Predicting drug resistance in multiple pathogenic bacteria using different approaches of Multi-task learning

by

Mohammad H. Rezaie

B.Sc., Shahed University, 2019

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Summer 2022

Declaration of Committee

Name:	**Mohammad H. Rezaie**
Degree:	**Master of Science**
Thesis title:	**Predicting drug resistance in multiple pathogenic bacteria using different approaches of Multi-task learning**
Committee:	**Chair:** Mo Chen Assistant Professor, Computing Science
	Maxwell Libbrecht Supervisor Assistant Professor, Computing Science
	Kay C. Wiese Committee Member Professor, Computing Science
	Leonid Chindelevitch Examiner Adjunct Professor, Computing Science

Abstract

Motivation: Drug resistance is becoming an increasingly serious risk to human health around the world. Using techniques that predict drug resistance across different bacterial species that utilize whole-genome sequencing (WGS) data, doctors may administer the appropriate antibiotics to each patient, reducing the chance of drug resistance. Currently available machine learning techniques for this purpose transform whole genome sequence (WGS) data from a specific bacterial isolate into features corresponding to single-nucleotide polymorphisms (SNPs) or short sequence segments of a defined length K (K-mers). We present a novel technique for predicting drug resistance in multiple bacterial species based on gene burden. Our multi-input multi-output network predicts resistance of multiple species to multiple antibiotic drugs.

Results: On a large dataset of isolates from six species, we find that using these strategies yields a statistically significant improvement over state-of-the-art methods, and that this improvement is driven by our method's ability to account for the order of the genes in the genome and jointly training on multiple bacterial species.

Keywords: infectious disease, deep learning, antimicrobial resistance, tuberculosis, next-generation sequencing

Dedication

I'd like to dedicate my work to my family and my beloved friends.

Acknowledgements

This project would not have been possible without the support of many people. Many thanks to my advisers, Dr. Maxwell Libbrecht and Dr. Leonid Chindelevitch who read my numerous revisions and helped make some sense of the confusion.

Thanks to the Simon Fraser University for Completion Fellowship, providing me with the financial means to complete this project. And finally, thanks to my parents, and numerous friends who endured this long process with me, always offering support and love.

Table of Contents

List of Tables

List of Figures

Chapter 1

Multi-task learning without shared features improves prediction of drug resistance in pathogenic bacteria

1.1 Importance of drug resistance

Antibiotics are one of the mainstays of modern medicine. The phenomenon of drug resistance occurs when an infectious organism (also known as a pathogen) develops a mechanism to bypass the action of one or more medications commonly used for treatment [26]. Furthermore, the transfer of resistance genes between humans and non-human animals may have exacerbated the situation [27]. As a result, antibiotic resistance has become a global public health issue, as it can lead to increased disease burden, morbidity, and mortality [11, 13]. This issue has grown in importance given the fact that the rate of generating novel antibiotic medications has dramatically slowed down over the past four decades [30]. Many antibiotics lose their treatment effectiveness against organisms that have developed antibiotic resistance, necessitating the use of third-line medications that are frequently used as a last resort. Drug resistance continues to be a barrier to providing targeted and appropriate treatment. For instance, despite the global scope of the problem, due to a lack of laboratory-based resources, only about half of the countries with a high burden of multi-drug resistant tuberculosis (MDR-TB) have the advantage of innovative diagnostic capabilities [36]. Tuberculosis (TB), which is caused by the bacteria Mycobacterium tuberculosis, is still a serious global public health issue, with over 10.0 million people affected and a projected 1.6 million fatalities in 2017. Drug resistance is becoming more common, posing a severe threat to efficient TB control. Rifampicin (RIF), isoniazid (INH), ethambutol (EMB), and pyrazinamide (PZA) are the four medications used in first-line anti-TB therapy (World Health Organization, 2017). Multidrug-resistant tuberculosis (MDR-TB) bacteria are resistant to at least RIF and INH, and there were more than 550,000 new resistant cases in

2017. Extensive drug resistance (XDR-TB) is defined as resistance to second-line medicines, such as fluoroquinolones [ciprofloxacin (CIP), ofloxacin (OFL), or moxifloxacin (MOX), as well as injectables (INJ) amikacin (AMK), kanamycin (KAN), and capreomycin (CAP)] and injectables (INJ) amikacin (AMK). Traditional TB treatment regimens are lengthy (>6 months) and involve the use of multiple medications at the same time. Drug-resistant tuberculosis requires even longer time and medications with severe side effects and reduced efficacy. Machine learning (ML) methods, try to predict drug resistance by employing models learned directly on coupled WGS and drug susceptibility testing (DST) data [37, 18, 8, 7]. Machine learning (ML) is the process of machines learning without being explicitly programmed. It focuses on making data-driven predictions and has a variety of bioinformatics applications.

Bioinformatics is the study of how to handle biological data using computational and mathematical methods. In recent years, biological data has risen at an exponential rate, resulting in two concerns. Two main concerns are gathering efficient amount of data and also how much meaningful knowledge we can extract from the dataset. The second problem may be overcome using machine learning, which can produce knowledge from heterogeneous data. Deep learning, a machine learning technology, is used to enable feature learning automatically. By merging various features from the dataset, a new collection of features is created. Algorithms can now make complicated predictions on vast datasets using this method. Microarrays, evolution, systems biology, genomics, text mining, and proteomics are just a few of the bioinformatics subfields where machine learning is now being used.

1.2 A large number of approaches have been developed for predicting antibiotic drug resistance

A number of approaches have been developed for predicting antibiotic drug resistance. The most widely-used methods are rule-based methods, rule based machine learning methods are any algorithm which try to identify and learn the "rules" in data. The defining characteristic of a rule-based machine learner is the identification and utilization of a set of relational rules that collectively represent the knowledge captured by the system. This is in contrast to other machine learners that commonly identify a singular model that can be universally applied to any instance in order to make a prediction. [3, 25, 20].

However, rule-based methods have several drawbacks. First, their sensitivity is restricted since they use only a few genetic loci every test, ranging from 1-6 loci per test [33, 21]. They also lack the ability to detect the majority of rare gene variants in the targeted locus, such as promoter variants, deletions, and insertions[9]. They usually rely just on presents or absence of variants and fail to uncover interactions between genome variants [10, 23]. These tests' limitations highlight the need for more comprehensive drug resistance prediction and testing.

1.3 Background on Machine learning and Deep learning

The group of computer algorithms that can be improved by learning patterns of data is known as machine learning (ML); it is also known as artificial intelligence. Machine learning algorithms try to construct a model using training data to achieve predictions. Machine learning algorithms have variety of applications, such as automatic driving, face recognition, drug resistance prediction in which applications, traditional computer algorithms will not work. One of the related field of studies is data mining which tries to explore the data. Deep learning is a kind of machine learning technology that learns patterns in data using artificial neural networks. Three forms of learning are supervised, unsupervised and semi-supervised learning. Deep learning architectures, such as deep belief networks, deep neural networks, deep reinforcement learning, recurrent neural networks (RNN), and convolutional network are various type of deep learning algorithms which can be used in various types of applications such as image recognition, computer vision, and etc.

1.4 Machine learning-based methods can learn more accurate rules with complicated interactions between variations.

The wide-n-deep neural network (WnD) [3], and DeepAMR [37], are examples of contemporary state-of-the-art models. WnD project uses 3,601 strains of *Mycobacterium Tuberculosis (TB)* and deep neural network architecture to predict phenotypic drug resistance to 10 anti-TB drugs. DeepAMR [37] also used 8388 TB isolates from 16 countries on six continents to develop an end-to-end multi-task model with deep denoising auto-encoder (DeepAMR) for multiple drug classification. Moreover they used conventional ML models such as Support vector machine (SVM) [37, 24], Logistic Regression (LR) [3, 17], Random forest (RF) [29, 3] in their works as base-line models. Machine learning methods have been used in the past to help in digital X-ray analysis, drug development, and assessing anti-TB characteristics of substances in the context of tuberculosis. Researchers have looked into using random forest classification and GBT models to predict pathogen drug resistance. In the case of tuberculosis, instead of using a single statistical model for all medications, several statistical models were used to different drugs within the same study.

1.5 Although machine learning models have great performance, present machine learning approaches have several flaws.

While reviewing a wide range of research in this field, we discovered some flaws that we address in this thesis: They are all based on SNPs or k-mers, and they are all trained on only

one species. Training models based on SNPs or k-mers includes a number of features that contribute to overfitting, and as a result, studies are unable to train complicated models. Existing techniques are also trained separately for each species.

