Gene Resistace Antibiotic

Taliya Shitreet, Renana Rimon, Tahel Zeharia

[1]School of Computer Science, Ariel University, Israel

Taliyashitreet@gmail.com, renana1414@gmail.com, tahelest@gmail.com

Abstract

Antibiotic resistance presents a significant global health challenge, emerging from bacterial evolution and the excessive use of antimicrobial agents. This issue has spurred the necessity for innovative strategies to address the intricate interplay between genes and antibiotics. When confronted with bacterial infections, the emergence of antibiotic-resistant strains undermines treatment effectiveness and can lead to complications. The comprehension of the likelihood of antibiotic resistance in a bacterial strain, based on its genetic composition, holds paramount importance for well-informed clinical decisions and the development of novel drugs.

The research pursued in this study aims to classify the extent of gene resistance to antibiotics. Existing databases encompass a partial classification of bacterial resistance, categorization of gene resistance to antibiotic families, and limited gene-to-drug resistance associations.

To fulfill this research objective, established databases are harnessed and machine learning techniques are employed, encompassing both supervised and unsupervised methods. While the unsupervised approach employs a bicluster model to explore bacterial and genetic relationships, the supervised method integrates a Multilayer Perceptron network with Relu and sigmoid functions for deep learning. Our efforts involve constructing and training specialized models for individual drugs using diverse datasets. However, challenges stemming from inconsistent data hindered optimal results.

In response to these challenges, a statistical model is introduced, capitalizing on existing data to prognosticate gene resistance probabilities. The model demonstrates a reasonable degree of predictive capability, based on the available data.

1. Introduction

ntibiotic resistance is a global health concern resulting from the development of bacteria and the misuse of antimicrobial substances. This phenomenon, arising from decades of unrestricted antibiotic use, has created a significant challenge to public health.

When a bacterial infection occurs, and the bacteria are resistant to antibiotics, a problematic situation arises – the antibiotics fail to eliminate the bacteria, and the infection persists, potentially becoming more complicated. Moreover, in some cases, other bacteria in the body may develop antibiotic resistance. Consequently, when needed, the bacteria are no longer responsive to the antibiotic, necessitating the development of new antibiotics.

The bacterial genome consists of various genes, each capable of conferring resistance to specific drugs. If it is known that a particular gene in the genome provides resistance to a drug, there is a high probability that the bacteria containing this gene will be resistant to that drug. Similarly, when given a specific bacterial strain and its constituent genes, if none of the genes confer drug resistance, the bacteria are not resistant, and the drug can effectively treat the patient.

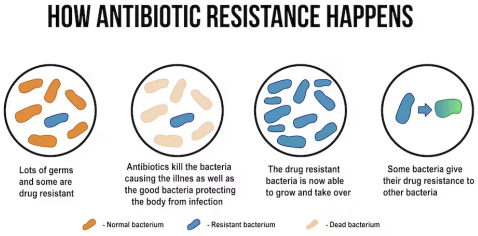

Figure 1: *the process of the antibiotics resistance arises*

For a new bacterial strain that is not yet known in the medical world whether it is resistant or susceptible to a particular drug, given a list of its genes, we aim to determine the probability of the bacteria being resistant to the drug. Therefore, we need to assess the likelihood of each gene conferring drug resistance.

The research objective is to classify the level of resistance of a gene to antibiotics. Currently, there is no comprehensive database that categorizes the resistance level of genes to drugs. The existing databases include:

A partial database that classifies the resistance of bacteria to specific antibiotics through laboratory tests.

A database classifying the resistance of a gene to an antibiotic family. In other words, with a high prob-

ability, the gene will confer resistance to drugs from this family, but not necessarily to all drugs within the family. Although, there is a high likelihood that it will not confer resistance to drugs from other antibiotic families. A highly limited database that classifies the resistance of a gene to a specific drug. To achieve the research goal, we will utilize the existing databases along with machine learning and deep learning tools, employing supervised and unsupervised methods.

In the unsupervised approach, we will use the bicluster model [1] to explore the relationship between different bacteria and genes. The model's objective is to classify examples into distinct groups based on Hamming distances. While the model successfully grouped the data into different clusters with strong similarities within each cluster, unfortunately, this division does not contribute to the research goal since the reaserch aim is to classify gene resistance, not bacterial. We will attempt to create a model specifically for individual bacteria to isolate the genes that have the most significant impact on resistance, but this attempt also did not succeed.

