Utrecht University

Bachelor Thesis Artificial Intelligence

Machine learning to diagnose and prognose schizophrenia: using three different approaches to examine the models' misclassifications

Myrthe Hemker
6007457

15 ECTS

Supervisor:
Dr. Hugo Schnack

Second reader:
Prof. dr. ir. Jan Broersen

July 1, 2020

Abstract

At present, machine learning models that diagnose and prognose schizophrenia are not used in clinical practice yet, since the performances of these models are not high enough. So far, very little research has been done on the misclassifications of machine learning models that diagnose or prognose schizophrenia. However, to improve the performances of the ML models, the misclassifications could be analysed in more detail in order to find possible reasons for incorrect classifications. In this thesis, three approaches are proposed to analyse the models' misclassifications. In the first approach, we analysed whether there were features that showed significant differences between the correct and incorrect classifications. The second approach determined whether the same or other participants are misclassified in various models, and the third approach analysed the influence of the model's threshold on the misclassifications. We have demonstrated these different approaches on two real-world datasets. The results showed that ML models that diagnose and prognose schizophrenia had features that reported significantly different values for the misclassified sample compared with the correctly classified sample. Comparing models demonstrated that there are specific participants that are frequently misclassified. Furthermore, our findings indicated that the incorrect predictions in the prognostic models could be caused by the definition of the remission criteria. These findings indicate that analysing the misclassified sample can help to identify aspects that influence the performance, which, eventually, can help to improve the performances of the machine learning models.

Contents

Chapter 1

Introduction

Of all psychiatric disorders, one of the most mysterious psychiatric disorder is schizophrenia (Kapur & van Os, 2009; Sun et al., 2009). It has a long history that exceeds over a century of publications (Kraepelin, 1893; Bleuler, 1911). The diagnostic criteria have since undergone multiple changes (Rittmannsberger, 2012). However, with even over a century of medical attention to the disease, many characteristics of the disease remain elusive. Although many indicators have been tried, so far, none of these has been definitively proven to possess the desired diagnostic performance (Jablensky, 2010). As diagnosis is concerned with the current state, the challenges become even more complicated in prognosis as it requires looking into the future. For example, up to one out of two patients who were initially diagnosed with schizophrenia may develop a wide variety of unfavourable disease outcomes, even if they underwent reputable pharmacological and psychosocial treatments (Koutsouleris et al., 2016).

An important factor for these challenges is that schizophrenia is very heterogeneous, which means that there is a broad variety in symptom presentation, functional outcome, and clinical course. This makes diagnosis and prognosis difficult (Messinger, 2013). An objective measure to diagnose and prognose schizophrenia would be valuable (Nieuwenhuis et al., 2012). This could improve clinical outcomes of patients by earlier detection of diagnosis and prognosis, resulting in possibly better treatment by earlier implementing target interventions such as optimising pharmacological treatment or assertive case-management (Mourao-Miranda et al., 2012). However, many decades of experience showed that without additional support, diagnosis and prognosis seem to be considerably difficult.

A powerful tool to support human diagnosis and prognosis is machine learning (ML). ML is an application originated from the artificial intelligence (AI) field that gives systems the ability to learn from data and to detect patterns in the data (Maity & Das, 2017). In the last decades, the interest of ML has grown tremendously, and also beyond AI, due to an exponential increase in the capability of computer systems to process and store data (Tandon & Tandon, 2018). Due to this increased capacity and effectiveness, the field of psychiatry has started to employ ML as well (Janssen, Mourão-Miranda, & Schnack, 2018). In this field, ML techniques attempt to diagnose and prognose psychiatric disorders (Schnack, 2020).

One major advantage of ML, compared to conventional statistics, is that it allows predicting at individual rather than group level, which is very valuable for clinical use (Orru, Pettersson-Yeo, Marquand, Sartori, & Mechelli, 2012). Besides, ML models are able to simultaneously evaluate multiple variables, also known as features, which allow identifications of associations between multiple data variables (Zhang, 2017).

Although these advantages seem very promising for clinical use, ML models that attempt to diagnose and prognose schizophrenia are not used in clinical practice yet, since the performances of these models are not high enough (Iwabuchi, Liddle, & Palaniyappan, 2013). To improve these performances, it would be valuable to understand how the ML model works and what type of misclassifications occur in a model (Liu, Wang, Liu, & Zhu, 2017). However, most users often treat these models as a black box due to its unclear working mechanism and incomprehensible functions (Fekete, 2013; Mühlbacher, Piringer, Gratzl, Sedlmair, & Streit, 2014). Nonetheless, without a clear understanding of the mechanism of the model and the factors that cause incorrect prediction, the development of high-performance models typically relies on a time-consuming trial-and-error process (Liu et al., 2017). More transparency of the systems could help to understand

and analyse the incorrect predictions.

So far, very little attention has been paid to identify the misclassified sample of the ML models, neither in ML models that diagnose or prognose schizophrenia, nor in psychiatry in general. Outside of the psychiatry, more research has been done on the misclassified sample. For instance, ML studies in medicines have categorised the models' misclassifications on the nature of the error by reviewing the misclassifications by two study investigators independently (Yadav et al., 2016).

More focus on the subjects that were diagnosed and prognosed incorrectly by the ML models could help to address the issue of the possible sources of the incorrect classifications (Di Carlo et al., 2019). For this reason, this thesis aims to explore who the misclassifications are, and what techniques can be used to identify the specifics of these misclassifications. It would be valuable to investigate whether the misclassifications have some specific features (Alsallakh, Hanbury, Hauser, Miksch, & Rauber, 2014). Furthermore, it would be valuable to analyse whether the same or different participants are misclassified by different models.

To tackle the aforementioned challenges, three different approaches are developed for this thesis. In the first approach, it is being evaluated whether certain variables report different values for misclassifications compared to the correct classifications. In the second approach, the misclassifications of the different models will be evaluated on their overlap in misclassifications. In the third approach, we analysed the influence of the threshold in classifiers on the misclassifications. These three approaches could help to understand the misclassifications and, eventually, could help to improve the ML models. An additional beneficial outcome, if this methodology is successful, is that it could be applied to tackle other psychiatric diagnostic and prognostic issues. It could even be helpful beyond psychiatry and might be employed in other medical and non-medical applications.

In the next chapter (Chapter 2), the theoretical background of the thesis is given. First, the traditional approach is reviewed. Subsequently, we elaborate on the three new approaches that result from our focus on misclassifications. To analyse the misclassifications with these approaches, ML models have been created. These ML models have been trained on two different real-world datasets obtained by University Medical Centre Utrecht. In chapter 3, we describe the two clinical datasets that are used. Furthermore, this chapter gives a description of the pre-processing procedure of the ML models, the selections of features used in training the employed ML algorithms, and how has been dealt with incomplete and unbalanced datasets. The ML models used are explicitly explained since understanding what happens in the model could help to understand when a participant is misclassified. The results of the analysed models using the three approaches are described in chapter 4. In chapter 5 of the thesis, the results are discussed. In the last chapter (chapter 6), the main conclusions are provided. Finally, three appendices are added to the thesis. The first appendix details on the theory that was applied in this study. In the second appendix, additional tables and figures are provided that contain more information on the data and feature selections that were used. The final appendix consists of a manual of the `R` scripts that were used to create the models and to analyse the misclassifications. These scripts could also be used for other ML studies.

Chapter 2

Approaches to examine the misclassifications

This chapter begins with discussing the current procedure to evaluate ML models that diagnose and prognose schizophrenia. This procedure focuses on evaluating the models on their performances. The chapter will then describe the approaches designed to identify specific characteristics of the participants that are misclassified by the ML models.

2.1 Current procedure: focus on performance measures

In the current procedure, ML models are evaluated based on their performances. A commonly used performance metric for classifiers is the classification accuracy, in which the number of correct classifications is divided by the total number of participants that have been classified. Although the accuracy is a wildly used metric, the accuracy can be misleading, especially when a dataset is imbalanced, meaning that one class occurs more in a dataset than the other class(es) (Kassraian-Fard, Matthis, Balsters, Maathuis, & Wenderoth, 2016). For instance, suppose that a binary classifier is trained and tested on datasets in which 80 percent had a positive class label, and only 20 percent had a negative class label. This classifier could simply predict the value of the majority class, in this case, the positive class, for all predictions and still achieve high classification accuracy. Thus, a high classification accuracy can just mean that the classifier is by default predicting the class that appears most in the dataset (Weiss, 2004).

Therefore, it is useful also to use other performance metrics to evaluate the performance of a classifier. Before describing other performance metrics, the following terminology is needed. Participants that are correctly identified to have a condition (e.g. as patients with schizophrenia) are the so-called true positives (TPs). Participants that are correctly classified not to have a condition (e.g. as healthy controls) are the so-called true negatives (TNs). A binary classifier has two types of errors. When a participant with a condition is incorrectly predicted not to have the condition (e.g. patient with schizophrenia as healthy control), it is called a false negative (FN). When a participant without a condition is incorrectly predicted to have a condition (e.g. healthy control as a patient with schizophrenia), it is called a false positive (FP) (Kassraian-Fard et al., 2016). This could also be illustrated by a confusion matrix such as in figure 2.1.

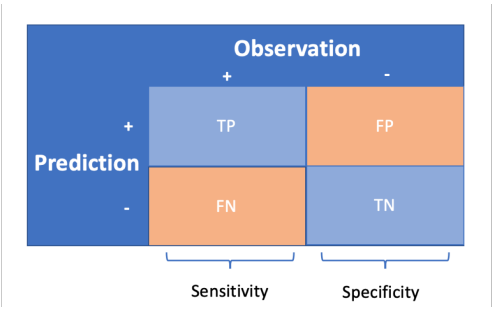


Figure (2.1) Illustration of a confusion matrix

Using the notation #FN, #FP, #TP and #TN for the number of false negatives, false positives, true positives, and false negatives, it follows that

$$\text{Accuracy} = \frac{\#TP + \#TN}{\#TP + \#TN + \#FP + \#FN} \tag{2.1}$$

Other performance metrics that can be used to assess the quality of a model are the metrics sensitivity and specificity. The sensitivity, also known as recall, measures the percentage of actual positives that are correctly classified as such (e.g. the percentage of patients with schizophrenia that are correctly classified as a patient). The specificity represents the performance of those with a negative class label (e.g. the percentage of healthy controls that are correctly identified as a healthy subject). The formulas for sensitivity and specificity are:

$$\text{Sensitivity} = \frac{\#TP}{\#TP + \#FN} \tag{2.2}$$

and

$$\text{Specificity} = \frac{\#TN}{\#FP + \#TN} \tag{2.3}$$

Hence, a classifier that only predicts positive class labels would result in a sensitivity of 1 and a specificity of 0, whereas a classifier that only predicts negative class labels would result in a specificity of 1 and a sensitivity of 0. Ideally, we would like to have a model that scores high on both metrics. For this reason, the balanced accuracy is used, which is simply the average of these two values. This metric is thus robust to imbalanced datasets, since a model that predicts only one class leads to a balanced accuracy of only 0.5. The formula for the balanced accuracy is:

$$\text{Balanced accuracy} = \frac{\frac{\#TP}{\#TP+\#FN} + \frac{\#TN}{\#TN+\#FP}}{2} \tag{2.4}$$

For any classifier, there is usually a trade-off between the sensitivity and specificity. This trade-off can be visualised by a receiver operating characteristic (ROC) curve. The ROC curve illustrates how well a model can distinguish the two classes for various thresholds. It plots the true positive rate (TPR) against the false positive rate (FPR) in which the TPR is equal to the sensitivity and the FPR is equal to 1 minus the specificity. An ideal classification would yield a TPR of 1 and an FPR of 0 (Kassraian-Fard et al., 2016).

When the ROC curve gets closer to the top-left corner, the higher the balanced accuracy, hence, the better a model can separate the two classes correctly. The coordinate of the sensitivity and specificity with the smallest euclidean distance to the top-left corner could be selected as a new threshold. Figure 2.2 illustrates two different ROC curves. The Area Under the Curve (AUC) of the ROC curve could be used as a metric to measure the degree of the separability of a model. The higher the AUC, the better the model is able to distinguish two classes. Ideally, a classifier would have an AUC of 1. This is also illustrated in figure 2.2d.

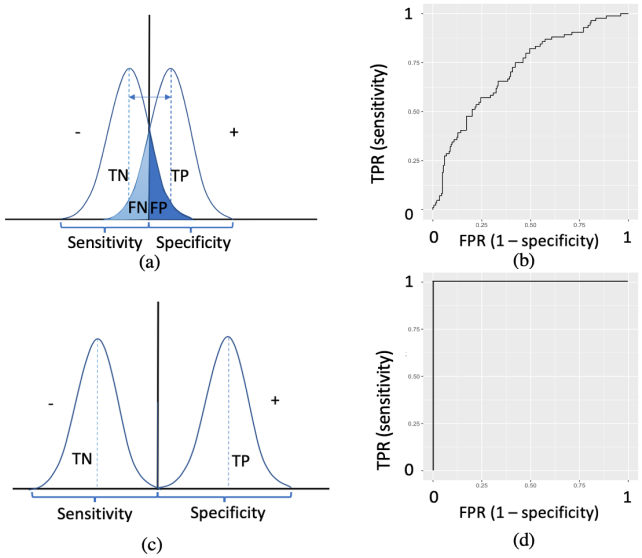


Figure (2.2) (a) Illustration of data that is not perfectly separable resulting in some overlap. (b) illustration of an ROC curve of (a) with an AUC of 0.75. (c) illustrates a dataset that is perfectly separable resulting in 100% sensitivity and 100% specificity. (d) An ROC curve of a perfectly separable dataset (c) with AUC 1.

In summary, different performance metrics can be used to evaluate the performance of a model. These metrics can deal with certain problems (e.g. imbalanced datasets) and focus on the size of the sample that is misclassified by the model. However, except for the number of misclassifications, these metrics do not tell us anything about the specifics of the misclassified participants. Nonetheless, the specifics of the misclassified participants could help us to understand why specific misclassifications occur in the models. In the next section, we propose the three promising approaches to study the misclassified participants in more detail. These approaches use systematic identification of the characteristics and causes of the misclassifications which could help to understand why the models did not reach the desired performance. This could help us to make better decisions which may improve the models.

2.2 Approach 1: effect size evaluation of data features

In the first approach, the aim is to determine whether there are features that have significantly different values for the misclassified participants compared to the correctly classified participants. These features can play a critical role in identifying the misclassified participants and can help us understand why the specific participants are misclassified (Alsallakh et al., 2014). In this approach, we have chosen only to analyse the features that were not used for training, since these features are not analysed by the model yet. The misclassified participants are compared to the correctly classified participants to check whether the specific values are a characteristic of the misclassified sample rather than the whole sample. For instance, when a dataset consists of 90 percent males and only 10 percent females, presumably more misclassifications are male. However, the fact that the misclassified participants are mostly male would not be a characteristic of the misclassifications. Hence, the misclassified participants need to be compared to the correctly classified participants. The misclassified participants could be compared either to the correctly classified participants with the same class label given by an expert or to the correctly classified participants with the same prediction label indicated by the model. This difference can clearly be illustrated by a confusion matrix such as in figure 2.3. The misclassified participants can be compared either along the axis of observation (vertical arrows) or along the axis of prediction (horizontal arrows). Comparing the correctly and incorrectly classified participants with the same observation can be used to identify specific characteristics of the misclassifications, whereas comparison between the correctly and incorrectly classified participants with the same prediction can help to identify variables that can be used to discriminate the correct and incorrect classifications.

Since this thesis aims to identify the specifics of the misclassified participants, we focus on the comparison along the axis of observations.

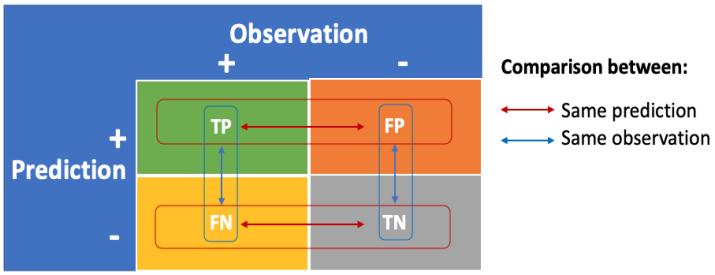


Figure (2.3) Confusion Matrix. The red horizontal arrows illustrate the comparison between the correct and incorrect classifications with the same prediction indicated by a model. The blue vertical arrows illustrate the comparison between the correct and incorrect classifications with the same class label given by an expert.

To analyse whether features have significantly different values for the misclassifications it is a possibility to investigate them one-by-one. For instance, one may suspect that males or females are more likely to be incorrectly classified, since the variable sex influences symptoms and responses to the treatment of schizophrenia (Abel, Drake, & Goldstein, 2010). However, analysing single features is time-consuming (Kononenko, 2001).

Therefore, we created a dedicated, time-efficient way to systematically evaluate and rank all the remaining variables on their effect size. The effect size is a metric that computes the difference between two groups with respect to a variable (Becker, 2000). The larger the effect size, the more significant the difference. In this case, the two groups are the correct and incorrect classifications of either the positive class or the negative class.

The effect size can be computed by dividing the difference between the correct and incorrect classifications by the pooled standard deviation. The formula to compute the effect size is:

$$d = \frac{\Delta x}{SD_{pooled}} \tag{2.5}$$

When analysing the incorrect and correct classifications with a positive class label, Δx is the difference with respect to a feature x between the false negatives and true positives of the sample. Thus, the formula for Δx for the positive class is:

$$\Delta x = x_{TP} - x_{FN} \tag{2.6}$$

And for the incorrect and correct classifications with a negative class label Δx can be defined as:

$$\Delta x = x_{TN} - x_{FP} \tag{2.7}$$

SD_{pooled} stands for the pooled standard deviation. The pooled standard deviation for the positive class is formulated as follows:

$$SD_{pooled} = \sqrt{S_{FN}^2 + S_{TP}^2} \tag{2.8}$$

whereas the pooled standard deviation for the negative class is defined as:

$$SD_{pooled} = \sqrt{S_{FP}^2 + S_{TN}^2} \tag{2.9}$$

where $S_{FN}^2, S_{TP}^2, S_{FP}^2$ and S_{TN}^2 are the within-sample variances of the variable x for the false negatives, true positives, false positives, and true negatives, respectively. Therefore, the larger Δx and/or the smaller the pooled standard, the larger the effect size. Figure 2.4 illustrates some features with different effect sizes.

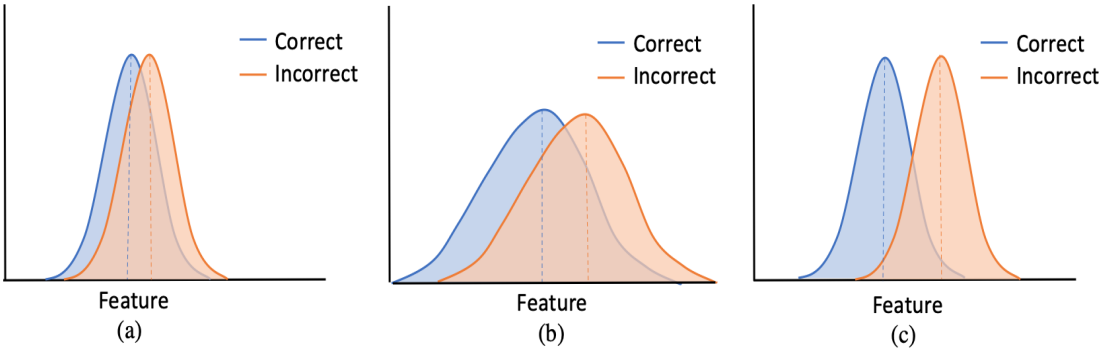


Figure (2.4) Effect sizes. (a) illustrates a small effect size. The difference between the correct and the incorrect classifications is very small. (b) illustrates a relatively large effect size. Although Δx is relatively large, there is a considerable amount of overlap due to a large pooled standard deviation. (c) illustrates a large effect size. The difference between the two samples is large. Besides, the pooled standard deviation is relatively small. As a result, the effect size is large which results in relatively little overlap.

To help with the interpretation of the effect sizes, Cohen (1988) proposed that effect size values of 0.2, 0.5, and 0.8 represented small, medium, and large effect sizes, respectively, perhaps more meaningfully described as hardly, subtle and obviously statistical (Fritz, Morris, & Richler, 2012). If the effect size is below 0.2, there is almost complete overlap between the correct and incorrect classifications, i.e. the feature has no significantly different values for the incorrect classification compared to the correct classifications. In table 2.1 an overview of Cohen's qualifications is given.