Chapter 2

Data Pipeline and GEM-DR

We acquired our data from Pathosystems Resource Integration Center (PATRIC) that provided full genome sequencing of our isolates [12] for researchers. We collected and analysed for this study, a total of 21,407 isolates from six different species (*Streptococcus pneumoniae*: 5805 isolates, *Staphylococcus aureus*: 1989 isolates,*Salmonella enterica*: 2218 isolates, *Acinetobacter baumannii*: 1415 isolates, *Escherichia coli (E. coli*): 2020 isolates, *Mycobacterium Tuberculosis (TB*): 7960 isolates). The PATRIC website has been used to retrieve all of the aforementioned isolates' short reads including whole-genome sequencing and their associated pehontypes. [6]. Different species had various proportion of susceptible isolates compered to Resistance isolate, varying from 13.0% to 92.0% (*Streptococcus pneumoniae*: 23.0% to 92.0% , *Staphylococcus aureus*: 27.0% to 88.0%, *Salmonella enterica*: 47.0% to 91.0%, *Acinetobacter baumannii*: 13.0% to 55.0% , *Escherichia coli (E. coli)*: 27.0% to 88%, *Tuberculosis (TB)*: 55% to 87%)

2.1 Pre-processing

2.1.1 SNPs Dataset

To get SNP information from our downloaded dataset, we first had to remove those samples with missing phenotype. After that, we mapped the reference genome to each phenotype's raw dataset. The National Center for Biotechnology has provided all of the reference genomes (NCBI). We also use the Snippy tool to identify the variants.

Snippy was run on each species independently, and SNPs were called using the reference genome. We utilized .csv file to collect our SNPs as an outcome of the various outputs. As a result, each read has its own .csv file. We combined all of the.csv files into a single table that contained all of the readings as well as each read's SNPs.

As a result, we acquired six datasets for six species. Each dataset comprises the number of isolates multiplied by the number of unique SNPs variants in all readings). For example, we have a dataset matrix that includes 2020 isolates (rows) and 314,562 (columns) which is number of all unique SNPs for E.coli. The advantage of utilising gene burden features

rather than SNP-based features is that gene burden data decreases the number of features significantly, alleviating the "curse of dimensionality." When the number of available isolates is limited in comparison to the number of SNPs, utilising gene load data may result in more accurate models with less overfitting. When considerably higher sample numbers are available, the gene burden-based approach may lose this benefit.

2.1.2 Gene burden features

We developed a gene burden dataset after compiling SNP datasets for each species. In each isolate, we counted the number of SNPs for each gene. Using the reference genome, there were at least 10 SNPs in one isolate. in 3,585 genes in *Acinetobacter baumannii*, 4,140 genes in E.coli, 4569 genes in *Salmonella enterica*, 2,659 genes in *Staphylococcus aureus*, 2,043 genes in *Streptococcus pneumoniae*, and 3,960 genes in *TB*. Each species is represented as a feature matrix with a particular dataset, with every row representing one isolate and each column representing counted number of SNPs in one gene. As a result, each cell indicates the number of variations that occurred inside a single gene in a particular isolate.

2.1.3 Defining labels for data

Each isolate from a given species was labelled using available drug labels from the same species. All of the isolates have labels for just one species, and no drug resistance labels for the other species are given. Among all of the species, only a few contain all of the drug labels, and the majority of the isolates are lacking part of the labels.

Acinetobacter			E.coli		
Drug	# of Resistance	Total #	Drug	# of Resistance	Total#
Ciprofloxacin	945	1076	Amikacin	95	836
Gentamicin	913	1052	Amoxicillin	597	989
Imipenem	572	1052	Ampicillin	739	1005
Tobramycin	635	966	Aztreonam	148	531
Amikacin	529	966	Cefalotin	195	366
Ceftazidime	746	803	Cefalotin	179	826
Ceftriaxone	623	780	Cefotaxime	258	1426
Levofloxacin	661	857	Cefoxitin	97	477
Aztreonam	723	763	Ceftazidime	197	1393
Cefotaxime	698	728	Ceftriaxone	84	99
Meropenem	334	543	Cefuroxime	210	1282
			Ciprofloxacin	320	1392
			Ertapenem	53	461
			Gentamicin	149	1392
			Meropenem	33	479
Salmonella			Staphylococcus		
Drug	# of Resistance	Total#	Drug	# of Resistance	Total#
ampicillin	820	2053	Gentamicin	146	1256
streptomycin	842	2031	Erythromycin	430	1255
tetracycline	1087	2030	Methicillin	689	1481
chloramphenicol	341	2026	Fusidic	74	968
ciprofloxacin	13	1972	Penicillin	876	1016
gentamicin	190	1893	Tetracycline	179	1097
sulfisoxazole	597	1774			
ceftriaxone	348	1769	Ciprofloxacin	423	1095
ceftiofur1	342	1760	Rifampin	14	1006
cefoxitin	281	1733	Clindamycin	317	656
trimethoprim	34	257	Trimethoprim	13	467
kanamycin	41	704	Oxacillin	27	168
Streptococcus			TB		
Drug	# of Resistance	Total #	Drug	# of Resistance	Total#
Chloramphenicol	220	2097	Amikacin	573	2033
Clindamycin	38	428	Capreomycin	552	1991
Cotrimoxazole	878	1938	Ciprofloxacin	37	443
Erythromycin	912	2514	Ethambutol	1407	6096
Penicillin	1158	2299	Ethionamide	498	1516
Tetracycline	785	2079	Isoniazid	3445	7734
Trimethoprim	1980	2561	Kanamycin	697	2436
			Moxifloxacin	129	961
			Ofloxacin	800	2911
			Pyrazinamide	754	3858
			Rifampicin	2968	7715
			Streptomycin	2104	5125

Table 2.1: Summary of the number of isolates and the label distribution in our data.

In this work we tried to share the resistance profile across species and also use multiple datasets for training our network. Since for each species we have not enough data to efficiently train a neural network, using 6 different datasets help us to solve this challenge. We implemented gene burden features (that we previously applied to TB [31]). After gathering SNP datasets for each species, we created a gene burden dataset. We counted the number of SNPs in each gene for each isolate. The gene structure is what we wanted to include in this study. Gene burden features decrease the number of factors and increase generalisation. In this work we trained our model across multiple species. We utilise masked loss to do this, which allows us to use isolates that lack certain drug resistance response markers. We can accomplish this despite the fact that we don't have any features in common with other species. It turns out that simply sharing logic is beneficial.

2.2 Method

2.2.1 Addressing missing labels in large dataset

The absence of labelling was one of the most significant problems we faced while working on this project.

We came up a lot of isolates that did not have labels for some of the medicines they were connected with. We utilized a masked loss function instead of a regular loss function to deal with this problem and get the most out of the data. This loss function was implemented to ignore and not compute the loss for missing labels. As a result, we have been informed that missing labels have no bearing on the network's weights. The challenge of using the multi-task model in our dataset is that many isolates lack labels for some of the drugs. For this reason we replace the usual loss function with a masked loss function. The masked loss function ignores the missing labels in calculating the loss, and therefore, these missing labels do not affect the network weights. We use the binary cross-entropy as the loss function. Specifically, we use the loss function

$$\text{Loss } = \sum_{i=1}^{I} \sum_{d=1}^{D} \mathbb{1}(X_{i,d} \text{ is available}) H(X_{i,d}, Y_{i,d}) \tag{2.1}$$

where I and D are the numbers of isolates and drugs respectively, $X_{i,d}$ and $Y_{i,d}$ are the true and predicted resistance values, H is the binary cross-entropy function, and $\mathbb{1}$ is the indicator function.

2.2.2 GEM-DR Input and output

In GEM-DR , there are six separate input sections, each of which is assigned to a different species. Furthermore, each input section contains a set number of nodes equal to the species' number of genes. Throughout this procedure, GEM-DR makes use of all six species

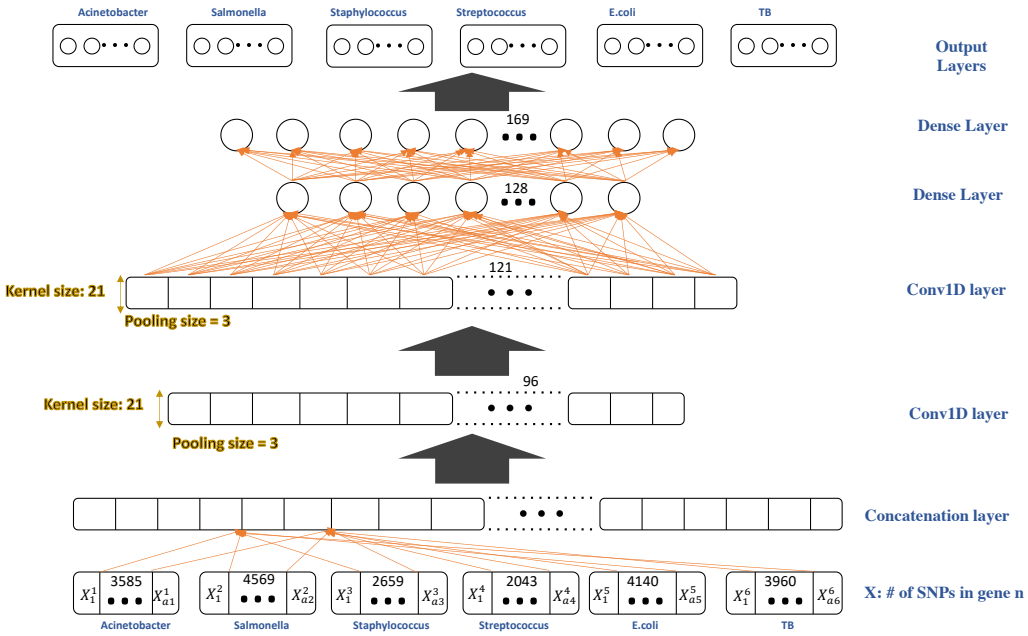

Figure 2.1: Model architecture of GEM-DR.

gene burden base features for learning. The GEM-DR output layers have six separate parts of input, each of which includes multiple drug resistance prediction nodes. Streptococcus pneumoniae input layer section had 2043 neurons, *Staphylococcus aureus* input layer section had 2659 neurons, *Salmonella enterica* input layer section had 4569 neurons, *Acinetobacter baumannii* input layer section had 3585 neurons, *E. coli* input layer section had 4140 neurons, and *Tuberculosis* input layer section had 7960 neurons. As a result, a total of 20,956 neurons were used in six separate section in GEM-DR input layer.

