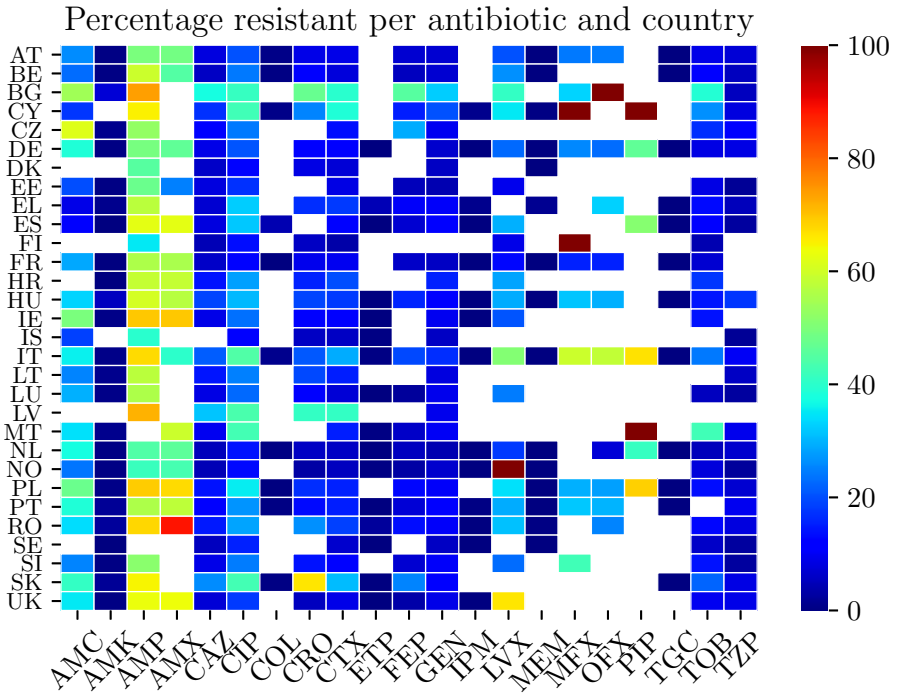
CHALMERS
UNIVERSITY OF TECHNOLOGY



Neural networks for predicting antibiotic resistance

- an analysis of performance

Master's thesis in Engineering Mathematics and Computational Science

Beatrice Skyman

Department of Mathematical Sciences
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2020

Master's thesis 2020

Neural networks for predicting antibiotic resistance

- an analysis of performance

Beatrice Skyman

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Applied Mathematics and Statistics
Biomathematics and Biostatistics
Chalmers University of Technology
Gothenburg, Sweden 2020

Neural networks for predicting antibiotic resistance
- an analysis of performance
Beatrice Skyman

Supervisor: Anna Johnning, Department of Mathematical Sciences
Supervisor: Erik Kristiansson, Department of Mathematical Sciences
Examiner: Erik Kristiansson, Department of Mathematical Sciences

Master's Thesis 2020
Department of Mathematical Sciences
Applied Mathematics and Statistics
Biomathematics and Biostatistcs
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Heatmap over percentage of antibiotic resistance in different countries.

Typeset in LaTeX
Gothenburg, Sweden 2020

Neural networks for predicting antibiotic resistance
- an analysis of performance

BEATRICE SKYMAN
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

This thesis investigate the possibility of predicting bacteria resistant to antibiotic treatment. The investigation is a part of developing a diagnostic tool for hospitals to use when deciding which antibiotic a patient should be treated with. This tool will be more and more relevant as antibiotic resistance and multidrug-resistance is spreading in the world with an accelerating rate.

Neural networks are used for predicting the probability of a bacterium being resistant to an antibiotic. The architecture of these networks is investigated and their outcome and performance are analysed. The distribution of the predictions are investigated and then different classification limits are tested. A classification limit gives a number of unclassified samples and errors, and these samples are investigated with regards to age, gender and country. Finally, the performance of the model is measured as the number of tested antibiotics is reduced.

When investigating the architecture of the neural networks the result are quite similar regardless to the number of layers and neurons. For the predictions, some antibiotics are more easily separated than other. This leads to a big variation in the number of unclassified samples and error between different antibiotics. When analysing the unclassified samples and errors only country could affect the predictions. The performance of the model when having more than four antibiotics tested as input is high. The conclusion is therefore that predicting antibiotic resistance using neural networks is possible and could potentially be used to replace measurements in the hospital laboratory.

Keywords: antibiotic resistance, neural networks, machine learning, predicting performance, diagnostic

Acknowledgements

Firstly, I want to express my thanks to my supervisors Anna Johnning and Erik Kristiansson. Thank you for providing the idea for this thesis. I am very grateful for your support, guidance and all interesting discussions we had during this spring.

I would also like to thank Love Westlund Gotby and my opponents Klara Granbom and Hampus Lane for all valuable feedback on this thesis.

Finally, I want to thank all my Chalmers friends for their support, not only during this thesis but also during my five years here.

Beatrice Skyman, Gothenburg, June 2020

Contents

List of Figures

List of Tables

Glossary

Notation	Description
major error	A bacterium incorrectly predicted as resistant when being susceptible. IX, 10, 11, 23, 30, 32, 33
multidrug-resistant	A bacterium is resistant to more than one antibiotic. 2
pathogen	A bacterium species causing disease. 5
resistant	If a bacterium is resistant to a specific antibiotic then that antibiotic will not kill the bacterium nor stop it from reproducing. I, IX, 1, 2, 5–11, 14–18, 21, 25–27, 29, 31, 32
specimen	The source of the sample, for example it could be a blood sample. 5, 31
susceptible	If a bacterium is susceptible to a specific antibiotic then that antibiotic will kill the bacterium or stop it from reproducing. IX, 5, 6, 8–11, 16, 17, 26, 31, 32
very major error	A bacterium incorrectly predicted as susceptible when being resistant. III, IX, 10, 11, 18, 19, 22, 23, 27–30, 32, 33

Acronyms

Notation	Description
AMC	Amoxicillin-clavulanic acid. 11, 13, 23, 30, 31
AMK	Amikacin. 13, 23
AMP	Ampicillin. I, III, 8, 11, 13, 17–19, 21, 23, 26–29, 31, 32
AMX	Amoxicillin. 13, 23
CAZ	Ceftazidime. 11, 13, 23, 29
CIP	Ciprofloxacin. I, III, 11, 13, 16–18, 23, 26, 27
COL	Colistin. 13, 23
CRO	Ceftriaxone. 13, 23
CTX	Cefotaxime. 11, 13, 23, 29
E. coli	Escherichia coli. 5, 14, 25
ETP	Ertapenem. 13, 23, 29
FEP	Cefepime. 13, 23
GEN	Gentamicin. 11, 13, 23
IPM	Imipenem. 13, 23
LVX	Levofloxacin. 11, 13, 23
MEM	Meropenem. 13, 23
MFX	Moxifloxacin. 13, 23, 25
OFX	Ofloxacin. 13, 23
PIP	Piperacillin. I, III, 11, 13, 17, 18, 23, 25–27, 29
TGC	Tigecycline. 13, 23, 31
TOB	Tobramycin. 11, 13, 23, 25
TZP	Piperacillin-tazobactam. 11, 13, 23, 30

1

Introduction

This chapter presents the purpose of this thesis. Firstly, some background is given, explaining today's situation regarding antibiotic resistance in the world. Secondly, the aim and the importance of the topic is presented. Lastly, the method and what is included in the thesis is presented.

1.1 Background

One of the biggest threats to global health today is antibiotic resistance. According to the *World Health Organization*, WHO, antibiotic resistance is widely spread over the world. An antibiotic is a drug used not to only to treat bacterial infections but also to prevent them [1]. A bacterium being resistant to a specific antibiotic means treating a patient with that antibiotic will be ineffective. Antibiotics can use different ways to kill bacteria or prevent them from reproducing. Bacteria can be naturally resistant to an antibiotic. An example is when an antibiotic attacks the process creating cell membrane, if the bacteria lack this specific process, the antibiotic will not work. Otherwise it could develop one or several defence mechanisms. A defence mechanism could for example be to decrease the uptake of the antibiotic or deactivate the antibiotic [2].

Bacteria can develop new resistance mechanism naturally. Today, antibiotics are widely used in both healthcare and food production, but unfortunately, antibiotics are also misused. The misuse leads to over consumption which accelerate the development of antibiotic resistant bacteria [1]. A bacterium can be resistant to more than one antibiotic, either because some resistance mechanism gives resistance to several antibiotics or because the bacterium have developed several resistance mechanisms. This phenomena is called multidrug-resistance and implicates few or non functional treatments [3]. When bacteria are resistant to several antibiotics resistant patterns can arise. These patterns could, for example, be that developing resistance against *Antibiotic A* often implicate being resistant to *Antibiotic B*.

With more antibiotic resistant bacteria infections like pneumonia and tuberculosis will be harder to treat. Additionally, taking the increase of multidrug-resistant bacteria into account the diseases will be even harder to treat. This is the reason

why the increase of antibiotic resistance is one of the biggest current threats to global health.

1.1.1 The process of finding an efficient antibiotic

When a patient has a bacterial infection and is in need of an antibiotic treatment, a physician makes an empirical guess regarding which antibiotic to use [4]. Meanwhile, a sample from the patient is sent to a lab where a few antibiotics are tested on the sample to see if the bacteria is resistant to any of them. To get the result from one antibiotic, it takes more than 36 hours. The process also takes a lot of resources, meaning, what they think are just enough antibiotics are tested. If the bacterium is resistant to all tested antibiotics, some other antibiotics are tested. This process goes on until a test shows non-resistant. This can therefore take valuable time, time a seriously ill patient does not have.

1.2 Aim of this thesis

Today, the process of finding an efficient antibiotic is very slow in the cases of multidrug-resistant bacteria. The aim of this thesis is therefore to investigate the possibility to speed up this process by predicting antibiotic resistance from the result of only a few antibiotics. This could, if successful, make it unnecessary to perform some of the antibiotic tests, which can lead to a patient getting an efficient treatment faster while using less resources.

1.2.1 The importance of this topic

As mentioned before, there is a problem with antibiotic resistance and the problem is increasing [1]. Especially problematic are multidrug-resistant bacteria where none or only a few antibiotics have effect. If a patient has a multidrug-resistant bacterial infection it is harder to treat and to find an efficient antibiotics. The process of finding a functional antibiotics can, in some cases, take a lot of time. Being able to predict antibiotic resistance would therefore speed up the process, save money, resources and lives.

1.2.2 Specification

The aim of the thesis is to speed up the process of finding an antibiotic that the analysed bacteria is not resistant to and which then can be given to a patient. This will be done using neural networks. A part during the development is to answer the question on how an optimal topology of a network would look like. This includes the number of layers and their sizes.

When a neural network has been chosen, the performance will be evaluated. This will be done by considering only the test result from a few antibiotics and see if some combinations performs better than others. The importance of the number of test results as inputs will also be investigated.

1.3 Methods

In this thesis the prediction of antibiotic resistance is done using neural networks. Before the predictions is done, the data, which is provided by ECDC extracted from The European Surveillance System – TESSy, is gone through [5]. Then relevant information from the data is extracted. This follows by an investigation regarding the architecture of a neural network. After deciding one neural network the outcome of the model is analysed regarding other aspects than test results, such as age of the patient and country. Lastly, the performance of the model is studied, for example how the performance change when giving the model less information.

1.3.1 Limitations

This master's thesis is a continuation of a previous thesis, where if neural network is a good method to predict antibiotic resistance was investigated. The neural networks model performed well and got an average error rate of 5 % [6]. Therefore, this thesis will continue the work on neural networks and not investigate other prediction models.

The previous mentioned master's thesis also briefly looked into the effect on adding country as input of the neural network. There are also more variables from the given data that could be added to the model, variables like a patient's age and gender and which bacteria it is. These aspects could be added as inputs to the model but due to a limitation will not be added. However, the analysis of the performance of the model will consider these aspects, for example to see if samples from some countries are easier to predict than others.

It is sometimes hard for antibiotics to tell harmful bacteria apart from the harmless. This means that antibiotics sometimes kills harmless bacteria, such as those in your digestive system. This phenomena is called a side effect and can affect the patient receiving the treatment. Different antibiotics have different side effects, some more severe than others [3]. This means that a doctor prescribing an antibiotic to a patient often prefers some antibiotics above others. It would be interesting to add the aspect of side effect when trying different combinations of input of the model. However, this thesis will have a mathematical perspective and therefore, the issue of prioritising some antibiotics due to side effects will not be considered in this thesis.

2

Theory and Method

In this chapter the investigation will be thoroughly explained step by step. In short, the investigation consisted of four parts. First, some data was extracted of the whole data set. Next, an architecture for the neural network model was found. Then, the outcome of the model was analysed and a method for classifying the output was chosen. Lastly, performance of the model was analysed for different combinations of input.