In the supervised approach with deep learning, we will use an MLP (Multilayer Perceptron) network [2] with the Relu and sigmoid functions. During training, we constructed 114 models, each trained for a specific drug, given the large and diverse datasets. The challenge in building these models was significant as the task varied between training and testing. During training, the model received data of a combination of genes, representing a bacterial strain, whereas during testing, the goal was to classify a single gene. The classification provided by the model was a probability value between zero and one, with one indicating gene resistance and zero indicating gene sensitivity to the drug. This method did not yield the desired results optimally, as the data was not sufficiently consistent.

As a result, the final model presented in the article is a statistical model. The model utilizes existing data and statistically predicts the likelihood of a gene's resistance to a drug. The final model successfully predicts the gene resistances reasonably well based on the available data.

2. Related Work

In recent years, the rise of antibiotic resistance genes (ARGs) and their implications for public health have garnered significant attention within the scientific community. With antibiotic resistance recognized as a global problem [3], there is a growing concern about its impact on morbidity, mortality, and treatment costs for infectious diseases.

Porooshat Dadgostar [4] sheds light on the potential consequences of widespread antibiotic resistance on healthcare systems, patient outcomes, and the global economy. It emphasizes the urgent need for innovative approaches, such as the development of accurate classification models for gene-antibiotic resistance associations, as pursued in our project.

Antimicrobial susceptibility testing plays a crucial role in guiding bacterial infection treatment, especially amid the growing threat of antibiotic resistance. Commonly used methods rely on phenotypic resistance detection, assessing bacterial growth in the presence of antibiotics. Although highly sensitive, these methods necessitate bacterial isolation and extended incubation times. Advances in understanding antibiotic resistance molecular mechanisms have spurred the development of rapid detection techniques [5].

Numerous studies have focused on understanding the mechanisms of bacterial resistance to antibiotics and devising effective strategies to combat this global health threat. For instance, Andre Brincat and Markus Hofmann [6] presents a text mining pipeline that employs deep learning models to automatically extract gene-antibiotic resistance relations from English biomedical abstracts. This approach accelerates the curation process by predicting genes linked to antibiotic resistance, providing valuable insights for clinical decisions and guiding future antibiotic research.

Additionally, researchers have explored the prediction of antibiotic resistance based on genetic markers and sequence data. In contrast to [7], which examines the performance of the protein-based AMRFinder tool, our work focuses on the development of a novel statistical model for predicting the likelihood of gene-antibiotic resistance associations. Unlike AMRFinder's identification of AMR genes in whole-genome sequences using protein-based hidden Markov models, our model capitalizes on existing data to predict gene resistance likelihood, demonstrating reasonable predictive capabilities based on the available information.

Furthermore, according to [8], the successful detection of resistance in clinically relevant bacterial species through MALDI-TOF MS technology has enabled early recognition of drug resistance and potential optimization of antibiotic therapy for better patient outcomes. However, in contrast to this application of MALDI-TOF MS that primarily focuses on detecting bacterial antimicrobial resistance, our innovative approach involves a statistical model that capitalizes on existing data to predict the likelihood of gene resistance.

Our project pursued the ambitious goal of developing a computational model capable of predicting antibiotic resistance based on gene-antibiotic associations. Through the integration of available genomic information, our model represents a novel approach in tackling the complex relationship between genes and antibiotics, aiming to offer insights into the challenges

posed by antibiotic resistance. As we continue to refine and enhance our model, we strive to contribute further to the ongoing battle against antimicrobial resistance and its impact on public health.

3. Datasets

The work process involved dealing with multiple data tables. Initially, we extracted essential data from the National Library of Medicine's website.

Specifically, the "organism_genotype_phenotype" table contains 22,097 rows. Each row represents an experiment conducted on a specific bacterial strain (isolate), and for each bacterial group (organism group), there can be multiple records of different strains. There are 37 distinct bacterial groups, and there is a different number of strains that have been tested in the laboratory and appear in the data.[Table 1].

The table [Table 2] includes two key columns: "AMR genotype" describes the list of genes found in the bacterial genome, and "AST phenotypes" contains information about the drugs tested for that particular bacterium. The drug entries are classified as R (Resistance), indicating that the bacterium has developed resistance to the drug, rendering it ineffective. Alternatively, entries marked as S (Susceptible) indicate that the bacterium is sensitive to the drug, and its usage can be beneficial in treating the disease. There are also entries marked as I (Intermediate), indicating that the bacterium is partially resistant and partially sensitive to the drug.