2.2.3 Model Architecture

Furthermore, as the following hidden layer, we have merged all of the input layers together, followed by a batch normalisation layer. All of the species' resistance profiles are shared in this hidden layer, which aids our model in learning drug resistance patterns in diverse species. As the following layer, we utilized the reshape layer to prepare the data and input it into two CNN layers. We utilized 96 filters, a kernel size of 21, two strides, the same padding, and a random uniform as the kernel initializer for the first CNN layer, which was followed by a relu activation layer. We use CNN layers to take into account all of the interactions between gene neighbours. After each CNN layer, we utilized average pooling as well. Except for the number of filters, which is 121, we utilized the identical settings for the second CNN layer. The information was then fed into Dense layers using the flatten layer. We utilized two Dense layers, each with 128 and 169 neurons. Kernel and bias initializers

Hyperparameter	Optimal Value
CNN: Number of Layers	96
CNN: Kernel Size	21
CNN: Number of Filters	121
CNN: Pooling Size	2
Dense: Number Layers	128
Dense: Number Units	169
Optimizer	SGD
Learning rate	0.394

Table 2.2: Optimized hyper-parameters that has been used in GEM-DR

for dense layers are random uniform and Truncated normal. There is a relu activation layer in both dense layers.

We used the binary cross-entropy loss function since our job was multi-task classification and each label had a binary value of being resistant (one) or susceptible (zero):

$$LOSS = \sum_{i=1}^{I} \sum_{d=1}^{D} 1(X_{i,d} \text{ is present}) H(X_{i,d}, Y_{i,d}) \tag{2.2}$$

2.2.4 Bayesian hyper-parameter optimization

To choose the optimal values for several hyperparameters involved in building the model, Bayesian optimization approach was applied. In the process of Bayesian optimization, a probabilistic model is formed to map combinations of hyperparameters (H) to probability values of achieved score on the objective function (AUC-ROC) [32]. For this purpose, a surrogate function (Gaussian Process Regression) was used to compute $P(\text{AUC-ROC}|H)$ and maximize the objective function in an iterative manner. Hyperparameter optimization was implemented through the Scikit-optimize Python package [14]. In this procedure we used 10-fold cross validation in order to cross validate our data. We split data into %70 train data, %20 test data and %10 validation data. Also for validation of our method we used optimized base-line models which have been optimized with same fractions of data. The difference between optimizing our original model and othe base-line models is we could cross-validate GEM-DR with all 6 species at once yet for the other base-line models we should cross validate each model with one species at a time. Table 2.2 presents chosen optimal hyperparameters that were ultimately used to build the final model.

2.2.5 Evaluation

We compared our GEM-DR model against a wide number of current approaches that are considered to be state-of-the-art. We didn't include catalog techniques or other methods

that have been shown to perform considerably worse than the ones described here. For the comparison, the following approaches were used: (1) WnD (Wide-n-Deep neural network) [4], This method is optimized for predicting drug resistance in TB. (2) SVM (Support Vector Machines)[24] are a widely used ML model for binary classification, with applications in many areas. (3) LR (Logistic regression) [17] using this method recent studies have demonstrated that it is superior in predicting treatment resistance in tuberculosis. (4) Gaussian Naive Bayes(GNB) [19, 39] since it has been widely used in a variety of bioinformatics applications. (5) Random Forest (RF) [29], in bioinformatics, RF has a variety of applications for dealing with large data and whole genome sequences. In this work for the aim of comparing our model performance with the state-of-the-art models we have implemented and trained the widendeep neural network , support vector machine (SVM), logestic regression (LR), and also Random forest (RF) [38, 17]. As we mentioned in the optimization section, we used same procedure and dataset for training and testing the base-line models. We also ran the Bayesian hyperparameter Optimization to choose the best parameters for all the models. We used the following parameters for each model. logistic regression: $C = 0.1$ and f_2 penalty. Random forest: 120 estimators, a minimum sample split of 4, and a maximum depth of 30. Support vector machine: C = 0.1 and linear kernel have been used. WnD model we used five layers, 490, 529,571,617,666 nodes have been used respectively for the five layers. Also each layer has been followed by a dropout layer with 0.12 rate. All of the layers have L1 kernel regularizer as well. To evaluate the accuracy of our predictions, we used the AUC-ROC and AUC-PR values. The AUC-ROC metric is the area under the plot of the true positive rate (TPR) against the false positive rate (FPR) at different classification thresholds. The AUC-PR is similar to AUC-ROC, but the plot is that of precision ($\frac{TP}{TP+FP}$) against recall ($\frac{TP}{TP+FN}$) at different classification thresholds.

2.2.6 Implementation

For training we used Python 3.7 for implementing our neural network and all of the state-of-the-arts methods as well. We utilized Keras [5] and Tensorflow on top of that. We also used Scikit-learn [28] library to implement other machine learning methods.

2.3 Results

2.3.1 Does GEM-DR outperform state-of-the-art ML models in most of the drug labels prediction?

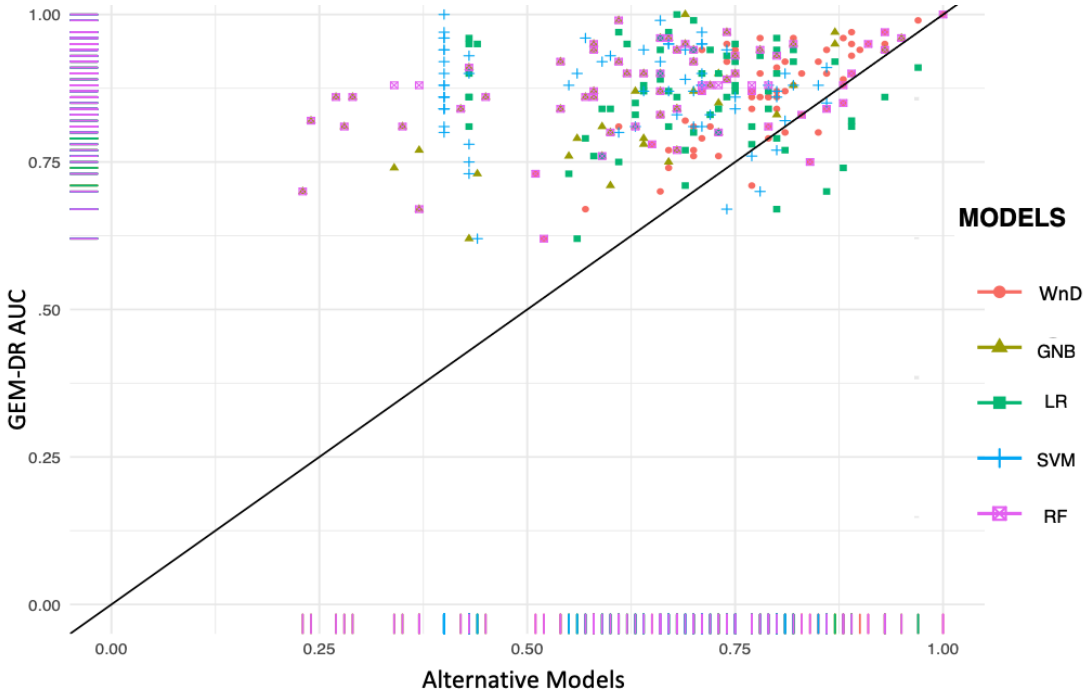

Figure 2.2: Overall performance comparison between GEM-DR and state-of-the-art ML models (WnD [4], SVM [24], LR [17], GNB [19, 39], and RF [29]); The GEM-DR AUC is compared to the AUC of the other baseline ML models in terms of overall performance. Each shape reflects drug resistance prediction AUC performance of GEM-DR in the X axis and the other model AUC performance in the Y axis. If GEM-DR outperforms the baseline models in that particular drug, the shape placement is above the $Y = X$ line, but if the baseline model exceeds GEM-DR in AUC-ROC performance, the shape placement is below the $Y = X$ line.