2.1 Filtering and restructuring of data

The data from TESSy consisted of eleven files from years between 2000 and 2017. Each file had results from several different pathogens [5]. Due to the size of the data, a limitation was done by only using data from 2017 and the bacteria Escherichia coli (E. coli). The data consisted of test results for those antibiotics that had been tested. When an antibiotic was tested on a bacteria sample the result could be one of three classes: resistant, intermediate and susceptible, for short R, I, and S, respectively. A generalisation was made by merging S and I into one class, S, meaning heron the test results will be one of two classes. The data also contained information about the hospital, which country the result was registered in and which specimen the samples came from, but also some information about the patient, like age and gender. The data now looked as in Table 2.1. A ”-” in the table meant that antibiotic had not been tested.

Table 2.1: An example of data extracted from the original data files from TESSy. There is, for example, information about the patient's age and gender, In the end of the rows, there is the test result for all antibiotics as R or S, but if an antibiotic had not been tested it was marked with ”-”.

Sample	Country	Age	Gender	Antibiotic 1	Antibiotic 2
1	Sweden	32	M	R	S
2	Germany	101	F	S	-

Different samples had a different number of antibiotics tested on it, coincidentally as some antibiotics are more common to test than other. To narrow the research

a bit more only antibiotics used on more than 10 % of the samples were considered and samples were less than 4 antibiotics had been tested were removed.

2.1.1 Construction of data sets for the neural networks

The first thing to do when developing a neural network was to structure the input and the output data. Each antibiotic had their own network since the test results from other antibiotics could affect each antibiotic differently. This meant, for each antibiotic's network there had to be two unique data sets, one for input data and the other for output. Each antibiotic was represented by two variables: one for resistant, R and one for susceptible, S. If a sample was resistant to an antibiotic the variable R was set to 1 and S to 0, and the other way around if susceptible. However, if an antibiotic was not tested both variables was set to 0.

For *Antibiotic A*, for example, the two data sets only consisted of samples where A had been tested. The input data set was the test result of all antibiotics except *Antibiotic A*. Then, the output was the test result for A. In both data sets, each antibiotic was represented by two variables, as described above. An example of a part of a input data set can be seen in Table 2.2.

Table 2.2: Example of a part of a input data set. The first sample was resistant to antibiotic 1, susceptible to antibiotic 2 and antibiotic 3 was not tested.

Antibiotic / Sample	1		2		3	
	R	S	R	S	R	S
1	1	0	0	1	0	0
2	0	1	0	0	0	1

To get an overview of the samples and the relation to countries a heat map was made for the number of samples each antibiotic had been tested on in each country. The ratio of resistant samples for different antibiotics varied a lot. Therefore, it was interesting to see if the ratio varied in different countries. This was done by constructing a heatmap over the ratios.

2.2 Investigating architecture of a neural network

A standard type of a neural network called multi layer perceptron (MLP) consists of a number of layers, each with a given number of neurons in it. Figure 2.1 shows an example of such an MLP. As in the figure, the layers are ordered sequentially so that one layer comes after another one. The first layer is the input layers, then there are zero or more so called hidden layers, and finally there is the output layer. A layer is connected to the next layer by the neurons, meaning all neurons in, for example, the first layer is connected to all neurons in the second layers. On these connections, there are so called weights and on each neuron, there is a bias. The input data is weighted and then the bias is added on each neuron. After a layer it is common to use an activation function, usually to add non-linearity to the model

[7]. This process is repeated for each layers. The output of a neural network model could for example be a prediction of a fund's risk or, as in this case, a prediction of a sample being resistant to a specific antibiotic. The size of a layer and how many to use is optional.

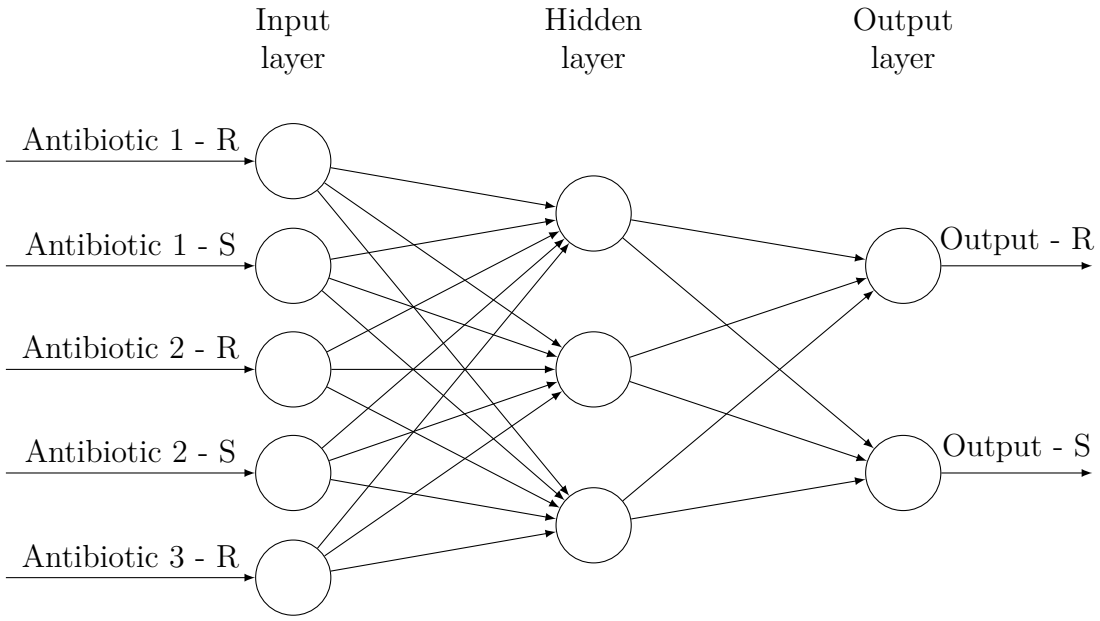


Figure 2.1: An example of a part of one of the neural networks. The input layer consist of test results for all antibiotics except the one being predicted. The output layer will give the probability of the input sample being resistant and susceptible, respectively, to the given antibiotic.

A neural network model has a learning process that mathematically means that the weights are optimised. In this thesis, supervised learning was used and meant the model could learn from the test results. A comparison between predictions and test results are done by a cost function, which, simplified, measures how far apart they are. The weights are then optimised so that the cost function is minimised. Then, usually *gradient descent* is used. This means updating the weight in the direction where the gradient is steepest, since the cost function should be minimised [7].. However, here the stochastic gradient descent was used which is an stochastic approximation of the gradient descent method [8].

In this thesis there was one neural network for each antibiotic. Each network started with one input layer, on the form described in Section 2.1.1 and in Table 2.2. This meant each antibiotic was represented by two nodes, one for resistance and one for susceptibility. The same was applied for the output layer, which only represented one antibiotic. In Figure 2.1 the output layer and a smaller part of the input layer is shown. As seen in the figure, there is one additional layer called hidden layer. During the next part different number of hidden layers were considered with different number of neurons.

There are very many parameters when working with neural networks, such as number of layer and neurons, activation functions and cost functions. Some limitations were therefore made to make this part more apprehensibly. The first thing was to only use one antibiotic during the investigation. Many antibiotics had a low resistant rate and if so the model could set all predictions to susceptible and still get a very good accuracy. Therefore, the antibiotic chosen was AMP where the rate was near 50 %. Another limitation was to use *ReLU* as activation function for all layers except the output layer. The ReLU function was defined as $f(x) = \max(0, x)$ with the addition of bounding the maximum function value to 6 [9]. For the output layer, softmax was used as activation function and was defined as

$$f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}} \tag{2.1}$$

where x_i represents the value at neuron i and K is the number of neurons in the considered layer. This function gave the probability of the sample being resistant or susceptible [10]. This means the sum of all $f(x_i)$ is 1. As cost function *Categorical cross entropy* was used and that was because it can handle outcome of the type categorical probabilities [11].

The remaining parameters were the number of layers and the number of neurons in each layer. To analyse how different number of layers and neurons affect the predictions a simplified algorithm was used. This algorithm proceeded from the optimal network from the previous number of layers and can be seen in Algorithm 1. The algorithm then built a network like this with 10 layers and a maximum of 20 neurons per layer.

Algorithm 1: Greedy algorithm

Result: Best size of layer i given a fixed size of all previous layers

```
architecture = [];                    // Size of layer i is placed at architecture[i].
for Layer i = 1:10 do
    lowest_cost_function = infinity;
    for neuron j = 1:20 do
        test_architecture = architecture[1:(i-1)] + [j];          // Use the found
         architecture for the previous layers and have j neurons in layer i.
        cost_function ( test_architecture );  // Calculate the cost function for
         the given architecture.
        if cost_function < lowest_cost_function then
            lowest_cost_function = cost_function;
            architecture[i] = j;     // If this architecture gives the lowest cost
             function, save it.
        end
    end
end
```

The input layer was of size 40 and the greedy algorithm only considered up to 20 neurons per layer. Therefore, some arbitrary networks with larger layers were tested.

This was done so that each hidden layer in the network had fewer neurons than the previous layer. They had four to five hidden layers and used the same cost function and activation functions as described above. For further investigation, a neural network consisting of 4 hidden layers with 30, 22, 15 and 7 neurons respectively was used.

All steps in developing the model was done in Python. The packages used for the neural networks were Tensorflow and Keras. When training the model, 80 % of the data was used for training and 20 % was used for validation. The data was randomly split in these two sets. The model learned using training data and then its performance was evaluated on the validation data. The model trained during 200 epochs.

2.3 Analysing the results of the model

An architecture had now been chosen and the neural networks was trained. Then, the model could be analysed starting with looking at the distribution of the predictions, for example if it was a clear separation between the resistant and the susceptible samples. Then, the different methods used to classify the outcome is presented. The classification resulted in errors and unclassified samples, which is further explained in Section 2.3.2. The errors and unclassified samples were looked into to see if they had some common attributes. Lastly, unique classification limits was found for each antibiotic. These limits was then used to investigate different combinations of antibiotics as input of the model. The different parts will be more detailed described below.

2.3.1 Visualise predictions

Since the activation function on the last layer in the neural network model was softmax, the output variables were between 0 and 1 and the sum of them was equal to 1, according to Equation 2.1. One way to visualise the predictions for an antibiotic was to make a histogram over the samples. To put more information into the visualisation the resistant and susceptible samples were visualised in different colours, blue for the resistant and yellow for the others.

Because the output takes a number in the range 0 to 1 this can lead to a compact distribution. This means many samples can take approximately the same values. One way to overcome this was to use a logit transformation, after the activation function, to stretch out the predication. The logit transform was defined as

$$g(x) = \log\left(\frac{x}{1-x}\right)$$

and therefore gave values between the negative and positive infinity. For the transformation base-10 logarithm was used.

2.3.2 Analyse limits for classification of the predictions

When measuring the performance of this neural network model, having the output as classes, instead of continuous variables, would make it easier. For example, the number of errors cannot be calculated without the output as classes. Therefore, classification limits were used and only the resistant variable of the output was considered. There were two limits. The first one marked the susceptible samples, meaning all samples with output variable below that limit was classed as susceptible. If the variable was larger than the second limit, then it was classified as resistant. This meant some samples were stuck between these two limits and they were called unclassified. This method was used because the resistant and susceptible groups were often separated. However, often some parts of the two groups were overlapping and that is the reason for using two limits. The overlapping parts became the unclassified sample. The classification could have been done using only one limits, meaning none unclassified samples. The problem then would be that many samples would be incorrect predicted.

Two approaches for classification were compared. The first approach was to set several symmetric limits close to 0 and 1, respectively, and see how many errors and unclassified samples each limit got. The errors were divided into two groups: major and very major error. Very major error is when a resistant sample is classified as susceptible and major error is the opposite. All limits were evaluate separately for each antibiotic. The other approach was to compare the prediction to the proportion of resistant samples, heron called the resistant ratio. The idea was that if the model does not know whether a sample is resistant or not, then the prediction would be close to the resistant ratio. To handle this problem, the chosen approach was to use fix steps from the ratio but on the logit transformed prediction as described in Section 2.3.1. The ratio was of course also logit transformed.

2.3.3 Analyse unclassified samples and errors

When using the approach of having two limits, samples could be between these limits, so called unclassified samples. To understand the model more, these samples and errors are important to analyse. For this analysis the logit transformation method, as described in Section 2.3.2, with a fixed limit 1 unit of length from the ratio. This was analysed by looking at the distribution of age, gender and country in these samples and compare it with the distributions in the whole data set. Again, this was done for each antibiotic separately because different samples were unclassified for different antibiotics. Another interesting aspect was the number of other antibiotics tested on the unclassified samples. This was also analysed in the same way as age, gender and country.

2.3.4 Analyse correlation between variables and test result

As mentioned in Section 1.1, patterns sometimes arise between different antibiotics meaning that, for example, a bacterium being resistant to *Antibiotic A* always implies being resistant to *Antibiotic B* but not the other way around. This was

visualised by a correlation plot of all test results, separately for each antibiotic.