The table was further divided based on the drug resistance criterion, and two new columns were added instead of "AST phenotypes." These new columns are "drug," which denotes the name of the drug, and "resistance," where R is represented as 1, S as 0, and I as 0.5.

As a result, the table now comprises 316,071 rows and 5 columns. Among these, 234,401 entries are classified as 0 (indicating no resistance), 71,001 entries as 1 (indicating resistance), and 10,661 entries as 0.5 (indicating intermediate resistance).

Another table, called "resistance_validation," was extracted from the National Library of Medicine's website. This table partially serves our research objectives, as each row in the table represents a gene, the corresponding drug for which the gene confers resistance, and the drug class to which the drug belongs. The table comprises 1,252 rows. [Table 3]

Another central table used in the course of the research is the "drug_anti_r" table, which is a massive table with 14 columns and 2,920,586 rows. Its primary use was to extract information from the columns "aro term" and "drug class." For each gene (aro term), this table reveals to which antibiotic family it confers resistance. [Table 4]

Table 1: *Number of isolates in each organism group*

Organism group	#Isolate
Salmonella enterica	8573
E.coli and Shigella	6326
Campylobacter jejuni	2831
Acinetobacter baumannii	1113
Klebsiella pneumoniae	921
Staphylococcus aureus	638
Pseudomonas aeruginosa	598
Neisseria gonorrhoeae	477
Streptococcus pneumoniae	304
Enterobacter cloacae	66
Enterobacter hormaechei	39
Serratia marcescens	26
Providencia alcalifaciens	24
Klebsiella oxytoca	23
Enterococcus faecalis	20
Citrobacter freundii	17
Enterococcus faecium	16
Enterobacter roggenkampii	15
Pasteurella multocida	10
Vibrio parahaemolyticus	9
Stenotrophomonas maltophilia	7
Enterobacter kobei	6
Morganella morganii	6
Burkholderia cepacia complex	5
Enterobacter asburiae	5
Pluralibacter gergoviae	5
Corynebacterium striatum	3
Enterobacter ludwigii	3
Listeria monocytogenes	2
Vibrio cholerae	2
Enterobacter bugandensis	1
Cronobacter	1
Pseudomonas putida	1
Clostridioides difficile	1
Kluyvera_intermedia	1
Streptococcus agalactiae	1
Aeromonas salmonicida	1

Table 2: *Antimicrobial Resistance Data. Each row represents the antimicrobial resistant genes for each organism group, along with the listed genes on its genome and its resistance or sensitivity to specific drugs.*

Organism group	**Isolate**	**AMR genotypes***	**Drug****	**Resistance*****
Pseudomonas aeruginosa	PDT001677722.1	aac(6'), aadA6, aph(3')-IIb	ceftazidime-avibactam	1
Staphylococcus aureus	PDT001440611.1	abc-f, ant(6)-Ia, aph(3')-IIIa, blaI	sulfamethoxazole	0
Salmonella enterica	PDT000085904.2	mdsA, mdsB, tet(B)	azithromycin	0
Salmonella enterica	PDT000050264.3	aph(3")-Ib, aph(6)-Id, mdsA, mdsB, tet(B)	streptomycin	1
Salmonella enterica	PDT000135089.2	aph(3')-Ia, blaTEM-1, mdsA, mdsB	ciprofloxacin	0
Neisseria gonorrhoeae	PDT001615856.1	farB, mtrA, mtrC, norM	penicillin	0.5

* List of antimicrobial resistant genes encoded by the isolate (Computed from computational analysis results).

** Drugs tested in an antimicrobial susceptibility test (AST).

*** AST test results indicating the resistance level.

Table 3: *Gene and a drug they confer resistance to*

Gene	**Subclass**
marr	chloramphenicol
aph2iva	gentamicin
soxr	quinolone
baes	temocillin
npma	aminoglycoside
blarasa	cephalosporin
pona	beta-lactam

All these tables underwent a preprocessing and data cleaning process to organize the data in a way that is compatible with various databases.