2.3.2 GEM-DRoutperformed state-of-the-art baseline models

In this work we assessed the model by putting in comparison with the state-of-the-art models which have been achieved great results as our best knowledge so far.

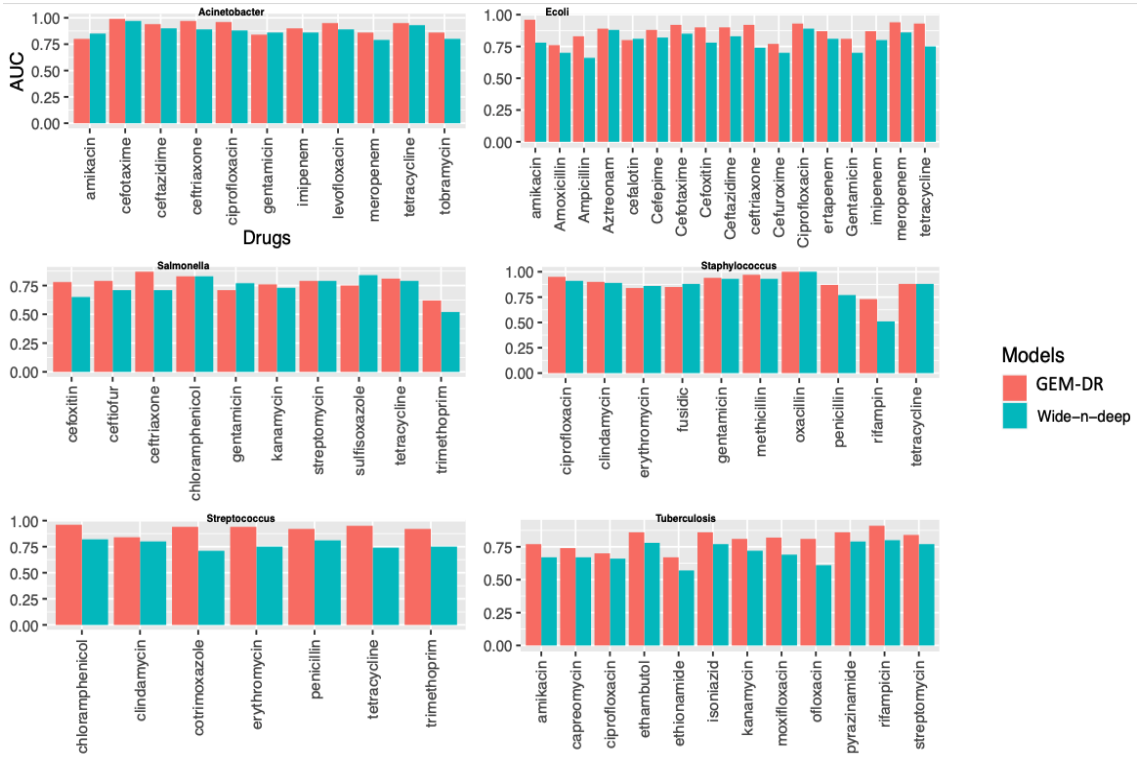

Figure 2.3: The outcomes of GEM-DR and WnD performance in predicting drug resistance are shown in these graphs. In comparison to feeding all the species at once to GEM-DR, we tried training WnD on each species.

To train and test these baseline models, we used the same data that we used to train and test our model. We utilized WnD, SVM, RF, and LR for comparison and assessment of our work outcomes, using the AUC-ROC measure for this comparison.

GEM-DR outperformed in 59 drugs out of 68 drugs. In most cases, GEM-DR produces significantly better results than the WnD approach (Figure 2). For 9 of the 11 drugs tested in *Acinetobacter baumannii*, GEM-DR outperformed the Wide and Deep Model. Wnd model, Amikacin and Gentamicin performed better in terms of AUC . We got improved AUC ratings in E.coli for 15 of the 17 medications; Cefalotin and Ertapenem did slightly better with the WnD model. Our approach outperformed the WnD model in *Salmonella enterica* for 8 out of 10 medications, including chloramphenicol and sulfisoxazole. Moreover, our perdition for 8 out of 10 drugs in *Staphylococcus aureus*, was better. Using WnD model Erythromycin, Fusidic had better predicted results. In *Streptococcus pneumoniae* we outperformed WnD model in all of the drugs. Finally, our model performed substantially better than WnD model in all of the drugs.

2.3.3 Training multiple species together comes with better prediction performance

Further, we determined the reason for the AUC-ROC improvement to be sharing drug resistance profiles across species. After co-training with other species, GEM-DR average performance increased for 5 out of 6 species. We first trained GEM-DR with one species at a time to show the effect of co-training species and drug resistance profile sharing, then compared our findings to the GEM-DR results after co-training six species together, and presented the average results. After adding other species to *Salmonella enterica*, GEM-DR performance did not increase (Fig 3.4).

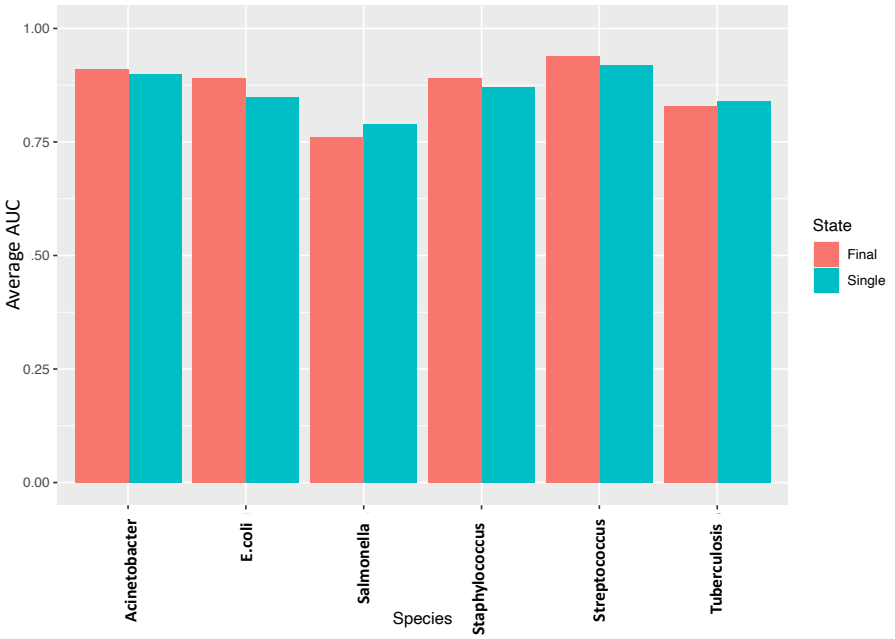

Figure 2.4: We initially trained GEM-DR with one species at a time to illustrate the effect of co-training species and drug resistance profile sharing among them, and then compared those findings with the GEM-DR results after co-training 6 species together, and displayed the average results in this figure.

Which co-training species partner specifically helps the improvement of training procedure?

To select the best co-training partner for each species, we ran an experiment in which we co-trained every two species together and compared the results to solo training performance results. We keep track of the AUC-ROC scores of joint species training to learn more about the influence of sharing several species' resistance profiles on model performance. We hypothesized that using a multi-task model could improve the accuracy of our model, especially for drugs with relatively limited labeled data available, driven by shared mechanisms of resistance. If true, this hypothesis would imply that the patterns learned from one drug can compensate for the lack of training data for another drug. We observed that using the multi-task model improved the performance of the gene burden-based GEM-DR. Furthermore, we see that the gene burden-based GEM-DR trained on a multiple-species displays good performance and can accurately predict drug resistance to that specific drug. *Acinetobacter baumannii* improves performance the most when combined with *Streptococcus pneumoniae*(Fig 5). Many species have improved the *E. coli* drug resistance prediction profile, with *Acinetobacter baumannii* being the best combination. When co-trained with *Streptococcus pneumoniae* and *Acinetobacter baumannii*, *Salmonella enterica* outperformed solo training.

