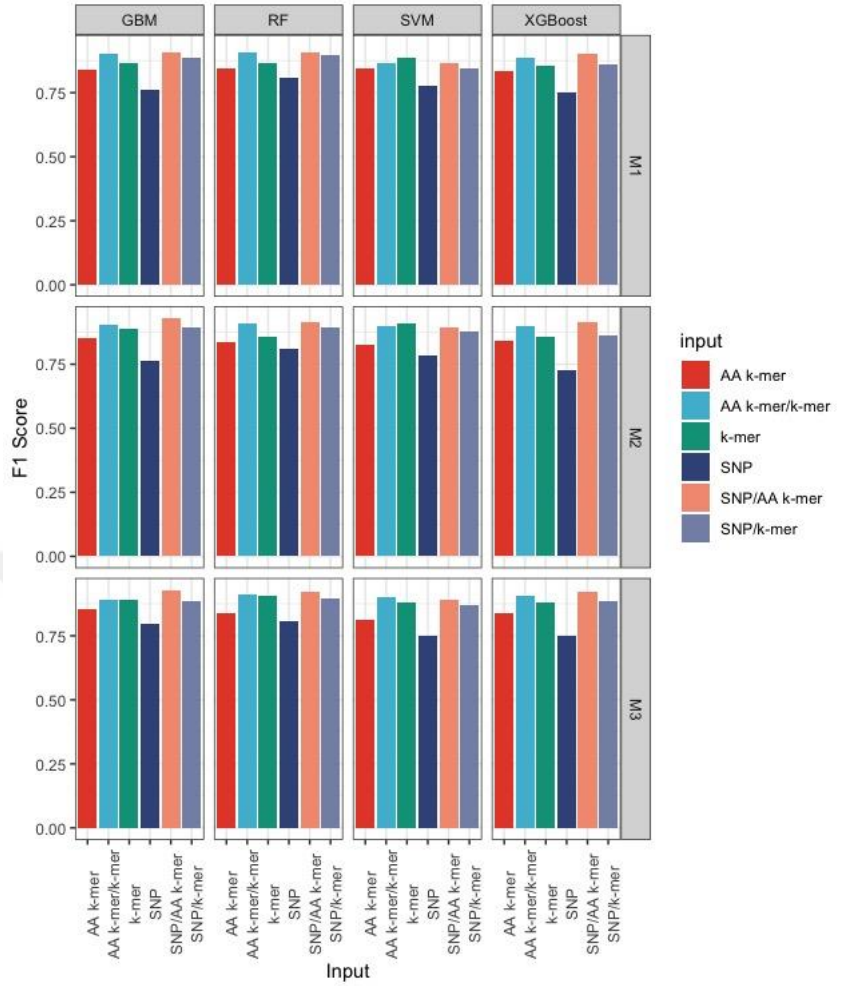

Figure A2. Comparison of F1 Scores of all erythromycin S.pneumoniae models.

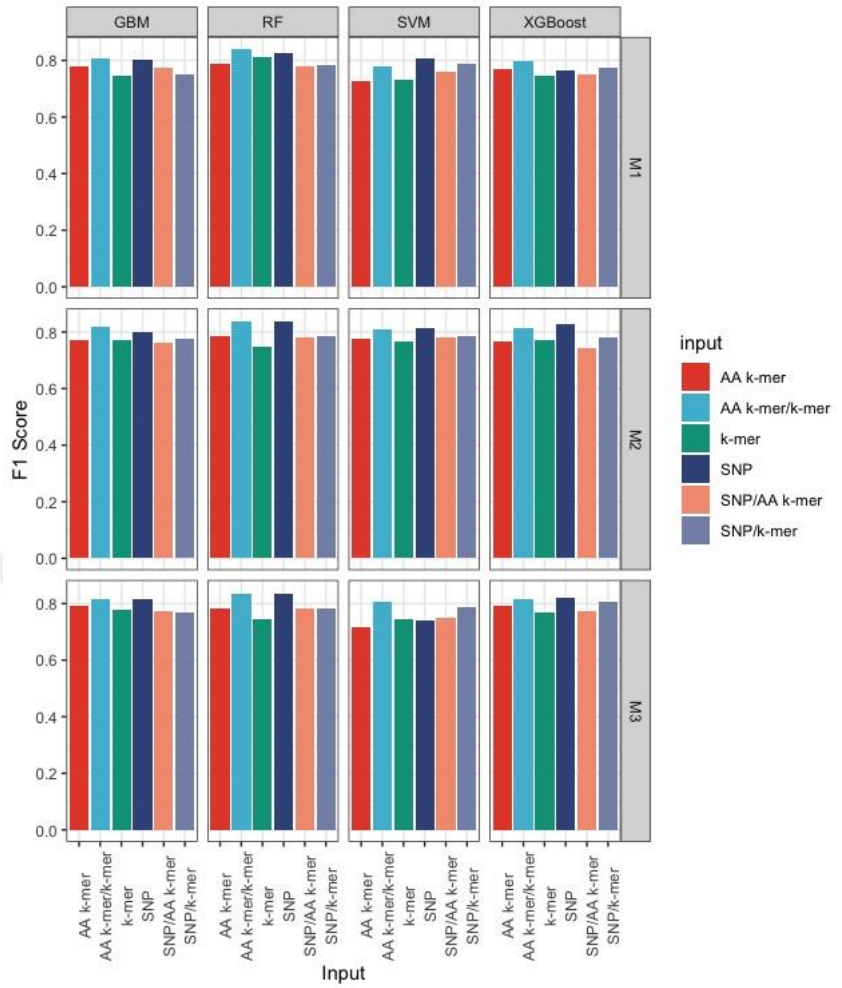

Figure A3. Comparison of F1 Scores of all tetracycline S.pneumoniae models.

Appendix 3

Table A9: Penicillin Results of S.pneumoniae

Algorithm	Model#	input	F1-score	Kappa	Accuracy
Random Forest	M1	k-mer	0.9524	0.9423	0.9833
Random Forest	M2	k-mer	0.944	0.9322	0.9806
Random Forest	M3	k-mer	0.944	0.9322	0.9806
Support Vector Machine	M1	k-mer	0.9516	0.9415	0.9833
Support Vector Machine	M2	k-mer	0.9268	0.9118	0.975
Support Vector Machine	M3	k-mer	0.9516	0.9415	0.9833
Stochastic Gradient Boosting	M1	k-mer	0.9431	0.9314	0.9806
Stochastic Gradient Boosting	M2	k-mer	0.9268	0.9314	0.9806
Stochastic Gradient Boosting	M3	k-mer	0.9375	0.924	0.9778
Extreme Gradient Boosting	M1	k-mer	0.8976	0.8757	0.9639
Extreme Gradient Boosting	M2	k-mer	0.8992	0.8772	0.9639
Extreme Gradient Boosting	M3	k-mer	0.9147	0.8961	0.9694
Random Forest	M1	AA k-mer	0.8889	0.8675	0.9639
Random Forest	M2	AA k-mer	0.8983	0.8785	0.9667
Random Forest	M3	AA k-mer	0.9076	0.8893	0.9694
Support Vector Machine	M1	AA k-mer	0.9256	0.9106	0.975
Support Vector Machine	M2	AA k-mer	0.9256	0.9106	0.975
Support Vector Machine	M3	AA k-mer	0.8983	0.8785	0.9667
Stochastic Gradient Boosting	M1	AA k-mer	0.9076	0.8893	0.9694
Stochastic Gradient Boosting	M2	AA k-mer	0.8889	0.8675	0.9639
Stochastic Gradient Boosting	M3	AA k-mer	0.8889	0.8544	0.9611
Extreme Gradient Boosting	M1	AA k-mer	0.8793	0.8563	0.9611
Extreme Gradient Boosting	M2	AA k-mer	0.8983	0.8785	0.9667
Extreme Gradient Boosting	M3	AA k-mer	0.8889	0.8675	0.9639
Random Forest	M1	SNP	0.8364	0.8076	0.95
Random Forest	M2	SNP	0.8319	0.801	0.9472
Random Forest	M3	SNP	0.8393	0.8102	0.95
Support Vector Machine	M1	SNP	0.8	0.7648	0.9389
Support Vector Machine	M2	SNP	0.8182	0.7862	0.9444

Table A9: Penicillin Results of S.pneumoniae (continue)

Support Vector Machine	M3	SNP	0.8596	0.8336	0.9556
Stochastic Gradient Boosting	M1	SNP	0.7931	0.7537	0.9333
Stochastic Gradient Boosting	M2	SNP	0.8214	0.7891	0.9444
Stochastic Gradient Boosting	M3	SNP	0.8214	0.7891	0.9444
Extreme Gradient Boosting	M1	SNP	0.6526	0.6062	0.9083
Extreme Gradient Boosting	M2	SNP	0.8319	0.801	0.9472
Extreme Gradient Boosting	M3	SNP	0.8393	0.8102	0.95
Random Forest	M1	SNP/AA k-mer	0.9058296	0.8663	0.9443
Random Forest	M2	SNP/AA k-mer	0.9155556	0.8796	0.9496
Random Forest	M3	SNP/AA k-mer	0.9196429	0.8857	0.9523
Support Vector Machine	M1	SNP/AA k-mer	0.8675799	0.8135	0.9231
Support Vector Machine	M2	SNP/AA k-mer	0.8949772	0.852	0.939
Support Vector Machine	M3	SNP/AA k-mer	0.8930233	0.8505	0.939
Stochastic Gradient Boosting	M1	SNP/AA k-mer	0.9049774	0.8656	0.9443
Stochastic Gradient Boosting	M2	SNP/AA k-mer	0.9285714	0.8984	0.9576
Stochastic Gradient Boosting	M3	SNP/AA k-mer	0.9292035	0.8989	0.9576
Extreme Gradient Boosting	M1	SNP/AA k-mer	0.9026549	0.861	0.9416
Extreme Gradient Boosting	M2	SNP/AA k-mer	0.9155556	0.8796	0.9496
Extreme Gradient Boosting	M3	SNP/AA k-mer	0.9196	0.8857	0.9523
Random Forest	M1	SNP/k-mer	0.8956522	0.8499	0.9363
Random Forest	M2	SNP/k-mer	0.8917749	0.844	0.9337
Random Forest	M3	SNP/k-mer	0.8965517	0.8506	0.9363
Support Vector Machine	M1	SNP/k-mer	0.8448276	0.7759	0.9045
Support Vector Machine	M2	SNP/k-mer	0.8776371	0.8217	0.9231
Support Vector Machine	M3	SNP/k-mer	0.8717949	0.8142	0.9204
Stochastic Gradient Boosting	M1	SNP/k-mer	0.887931	0.8382	0.931
Stochastic Gradient Boosting	M2	SNP/k-mer	0.8956522	0.8499	0.9363