Apart from the test results, the correlation between some variables were measure, starting with the age of the patient. The two other variables were the number of other antibiotics tested and the number of antibiotics a sample was resistant to. A conjecture was that resistance to more antibiotics increase the probability of resistance to other antibiotics. A hypothesis about the correlation of the number of other antibiotics tested was not equally straightforward. Nevertheless, both variables are interesting to analyse in a correlation aspect. When calculating the correlation, the Pearson correlation coefficient was used.

2.4 Antibiotic-unique classification limits

The model's predictions for different antibiotics could varied in how much separation there were between the two classes, *R* and *S*. One could therefore argue that antibiotic-unique limits for determine the classes should be used. When finding these limits several approaches can be used, such as minimising the number of unclassified samples and the number of errors. To classify a bacteria sample as susceptible when it really is resistant is a severe mistake and it is therefore important to minimise when developing a model. Hence, limits were set symmetric around the resistant ratio so that the very major error rate would not exceed 5 %. During this procedure the logit transformed predictions were used. After finding these limits the number and percentage of unclassified samples and major errors could be analysed to see how good these limits were depending on other aspects than very major errors.

2.5 Performance of the model

Performance can be measured in several ways. A naive way is to just count the errors, preferably in percent. The problem with this naive measure is that the unclassified samples are not considered. Therefore, a score system , where every correct prediction gives one point, every wrong prediction gives one minus point and every unclassified samples gives zero points, was used. The sum divided by the total number of samples solves the problem of not considering unclassified samples in the naive measure.

The procedure this thesis aims to make faster will use the test result of some antibiotics to predict resistance of the other antibiotics. Therefore the performance of different combination of antibiotics as inputs was investigated. To narrow the investigation a bit only ten of the 21 antibiotics were considered when finding the best combination to use as inputs. This meant that all the other antibiotics had a zero-zero input, as if they had not been tested on that sample, and their predictions were not evaluated. The ten antibiotics chosen were *AMC, AMP, CAZ, CIP, CTX, GEN, LVX, PIP, TOB and TZP*. They were chosen because they did not have extremely few resistant samples and used different resistance mechanisms.

The combinations were the input of the model and consisted of zero to nine antibiotics. This meant that only samples having test results on all antibiotics in the combination was considered when calculating the performance. The remaining input antibiotics was set to zero-zero, as if they have not been tested. Then the score of each combination for every antibiotic was calculated. When analysing the performance, the mean performance of the number of input antibiotics. The purpose was to see if more antibiotics tested improved the performance. The evaluation was also done for all antibiotic separately, meaning finding the best combination for every number of inputs on every antibiotic respectively. Since some antibiotics could be harder to predict it was relevant to analyse the model's performance on antibiotics separately.

2.5.1 Finding the best worst-case scenario

In the previous paragraph how to find the best combination for every antibiotics was described. However, it may be more interesting to know the best worst-case scenario. This since only looking at the overall best combination does not tell if all other combinations are performing okay, well or bad. The best worst-case scenario was computed for every number of inputs, i.e. one to nine. The steps were the following, here for the example having 3 antibiotics as inputs.

1. Find all combinations of 3 antibiotics
2. For each combination:
   (a) For every antibiotic left of the ten chosen, see Section 2.5, calculate the score when having this combination as input
   (b) Save the worst score
3. Take the combination with the best score, of those saved worst score.

The result from this will be one combination per number of inputs accompanied by the score.

3

Results

In the following chapter the result found during the investigation part will be presented. During the process, 21 different antibiotics were considered that is to many too present in this report. Therefore, mainly the result from three antibiotics, *AMP*, *CIP* and *PIP*, will be shown. These are chosen to show the wideness of the results.

3.1 Extraction of data

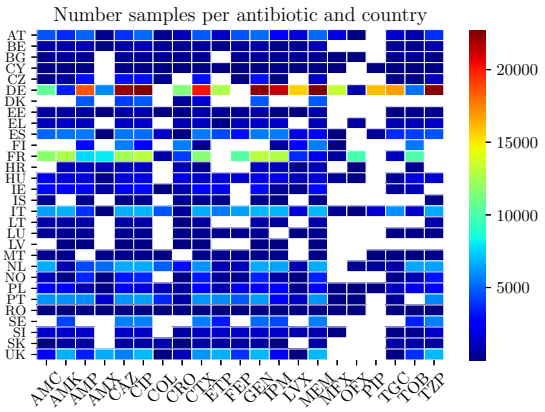
Table 3.1: Number of samples per antibiotic and distribution of resistant and susceptible samples.

Antibiotic	Number samples	Number R	Number S	% R	% S
AMC	68 996	22 595	46 401	32.75	67.25
AMK	67 198	630	66 568	0.94	99.06
AMP	83 626	44 832	38 794	53.61	46.39
AMX	30 966	16 919	14 047	54.64	45.36
CAZ	10 7270	10 129	97 141	9.44	90.56
CIP	110 511	24 108	86 403	21.82	78.18
COL	19 259	187	19 072	0.97	99.03
CRO	37 780	4 169	33 611	11.03	88.97
CTX	95 809	11 944	83 865	12.47	87.53
ETP	51 531	100	51 431	0.19	99.81
FEP	44 305	4 264	40 041	9.62	90.38
GEN	107 689	9 242	98 447	8.58	91.42
IPM	78 265	60	78 205	0.08	99.92
LVX	34 883	80 85	26 798	23.18	76.82
MEM	97 512	66	97 446	0.07	99.93
MFX	17 450	4 703	12 747	26.95	73.05
OFX	12 454	2 191	10 263	17.59	82.41
PIP	19 257	9 348	9 909	48.54	51.46
TGC	36 248	56	36 192	0.15	99.85
TOB	52 554	4 882	47 672	9.29	90.71
TZP	78 724	6 005	72 719	7.63	92.37

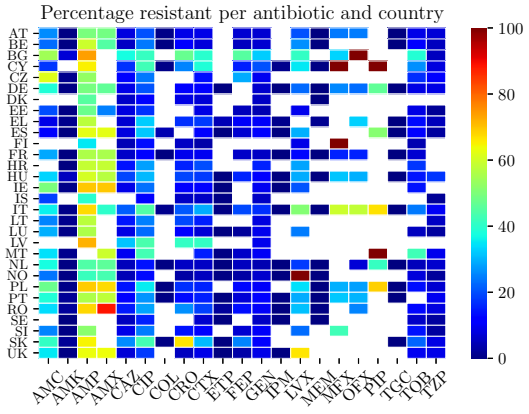
Relevant data was extract from TESSy's register. This was done both by only

choosing E. coli samples from 2017 and removing samples that have tested three or less antibiotics. After filtering, 116 541 samples were left. The distribution over these samples per antibiotic and number and ratio of resistant samples can be seen in Table 3.1.

An overview of the samples in the form of heat maps can be seen in Figure 3.1. Figure 3.1a shows on how many samples an antibiotic had been tested on in each country. A white cell means that antibiotic had not been tested on any samples in that country. Figure 3.1b shows the percentage of resistant samples, which means values in the heatmap are between 0 and 100 %. Again, a white cell means no resistant samples for that specific antibiotic in that country.



(a) Distribution of number of samples per antibiotic and country. A white cell indicates that an antibiotics is not used in a given country.



(b) Distribution of percentage of resistant samples per antibiotic and country. A white cell indicates that none samples are resistant to a given antibiotic and country.

Figure 3.1: Two heat maps of the distribution of samples and resistance ratio in different countries. The white parts represents when an antibiotic is not used at all in a country.

Next, unique data sets were constructed for the antibiotics specific neural networks. For *Antibiotic A*'s network the samples used to train and evaluate the model had to have a test result from *Antibiotic A*. This meant a sample can only be used in the training set to those antibiotic-networks the sample have been tested for. The number of samples used in training and validation for each network was the same as *Number samples* in Table 3.1.

3.2 Investigation of the network architecture

The first part of the investigation regarding a neural network's architecture was the simplified algorithm, explained in Section 2.2, where different number of layers and neurons were tested with up to ten layers and with a maximum of 20 neurons in

each layer. The result of the algorithm can be seen in Table 3.2. The cost function, *Categorical crossentropy*, can also be seen in the table. Note that the cost function on, for example, layer 3 is for the model having 18 neurons in the first layer, 18 neurons in the second layer and 15 neurons in the third layer.

Table 3.2: Result of the simplified greedy algorithm introduced in Section 2.2.

Layer	Number neurons	Categorical crossentropy
1	18	0.3115
2	18	0.3080
3	15	0.3066
4	15	0.3057
5	8	0.3062
6	6	0.3062
7	19	0.3064
8	17	0.3060
9	12	0.3063
10	11	0.3062

Some arbitrary architectures for the neural network considering layers of bigger size were also tested. Several architectures were tried and the one that performed best was the model found in Table 3.3.

Table 3.3: Result of the best, non-greedy and arbitrary, chosen model. Note that the categorical crossentropy is, in general, lower than for greedy algorithm.

Layer	Number neurons	Categorical crossentropy
1	30	0.3106
2	22	0.3062
3	15	0.3064
4	7	0.3050

3.3 Analysis of the outcome of the model

In this section the results of the outcome will be presented. This includes the distribution of predictions and how different limits affect the number of errors. The results of the analysis of the unclassified samples is shown with regard to age, country, gender and the number of other antibiotics tested. Then, the correlation between the test results and number of samples is presented.

3.3.1 Visualisation of the predictions

The output of the model was two continuous values between zero and one for the probability of resistance and susceptibility. A plot of the resistant variable for three

antibiotics can be seen in Figure 3.2. The true resistant samples are coloured blue and the susceptible ones are coloured yellow..

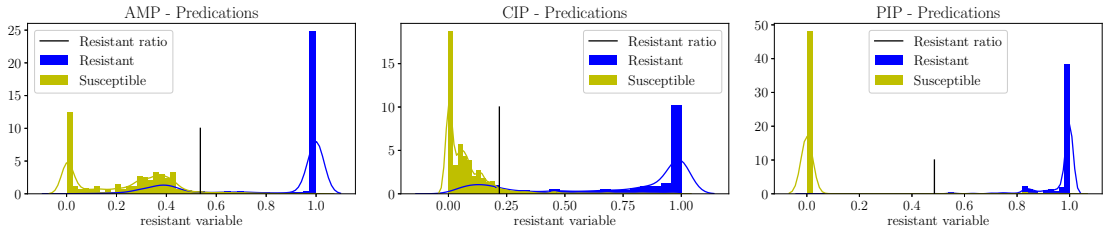


Figure 3.2: Here are distributions of the output of the model for three antibiotics: AMP, CIP and PIP. The blue parts are the resistant samples and the yellow are susceptible. The predictions shown are only the resistant variable, which means a prediction close to 1 represents the model predict the sample being resistant.

As mentioned in Section 2.3.1, a logit transformation of the prediction could be relevant to apply. A visualisation of the logit transformed predictions can be seen in Figure 3.3. To get a point of reference when analysing these plots the logit transformed resistant ratio is plotted as a black line.

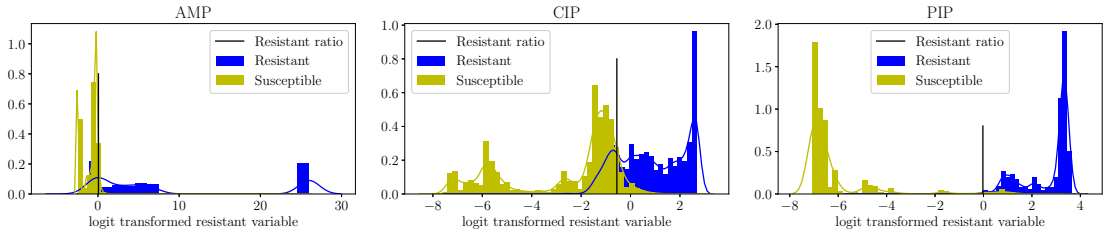


Figure 3.3: Distribution plots of the logit transformed output for three antibiotics: AMP, CIP and PIP. The blue parts are the resistant samples and the yellow are susceptible. The logit transformed resistant ratio is shown as a black line.

The difference in prediction depending on country is visualised. This means that all samples are grouped by country. Each plot shows an antibiotic's predictions for the respective country. The result for antibiotic *CIP* can be seen in Figure 3.4. Countries with less than 500 samples are marked with a red title.