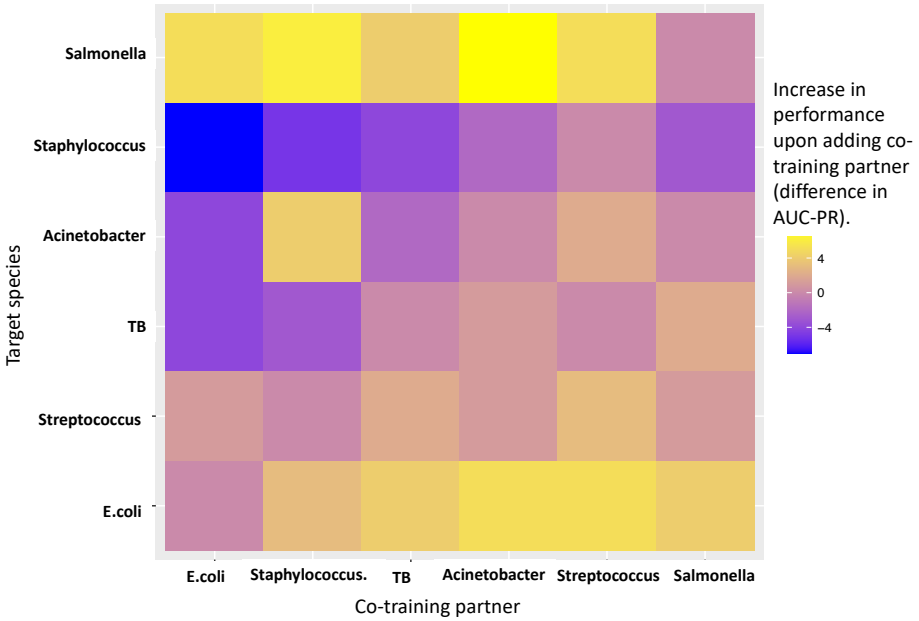

Figure 2.5: We conducted an experiment in which we co-trained every two species together and compared the outcomes to solo training performance results in order to determine the optimum co-training partner for each species.

Staphylococcus aureus was a suitable training partner for *Staphylococcus aureus*. Pair training with *Staphylococcus aureus* enhanced the AUC-ROC performance of *Streptococcus pneumoniae*, and lastly, combined training of *TB* with *Acinetobacter baumannii* and *Salmonella enterica* improved the performance of GEM-DR. As a result, we can conclude that exchanging resistance profiles between species and training will enhance prediction (Fig 3.5).

Chapter 3

Drug resistance prediction in E.coli using pathways-based approach

3.1 Background

Escherichia coli, also known as E. coli is a Gram-negative, facultative anaerobic, rod-shaped, coliform bacterium of the genus Escherichia that is commonly found in the lower intestine of warm-blooded organisms. Most E. coli strains are harmless, but some serotypes (EPEC, ETEC etc.) can cause serious food poisoning in their hosts, and are occasionally responsible for food contamination incidents that prompt product recalls. The harmless strains are part of the normal microbiota of the gut, and can benefit their hosts by producing vitamin K2, and preventing colonisation of the intestine with pathogenic bacteria, having a mutualistic relationship. E. coli is expelled into the environment within fecal matter. The bacterium grows massively in fresh fecal matter under aerobic conditions for three days, but its numbers decline slowly afterwards.

E. coli and other facultative anaerobes constitute about 0.1% of gut microbiota, and fecal–oral transmission is the major route through which pathogenic strains of the bacterium cause disease. Cells are able to survive outside the body for a limited amount of time, which makes them potential indicator organisms to test environmental samples for fecal contamination. A growing body of research, though, has examined environmentally persistent E. coli which can survive for many days and grow outside a host.

The bacterium can be grown and cultured easily and inexpensively in a laboratory setting, and has been intensively investigated for over 60 years. E. coli is a chemoheterotroph whose chemically defined medium must include a source of carbon and energy. E. coli is the most widely studied prokaryotic model organism, and an important species in the fields of biotechnology and microbiology, where it has served as the host organism for the majority of work with recombinant DNA. Under favorable conditions, it takes as little as 20 minutes to reproduce.

The majority of E. coli are harmless and even beneficial for the digestive system. However, some of them can cause serious food poisoning and diarrhea, not to mention enormous economic losses due to contaminated food that came into contact with E. coli [35, 34]. E. coli is normally used as an indicator for anti-microbial resistance in animal and meat industry . E. coli can transfer their DNA using bacterial conjugation or transduction, which means they can spread horizontally across an existing population. One type of E. coli that leads to diarrhea is the Shiga toxin-producing E. coli (STEC). The transduction of STEC can be used to produce Escherichia coli O157:H7 (E. coli O157:H7). E. coli O157:H7–the Shiga toxin-producing strain of E. coli– has been identified for the first time as a human pathogen in 1982 [2, 22]. E. coli O157:H7 is responsible for kidney failure and hemorrhagic diarrhea which can be deadly in children younger than five years old [15]. E. coli O157:H7, has been estimated to cause trouble for three to eight of every 100,000 people. The main source of bacteremia in England was E. coli in 2011, for which an incidence of 50.7 cases per 100,000 population [1].
Given the availability of DNA sequenced data from PATRIC database from E. coli, we developed three machine learning models for 2020 isolates gathered from PATRIC online database to classify E. coli. resistance against 17 drugs. Further, we proposed a pathways-based multi-task learning method to use all the information about genes in E.coli, comparing its results and the other ML methods we found that due top the lack of data in this area we cannot use just this information to train our network and it does not improve the results of drug resistance prediction in E. coli.

3.2 Materials and Methods

3.2.1 Data and Data Prepossessing

The Pathosystems Resource Integration Center (PATRIC) is the all-bacterial Bioinformatics Resource Center (BRC) (http://www.patricbrc.org). This impressive work has been made by the cumulative effort of two institutions of the original National Institute of Allergy and Infectious Diseases-funded BRCs. They have provided variety type of bioinformatics data for researchers [e.g. transcriptomics,three-dimensional protein structures and sequence typing data, and protein–protein interactions (PPIs)]. This source currently includes more than 100,00 sequenced genomes.Moreover, these genomes has been annotated Strongly with the Rapid Annotations using Subsystems Technology (RAST). Further, There are also Multiple versions of annotations provided for sequenced genomes. One of the most important features of PATRIC is Availability of online comparison between various annotations. For the aim of comparative transcriptomic analysis, researchers can use both private and public data in the same time using provided online tools on PATRIC. All of these data and associated visualizations tools are freely available online. All of the data which have been used in this work are based on resources available on PATRIC 2020 webasite version. Using PATRIC

FTP server, we have downloaded over 2000 E. coli sequenced genomes associated with E. coli and their annotations. Furthermore, we have used snippy tool. Using snippy tool we have extracted SNPs from all of the sequenced genomes we have gathered before. Associated annotation table contains variety of annotations such as taxon id, the antibiotic which has been tested for reaction of isolate to it, resistant phenotype, MIC measurement value, etc. We have used clinical breakpoints of EUCAST (European Committee on Antimicrobial Susceptibility Testing) v. 10.0 (2020). Some of the isolates do not have resistant phenotype and instead they have measured value of MICs. Hence, we have used EUCAST clinical breakpoints table to identify their resistance phenotype.

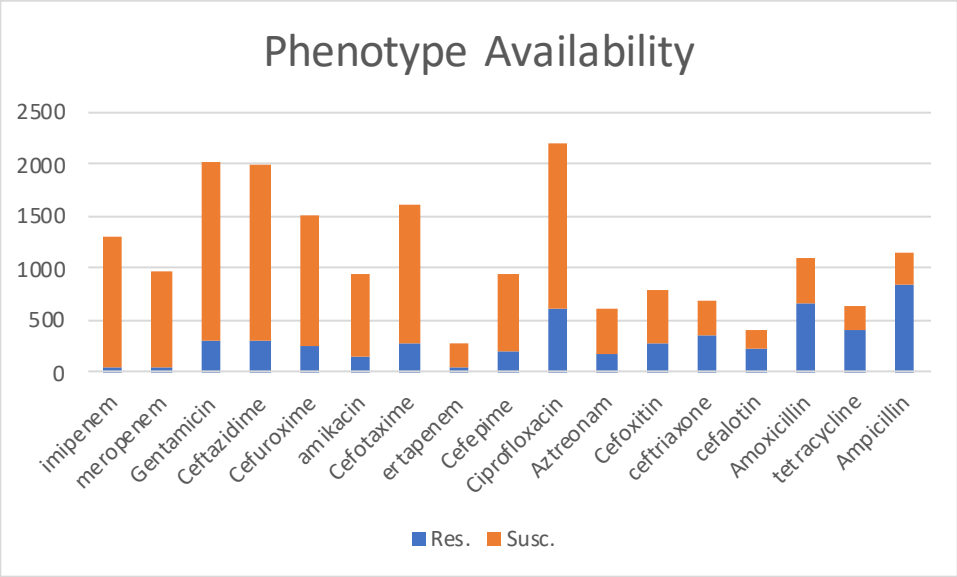

Figure 3.1: Availability of 17 drugs which have been used in this work.

Our study includes 2020 sequenced genes of E. coli isolates. As it is shown in Fig.1 we have Used 17 drugs resistance testing labels for our work. Of the 2020 isolates were have been tested against 17 drugs, Gentamicin has the biggest number of labels available with the number of 2020. Further, Ampicillin has the biggest number of resistance isolates against it with number of 849 of 1144 and Imipenem has the biggest number of Susceptible isolates against it with the number of 1257.

3.2.2 Methods

Our study includes 2020 sequenced isolates. It worth to mentioned that not every isolate was tested against all drugs. We performed three different machine learning models in order to explore the variety of structures within the genomics variations in our isolates. We have applied Logistic Regression, Random Forest, and Support Vector Machine with L2 regularisation (SVM) models in our work.