Table A9: Penicillin Results of S.pneumoniae (continue)

Stochastic Gradient Boosting	M3	SNP/k-mer	0.8841202	0.8324	0.9284
Extreme Gradient Boosting	M1	SNP/k-mer	0.8609865	0.8026	0.9178
Extreme Gradient Boosting	M2	SNP/k-mer	0.8622222	0.8036	0.9178
Extreme Gradient Boosting	M3	SNP/k-mer	0.887	0.8374	0.931
Random Forest	M1	AA k-mer/k-mer	0.9090909	0.869	0.9443
Random Forest	M2	AA k-mer/k-mer	0.9090909	0.869	0.9443
Random Forest	M3	AA k-mer/k-mer	0.9090909	0.869	0.9443
Support Vector Machine	M1	AA k-mer/k-mer	0.8669528	0.8075	0.9178
Support Vector Machine	M2	AA k-mer/k-mer	0.8965517	0.8506	0.9363
Support Vector Machine	M3	AA k-mer/k-mer	0.8995633	0.8558	0.939
Stochastic Gradient Boosting	M1	AA k-mer/k-mer	0.9043478	0.8624	0.9416
Stochastic Gradient Boosting	M2	AA k-mer/k-mer	0.9043478	0.8624	0.9416
Stochastic Gradient Boosting	M3	AA k-mer/k-mer	0.8898678	0.8424	0.9337
Extreme Gradient Boosting	M1	AA k-mer/k-mer	0.8888889	0.8416	0.9337
Extreme Gradient Boosting	M2	AA k-mer/k-mer	0.8995633	0.8558	0.939
Extreme Gradient Boosting	M3	AA k-mer/k-mer	0.9043	0.8624	0.9416

Table A10: Erythromycin Results of *S.pneumoniae*

Algorithm	Model#	input	F1-score	Kappa	Accuracy
Random Forest	M1	k-mer	0.8636364	0.8075	0.9204
Random Forest	M2	k-mer	0.8584475	0.8006	0.9178
Random Forest	M3	k-mer	0.9066667	0.867	0.9443
Support Vector Machine	M1	k-mer	0.8888889	0.839	0.931
Support Vector Machine	M2	k-mer	0.9082969	0.8683	0.9443
Support Vector Machine	M3	k-mer	0.8778281	0.8272	0.9284
Stochastic Gradient Boosting	M1	k-mer	0.8672566	0.8104	0.9204
Stochastic Gradient Boosting	M2	k-mer	0.8898678	0.8416	0.9337

Table A10: Erythromycin Results of *S.pneumoniae* (continue)

Stochastic Gradient Boosting	M3	k-mer	0.8898678	0.8424	0.9337
Extreme Gradient Boosting	M1	k-mer	0.855814	0.7985	0.9178
Extreme Gradient Boosting	M2	k-mer	0.8584071	0.7978	0.9151
Extreme Gradient Boosting	M3	k-mer	0.88	0.829	0.9284
Random Forest	M1	AA k-mer	0.8425532	0.7714	0.9019
Random Forest	M2	AA k-mer	0.8376068	0.7647	0.8992
Random Forest	M3	AA k-mer	0.8389831	0.7658	0.8992
Support Vector Machine	M1	AA k-mer	0.8448276	0.7759	0.9045
Support Vector Machine	M2	AA k-mer	0.8255319	0.7467	0.8912
Support Vector Machine	M3	AA k-mer	0.8103448	0.7261	0.8833
Stochastic Gradient Boosting	M1	AA k-mer	0.8370044	0.7668	0.9019
Stochastic Gradient Boosting	M2	AA k-mer	0.8497854	0.7827	0.9072
Stochastic Gradient Boosting	M3	AA k-mer	0.8547009	0.7894	0.9098
Extreme Gradient Boosting	M1	AA k-mer	0.8347826	0.7623	0.8992
Extreme Gradient Boosting	M2	AA k-mer	0.8434783	0.7748	0.9045
Extreme Gradient Boosting	M3	AA k-mer	0.8376068	0.76	0.8992
Random Forest	M1	SNP	0.8056	0.7277	0.8886
Random Forest	M2	SNP	0.8111	0.7349	0.8912
Random Forest	M3	SNP	0.8056	0.7277	0.8886
Support Vector Machine	M1	SNP	0.7793	0.6929	0.8753
Support Vector Machine	M2	SNP	0.7826	0.7013	0.8806
Support Vector Machine	M3	SNP	0.75	0.6557	0.8621
Stochastic Gradient Boosting	M1	SNP	0.7623	0.6625	0.8594
Stochastic Gradient Boosting	M2	SNP	0.7624	0.6771	0.8727
Stochastic Gradient Boosting	M3	SNP	0.7964	0.712	0.8806
Extreme Gradient Boosting	M1	SNP	0.7522	0.6462	0.8515

Table A10: Erythromycin Results of *S.pneumoniae* (continue)

Extreme Gradient Boosting	M2	SNP	0.7248	0.613	0.8408
Extreme Gradient Boosting	M3	SNP	0.75	0.6444	0.8937
Random Forest	M1	SNP/AA k-mer	0.9058296	0.8663	0.9443
Random Forest	M2	SNP/AA k-mer	0.9155556	0.8796	0.9496
Random Forest	M3	SNP/AA k-mer	0.9196429	0.8857	0.9523
Support Vector Machine	M1	SNP/AA k-mer	0.8675799	0.8135	0.9231
Support Vector Machine	M2	SNP/AA k-mer	0.8949772	0.852	0.939
Support Vector Machine	M3	SNP/AA k-mer	0.8930233	0.8505	0.939
Stochastic Gradient Boosting	M1	SNP/AA k-mer	0.9049774	0.8656	0.9443
Stochastic Gradient Boosting	M2	SNP/AA k-mer	0.9285714	0.8984	0.9576
Stochastic Gradient Boosting	M3	SNP/AA k-mer	0.9292035	0.8989	0.9576
Extreme Gradient Boosting	M1	SNP/AA k-mer	0.9026549	0.861	0.9416
Extreme Gradient Boosting	M2	SNP/AA k-mer	0.9155556	0.8796	0.9496
Extreme Gradient Boosting	M3	SNP/AA k-mer	0.9196429	0.8857	0.9523
Random Forest	M1	SNP/k-mer	0.8956522	0.8499	0.9363
Random Forest	M2	SNP/k-mer	0.8917749	0.844	0.9337
Random Forest	M3	SNP/k-mer	0.8965517	0.8506	0.9363
Support Vector Machine	M1	SNP/k-mer	0.8448276	0.7759	0.9045
Support Vector Machine	M2	SNP/k-mer	0.8776371	0.8217	0.9231
Support Vector Machine	M3	SNP/k-mer	0.8717949	0.8142	0.9204
Stochastic Gradient Boosting	M1	SNP/k-mer	0.887931	0.8382	0.931
Stochastic Gradient Boosting	M2	SNP/k-mer	0.8956522	0.8499	0.9363
Stochastic Gradient Boosting	M3	SNP/k-mer	0.8841202	0.8324	0.9284
Extreme Gradient Boosting	M1	SNP/k-mer	0.8609865	0.8026	0.9178

Table A10: Erythromycin Results of *S.pneumoniae* (continue)

Extreme Gradient Boosting	M2	SNP/k-mer	0.8622222	0.8036	0.9178
Extreme Gradient Boosting	M3	SNP/k-mer	0.8869565	0.8374	0.931
Random Forest	M1	AA k-mer/k-mer	0.9090909	0.869	0.9443
Random Forest	M2	AA k-mer/k-mer	0.9090909	0.869	0.9443
Random Forest	M3	AA k-mer/k-mer	0.9090909	0.869	0.9443
Support Vector Machine	M1	AA k-mer/k-mer	0.8669528	0.8075	0.9178
Support Vector Machine	M2	AA k-mer/k-mer	0.8965517	0.8506	0.9363
Support Vector Machine	M3	AA k-mer/k-mer	0.8995633	0.8558	0.939
Stochastic Gradient Boosting	M1	AA k-mer/k-mer	0.9043478	0.8624	0.9416
Stochastic Gradient Boosting	M2	AA k-mer/k-mer	0.9043478	0.8624	0.9416
Stochastic Gradient Boosting	M3	AA k-mer/k-mer	0.8898678	0.8424	0.9337
Extreme Gradient Boosting	M1	AA k-mer/k-mer	0.8888889	0.8416	0.9337
Extreme Gradient Boosting	M2	AA k-mer/k-mer	0.8995633	0.8558	0.939
Extreme Gradient Boosting	M3	AA k-mer/k-mer	0.9043478	0.8624	0.9416

Table A11: Tetracycline Results of *S.pneumoniae*

Algorithm	Model#	input	F1-score	Kappa	Accuracy
Random Forest	M1	k-mer	0.8131868	0.6607	0.8317
Random Forest	M2	k-mer	0.7489712	0.5895	0.7987
Random Forest	M3	k-mer	0.746888	0.5892	0.7987
Support Vector Machine	M1	k-mer	0.7315175	0.5381	0.7723
Support Vector Machine	M2	k-mer	0.7659574	0.5623	0.7822
Support Vector Machine	M3	k-mer	0.7447699	0.5889	0.7987
Stochastic Gradient Boosting	M1	k-mer	0.745098	0.6487	0.8251
Stochastic Gradient Boosting	M2	k-mer	0.7715356	0.559	0.7822
Stochastic Gradient Boosting	M3	k-mer	0.780303	0.6128	0.8086
Extreme Gradient Boosting	M1	k-mer	0.745098	0.5646	0.7855
Extreme Gradient Boosting	M2	k-mer	0.7715356	0.5932	0.7987