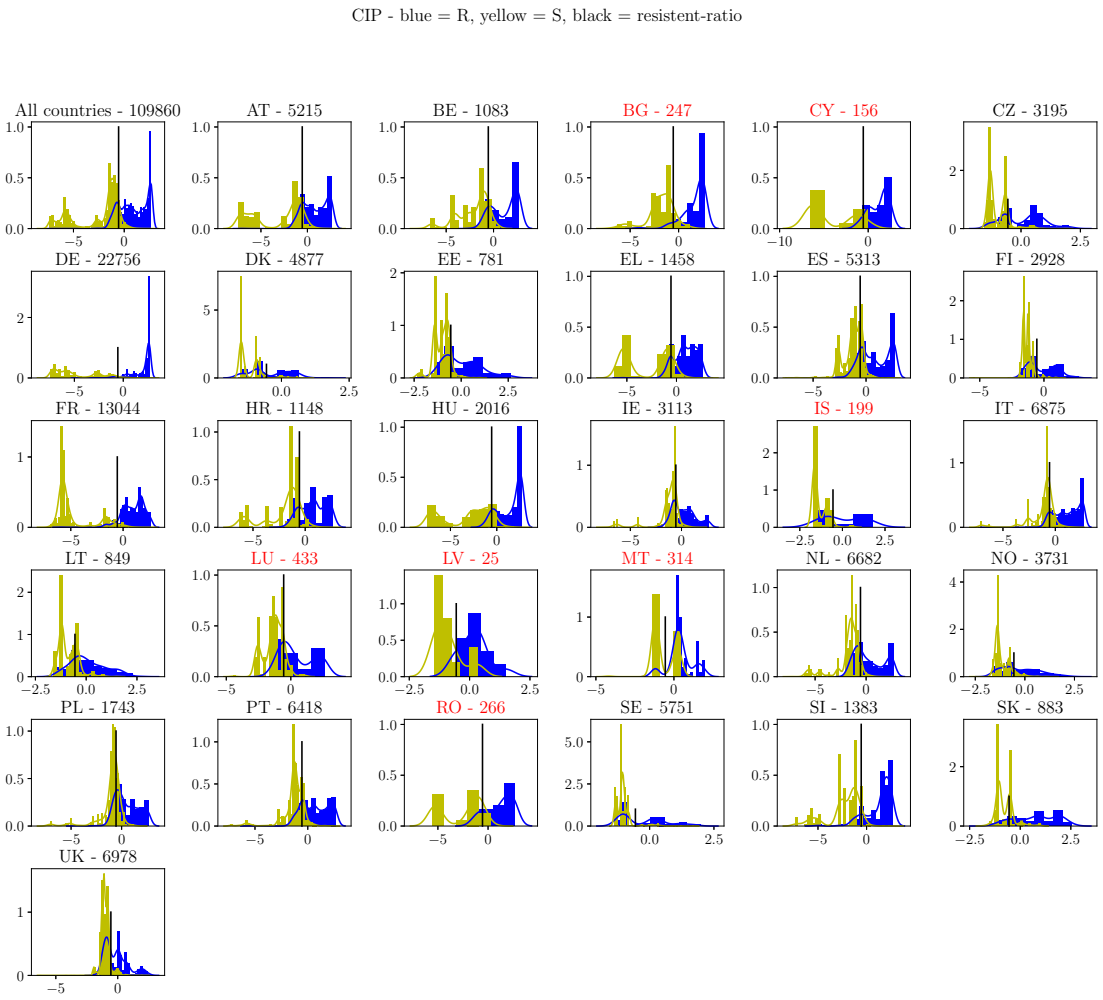


Figure 3.4: Logit transformed predictions of antibiotic CIP for every country. The black line represents the resistant ratio of all samples, i.e. it is not unique for every country. For some countries the predictions are well separated, for examples see Germany, *DE*, and for some countries the resistant and the susceptible samples are overlapping, see Spain, *ES*.

3.3.2 Investigate limits for classification of the predictions

Since the output of the model was a continuous variable it is preferable to set limits where a sample is classified as resistant or not. Here, two limits are used; one, where samples predicted under it are classified as susceptible and one, where samples predicted over it are classified as resistant. This meant some samples could be between these limits and are thus unclassified. The first analysis was on the outcome of the model, i.e. not transformed. Different limits were set close to 0 and 1. Then the major and very major error rate were calculated for every pair of limits together with the percentage of samples that were unclassified. The result for AMP, CIP and PIP can be seen in Figure 3.5.

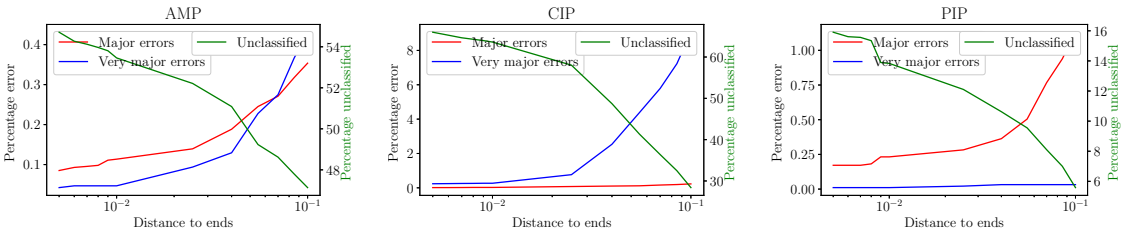


Figure 3.5: Investigation of different symmetric limits on the output for the three antibiotics: AMP, CIP and PIP.. The percentage of errors and unclassified samples are shown as a function of the distance to 0 and 1, respectively. The results are clearly very different for the three antibiotics, both regarding the percentage of errors and unclassified samples, but also the increase and decrease depending on the distance.

The other approach was to use the logit transformed predictions and starting from the resistant ratio, then taking a fixed step from the ratio as limits. The result for AMP, CIP and PIP using this method is shown in Figure 3.6. The percentage of error and unclassified is this time a function of the distance from the ratio.

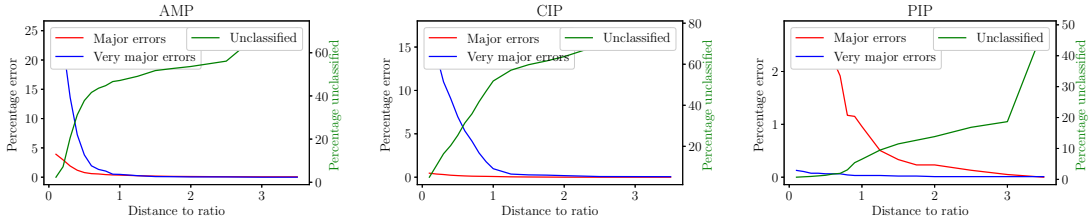


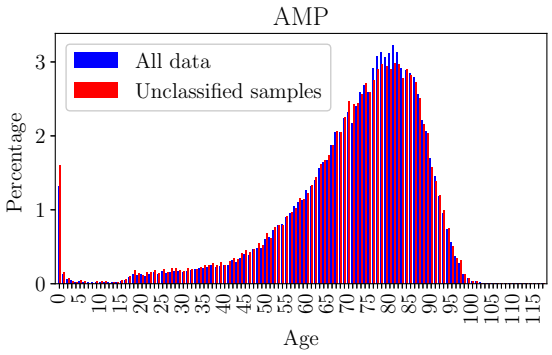
Figure 3.6: Percentage of unclassified samples and errors divided in: major and very major error depending on the distance from the limits to the resistance ratio, logit transformed for the antibiotics AMP, CIP and PIP. As in Figure 3.5, the results differs between the antibiotics. If comparing with other approach, again see Figure 3.5, the graphs here do not increase and decrease in the same ways as the others..

3.3.3 Analysing unclassified samples and very major errors

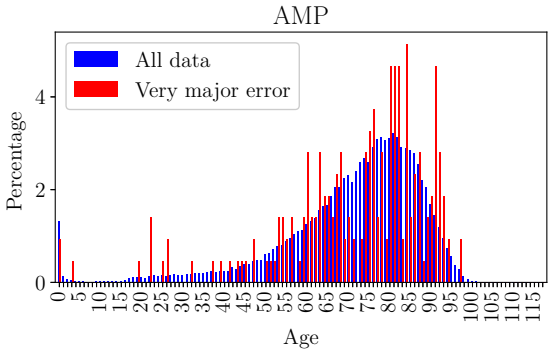
As seen in the Figures 3.5 and 3.6 there were both errors and samples that got unclassified. To improve the model it is important to understand why samples get unclassified or not predicted correctly. This section will show the results from analysing the unclassified samples, where the predictions were logit transformed and limits for the classification were set to 1 unit of length of each side of the resistant ratio. The very major error cases was also analysed. The attributes age, gender, country and number of antibiotics tested, were investigated to see if these samples had things in common. The figures in this section only considers the antibiotic, AMP,that had 39 388 unclassified samples and 214 very major errors. This corresponds to 47 % and 0.3 %, respectively, of the whole data set for AMP.

3.3.3.1 Age

The first interesting parameter was age. For most samples the patient's age was known. The analysis consisted of a comparison of the age distribution for all data and for the unclassified and very major error respectively. A comparison for antibiotic *AMP* can be seen in Figure 3.7.



(a) The distribution of age on the whole data set for AMP compared with the distribution for the unclassified samples. The distributions are very similar apart from a minor decrease around 80 for the unclassified samples.



(b) The distribution of age on the whole data set for AMP compared with the very major errors. Even with very few samples, the very major errors seems to follow the whola data sets distribution.

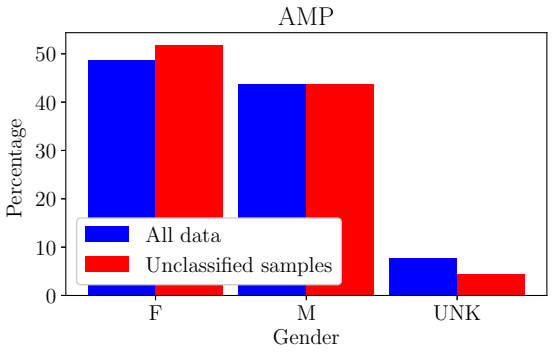
Figure 3.7: Two comparisons of the age distribution for antibiotic AMP, for the whole data set against the unclassified samples and the very major errors, respectively. The classification limits are set symmetrically to 1 unit of length around the ratio.

3.3.3.2 Gender

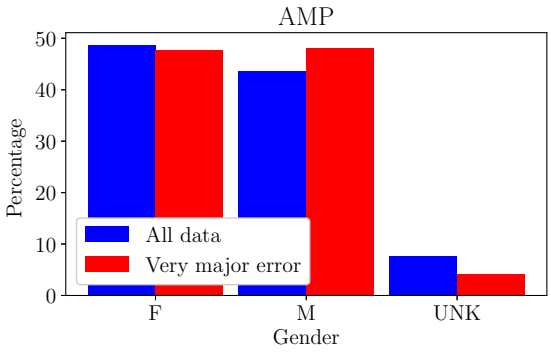
For each bacteria sample there was information about the patient's gender, they were called *F* for female and *M* for male. However, sometimes the gender was unknown, which here is called *UNK*. This analysis shows if one of the genders are more common among the very major errors or the unclassified samples. The distribution comparisons for AMP is shown in Figure 3.8. The differences are quite small, meaning a specific gender did not seem to be more common.

3.3.3.3 Country

Antibiotics were not equally common in all countries and for some countries only a small number of antibiotics are reportedly used. As seen in Figure 3.1b antibiotic resistance is more spread in some countries. The result of investigating the distribution of country can be seen in Figure 3.9.

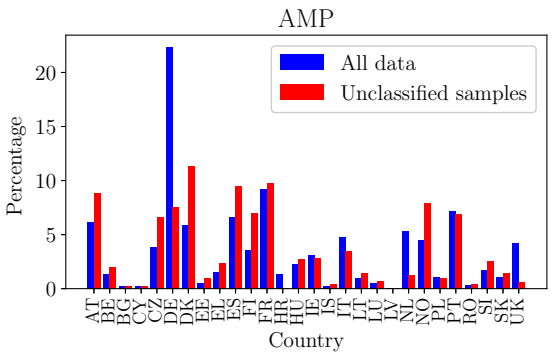


(a) The unclassified samples analysed in regard with gender. Comparison of gender's distribution on the whole data set and the distribution of the unclassified samples. The differences between the two groups are very small.

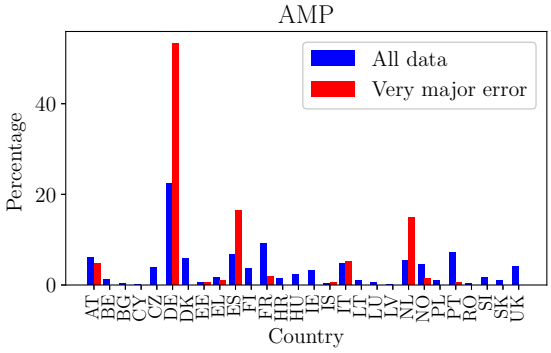


(b) The very major errors analysed in regard with gender. Comparison of gender's distribution on the whole data set and the distribution of the unclassified samples. The distributions look very similar.

Figure 3.8: Analysing if genders affect the probability of a sample being correctly classified. The figures present the distribution of gender in the whole data set for AMP compared with the unclassified samples and the very major errors, respectively. The classification limit is set to 1 unit of length symmetric around the ratio.