Effect size	**Cohen's qualification**
0-0.2	-
0.2-0.5	Small
0.5-0.8	Medium
> 0.8	Large

Table (2.1) Effect sizes and their Cohen's qualification.

To analyse whether a model has features with significantly different values for incorrectly classified sample, each remaining variable could be analysed whether it is has a small, medium or large effect size (ES). Table 2.2 gives an illustration.

	Remaining variables	**Large ES**	**Medium ES**	**Small ES**
Model 1	100	5 (5%)	20 (20%)	45 (45%)
Model 2	80	0 (0%)	12 (15%)	32 (40%)
Model 3	100	8 (8%)	25 (25%)	30 (30%)
Model 4	50	5 (10%)	10 (20%)	15 (30%)

Table (2.2) Illustration of four models that are analysed on the remaining variables. For each remaining variable there is observed whether a variable has a small, medium, or large effect size between the correct and incorrect classifications.

In this example, models 1, 3 and 4 seem to have some features with large effect sizes. Model 2 does not have any large effect sizes. The features with large effect sizes can be further analysed and visualised. Features with continuous values can be visualised by density plots, whereas features with discrete values can be visualised by histograms. The density plots illustrate the distribution of a variable. Figure 2.5 gives an example of what such a density plot or histogram may look like.

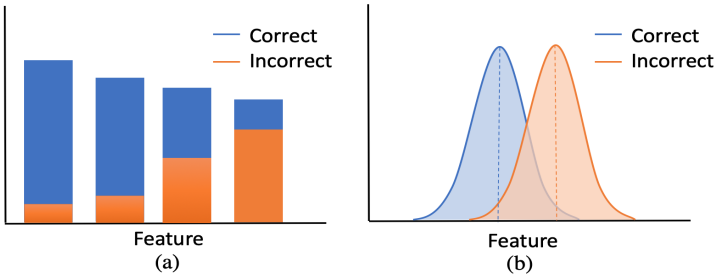


Figure (2.5) Illustrations of features that reported large effect sizes between the correct and incorrect classifications. (a) illustrates a histogram of a feature with discrete values. (b) illustrates a density plot of a feature with continuous values.

When several models are trained on the same sample but on different variables, the reported features with large effect sizes could also be analysed whether the same or different features are reported.

2.3 Approach 2: direct model comparison

The second approach is used to determine whether the same or different participants are misclassified in various models. This could help to identify factors that influence the misclassifications. If the same set of subjects is misclassified in different models, there seems to be something specifically different with these misclassified subjects compared to the other participants. In that case, one could focus on the specifics of the misclassified sample in order to find out what makes that this same sample is misclassified in all the different models. When different participants are misclassified in various models, the focus should be shifted to the selected variables, since this would suggest that the choice of features affects which participants are misclassified. To determine whether the same or different participants are misclassified, we compared different models that are trained on the same sample of participants but on different input features. There are different ways to compare models on their misclassifications.

The first way is by analysing the degree of overlap between two models with confusion matrices and Venn-diagrams. In table 2.3, an example of a confusion matrix of the negative class is given. In this example, the number of participants that are misclassified in both models is represented by a. The values b and c represent the number of participants that are misclassified in a single model. b represents the participants that are misclassified by model 1 and correctly classified by model 2. c represents the reverse: the participant that are misclassified in model 2 and correctly classified in model 1. The participants who are not misclassified in a single model are represented by d. The values of the confusion matrix are also visualised by a Venn-diagram in figure 2.6.

		Model 2	
		FP	TN
Model 1	FP	a	b
	TN	c	d

Table (2.3) Confusion matrix - Negative class

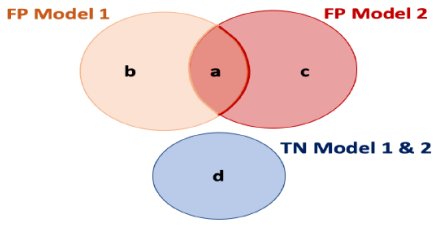


Figure (2.6) Venn-diagram - Negative class

The degree of overlap of misclassifications of two models could also be computed in percentages. These percentages can include or exclude the participants that are classified correctly by both models. The degree of overlap when including these participants can be defined as

$$\frac{a}{a+b+c+d} \times 100 \tag{2.10}$$

and

$$\frac{a}{a+b+c} \times 100 \tag{2.11}$$

when only taking the misclassifications of the two models into account. The Venn-diagrams could be expanded, to compare more than two models with each other. Another option is to use a lower triangular heatmap which can illustrate the degree of overlap of misclassifications of multiple models. A lower triangular heatmap shows the degree overlap by colours. Figure 2.7 gives an illustration. In this example, the lighter the colour, the more overlap of misclassifications there is.

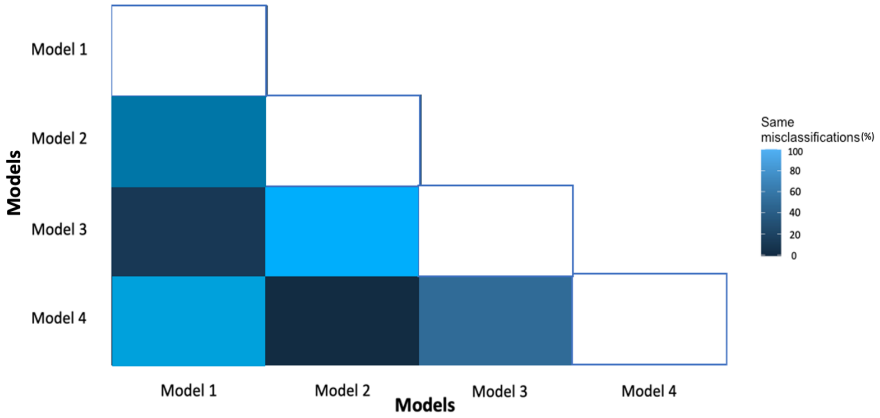


Figure (2.7) A lower triangular heatmap that illustrates the degree of overlap in misclassifications of four different models. In this illustration, model 2 and model 3 seem to have the most overlap and model 2 and 4 the least.

Another way to determine whether the same or different participants are misclassified in various models is by counting the number of models in which each participant is misclassified. To get a better understanding, we can visualise this with a histogram. In this histogram, each bin k represents a sample X_k, where X_k represents the number of participants that are misclassified in k models. When we analyse m models, $0 \leq k \leq m$. Thus, when $X_0 = 7$, it would mean that seven participants of the sample X were not misclassified in a single model, whereas $X_m = 1$ would mean that one participant was misclassified in all the evaluated models. An example is illustrated below in which three models are evaluated on their misclassifications. In this example, seven participants are correctly classified in all the models, eight participants are only misclassified in a single model, three participants are misclassified in two of the three models, and one participant is misclassified in all of the three models.

Number of models in which the participants are misclassified	**Number of participants**
0	7
1	8
2	3
3	1

Table (2.4) The overlap of misclassifications of three models.

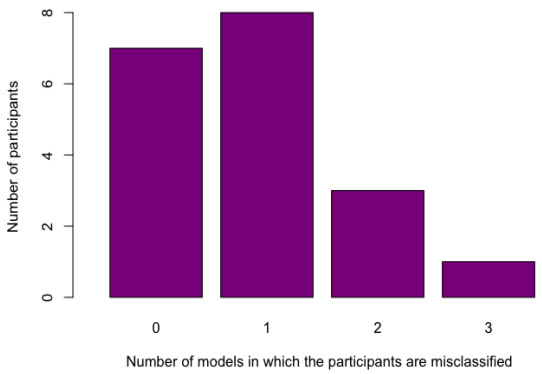
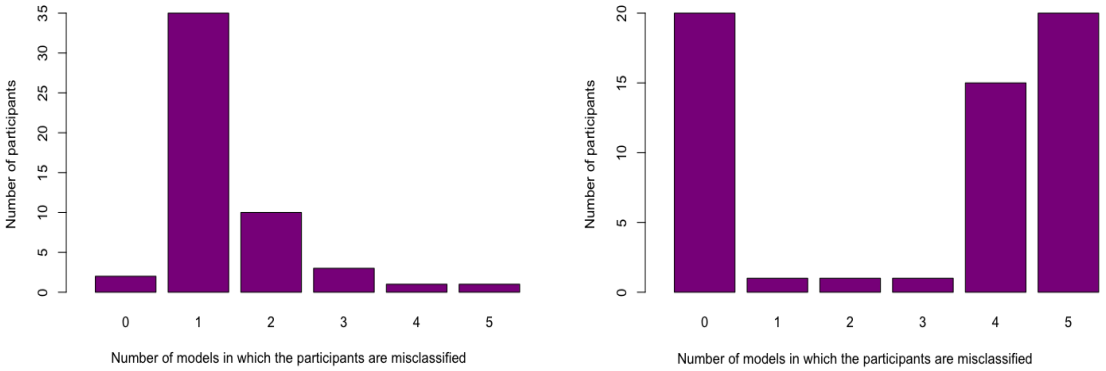


Figure (2.8) Histogram of the three models.

When different participants are misclassified in various models, it would be unlikely that the same participants are misclassified in multiple models. In this case the results would show that most subjects are only misclassified in a few models ($0 < k < \frac{1}{2}m$). Otherwise, when the same participants are misclassified in various models, the results would show that one part of the sample is misclassified in multiple models ($\frac{1}{2}m \leq k \leq m$), and the remaining part of the sample is not misclassified in a single model ($k = 0$). Figure 2.9 demonstrates two almost ideal illustrations in

which figure 2.9a illustrates a histogram when different participants are misclassified in five different models. Figure 2.9b gives an almost ideal illustrations of the results when the same participants are misclassified in five different models.



(a) **(b)**

Figure (2.9) Two illustrations in which five models are evaluated on their misclassifications. (a) illustrates a histogram when different participants are misclassified in five different models. (b) illustrates a histogram when the same participants are misclassified in five different models.

To analyse in which models the sample is mostly misclassified, the histograms could be further expanded to ones where colours indicate in which models the participants were incorrectly classified. This type of visualisation helps to analyse if specific models have little overlap in misclassifications. Figure 2.11 gives an example of such a histogram. In this example, the participants that are misclassified in a single model are misclassified either in the model trained on feature set 1 or in the model trained on feature set 2. Furthermore, the figure shows that all the participants that are misclassified in two of the three models are misclassified in the models trained on feature set 2 and 3. In this example, the model that was trained on feature set 1 seem to have little overlap in misclassifications.

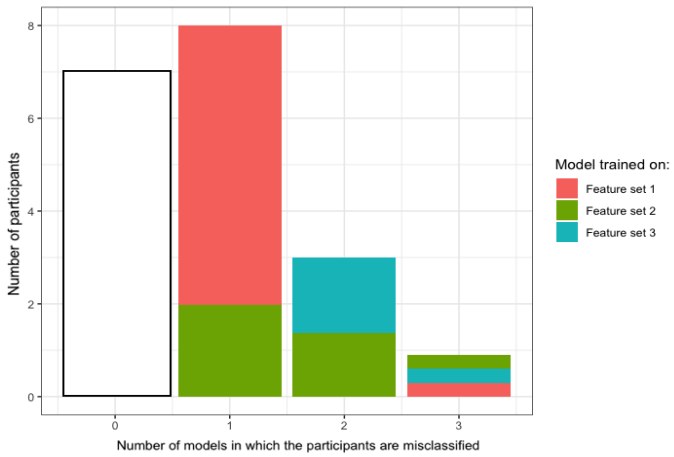


Figure (2.10) Illustration of a histogram that specifies in which models the misclassifications occurred. The colours present in which specific models the participants were misclassified. The size of the colours represents the proportion.

2.4 Approach 3: evaluation of the threshold

Most studies that attempt to diagnose or prognose schizophrenia with machine learning chose to use a classifier with discrete target values rather than a regression model with continuous target values since discrete target values are less ambiguous (Janssen et al., 2018). However, a disadvantage of discrete target values is that it loses some information, such as the severity of the disease. Many classification algorithms work internally with continuous values. These continuous output values are converted to discrete output values using a threshold. Participants with values further away from the threshold are less likely to be misclassified, whereas participants with values closer to this threshold are more likely to be misclassified (Bolin & Finch, 2014). To analyse whether the misclassifications are the ones close to the threshold, the continuous output value of the ML models can be used to compute the distance to the threshold. The smaller the distance between the threshold and the continuous value of the misclassifications, the lower the degree of misclassifications seem to be.

To quantify the distance to the threshold of all the misclassifications, the absolute mean distance can be used. A distinction could be made between the false positives and the false negatives of the model. Furthermore, the distance to the threshold of the correct classifications could be observed. The distance to the threshold could also be illustrated by scatter plots. An illustration is given below.

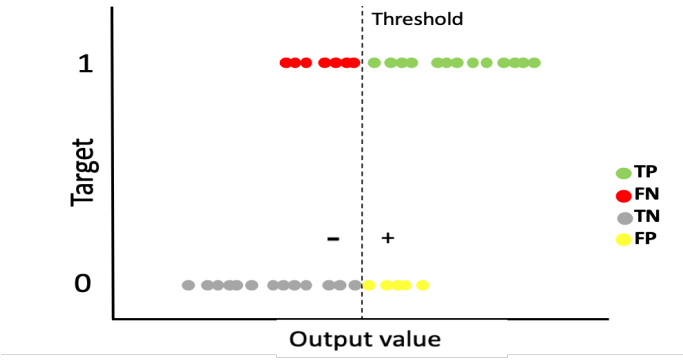


Figure (2.11) Illustration of a scatter plot that visualises the distance to the threshold of the true positives, false negatives, true negatives and false positives of an ML model. The values on the y-axis represent the continuous output values of the ML model. Each point represents a participant. The threshold is illustrated by the dotted line. Participants on the left of the dotted line are predicted to have a negative class label, whereas participants on the right of the dotted line are predicted to have a positive class label.

Another way to analyse the influence of the threshold is by analysing the severity of symptoms. For instance, Di Carlo et al. (2019) showed that healthy subjects that were incorrectly predicted by an ML model as patients with schizophrenia had relatively higher positive schizotypy scores compared to the correctly classified healthy subjects meaning that the misclassified healthy subjects shared relatively more subclinical traits and biological phenotypes with patients with schizophrenia. This schizotypy score of the healthy controls was measured by the Schizotypal Personality Questionnaire (SPQ). Note that in this study, only the healthy controls were analysed on their misclassifications. However, it would also be valuable to observe whether the patients with schizophrenia with less severe symptoms were more likely to be misclassified than the patients with more severe symptoms.

To quantify the severity of the symptoms, clinical questionnaires could be used. For instance, the Positive and Negative Syndrome Scale (PANNS) is a questionnaire that rates the symptom severity of patients with schizophrenia with a scale ranging from 1 to 7. The severity of a symptom is rated 1 when the symptom is absent, 3 when mildly present and 7 when extremely present.

This type of analyses could also be done for prognostic models. In this case, there can be determined whether the participants with values close to the threshold of the remission criteria were more likely to be misclassified. In the prognostic classifiers, the severity of symptoms is already

taken into account. However, no distinction is made in the degree of the severity of the symptoms. According to the prognostic classifier, a patient has either reached remission or not. To quantify the severity of the symptoms to a more continuous value, we could, for instance, compute the mean of the items that determine whether a patient has reached remission. This mean remission score could be computed for the misclassified sample as well as for the correctly classified sample.

Additionally, this remission score can be plotted against the continuous values of the ML model. This could also help to evaluate the influence of the loss of information. An example is shown in figure 2.12. In this figure, the continuous target score (e.g. mean of the remission criteria of a prognostic model) is plotted against the continuous value of the ML model. In this example, the misclassifications seem to be influenced by the loss of information.

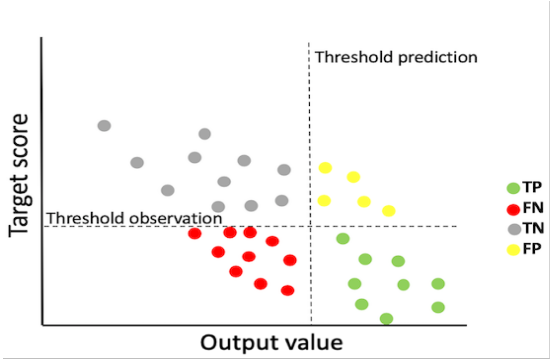


Figure (2.12) Illustration of a model in which the target score is plotted against the continuous output value of the model. Each point represents a participant. The different colours show whether a participant is a true positive, false negative, true negative, or false positive. The horizontal dotted line represents the threshold for the observations. The participants with a target score below this threshold get a negative class label, whereas participants with a target score above this threshold get a positive class label. This vertical dotted line represents the threshold of the ML model. Participants on the left of this threshold are predicted to have a negative class label, whereas participants on the right of the threshold are predicted to have a positive class label.

Chapter 3

Prediction models

To analyse and identify the specifics of misclassifications with the approaches from the previous chapter, two real-world datasets have been used. These two clinical datasets have been used to train different ML models that diagnose and prognose schizophrenia. This chapter starts with a brief overview of the process of ML. Subsequently, we discuss the steps and choices of our ML models, such as the choice of the ML algorithm and the choice of the feature selections.

A pipeline in machine learning

To predict an outcome with a (supervised) ML model, several steps need to be taken. These steps include data collection, extraction of certain features from the data, and specification of the output variable, also known as the target (Tandon & Tandon, 2018). This pipeline in ML is illustrated in figure 3.1. Different types of measurement instruments can be used for data collection. For instance, one widely used measurement instrument for ML models that diagnose or prognose schizophrenia is brain imaging, also called neuroimaging. Other measurement instruments that have been used are clinical observations, demographic information or blood samples (Koutsouleris et al., 2015; Schnack, 2020). The choice of data, features, and target can have a significant impact on the ML models. For instance, when the features used for classification are not optimally representative for a particular target, the features weaken and limit the prediction performances.

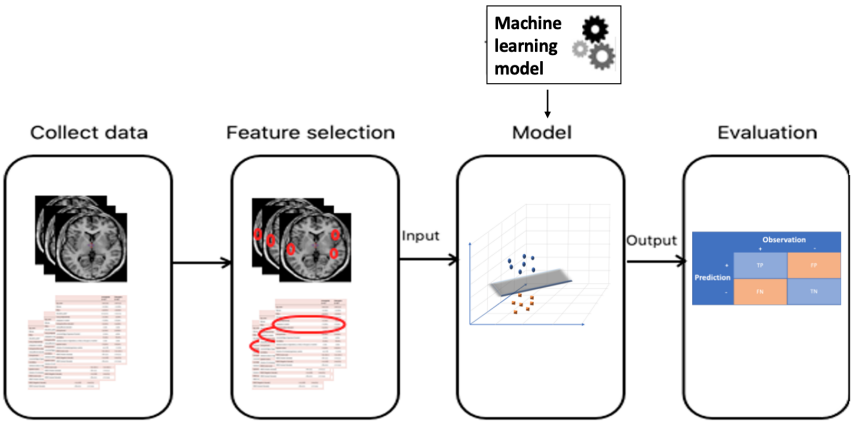


Figure (3.1) Pipeline of machine learning.

3.1 Data

In this section, the two clinical datasets, that are used for this study, are discussed. The two datasets are very different on aspects, such as the choice of the sample, the variables that are selected, how the information is collected, and the choice of the target selection. More details of the dataset can also be found in table B.1 and table B.2 of appendix B.

Dataset 1

The first dataset, the so-called OPTiMiSE dataset, consists of information on clinical variables and demographic information such as age, sex, living situation, education of the participant, and background of the parents (e.g. education degree). Patients (n = 386) were all aged between 16 and 44 and had to meet the criteria of the DSM-IV (Diagnostic and Statistical Manual of Mental Disorders) for schizophrenia, schizophreniform disorder, or schizoaffective disorder. Clinical variables include clinical questionnaires such as Positive and Negative Syndrome Scale (PANNS), Clinical Global Impression of Severity / Improvement (CGI), Calgary Depression Rating Scale for Schizophrenia (CDSS), Udvalg for Kliniske undersogelser Side Effect Rating Scale (UKU), and Subjective Well-being under Neuroleptics (SWN). The patients had follow-up interviews taken one, two, and four weeks after onset. After four weeks, there was observed if symptomatic remission had been achieved. Symptomatic remission was achieved when eight particular symptoms rated by the PANSS (items P1, P2, P3, N1, N4, N6, G5, and G9) were, at most, only mildly present (maximum rating of three), i.e. the symptoms did not interfere with daily life functioning anymore. The patients were recruited in 27 centres, located in 14 European countries (Austria, Belgium, Bulgaria, Czech Republic, Denmark, France, Germany, Italy, the Netherlands, Poland, Romania, Spain, Switzerland, and the United Kingdom) and Israel. This dataset has been used to prognoses patients with schizophrenia or other schizophrenia-like diagnoses. Further details can be found in Kahn et al. (2018).