Table A11: Tetracycline Results of *S.pneumoniae* (continue)

Extreme Gradient Boosting	M3	k-mer	0.77	0.5861	0.7954
Random Forest	M1	AA k-mer	0.7900356	0.6086	0.8053
Random Forest	M2	AA k-mer	0.7844523	0.5956	0.7987
Random Forest	M3	AA k-mer	0.7844523	0.5956	0.7987
Support Vector Machine	M1	AA k-mer	0.7279693	0.5255	0.7657
Support Vector Machine	M2	AA k-mer	0.7777778	0.5765	0.7888
Support Vector Machine	M3	AA k-mer	0.7175573	0.5056	0.7558
Stochastic Gradient Boosting	M1	AA k-mer	0.7793103	0.5769	0.7888
Stochastic Gradient Boosting	M2	AA k-mer	0.7697842	0.575	0.7888
Stochastic Gradient Boosting	M3	AA k-mer	0.7943262	0.6154	0.8086
Extreme Gradient Boosting	M1	AA k-mer	0.7692308	0.5808	0.7921
Extreme Gradient Boosting	M2	AA k-mer	0.765343	0.5682	0.7855
Extreme Gradient Boosting	M3	AA k-mer	0.7929825	0.6092	0.8053
Random Forest	M1	SNP	0.8239437	0.6687	0.835
Random Forest	M2	SNP	0.8395904	0.6896	0.8449
Random Forest	M3	SNP	0.8356164	0.6829	0.8416
Support Vector Machine	M1	SNP	0.8083624	0.636	0.8185
Support Vector Machine	M2	SNP	0.8137931	0.643	0.8218
Support Vector Machine	M3	SNP	0.7401575	0.5577	0.7822
Stochastic Gradient Boosting	M1	SNP	0.8041958	0.6292	0.8152
Stochastic Gradient Boosting	M2	SNP	0.8014184	0.6286	0.8152
Stochastic Gradient Boosting	M3	SNP	0.8178694	0.6497	0.8251
Extreme Gradient Boosting	M1	SNP	0.7670251	0.5685	0.7855
Extreme Gradient Boosting	M2	SNP	0.8287671	0.6697	0.835
Extreme Gradient Boosting	M3	SNP	0.82	0.6432	0.8218

Table A11: Tetracycline Results of *S.pneumoniae* (continue)

Random Forest	M1	SNP/AA k-mer	0.7773852	0.5811	0.7914
Random Forest	M2	SNP/AA k-mer	0.7816901	0.5879	0.7947
Random Forest	M3	SNP/AA k-mer	0.7816901	0.5879	0.7947
Support Vector Machine	M1	SNP/AA k-mer	0.7609428	0.5305	0.7649
Support Vector Machine	M2	SNP/AA k-mer	0.779661	0.5699	0.7848
Support Vector Machine	M3	SNP/AA k-mer	0.75	0.5105	0.755
Stochastic Gradient Boosting	M1	SNP/AA k-mer	0.7746479	0.5746	0.7881
Stochastic Gradient Boosting	M2	SNP/AA k-mer	0.7615658	0.5542	0.7781
Stochastic Gradient Boosting	M3	SNP/AA k-mer	0.7730496	0.5743	0.7881
Extreme Gradient Boosting	M1	SNP/AA k-mer	0.7508772	0.5283	0.7649
Extreme Gradient Boosting	M2	SNP/AA k-mer	0.7446809	0.5211	0.7616
Extreme Gradient Boosting	M3	SNP/AA k-mer	0.7724138	0.5624	0.7815
Random Forest	M1	SNP/k-mer	0.7819549	0.6118	0.8079
Random Forest	M2	SNP/k-mer	0.7867647	0.6127	0.8079
Random Forest	M3	SNP/k-mer	0.7810219	0.5996	0.8013
Support Vector Machine	M1	SNP/k-mer	0.7896679	0.6192	0.8113
Support Vector Machine	M2	SNP/k-mer	0.7841727	0.6003	0.8013
Support Vector Machine	M3	SNP/k-mer	0.7862595	0.6245	0.8146
Stochastic Gradient Boosting	M1	SNP/k-mer	0.75	0.534	0.7682
Stochastic Gradient Boosting	M2	SNP/k-mer	0.7756654	0.6046	0.8046
Stochastic Gradient Boosting	M3	SNP/k-mer	0.7692308	0.5795	0.7914
Extreme Gradient Boosting	M1	SNP/k-mer	0.7720588	0.586	0.7947
Extreme Gradient Boosting	M2	SNP/k-mer	0.7810219	0.5996	0.8013
Extreme Gradient Boosting	M3	SNP/k-mer	0.8042705	0.634	0.8179

Table A11: Tetracycline Results of *S.pneumoniae* (continue)

Random Forest	M1	AA k-mer/k-mer	0.8413793	0.6959	0.8482
Random Forest	M2	AA k-mer/k-mer	0.8373702	0.6891	0.8449
Random Forest	M3	AA k-mer/k-mer	0.8333333	0.6824	0.8416
Support Vector Machine	M1	AA k-mer/k-mer	0.7789474	0.5827	0.7921
Support Vector Machine	M2	AA k-mer/k-mer	0.8083624	0.636	0.8185
Support Vector Machine	M3	AA k-mer/k-mer	0.8071429	0.6416	0.8218
Stochastic Gradient Boosting	M1	AA k-mer/k-mer	0.8054608	0.6236	0.8119
Stochastic Gradient Boosting	M2	AA k-mer/k-mer	0.8194444	0.6559	0.8284
Stochastic Gradient Boosting	M3	AA k-mer/k-mer	0.8163265	0.6435	0.8218
Extreme Gradient Boosting	M1	AA k-mer/k-mer	0.7959866	0.598	0.7987
Extreme Gradient Boosting	M2	AA k-mer/k-mer	0.8122867	0.6368	0.8185
Extreme Gradient Boosting	M3	AA k-mer/k-mer	0.8179	0.6497	0.8251

Appendix 4

Table A12: Erythromycin Results of *S.aureus*

Algorithm	Model#	input	Accuracy	Kappa	F1-score
Random Forest	M1	k-mer	0.9891	0.9752	0.9744
Random Forest	M2	k-mer	0.9891	0.9759	0.9744
Random Forest	M3	k-mer	0.9889	0.9757	0.9744
Support Vector Machine	M1	k-mer	0.9832	0.9645	0.9744
Support Vector Machine	M1-cost	k-mer	0.9749	0.9468	0.9744
Support Vector Machine	M2	k-mer	0.9923	0.9831	0.9744
Support Vector Machine	M3	k-mer	0.9923	0.9831	0.9744
Stochastic Gradient Boosting	M1	k-mer	0.9763	0.9508	0.9744
Stochastic Gradient Boosting	M3	k-mer	0.9749	0.9463	0.9744
Extreme Gradient Boosting	M1	k-mer	0.9587	0.9066	0.9500
Extreme Gradient Boosting	M2	k-mer	0.9708	0.9330	0.9744

Table A12: Erythromycin Results of *S.aureus* (continue)

Extreme Gradient Boosting	M3	k-mer	0.9708	0.9330	0.9744
Random Forest	M1	AA k-mer	0.9342	0.8499	0.9412
Random Forest	M2	AA k-mer	0.9784	0.9526	0.9412
Random Forest	M3	AA k-mer	0.9408	0.8706	0.9444
Support Vector Machine	M1	AA k-mer	0.9603	0.9198	1.0000
Support Vector Machine	M2	AA k-mer	0.9587	0.9124	1.0000
Support Vector Machine	M2	AA k-mer	0.9686	0.9339	0.9444
Support Vector Machine	M3	AA k-mer	0.9673	0.9324	1.0000
Stochastic Gradient Boosting	M1	AA k-mer	0.9436	0.8825	0.9730
Stochastic Gradient Boosting	M2	AA k-mer	0.9554	0.9012	0.9730
Stochastic Gradient Boosting	M3	AA k-mer	0.9522	0.9009	0.9730
Extreme Gradient Boosting	M1	AA k-mer	0.9428	0.8759	0.9730
Extreme Gradient Boosting	M2	AA k-mer	0.9679	0.9333	0.9444
Extreme Gradient Boosting	M3	AA k-mer	0.9715	0.9420	1.0000
Random Forest	M1	SNP	0.8159	0.6079	0.8235
Random Forest	M2	SNP	0.8276	0.6308	0.8235
Random Forest	M3	SNP	0.8359	0.6490	0.8235
Support Vector Machine	M1	SNP	0.8520	0.6787	0.8485
Support Vector Machine	M2	SNP	0.8499	0.6767	0.8485
Support Vector Machine	M2	SNP	0.8450	0.6685	0.8485
Support Vector Machine	M3	SNP	0.8417	0.6664	0.8485
Stochastic Gradient Boosting	M1	SNP	0.6938	0.3526	0.7647
Stochastic Gradient Boosting	M3	SNP	0.7669	0.5140	0.8235
Extreme Gradient Boosting	M2	SNP	0.7572	0.4842	0.8235
Extreme Gradient Boosting	M3	SNP	0.7698	0.5158	0.8235
Random Forest	M1	SNP/AA k-mer	0.9454	0.8795	1.0000
Random Forest	M2	SNP/AA k-mer	0.9760	0.9446	1.0000

Table A12: Erythromycin Results of *S.aureus* (continue)