(a) The distribution of country on the whole data set compared with the distribution for the unclassified samples. Note how less common Germany, *DE*, is among the unclassified samples than in the whole data set for AMP.

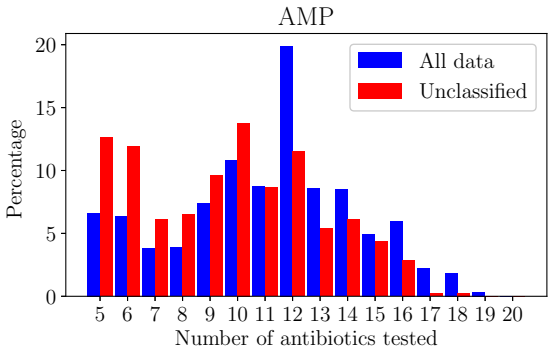


(b) The distribution of country on the whole data set compared with the very major error samples. Note that Germany, *DE*, who was less common in among the unclassified samples, now is more common among the very major errors.

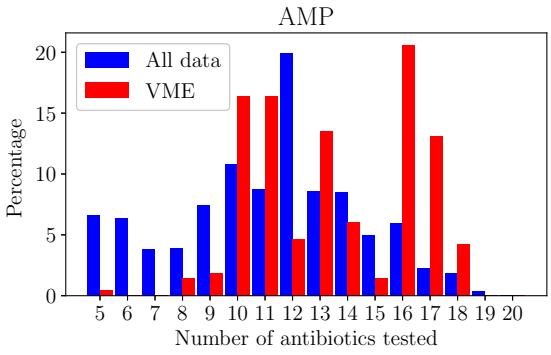
Figure 3.9: This figures investigate the countries influence on a sample being classified correctly. The distributions for AMP differs for most countries, but usually very little.The classification limit is set to 1 unit of length symmetric around the ratio.

3.3.3.4 Number of antibiotics tested

The last parameter investigated was the number of antibiotics tested on a sample. An hypothesis is that with only a few antibiotics tested the model has less information and therefore has a harder time predicting. The result of this analysis can be seen in Figure 3.10.



(a) Here, the distribution of the number of antibiotics tested on every sample for the whole data set is compared with the distribution in the set of unclassified samples. Note that the distribution for the unclassified samples are shifted to the left. This means, in general, that samples with less antibiotics tested had a higher probability of not being classified.



(b) The distribution of number of antibiotics on the whole data set compared with the very major error samples. 'VME' stands for very major error. Here, testing many antibiotics is more common among the very major errors compared with the whole data set.

Figure 3.10: Comparing distributions of the number of antibiotics tested for AMP. The distributions in the group of unclassified samples and the group of very major errors are compared with the distribution in the whole data set, respectively. The classification limit is set to 1 unit of length symmetric around the ratio.

3.3.4 Analysis of correlation

The correlation between the input and output of the model was analysed. The correlation plot for antibiotic AMP can be seen in Figure 3.11. Apart from the test results, also the age, the number of antibiotics tested and the number of antibiotics a samples was resistant to were added to the correlation plot. When calculating the correlation, all samples were considered, even those where an antibiotic had not been tested. The number of antibiotics tested on a sample and the number of antibiotics the sample was resistant to were particularly interesting. Theses variables together with age and the test result for an antibiotic are shown in a smaller correlation matrix that can be seen in Figure 3.12.

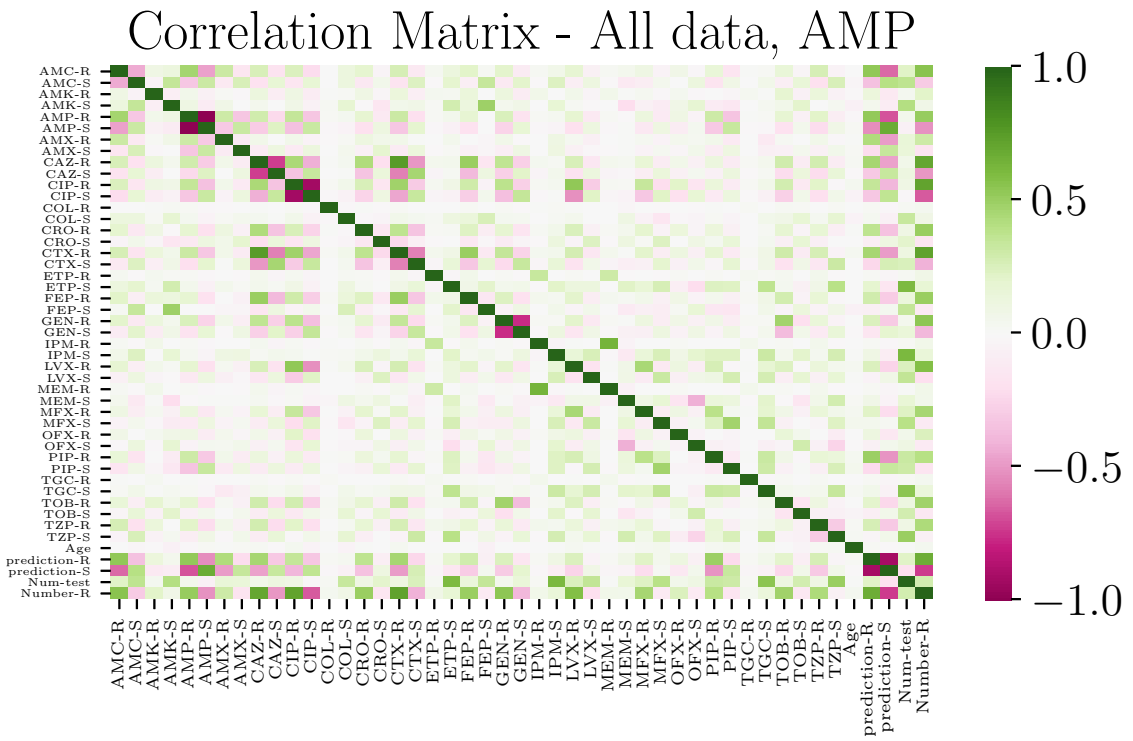


Figure 3.11: Correlation plot of all input and output of the neural network for AMP. The variable age, the predictions, *R* and *S*, the number of other antibiotics tested on a sample and the number of resistant test results are also shown. In the figure it is clear that some test results, antibiotics, affect the predictions, *prediction_R* and *prediction_S*, more than others. Note the very low correlation between age and all other variables.

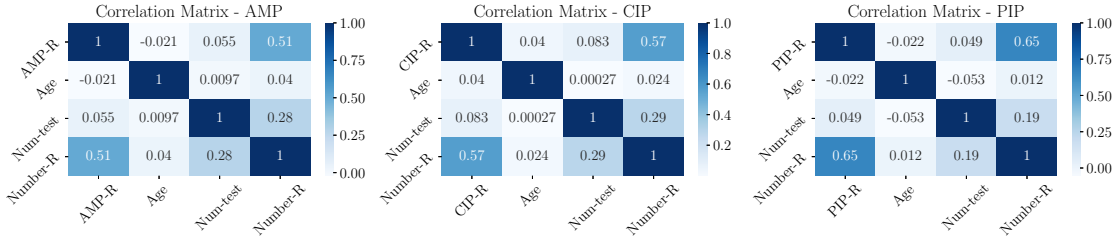


Figure 3.12: The correlation of some variables for AMP, CIP and PIP. As in Figure 3.11, the correlation between age and the other variables are very low. On the other hand, the correlation of number of antibiotics a sample is resistant to and the test result for the given antibiotic is over 0.5 for all three antibiotics.

3.4 Unique classification limits for antibiotics

As mentioned before, unique limits for classification were produced. The limits were set so that the very major error rate for an antibiotic did not exceed 5 %. In Table 3.4 the unique limits are listed. When fixating the very major error rate the number

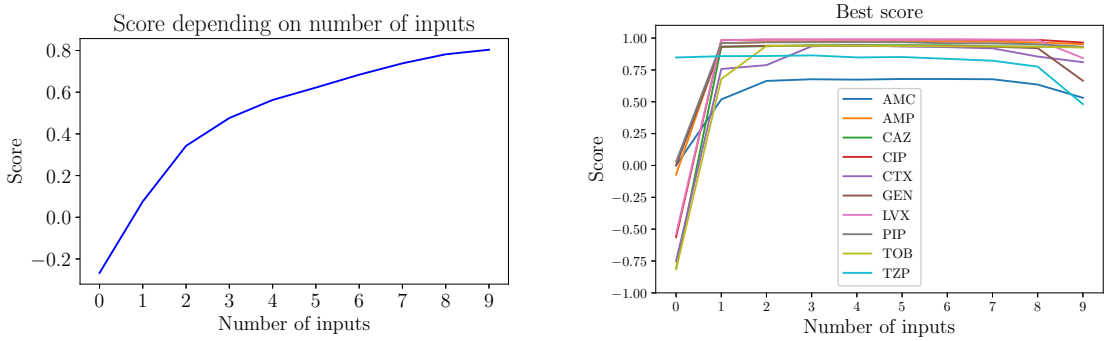
of unclassified samples and major errors were not considered. They are therefore also listed in the table. Other limits were also produced to not exceed 1 % very major error and the result can be seen in Table C.1 in Appendix.

Table 3.4: Classification limits for every antibiotic, where the limits are set at the ratio plus minus delta. For every limit there is a percentage of unclassified samples and errors. Note how the percentage of unclassified samples differs between different antibiotics, for example MEM has 100 % while PIP has 0 %. Also, note that the percentage of major error are less than 5 % for most antibiotics.

Antibiotic	Ratio	Delta	% unclassified	% very major	% major
AMC	-0.31	0.29	16.15	4.96	9.97
AMK	-2.02	1.58	26.71	4.94	0.61
AMP	0.06	0.48	36.57	4.46	0.61
AMX	0.08	0.40	16.56	4.91	0.77
CAZ	-0.98	0.00	0	2.88	3.71
CIP	-0.55	0.65	32.69	5.00	1.90
COL	-2.01	2.30	79.64	4.81	0.02
CRO	-0.91	0.00	0	2.50	3.00
CTX	-0.85	0.19	3.327	4.98	3.32
ETP	-2.72	2.90	20.62	4.04	0.02
FEP	-0.97	0.00	0	2.02	4.31
GEN	-1.03	0.55	23.66	4.95	4.67
IPM	-3.11	1.18	98.94	1.67	0.33
LVX	-0.52	0.00	0	2.23	1.13
MEM	-3.17	0.14	100	0.00	0.00
MFX	-0.43	0.62	22.81	4.66	0.31
OFX	-0.67	0.51	12.88	4.84	0.43
PIP	-0.03	0.00	0	0.28	3.33
TGC	-2.81	2.39	99.90	0.00	0.05
TOB	-0.99	0.03	0.46	4.97	5.65
TZP	-1.08	0.28	18.07	4.98	10.1

3.5 Performance of the model

When analysing the model's performance ten antibiotics were chosen. Only predictions of these antibiotics were considered and the remaining antibiotics were set as *not tested* in the input data frames. From these ten antibiotics different combination of inputs were made and the score for all combinations on all antibiotics, respectively, was calculated. The combinations were grouped by the number of antibiotics in the combination. The mean value of all scores in each group was then calculated and the result can be seen in Figure 3.13a. Then, the best score for all number of antibiotics tested in the input was analysed and done individually for all antibiotics. The result is shown in Figure 3.13b.



(a) The mean performance score as a function of the number of antibiotics as input of the model. The model gets, in general, a higher score when having more inputs.

(b) Best score for every number of antibiotics as input individual for every antibiotics. There are some differences between the antibiotics. However, many antibiotics have a really high score for one to eight antibiotics as inputs.

Figure 3.13: Two measurements of the model's performance with regard to the number of antibiotics as input to the model.

3.5.1 The best worst-case scenario

For every number of antibiotics as inputs a combination and the corresponding score was found as the best worst-case scenario. The method used is explained in Section 2.5.1 and the result can be seen in Table 3.5.

Table 3.5: This table presents the best worst-case scenario of each number antibiotics as a score. At the right the combination corresponding to each score is shown.Note how high the score are for most number of inputs and also how more common some antibiotics are in the found combination.

	Score	AMC	AMP	CAZ	CIP	CTX	GEN	LVX	PIP	TOB	TZP
9	0.9649	AMC	AMP	CAZ		CTX	GEN	LVX	PIP	TOB	TZP
8	0.9536	AMC		CAZ		CTX	GEN	LVX	PIP	TOB	TZP
7	0.9287	AMC			CIP	CTX	GEN		PIP	TOB	TZP
6	0.9034	AMC			CIP	CTX			PIP	TOB	TZP
5	0.8374	AMC			CIP	CTX			PIP	TOB	
4	0.7578	AMC			CIP				PIP	TOB	
3	0.4783	AMC			CIP			LVX			
2	0.2813		AMP			CTX					
1	-0.4294								PIP		

4

Discussion

In this chapter there will be a discussion of the results and findings. It will start with the results found in the previous section and reflect over the chosen method. Lastly some improvements and further work will be presented.