Dataset 2

The second dataset, the so-called GROUP dataset, contains information of the brain of 204 healthy controls and 84 patients with schizophrenia or other schizophrenia-like diagnoses and were all aged between 16 and 50 years old. The patients have also met the DSM-IV criteria for a nonaffective psychotic disorder, diagnosis of schizophrenia, schizoaffective disorder or schizophreniform disorder. This dataset has been used for ML models that diagnose patients with schizophrenia or other schizophrenia-like diagnoses. The participants were recruited at University Medical Center Utrecht.

The GROUP dataset consists of brain volumes and other brain metrics which were obtained by MRI scans. The dataset includes subcortical volumes (e.g. volume of white matter), volumes of larger structures of the brain (e.g. volume of the total brain), and Regions of Interest (ROIs)[1]. For each ROI, the mean cortical thickness, the volume, and the surface area was measured. In addition to the brain features, the variables sex, age, and the total IQ score of the participants were available. More details of the data can be found in Kubota et al. (2015).

3.2 Pre-processing

Both datasets contained features with different ranges. For instance, the OPTiMiSE dataset contains the two features age and income. The feature age ranges from 16 to 44, while the variable income ranges from 0 to 340000. Consequently, when we do further analysis, the attributed income will intrinsically influence the result more due to its larger values. However, it is not necessarily the case that this variable is more important as a predictor. For this reason, the features with different ranges were normalised. This way the features with numeric values of a dataset were converted to a standard scale, without distorting differences in the ranges of the values.

The features were normalised with z-transformation, in which each data point was scaled by the following formula:

$$\frac{x_i - \bar{x}}{SD}$$

[1]ROIs are selections of different brain regions (Poldrack, 2007).

where x_i represents the original values of a variable x, $\bar{x}$ is the mean of the variable, and SD stands for the standard deviation of the variable. As a result, each feature was transformed to have a mean of zero and a standard deviation of 1. Note that this is not done for features with binary values.

Furthermore, before training the binary classifiers that diagnose or prognose schizophrenia, the positive and negative class have to be defined. The diagnostic models selected the patients with schizophrenia as the positive class, whereas the prognostic models selected the patients who reached remission as the positive class.

3.3 Feature selections

As mentioned earlier, the performance of models highly depends on what features are used to train the models. The stronger the relationship between the features and target, the higher the performance. However, not only the feature itself but also the number of features affects the performances since the number of features influences the bias-variance trade-off; a very low number of features results in a high bias and a low variance. Consequently, a model is more sensitive to underfit, meaning that a model is unable to capture the underlying pattern of the data. On the other hand, a high number of features results in a low bias and a high variance such that the model is more sensitive to overfit, meaning that a model is unable to generalise well on unseen data (Schnack, 2020).

To evaluate the effect of the features and to analyse approach 2, different models have been trained on different feature selection of the same sample. First of all, two models are trained each on all the features of one dataset. To reduce the number of features, different combinations of features have been selected. Some of the feature selections have been based on categories. For instance, from the OPTiMiSE dataset, we selected the categories (clinical) questionnaires, demographic information, and information on psychiatric diagnoses and current treatment as feature selections. Some feature selections were subcategories. For example, of the category clinical questionnaires, three different questionnaires (PANNS, CDSS, and UKU) have been selected. Table 3.1a presents an overview of the different feature selections that have been selected from the OPTiMiSE dataset.

From the GROUP dataset, we selected the categories subcortical volumes and the volumes of large brain areas. The GROUP dataset also contains the category ROIs. From this category, three different subsets have been selected: the cortical thickness of the ROIs, the volumes of the ROIs, and the surface areas of the ROIs. Furthermore, feature selections have also been selected based on the side of the brain. One feature selection only contained variables of the left hemisphere and another feature selection only variables of the right hemisphere. The different feature sets selected from the GROUP dataset are represented in table 3.1b.

Feature selection
1. All features (231)
2. Demographic (25)
3. Psychiatric diagnoses and current treatment (65)
4. Questionnaires (133)
5. PANNS (30)
6. UKU (60)
7. CDSS (9)
8. Expert selection (6)
9. Elastic net (28)

(a) OPTiMiSE

Feature selection
1. All features (245)
2. Left hemisphere (121)
3. Right hemisphere (121)
4. Subcortical volumes (15)
5. Volumes of larger structures (11)
6. ROI - Thickness (68)
7. ROI - Volume (68)
8. ROI - Area (68)
9. Elastic net (41)

(b) GROUP

Table (3.1) Feature selections of the GROUP and OPTiMiSE dataset

A feature selection could also be based on a knowledge-based selection in which features are selected based on previous research. According to previous research, prognostic predictors are variable such as the CGI severity, the total PANNS score, the duration of the current psychotic episode, the current occupation, sex, and duration of education (Abel et al., 2010; Kay, Fiszbein, & Opler, 1987; Loebel et al., 1992; Merinder, 2000; Cooper, 1961). Hence, from the OPTiMiSE dataset, we selected also these variables as feature selection.

Another way to select features is with linear regularised models. Linear regularised models selects features based on their link with the target. Linear regularised models are especially useful when there are many correlated predictor variables (Friedman, Hastie, & Tibshirani, 2010). In this study, we used the elastic net model as linear regularised model. The elastic net model is a compromise between the lasso penalty and ridge penalty in which the lasso penalty tends to pick only one feature that has the strongest link with target and does not take notice of the other features, whereas the ridge penalty tends to select multiple variables instead of only one feature (Zou & Hastie, 2005). The variables selected by the elastic net model can be found in table B.3 of appendix B.

3.4 ML algorithms

To create an ML model, we used soft-margin linear Support Vector Machines (SVMs) as supervised ML algorithm since this ML algorithm is frequently used in psychiatry, especially in psychiatric neuroimaging (Mourao-Miranda et al., 2012; de Wit et al., 2017). However, note that any other ML algorithm could have been chosen for classification, given the fact that the misclassifications are the main focus. An advantage of linear SVMs is that they are relatively easy to interpret due to their interpretable weight vector that could easily show the weights for each feature. Besides, SVMs can handle noisy and high-dimensional data; this helps to prevent overfitting. In addition, SVMs are relatively memory efficient.

SVMs are based on statistical learning theories of Vapnik (1999). The key idea of (linear) SVMs is to find an optimal (linear) decision boundary that has a maximum margin, meaning that the boundary is maximally far from the data samples of the two classes (Kassraian-Fard et al., 2016). This optimal boundary is defined by a so-called Optimal Separation Hyperplane (OSH). The OSH separates the subjects based on their positions in an m-dimensional feature space, where m is the number of features. The position of a subject in the m-dimensional feature space is determined by their values of the features. If the number of features is two, the hyperplane is just a straight line, whereas the hyperplane is a two-dimensional subspace when the number of input features is three. This also illustrated in figure 3.2.

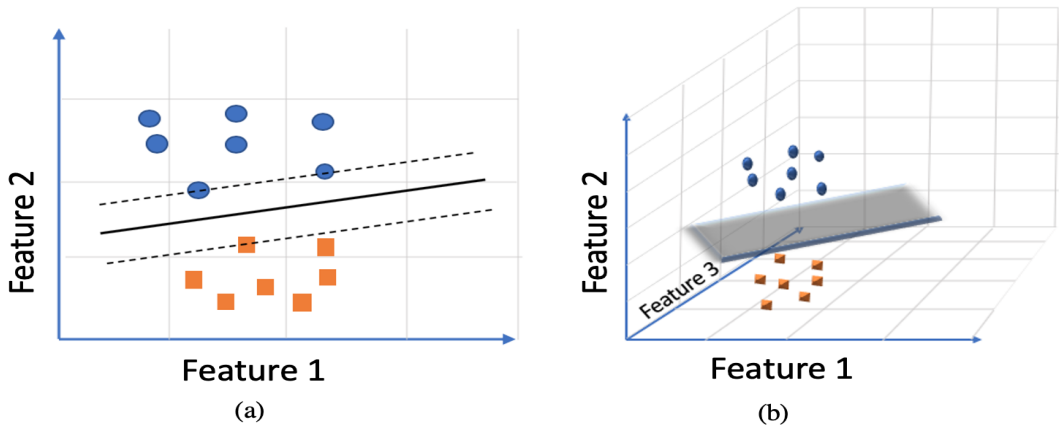


Figure (3.2) SVM. (a) is an illustration the optimal boundary in a two-dimensional feature space. (b) illustrates a two-dimensional optimal separation hyperplane in a three-dimensional feature space.

Of all the data points, only the data points that are close to the hyperplane influence the position and orientation of the hyperplane. These data points are called support vectors. Figure 3.3a illustrates some possible boundaries in a 2-dimensional feature space. 3.3b illustrate the OSH with the maximum distance between the two classes. In this figure the support vectors are encircles.

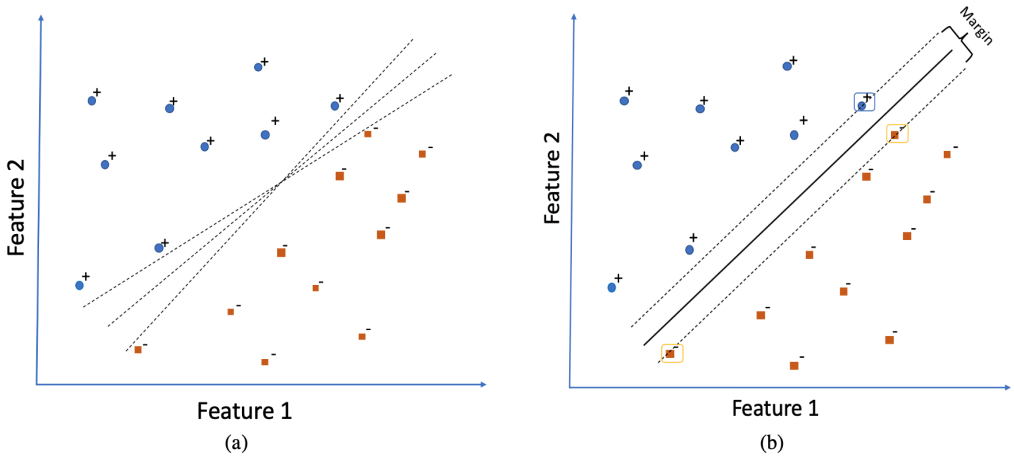


Figure (3.3) (a) illustration of possible hyperplanes in a two-dimensional feature space in which (b) illustrates the OSH with the largest margin.

In support vector machines, there are two types of margins: hard margins and soft margins. Hard margins work well when the classes can be linearly separated. However, in most cases, the data is not linearly separable. In those cases, soft margins are used. A soft margin allows some training data points to be misclassified. Consequently, some support vectors will lie within the margin (Franke, Ziegler, Klöppel, & Gaser, 2010). This is controlled by the so-called cost parameter C. To find the optimal value for this cost parameter, different values can be evaluated. Further details of the soft and hard margin can be found in section 1 and 2 from appendix A.

To enable an unbiased estimation of the model's generalisability to new patients, we employed a nested cross-validation approach since nested cross-validation prevents that information would leak between the patients that were used for training and validating the models (Koutsouleris et al., 2016). Both datasets were divided into ten outer folds. Each outer fold had left out one non-overlapping test fold. These left-out folds were used for evaluation. The remaining parts of the outer folds were used to train the models. Each training set in the outer loop was again divided into five inner folds in which the cost parameter (C) was optimised. More details of (nested) cross-validation can be found in section 3 from appendix A. The cost variables used for parameter optimisation were powers of 2, ranged from 2^{-7} to 2^{5}. For each outer fold, the cost parameter with the best performance in the inner loop was reported. The average of all the ten reported values was computed by the geometric mean since the values were powers of two. Additionally, the median of the reported values was computed. The geometric mean was used when there was a great variation in the reported values; otherwise, the median was used as the final cost parameter value.

3.5 Missing Values

It very often happens that some information about patients in patient records lacks certain data (Kononenko, 2001). In the GROUP dataset, only the variable total IQ score had some missing values. However, in the OPTiMiSE dataset, multiple variables had some missing values. We initially, removed potentially difficult-to-obtain variables of the OPTiMiSE dataset (those with $\geq 15\%$ missing values). To deal with the remaining missing values, the ML algorithm K-nearest neighbouring was employed to estimate the missing values (Troyanskaya et al., 2001). K-nearest neighbouring replaces the missing values in a dataset by the values of the k points that are closest to the other variables. This technique is also employed by others that prognose schizophrenia with ML (Koutsouleris et al., 2016).

3.6 Class weights

Datasets could be imbalanced, meaning that datasets contain imbalanced class sizes. For instance, the GROUP dataset includes 84 patients with schizophrenia (SZ) or other (SZ-like) diagnoses and 204 healthy controls. As a result of this imbalance, an ML model could only predict one outcome and still reach a high accuracy. To handle these imbalanced datasets, class weights could be added in which the misclassifications of the smaller class were heavier penalised. Thus, for instance, in the GROUP dataset, the misclassified patients with schizophrenia should be penalised $\frac{204}{84}$ times more than the misclassified healthy controls.

3.7 Performance measures

Before analysing the different approaches, the models are analysed on their performances. Different performance metrics have been used. First of all, the accuracy has been computed for both datasets. Other performance metrics that have been used are the sensitivity and specificity. The sensitivity of the diagnostic models represents the percentage of patients that were correctly classified by the model. The sensitivity of the prognostic models represents the percentage of patients that were correctly classified to reach remission. The specificity of the diagnostic models represents the percentage of healthy subjects that were classified as such, and the specificity of the prognostic models represents the percentage of patients that were correctly classified not to reach remission. A lower specificity of the diagnostic models would mean that more healthy participants were incorrectly diagnosed by the models. On the other hand, a lower specificity of the prognostic models would mean that more patients who did not reach remission were incorrectly prognosed by the models. Furthermore, the balanced accuracy and the AUC have been measured.

Chapter 4

Analyses

In this chapter, we start with briefly evaluating the performance of the models from the previous chapter. Subsequently, the models' misclassifications are analysed using the three approaches from Chapter 2. The purpose of this chapter is to identify the specifics of the models' misclassifications and to show whether the approaches were able to identify the misclassified participants.

4.1 Performance of the models

Prognostic models

Table 4.1 presents an overview of the performances of the prognostic models trained on different subsets of the OPTiMiSE dataset.

Models trained on (No. variables)	Accuracy	AUC	Sensitivity	Specificity	Balanced Accuracy
All variable (231)	0.601	0.573	0.659	0.489	0.574
Demographic (16)	0.516	0.560	0.467	0.611	0.539
Diagnoses and treatment (65)	0.593	0.550	0.620	0.542	0.581
Clinical questionnaires (133)	0.630	0.607	0.678	0.534	0.606
PANNS (30)	0.640	0.645	0.655	0.611	0.633
CDSS (9)	0.611	0.570	0.765	0.313	0.539
UKU (60)	0.583	0.512	0.718	0.321	0.519
Knowledge-based selection (6)	0.586	0.630	0.549	0.657	0.630
Elastic net (28)	0.655	0.671	0.659	0.649	0.654

Table (4.1) Performances of the prognostic models trained on different feature selections of the OPTiMiSE dataset. Each row presents the performance of one model. The leftmost column shows on which features and the number of features each model is trained.

As can be seen from the table above, the model trained on the variables selected by the elastic net model resulted in the highest balanced accuracy, followed by the model trained on the PANNS variables and the knowledge-based selection. In most models, the accuracy and balanced accuracy ranged between 50 and 60 percent. In the models trained on the demographic variables and the knowledge-based variables, the sensitivity was higher than the specificity. In the other models, the specificity was higher. The models trained on the CDSS and UKU questionnaires reported both a very low specificity (< 0.33).

Table 4.2 presents the number of false positives, false negatives, true positives, and true negatives of the prognostic models. In these prognostic models, the false positives represent the patients who did not reach remission but were incorrectly predicted to achieve remission after four weeks. The false negatives represent the participants who reached remission but were incorrectly predicted not to meet the remission criteria.

	All	Demographic	Diagnose & Treatment	Clinical Questionnaires	PANNS	CDSS	UKU	Expert	Elastic net
FN	87	119	97	82	88	60	72	115	87
TP	168	136	158	173	167	195	183	140	168
FP	67	51	60	61	51	90	89	45	46
TN	64	80	71	70	80	41	42	86	85

Table (4.2) Number of false positives, false negatives, true positives and true negatives of the prognostic models trained on different feature selections of OPTiMiSE dataset. Each column presents a different model.

As shown in table 4.2, the number of false negatives of the prognostic models varies between the 60 and 119 subjects and the false positives between the 45 and 90 subjects.

Diagnostic models

Table 4.3 presents an overview of the performances of the different diagnostic models trained on different subsets of the GROUP dataset.

Models trained on (No. variables)	Accuracy	AUC	Sensitivity	Specificity	Balanced Accuracy
All variables (245)	0.701	0.741	0.595	0.775	0.685
Subcortical Volume (15)	0.635	0.637	0.595	0.652	0.624
Large Volume (11)	0.629	0.620	0.548	0.662	0.605
ROI - Thickness (68)	0.624	0.634	0.571	0.642	0.607
ROI - Area (68)	0.693	0.695	0.655	0.681	0.668
ROI - Volume (68)	0.653	0.697	0.560	0.691	0.625
Right hemisphere (121)	0.698	0.706	0.607	0.735	0.671
Left hemisphere (121)	0.722	0.750	0.667	0.745	0.706
Elastic net (41)	0.774	0.822	0.714	0.799	0.757

Table (4.3) Performance of the diagnostic models trained on different feature selections of the GROUP dataset. Each row presents the performance of one model. The leftmost column shows on which and the number of features each model is trained.

It can be seen from the data in table 4.3 that the balanced accuracy of most diagnostic models ranged between 60 and 70 percent. In all the diagnostic models the specificity was higher than sensitivity indicating that there is a higher chance that a patient with schizophrenia is misclassified as a healthy control than a healthy control is misclassified as a patient with schizophrenia.

Again, the models trained on the features selected by the elastic net model yielded the highest performance, followed by the model trained on the left hemisphere variables and the model trained on all the variables of the GROUP dataset. The number of false positives, false negatives, true positives, and true negatives of the diagnostic models is shown in table 4.4. In these models, the false negatives represent the patients with schizophrenia that were incorrectly predicted as healthy subjects. The false positives represent the healthy controls that were incorrectly predicted as patients with schizophrenia.

	All	Subcortical volumes	Large volumes	Thickness ROI	Area ROI	Volume ROI	Right Hemisphere	Left Hemisphere	Elastic net
FN	39	34	38	36	29	37	33	28	24
TP	45	50	46	48	55	47	51	56	60
FP	43	71	69	73	65	63	54	52	41
TN	161	133	135	131	139	141	150	152	163

Table (4.4) Number of false positives, false negatives, true positives and true negatives of the diagnostic models trained on different feature selections of GROUP dataset. Each column presents a different model.

In table 4.4 we can see that the number of false negatives of the models varied between the 24 and 39, and the false positives between the 41 and 73.

The different results of the diagnostic and prognostic models showed that relatively more participants are misclassified in the prognostic models. More details of the misclassified participants will be discussed in the next sections.

4.2 Approach 1

In this section, we start with analysing whether the diagnostic and prognostic models reported some features with small, medium, or large effect sizes between the true and false class labels. If the models reported features with large effect sizes, these features were presented in tables, and some were further illustrated by density plots or histograms.

Prognostic models

Table 4.5 and 4.6 present the number of remaining features with large, medium, and small effect sizes of all the nine prognostic models. Table 4.5 shows all the effect sizes between the true negatives and false positives of all the prognostic models and table 4.6 the effect sizes between the true positives and the false negatives. As mentioned earlier, we only analysed the features that were not used for training. Note that the models trained on more variables have less remaining features. In particular, the model trained on all the features did not have any remaining features, and thus reported no remaining features with a large, medium, or small effect size.