Random Forest	M3	SNP/AA k-mer	0.9540	0.8978	1.0000
Support Vector Machine	M1	SNP/AA k-mer	0.9672	0.9312	1.0000
Support Vector Machine	M2	SNP/AA k-mer	0.9659	0.9270	1.0000
Stochastic Gradient Boosting	M2	SNP/AA k-mer	0.9753	0.9479	1.0000
Stochastic Gradient Boosting	M3	SNP/AA k-mer	0.9788	0.9548	1.0000
Extreme Gradient Boosting	M1	SNP/AA k-mer	0.9826	0.9626	1.0000
Extreme Gradient Boosting	M2	SNP/AA k-mer	0.9958	0.9912	1.0000
Extreme Gradient Boosting	M3	SNP/AA k-mer	0.9958	0.9912	0.9714
Random Forest	M1	SNP/k-mer	0.9893	0.9770	0.9730
Random Forest	M2	SNP/k-mer	0.9893	0.9770	0.9730
Random Forest	M3	SNP/k-mer	0.9921	0.9829	0.9730
Support Vector Machine	M1	SNP/k-mer	0.9923	0.9831	1.0000
Support Vector Machine	M2	SNP/k-mer	0.9923	0.9831	1.0000
Support Vector Machine	M3	SNP/k-mer	0.9917	0.9824	1.0000
Stochastic Gradient Boosting	M1	SNP/k-mer	0.9692	0.9337	1.0000
Stochastic Gradient Boosting	M2	SNP/k-mer	0.9791	0.9546	0.9730
Stochastic Gradient Boosting	M3	SNP/k-mer	0.9874	0.9740	1.0000
Extreme Gradient Boosting	M1	SNP/k-mer	0.7850	0.5254	0.9231
Extreme Gradient Boosting	M2	SNP/k-mer	0.9711	0.9351	0.9730
Extreme Gradient Boosting	M3	SNP/k-mer	0.9707	0.9364	0.9730
Random Forest	M1	AA k-mer/k-mer	0.9675	0.9318	0.9730
Random Forest	M2	AA k-mer/k-mer	0.9759	0.9488	0.9730
Random Forest	M3	AA k-mer/k-mer	0.9756	0.9485	0.9730
Support Vector Machine	M1	AA k-mer/k-mer	0.9756	0.9468	1.0000
Support Vector Machine	M2	AA k-mer/k-mer	0.9826	0.9614	0.9730
Support Vector Machine	M3	AA k-mer/k-mer	0.9769	0.9530	1.0000
Stochastic Gradient Boosting	M1	AA k-mer/k-mer	0.9002	0.8033	0.9091
Stochastic Gradient Boosting	M2	AA k-mer/k-mer	0.9749	0.9464	0.9730

Table A12: Erythromycin Results of *S.aureus* (continue)

Stochastic Gradient Boosting	M3	AA k-mer/k-mer	0.9840	0.9667	0.9412
Extreme Gradient Boosting	M1	AA k-mer/k-mer	0.9436	0.8839	1.0000
Extreme Gradient Boosting	M2	AA k-mer/k-mer	0.9797	0.9584	0.9474
Extreme Gradient Boosting	M3	AA k-mer/k-mer	0.9837	0.9645	0.9730

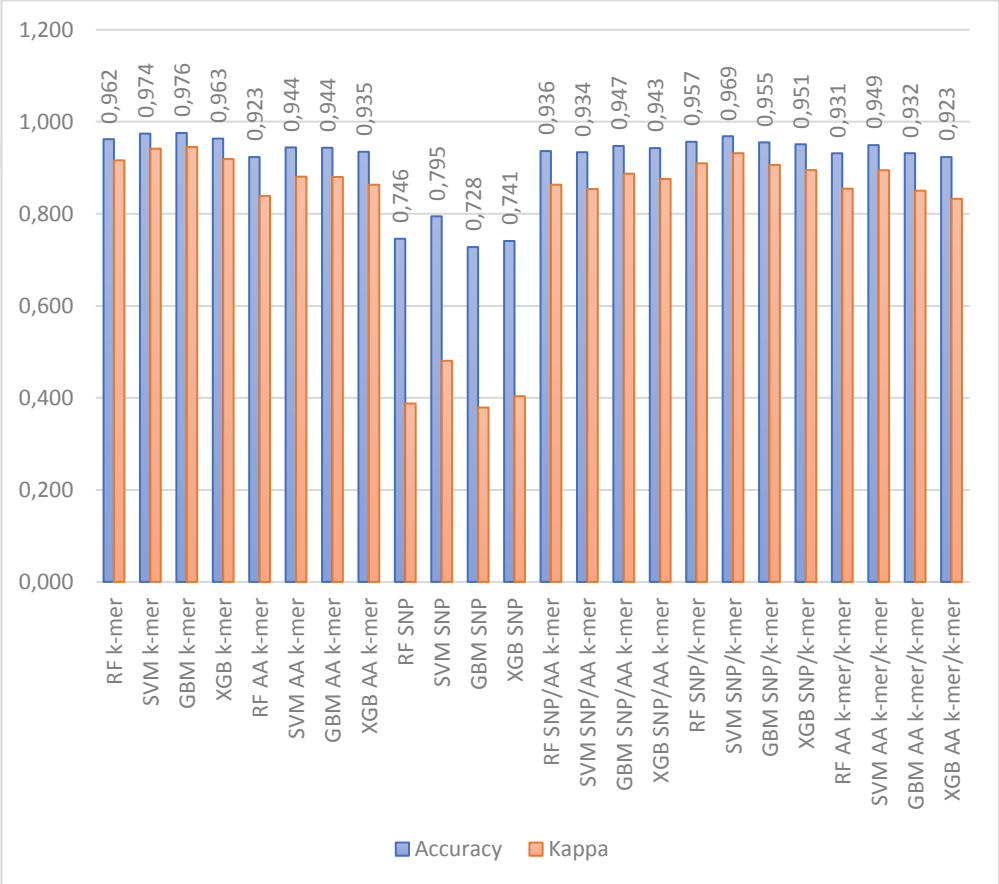

Figure A4. Comparison of Accuracy of all erythromycin *S.aureus* models.

Table A13: Tetracycline Results of *S.aureus*

Algorithm	Model#	input	Accuracy	Kappa	F1-score
Random Forest	M1	k-mer	0.9891	0.9752	0.9744
Random Forest	M2	k-mer	0.9891	0.9759	0.9744
Random Forest	M3	k-mer	0.9889	0.9757	0.9744
Support Vector Machine	M1	k-mer	0.9832	0.9645	0.9744
Support Vector Machine	M1-cost	k-mer	0.9749	0.9468	0.9744
Support Vector Machine	M2	k-mer	0.9923	0.9831	0.9744
Support Vector Machine	M3	k-mer	0.9923	0.9831	0.9744
Stochastic Gradient Boosting	M1	k-mer	0.9763	0.9508	0.9744
Stochastic Gradient Boosting	M3	k-mer	0.9749	0.9463	0.9744
Extreme Gradient Boosting	M1	k-mer	0.9587	0.9066	0.9500
Extreme Gradient Boosting	M2	k-mer	0.9708	0.9330	0.9744
Extreme Gradient Boosting	M3	k-mer	0.9708	0.9330	0.9744
Random Forest	M1	AA k-mer	0.9342	0.8499	0.9412
Random Forest	M2	AA k-mer	0.9784	0.9526	0.9412
Random Forest	M3	AA k-mer	0.9408	0.8706	0.9444
Support Vector Machine	M1	AA k-mer	0.9603	0.9198	1.0000
Support Vector Machine	M2	AA k-mer	0.9587	0.9124	1.0000
Support Vector Machine	M2	AA k-mer	0.9686	0.9339	0.9444
Support Vector Machine	M3	AA k-mer	0.9673	0.9324	1.0000
Stochastic Gradient Boosting	M1	AA k-mer	0.9436	0.8825	0.9730
Stochastic Gradient Boosting	M2	AA k-mer	0.9554	0.9012	0.9730
Stochastic Gradient Boosting	M3	AA k-mer	0.9522	0.9009	0.9730
Extreme Gradient Boosting	M1	AA k-mer	0.9428	0.8759	0.9730
Extreme Gradient Boosting	M2	AA k-mer	0.9679	0.9333	0.9444
Extreme Gradient Boosting	M3	AA k-mer	0.9715	0.9420	1.0000
Random Forest	M1	SNP	0.8159	0.6079	0.8235
Random Forest	M2	SNP	0.8276	0.6308	0.8235
Random Forest	M3	SNP	0.8359	0.6490	0.8235
Support Vector Machine	M1	SNP	0.8520	0.6787	0.8485
Support Vector Machine	M2	SNP	0.8499	0.6767	0.8485
Support Vector Machine	M2	SNP	0.8450	0.6685	0.8485
Support Vector Machine	M3	SNP	0.8417	0.6664	0.8485
Stochastic Gradient Boosting	M1	SNP	0.6938	0.3526	0.7647
Stochastic Gradient Boosting	M3	SNP	0.7669	0.5140	0.8235

Table A13: Tetracycline Results of *S.aureus* (contionue)