4.1 The data

The amount of data available for this project was large. However, the used data only considered one bacteria, *E. coli*, from one year. Using only data from one bacteria is a reasonable limitation, since considering several bacteria would need an investigation whether different bacteria could use the same neural network or not. If not so, the number of networks would increase by about twenty for every new bacteria.

After filtering the data 116 541 samples were left. All antibiotic were not tested on every sample, the number of samples per antibiotic can be seen in Table 3.1. The difference in number of samples are up to one order of magnitude. In this table the number and percentage of resistant samples can also be seen. Again, the differences between the antibiotics are big, from barely 0.1 % resistant samples to over 50 %. The differences are important to have in mind later when analysing the model's performance. The next thing looked into was the number of samples in relation to countries. In Figure 3.1a the distributions of the number of samples per country is shown. France and Germany stand out as having a lot more samples than other countries and they use nearly all kinds of antibiotics, compared with Denmark that only use seven antibiotics. There are some antibiotics, for example MFX, ofx and PIP, that are used only in a few countries. In Figure 3.1b that shows the ratio of resistant samples in different countries, it is clear that the percentage are higher for some antibiotics, which can be substantiated by Table 3.1, but also that resistance is more widely spread in some countries. For example resistance to TOB seemed to be more common in Bulgaria and Malta than the rest of the countries. However, from Figure 3.1a both these countries seems to have a very low number of samples, which then means the colours in Figure 3.1b are not always reliable.

4.2 The architecture of the neural network model

During the analysis of the architecture of the neural network two approaches were used. The first one was a simplified and greedy algorithm. The result for this approach can be seen in Table 3.2. The cost function, *Categorical crossentropy*, decreases when using more layers, but is quite stable after 3 layers. Nevertheless, the best performance is with 4 layers, even tough the difference is very small. The second approach was to test some arbitrary models with larger layers in the beginning of the network and in with fewer layers. The result of the best model can be found in Table 3.3. This was the one network chosen to be used in the rest of the thesis. The cost function for the first layers in the chosen model is a bit better than for the first layer of the greedy algorithm. After the first layer the performance goes up and down a bit until the end where the cost function takes the lowest value of both approaches.

4.3 Analysis of the outcome of the model

After a neural network model was chosen the outcome of the model was analysed. Firstly, only the predictions were looked at, how separated they were and if a transformation was possible to make them more separated. Secondly, how to classify the output was analysed. After the classification some samples were unclassified and some were missclassified. Therefore, these samples were looked into with regards to age, gender, country and the number of antibiotics tested. The last thing analysed was the correlation of some of the samples' variables. In this section the aspects and the results of this investigation will be discussed.

4.3.1 Predictions

Firstly, the model's prediction was plotted, both as it was and after logit transformation. The distributions for antibiotic AMP, CIP and PIP can be seen in Figure 3.2 and 3.3. It is clear from the figures that predictions of some antibiotics are very easy to separate, for example see PIP, and others are harder. For both AMP and CIP the predictions for the resistant and susceptible samples are overlapping. However the main part of the resistant samples are on the right hand side of the resistant ratio and the main part of the susceptible samples are on the left hand side. The reason for logit transforming the predictions was to stretch out this distribution. Comparing the two ways of visualising, it is less overlapping after the logit transformation.

The logit transformed predictions were also analysed per country, see Figure 3.4. Before looking at the different distributions it is important to notice the different proportion of samples in every country, for example Germany had 22756 samples but Latvia only had 25 samples. Regarding the separation of the predictions Germany (DE), again, stands out. The country seems to have nearly all samples predicted correctly. For Spain (ES) that have over 5 000 samples the separation is not so good and the resistant and susceptible samples are overlapping around the resistant ratio.

The conclusion is that the country parameter has an impact on how separated the predictions are.

4.3.2 Analysis of limits for classification

Since the predictions are not always distinctively separated, using different limits for classification was investigated. This was done both with and without the logit transformation of the predictions. However, the approaches were different in the two cases. Firstly, limits for the predictions without logit transformed was tried. The approach was to set the limits close to 0 and 1 respectively. The result for three antibiotics can be seen in Figure 3.5. The results are as expected, the number of errors is increasing and the number of unclassified samples are decreasing when the limits gets further from the ends. However, the result are different for different antibiotics, for example the percentage of errors differed one order of magnitude between AMP and PIP. The number of unclassified samples is expected according to the predictions figure, see Figure 3.2. It is not surprising that PIP has a lower number of unclassified than the others since its predictions are far more separated.

The second approach was to use the logit transformed predictions and, instead of starting from the ends, start from the resistant ratio and set symmetric limits around it. The result can be seen in Figure 3.6. Again, the result are different for the three antibiotics showed. When the distance is 0 from the ratio, naturally no samples are unclassified. An interesting observation is the low percentage of errors for PIP. Setting the limits further away from the ratio results in an decreasing the error rate. However, setting the limits so that there are no errors results in that over 50 % of the samples are unclassified for all three antibiotics.

For both approaches the percentage of unclassified samples and errors was studied instead of total number and the reason for this is that the three antibiotics do not have the same amount of samples. Comparing the two approaches the result are quite similar for both AMP and PIP. However, looking at 5 % very major errors in Figure 3.5 for CIP, it corresponds to about 40 % unclassified samples without the logit transformation but only 30 % for the logit transformed, in Figure 3.6. The difference is not very big but a small advantage for the logit transformed. Therefore, this approach was chosen.

4.3.3 The unclassified samples and the very major errors

To understand the model more, an investigation of the unclassified samples and the very major errors was made. The question here was whether some other factors from the samples would lead to the sample being wrongly predicted or unclassified. The method used was to plot the distribution of all samples and compare with the unclassified samples and the errors respectively. During this analysis the limits for the predictions were set to one unit of length from the resistant ratio on each side.

4.3.3.1 Age

The first variable investigated was age. The result for AMP can be seen in Figure 3.7. Starting with the unclassified samples, the difference was very small, only a minor decrease around the age of 80. For the very major errors, there were only 214 samples and since the samples were group by years and not in bigger intervals, like for example age $30 - 40$, the amount was to small to say something definite. Nevertheless, the distribution of the age of the very major errors seemed to be quite similar to the distribution of all data. Since the differences were so small in both cases, the age appeared not to influence the prediction.

4.3.3.2 Gender

A patient's gender was divided in three classes: female, male and unknown. The result for AMP can be seen in Figure 3.8. From the figure approximately 10 % of the samples had the gender set to unknown. For the rest of the samples, female was more common than male. For both the very major error and the unclassified samples the class unknown was less common, but in both cases the difference was very small. The difference for the classes female and male was only a few percentage points. The conclusion is that gender was not a factor in which samples got unclassified or missclassified.

4.3.3.3 Country

As mentioned before, the number of samples per country and antibiotics varied a lot, see Figure 3.1a. The variable country was investigated as the previous variables age and gender and the result for AMP is found in Figure 3.9. For the unclassified samples, most countries were roughly equally common as in the whole data set, but some of the countries were not, for example Denmark and Germany. Denmark was more common among the unclassified samples and Germany was much less common. One could argue that samples from Germany was easier to predict since less of them were unclassified. However, for the very major error Germany was a lot more common than for all data and therefore it is important to look at both figures. The big differences for Germany and Denmark could indicate that the country affect the model's prediction.

The reason for the differences in distributions could be the accuracy in the different countries reporting. If so, adding country as an input variable could confuse the model if the country becomes more thorough. Also, if one hospital in a country is much more thorough than the others, then the output of the model will probably be less accurate for that hospital. Another argument is that bacteria does not care which country the patient are from. Travelling is very common today and a bacterial disease can be caught anywhere. Therefore, only the test result as input of the model would probably be less biased than using a country variable as well.

4.3.3.4 Number of antibiotics tested

The last variable to investigate was the number of antibiotics tested on each sample. The result for AMP can be seen in Figure 3.10. A hypothesis was that if giving less information to the model, i.e. fewer test results, it would be harder for the model to predict correctly. This seemed to agree for the unclassified samples, where particularly 5 and 6 number of antibiotics tested stood out as much more common. However, for the very major error it was the opposite, 5 and 6 number of antibiotics tested were barely represented at all. One reason for this could be the low amount of samples in total having over 16 antibiotics tested. This gives less data to train on, which then makes it harder for the model to predict that kind of samples.

4.3.4 Correlation of the samples' variables

The result for AMP can be seen in Figure 3.11. Considering the test results, most of them have a correlation close to 0. Looking at *CAZ_R* and *CTX_R* the correlation is almost 1 meaning that the test results for these two often were the same. A problem with the plot is that it also takes the *not tested* cases in account. This could be why many boxes are lightly coloured, indicating a low correlation. Regarding the predictions, the correlation between the predictions and the test results is bigger than between only the test results. The correlation tells how much different test results impact the predictions and in the plot there are some antibiotics that have a much higher correlation than the others.

Three variables were chosen for further investigation, these were age, the number of antibiotics tested and the number of resistant test results. These three together with the test result of the resistant variable were plotted in a correlation matrix. The result for three antibiotics can be seen in Figure 3.12. The highest correlation is between the resistant variable and the number of resistant test results. This means that the probability of a sample being resistant to a specific antibiotic increase if it is resistant to several others. The correlation between the number of other antibiotics tested and the resistant variable is between 0.2 and 0.3 for all three examples. Comparing with the number of resistant test result the latter have more impact on the outcome. For the variable age, the correlation is not so high with any of the other variables. The reason for this could be that age was a discrete number between 0 and around 115 compared to the other variables that were continuous values from 0 to 1. Also, the age is well distributed which makes it harder to find a correlation.

4.4 Unique limits for classification

Because the separation of the predictions varied a lot between different antibiotics unique limits were produced. These limits were set so that the very major error rate did not exceed 5 %. The limits can be seen in Table 3.4. The limits differ a lot between 0 for PIP up to 2.90 for ETP. As mentioned before, a reason for the variation is the variation in the separation of predictions. The method also only consider the number of very major errors meaning that the limits could be very

different if the number of unclassified samples or the total number of errors were considered instead.

For each unique limit the percentage of errors and unclassified samples can also be seen in Table 3.4. Even here, the results varies a lot, especially among the unclassified where some has 0 % unclassified samples and others up to 100 %. An intresting thing is that the percentage of major error are almost always less than 5 % and often a lower rate than the very major error. This implied that setting a fix limit for the very major error rate did not make the major error rate increase a lot. The low number of major error and very major error can sometimes be explained by the high number of unclassified samples, which means one can not only focus on the errors.

4.5 Performance of the model

When analysing the performance of the model the first thing investigated was the importance of the number of antibiotics tested. The score for all combination was measured. For every number of antibiotics as input the mean value of the score were calculated and plotted, see Figure 3.13a. From the figure it is clear that having more test results as input makes the model performed better. However, the increasing faded from the middle and upwards. The conclusion is that having more inputs is better for the model.

Then, the differences between the performance of the antibiotics was investigated. This time, only the best score for every number of antibiotics was considered. The result can be seen in Figure 3.13b. For zero number of inputs the model did not perform well for any antibiotic except TZP. When adding just one input all antibiotics increase their performance and stays on the same level until 9 inputs where all decreases slightly. A naive way of thinking would be that more information make it easier for the model to predict. It seem true for the mean, see Figure 3.13a, but not when taking the best combination. It could be because test results from some antibiotics confuses the model and that it is better to not know some test results.

4.5.1 The best worst-case scenario of the model

The last investigation was to find the best worst-case scenario. How this was computed is explained in Section 2.5.1. For every number of inputs the best worst-case scenario was found with regards to score. In Table 3.5 the result is shown as score and which combination of inputs it belongs to. The increase of these scores was quite similar to the increase in Figure 3.13a, regarding the stagnation at the end. According to the table, the model performs quite well from 4 antibiotics as input and really well from 6 inputs. Looking at the combinations it is clear that some antibiotics are more common than others, for examples AMC are in all but two combinations. An antibiotic that is hard for the model to predict would probably not be the one predicted in the combination for the best worst-case scenario, therefore the probability of that antibiotic would be in the combination instead increases.

From Figure 3.13b AMC had the lowest performance almost all times which could be the reason for the high occurrence in the combinations of the best worst-case scenarios.

4.6 Improvement and further work

Some things in the method of this thesis were simplified and some things could have been done in several other ways. Therefore, a discussion about some improvements follows here, starting with the data set and the network and ending with the classification of the predictions.