Models trained on (No. variables)	No. remaining variables	Large ES	Medium ES	Small ES
All (231)	0	-	-	-
Diagnose & Treatment (65)	166	0 (0.0%)	2 (1.2%)	40 (24.1%)
Demographic (25)	206	0 (0.0%)	1 (0.5%)	62 (30.1%)
All questionnaires (133)	98	0 (0.0%)	1 (1.0%)	26 (26.5%)
PANNS (30)	201	0 (0.0%)	2 (1.0%)	61 (30.3%)
CDSS (9)	222	2 (0.9%)	6 (2.7%)	111 (50%)
UKU (60)	171	2 (1.2%)	16 (9.4%)	63 (36.8%)
Expert (6)	225	10 (4.4%)	19 (8.4%)	50 (22.2%)
Elastic net (28)	203	0 (0.0%)	5 (2.5%)	53 (26.1%)

Table (4.5) Reported features with large, medium, and small effect sizes between the false positives and true negatives of the prognostic models. The remaining variables represent the number of features on which the approach was applied. These features are not used for training.

From the table above, we can see that three prognostic models reported features with large effect sizes, and each model reported features with medium and small effect sizes. The three models that reported large effect sizes are the models trained on the CDSS questionnaire, UKU questionnaire, and the knowledge-based variables.

Models trained on (No. variables)	No. remaining variables	Large ES	Medium ES	Small ES
All (231)	0	-	-	-
Diagnose & Treatment (65)	166	0 (0.0%)	0 (0.0%)	50 (30.1%)
Demographic (25)	206	0 (0.0%)	1 (0.5%)	64 (31.1%)
All questionnaires (133)	98	0 (0.0%)	4 (4.1%)	20 (20.4%)
PANNS (30)	201	0 (0.0%)	7 (3.5%)	45 (22.4%)
CDSS (9)	222	5 (2.3%)	14 (6.3%)	88 (39.6%)
UKU (60)	171	0 (0.0%)	0 (0.0%)	64 (37.4%)
Expert (6)	225	13 (5.8%)	32 (14.2%)	55 (24.4%)
Elastic net (28)	203	2 (1.0%)	10 (4.9%)	65 (32.0%)

Table (4.6) Reported features with large, medium, and small effect sizes between the false negatives and the true positives of the prognostic models. The remaining variables represent the number of features on which the approach was applied. These features are not used for training.

Table 4.6 shows that all the prognostic models with a positive target label reported remaining features with small effect sizes, six models reported features with medium effect sizes and three models reported features with large effect sizes. The three models that reported large effect sizes are the model trained on the knowledge-based variables, the model trained on CDSS variables and the models trained on the variables that were selected by the elastic net model.

The features that reported large effect sizes could be further analysed in order to identify more of the specifics of the incorrect classifications. The reported features with large effect sizes are listed below in table 4.7.

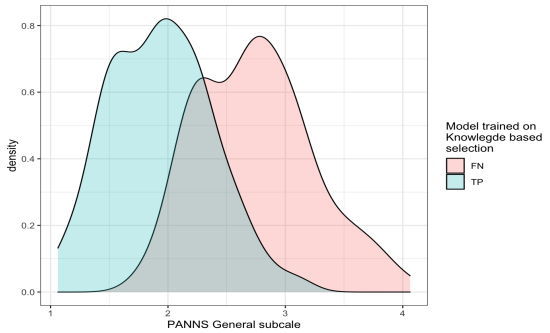
Positive class - remission achieved		**Negative class - no remission achieved**	
Feature	**Model with large feature**	**Feature**	**Model with large feature**
PANNS general subscale	expert	PANNS general subscale	expert
PANNS negative subscale	expert	PANNS negative subscale	expert
PANNS positive subscale	expert	PANNS positive subscale	expert
PANNS N4	expert	PANNS N4	expert
PANNS G13	expert	PANNS G13	expert
PANNS N1	expert	PANNS N2	expert
PANNS N3	expert	PANNS N7	expert
PANNS N6	expert	PANNS G11	expert
PANNS N7	expert	PANNS G15	expert
PANNS G11	expert	PANNS N3	expert
PANNS G15	expert	SWN 20	CDSS
PSP B	expert	PANNS G6	CDSS
PANNS N2	expert	PANNS G12	UKU
PANNS G6	CDSS	PANNS total score	UKU
SWN 4	CDSS		
SWN 9	CDSS		
UKU PS 1 5DGR	CDSS		
PANNS P3	CDSS		
MINI C - Suicidality	Elastic net		
CGI severity	Elastic net		

Table (4.7) Reported features of the prognostic models with large effect sizes. Additionally, there is reported in which model the large effect size was reported.

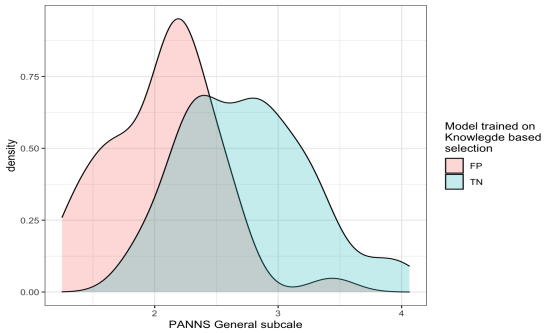
As can be seen from table 4.7, most of the large effect sizes were reported by the model trained on the knowledge-based variables. Most of the features with a large effect size are PANNS variables. This applies to both the positive class and the negative class. Some features had a large effect size for both classes. However, no feature with large effect size was reported by multiple models.

The features with large effect sizes could also be visualised with density plots or histograms. This type of visualisation could help with the interpretability of the results. The features with continuous values are illustrated with density plots, whereas features with discrete values are visualised with histograms. Figure 4.1 shows six features that reported large effect sizes. These features differ in some aspects: some features were PANNS variables (a-d) and some features were SWN[1] variables (e-f), some features had continuous values (a-b) and some features had discrete values (c-f). Furthermore, the features were also reported by different models.

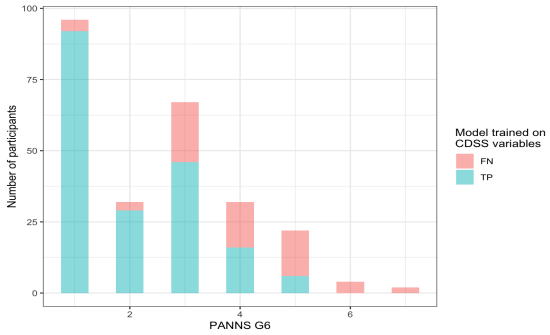
[1]SWN is a questionnaire about subjective wellbeing.



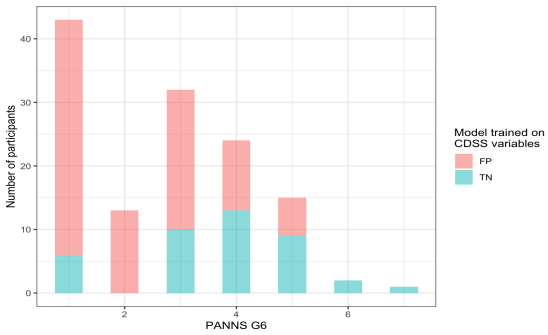
(a) Large effect size
Model trained on knowledge-based variables
Positive class label - remission achieved



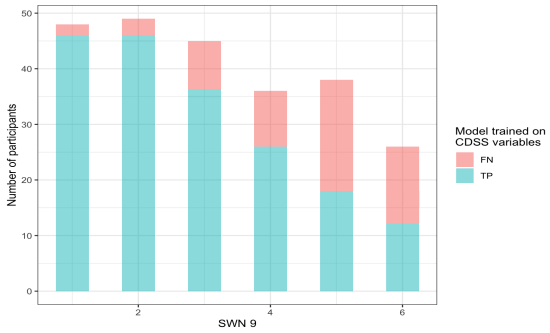
(b) Large effect size
Model trained on knowledge-based variables
Negative class label - no remission achieved



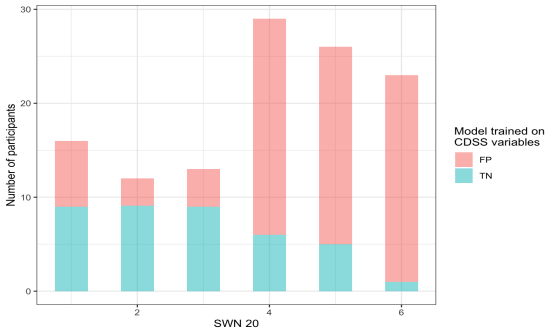
(c) Large effect size
Model trained on CDSS variables
Positive class label - remission achieved



(d) Large effect size
Model trained on CDSS variables
Negative class label - no remission achieved



(e) Large effect size
Model trained on CDSS variables
Positive class label - remission achieved



(f) Large effect size
Model trained on CDSS variables
Negative class label - no remission achieved

Figure (4.1) Illustrations of the features that reported large effect sizes. The features with continuous values are presented by density plots (a-d) and the features with discrete values (e-f) are presented by histograms. All features with continuous values are features that are reported by the model trained on knowledge-based variables (a-b). All the histograms are features reported by the model trained on the CDSS variables (c-f). The left figures (a,c,e) illustrate the difference between the correct and incorrect classification with a positive class label, the participants who met the remission criteria after four weeks. The figures on the right (b,d,f) visualise the difference between the correct and incorrect classifications with a negative class label, the patients who did not achieve remission. The top figures show the feature *PANNS general subscale* for the true and false class labels of the model trained on the knowledge-based variables of the positive class (a) and of the negative class labels (b). In figure (c) and (d) the feature *PANNS G6* is illustrated for both the positive class (c) as well as the negative class (d) of the model trained on the CDSS variables. Figure (e) illustrates the feature *SWN 9* for the positive class of the model trained on the CDSS variables. Figure (f) visualises the feature *SWN 20* of the patients with a negative class label.

Figures 4.1(a-d) show that in the models trained on the knowledge-based variables and CDSS variables the false negatives, the patients who achieved remission but were incorrectly predicted not to achieve remission, had relatively higher PANNS scores at baseline compared to true positives, the participants who were correctly predicted to achieve remission. On the other hand, the

false positives, the patients who did not reach remission, but were incorrectly predicted to reach remission, scored lower on the PANNS scores at baseline compared to the true negatives, the participants that were correctly predicted not to reach remission. Furthermore, figures 4.1(e-f) show that the misclassified participants of both class labels had relatively higher SWN scores compared to the correctly classified participants.

Diagnostic models

The number of the remaining features with a small, medium, and large effect size of all the diagnostic models are presented in table 4.8 and 4.9. Table 4.8 presents the effect sizes between the correct and incorrect classifications with a negative class label, and table 4.9 the effect sizes between the correct and incorrect classifications with a positive class label.

Models trained on (No. variables)	No. remaining variables	Large ES	Medium ES	Small ES
All (245)	0	-	-	-
Left hemisphere (121)	124	0 (0.0%)	4 (3.2%)	35 (28.2%)
Right hemisphere (121)	124	0 (0.0%)	0 (0.0%)	19 (15.3%)
Subcortical volumes (15)	230	0 (0.0%)	0 (0.0%)	45 (19.6%)
Large volumes (11)	234	0 (0.0%)	2 (0.9%)	81 (34.6%)
ROI - Thickness (68)	177	0 (0.0%)	3 (1.7%)	66 (37.3%)
ROI - Volume (68)	177	0 (0.0%)	0 (0.0%)	31 (17.5%)
ROI - Area (68)	177	0 (0.0%)	0 (0.0%)	29 (16.4%)
Elastic net (41)	204	0 (0.0%)	0 (0.0%)	52 (25.5%)

Table (4.8) Reported features with large, medium, and small effect sizes between the false positives and the true negatives of the diagnostic models. The remaining variables represent the number of features on which the approach was applied. These features are not used for training.

It can be seen from the data in table 4.8 that the participants with a negative class label, the healthy controls, reported not a single feature with a large effect size. Three of the diagnostic models reported features with medium effect sizes, and all models reported features with small effect sizes.

Models trained on (No. variables)	No. remaining variables	Large ES	Medium ES	Small ES
All (245)	0	-	-	-
Left hemisphere (121)	124	3 (2.4%)	12 (9.7%)	43 (34.7%)
Right hemisphere (121)	124	0 (0.0%)	2 (1.6%)	26 (21.0%)
Subcortical volumes (15)	230	0 (0.0%)	3 (1.3%)	95 (41.3%)
Large volumes (11)	234	0 (0.0%)	35 (15.0%)	124 (53.0%)
ROI - Thickness (68)	177	0 (0.0%)	2 (1.1%)	50 (28.2%)
ROI - Volume (68)	177	0 (0.0%)	3 (1.7%)	53 (30.0%)
ROI - Area (68)	177	1 (0.6%)	29 (16.4%)	98 (55.4%)
Elastic net (41)	204	0 (0.0%)	12 (5.9%)	69 (33.8%)

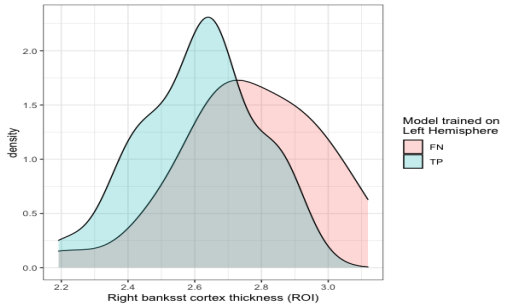
Table (4.9) Reported features with large, medium, and small effect sizes between the false negatives and the true positives of the diagnostic models. The remaining variables represent the number of features on which the approach was applied. These features are not used for training.

As shown in the table above, two models reported features with large effect sizes between the correctly and incorrectly classified patients with schizophrenia. The model trained on the left hemisphere reported three features with large effect sizes. The model trained on the surface areas of the ROIs reported one feature with a large effect size. The four features are listed below in table 4.10.

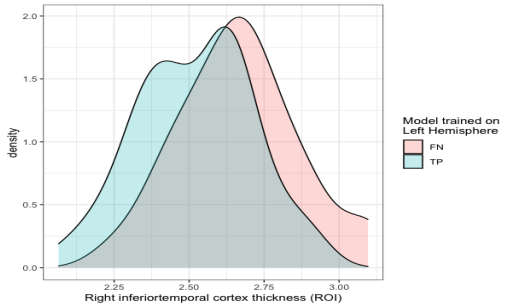
Positive class - patients with schizophrenia		Negative class - healthy controls	
Feature	**Model with large feature**	**Feature**	**Model with large feature**
Right banksst cortex thickness	Left hemisphere	-	-
Right inferior temporal cortex thickness	Left hemisphere	-	-
Right superior temporal cortex thickness	Left hemisphere	-	-
Right entorhinal cortex volume	ROI - Area	-	-

Table (4.10) Reported features with large effect sizes of the diagnostic models. Additionally, there is reported in which model the large effect size was reported.

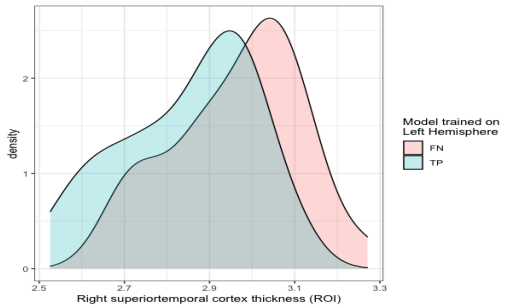
As can be seen from table 4.10, different features resulted in large effect sizes. All the features that reported large effect sizes were from the right hemisphere. Figure 4.2 visualises the four reported features with large effect sizes.



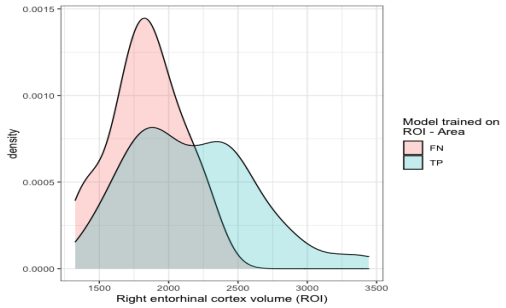
(a) Large effect size
Model trained on left hemisphere
Positive class label - patients with schizophrenia



(b) Large effect size
Model trained on left hemisphere
Positive class label - patients with schizophrenia



(c) Large effect size
Model trained on left hemisphere
Positive class label - patients with schizophrenia



(d) Large effect size
Model trained on surface areas ROI
Positive class label - patients with schizophrenia

Figure (4.2) Density plots of the features with large effect sizes. Figures (a,b,c) visualised the features with large effect sizes reported by the model trained on variables of the left hemisphere. Figure (d) illustrates the feature with the large effect size reported by the model trained on the surface areas of the ROIs. All figures illustrate the difference between the correct and incorrect classified patients with schizophrenia. (a) illustrates the feature *right banksst cortex thickness*, (b) the feature *right inferiortemportal cortex thickness*, and (c) the feature *right inferior temporal cortex thickness*. (d) *right entorhinal cortex volume* illustrates the features with a large effect size reported by the model trained on the ROI surface areas.

From the data in figure 4.2, we can see that the sizes of the brain regions of the misclassified patients that were reported by the model trained on the left hemisphere variables were relatively larger compared to the correctly classified patients, whereas the size of the brain region of the misclassified patients reported by the model trained on the surface areas of the ROIs was relatively smaller compared to the correctly classified patients.

4.3 Approach 2

In the second approach, we first evaluate the overlap when two models are compared on their misclassifications. Subsequently, we extend the analysis to compare multiple models with regards to their misclassifications.

Overlap of misclassified patients of two models

To evaluate the degree of overlap in misclassifications of two models, confusion matrices and Venn-diagrams have been used. As an illustration, two prognostic models and two diagnostic models have been selected to show this. From the prognostic models, we selected two models that were trained on two different questionnaires: the model trained on the UKU questionnaire and model trained on the CDSS questionnaire. From the diagnostic models, the model trained on the left hemisphere and the model trained on the right hemisphere are compared. All these models showed approximately similar performances.

The confusion matrices and the Venn-diagrams of the selected models are listed below. Table 4.11 and 4.12 and figure 4.3 and 4.4 illustrate the overlap of the misclassified patients of the two prognostic models and table 4.13 and 4.14 and figure 4.5 and figure 4.6 the overlap of the two diagnostic models.

Prognostic models

		UKU	
		FP	TN
CDSS	FP	66	24
	TN	23	18

Table (4.11) Confusion matrix - Negative class

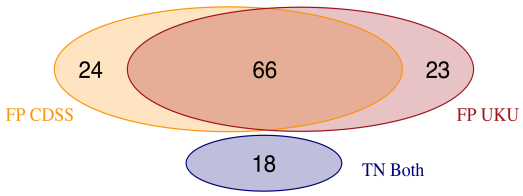


Figure (4.3) Venn-diagram - Negative Class

		UKU	
		FN	TP
CDSS	FN	16	44
	TP	56	139

Table (4.12) Confusion matrix - Positive class

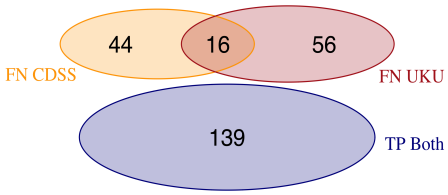


Figure (4.4) Venn-diagram - Positive Class

As shown in table 4.11 and figure 4.3, 24 of the 90 participants who were misclassified in the model trained on the CDSS variables were correctly classified by the model trained on UKU variables. 23 of the 89 misclassified participants in the model trained on the UKU variables were correctly classified in the model trained on the CDSS variables. In total, 66 of the 131 (50.3%) participants who did not meet the remission criteria were misclassified in both prognostic models. When only taking the misclassifications into account, 58.4% was misclassified in both models.

Table 4.12 and figure 4.4 show that of the 60 patients that were misclassified by the model trained on the CDSS questionnaire, 44 of them were correctly classified by the model trained on the UKU variables. Of the 72 misclassified participants in the model trained on the UKU questionnaire, 56 were not misclassified in the model trained on the CDSS variables. In total, 16 of the 255 (7.1%) patients who met the remission criteria were misclassified in both datasets. This percentage was 18.6% when excluding the participants who were correctly classified in both models.