Extreme Gradient Boosting	M2	SNP	0.7572	0.4842	0.8235
Extreme Gradient Boosting	M3	SNP	0.7698	0.5158	0.8235
Random Forest	M1	SNP/AA k-mer	0.9454	0.8795	1.0000
Random Forest	M2	SNP/AA k-mer	0.9760	0.9446	1.0000
Random Forest	M3	SNP/AA k-mer	0.9540	0.8978	1.0000
Support Vector Machine	M1	SNP/AA k-mer	0.9672	0.9312	1.0000
Support Vector Machine	M2	SNP/AA k-mer	0.9659	0.9270	1.0000
Stochastic Gradient Boosting	M2	SNP/AA k-mer	0.9753	0.9479	1.0000
Stochastic Gradient Boosting	M3	SNP/AA k-mer	0.9788	0.9548	1.0000
Extreme Gradient Boosting	M1	SNP/AA k-mer	0.9826	0.9626	1.0000
Extreme Gradient Boosting	M2	SNP/AA k-mer	0.9958	0.9912	1.0000
Extreme Gradient Boosting	M3	SNP/AA k-mer	0.9958	0.9912	0.9714
Random Forest	M1	SNP/k-mer	0.9893	0.9770	0.9730
Random Forest	M2	SNP/k-mer	0.9893	0.9770	0.9730
Random Forest	M3	SNP/k-mer	0.9921	0.9829	0.9730
Support Vector Machine	M1	SNP/k-mer	0.9923	0.9831	1.0000
Support Vector Machine	M2	SNP/k-mer	0.9923	0.9831	1.0000
Support Vector Machine	M3	SNP/k-mer	0.9917	0.9824	1.0000
Stochastic Gradient Boosting	M1	SNP/k-mer	0.9692	0.9337	1.0000
Stochastic Gradient Boosting	M2	SNP/k-mer	0.9791	0.9546	0.9730
Stochastic Gradient Boosting	M3	SNP/k-mer	0.9874	0.9740	1.0000
Extreme Gradient Boosting	M1	SNP/k-mer	0.7850	0.5254	0.9231
Extreme Gradient Boosting	M2	SNP/k-mer	0.9711	0.9351	0.9730
Extreme Gradient Boosting	M3	SNP/k-mer	0.9707	0.9364	0.9730
Random Forest	M1	AA k-mer/k-mer	0.9675	0.9318	0.9730
Random Forest	M2	AA k-mer/k-mer	0.9759	0.9488	0.9730
Random Forest	M3	AA k-mer/k-mer	0.9756	0.9485	0.9730
Support Vector Machine	M1	AA k-mer/k-mer	0.9756	0.9468	1.0000
Support Vector Machine	M2	AA k-mer/k-mer	0.9826	0.9614	0.9730
Support Vector Machine	M3	AA k-mer/k-mer	0.9769	0.9530	1.0000
Stochastic Gradient Boosting	M1	AA k-mer/k-mer	0.9002	0.8033	0.9091
Stochastic Gradient Boosting	M2	AA k-mer/k-mer	0.9749	0.9464	0.9730
Stochastic Gradient Boosting	M3	AA k-mer/k-mer	0.9840	0.9667	0.9412
Extreme Gradient Boosting	M1	AA k-mer/k-mer	0.9436	0.8839	1.0000
Extreme Gradient Boosting	M2	AA k-mer/k-mer	0.9797	0.9584	0.9474

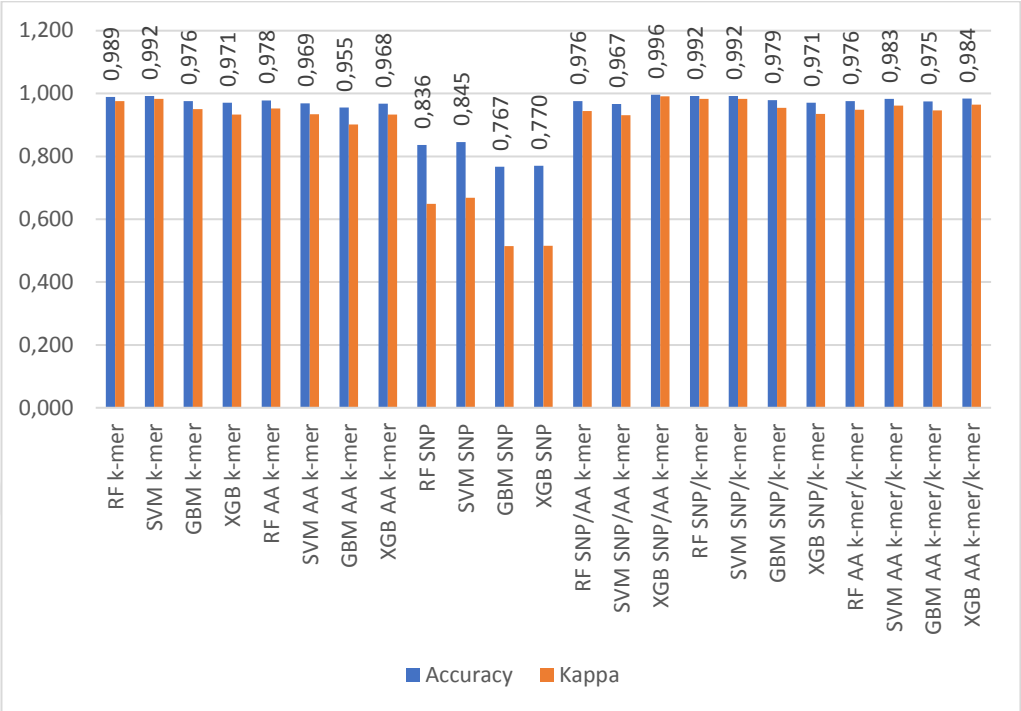

Figure A5. Comparison of Accuracy of all Tetracycline *S.aureus* models.

Table A14: Gentamicin Results of *S.aureus*

Algorithm	Model#	input	Accuracy training	Kappa training	F1-score
Random Forest	M1	k-mer	0.983	0.959	1.000
Random Forest	M2	k-mer	0.984	0.963	1.000
Random Forest	M3	k-mer	0.984	0.963	1.000
Support Vector Machine	M1	k-mer	0.992	0.983	1.000
Support Vector Machine	M1-cost	k-mer	0.992	0.982	1.000
Support Vector Machine	M2	k-mer	0.992	0.982	1.000
Support Vector Machine	M3	k-mer	0.992	0.980	1.000
Stochastic Gradient Boosting	M2	k-mer	0.984	0.961	1.000
Stochastic Gradient Boosting	M3	k-mer	0.996	0.990	1.000
Extreme Gradient Boosting	M1	k-mer	0.976	0.943	1.000
Extreme Gradient Boosting	M2	k-mer	0.984	0.961	1.000

Table A14: Gentamicin Results of *S.aureus* (continue)

Extreme Gradient Boosting	M3	k-mer	0.984	0.961	1.000
Random Forest	M1	AA k-mer	1.000	1.000	1.000
Random Forest	M2	AA k-mer	1.000	1.000	1.000
Random Forest	M3	AA k-mer	1.000	1.000	1.000
Support Vector Machine	M1	AA k-mer	0.992	0.982	1.000
Support Vector Machine	M2	AA k-mer	0.992	0.982	1.000
Support Vector Machine	M3	AA k-mer	0.992	0.983	1.000
Stochastic Gradient Boosting	M1	AA k-mer	0.951	0.876	1.000
Stochastic Gradient Boosting	M2	AA k-mer	0.991	0.977	1.000
Stochastic Gradient Boosting	M3	AA k-mer	0.992	0.980	1.000
Extreme Gradient Boosting	M1	AA k-mer	0.991	0.974	1.000
Extreme Gradient Boosting	M2	AA k-mer	0.992	0.980	1.000
Extreme Gradient Boosting	M3	AA k-mer	0.991	0.977	1.000
Random Forest	M1	SNP	0.804	0.450	0.500
Random Forest	M2	SNP	0.854	0.606	0.500
Random Forest	M3	SNP	0.854	0.606	0.500
Support Vector Machine	M1	SNP	0.851	0.612	0.500
Support Vector Machine	M2	SNP	0.860	0.627	0.500
Support Vector Machine	M3	SNP	0.860	0.630	0.500
Support Vector Machine	M4	SNP	0.843	0.589	0.500
Stochastic Gradient Boosting	M1	SNP	0.678	0.172	0.640
Stochastic Gradient Boosting	M2	SNP	0.674	0.029	0.640
Stochastic Gradient Boosting	M3	SNP	0.676	0.266	0.640
Extreme Gradient Boosting	M3	SNP	0.732	0.223	0.200
Random Forest	M1	SNP/AA k-mer	1.000	1.000	1.000
Random Forest	M2	SNP/AA k-mer	0.991	0.976	1.000
Random Forest	M3	SNP/AA k-mer	0.997	0.991	1.000
Extreme Gradient Boosting	M1	SNP/AA k-mer	0.992	0.983	1.000

Table A14: Gentamicin Results of *S.aureus* (continue)

Extreme Gradient Boosting	M2	SNP/AA k-mer	0.992	0.980	1.000
Extreme Gradient Boosting	M3	SNP/AA k-mer	0.992	0.980	1.000
Random Forest	M1	SNP/k-mer	0.992	0.980	0.947
Random Forest	M2	SNP/k-mer	0.992	0.980	0.947
Random Forest	M3	SNP/k-mer	0.992	0.980	0.947
Support Vector Machine	M1	SNP/k-mer	0.992	0.980	0.947
Support Vector Machine	M2	SNP/k-mer	0.967	0.929	0.947
Support Vector Machine	M3	SNP/k-mer	0.992	0.980	0.947
Stochastic Gradient Boosting	M2	SNP/k-mer	0.991	0.977	0.947
Stochastic Gradient Boosting	M3	SNP/k-mer	0.992	0.980	0.947
Extreme Gradient Boosting	M1	SNP/k-mer	0.958	0.892	0.889
Extreme Gradient Boosting	M2	SNP/k-mer	0.967	0.916	0.947
Extreme Gradient Boosting	M3	SNP/k-mer	0.970	0.926	0.947
Random Forest	M1	AA k-mer/k-mer	1.000	1.000	1.000
Random Forest	M2	AA k-mer/k-mer	0.995	0.988	1.000
Random Forest	M3	AA k-mer/k-mer	0.992	0.983	0.941
Support Vector Machine	M1	AA k-mer/k-mer	0.992	0.983	1.000
Support Vector Machine	M2	AA k-mer/k-mer	0.992	0.982	1.000
Support Vector Machine	M3	AA k-mer/k-mer	0.992	0.982	1.000
Stochastic Gradient Boosting	M1	AA k-mer/k-mer	0.917	0.790	0.875
Stochastic Gradient Boosting	M2	AA k-mer/k-mer	1.000	1.000	0.941
Stochastic Gradient Boosting	M3	AA k-mer/k-mer	1.000	1.000	0.941
Extreme Gradient Boosting	M1	AA k-mer/k-mer	1.000	1.000	0.941
Extreme Gradient Boosting	M2	AA k-mer/k-mer	0.996	0.990	0.941
Extreme Gradient Boosting	M3	AA k-mer/k-mer	1.000	1.000	0.941