4.6.1 The data

When extracting the data some antibiotics were removed due to too few test results. However, in this process the number of resistant test results were not considered. For example TGC has only 56 resistant samples, equal to 0.08 % of TGC's data set, which perhaps is too few for a model to learn from. There are four antibiotics with less than 100 resistant samples, which all represents less than 1 % of the samples of that antibiotic. Predicting these four antibiotics could result in the model setting all input samples to susceptible and still get a very high accuracy. Therefore, developing models for these antibiotic could be challenging. However, this does not necessarily mean that they should not be a part of the input for the other antibiotics. A pattern, where being resistant to one of these four antibiotics always implies a resistance to another antibiotics, could still exist which then should be an important feature of the input. Therefore, the importance of different antibiotics in the model should be studied more.

An unstudied and possible issue of the data is the accuracy. Test results are reported by staff at a hospitals and is voluntary. Looking at the number of samples from each country and compared with the population size it indicates that not all test results are reported. One could only speculate in which test results that are reported. To get a more accurate training data it is important that all test result is reported regardless specimen, hospital or country. Another thing about the data is the lack of information for some categories. For example in Figure 3.8 nearly 10 % of the AMP samples do not have information about the patient's gender.

4.6.2 The neural network model

After extracting relevant parts of the data the data frames of input and output were constructed for each antibiotic. In this thesis the test results were represented by two variables, as explained in Table 2.2. This could have been done in several other ways, for example having three variables instead where the third represent *not tested*. Another way could be to only have one variable where 1, 0 and -1 represents resistant, not tested and susceptible respectively. One reason for not using the last option is that the different classes not necessarily are ordered with equal separation. A less naive approach could then be to set "not tested" to a number between -1

and 1 that corresponds to the resistant ratio. Nevertheless, such a model, or the one with three variables, could be a better approach than using two variables and the question is therefore in need of further investigation.

The approach of analysing different number of layers and neurons in the neural network is very simplified. However, the result from both this approach and when testing some larger, arbitrary networks is nearly the same, see Table 3.2 and Table 3.3. Since many parameters are fixated during the investigation of the networks architecture, it would be interesting to analyse the model more regarding these parameters. Two examples of interesting parameters are activation function and drop out rate, which is when some neurons in the network is dropped out during the training to reduce overfitting. The most important parameter would be the cost function. The output of the model is two continuous variables between 0 and 1 that represented the probability of the sample being resistant or susceptible. Since the output hardly ever is exactly 0 or 1, every sample will add to the cost. However, using the method with logit transformed predictions and limits which decides if a sample is resistant or not, see Section 2.4, could be a better way to compute the cost. The output would then be resistant, susceptible or unclassified. The cost function should then punish the unclassified and the errors. For example by the scoring system that is used when evaluating the performance of the model, see Section 2.5.

Along with the cost function is the question on how to estimate the performance of the model. Is it only important to minimise the errors, or are very major error worse than major errors and if so how much worse. Now, the cost function only considered how far from the truth the prediction is. However, a very major error have a more severe consequence than a major error. Therefore it would be interesting to investigate how the two types of errors could be penalised differently and how that would impact the performance of the model. Creating a new cost function that takes this into account would probably improve the performance.

When the architecture was investigated, only the antibiotic AMP was considered. At that time the variation in difficulty level for separating the prediction was not known, but after the model was developed the fact was clear: some antibiotics are harder to predict than others. Therefore, it would be interesting to use other antibiotics for this part and see how the architecture varies. Perhaps a better idea would be to have an unique architecture for each antibiotic.

4.6.3 Limits for classification

The unique limits discussed in Section 4.4 only considered the percentage of very major errors and was sat symmetric around the resistant ratio. From the prediction figures, for examples see Figure 3.2, the susceptible samples seemed to be easier to predict correct then the resistant samples. Setting limits symmetric could give more unclassified samples than errors it removes. Therefore, it would be interesting to let the right limit be set so that the major error rate does not exceed 5 %. This means the new limits would not be symmetric.

5

Conclusion

A general conclusion during the work with this thesis is how different the model behave for different antibiotics. It starts with how well separated the predictions are which then leads to different classification limits and percentage of errors. However, for most antibiotics the model performs very well having unique limits and with regards to the number of unclassified samples and errors. When measuring the best worst-case scenario the results implies that the model works very good when having four or more test results as inputs.

The purpose of this thesis is to investigate the possibility of speeding up the process of finding a functional antibiotic treatment. From the results it appears that using a model like this one could be a helpful tool in this process. There are some things that could improve the model, such like training the model with a cost function that punish very major errors more than major errors and trying unique architectures for different antibiotics. A conclusion is therefore to develop the model more to get even better results.

6

Disclaimer

The views and opinions of the authors expressed herein do not necessarily state or reflect those of the ECDC. The accuracy of the authors' statistical analysis and the findings they report are not the responsibility of ECDC. ECDC is not responsible for conclusions or opinions drawn from the data provided. ECDC is not responsible for the correctness of the data and for data management, data merging and data collation after provision of the data. ECDC shall not be held liable for improper or incorrect use of the datam

Bibliography

[1] World Health Organization. Antibiotic resistance. https://www.who.int/news-room/fact-sheets/detail/antibiotic-resistance, February 2018.

[2] ReAct. Resistance mechanisms. https://www.reactgroup.org/toolbox/understand/antibiotic-resistance/resistance-mechanisms-in-bacteria/.

[3] ReAct. Resistance mechanisms. https://www.reactgroup.org/toolbox/understand/antibiotic-resistance/multidrug-resistant-bacteria/.

[4] J. O'Neill. Tackling drug-resistant infections globally: Final report and recommendations. https://amr-review.org/sites/default/files/160525_Final%20paper_with%20cover.pdf, May 2016.

[5] TESSy The European Surveillance System. Antimicrobial resistance (amr) reporting protocol 2018. https://www.ecdc.europa.eu/sites/default/files/documents/EARS-Net%20reporting%20protocol%202018.%20docx.pdf, February 2018.

[6] Hampus Lane. Predicting Antibiotic Resistance Phenotypes using Neural Networks. Master's thesis, Chalmers University of Technology, Sweden, 2020.

[7] L. J. LandauJ. G. Taylor. *Concepts for Neural Networks*. Springer, London, 1998.

[8] Xin Qian and Diego Klabjan. The impact of the mini-batch size on the variance of gradients in stochastic gradient descent, 2020.

[9] Hans Christian Jung, Johannes Maly, Lars Palzer, and Alexander Stollenwerk. Quantized compressed sensing by rectified linear units, 2019.

[10] Behnam Asadi and Hui Jiang. On approximation capabilities of relu activation and softmax output layer in neural networks, 2020.

[11] Xinshao Wang, Elyor Kodirov, Yang Hua, and Neil Robertson. Instance cross

entropy for deep metric learning, 2019.

A

Additional ways to analyse limits for classification of the predictions

When the predictions are analysed two types of classification limits are used, as described in Section 2.3.2. However, the classification could be done in several ways. Two other methods are therefore investigated. The first one proceeds from the resistant ratio and move fix steps from it. This method is similar to the one used in the report but without the logit transformation. In Figure A.1 the result for AMP, CIP and PIP can be seen. The ratio is unique for all antibiotics which mean that some antibiotics could not increase their step length as much as others.

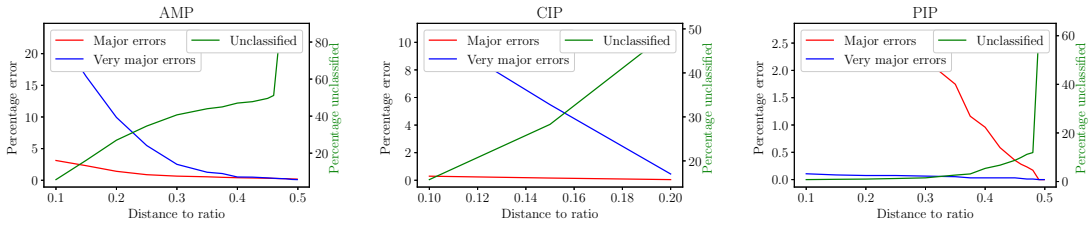


Figure A.1: Percentage of unclassified samples and errors depending on the distance to the ratio for three antibiotics: AMP, CIP, PIP. Note how difference between the antibiotics.

The other method also proceeds from the resistant ratio, but uses fix percentage steps of the distance between the ratio and the edge, meaning 0 and 1. The result for AMP, CIP and PIP can be seen in Figure A.2.

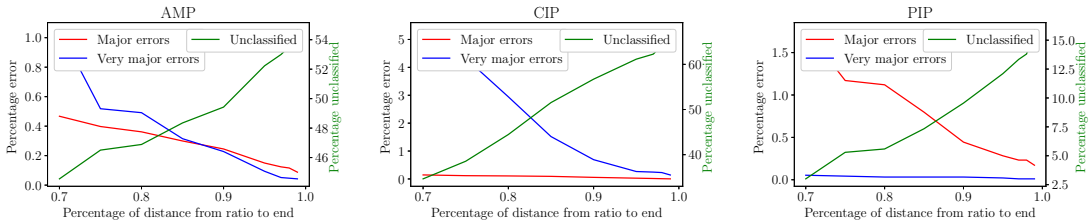


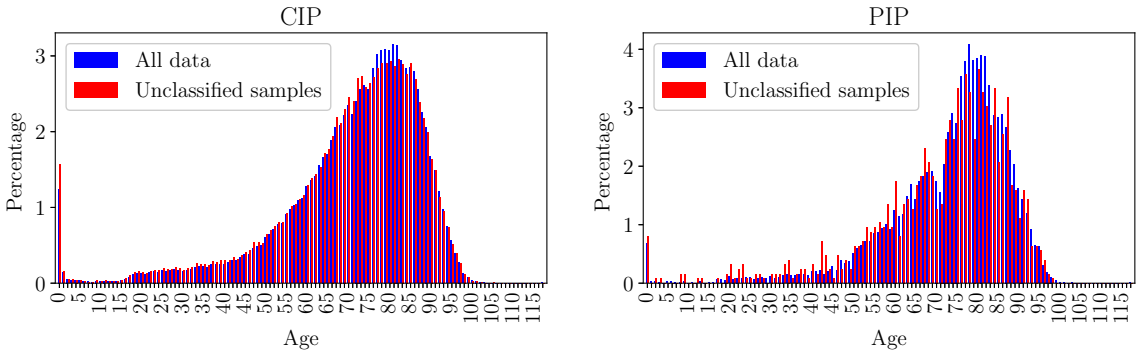
Figure A.2: Percentage of unclassified samples and errors as a function of the percentage of distance from the ratio to the end. Here is the results for antibiotics AMP, CIP, PIP. Interesting is the differences between the antibiotics.

B

Additional figures for analysing unclassified samples and errors

In Section 3.3.3, the results from analysing unclassified samples and errors for antibiotic AMP are presented. In this chapter the results for two more antibiotics are presented, CIP and PIP. The number of unclassified samples are 57466 and 1348 for CIP and PIP, respectively. For number of very major error, CIP have 235 which is a very major error rate of 0.97 %. However, PIP have 3 very major errors, which is a 0.03 % rate. Therefore, no conclusions can be made from the results of PIP's very major error.

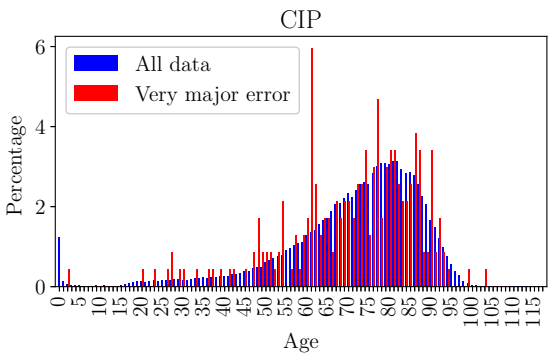
B.1 Age



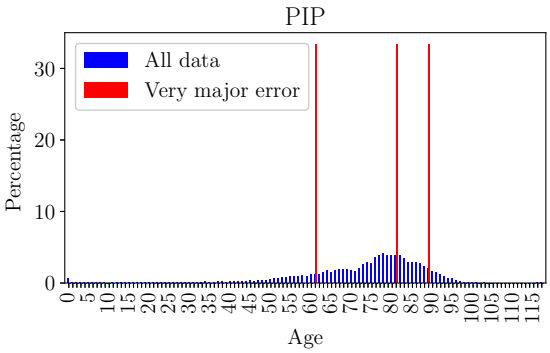
(a) Differences in the distributions are very small, with only a minor decrease for the unclassified samples around the age of 80.

(b) The distribution of PIP's data is not as smooth as CIP's data, but here also, the unclassified samples have a minor decrease around 80.

Figure B.1: The distribution of age on the whole data set compared with the distribution of unclassified samples for antibiotic CIP and PIP. The classification limit is set to 1 unit of length around their respective resistant ratio.