While the evaluation procedure we have considered the same distribution for each set of evaluation of each drug which means we tried to hold the same portion of resistance isolates with susceptible ones in terms of individual drugs. As we mentioned before our feature set includes all of the SNPs which have been extracted form sequenced data with snippy tool. This feature set contains 956000 SNPs which have been occurred in 2020 isolates.

3.2.3 Pathways-based method

We proposed a pathways based network to use all the available information exist in E.coli. In the input all the genes associated with specific pathway are connected to one neuron in the following hidden layer and the input layer is not fully connected with the first hidden layer. After that the neurons in the first hidden layer are fully-connected to the next shared layer. and then we have 17 neurons representing 17 drugs labels in E. coli.

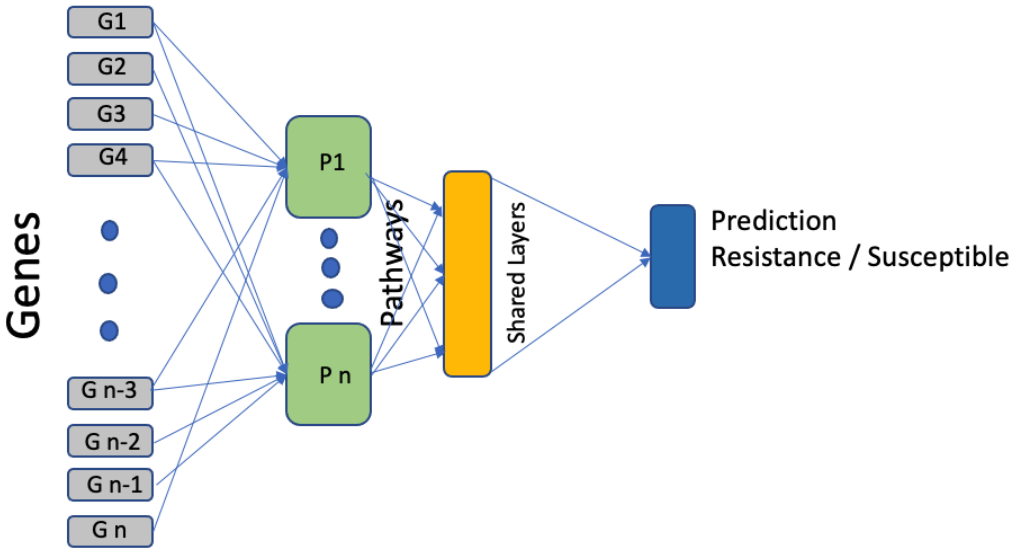

Figure 3.2: Multi-task learning using pathways background in E. coli genome

3.2.4 Architecture

368 In this study, there are total of 4140 separate input sections regarding the number of genes we have in E. coli, each of which is assigned to a genes with a different number SNPs associated with. Furthermore, each input section contains a number of SNPs occured in that. Throughout this procedure, this model makes use of all of 368 pathwyas based features for learning. This model output layers have 17 separate parts, each of which indicates the resistance/susceptibility of the specific drug to E. coli(Amikacin, Amoxicillin, Ampicillin, Aztreonam, Cefalotin, Cefepime, Cefotaxime, Cefoxitin, Ceftazidime, Ceftriaxone, Cefuroxime, Ciprofloxacin, Ertapenem, Gentamicin, Imipenem, Meropenem, Tetracycline).

Furthermore, as the following hidden layer after input section, we have 368 nodes each of them associated to a specific pathway. Each of the nodes in this layer is connected to the input nodes which are associated to that specific pathway. This layer is followed by a batch normalisation layer. After this layer all the information is merged in the following concatenation layer. The information was then fed into Dense layers using the flatten layer. We utilized two Dense layers, each with 256 and 128 neurons. Kernel and bias initializers for dense layers are random uniform and Truncated normal. There is a relu activation layer in both dense layers.

We used the binary cross-entropy loss function since our job was multi-task classification and each label had a binary value of being resistant (one) or susceptible (zero):

$$LOSS = \sum_{i=1}^{I} \sum_{d=1}^{D} 1(X_{i,d} \text{ is present}) H(X_{i,d}, Y_{i,d}) \tag{3.1}$$

For the aim of training and optimization, we used 10-fold cross validation in order to corss validate our data. We split data into 70% train data, 20% test data and 10% validation data. Also for validation of our method we used optimized base-line models which have been optimized with same fractions of data.

3.3 Results

After applying Logistic Regression we have the best AUC for Imipenem with 100%, and the lowest AUC is for Amoxicillin with 68%. This result has been slightly improved in (B) Random Forest with the lowest amount of AUC for Amoxicillin with 70%. And lowest results was for (c) SVM with 66% AUC for Amoxicillin. Hence, Random Forest had the overall performance in predicting susceptibility/ resistance against 17 drugs. We have separated our data into three exclusive sections (training, testing, validation) data with 70% , 20%, 10% portions accordingly. It is worth to mentioning that we have not used validation set in our either training, testing procedure.

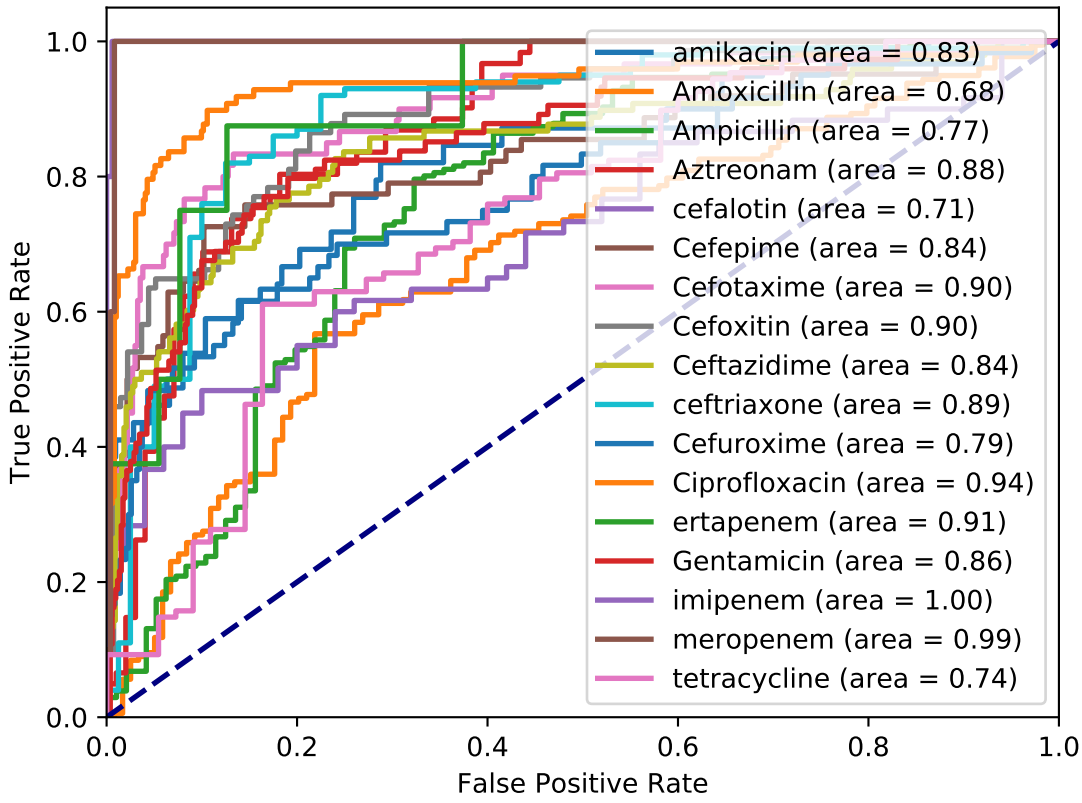

Figure 3.3: Logistic Regression

Using Logistic regression we achieved best prediction result with Imipenem, Meropenem, and Ciprofloxacin respectively with 100%, 99% and 94% AUC-ROC.