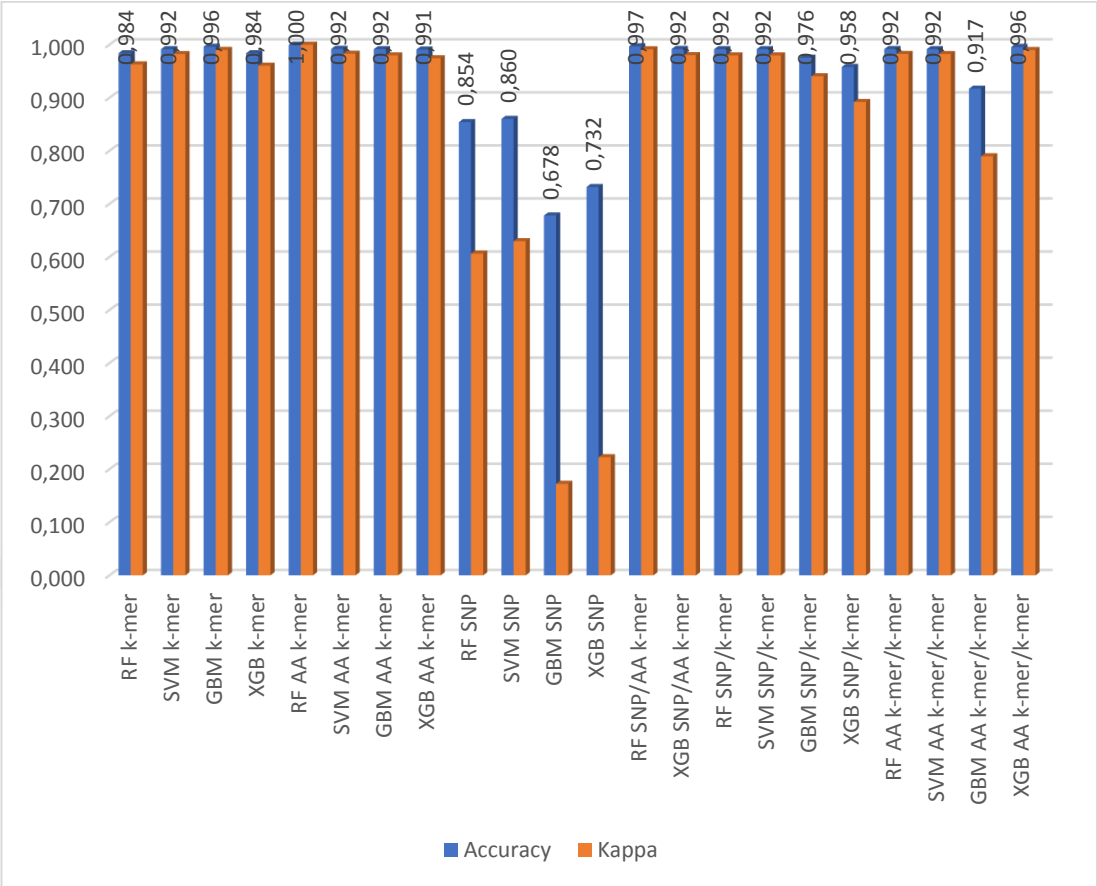

Figure A6. Comparison of Accuracy of all Gentamicin *S.aureus* models.

Table A15: Ciprofloxacin Results of S.aureus

Algorithm	Model#	Input	Accuracy	Kappa	F1-score
Random Forest	M1	k-mer	0.9508547	0.8923563	0.974359
Random Forest	M2	k-mer	0.9754274	0.9439746	0.974359
Random Forest	M3	k-mer	0.9536325	0.899072	0.974359
Support Vector Machine	M1	k-mer	0.9589744	0.9090432	0.974359
Support Vector Machine	M1-cost	k-mer	0.9673077	0.9290432	0.974359
Support Vector Machine	M2	k-mer	0.950641	0.893504	1
Support Vector Machine	M3	k-mer	0.9679487	0.9285867	0.974359
Stochastic Gradient Boosting	M1	k-mer	0.9442308	0.8842408	0.9473684
Stochastic Gradient Boosting	M2	k-mer	0.9516026	0.8938656	0.9473684
Stochastic Gradient Boosting	M3	k-mer	0.9637821	0.9220376	0.9230769
Extreme Gradient Boosting	M1	k-mer	0.9102564	0.8022678	0.9473684

Table A15: Ciprofloxacin Results of S.aureus (continue)

Extreme Gradient Boosting	M2	k-mer	0.9596154	0.9110572	0.9189189
Extreme Gradient Boosting	M3	k-mer	0.9560897	0.9064137	0.9189189
Random Forest	M1	AA k-mer	0.909965	0.8090308	0.9444444
Random Forest	M2	AA k-mer	0.9157731	0.8218631	0.9444444
Random Forest	M3	AA k-mer	0.9153069	0.8205268	0.9444444
Support Vector Machine	M1	AA k-mer	0.9346154	0.8564738	1
Support Vector Machine	M2	AA k-mer	0.9274476	0.8466029	1
Support Vector Machine	M2	AA k-mer	0.95	0.8958333	1
Support Vector Machine	M3	AA k-mer	0.9115385	0.8061755	1
Stochastic Gradient Boosting	M1	AA k-mer	0.8929487	0.7731145	0.9189189
Stochastic Gradient Boosting	M2	AA k-mer	0.9176282	0.8174268	1
Stochastic Gradient Boosting	M3	AA k-mer	0.9348776	0.8592301	0.974359
Extreme Gradient Boosting	M1	AA k-mer	0.8793706	0.7384871	0.9444444
Extreme Gradient Boosting	M2	AA k-mer	0.9219988	0.8270912	0.95
Extreme Gradient Boosting	M3	AA k-mer	0.9012821	0.7782273	1
Random Forest	M1	SNP	0.8685315	0.7036529	0.8888889
Random Forest	M2	SNP	0.8764375	0.7217033	0.8823529
Random Forest	M3	SNP	0.8792152	0.7275856	0.8125
Support Vector Machine	M1	SNP	0.8596154	0.6909098	0.75
Support Vector Machine	M2	SNP	0.8828671	0.7401725	0.7741935
Support Vector Machine	M2	SNP	0.8857809	0.7661961	0.8125
Support Vector Machine	M3	SNP	0.87669	0.724905	0.8125
Stochastic Gradient Boosting	M1	SNP	0.8982517	0.76384874	0.8125
Stochastic Gradient Boosting	M2	SNP	0.8143939	0.5697198	0.8125
Stochastic Gradient Boosting	M3	SNP	0.8197552	0.6029516	0.8125
Extreme Gradient Boosting	M1	SNP	0.7668415	0.47269451	0.9189189
Extreme Gradient Boosting	M2	SNP	0.8445513	0.6495401	0.8125
Extreme Gradient Boosting	M3	SNP	0.8638112	0.7007827	0.8125

Table A15: Ciprofloxacin Results of S.aureus (continue)

Random Forest	M1	SNP/AA k-mer	0.9238345	0.8359389	0.8823529
Random Forest	M2	SNP/AA k-mer	0.9266123	0.8411676	0.8823529
Random Forest	M3	SNP/AA k-mer	0.9266123	0.8411676	0.8823529
Support Vector Machine	M1	SNP/AA k-mer	0.9338578	0.8516393	0.9473684
Support Vector Machine	M2	SNP/AA k-mer	0.9665501	0.9303992	0.9444444
Support Vector Machine	M3	SNP/AA k-mer	0.950641	0.8978228	0.8888889
Stochastic Gradient Boosting	M1	SNP/AA k-mer	0.9004079	0.7973656	0.7741935
Stochastic Gradient Boosting	M2	SNP/AA k-mer	0.9501457	0.8928147	0.972973
Stochastic Gradient Boosting	M3	SNP/AA k-mer	0.9435897	0.8805146	0.9444444
Extreme Gradient Boosting	M1	SNP/AA k-mer	0.9166667	0.807535	0.9142857
Extreme Gradient Boosting	M2	SNP/AA k-mer	0.9439103	0.8805264	0.9142857
Extreme Gradient Boosting	M3	SNP/AA k-mer	0.9306527	0.8545991	0.9444444
Random Forest	M1	SNP/k-mer	0.9463869	0.8817437	0.9473684
Random Forest	M2	SNP/k-mer	0.9574981	0.9069234	0.9473684
Random Forest	M3	SNP/k-mer	0.9607809	0.9128373	0.9473684
Support Vector Machine	M1	SNP/k-mer	0.9666667	0.9256863	0.972973
Support Vector Machine	M2	SNP/k-mer	0.9665501	0.9291787	0.9473684
Support Vector Machine	M3	SNP/k-mer	0.9582168	0.9091787	0.972973
Stochastic Gradient Boosting	M1	SNP/k-mer	0.8934732	0.7707664	0.9473684
Stochastic Gradient Boosting	M2	SNP/k-mer	0.9466783	0.8830913	0.974359
Stochastic Gradient Boosting	M3	SNP/k-mer	0.962704	0.9224478	0.9473684
Extreme Gradient Boosting	M1	SNP/k-mer	0.8678322	0.7047151	0.9
Extreme Gradient Boosting	M2	SNP/k-mer	0.9507867	0.896337	0.9473684
Extreme Gradient Boosting	M3	SNP/k-mer	0.9543706	0.8993327	0.9473684
Random Forest	M1	AA k-mer/k-mer	0.9234266	0.838222	0.9189189
Random Forest	M2	AA k-mer/k-mer	0.9262044	0.8440554	0.9189189

Table A15: Ciprofloxacin Results of S.aureus (continue)

Random Forest	M3	AA k-mer/k-mer	0.9315462	0.8568061	0.8947368
Support Vector Machine	M1	AA k-mer/k-mer	0.9435897	0.8741903	0.972973
Support Vector Machine	M2	AA k-mer/k-mer	0.9519231	0.9008112	0.9230769
Support Vector Machine	M3	AA k-mer/k-mer	0.9429487	0.8836592	0.9230769
Stochastic Gradient Boosting	M1	AA k-mer/k-mer	0.9448718	0.8816903	0.95
Stochastic Gradient Boosting	M2	AA k-mer/k-mer	0.9339161	0.8538242	0.9268293
Stochastic Gradient Boosting	M3	AA k-mer/k-mer	0.9268648	0.8447057	0.95
Extreme Gradient Boosting	M1	AA k-mer/k-mer	0.6482517	0	0.7916667
Extreme Gradient Boosting	M2	AA k-mer/k-mer	0.9265443	0.8387757	0.95
Extreme Gradient Boosting	M3	AA k-mer/k-mer	0.9291084	0.8406923	0.9230769