Throughout the process, there was a need to create additional datasets. For each drug, there was its basic Training dataset, consisting of rows of vectors, where each row represents a bacterium that is a combination of genes. Each index in the vector represents a gene, and the entry in the vector is 1 if the gene is present in the bacterium's genome, otherwise it is 0. The label for each row in the data (for each vector) is 1 if the bacterium is found to be resistant to the drug, 0 if it is found to be susceptible, and 0.5 otherwise.

Another table we created contains the entire results of the process. This table has all the genes as its rows and all the drugs as its columns. Each cell in the table contains a probability value between 0 and 1, representing the level of drug resistance for that particular gene. If a cell contains -1, the information is not relevant. The table consists of 1209 rows and 114 columns.

4. Methodologies and interim results

In order to utilize deep learning models and machine learning, the study required working with appropriate data for model training. As part of this process, various model training variations were explored to arrive at the most relevant outcome for the task.

In general, the decision was made to train 114 distinct models, one for each drug. Each model was constructed as a deep learning model with four layers: the initial layer comprised 256 neurons, the second had 64 neurons, and the third had 16 neurons. Between each of these layers, the data underwent processing through a ReLU activation layer [9]. The final layer consisted of one neuron utilizing a sigmoid activation function to generate an output value between 0 and 1, representing the gene's resistance level to the specific drug. The binary cross-entropy loss function [10] as a hyperparameter was employed as it was well-suited for binary classification, encouraging the model to produce output values in the desired range of 0 to 1. Additionally, the Mean Squared Error (MSE) as a hyperparameter metric was utilized to transform the problem into a regression task, interpreting model predictions as continuous estimates rather than binary variables. This combined approach aimed to ensure accurate outputs, particularly when resistance values were close to 0 or 1, without being overly influenced by extreme values.

The training data for the models consisted of vectors composed of ones and zeros (the nature of which

Table 4: *resistance gene name (left) and its drug class.*

aro_term	**drug_class**
ANT(2")-Ia	aminoglycoside antibiotic
qacEdelta1	streptogramin antibiotic
sul1	sulfonamide antibiotic
adeF	fluoroquinolone antibiotic; tetracycline antibiotic

will be further elaborated). The label for each vector was assigned as follows: 1 for bacterium showing resistance to the drug, 0 for susceptible bacteria, and 0.5 for those exhibiting intermediate resistance based on laboratory test results.

The primary objective of the models was to accurately predict the data. For assessing the model's performance, a separate test dataset was created, comprising rows and vectors. Each vector represented a gene, where each index in the vector represented a different gene. For a specific gene in the input, the corresponding entry in the vector was set to 1, while for all other gene entries, the value was set to 0. The aim was to classify the model for each One- Hot vector, predicting the resistance level of that gene to the drug.

4.1. Unsupervised Method - Clustering Method

In order to achieve a comprehensive understanding of the data, an unsupervised learning method called biclustering was utilized. Biclustering is a data mining technique that enables simultaneous clustering of both rows and columns in a matrix.

The input data for the model consisted of the training dataset mentioned earlier, specifically for a particular drug, such as "ciprofloxacin," with the exclusion of the label column.

The model was trained with different clusters size: 2, 3, 7, and 50. Initially, the hope was that training with 2 clusters would enable the model to distinguish between examples labeled as 1 and those labeled as 0. However, this attempt did not yield satisfactory results. Another effort was made with 3 clusters, aiming to classify them into 0, 0.5, and 1, mirroring the original labels, but again, the outcome was not successful.

Following multiple attempts, the model was finally trained with 50 clusters, each intended to represent a group of specific examples with similar gene combinations and associated with the same original label. The objective was to investigate examples grouped in the same cluster, thereby identifying prevalent genes, which might indicate potential resistance.

The examples used for model training were isolate strains of bacteria subjected to laboratory testing. For each bacterium, thousands of different isolate strains were examined. When evaluating the 50 clusters classified by the model, certain similarities between various gene combinations were detected. However, in almost all clusters, the majority of records belonged to the same bacterium.

For instance[Table 5], in cluster 0, all records corresponded to isolate strains of "Acinetobacter baumannii." Furthermore, the majority of labels in this cluster were also labeled as 1, consistent with the majority of labels for "Acinetobacter baumannii" regarding the drug. Similar patterns were observed in clusters 9, 24, 31, 33, 37, and 48, where "Acinetobacter baumannii" appeared with the highest frequency. It can be observed that the bacteria appearing with a high percentage in most clusters are those for which there are already numerous examples in the original data[Table 1]. In other words, classifying clusters in this manner was not efficient for the research objective because the sensitivity/resistance of the bacteria to the antibiotic is already known. The goal is to identify specific gene resistances instead.