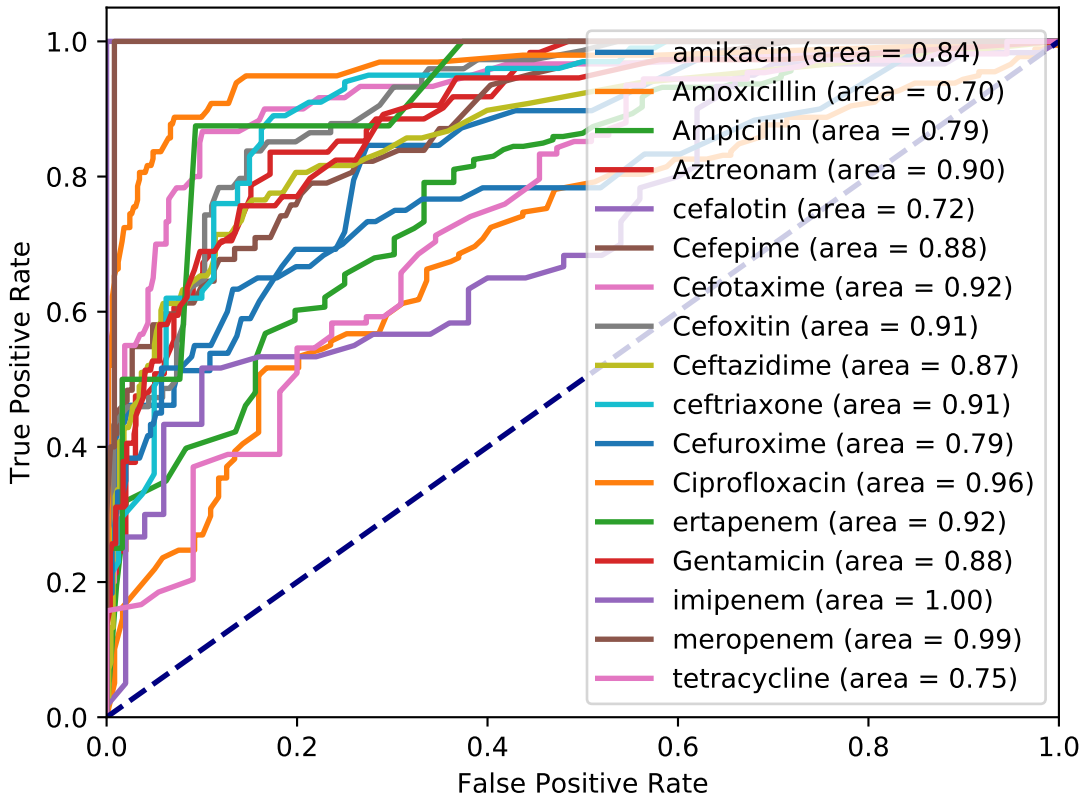

Figure 3.4: Random Forest

Using Random Forres we achieved best prediction result with Imipenem, Meropenem, and Ciprofloxacin respectively with 100%, 99% and 96% AUC-ROC.

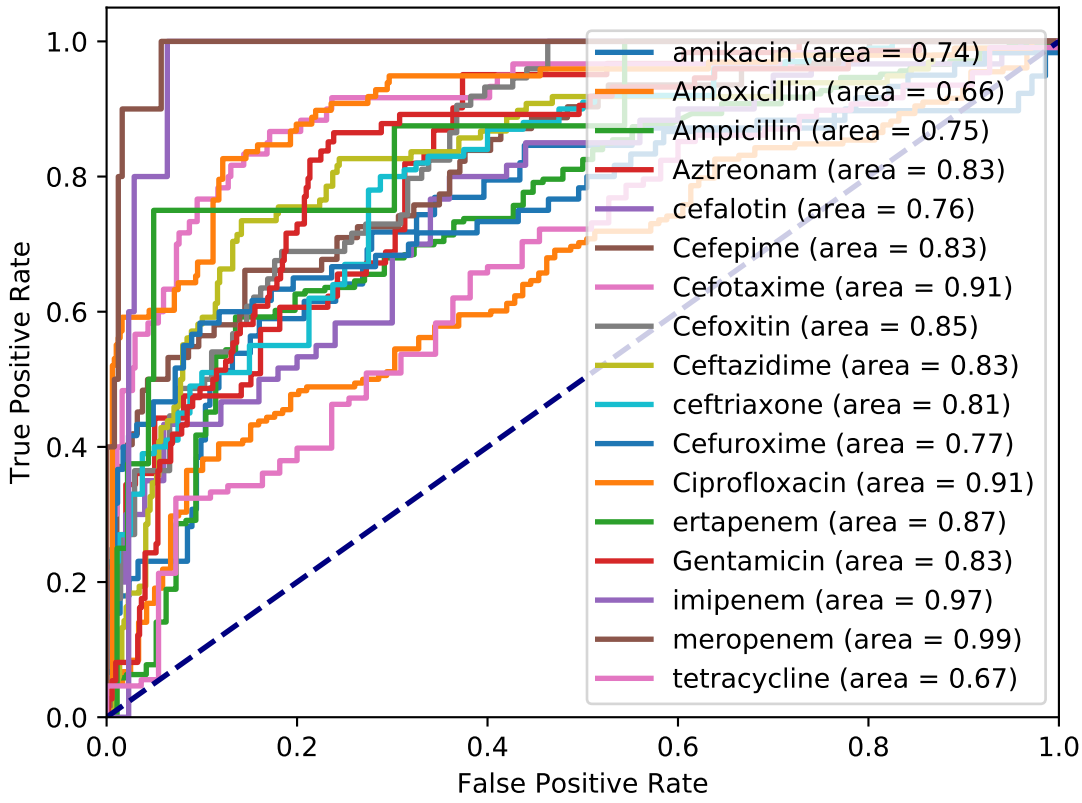

Figure 3.5: Support Vector Machine with L2 regularisation

Using Support Vector machine (SVM) we achieved best prediction result with Imipenem, Meropenem, and Ciprofloxacin respectively with 99%, 97% and 91% AUC-ROC.

3.4 Drug resistance prediction in E. coli using pathway information

In this section we used all the known pathways background which have been gathered from EcoCyc [16]. In this model all the genomes which are associated to a pathways are connected to just one neuron in the first hidden layer, which represents all the pathways in our datasets. This hidden layers is followed bye two more hidden layer which are fully connected.

We have trained and tested our pathway-based model with over 2020 E. coli isolates which have been gathered form PATRIC database.

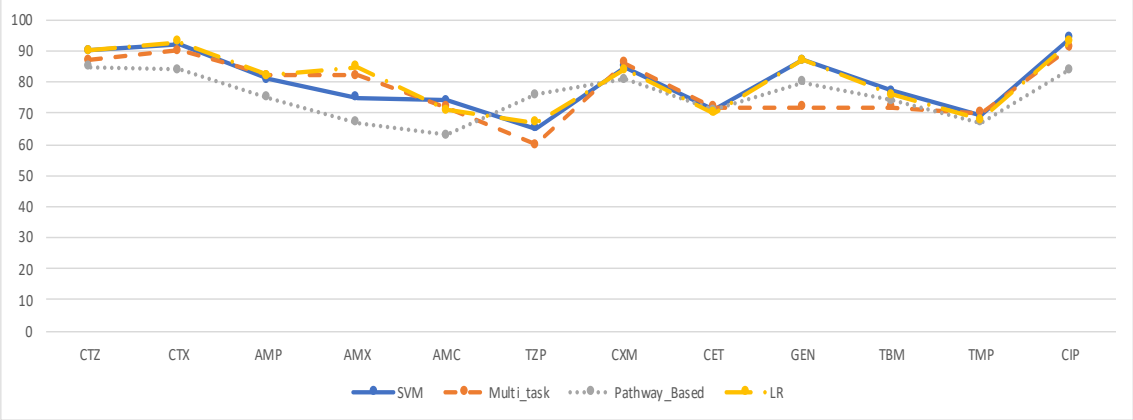

Figure 3.6: Overall comparison of drug resistance prediction in E. coli between pathways based mode, fully connected feedforward model, Logistic regression, and Support vectore machine

Using this comparison we found that since there is not enough of knowledge in pathways available yet, we can not use this approach as the main approach for predicting dug resistance in E. coli. In conclusion we found that the baseline models perform similarly or better than the pathways based model. In CTZ, CXM, CET, TBM, and TMP all the tested models have been performed similarly. We had a better results in TZP with pathways absed model comparing to baseline models. Baseline models ahve been also peformed similarly in CTZ, CTX, AMP, AMC, TZP, CXM, CET, TMP, and CIP.

Chapter 4

Conclusion

We suggest a unique application of GEM-DR, a deep learning architecture, for predicting drug resistance in six distinct species in this study (*Acinetobacter baumannii*, *Salmonella enterica*, *Staphylococcus aureus*, *Streptococcus pneumoniae*, and *TB*). For each of the informative species, we additionally presented 6 Gene burden base feature data sets. This design, which combines CNN and fully linked hidden layers, has the benefit of taking into account both gene mutation load and gene neighbouring. In a number of situations, GEM-DR surpasses existing state-of-the-art techniques. We discovered that when we utilise the gene burden, which is defined as the amount of mutations in each gene, rather than the complete SNP data, GEM-DR performs best. The gene burden allows us to examine the relevance of individual genes, while the model architecture allows to learn about how they are arranged in the genome. GEM-DR exceeded other state-of-the-art models in solo species training by introducing a new design. We demonstrated that co-training species and exploiting the advantage of sharing drug resistance profiles across species assist the model in identifying novel non-linear patterns that may be utilized to predict various types of drug resistance labels. This method also makes advantage of the masking loss function, which allows us to utilise the majority of the data, even if part of the drug labels are missing across species. Finally, we presented a unique state-of-the-art approach for predicting drug resistance that incorporates gene order and gene burden information. Our findings show that utilising the GEM-DR structure, gene burden-based prediction is highly effective. Apart from achieving good results on prediction, it is equally important to study the trained models and understand resistance-associated markers, something that seems to be a limitation of most previous work. One key challenge is to construct a reliable feature importance extraction method, preferably from a machine learning perspective, but possibly involving expertise-based feature engineering, to eliminate SNPs that may contribute to irrelevant genes being identified as important. By effectively eliminating noise, we are also able to include more relevant SNPs, which will potentially amplify the advantage of training models on a diverse dataset.