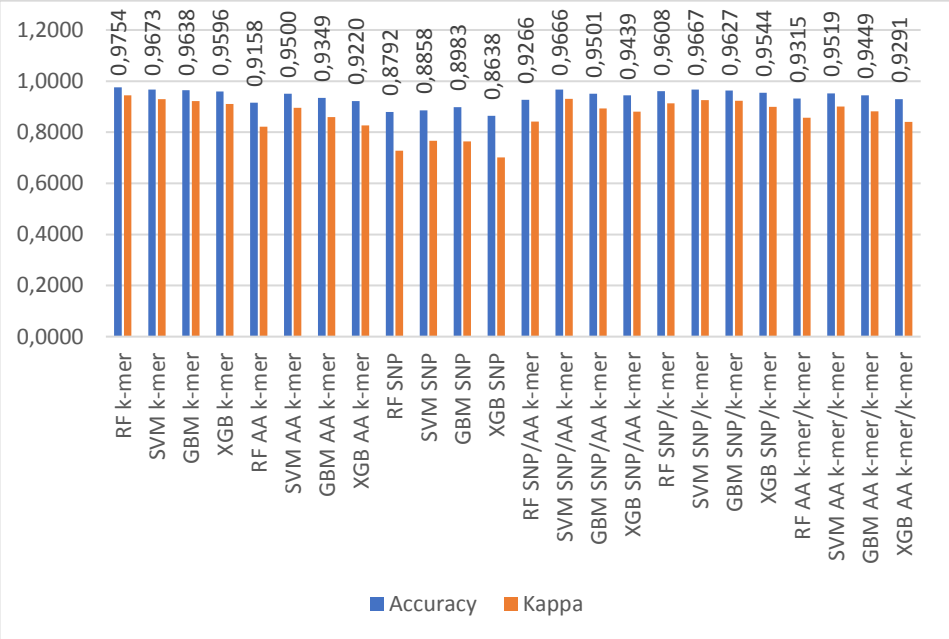

Figure A7. Comparison of Accuracy of all Ciprofloxacin S.aureus models.

Appendix 5

Table A16: Ampicillin Results of *E.faecium*

Algorithm	Model#	input	Accuracy	Kappa	F1-score
Random Forest	M1	k-mer	0.9533	0.8537	0.9231
Random Forest	M2	k-mer	0.9467	0.8045	0.9474
Random Forest	M3	k-mer	0.9267	0.7700	0.8649
Support Vector Machine	M1	k-mer	0.9100	0.6828	0.8889
Support Vector Machine	M1-cost	k-mer	0.9200	0.7300	0.8889
Support Vector Machine	M2	k-mer	0.9300	0.7198	0.9268
Support Vector Machine	M3	k-mer	0.9300	0.8059	0.8649
Stochastic Gradient Boosting	M1	k-mer	0.9100	0.7240	0.9474
Stochastic Gradient Boosting	M2	k-mer	0.9150	0.7590	0.9474
Stochastic Gradient Boosting	M3	k-mer	0.9250	0.7734	0.9474
Extreme Gradient Boosting	M2	k-mer	0.9200	0.7566	0.9744
Extreme Gradient Boosting	M3	k-mer	0.9300	0.7714	0.8889
Random Forest	M1	AA k-mer	0.9533	0.8355	0.9730
Random Forest	M2	AA k-mer	0.9567	0.8483	0.9730
Random Forest	M3	AA k-mer	0.9633	0.8634	0.9730
Support Vector Machine	M1	AA k-mer	0.9300	0.7221	1.0000
Support Vector Machine	M2	AA k-mer	0.9400	0.7692	1.0000
Support Vector Machine	M2	AA k-mer	0.9400	0.7958	0.9474
Support Vector Machine	M3	AA k-mer	0.9100	0.6452	1.0000
Stochastic Gradient Boosting	M1	AA k-mer	0.8700	0.6938	0.8485
Stochastic Gradient Boosting	M2	AA k-mer	0.9050	0.6823	0.8485
Stochastic Gradient Boosting	M3	AA k-mer	0.9100	0.7048	0.9444
Extreme Gradient Boosting	M1	AA k-mer	0.9300	0.7343	0.9730
Extreme Gradient Boosting	M2	AA k-mer	0.9150	0.6736	0.9474
Extreme Gradient Boosting	M3	AA k-mer	0.9500	0.8094	0.9189
Random Forest	M1	SNP	0.9200	0.6769	0.9474
Random Forest	M2	SNP	0.9200	0.6769	0.9474

Table A16: Ampicillin Results of *E.faecium* (continue)

Random Forest	M3	SNP	0.9233	0.6897	0.9474
Support Vector Machine	M1	SNP	0.9600	0.8231	0.9474
Support Vector Machine	M2	SNP	0.9900	0.9615	0.9474
Support Vector Machine	M3	SNP	0.9700	0.8846	0.9474
Stochastic Gradient Boosting	M1	SNP	0.9400	0.7462	0.9474
Stochastic Gradient Boosting	M2	SNP	0.9350	0.7287	0.9474
Stochastic Gradient Boosting	M3	SNP	0.9350	0.7402	0.9474
Extreme Gradient Boosting	M2	SNP	0.9200	0.6583	0.9474
Extreme Gradient Boosting	M3	SNP	0.9150	0.6269	0.9474
Random Forest	M1	SNP/AA k-mer	0.9633	0.8719	0.9474
Random Forest	M2	SNP/AA k-mer	0.9633	0.8719	0.9474
Random Forest	M3	SNP/AA k-mer	0.9700	0.8975	0.9474
Support Vector Machine	M1	SNP/AA k-mer	0.9800	0.9231	0.9474
Support Vector Machine	M2	SNP/AA k-mer	0.9900	0.9737	0.9730
Support Vector Machine	M3	SNP/AA k-mer	0.9500	0.7737	0.9474
Stochastic Gradient Boosting	M1	SNP/AA k-mer	0.9300	0.7915	0.9474
Stochastic Gradient Boosting	M2	SNP/AA k-mer	0.9650	0.8726	0.9474
Stochastic Gradient Boosting	M3	SNP/AA k-mer	0.9650	0.8775	0.9474
Extreme Gradient Boosting	M1	SNP/AA k-mer	0.9400	0.7837	0.9474
Extreme Gradient Boosting	M2	SNP/AA k-mer	0.9650	0.8715	0.9474
Extreme Gradient Boosting	M3	SNP/AA k-mer	0.9500	0.7962	0.9231
Random Forest	M1	SNP/k-mer	0.9500	0.8000	0.9231
Random Forest	M2	SNP/k-mer	0.9500	0.8000	0.9231
Random Forest	M3	SNP/k-mer	0.9500	0.8000	0.9474
Support Vector Machine	M1	SNP/k-mer	0.9800	0.9231	0.9730
Support Vector Machine	M2	SNP/k-mer	0.9900	0.9737	0.9730
Support Vector Machine	M3	SNP/k-mer	0.9300	0.9231	0.9730

Table A16: Ampicillin Results of *E.faecium* (continue)

Stochastic Gradient Boosting	M1	SNP/k-mer	0.9400	0.7935	0.9474
Stochastic Gradient Boosting	M2	SNP/k-mer	0.9500	0.8034	0.9474
Stochastic Gradient Boosting	M3	SNP/k-mer	0.9500	0.8022	0.9730
Extreme Gradient Boosting	M1	SNP/k-mer	0.9300	0.7452	0.9500
Extreme Gradient Boosting	M2	SNP/k-mer	0.9350	0.7385	0.9231
Extreme Gradient Boosting	M3	SNP/k-mer	0.9300	0.7210	0.9500
Random Forest	M1	AA k-mer/k-mer	0.9467	0.7968	0.9730
Random Forest	M2	AA k-mer/k-mer	0.9533	0.8301	0.9730
Random Forest	M3	AA k-mer/k-mer	0.9567	0.8401	0.9474
Support Vector Machine	M1	AA k-mer/k-mer	0.9200	0.6958	0.9744
Support Vector Machine	M2	AA k-mer/k-mer	0.9400	0.7231	0.9500
Support Vector Machine	M3	AA k-mer/k-mer	0.9300	0.7221	0.9500
Stochastic Gradient Boosting	M1	AA k-mer/k-mer	0.9000	0.6432	0.9474
Stochastic Gradient Boosting	M2	AA k-mer/k-mer	0.9150	0.7279	0.9474
Stochastic Gradient Boosting	M3	AA k-mer/k-mer	0.9450	0.8134	0.9730
Extreme Gradient Boosting	M1	AA k-mer/k-mer	0.8200	0.2973	0.9048
Extreme Gradient Boosting	M2	AA k-mer/k-mer	0.9650	0.8599	0.9189
Extreme Gradient Boosting	M3	AA k-mer/k-mer	0.9500	0.8083	0.9444

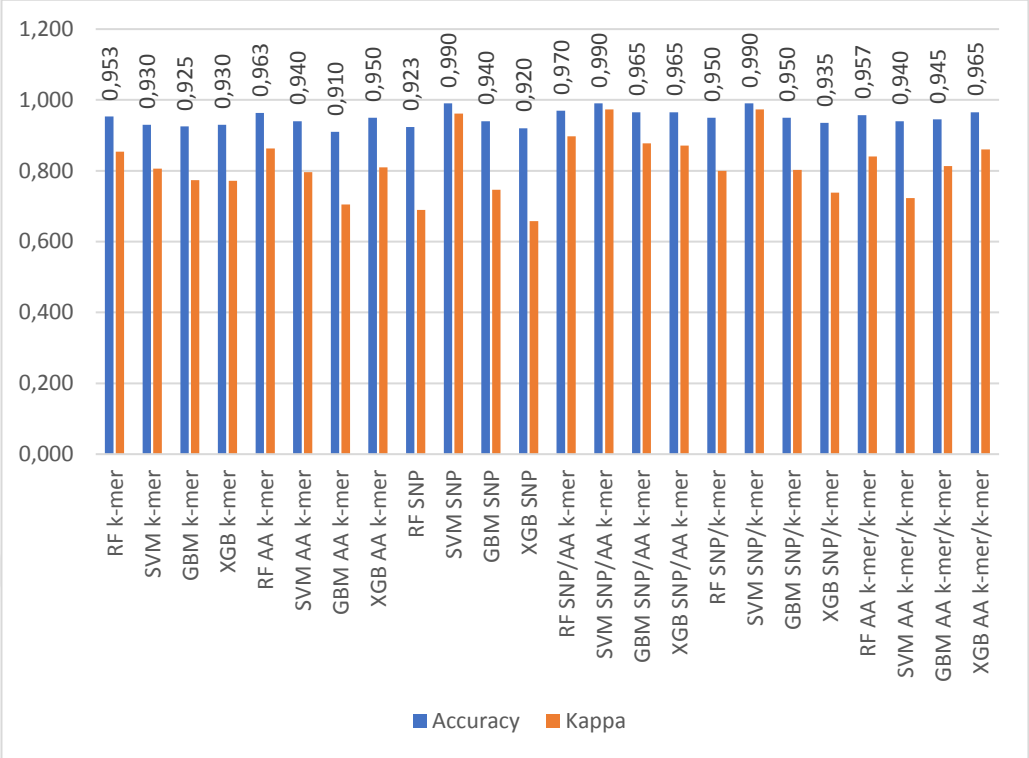

Figure A8. Comparison of Accuracy of all Ampicillin *E.faecium* models.