Diagnostic models

		Right hemisphere	
		FP	TN
Left hemisphere	FP	21	31
	TN	33	119

Table (4.13) Confusion matrix - Negative Class

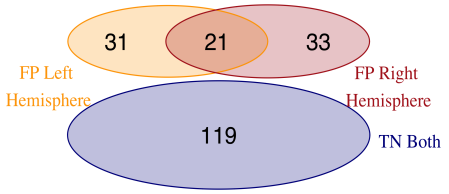


Figure (4.5) Venn-diagram - Negative Class

		Right hemisphere	
		FN	TP
Left hemisphere	FN	14	14
	TP	19	37

Table (4.14) Confusion matrix - Positive Class

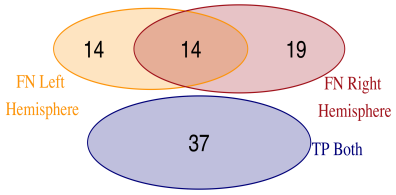
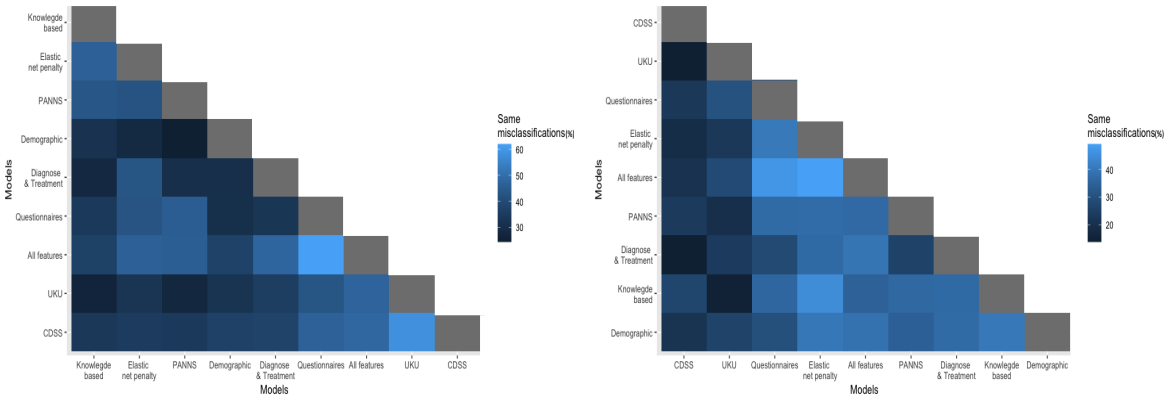


Figure (4.6) Venn-diagram - Positive Class

It can be seen from the data in table 4.13, that of all 204 healthy controls, 21 (10.3%) were misclassified in both models. When only taking the misclassified healthy controls into account, 24.7% was misclassified in both models. Of all the 84 patients with schizophrenia 14 patients (16.7%) were misclassified in both models. Of the 47 misclassified patients with schizophrenia, 29.7% was misclassified in both models. Both confusion matrices and Venn-diagrams showed that the overlap in misclassified participants was relatively low for both classes.

The degree of overlap of misclassified participants of all the models is illustrated by lower triangular heatmaps in figure 4.7 and in figure 4.8. Figure 4.7 illustrates prognostic models trained on the OPTiMiSE dataset and figure 4.8 the diagnostic models trained on the GROUP dataset. In these heatmaps, only the misclassifications are taken into account (see equation 2.11).

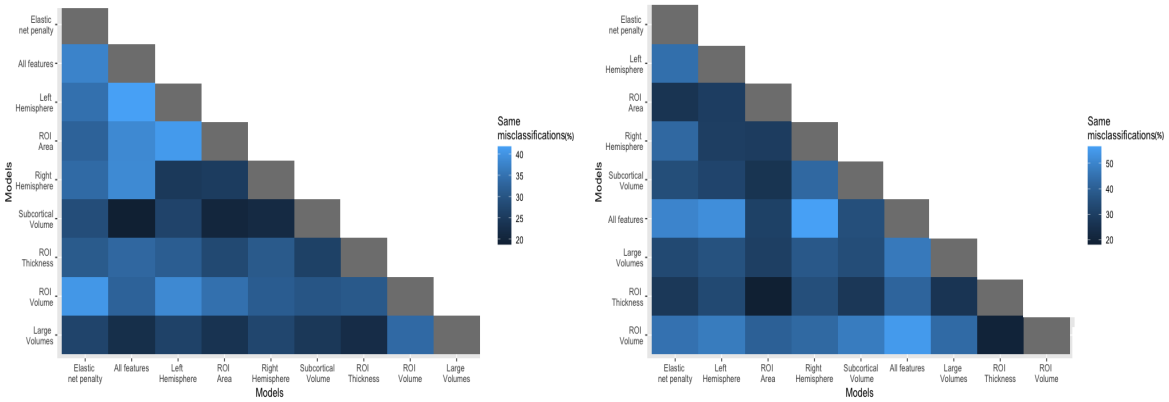


(a) Negative class - no remission achieved

(b) Positive class - remission achieved

Figure (4.7) Lower triangular heatmaps that visualise the overlap of prognostic models trained on subsets of the OPTiMiSE dataset.

As shown in figure 4.7a, of all the prognostic models from the OPTiMiSE dataset, the least overlap of the false positives, the patients who did not achieve remission but were incorrectly predicted to reach remission, was reported between the model trained on the PANNS variables and model trained on the demographical variables (24.4%). The most overlap of the false positives was reported between the model trained on all the variables and the model trained on all the clinical questionnaires (62.0%). As shown in figure 4.7b, the least overlap of false negatives, the patients who achieved remission but who were incorrectly predicted not to achieve remission, was reported between the models trained on diagnosis and treatment variables and CDSS variables (15.8%) and the models trained on the UKU variables and CDSS variables (15.9%), whereas the most overlap was reported between the models trained on variables selected by the elastic net model and the model trained on all the variables of the OPTiMiSE dataset (48.2%).



(a) Negative class - healthy controls **(b)** Positive class - patients with schizophrenia

Figure (4.8) Lower triangular heatmaps that visualise the overlap of diagnostic models trained on subsets of the GROUP dataset.

Of all the diagnostic models from the GROUP dataset, the least overlap of the false positives, the healthy controls that were incorrectly predicted, was reported between the model trained on all the variables of the GROUP dataset and the model trained on the subcortical volumes (18.7%). The most overlap was reported between the models trained on all the variables and the models trained on the left hemisphere variables (41.8%). Of the false negatives, the misclassified patients with schizophrenia, the least overlap was reported between the models trained on the cortical thickness and the volumes of the ROIs (18.8%). The most overlap of false negatives was shown between the model trained on the right hemisphere variables and the model trained on all the variables of the GROUP dataset (56.5%).

As can be seen in the figures above, the misclassified healthy controls of diagnostic models had, in general, less overlap compared to the misclassified patients with schizophrenia of the diagnostic models. The prognostic models trained on variables of the OPTiMiSE dataset had relatively more overlap compared to the diagnostic model.

Overlap of misclassified patients of multiple models

To compare multiple models on their misclassifications, it was observed in how many models the participants were misclassified. When the participant was not misclassified in a single model, the number of models in which the participant was misclassified was zero. When the participant was misclassified in all models, the participant was misclassified in nine models.

Prognostic models

In figure 4.9 two histograms are illustrated. The histograms visualise the number of prognostic models in which the participants were misclassified. Figure 4.9a illustrates this for the patients who did not reach remission and figure 4.9b illustrates this for the patients who did achieved remission.

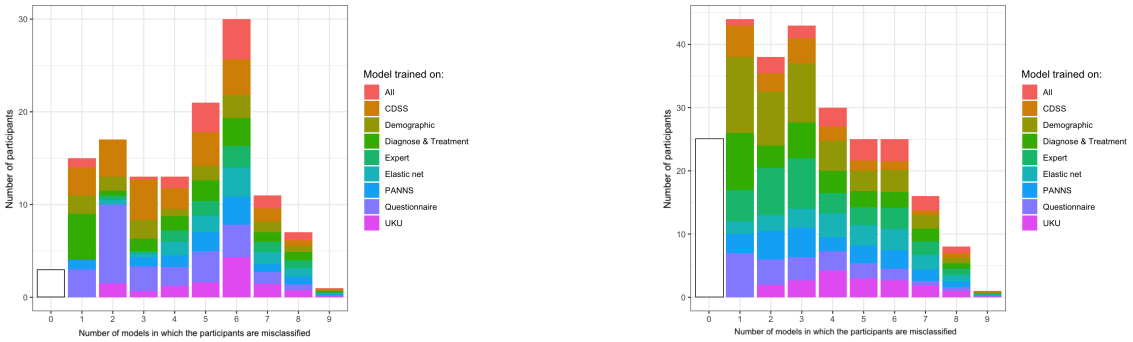
(a) Negative class - no remission achieved **(b)** Positive class - remission achieved

Figure (4.9) Number of participants that are misclassified in k prognostic models where $0 \leq k \leq 9$. (a) illustrates all 131 patients who had not achieved remission after four weeks. (b) illustrates all the 255 patients who had achieved remission after four weeks.

Figure 4.9a shows that, for instance, of all the patients who did not reach remission, 15 patients were misclassified in one of the nine models, 17 patients were misclassified in two of the nine models, and only one participant was misclassified in all the models. Most patients who did not achieve remission were misclassified in six models. Interestingly, the number of participants that are misclassified in seven models is relatively low. Most patients who reached remission, were misclassified in one or three models (figure 4.9b).

Figure 4.10 shows in which models the misclassifications occurred; each colour indicates one of the nine models. Given the total number of misclassified participants in k models, the colour indicates the contribution of a particular model to this total. Essentially, each bin is scaled down such that is corresponds to the number of participants instead of the total number of misclassifications.

For example, 17 participants were misclassified in two models. Hence, in total, 34 misclassifications occurred, which can be decomposed into the contribution of each model. Figure 4.10a shows that half of these misclassifications occurred in the model trained on all the questionnaire items. This means that of all participants who were misclassified in two models, all were misclassified in the models trained on the questionnaire items, and additionally in another model. The remaining colours show in which other models the participants were misclassified.



(a) Negative class - no remission achieved **(b)** Positive class - remission achieved

Figure (4.10) Number of the prognostic models in which the participants were misclassified. The colours present in which models the participants were misclassified. The size of the colours represents the proportion. (a) illustrates all 131 patients who had not achieved remission after four weeks. (b) illustrates all the 255 patients who had achieved remission after four weeks.

Figure 4.10 shows that participants who did not meet the remission criteria and that were only misclassified in a single model, were mostly misclassified in the model trained on the diagnoses and treatment variables. The participants who achieved remission and that were only misclassified in a single model were mostly misclassified in the model trained on the demographic variables. These models seem to have less overlap in misclassification compared to the other models. Another way to visualise in which models the different participants were misclassified is by heatmaps. In table B.1 of the appendix, the heatmaps of these nine prognostic models can be found.

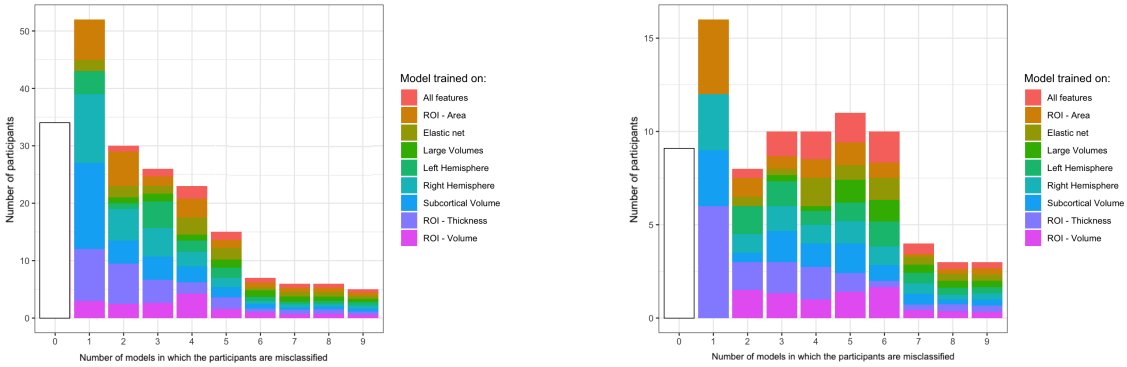
Diagnostic models

In figure 4.11, the number of diagnostic models in which the participants were misclassified is shown. Additionally, figure 4.12 illustrates in which models the participants were misclassified.

(a) Negative class - healthy controls **(b)** Positive class - patients with schizophrenia

Figure (4.11) Number of participants that are misclassified in k diagnostic models where $0 \leq k \leq 9$. (a) represents the 204 healthy controls. (b) represents the 84 patients with schizophrenia.

Figure 4.11a shows that most healthy controls were misclassified in one of the nine models. Of all the healthy controls, relatively few participants were misclassified in more than half of all the models. As shown in figure 4.11b, the patients of the GROUP dataset were also mostly misclassified in a single model. However, relatively more patients were misclassified in multiple models compared to the healthy controls.



(a) Negative class - healthy controls **(b)** Positive class - patients with schizophrenia

Figure (4.12) Number of the diagnostic models in which the participants were misclassified. The colours present in which specific models the participants were misclassified. The size of the colours represents the proportion. (a) represents the 204 healthy controls. (b) represents the 84 patients with schizophrenia.

Figure 4.12 shows that the participants that were only misclassified in one model were mostly misclassified in the model trained on the subcortical volumes, the model trained on the right hemisphere variables, and the models trained on the cortical thickness and surface areas of the ROIs. In table B.2 of appendix B this is illustrated by heatmaps.

4.4 Approach 3

In this section, we first show the mean distance to the threshold of the correct and incorrect classifications. Some of these distances have been illustrated by scatter plots. Furthermore, the misclassified patients in the prognostic models were evaluated on the mean score of the remission criteria. Unfortunately, there was no further information available on the severity of symptoms of the GROUP sample. Otherwise, the correct and incorrectly classified patients with schizophrenia of the diagnostic models could have been analysed on the severity of the symptoms.

Distance to threshold

The ML models used in this study reported the class probability of the positive class label for each participant. When this value was equal or greater than the threshold, the participant was predicted to have a positive label. Otherwise, the participant was predicted to have a negative class label. To compute the distance to threshold, the difference between the class probability of the positive class and the threshold is computed.

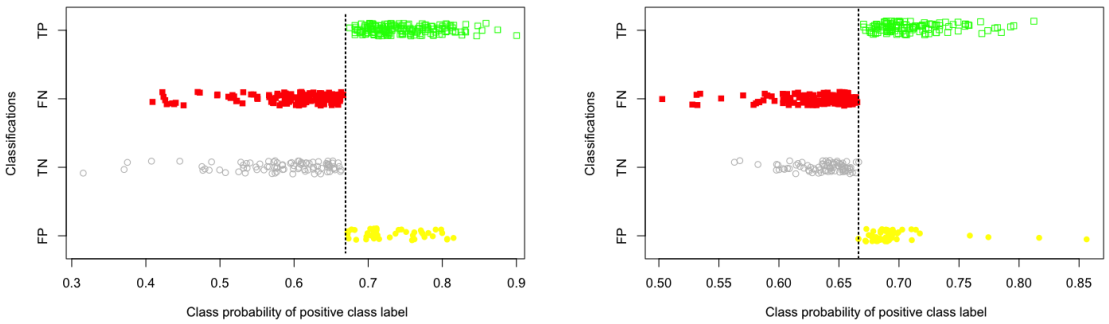
Prognostic models

Most prognostic models reported a threshold between 0.6 and 0.7. In table 4.15, an overview of the absolute mean distance to the threshold of the false positives, false negatives, true positives, and the true negatives of all the prognostic models is presented.

Model trained on	Mean FN	Mean FP	Mean TP	Mean TN
All variables	0.053	0.065	0.067	0.054
Diagnoses & treatment	0.028	0.024	0.016	0.027
Demographics	0.038	0.032	0.040	0.030
All questionnaires	0.085	0.069	0.059	0.111
PANNS	0.071	0.075	0.073	0.086
CDSS	0.019	0.012	0.012	0.022
UKU	0.015	0.005	0.007	0.015
Knowledge-based	0.064	0.062	0.076	0.082
Elastic net	0.110	0.099	0.104	0.131

Table (4.15) Absolute mean distance to the threshold of the prognostic models. Each row presents the absolute mean distance of the false negatives, the false positives, the true positives, and true negatives of a model. The difference between the class probability of the positive class and the threshold was used to compute the distance to the threshold.

As can be seen from table 4.15, the prognostic models reported different mean distances to the threshold for the false positives, false negatives, true positives, and true positives. In some models such as the model trained on the knowledge-based variables, the mean distance to the threshold of the incorrect classifications was smaller than the mean distance to the threshold of the correct classifications. Other models reported that the mean distance to the threshold of the correct classifications was smaller (e.g. the model trained on the demographic variables reported that the mean distance to the threshold of the true negatives was smaller compared to the mean distance to the threshold of the false negatives). Hence, table 4.15 shows that the mean distance to the threshold of the incorrect classifications was not always smaller than the mean distance to the threshold of the correct classifications of the prognostic models. To get a better understanding, we also plotted the continuous values of two prognostic models.



(a) Model trained on knowledge-based variables **(b)** Model trained on demographic variables

Figure (4.13) The distance to the threshold of the true positives, false negatives, true negatives, and false positives of two prognostic models. The values on the y-axis represent the class probability of the positive class of the participants. Each point represents a participant. The threshold is illustrated by the dotted line. Participants on the left of the threshold are predicted not to reach remission, whereas participants on the right of the threshold are predicted to reach remission.

Figure 4.13 shows that the misclassified as well as the correctly classified participants of the model trained on the demographic variables had, in general, a smaller absolute mean distance to the threshold than the model trained on the knowledge-based variables.

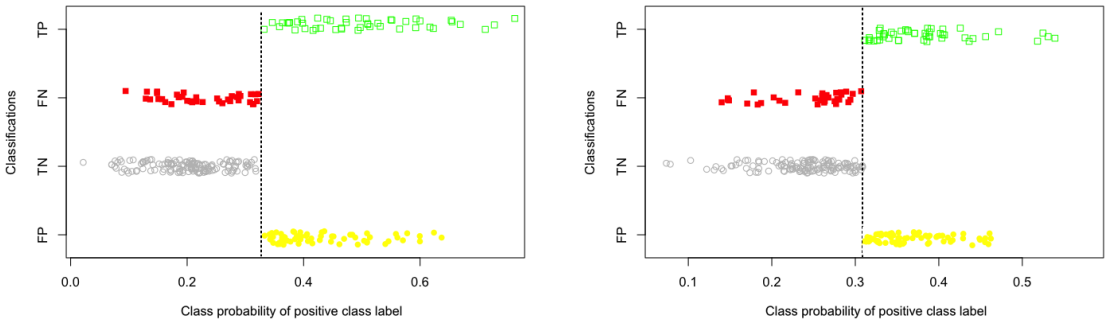
Diagnostic models

Most diagnostic models reported a threshold between 0.3 and 0.4. In table 4.16, the absolute mean distance to the threshold of the false negatives, false positives, true positives, and true negatives of the diagnostic models is shown.

Model trained on	Mean FN	Mean FP	Mean TP	Mean TN
All variables	0.053	0.065	0.067	0.054
Subcortical volumes	0.090	0.080	0.088	0.094
Large volumes	0.055	0.059	0.063	0.054
ROI - volume	0.095	0.100	0.167	0.117
ROI - thickness	0.061	0.057	0.074	0.070
ROI - area	0.055	0.137	0.124	0.071
Right hemisphere	0.124	0.089	0.128	0.130
Left hemisphere	0.085	0.133	0.168	0.121
Elastic net	0.120	0.190	0.280	0.172

Table (4.16) Absolute mean distance to the threshold of the diagnostic models. Each row presents the absolute mean distance of the false negatives, the false positives, the true positives, and true negatives of a model. The difference between the class probability of the positive class and the threshold was used to compute the distance to the threshold.

As can be seen in table 4.16, the absolute mean distance to the threshold of the correctly and incorrectly classified patients differed little from each other. In most models, the mean distance to the threshold was smaller for the misclassified patients compared to the correctly classified patients. The diagnostic models trained on the volumes and cortical thickness of the ROIs are shown as an illustration in figure 4.14.



(a) Model trained on the volumes of the ROIs **(b)** Model trained on cortical thickness of the ROIs

Figure (4.14) The distance to the threshold of the true positives, false negatives, true negatives and false positives of two diagnostic models. The values on the y-axis represent the class probability of the positive class of the participants. Each point represents a participant. The threshold is illustrated by the dotted line. Participants on the left of the threshold are classified by the models as healthy controls, whereas participants on the right of the threshold are classified as patients with schizophrenia.

As shown in the figures above, the distance to the threshold of misclassifications of both diagnostic models is relatively smaller compared to the distance of the correct classifications. The model trained on the volumes of the ROIs reported relatively more variation in distances compared to the model trained on the cortical thickness of the ROIs.