Nevertheless, this clustering approach did not efficiently serve the research goal of identifying specific genes responsible for resistance. Numerous clusters included strains from the same bacterium, making it challenging to pinpoint relevant genes.

Therefore, an additional approach using the biclustering model was employed, this time on specific bacterial data. To validate the results, the antibiotic "Amikacin" was chosen, for which true results were available in the validation data. The bacteria "Acinetobacter baumannii" were also tested for this antibiotic. The results of the clustering model on the bacteria did not yield significant findings; the clusters were mixed with records labeled as both sensitive and resistant, showing no meaningful distinction between them. Nevertheless, an investigation was conducted to identify genes that confer resistance despite this outcome.

The investigation was performed for each subgroup of labels within a cluster. For example, for a model with 2 clusters, there were 4 sub-groups. In the initial test, it was discovered that the majority of influential genes in a sub-group, meaning genes that appeared in most records within that sub-group, were the same as the genes present across all sub-groups. Consequently, these genes were characteristic of the bacteria, regardless of resistance or sensitivity. Therefore, these columns were removed from the training table, and the model was retrained to allow for more accurate clustering and to find similarities between genes

Table 5: *Clusters distribution for the Ciprofloxacin antibiotic shows the most frequently occurring '#organism group' and its corresponding Max Value_x, which represents the percentage of occurrences of that bacterium in the cluster. Additionally, for each cluster, the label that appears with the highest percentage and its Max Value_y, which represents the percentage of occurrences of that label in the cluster, are also provided.*

Cluster	#Organism group	Max Value_x	label	Max Value_y
0	Acinetobacter baumannii	1.000000	1.0	0.911765
9	Acinetobacter baumannii	0.751174	1.0	0.624413
24	Acinetobacter baumannii	0.974790	1.0	0.941176
31	Acinetobacter baumannii	0.982759	1.0	0.741379
33	Acinetobacter baumannii	0.426901	1.0	0.678363
37	Acinetobacter baumannii	0.906103	1.0	0.929577
48	Acinetobacter baumannii	0.545455	1.0	0.636364
3	Campylobacter jejuni	0.884647	0.0	0.817427
14	Campylobacter jejuni	0.563981	0.0	0.796209
22	Campylobacter jejuni	0.586207	0.0	0.676393
28	Campylobacter jejuni	0.409836	0.0	0.629508
35	Campylobacter jejuni	0.941284	0.0	0.642202
36	Campylobacter jejuni	0.435294	0.0	0.614706
41	Campylobacter jejuni	0.688596	0.0	0.815789
5	Campylobacter jejuni	0.450000	1.0	0.538462
30	Campylobacter jejuni	0.469388	1.0	0.632653
8	E.coli and Shigella	0.592105	0.0	0.671053
10	E.coli and Shigella	0.588384	0.0	0.868687
13	E.coli and Shigella	0.955171	0.0	0.959497
15	E.coli and Shigella	0.580808	0.0	0.752525
26	E.coli and Shigella	0.883278	0.0	0.865894
40	E.coli and Shigella	0.370474	0.0	0.841226
42	E.coli and Shigella	0.354610	0.0	0.730496
43	E.coli and Shigella	0.735294	0.0	0.781513
19	E.coli and Shigella	0.774011	1.0	0.502825
32	E.coli and Shigella	0.691964	1.0	0.723214
29	Klebsiella pneumoniae	0.367925	0.0	0.606918
25	Klebsiella pneumoniae	0.538462	1.0	0.826923
34	Klebsiella pneumoniae	0.802632	1.0	0.960526
39	Klebsiella pneumoniae	0.421053	1.0	0.684211
47	Klebsiella pneumoniae	0.500000	1.0	0.870370
49	Klebsiella pneumoniae	0.484848	1.0	0.555556
23	Neisseria gonorrhoeae	0.455899	0.0	0.570447
46	Pseudomonas aeruginosa	0.845361	0.0	0.443299
6	Pseudomonas aeruginosa	0.929293	1.0	0.707071
17	Pseudomonas aeruginosa	1.000000	1.0	0.991770
27	Pseudomonas aeruginosa	0.914286	1.0	0.657143
1	Salmonella enterica	0.984717	0.0	0.963956
2	Salmonella enterica	0.318841	0.0	0.681159
4	Salmonella enterica	0.784615	0.0	0.871795
11	Salmonella enterica	0.955890	0.0	0.982875
12	Salmonella enterica	0.760188	0.0	0.904389
16	Salmonella enterica	0.727003	0.0	0.973294
18	Salmonella enterica	0.734361	0.0	0.975521
20	Salmonella enterica	0.659574	0.0	0.920213
21	Salmonella enterica	0.684564	0.0	0.644295
45	Salmonella enterica	0.601542	0.0	0.719794
7	Salmonella enterica	0.682353	1.0	0.741176
44	Salmonella enterica	0.712821	1.0	0.769231
38	Streptococcus pneumoniae	0.500000	0.0	0.700000