Table A17: Vancomycin Results of *E.faecium*

Algorithm	Model#	input	Accuracy	Kappa	F1-score
Random Forest	M1	k-mer	0.909	0.798	0.889
Random Forest	M2	k-mer	0.909	0.798	0.889
Support Vector Machine	M1	k-mer	0.912	0.804	0.889
Support Vector Machine	M1-cost	k-mer	0.910	0.806	0.889
Support Vector Machine	M2	k-mer	0.870	0.805	0.889
Support Vector Machine	M3	k-mer	0.871	0.800	0.889
Stochastic Gradient Boosting	M1	k-mer	0.862	0.703	0.947
Stochastic Gradient Boosting	M2	k-mer	0.892	0.767	0.824
Stochastic Gradient Boosting	M3	k-mer	0.901	0.788	0.889
Extreme Gradient Boosting	M1	k-mer	0.883	0.748	0.824
Extreme Gradient Boosting	M2	k-mer	0.885	0.754	0.824

Table A17: Vancomycin Results of *E.faecium* (continue)

Extreme Gradient Boosting	M3	k-mer	0.897	0.779	0.824
Random Forest	M1	AA k-mer	0.940	0.877	0.889
Random Forest	M2	AA k-mer	0.944	0.886	0.889
Random Forest	M3	AA k-mer	0.944	0.884	0.889
Support Vector Machine	M1	AA k-mer	0.891	0.786	0.889
Support Vector Machine	M2	AA k-mer	0.912	0.808	0.889
Support Vector Machine	M2	AA k-mer	0.899	0.773	0.889
Support Vector Machine	M3	AA k-mer	0.911	0.807	0.889
Stochastic Gradient Boosting	M1	AA k-mer	0.841	0.661	0.750
Stochastic Gradient Boosting	M2	AA k-mer	0.915	0.825	0.842
Stochastic Gradient Boosting	M3	AA k-mer	0.935	0.860	0.889
Extreme Gradient Boosting	M1	AA k-mer	0.909	0.800	0.889
Extreme Gradient Boosting	M2	AA k-mer	0.956	0.909	0.842
Extreme Gradient Boosting	M3	AA k-mer	0.949	0.898	0.842
Random Forest	M1	SNP	0.759	0.520	0.632
Random Forest	M2	SNP	0.756	0.514	0.632
Support Vector Machine	M1	SNP	0.736	0.462	0.632
Support Vector Machine	M2	SNP	0.780	0.580	0.636
Support Vector Machine	M3	SNP	0.729	0.432	0.636
Stochastic Gradient Boosting	M1	SNP	0.778	0.536	0.632
Stochastic Gradient Boosting	M2	SNP	0.753	0.490	0.874
Stochastic Gradient Boosting	M3	SNP	0.764	0.506	0.588
Extreme Gradient Boosting	M2	SNP	0.781	0.548	0.556
Extreme Gradient Boosting	M3	SNP	0.760	0.505	0.588
Random Forest	M1	SNP/AA k-mer	0.936	0.872	0.900
Random Forest	M2	SNP/AA k-mer	0.949	0.899	1.000
Random Forest	M3	SNP/AA k-mer	0.960	0.920	0.952
Support Vector Machine	M1	SNP/AA k-mer	0.929	0.857	0.952

Table A17: Vancomycin Results of *E.faecium* (continue)

Support Vector Machine	M2	SNP/AA k-mer	0.950	0.895	0.857
Support Vector Machine	M3	SNP/AA k-mer	0.949	0.895	0.909
Stochastic Gradient Boosting	M1	SNP/AA k-mer	0.901	0.789	0.842
Stochastic Gradient Boosting	M2	SNP/AA k-mer	0.966	0.930	0.952
Stochastic Gradient Boosting	M3	SNP/AA k-mer	0.960	0.919	0.909
Extreme Gradient Boosting	M1	SNP/AA k-mer	0.879	0.731	1.000
Extreme Gradient Boosting	M2	SNP/AA k-mer	0.954	0.902	0.952
Extreme Gradient Boosting	M3	SNP/AA k-mer	0.929	0.854	0.952
Random Forest	M1	SNP/k-mer	0.922	0.836	0.947
Random Forest	M2	SNP/k-mer	0.942	0.885	0.947
Random Forest	M3	SNP/k-mer	0.942	0.886	0.947
Support Vector Machine	M1	SNP/k-mer	0.912	0.824	0.900
Support Vector Machine	M2	SNP/k-mer	0.939	0.878	0.947
Support Vector Machine	M3	SNP/k-mer	0.930	0.857	0.947
Stochastic Gradient Boosting	M1	SNP/k-mer	0.901	0.800	0.889
Stochastic Gradient Boosting	M2	SNP/k-mer	0.899	0.785	0.889
Stochastic Gradient Boosting	M3	SNP/k-mer	0.908	0.814	0.889
Extreme Gradient Boosting	M1	SNP/k-mer	0.900	0.800	0.889
Extreme Gradient Boosting	M2	SNP/k-mer	0.913	0.820	0.889
Extreme Gradient Boosting	M3	SNP/k-mer	0.909	0.813	0.889
Random Forest	M1	AA k-mer/k-mer	0.942	0.882	0.900
Random Forest	M2	AA k-mer/k-mer	0.942	0.881	0.857
Random Forest	M3	AA k-mer/k-mer	0.948	0.894	0.857
Support Vector Machine	M1	AA k-mer/k-mer	0.908	0.812	0.900
Support Vector Machine	M2	AA k-mer/k-mer	0.920	0.840	0.900
Support Vector Machine	M3	AA k-mer/k-mer	0.909	0.808	0.900
Stochastic Gradient Boosting	M1	AA k-mer/k-mer	0.900	0.780	0.778

Table A17: Vancomycin Results of *E.faecium* (continue)

Stochastic Gradient Boosting	M2	AA k-mer/k-mer	0.924	0.846	0.857
Stochastic Gradient Boosting	M3	AA k-mer/k-mer	0.959	0.919	0.857
Extreme Gradient Boosting	M1	AA k-mer/k-mer	0.872	0.728	0.889
Extreme Gradient Boosting	M2	AA k-mer/k-mer	0.939	0.876	0.800
Extreme Gradient Boosting	M3	AA k-mer/k-mer	0.944	0.883	0.900

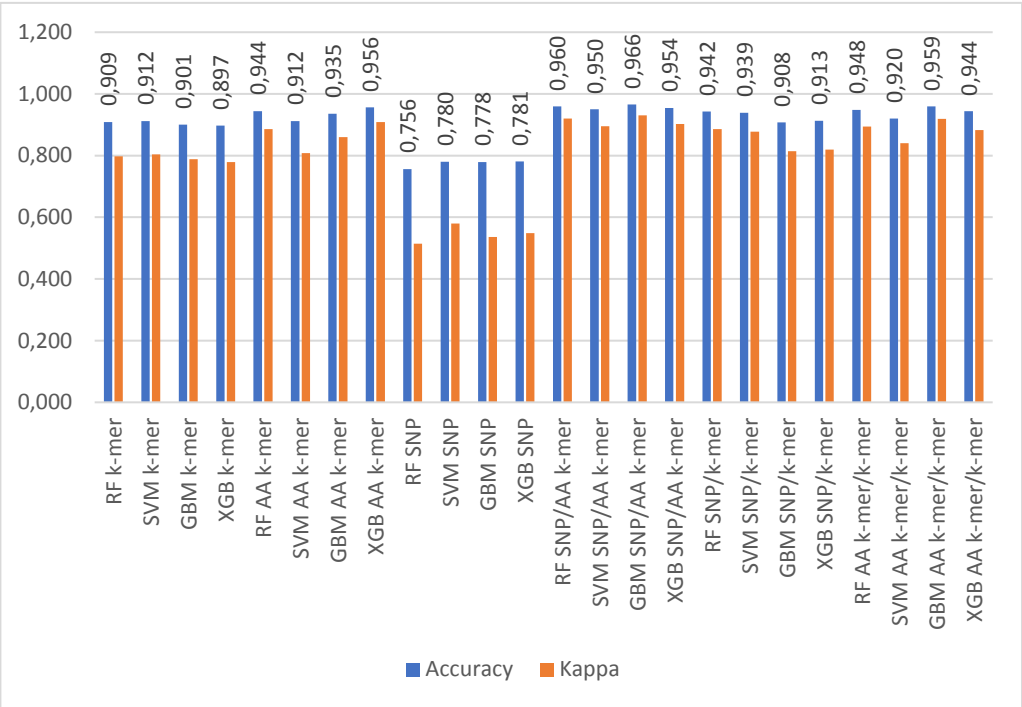

Figure A9. Comparison of Accuracy of all Vancomycin *E.faecium* models.

9 CURRICULUM VITAE