Our work can be seen as an exploratory study. To the best of our knowledge, it is the first one to investigate the issues of sparse and imbalanced dataset and the identification of drug resistance markers at different resolutions by analyzing machine learning models trained on 6 major datasets.

We hope that this study will inform future work on drug resistance in pathogenic bacteria and the application of machine learning to the drug resistance problem. Another Challenge of this work is identifying exact SNPs contributed to the prediction. Since the nature of our model has been trained by gene burden features, using interpretation models we are potentially able to find the significant contributed genes to the prediction; yet we are unable to identify the exact SNPs in identified genes. To do that, we may change the model architecture in a different manners in the future.

Because GEM-DR has been trained by different drug resistance patterns, this study may also be utilized as a transform learning task in other areas or drugs. The goal of this study was to develop the most accurate drug resistance prediction model possible. However, the lack of interpretability, as well as the absence of interpretability in most other neural network-based techniques, is an area that can be improved in the future. As a result, we want to add information to the GEM-DR technique in the future to make it more interpretable.

We Also conducted a study on E. coli pathways. In that work we utilized 368 pathways which include 4140 genes in E.coli. We used the shared information between the associated gens in each pathway to predict drug resistance for 17 drugs in E. coli. Comjparing our results with 3 baseline models (feedforward multi task learning, LR, SVM) we found that our model performed similarly to the base line model and in some cases lower than them. Our results indicates that due to the lack of information in this area we could not achieve better performance than our baseline models. We believe having more information in this area we can conduct more powerful studies and achieve better performance using pathways information in E. coli.

Bibliography

[1] JK Abernethy et al. "Thirty day all-cause mortality in patients with Escherichia coli bacteraemia in England". In: *Clinical Microbiology and Infection* 21.3 (2015), 251–e1.

[2] Harald Brüssow, Carlos Canchaya, and Wolf-Dietrich Hardt. "Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion". In: *Microbiol. Mol. Biol. Rev.* 68.3 (2004), pp. 560–602.

[3] Michael L Chen et al. "Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction". In: *EBioMedicine* 43 (2019), pp. 356–369.

[4] Michael L Chen et al. "Deep learning predicts tuberculosis drug resistance status from whole-genome sequencing data". In: *BioRxiv* (2018), p. 275628.

[5] Francois Chollet et al. "keras. GitHub repository". In: *https://github. com/fchollet/keras>. Accessed on* 25 (2015), p. 2017.

[6] James J Davis et al. "The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities". In: *Nucleic acids research* 48.D1 (2020), pp. D606–D612.

[7] Wouter Deelder et al. "Machine learning predicts accurately Mycobacterium tuberculosis drug resistance from whole genome sequencing data". In: *Frontiers in genetics* 10 (2019), p. 922.

[8] Alexandre Drouin et al. "Interpretable genotype-to-phenotype classifiers with performance guarantees". In: *Scientific reports* 9.1 (2019), pp. 1–13.

[9] Maha R Farhat et al. "Genetic determinants of drug resistance in Mycobacterium tuberculosis and their diagnostic value". In: *American journal of respiratory and critical care medicine* 194.5 (2016), pp. 621–630.

[10] Maha R Farhat et al. "Gyrase mutations are associated with variable levels of fluoroquinolone resistance in Mycobacterium tuberculosis". In: *Journal of clinical microbiology* 54.3 (2016), pp. 727–733.

[11] C Gagliotti et al. "Escherichia coli and Staphylococcus aureus: bad news and good news from the European Antimicrobial Resistance Surveillance Network (EARS-Net, formerly EARSS), 2002 to 2009". In: *Eurosurveillance* 16.11 (2011), p. 19819.

[12] Joseph J Gillespie et al. "PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species". In: *Infection and immunity* 79.11 (2011), pp. 4286–4298.

[13] Stephan Harbarth and Matthew H Samore. "Antimicrobial resistance determinants and future control". In: *Emerging infectious diseases* 11.6 (2005), p. 794.

[14] Tim Head et al. *scikit-optimize/scikit-optimize: v0.5.2*. Version v0.5.2. Mar. 2018. DOI: 10.5281/zenodo.1207017. URL: https://doi.org/10.5281/zenodo.1207017.

[15] Helge Karch, Phillip I Tarr, and Martina Bielaszewska. "Enterohaemorrhagic Escherichia coli in human medicine". In: *International Journal of Medical Microbiology* 295.6-7 (2005), pp. 405–418.

[16] Peter D Karp et al. "The ecocyc database". In: *Nucleic acids research* 30.1 (2002), pp. 56–58.

[17] Samaneh Kouchaki et al. "Application of machine learning techniques to tuberculosis drug resistance analysis". In: *Bioinformatics* 35.13 (2019), pp. 2276–2282.

[18] Samaneh Kouchaki et al. "Multi-label random forest model for tuberculosis drug resistance classification and mutation ranking". In: *Frontiers in microbiology* 11 (2020), p. 667.

[19] Pedro Larranaga et al. "Machine learning in bioinformatics". In: *Briefings in bioinformatics* 7.1 (2006), pp. 86–112.

[20] Francis N Lauener et al. "Genetic determinants and prediction of antibiotic resistance phenotypes in Helicobacter pylori". In: *Journal of clinical medicine* 8.1 (2019), p. 53.

[21] Daphne I Ling, Alice A Zwerling, and Madhukar Pai. "GenoType MTBDR assays for the diagnosis of multidrug-resistant tuberculosis: a meta-analysis". In: *European Respiratory Journal* 32.5 (2008), pp. 1165–1174.

[22] Paul S Mead and Patricia M Griffin. "Escherichia coli O157: H7". In: *The Lancet* 352.9135 (1998), pp. 1207–1212.

[23] Hanna Nebenzahl-Guimaraes et al. "Systematic review of allelic exchange experiments aimed at identifying mutations that confer drug resistance in Mycobacterium tuberculosis". In: *Journal of antimicrobial chemotherapy* 69.2 (2014), pp. 331–342.

[24] William S Noble. "What is a support vector machine?" In: *Nature biotechnology* 24.12 (2006), pp. 1565–1567.

[25] Ákos Nyerges et al. "Directed evolution of multiple genomic loci allows the prediction of antibiotic resistance". In: *Proceedings of the National Academy of Sciences* 115.25 (2018), E5726–E5735.

[26] World Health Organization et al. *Antimicrobial resistance global report on surveillance: 2014 summary*. Tech. rep. World Health Organization, 2014.

[27] World Health Organization et al. *The world health report 2007: a safer future: global public health security in the 21st century*. World Health Organization, 2007.

[28] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[29] Yanjun Qi. "Random forest for bioinformatics". In: *Ensemble machine learning*. Springer, 2012, pp. 307–323.

[30] AD Russell. "Antibiotic and biocide resistance in bacteria: introduction". In: *Journal of applied microbiology* 92 (2002), 1S–3S.

[31] Amir Hosein Safari et al. "Predicting drug resistance in M. tuberculosis using a Long-term Recurrent Convolutional Network". In: *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. 2021, pp. 1–10.

[32] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. *Practical Bayesian Optimization of Machine Learning Algorithms*. 2012. arXiv: `1206.2944 [stat.ML]`.

[33] Elisa Tagliani et al. "Diagnostic performance of the new version (v2. 0) of GenoType MTBDRsl assay for detection of resistance to fluoroquinolones and second-line injectable drugs: a multicenter study". In: *Journal of clinical microbiology* 53.9 (2015), pp. 2961–2969.

[34] GuoBao Tian et al. "Detection of resistance to β-lactams and characterization of extended-spectrum lactamases in Escherichia coli isolates from swine." In: *Zhongguo Yufang Shouyi Xuebao/Chinese Journal of Preventive Veterinary Medicine* 33.10 (2011), pp. 776–780.

[35] Richard L Vogt and Laura Dippold. "Escherichia coli O157: H7 outbreak associated with consumption of ground beef, June–July 2002". In: *Public health reports* 120.2 (2005), pp. 174–178.

[36] "WHO, Multidrug and extensively drug-resistant TB (MDR-TB) 2010 Global Report on Surveillance and Response". In: (2010).

[37] Yang Yang et al. "DeepAMR for predicting co-occurrent resistance of Mycobacterium tuberculosis". In: *Bioinformatics* 35.18 (2019), pp. 3240–3249.

[38] Yang Yang et al. "Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data". In: *Bioinformatics* 34.10 (2018), pp. 1666–1671.

[39] Malik Yousef et al. "Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier". In: *Bioinformatics* 22.11 (2006), pp. 1325–1334.