Severity of symptoms

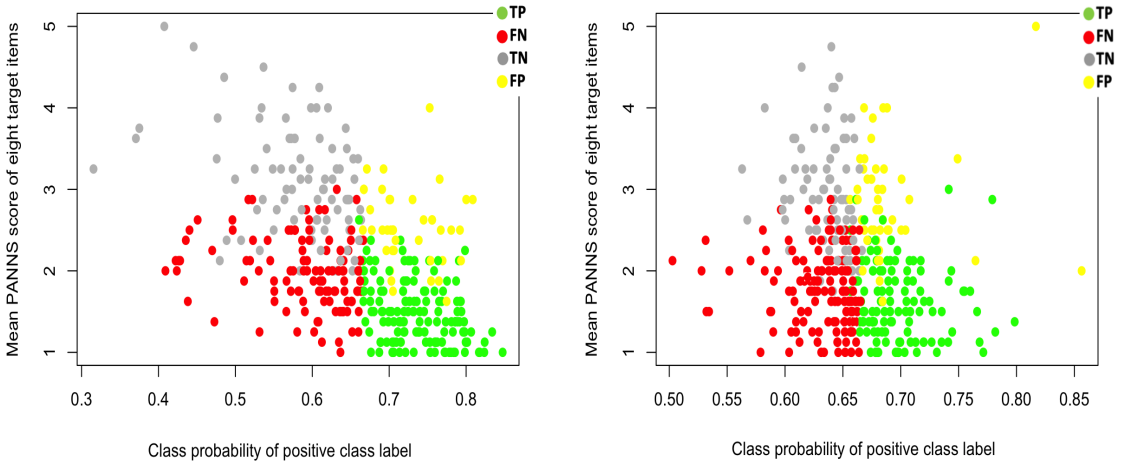
As mentioned earlier, whether a patient achieved remission was determined by eight PANNS variables. According to the definition of the remission criteria, a patient had achieved remission when all the eight items (items P1, P2, P3, N1, N4, N6, G5, and G9) had at most a score of three, meaning that these eight variables were at most mildly present. The mean of these eight variables is computed for the false positive, the false negatives, the true positives, and the true negatives of each model. Table 4.17 provides an overview of all the nine prognostic models.

Model trained on	Mean FN	Mean TP	Mean FP	Mean TN
All variables	1.9	1.6	2.9	3.0
Diagnoses & treatment	1.8	1.7	2.9	3.0
Demographics	1.8	1.6	2.8	3.0
All questionnaires	1.9	1.6	2.9	2.9
PANNS	1.9	1.6	2.7	3.0
CDSS	1.8	1.7	2.9	3.0
UKU	1.6	1.7	2.8	3.0
Knowledge-based	2.0	1.5	2.5	3.1
Elastic net	1.9	1.6	2.7	3.1

Table (4.17) Mean PANNS score of eight target items (items P1, P2, P3, N1, N4, N6, G5, and G9). These eight items determined whether a patients achieved remission after four weeks. Each row presents the mean score of the false positives, true positives, false negatives, and true negatives of a prognostic model.

What stands out in the table is that in all models, the mean score of the false negatives was relatively higher than mean score of the true positives. On the other hand, the false positives reported relatively lower mean scores compared to the true negatives.

Furthermore, as an illustration, we plotted the class probability of the positive class label against the mean of the eight target variables for two models.



(a) Model trained on knowledge-based variables **(b)** Model trained on demographic variables

Figure (4.15) Plot of class probability of the positive class label against the mean score of the eight variables that determine whether a patient reached remission. The colours show whether a patient is a true positives, false negative, true negative, or false positive.

As can be seen from the figures above, there is a stronger link between the class probability of the positive class and the target score of the model trained on the knowledge-based variables compared to the model trained on the demographic variables. This could be due to the fact that the knowledge-based model is trained on the total PANNS score at baseline.

Chapter 5

Discussion

In this chapter, we first discuss the performances of the ML models that were used in this study. Thereafter, we analyse who the model's misclassifications are based on the results of the previous chapter. We then go on to the advantages and limitations of our approaches. Furthermore, we discuss what other factors can influence the performances, what other factors have to be taken into consideration, and suggestions for further research.

5.1 Performance of the models

In this study, we created two types of models: diagnostic models that classified whether a subject was a patient with schizophrenia or a healthy control, and prognostic models that classified whether a patient had achieved remission after four weeks or not. The diagnostic models reported balanced accuracies in the range of 60.5-75.7%, whereas the prognostic models reported balanced accuracies in the range of 51.9-65.4%. Hence, the prognostic models reported relatively more misclassifications. A possible explanation for this result might be that prognostic targets could be more challenging than predictions of the current state, such as diagnoses, since future outcomes could be influenced by unknown factors (Schnack, 2020). In addition, the diagnostic and prognostic models used different variables for training. The diagnostic models used variables obtained by MRI-scans, whereas the prognostic models were trained on sociodemographic and clinical variables. The higher performances of the diagnostic models indicate that the variables of the diagnostic models had a stronger link with the target than the variables of the prognostic models.

Within the diagnostic and prognostic models, the models also demonstrated some differences in performances. Although these models were trained on the same sample, some models reported higher performances than others. This highlights the importance of feature selections. This importance is clearly shown by the models that were trained on the feature selections of the elastic net model. For both diagnostic and prognostic models, the models that were trained on the features that were selected by the elastic net model yielded the highest performances.

However, note that the performances of these models that were trained on the feature selections of the elastic net models might be over-optimistic given the fact that information between patients used for training and validating the models could have been leaked. To enable an unbiased estimation of the variables, nested cross-validation could be used to prevent information leaking between participants used for training and validating the models (Koutsouleris et al., 2016). Further studies could analyse the variables that were selected by the elastic net models in more detail, given the fact that these variables seem to have a stronger link with the target compared to the other variables.

5.2 Approach 1

The purpose of our first approach was to determine whether there were features that showed significant differences between the correctly and incorrectly classified samples. To determine whether a feature had significantly different values for the misclassifications, we used the effect size. As mentioned earlier, to quantify the differences, we used Cohen's qualifications of the small (merely statistical), medium (subtle statistical), and large (obvious statistical) effect size. We are particularly interested in the large effect sizes.

The results demonstrated that there are models that have features with large, medium, and small effect sizes indicating that there are features with significantly different values for the misclassified sample. For instance, the prognostic models trained on the knowledge-based variables and the CDSS variables showed that the patients who did not achieve remission but were incorrectly predicted by the prognostic models to achieve remission had lower PANNS scores, compared to the patients who were correctly predicted not to achieve remission. This suggests that these misclassified patients had less severe symptoms at baseline. On the other hand, patients who achieved remission but were incorrectly predicted not to achieve remission had relatively higher PANNS scores than the patients who were correctly predicted to achieve remission, indicating that these misclassified patients had more severe symptoms at baseline. This implies that the patients who did not achieve remission but that were incorrectly predicted to achieve remission were patients with less severe symptoms at baseline that recovered little in the four remaining weeks, whereas the patients that reached remission but were incorrectly predicted not to reach remission seem to be patients with more severe symptoms at baseline that showed significant improvement after baseline. However, bear in mind that these symptoms, rated by the PANNS questionnaire, also formed the basis for the remission definition. Furthermore, the reported features with large effect sizes of the model trained on the knowledge-based variables must be interpreted with caution, because these reported features were mostly implicitly used for training since one of the knowledge-based variables was the total PANNS score. Therefore, the reported PANNS variables with large effect sizes were subscales of this total score.

Overall, the results showed that the prognostic models had relatively more features with large and medium effect sizes than the diagnostic models. In fact, the diagnostic models reported not a single feature with a large effect size between the correct and incorrectly classified healthy controls. However, the diagnostic models did report multiple features with large and medium effect sizes between the correctly and incorrectly classified patients with schizophrenia. These features showed that the misclassified patients had different sizes in some specific brain regions than the correctly classified patients. These results seem to indicate that the ML models misclassify a homogeneous subgroup in a heterogeneous patient sample. From a clinical point of view, it could be the case, for instance, that the patients who were incorrectly classified as healthy subjects may have less severe symptoms. For this reason, we could consider to use more than two target values in which, for example, a distinction could be made in the type of diagnoses. As mentioned earlier, all the patients had met the criteria of the DSM-IV for schizophrenia, schizophreniform disorder, or schizoaffective disorder. Nonetheless, these different diagnoses differ in some aspects. For example, schizophreniform disorder is similar to schizophrenia. However, functional impairment is not required, and the duration of the disturbance is by definition less than six months (Maj, Pirozzi, Formicola, Bartoli, & Bucci, 2000). By allowing misclassifications to be labeled with these other, perhaps less severe, types of diagnoses the model might perform better.

Note that in our approach, we compared the correct and incorrect classifications with the same class label. However, in order to find features that can be used to distinguish the incorrect from the correct predictions, the effect sizes with respect to the features between the correct and incorrect classifications with the same prediction should be computed. Features that are not used for training and that report large effect sizes between these correct and incorrect classifications could be added to the model and may improve the performance of the model. The features that were used for training the model could also be analysed in this case. When these features report large effect sizes, it would indicate that these features cause incorrect predictions since the models distinguish the subjects incorrectly. To improve these models, the variables should be removed. Further studies could analyse whether features reported small, medium, or large effect sizes between these correct and incorrect classifications and could evaluate whether these features were able to improve the performance of a model.

A limitation of this approach is that the reported effect sizes can be caused by noise. The

datasets contain relatively little participants. To reduce the noise, other large datasets with similar variables could be evaluated by this approach in order to analyse whether these models report similar features with large effect sizes between the correct and incorrect classifications. This is an important issue for future research. Another limitation is that this approach only analyses one feature at the time and no combinations of multiple features. It could be the case that the misclassifications can be better identified when analysing combinations of features. Further research should be undertaken to investigate this.

5.3 Approach 2

In the second approach, the purpose was to determine whether the same or different participants were misclassified in various models. This is in part answered by investigating the overlap in misclassifications. Comparing two models showed that in most models, the overlap of misclassified participants was relatively low, ranging from 20% to 50%. The results demonstrated that models that show similar performances misclassify different participants. For instance, the diagnostic models trained on the same brain regions but on different sides of the brain reported similar performances, however mostly misclassified different participants. Two prognostic models that were trained on two different questionnaires also reported different participants that were misclassified. These results also suggested that one participant can better be classified by one questionnaire (e.g. UKU) and the other participants by another questionnaire (e.g. CDSS).

Comparing multiple models on their misclassifications, showed that of the healthy subjects of the diagnostic models, relatively few subjects were misclassified in more than half of the models. The patients with schizophrenia in the diagnostic models are more often misclassified in multiple models. The prognostic models trained on the OPTiMiSE dataset reported a relatively high number of patients that were misclassified in multiple models. In particular, the patients with schizophrenia who did not reach remission showed a very high number of overlap in misclassifications. Future research could combine approach 1 and 2 in which there can be observed whether there are specific characteristics for the participants who were frequently misclassified compared to the participants that were rarely misclassified.

When only a few participants are misclassified in more than half of the models, the performance can be improved by creating a new model that takes the predictions of all the different models into account. If more than half of all the models predict a positive class, the final model predicts a positive model; otherwise, it predicts a negative class. Note that in this case the same participant should be used for training and that no information should leak between patients used for training and validating the models. Since the diagnostic models reported relatively few participants that were misclassified in more than half of the models, this type of model is created from all the diagnostic models. For each participant in the GROUP dataset, there is computed in how many models the participant was predicted to be a patient with schizophrenia. If a participant was predicted as a patient with schizophrenia in five or more models, the participant was classified as patients with schizophrenia. When a participants was predicted in less than five of the models to be patient with schizophrenia, the final model predicted that the participant was a healthy control. Note that in this case, the same participants have been used for training the model and the same participants have been used for validating the model. This final model showed promising results. The model reported an accuracy of 75.7% and a balanced accuracy of 72.0%.

A limitation of this approach is that the selection of the models can strongly influence the degree of overlap. For instance, the GROUP dataset compared models that had more overlap in training features. For example, the models trained on the left and right hemisphere had overlap in features with the other models. Some other selection may show different results of overlap. More research should be done on different combinations of models.

5.4 Approach 3

In the third approach, we analysed the effect of the threshold on the misclassified sample. We showed different ways to analyse this. First, we used a continuous value rather than the discrete output of the ML models. Here, we observed the difference between the continuous values and the threshold of the model. We analysed this for the misclassified as well as the correctly classified participants. The results demonstrated that the distance to the threshold of the misclassifications

was relatively small in most models. This indicated that the misclassifications are often influenced by the threshold. The distance to the threshold of the misclassifications and threshold was in most models relatively similar to the distance to the threshold of the correct classifications. In most models, the misclassifications did not lie closer to the threshold than the correct classifications. This indicates that the misclassifications in these models are not the ones closest to the threshold.

In this approach, we computed the absolute mean distance to quantify the distance to the threshold. However, the mean has some limitations. For instance, the mean is very sensitive to extremely low and extremely high values in the sample. In addition to the mean, other metrics could be used to analyse the distance to the threshold. Note that the approaches showed different ways to analyse the misclassifications, which in future research can be refined. For instance, further research could focus more on the misclassifications with the smallest distance to the threshold, since these misclassifications seem to be easier to improve than the other misclassifications (Alsallakh et al., 2014).

Furthermore, we used a continuous target score rather than the binary target score to analyse the prognostic models. As the target score, we used the mean score of the eight specific PANNS variables that were used to determine if a participant had achieved remission after four weeks. These results showed similar results as the first approach: the false positives of the prognostic models had a relatively lower mean PANNS score than the true negatives. On the other hand, the false negatives had a higher mean PANNS score than the true positives. However, in the first approach, the PANNS scores at baseline were used, while in this approach, the PANNS scores after four weeks were used. This indicates that misclassifications are not caused by the different values at baseline, but due to the definition of the target. According to Helldin et al. (2007), an explicit and unified definition of the concept of remission in schizophrenia is lacking. As mentioned earlier, according to this prognostic model, a patient has achieved remission when eight specific symptoms rated by the PANNS questionnaire were at most mildly present. This definition of remission is based on studies in the USA and Europe that produced a new proposal for remission (Kane, Leucht, Carpenter, & Docherty, 2003; Andreasen et al., 2005; Van Os et al., 2006). The premise was that no single item of the eight items should show a score above three (mildly present). However, if at least one of these items is above three, according to this definition, the patient has not met the remission criteria. Thus, although the patients show low PANNS scores for the remaining items, according to the definition, no remission is achieved. Note that in almost all the models, the mean of the false positive was lower than three. This indicates that most patients who did not meet the remission criteria but were predicted by the model to achieve remission have scored low on most of the remission criteria, but had at least one symptoms that was more than mildly present. Otherwise, the patients that were incorrectly predicted not to achieve remission may have scored relatively high on the eight PANNS items, but all eight items did not have a score above three. Therefore, to improve the model, the definition of remission could be reconsidered. Models could be trained on different definitions of remission in order to evaluate if better performances are achieved. Note that in this approach we only evaluated the mean of the eight items. Nevertheless, to evaluate the effect of the different items, the items could also have been analysed on all the variables individually or the sum of the eight values could have been used.

Further research could also analyse the severity of symptoms of the patients with schizophrenia in diagnostic models. There could be observed whether patients with less severe symptoms were more likely to be misclassified by a diagnostic model. Approach 2 and 3 could also be combined by observing if the participants that were frequently misclassified had more severe symptoms.

5.5 Possible causes

Worth noting is that when training and testing ML models that diagnose and prognose schizophrenia, many assumptions are made. For example, when selecting some variable, there is already assumed that these variables are predictors for the target. However, selecting variables is inherently susceptible to bias and noise. For instance, sociologically, bias may inadvertently be caused due to the fact that clinical research typically recruits patients that are exposed to psychiatric institutions, rather than individuals who also have mental problems but never are diagnosed. Consequently, subclinical individuals with schizophrenia, that never have been in contact with a psychiatrist might systematically evade research efforts (Bzdok & Meyer-Lindenberg, 2018). Besides, in many of the available variables, there could be an indirect link between the variable and target (Schnack, 2020). For instance, Van Oel et al. (2000) found that dermatoglyphic patterns (i.e. specific skin

marking on the hands and feet) are statistically related to schizophrenia. Nonetheless, this variable is probably an indirect effect. Presumably, this is caused by a genetic variation that influences the sensitivity of schizophrenia as well as dermatoglyphic patterns.

Furthermore, the choice of the target could include a bias. For instance, when creating a model that predicts whether a subject has schizophrenia or not, there is already assumed that a phenomenon such as schizophrenia exists. Besides, the target labels used for training and testing the ML models are assumed to be correct. However, the reliability of the labels used for training and testing is influenced by the quality of the expert. Unfortunately, there can safely be assumed that the accuracy of the prediction reduces around 10 percent when using experts for labelling (Schnack & Kahn, 2016). In fact, Regier et al. (2013) investigated that different clinicians gave different diagnoses, even if the same diagnostic system such as Diagnostic and Statistical Manual of Mental Disorders 5 (DSM-5) has been used, due to differences in interrater reliability. Therefore, in many cases, this means that the standard is rather silver than gold. Consequently, the target labels are less reliable, which weakens the link between the input features and target.

To overcome these limitations, unsupervised ML models can be used as an alternative in which the ML models try to detect underlying structures in the observations (Janssen et al., 2018). These underlying structures could be used to find (more) homogeneous subgroups in a heterogeneous patient sample. Subgroups could be based on similar clinical and/or biological characteristics. In this case, the underlying assumption is that a similar outcome between subjects is more probable when subjects have similar characteristics. One example of an unsupervised ML approach can be found in Jauhar et al. (2018), in which the authors used unsupervised ML algorithms for a dataset of 202 patients with functional psychosis. This study showed that unsupervised learning models were able to discriminate between different types of psychosis.

Another limitation in this study is that the OPTiMiSE dataset contained some missing values. To replace the missing values, the nearest neighbour-based imputation was used. However, these values may not always be correct. Nonetheless, these missing values were used for training and testing the ML models. This could cause incorrect predictions.

Besides, the datasets used for this study were relatively small. Not only this study, but also most other studies in psychiatry are often limited in deploying ML algorithms to diagnose or prognose psychiatric diseases due to the small sample sizes of today's datasets and their insufficient phenotypic detail (e.g., medical history, progression in treatment, symptoms, and response) (Bzdok & Meyer-Lindenberg, 2018).

5.6 Clinical utility

The different approaches helped to detect the causes and characteristics of the misclassified participants of the ML models. This insight could help to improve the performance of the models. However, to use the ML model for clinical practice, more than only good performance is required. In addition, it is desired that ML models are transparent on their diagnostic and prognostic knowledge and that the models can explain their predictions (Kononenko, 2001; Dwyer, Falkai, & Koutsouleris, 2018).

The clinicians should be able to analyse and understand the generated knowledge of the ML models in order to use the diagnostic and prognostic models in clinical practice. The clinicians want and need to know why the model makes certain predictions. Ideally, the automatically generated knowledge from an ML model would show a new point of view on a problem to a clinician and would be able to show new regularities and interrelations which a clinician did not see before in an explicit form. To take the suggestion of an unexpected solution for a physician into consideration, an ML model is required to explain its decisions.

The only situation in which clinicians would accept a black-box model is when a model outperforms all the other models and clinicians themselves by a very large margin. Nevertheless, such a situation is usually very unlikely (Kononenko, 2001). However, other fields than psychiatry that used machine learning show very promising results. For instance, Ciompi et al. (2017) trained ML models on chest CT scans to diagnose lung cancer. These models achieved performances similar to the best radiologists. Another example is a recent study of de Groof et al. (2020) in which they present an ML model that showed higher accuracies than 53 doctors from different European hospitals.

Another aspect that should be considered when aiming to apply these models into clinical practice is that the variables should not be too difficult to obtain. When ML models, for instance,

require too much training, or if they are not commercially viable, it is highly unlikely that these models will be used for clinical practice. Hence, it is very important to consider the real-world practicality of a model (Dwyer et al., 2018).

5.7 Further research

In addition to the suggestions that have been mentioned, further research should be undertaken to investigate the effect of different ML algorithms on the misclassifications. For this thesis, only one machine learning algorithm has been chosen for illustration. However, different algorithms may affect the misclassifications. Furthermore, instead of a binary classifier, models with multiple targets could be used to evaluate the misclassifications. Diagnostic models could, for instance, distinguish between different forms of the disease. For example, according to the DSM criteria, there are different forms of schizophrenia, such as schizophrenic, schizophreniform, or schizoaffective disorder. In the diagnostic models used for this study, the patients are treated as one group. However, these patients can have a wide variety of symptom presentation, functional outcome, and clinical course.