that may have less impact compared to the dominant ones in the bacteria. Unfortunately, this attempt also did not yield significant results.[Table 6]

Table 6: *For each cluster, the percentages of occurrence for each label*

Cluster / Label	0	1
0	0.57	0.42
1	0.54	0.45

In conclusion, the data proved to be inadequate for constructing a model to identify genes responsible for resistance, as there were no strong correlations between these genes and the label 1 (indicating resistance).

4.2. Supervised methods

4.2.1. Method 1

A vector was created for each row in the data, representing a bacterium. Each index in the vector corresponded to a specific gene, and its value was set to 1 if the gene was present in the bacterium's genome, otherwise, it was set to 0. The goal was for the model to learn from these gene combinations and their impact on the bacterium's resistance level to specific antibiotics. The model had to classify the resistance level for individual genes, not just gene combinations.

The initial model performed well, achieving an accuracy of over 90% for gene combinations. However, it struggled to learn the correlation between gene combinations and the individual gene resistance levels. Additional potentially valuable information was received from the biological institute, represented in the "resistance_validation" table [Table 3]. Utilizing this data, we incorporated information that facilitated the analysis of model outcomes. The table furnishes details regarding the drug resistance of individual genes, rather than resistance as a gene combination. This aspect provided validation for the obtained results. To assess the accuracy of the relevant test, the resistance percentages for genes that were expected to exhibit high resistance probabilities to specific drugs were examined, as obtained from the "drug_anti_r" dataset [Table 4. The hypothesis was that genes belonging to certain antibiotic families should demonstrate high resistance probabilities to drugs within the same family. However, unfortunately, the results did not align with the expectations in this initial attempt.

4.2.2. Method 2

In this experiment, the emphasis was on categorizing genes into specific antibiotic families and assessing their impact on sensitivity or resistance to a particular drug. Two groups were created: *R* group included all genes belonging to the antibiotic family being tested in the model, while *S* group consisted of genes that, based on simple logic, were found to be sensitive to the drug (the logic assumed that if a bacterium is sensitive to a specific drug, all genes in its genome are sensitive to that drug). Conditions were set as follows:

If the label of a row is 1:

- If the gene not belongs to the *R* group or does not belong to the *S* group, its entry was set to 0.

If the label of a row is 0:

- If the gene belongs to the *R* group, its entry was set to 0.
- If the gene belongs to the *S* group, its entry was set to 1.

These conditions were modified from the default condition, which set the entry to 1 if the gene was present in the genome.

Despite these efforts, this experiment also did not yield satisfactory results. For genes that did not belong to either the *S* or *R* groups, high accuracy percentages were obtained, which was not the expected outcome since these genes might not be classified as sensitive or resistant due to their lack of association with the antibiotic family. The non-optimal performance of the model can be attributed to its challenges in discerning a definitive pattern among multiple variables during training. These variables encompass various factors such as the presence or absence of genes, sensitivity or resistance status, and the affiliation of genes to specific drug families. The complexity arising from the interplay of these diverse factors has likely contributed to the model's difficulty in establishing a clear pattern for accurate predictions.

4.2.3. Method 3

It was decided to perform some separation of variables using a two-stage model. Firstly, if a gene queried belongs to the *S* group, it directly receives a classification of 0.0; otherwise, it receives a prediction from the model.