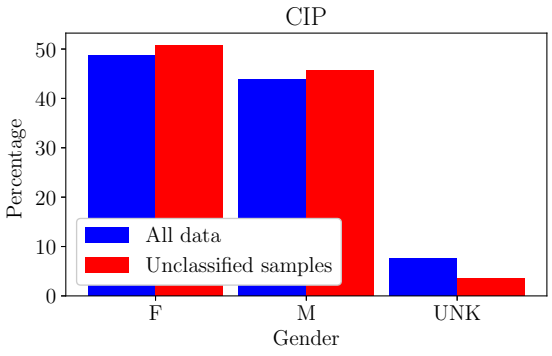
(a) The distribution of the very major errors follows the distribution of all data quite well.



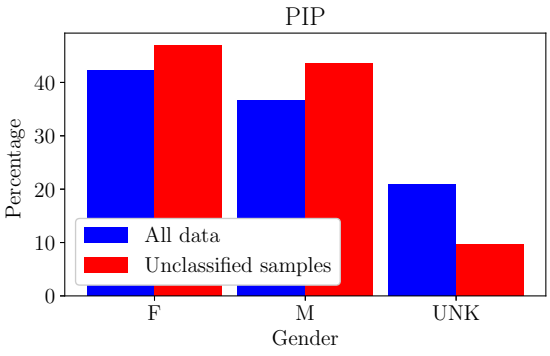
(b) The number of very major error for PIP is only 3, making it hard to tell if the distribution follows the whole data set.

Figure B.2: The distribution of age on the whole data set compared with the distribution of very major errors for antibiotic CIP and PIP. The classification limit is set to 1 unit of length around their respective resistant ratio.

B.2 Gender

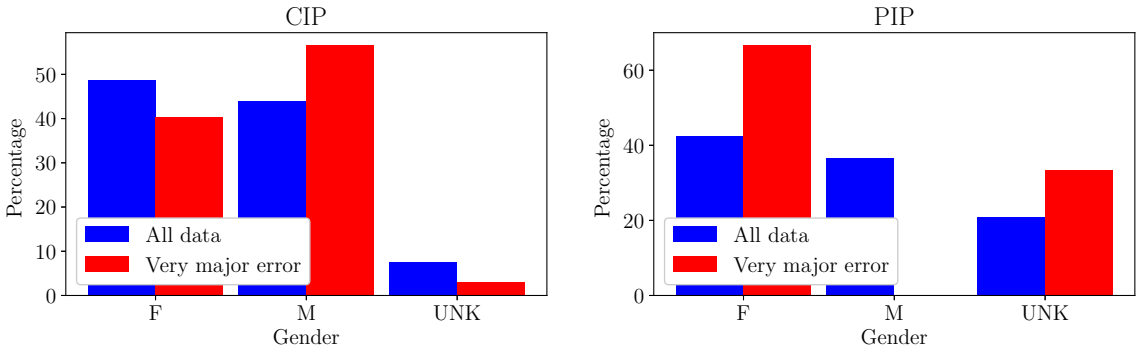


(a) The differences between the distribution of the whole data and the distribution of the unclassified samples are very small, around 3 percentage points for each gender.



(b) For PIP, the differences are around 5 percentage points for both female, *F*, and male, *M*. That results in a 10 percentage points difference for the unknown, *UNK*, samples.

Figure B.3: The distribution of gender on the whole data set compared with the distribution of unclassified samples for antibiotic CIP and PIP. The classification limit is set to 1 unit of length around their respective resistant ratio.

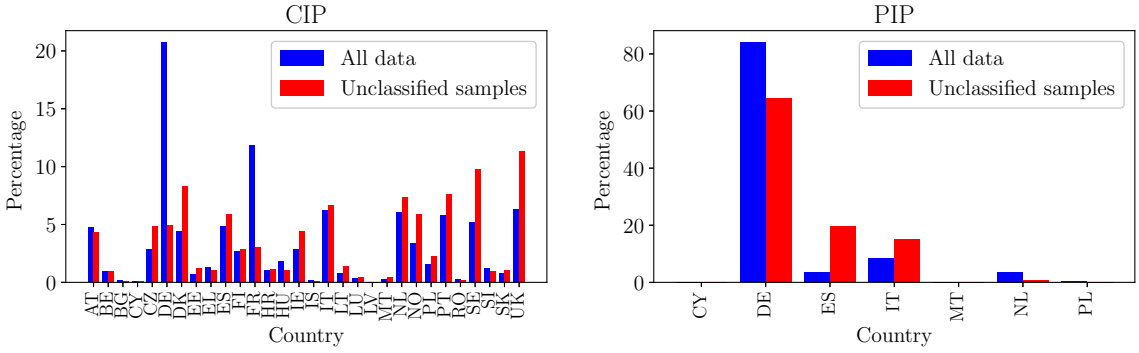


(a) Here, the differences are bigger than for the unclassified samples, especially for male, *M*, which has over 10 percentage points

(b) The number of very major error for PIP is only 3, making it hard to tell if the distribution follow the PIP's whole data set.

Figure B.4: The distribution of age on the whole data set compared with the distribution of very major errors for antibiotic CIP and PIP. The classification limit is set to 1 unit of length around their respective resistant ratio.

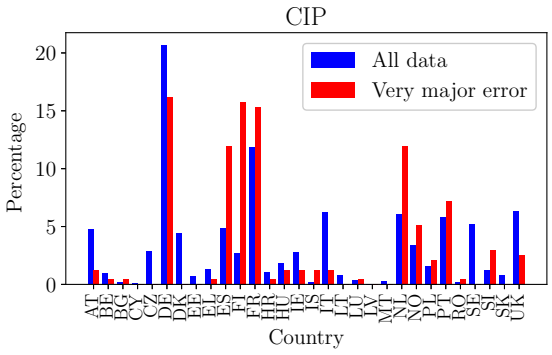
B.3 Country



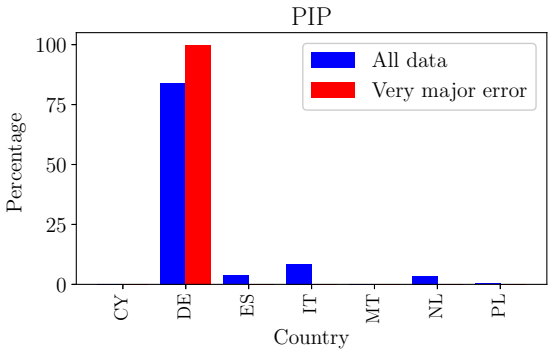
(a) The differences are quite big for some countries, especially Germany, *DE*, and France, *FR*, are less common amongst the unclassified samples.

(b) Mainly, Germany, *DE*, is less common amongst the unclassified samples while Spain, *ES*, and Italy, *IT* are more common.

Figure B.5: The distribution of country on the whole data set compared with the distribution of unclassified samples for antibiotic CIP and PIP. The classification limit is set to 1 unit of length around their respective resistant ratio.



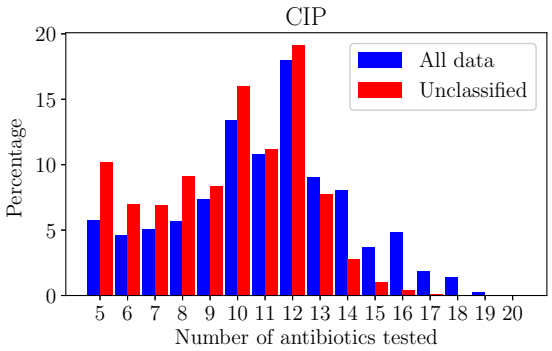
(a) Again, most countries have a differences between the two groups. However, in some cases, for example *UK*, the country is more common in the unclassified group but less common in amongst the very major errors.



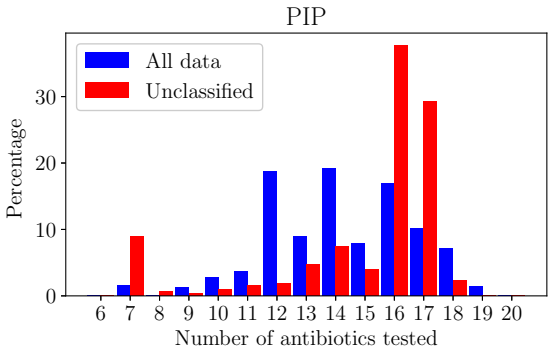
(b) The number of very major error for PIP is only 3, making it hard to tell if the distribution follow the whole data set.

Figure B.6: The distribution of country on the whole data set compared with the distribution of very major errors for antibiotic CIP and PIP. The classification limit is set to 1 unit of length around their respective resistant ratio.

B.4 Number of other antibiotics tested

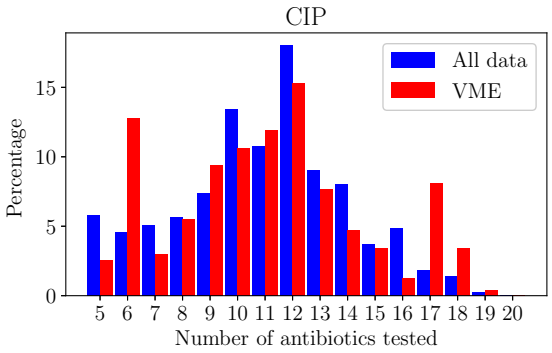


(a) For CIP, the distribution is shifted to left compared with the whole data set.

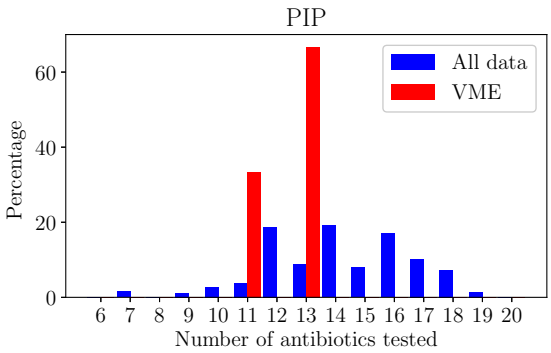


(b) Both 16 and 17 number of antibiotics as input are significantly more common in the unclassified group.

Figure B.7: The distribution of number of other antibiotics tested on the whole data set compared with the distribution of unclassified samples for antibiotic CIP and PIP. The classification limit is set to 1 unit of length around their respective resistant ratio.



(a) For most number of inputs, the differences are quite small. However, both 6 and 17 are more common amongst the very major errors.



(b) The number of very major error for PIP is only 3, making it hard to tell if the distribution follow the whole data set.

Figure B.8: The distribution of number of other antibiotics tested on the whole data set compared with the distribution of very major errors, here called *VME*, for antibiotic CIP and PIP. The classification limit is set to 1 unit of length around their respective resistant ratio.

C

Additional results having other classification limits

In Section 3.4, the classification limits when not letting the very major error rate not exceed 5 % is shown, see Table 3.4. However, one could argue that a model having 5 % very major errors could be to much errors. Therefore, the same procedure was done, but for a very major error rate of 1 % The result, as in the limits but also the percentage of errors and unclassified samples, can be seen in Table C.1.

Comparing with the result having 5 % very major errors, the obvious thing is that the limits now are further away from the ratio. This means, the percentage of unclassified samples have increased. Here, 7 antibiotics can classify less than 50 %. This is more than a doubling compared with 5 % very major error case. However, the percentage of major error are very low, under 1 %, for most antibiotics. The same goes for the percentage of very major error, which actually are 0 for 5 antibiotics.

Taking the percentage of unclassified samples in regard, the conclusion is to not have this very strict limits. These limits results in to high number of samples not being classified. However, it could be possible to use if the model got better on separating the resistant samples apart from the susceptible ones.

Table C.1: Unique limits for all antibiotic models together with percentage of unclassified and errors. The ratio and delta is logit-transformed. Specifically, look at the high percentage of unclassified samples there are for many antibiotics.

Antibiotic	Ratio	Delta	% unclassified	% very major	% major
AMC	-0.31	0.70	34.76	0.91	3.59
AMK	-2.02	2.59	99.66	0.00	0.08
AMP	0.06	0.80	46.02	0.64	0.37
AMX	0.08	0.69	25.72	0.93	0.38
CAZ	-0.98	0.93	9.34	1.00	1.77
CIP	-0.55	0.97	51.61	0.99	0.84
COL	-2.01	2.97	95.58	0.53	0.01
CRO	-0.91	0.94	8.09	0.95	0.22
CTX	-0.85	1.01	26.36	0.99	0.25
ETP	-2.72	3.58	99.89	0.00	0.01
FEP	-0.97	0.58	3.54	0.99	2.93
GEN	-1.03	1.26	57.50	0.81	0.98
IPM	-3.11	1.32	99.38	0.00	0.32
LVX	-0.52	0.85	5.39	0.80	0.59
MEM	-3.17	0.14	100	0.00	0.00
MFX	-0.43	1.41	44.23	0.98	0.07
OFX	-0.67	1.43	39.69	0.96	0.04
PIP	-0.03	0.00	0.00	0.28	3.33
TGC	-2.81	2.39	99.90	0.00	0.05
TOB	-0.99	1.14	19.57	0.99	0.05
TZP	-1.08	1.11	44.05	0.95	0.51