Furthermore, to improve the misclassification approaches and to make them even more valuable in the process of ML, it would be valuable if the approaches can be used in an interactive exploration environment. Note that the proposed approaches in this thesis were all a posteriori analyses. Ware et al. (2001) claimed that classifiers are able to compete with automated techniques if the users are able to integrate their domain knowledge in their classifier design.

Chapter 6

Conclusion

The aim of this study was to gain more insight into the misclassifications of the ML models. In this study, we investigated this for ML models that diagnosed and prognosed schizophrenia. Different approaches have been proposed to investigate the misclassifications. To the best of our knowledge, this is the first study that examines the misclassifications of these ML models in more detail, especially for schizophrenia. The analysis of the different approaches on the diagnostic and prognostic models that were trained on real-world datasets showed some promising results. The results demonstrated that in most models there are some features that showed significantly different values for the misclassified sample compared to the correctly classified sample with the same label. Follow-up research could investigate whether there are features that could also discriminate between the correct and incorrect classifications in order to improve the performances of the models directly. For example, this could be done through a further investigation of the differences between and within groups with the same label indicated by the models, rather than the groups based on the expert opinions. Furthermore, the results showed that in models, specific participants were frequently misclassified. This was the case especially in the prognostic models. These findings indicated that there is indeed something specifically different with this misclassified sample, compared to the correctly classified sample.

Rather than using the model as a black box, it seems to help to analyse the misclassifications, as it helps to identify the actual factors that could influence the misclassifications. For instance, our findings indicated that the incorrect prediction in the prognostic models could be caused by the definition of the remission criteria. Follow-up studies could evaluate whether changing these factors, such as the definition of the remission, could improve these models.

Overall, the different findings in the study could help to understand why a training process did not achieve the desired performance. Therefore, by investigating the errors, it can help to improve the ML models because it genuinely values the content-driven perspective from (medical) practitioners and researchers. The notion of examining the model's misclassification could also be used in other ML studies. It could be applied in many fields of interest, whether it concerns psychiatry or medicine in general or any other area of interest that deals with classifications. In all these fields, the analysis of the misclassified sample can help to understand and improve the ML models.

References

Abel, K. M., Drake, R., & Goldstein, J. M. (2010). Sex differences in schizophrenia. *International review of psychiatry*, *22*(5), 417–428.

Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H. T. (2012). *Learning from data* (Vol. 4). AMLBook New York, NY, USA:.

Alsallakh, B., Hanbury, A., Hauser, H., Miksch, S., & Rauber, A. (2014). Visual methods for analyzing probabilistic classification data. *IEEE transactions on visualization and computer graphics*, *20*(12), 1703–1712.

Andreasen, N. C., Carpenter Jr, W. T., Kane, J. M., Lasser, R. A., Marder, S. R., & Weinberger, D. R. (2005). Remission in schizophrenia: proposed criteria and rationale for consensus. *American Journal of Psychiatry*, *162*(3), 441–449.

Becker, L. A. (2000). Effect size (es). *Retrieved September*, *9*, 207.

Bleuler, E. (1911). *Dementia praecox: oder gruppe der schizophrenien.* F. Deuticke.

Bolin, J. E., & Finch, H. (2014). Supervised classification in the presence of misclassified training data: a monte carlo simulation study in the three group case. *Frontiers in psychology*, *5*, 118.

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(3), 223–230.

Ciompi, F., Chung, K., Van Riel, S. J., Setio, A. A. A., Gerke, P. K., Jacobs, C., & Ginneken, B. (2017). Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Scientific reports*, *7*, 46479.

Cohen, J. (1988). The effect size index: d. *Statistical power analysis for the behavioral sciences*, *2*(1).

Cooper, B. (1961). Social class and prognosis in schizophrenia. part ii. *British journal of preventive & social medicine*, *15*(1), 31.

de Groof, A. J., Struyvenberg, M. R., van der Putten, J., van der Sommen, F., Fockens, K. N., Curvers, W. L., & Baldaque-Silva, F. (2020). Deep-learning system detects neoplasia in patients with barrett's esophagus with higher accuracy than endoscopists in a multistep training and validation study with benchmarking. *Gastroenterology*, *158*(4), 915–929.

de Wit, S., Ziermans, T. B., Nieuwenhuis, M., Schothorst, P. F., van Engeland, H., Kahn, R. S., ... Schnack, H. G. (2017). Individual prediction of long-term outcome in adolescents at ultra-high risk for psychosis: Applying machine learning techniques to brain imaging data. *Human brain mapping*, *38*(2), 704–714.

Di Carlo, P., Pergola, G., Antonucci, L. A., Bonvino, A., Mancini, M., Quarto, T., & Blasi, G. (2019). Multivariate patterns of gray matter volume in thalamic nuclei are associated with positive schizotypy in healthy individuals. *Psychological medicine*, 1–9.

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual review of clinical psychology*, *14*, 91–118.

Fekete, J.-D. (2013). Visual analytics infrastructures: From data management to exploration. *Computer*, *46*(7), 22–29.

Franke, K., Ziegler, G., Klöppel, S., & Gaser, C. (2010). Estimating the age of healthy subjects from t1-weighted mri scans using kernel methods: exploring the influence of various parameters. *Neuroimage*, *50*(3), 883–892.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of experimental psychology: General*, *141*(1), 2.

Helldin, L., Kane, J. M., Karilampi, U., Norlander, T., & Archer, T. (2007). Remission in prognosis of functional outcome: a new dimension in the treatment of patients with psychotic disorders. *Schizophrenia Research*, *93*(1-3), 160–168.

Iwabuchi, S., Liddle, P. F., & Palaniyappan, L. (2013). Clinical utility of machine-learning approaches in schizophrenia: improving diagnostic confidence for translational neuroimaging. *Frontiers in psychiatry*, *4*, 95.

Jablensky, A. (2010). The diagnostic concept of schizophrenia: its history, evolution, and future prospects. *Dialogues in clinical neuroscience*, *12*(3), 271.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Janssen, R. J., Mourão-Miranda, J., & Schnack, H. G. (2018). Making individual prognoses in psychiatry using neuroimaging and machine learning. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(9), 798–808.

Kahn, R. S., van Rossum, I. W., Leucht, S., McGuire, P., Lewis, S. W., Leboyer, M., & Heres, S. (2018). Amisulpride and olanzapine followed by open-label treatment with clozapine in first-episode schizophrenia and schizophreniform disorder (optimise): a three-phase switching study. *The Lancet Psychiatry*, *5*(10), 797–807.

Kane, J. M., Leucht, S., Carpenter, D., & Docherty, J. P. (2003). The expert consensus guideline series. optimizing pharmacologic treatment of psychotic disorders. introduction: methods, commentary, and summary. *The Journal of clinical psychiatry*, *64*, 5–19.

Kapur, S., & van Os, J. (2009). Schizophrenia. *Lancet (London, England)*, *374*(9690), 635–645.

Kassraian-Fard, P., Matthis, C., Balsters, J. H., Maathuis, M. H., & Wenderoth, N. (2016). Promises, pitfalls, and basic guidelines for applying machine learning classifiers to psychiatric imaging data, with autism as an example. *Frontiers in psychiatry*, *7*, 177.

Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The positive and negative syndrome scale (panss) for schizophrenia. *Schizophrenia bulletin*, *13*(2), 261–276.

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, *23*(1), 89–109.

Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., Wobrock, T., ... Hasan, A. (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *The Lancet Psychiatry*, *3*(10), 935–946.

Koutsouleris, N., Riecher-Rössler, A., Meisenzahl, E. M., Smieskova, R., Studerus, E., Kambeitz-Ilankovic, L., & Falkai, P. (2015). Detecting the psychosis prodrome across high-risk populations using neuroanatomical biomarkers. *Schizophrenia bulletin*, *41*(2), 471–482.

Kraepelin, E. (1893). *Psychiatrie: ein kurzes lehrbuch für studirende und aerzte.* Abel.

Liu, S., Wang, X., Liu, M., & Zhu, J. (2017). Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*, *1*(1), 48–56.

Loebel, A. D., Lieberman, J. A., Alvir, J. M., Mayerhoff, D. I., Geisler, S. H., & Szymanski, S. R. (1992). Duration of psychosis and outcome in first-episode schizophrenia. *The American journal of psychiatry*.

Maity, N. G., & Das, S. (2017). Machine learning for improved diagnosis and prognosis in healthcare. In *2017 ieee aerospace conference* (pp. 1–9).

Maj, M., Pirozzi, R., Formicola, A. M., Bartoli, L., & Bucci, P. (2000). Reliability and validity of the dsm-iv diagnostic category of schizoaffective disorder: preliminary data. *Journal of affective disorders*, *57*(1-3), 95–98.

Merinder, L. B. (2000). Patient education in schizophrenia: a review. *Acta Psychiatrica Scandinavica*, *102*(2), 98–106.

Messinger, J. W. (2013). *Cognitive-affective processes in schizophrenia: The attentional-blink and olfactory hedonics* (Unpublished doctoral dissertation). Long Island University, The Brooklyn Center.

Mourao-Miranda, J., Reinders, A., Rocha-Rego, V., Lappin, J., Rondina, J., Morgan, C., & Doody, G. A. (2012). Individualized prediction of illness course at the first psychotic episode: a support vector machine mri study. *Psychological medicine*, *42*(5), 1037–1047.

Mühlbacher, T., Piringer, H., Gratzl, S., Sedlmair, M., & Streit, M. (2014). Opening the black box: Strategies for increased user involvement in existing algorithm implementations. *IEEE transactions on visualization and computer graphics*, *20*(12), 1643–1652.

Nieuwenhuis, M., van Haren, N. E., Pol, H. E. H., Cahn, W., Kahn, R. S., & Schnack, H. G. (2012). Classification of schizophrenia patients and healthy controls from structural mri scans in two large independent samples. *Neuroimage*, *61*(3), 606–612.

Orru, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., & Mechelli, A. (2012). Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neuroscience & Biobehavioral Reviews*, *36*(4), 1140–1152.

Poldrack, R. A. (2007). Region of interest analysis for fmri. *Social cognitive and affective neuroscience*, *2*(1), 67–70.

Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., & Kupfer, D. J. (2013).
American journal of psychiatry, *170*(1), 59–70.

Rittmannsberger, H. (2012). The diagnosis" schizophrenia": past, present and future. *Psychiatria Danubina*, *24*(4.), 408–0.

Sacchet, M. D., Prasad, G., Foland-Ross, L. C., Thompson, P. M., & Gotlib, I. H. (2015). Support vector machine classification of major depressive disorder using diffusion-weighted neuroimaging and graph theory. *Frontiers in psychiatry*, *6*, 21.

Schnack, H. G. (2020). Bias, noise, and interpretability in machine learning: from measurements to features. In *Machine learning* (pp. 307–328). Elsevier.

Schnack, H. G., & Kahn, R. S. (2016). Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters. *Frontiers in psychiatry*, *7*, 50.

Silver, H., & Shmoish, M. (2008). Analysis of cognitive performance in schizophrenia patients and healthy individuals with unsupervised clustering models. *Psychiatry research*, *159*(1-2), 167–179.

Sun, D., van Erp, T. G., Thompson, P. M., Bearden, C. E., Daley, M., Kushan, L., & Cannon, T. D. (2009). Elucidating a magnetic resonance imaging-based neuroanatomic biomarker for psychosis: classification analysis using probabilistic brain atlas and machine learning algorithms. *Biological psychiatry*, *66*(11), 1055–1060.

Tandon, N., & Tandon, R. (2018). *Will machine learning enable us to finally cut the gordian knot of schizophrenia.* Oxford University Press US.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., & Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, *17*(6), 520–525.

Van Os, J., Burns, T., Cavallaro, R., Leucht, S., Peuskens, J., Helldin, L., & Lachaux, B. (2006). Standardized remission criteria in schizophrenia. *Acta Psychiatrica Scandinavica*, *113*(2), 91–95.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, *10*(5), 988–999.

Ware, M., Frank, E., Holmes, G., Hall, M., & Witten, I. H. (2001). Interactive machine learning: letting users build classifiers. *International Journal of Human-Computer Studies*, *55*(3), 281–292.

Weiss, G. M. (2004). Mining with rarity: a unifying framework. *ACM Sigkdd Explorations Newsletter*, *6*(1), 7–19.

Yadav, K., Sarioglu, E., Choi, H.-A., Cartwright IV, W. B., Hinds, P. S., & Chamberlain, J. M. (2016). Automated outcome classification of computed tomography imaging reports for pediatric traumatic brain injury. *Academic emergency medicine*, *23*(2), 171–178.

Zhang, J. (2017). Multivariate analysis and machine learning in cerebral palsy research. *Frontiers in neurology*, *8*, 715.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301–320.

Appendix A

Additional Information

A.1 Hard Margin SVM

When diagnosing or prognosing schizophrenia with SVMs, the subjects are represented by a set of d features into a vector $\boldsymbol{x}_n$, where $\boldsymbol{x} \in \mathbb{R}^d$ of length d (Nieuwenhuis et al., 2012). Furthermore, each subject has a label y_n (e.g. patient with schizophrenia +1; healthy control -1). For each feature, an optimal weight has to be found for maximum separation of the classes. All optimal weight values are represented by the weight vector $\boldsymbol{w}$. In addition to the weight vector $\boldsymbol{w}$, an offset b has to be learned to maximise the margin between the hyperplane and each class (Sacchet, Prasad, Foland-Ross, Thompson, & Gotlib, 2015). The offset b, also known as the bias or threshold, is the distance to the origin of the hyperplane solution. Both $\boldsymbol{w}$ and b have to be learned based on the dataset containing different combinations of y and $\boldsymbol{x}$. Since the hard margin had to be linearly separable, the hyperplane, defined by the optimal b and $\boldsymbol{w}$, separates the data if and only if for $n = 1, ..., N$

$$y_n \dot{(} \boldsymbol{w}^T \boldsymbol{x}_n + b) > 0 \tag{A.1}$$

where $(\boldsymbol{w}^T \boldsymbol{x}_n + b)$ is the output of data point $\boldsymbol{x_n}$ of the SVM and y_n is the actual label of $\boldsymbol{x_n}$. This condition does not allow any misclassifications and assumes that the dataset is linearly separable with a hyperplane. Since it is always feasible for any separation hyperplane to select weights which could result in signals $y_n(w^T x_n + b)$ greater than or equal to 1 (Abu-Mostafa, Magdon-Ismail, & Lin, 2012), condition A.1 could also be defined as

$$y_n \dot{(} \boldsymbol{w}^T \boldsymbol{x}_n + b) = 1 \tag{A.2}$$

Beside separating the classes, the margin must also have a maximum margin. The size of the margin is defined by the formula:

$$\frac{2}{||\boldsymbol{w}||} \tag{A.3}$$

which is equal to

$$\frac{2}{\sqrt{\boldsymbol{w}^T \boldsymbol{w}}} \tag{A.4}$$

Thus, to find the optimal separation hyperplane, condition A.2, as well as the condition of maximum margin, have to be satisfied. Maximising the margin could be done by minimising $||\boldsymbol{w}||$, which is equivalent to minimising $\frac{1}{2}\boldsymbol{w}^T \boldsymbol{w}$. Hence, the optimisation problem for hard-margin SVMs that has to be solved in order to find the OSH could be defined as:

$$\text{minimise: } \frac{1}{2}\boldsymbol{w}^T \boldsymbol{w} \tag{A.5}$$

$$\text{subject to: } min_{n=1,...,N} = y_n(\boldsymbol{w}^T\boldsymbol{x}_n + b) = 1$$

This optimisation problem could be solved with algorithms such as gradient descent.

A.2 Soft Margin SVM

Nevertheless, in practice most datasets are not linearly separable. A soft margin allows a few misclassifications to a certain extent. This enables classification of non linearly separable datasets by adding an error measure ξ to condition A.2. The condition is now defined as:

$$y_n \cdot (\boldsymbol{w}^T\boldsymbol{x}_n + b) = 1 - \xi_n \tag{A.6}$$

If ξ_n is zero, the subject can be considered to be classified correctly. However, if $\xi_n > 0$, the subject is lying in an incorrect dimension leading to incorrect classification. Hence, the average error can be given as:

$$\frac{1}{N}\sum_{i=1}^{N}\xi_n \tag{A.7}$$

where N is the number of data points (e.g. subjects). The misclassification is controlled by the so-called cost parameter C. The parameter C determines to what extent we penalise an SVM for data points within the margin. Thus the optimisation problem for a soft-margin SVM is:

$$\text{minimise: } \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C \cdot \sum_{n=1}^{N}\xi_n \tag{A.8}$$

$$\text{subject to: } min_{n=1,...,N} = y_n(\boldsymbol{w}^T\boldsymbol{x}_n + b) = 1 - \xi_n$$

for all $n = 1, 2, ..., N$ where $\xi \geq 0$. Figure A.1 illustrates an optimal separation line with a small cost parameter value and an optimal separation line with a large cost parameter value.

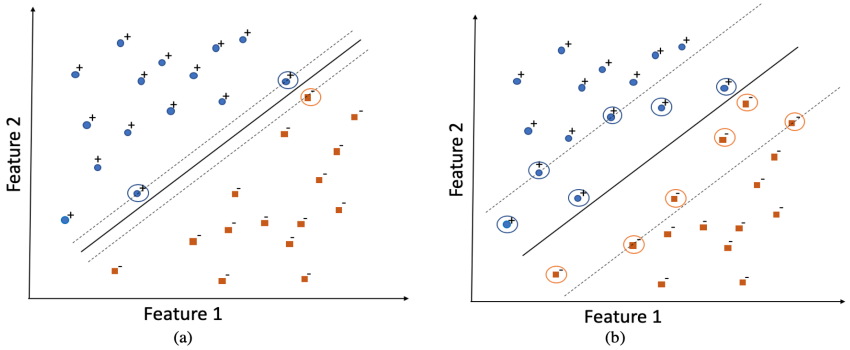


Figure (A.1) (a) Illustration of a high cost parameter value with a small margin. (b) illustrates a low cost parameter value with a large margin which allows some data points within the margin.

A.3 Cross-validation

Usually, when evaluating the performance of a machine learning model, the data is split into a training set and a test set. The training set is used to train the model, and the test set is used to evaluate the performance of the model.

However, if little data is available, this process can be challenging. Models trained on the small datasets tend to overfit, meaning that the model fits the training data too well since the

models are trained on only a few data points. As a result, the models do not generalise well on unseen data. Cross-Validation (CV) addresses this by repeatedly training and testing models on different parts of the data (Janssen et al., 2018).

A commonly used form of cross-validation is k-fold cross-validation that splits the data into k non-overlapping test sets of approximately equal sizes (James, Witten, Hastie, & Tibshirani, 2013). The test set k evaluates the model that is trained on the remaining $k-1$ subsets. The average of all the test folds could be used to evaluate the performance.

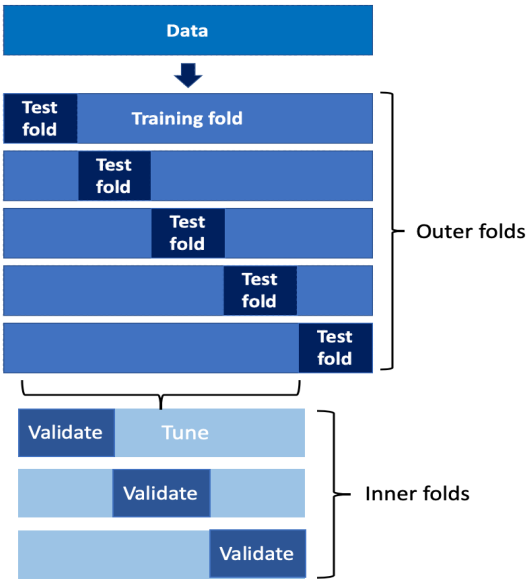


Figure (A.2) Illustration of nested cross-validation with five outer folds and three inner folds.

When training a model, the flexibility of the model should carefully be taken into account. If the model is too flexible, it will overfit. However, if the model is not flexible enough, it will underfit meaning that the model neither fits the training data, nor generalises well to unseen data. This flexibility is controlled by one or more parameters of the training algorithm. For SVM this parameter is the cost parameter C. This parameter can be optimised by evaluating different values and selecting the value that maximises the generalisability. Note that this optimisation should not be done on the test folds of the model since this would easily lead to overfitting.