In this method, the column count was duplicated. In other words, for each gene represented by a column at index x, an additional column exists at index x+len(columns) indicating whether the gene belongs or does not belong to the respective drug family. This addition of information provides the model with insights regarding the gene's affiliation to specific families.

Again, these results also provided high prediction accuracy for genes that do not belong to either the *S* or *R* groups, indicating a high number of false positives (FP). It seems that the model does not give enough importance to genes that do not belong to the family. The

model fails to grasp the patterns of gene influence on bacterial resistance.

In these attempts, it is important to mention that the results on the original test set are not the goal. The objective is not to predict the table on which we are training but rather to train on this table and evaluate the results on a different test set – one that our model has not seen before. We remove a lot of information from the data, so it is reasonable not to achieve good results on the data it was trained on. The test set on which the results were evaluated consists of individual genes rather than gene combinations. Each row in the test set represents a single gene, where the entire vector is filled with zeros except for the entry representing that specific gene.

4.3. Final Method - Logic Model

4.3.1. Methodology

Based on the fundamental understanding that only the genes belonging to the antibiotic family are relevant for predicting resistance, considering their frequency in the dataset, it was decided to reduce the dimensions of the data. We retained only the genes belonging to the *R* group, so the model would learn only from these genes, along with their corresponding labels. Here, a two-stage model was also used, but instead of using neural network layers, a logical model was applied.

The *logical model* operates on columns of the data that consist only of the *R*-group genes. The objective is to classify the genes within this group, determining which genes contribute to resistance and which do not. If a gene appears in the bacterial genome and the bacterium is sensitive to the antibiotic, it is likely that the gene contributes less to resistance. In this case, our model calculates the mean: if all bacteria in which a certain gene appears are resistant, the gene contributes 100% to resistance; if in 80% of the cases where the gene appears, the bacterium is resistant, then the gene contributes 80%, and so on. In this way, we expect to obtain more meaningful results.

Thus, the prediction is obtained differently by a function that takes the relevant *x_train* for the specific antibiotic and a list of genes, and returns the prediction. For each gene separately, the average occurrence of the gene is calculated with respect to label 1 (resistance).

4.3.2. Results

To truly evaluate the prediction results on the test set, we utilized the *validation* table [Table 3]. This table represents the validation for the test set, but only provides positive labels since it contains genes that confer resistance to specific antibiotics. However, we lack data that shows which genes in this table definitely do not confer resistance (sensitivity) to those antibiotics, and therefore, we do not have true labels of 0.

Table 7: *Example of results, 'Tobramycin' antibiotic*

Gene	Resistance probability
mdsA	0.5
dfrB1	0.67
dfrA32	1
dfrA5	0.22
dfrB5	0.4
dfrA7	0.16
dfrA19	0.91
mdsB	0.5
oqxA	0.87
dfrA27	0.85
dfrB3	1
dfrA14	0.81
dfrA15	1
dfrA16	0.7
dfrA25	1
mexA	0.63
dfrA17	0.48
oqxB	0.9
dfrA8	0.23
dfrA12	0.76
dfrA10	0.6
dfrA1	0.47
mexE	0.63

As a result, the most appropriate way to assess the model is using Recall, which measures how many of the genes in the validation the model successfully captures. Recall is calculated using a threshold of 0.8 for the prediction values.This threshold value was chosen after consultation with the Israel Institute for Biological Research, since our model is statistical, this means that above the threshold of 0.8 it is found in 80% of the rows in the training data that the same gene confers resistance to the specific drug, so there is no reason to take a lower threshold

It is important to note that there are only a total of 5 antibiotics with relatively valid data in the validation set (only the consistent data), the results are shown in [Table 8. this table shows relatively good results of the model when a gene is both related to the antibiotic family and belongs to the validation set. In such cases, the model tends to correctly detect the gene's resistance. However, there are instances where a gene is found in the validation set and belongs to the antibiotic family, but the majority of bacteria in which it appears are sensitive to the antibiotic. As a result, our model does not classify the gene as highly resistant, which is a logical and reasonable approach.[Table 7]

At the same time, there are contradictions in the data, where the validation provides genes that do

Table 8: *The table presents results for five antibiotics with relevant data. The "recall" column, calculated as the ratio of correctly identified genes in the predictions to the total number of genes in the validation set (column 3), indicates a slight miss for only a small number of genes (1 to 5). Columns 4 show genes that the model missed due to their infrequent appearance as "conferring resistance" in the training set. Column 5 reveals genes that the model missed because they were not classified as belonging to the antibiotic family.*