As shown in figure A.2 the test folds of the outer folds can be used to evaluate the generalisability, and the training sets of the outer folds are used for parameter tuning. These outer folds are again divided in k' inner folds. The parameter values are trained on $k'-1$ folds and validated on the remaining validation set k'. Each parameter is analysed on all the inner folds of each outer fold. This type of cross-validation is named nested cross-validation.

Appendix B

Additional Tables and Figures

Tables

Table (B.1) Information of the OPTiMiSE dataset

Categories	Subcategories	Number of features	Further information
Diagnoses & Treatment	Mini-International Neuropsychiatric Interview (MINI)	58	-
	Current state	7	e.g. DSM-IV classification, duration
Demographics	Demographics of Patient	15	e.g. race, education, income
	Demographics of mother and father of the patients	10	e.g. Race, education, income
Physical examinations	-	4	e.g. Blood pressure systolic, Blood pressure diastolic
Questionnaires	Positive and Negative Symptom Scale (PANNS)	34	Consists of the positive subscale (7), negative subscale (8) and general subscale (15). The average of each subscale and the total PANNS score was also reported (4)
	Global Impression (CGI)	2	-
	Calgary Depression Scale (CDSS)	9	-
	Personal and Social Performance scale (PSP)	5	Questionnaire about Personal and social functioning
	Global Impression (CGI)	2	-
	Subjective Wellbeing under Neuroleptic (SWN)	20	Questionnaire about Subjective wellbeing
	Udvalg for Kliniske Undersogelser (UKU)	60	Questionnaire about Adverse effects
Recreational drug use, alcohol use, caffeine use, smoking	-	5	-

Table (B.2) Information of the GROUP dataset

Categories	Subcategories	Number of features	Further information
Sociodemographic	-	3	age, IQ, sex
Subcortical volumes	-	15	e.g. hippocampus, amygdala
Large volumes	-	11	e.g. volume of the total brain
Regions Of Interest	Cortical thickness	68	e.g. inferior partietal, fusiform, etc.
	Surface area	68	e.g. inferior partietal, fusiform, etc.
	Volume	68	e.g. inferior partietal, fusiform, etc.
Additional features	Frontal, partietal, occipital, temporal	12	e.g. white area, pial area, etc.

Table (B.3) Variables selected by the elastic net model

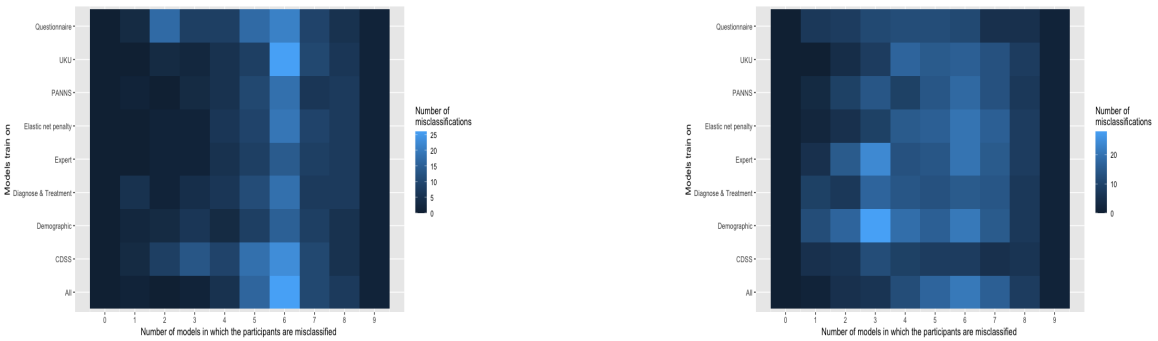
Features - prognostic model	Features - diagnostic model
Duration of psychosis	Left putamen cortex volume
Counry of birth	Left pallidum cortex volume
Main source of income	Left hippocampus cortex volume
Prognosis of psychiatrist	Right thalamus cortex volume
MINI - Substance induced mood disorder: Past	right putamen cortex volume
MINI - Major depressive episode with melancholic features: Current (2 weeks)	Right pallidum cortex volume
MINI - Manic episode: Current	Left bankssts cortex thickness
MINI - Manic episode: Past	Left lateral orbitofrontal cortex thickness
MINI - Social phobia (Social anxiety disorder): Current (past month)	Left lingual cortex thickness
MINI - Posttraumatic stress disorder: Current (past month)	Left parahippocampal cortex thickness
Schizophrenia: Current	Left parstriangularis cortex thickness
PANNS P2 - Conceptual Disorganization	Left posterior cingulate cortex thickness
PANNS N1 - Blunted affect	Left superior frontal cortex thickness
PANNS N2 - Emotional withdrawal	Right bankssts cortex thickness
PANNS N4 - Passive/ apathetic social withdrawal	Right entorhinal cortex thickness
PANNS N7 - Stereotyping thinking	Right isthmuscingulate cortex thickness
PANNS G5 - Mannerisms and posturing	Right lingual cortex thickness
PANNS G10 - Disorientation	Right pars opercularis cortex thickness
PSP - socially useful activities;including work and study	Right frontal pole cortex thickness
PSP - personal and social relationships	Left caudal middle frontal cortex volume
PSP score	Left pars orbitalis cortex volume
UKU - Autonomic Degree last 3 days	Left post central cortex volume
SWN 15 - Social distancing	Right caudal anterior cingulate cortex volume
Does patient smoke	Right cuneus cortex volume
PANNS Negative subscale	Right pars opercularis cortex volume
PANNS Positive subscale	Left caudal middle frontal cortex area
PANNS total score	Left middle temporal cortex area
Age (years)	Left pars opercularis cortex area
	Left pars orbitalis cortex area
	Left post central cortex area
	Left supramarginal cortex area
	Left insula cortex area
	Right bankssts cortex area
	Right inferior temporal cortex area
	Right lateral occipital cortex area
	Left lateral orbitofrontal cortex area
	Left parahippocampal cortex area
	Left frontal pole cortex area
	Left temporal pole cortex area

Figures

Heatmaps

In figure B.1 and B.2 heatmaps show the number of participants that are misclassified in k models where $0 \leq k \leq 9$. The heatmaps also specify in which models the participants were misclassified. The lighter the colour, the more participants were misclassified.

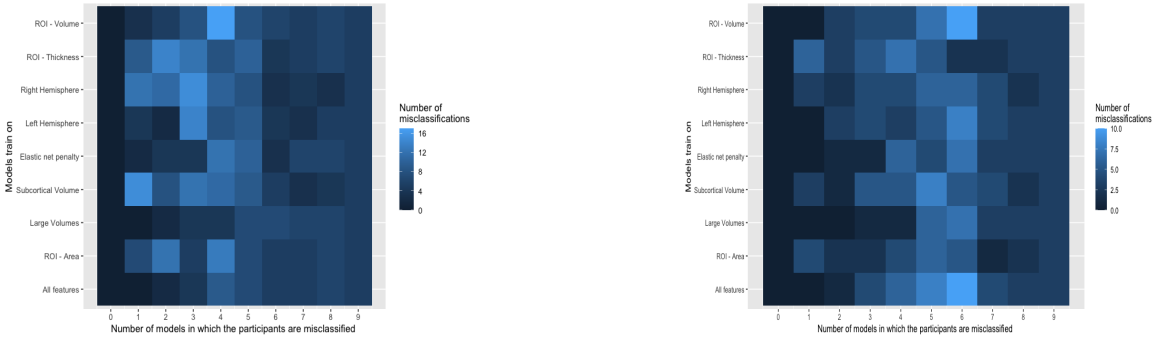
Prognostic models



(a) Negative class - no remission achieved **(b)** Positive class - remission achieved

Figure (B.1) Heatmap: number and names of prognostic models in which the participants were misclassified. Each row presents a different model and each column in how many models the participants were misclassified. (a) represents the 131 patients who did not meet the remission criteria. (b) represents the 255 patients who met the remission criteria.

Diagnostic models



(a) Negative class - healthy controls

(b) Positive class - patients with schizophrenia

Figure (B.2) Heatmap: number and names of diagnostic models in which the participants were misclassified. Each row presents a different model and each column in how many models the participants were misclassified. (a) represents the 204 healthy controls. (b) represents the 84 patients with schizophrenia.

Appendix C

Manual R scripts

This appendix contains instructions on the R scripts that were written for the purpose of this study. These R scripts could also be used for other machine learning studies since the functions are not dataset-specific. This study showed some approaches that could be used to get some insight into the misclassifications. In this appendix, we explain who these approaches can be used in R. We first show some basic instructions in R. This chapter will then go on to the instructions of the ML models and the approaches that were created for this study.

It is highly recommended to install RStudio. RStudio is a software package that makes it more convenient to use R as it highlights various components in your code (e.g., functions and specific statements such as if and else). Besides, RStudio has an interface that is easy to use.

C.1 Functions

In R many useful functions are already defined by scientists, including many statisticians. These functions are bundled in "packages" (a script or collection of scripts) that you can download and install. Available packages can be found on CRAN[1].
One of the packages that is used for this study is the package called caret.This package uses functions from the package e1071, also known as libSVM. These packages could be installed by the following commands:

```
> install.packages(''caret'')
> install.packages(''e1071'')
```

These commands have to be executed only once. After these commands, the packages are however not loaded yet. This could be done by the following commands:

```
> library(caret)
> library(e1071)
```

The functions that are specially made for this study can be found in the file named functions.R. The functions of this file can be used by loading the file from the computer. Before loading the functions from the file, the directory to this file had to be specified. This could be done by the following command:

```
> setwd("/path/to/my/directory")
```

For Windows, the command might look like:

```
> setwd("c:/Documents/my/working/directory")
```

Thereafter, the file can be downloaded by the command:

[1] https://cran.r- project.org/web/packages/

```
> source('functions.R')
```

C.2 Loading data

When the file that contains the data, on which the models should be trained, is saved in another folder of the computer than `functions.R`,the directory should be changed. This directory could be changed by `setwd` such as illustrated in the previous commands. The files of the dataset used for this study were saved as CSV (a comma-separated file). To load this file, the function `read_csv_files` can be used. The command for loading a CSV file is as follow:

```
> df <- read_csv_files("nameFile.csv")
```

The function `read_csv_files` reads the file as a `tibble`. Tibbles are data frames; however, they differ in some aspects from the default data frames in `R`. One of the advantages over the default data frame is when printing a data frame a `tibble` only shows the first ten rows and all the columns that fit on the screen. This is useful when one works with large datasets. When a model cannot be converted to a `tibble`, for instance, when a data frame has two identical columns, the dataset has to be modified first. In the case that a data frame has two identical columns, one of the columns in the dataset should be deleted first. If a file is not a CSV file, the built-in function `read.table` could be used to load a file.

C.3 Pre-processing

Before training a model on a dataset, some changes on the dataset often have to be made first. For instance, some variables might not be feasible for training a model. These variables should be removed from the dataset. Each variable is represented as a column in a dataset. A `variable` could be deleted from a dataset `df` by the following command:

```
> df$variable <- NULL
```

Another possibility is only to select some specific variables. For instance, when you only would like to use the first three columns, the sixth, fifteenth column from a data frame, the data frame `df` could be modified by:

```
> df <- df[c(1:3, 6, 15)]
```

Furthermore, in most cases, the participants had to meet some inclusion criteria. For example, when `inclusion criterium 1` had to be met, the participants who did not meet the criteria could be removed by:

```
> df <- df[df$inclusionCriterium1 == "correct",]
```

The datasets also had to be scaled. This had to be done before training. Scaling the dataset can be done by the function `scale`. When only specific variables have to be scaled, for instance, all variable except the first four variables, do:

```
> df[, -c(1:4)] <- scale(df[, -c(1:4)])]
```

In this study, K-nearest neighbouring is used to fill the missing values in the dataset. This is done with the function `preProcess` in which the `method` is set to `knnImpute:`

```
> knn = preProcess(df, method = "knnImpute")
> df_knn <- predict(knn, df)]
```

The function `predict` predicts the missing values based on other similar data points. The new data frame `df_knn` has no missing values.

C.4 Support Vector Machines

The function `svmNestedCross` is used to train a support vector machine with nested cross-validation. The function `svmNestedCross` has different variables (e.g. the number of inner and outer folds) that had to be specified. An example of such a linear support vector machine trained on all the variables of a dataset called `df` with the cost variables {0.125, 0.25, 0.5,2,4,8} used for tuning, in which the positive class appeared twice as often and where ten inner folds and twenty outer folds were used, is:

```
> svm <- svmNestedCross(data = df, target = Class ~.,
weight_value = (1/2), cost_values = c(2^(-3:3)),
nInnerFolds = 10, nOuterFolds = 20)
```

In this example, the different cost parameters were exponential. This way, a considerable range can be evaluated as a potential cost parameter. The value `weight_value` stand for the class weight. In this case, the class weight has the value `1/2`. This means that the positive class labels appeared twice as often than the negative class labels. This way the positive class has the class weight $\frac{1}{2}$ and the negative class a class weight of 1. If the negative class would appear twice as often in the dataset, the `weight_value` should be 2.

If the target variable of the dataset is not of the type `factor`, the function `svmNestedCross` will return an error. The target variable can be converted to a factor by:

```
> df$Class <- as.factor(df$Class)
```

The variables `nInnerFolds`, `nOuterFolds` and `cost_values` do not necessarily have to specified since `nInnerFolds` is set to five by default, `nOuterFolds` 10 by default and `cost_values` is set to `c(2^(-3:3))` by default.

The function `svmNestedCross` returns two values in which `svm[[1]]` represents the SVM model and `svm[[2]]]` the different cost parameters of the model. Additionally, the final cost variable used for training could be extracted by:

```
> svm[[1]]$bestTune
```

More details of the `svm` such as the class probability of the positive class label can be accessed with the following command:

```
> svm[[1]]$pred
```

C.5 Performances and Classifications

The performance of these models can be evaluated by the function `resultSVM`. This function prints different metrics such as the accuracy, specificity, sensitivity, and balanced accuracy. An example of an evaluation of an SVM could be:

```
> classifications <- resultSVM(svm = svm[[1]], data = df,
posClass_name = "SCZ", negClass_name = "HC")
```

In this case, the positive class represents the patients with schizophrenia (with label `SCZ`) and the negative class the healthy controls (with label `HC`). The variable `classifications` represents the true positives, true negatives, false positives and false negatives of the `svm`.
An ROC curve of the test folds of this example could be created by the following command:

```
> roc <- roc(predictor = svm [[1]]$pred$SCZ, response = svm [[1]]$pred$obs)
```

The Area Under the Curve could be measured by adding `$auc` after the ROC, for instance:

```
auc <- roc$auc
```

C.6 Analysing the Misclassifications

C.6.1 Individual level

When analysing a single model on one specific variable, the function `MisAnalyseA` could be used:

```
> MisAnalyseA(model = df$classifications, class = "+", variable= df$variable)
```

The function `MisAnalyseA` returns the mean and standard deviation of the `variable`. In this example, it returns the mean and standard deviation of the true positives and the false negatives since `class = "+"`. If `class` had the value `"-"`, it would have had returned the mean and standard deviation of the true negatives and the false positives. Besides, the effect size between the correct and incorrect classifications of the given class is returned. Additionally, the function returns a density plot.

The misclassifications of a model could also be analysed on all the remaining variables of the dataset by the function `MisAnalyseB`. This function computes the effect size of all the remaining variable in the dataset and returns the six features with the largest effect sizes. The function returns histograms when the features have discrete values and density plots when the features have continuous values. The misclassifications could also be evaluated on all the features of the dataset or only on the features that were used for training. As an illustration, an example of a model that is evaluated on the remaining variable is given:

```
> largest_effsize <- MisAnalyseB(df = df, df_train = df_M1,
name_class = "classifications_M1", features = "remaining")
```

In this case, `df_M1` represents the data frame with all the variables that were used for training. `df_M1` has to be a subset of `df`. The variables are evaluated on the remaining variables of `df`. `name_class` represents the name of the column that includes all the classifications (TPs, TNs, FPs, FNs) of the model that is evaluated. The variable `features` specifies whether the misclassifications are evaluated on the remaining features, on all the features of the dataset, or on the trained features. When `features = "all"` all the feature are evaluated. If `features = "training"`, only the features used for training are analysed. Use the subscript `[[i]]` to access the different density plots:

```
> largest_effsize[[1]]  # Feature with the largest effect size of the negative class
> largest_effsize[[2]] # Feature with the second largest effect size of negative class
```

`largest_effsize[[1]]`, `largest_effsize[[2]]` and `largest_effsize[[3]]`represent the three features with the largest effect size of the negative class and `largest_effsize[[4]]`, `largest_effsize[[5]]`, `largest_effsize[[6]]` represents the three features with the largest effect size of the positive class.

C.6.2 Comparison between models

Two models

To compare the misclassifications of two models, use the function `MisAnalyseCompare`. Here, an example of `MisAnalyseCompare`:

```
> MisAnalyseCompare(df = df,
name_class_M1 = "classifications_M1",
name_class_M2 = "classifications_M2",
pos_model_names = c("FN M1", "FN M2", "TP Both"),
neg_model_names = c("FP M1", "FP M2", "TN Both"),
name_file_pos = '/path/to/my/directory/figure1.png',
name_file_neg = '/path/to/my/directory/figure2.png',
distance = c(0.05,0.035, 0.10))
```

MisAnalyseCompare returns two confusion matrices in which the correct and incorrect classifications of the two models are compared. The first confusion matrix represents the participants from the positive class and the second confusion matrix represents the participants from the negative class. name_class_M1 and name_class_M2 represent the column names of the classifications of the two models. Besides, the function MisAnalyseCompare creates a Venn-diagram of the two models. The Venn-diagrams are directly saved as png on the computer. The names of the Venn-diagrams are specified by name_file_pos for the Venn-diagram of the positive class and by name_file_neg for the negative class. distance represents the distance between the label names, for instance, "FN M1", and the Venn-diagram itself. This distance could be modified depending on the sizes of the circles.

Multiple models

In case you want to compare more than two models on their misclassifications, other functions could be employed. For instance, the function MisAnalyse.All.frequency can be used to compute in how many models each participant is misclassified. Here, an example of MisAnalyse.All.frequency is given in which positive class had the value "1" and the negative class the value "-1":

```
> frequencyMisclassication <- MisAnalyse.All.number(df = df,
classNames = "classifications", targetName = "Class", targetValue_pos = 1,
targetValue_neg = -1)
```

These frequencies could also be further visualised by a histogram with the function MisAnalyse.All.Histogram. In MisAnalyse.All.frequency the false positives, as well as the false negatives, were returned. However, the function MisAnalyse.All.Histogram either returned the false positives or the false negatives. This has to be specified by name_Misclass. For instance, a histogram of the false positives of nine models could be created by the following lines:

```
> histogram_FN <-MisAnalyse.All.Histogram(df = df,
title = "Histogram False Positive",targetName = "Class",
targetValue = -1, name_Misclas = "FP", nModels = 9, labelnames = labelnames)
```

nModels represent the number of models, labelnames the names of the different models and title the title of the figure (by default the model has no title). The models in which the participants were misclassified could also be visualised by heatmaps with the function MisAnalyse.All.Heatmap:

```
> heatmap_FP <- MisAnalyse.All.Heatmap(df = df,
title = "Heatmap False Positive",targetName = "Class",
targetValue = -1, name_Misclas = "FP", nModels = 9, labelnames = labelnames)
```

A heatmap could also be used to visualise the degree of overlap of multiple models. An example of nine models that are evaluated on their overlap of the false positives is given below:

```
> df_overlap_FP <- MisAnalyse.All.overlap(df = df, targetValue = -1,
name_Misclas = "FP", nModels = 9, title = "Overlap False Positives ",
labelnames = labelnames)
```

C.6.3 Degree of misclassification

The absolute mean distance to the threshold of the true positives, true negatives, false positive and false negatives can be computed with the function MisAnalyse.distThreshold. An example is given below of a svm with patients with schizophrenia (label SCZ) as positive class and the healthy controls (label HC) as the negative class:

```
> distance_threshold <- MisAnalyse.distThreshold(svm = svm[[1]],
pos = "SCZ", neg = "HC"))
```

To compute the target score of the true positives, true negatives, false positives, and false positives in a dataset, the function `MisAnalyse.difThreshold` could be used. Here, an example of `MisAnalyse.difThreshold` in which the mean remission score of the true positives, true negatives, false positives, and false negatives is used to computed the target score.

```
> MisAnalyse.difThreshold(df = df,
name_class = "classifications_M1", targetScore = "meanRemissionScore"))
```

`name_class` represents the name of the column that includes all the classifications of the model that is evaluated. `targetScore` represent the name of the column that includes the target scores (e.g. mean remission score).