Antibiotic	Recall	Genes Detected / Genes in Validation	Genes with low Predictions	Genes not found in the Data
Tobramycin	83%	10/12	aac3ib: 0.27	aac3iva
Tetracycline	77%	17/22	tet35: 0.22, tet38: 0.11	tet34, mtrC, mtrR
Kanamycin	75%	3/4	aph3ib: 0.21	
Trimethoprim	83%	10/12		dfra23, dfra36
Vancomycin	87%	7/8	vang: 0.50	

not belong to the antibiotic family according to the "gene_anti_R" [Table 4. A reason for the discrepancies in the datasets according to the Israel Institute for Biological Research is that genes can be expressed in the bacteria or not. Also, another reason is that the tables were taken from different sources, so the names of genes/drugs in the biology can be presented in a slightly different way. Consequently, our model fails to correctly identify their resistance status. Overall, for the other antibiotics, we obtained unsatisfactory results mainly because of many inconsistencies in the data itself. The tables were collected from various sources, and this is the data provided to the process. For many antibiotics, we have seen cases where genes have been listed in the validation table, indicating their potential to confer resistance to the drug. However, in the x_train dataset, the same genes were found in bacteria that were sensitive to the same drug, leading to inconsistencies.

5. Conclusion and Future Work

The work yielded a logical model that leverages insights from a comprehensive analysis of international data on gene resistance to specific drugs. The provided model successfully classifies the data in a rational manner, taking into account the available information. Various attempts were made, including using a 4-layer deep learning model while considering the gene's associations with different antibiotic families. Although these attempts produced reasonable results based on the existing data, they fell short of ideal outcomes.

In conclusion, the logical model generated a table with 968 rows and 72 columns, where each row represents a gene and each column represents a drug, containing the predictive results of the logical model. The clustering method demonstrated that a deep learning model cannot discover a definitive pattern for determining gene resistance to drugs based on the current research data. However, several methods were tested, and with more consistent and informative data in the future about specific genes' resistance compared to gene combinations, more promising results and deeper insights could be achieved in the research.

6. References

[1] M. D. Noronha, R. Henriques, S. C. Madeira, and L. E. Zárate, "Impact of metrics on biclustering solution and quality: a review," *Pattern Recognition*, vol. 127, p. 108612, 2022.

[2] S. Sivasankari, J. Surendiran, N. Yuvaraj, M. Ramkumar, C. Ravi, and R. Vidhya, "Classification of diabetes using multilayer perceptron," in *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*. IEEE, 2022, pp. 1–5.

[3] World Health Organization, *The World Health Report*, Geneva, Switzerland, 1996.

[4] P. Dadgostar, "Antimicrobial resistance: implications and costs," *Infection and drug resistance*, pp. 3903–3910, 2019.

[5] M. R. Pulido, M. García-Quintanilla, R. Martín-Peña, J. M. Cisneros, and M. J. McConnell, "Progress on the development of rapid methods for antimicrobial susceptibility testing," *Journal of Antimicrobial Chemotherapy*, vol. 68, no. 12, pp. 2710–2717, 2013.

[6] A. Brincat and M. Hofmann, "Automated extraction of genes associated with antibiotic resistance from the biomedical literature," *Database*, vol. 2022, p. baab077, 2022.

[7] M. Feldgarden, V. Brover, D. H. Haft, A. B. Prasad, D. J. Slotta, I. Tolstoy, G. H. Tyson, S. Zhao, C.-H. Hsu, P. F. McDermott *et al.*, "Validating the amrfinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates," *Antimicrobial agents and chemotherapy*, vol. 63, no. 11, pp. 10–1128, 2019.

[8] W. Florio, L. Baldeschi, C. Rizzato, A. Tavanti, E. Ghelardi, and A. Lupetti, "Detection of antibiotic-resistance by maldi-tof mass spectrometry: An expanding area," *Frontiers in cellular and infection microbiology*, vol. 10, p. 572909, 2020.

[9] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[10] M. R. Rezaei-Dastjerdehei, A. Mijani, and E. Fatemizadeh, "Addressing imbalance in multi-label classification using weighted cross entropy loss function," in *2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME)*. IEEE, 2020, pp. 333–338.