Utrecht University

Bachelor thesis Artificial Intelligence

Machine learning assisted identification of metabolic syndrome in patients with schizophrenia using structural brain measures

Author:
Jip van der Rest
5994799

Supervisor:
Dr. Hugo Schnack

15 ECTS

January 3, 2021

Abstract

Both metabolic syndrome and schizophrenia are associated with decreased volume of total brain and specific brain areas. Moreover, both diseases are found to often come together. The aim of this study is to gain insight in the association of schizophrenia and metabolic syndrome in the brain and to get a better understanding of the place machine learning has in the medical field.

Therefore, a machine learning approach is used to classify subjects with and without metabolic syndrome in a group of schizophrenic patients using brain volume, cortical thickness and area of different regions of interest as features. Two common challenges in the medical field (small sample size and class imbalance) are analysed and to overcome these challenges, different feature selections are made, both knowledge-based selections and selections based on machine learning. A soft-margin support vector machine is trained on a real-world dataset ($n = 73$) using these feature selections. The results showed that the feature selections made by machine learning algorithms yielded better performances than the knowledge-based feature selections. However, before the models could be used in a clinical setting, further research should be done to the association between different regions of interest selected and metabolic syndrome.

Contents

Chapter 1

Introduction

1.1 Metabolic syndrome, schizophrenia and the brain

Decreased brain volume is associated with various diseases and syndromes. Two of them are metabolic syndrome (Tiehuis et al., 2014) and schizophrenia (Haijma et al., 2013; Cahn et al., 2002): both subjects with metabolic syndrome and subjects with schizophrenia are found to have smaller total brain volume than healthy controls. Metabolic syndrome is a cluster of cardiovascular risk factors that can lead to several diseases (e.g. diabetes type 2, cardiovascular disease).[1] Both metabolic syndrome and schizophrenia are also associated with volume decreases of specific brain areas.

Moreover, it is found that schizophrenia and metabolic syndrome often come together. Patients with schizophrenia have approximately a twofold risk of developing metabolic syndrome in comparison with the general population (Papanastasiou, 2013). Furthermore, they have a threefold risk to die from cardiovascular disease (Ringen, Engh, Birkenaes, Dieset, & Andreassen, 2014) and a twofold risk to die from a heart attack (Galassi, Reynolds, & He, 2006), two of the common health complications that associate with metabolic syndrome. The frequently occurring unhealthy lifestyle of patients with schizophrenia seems to contribute to both increased prevalence of metabolic syndrome and risk of death (Heald et al., 2017). Besides, the use of anti-psychotic medication may play a role in the development of metabolic syndrome in patients with schizophrenia (Papanastasiou, 2013).

These brain abnormalities in both schizophrenia and metabolic syndrome and the fact that both diseases often occur together raise the question whether there is a possible connection between schizophrenia and metabolic syndrome in the brain.

A recent study compared the brain structures of patients with schizophrenia and metabolic syndrome to these of a control group consisting of patients with schizophrenia but without metabolic syndrome. The brain volume was measured from MRI data and was compared between the two groups us-

[1] More information about metabolic syndrome and schizophrenia can be found in *Chapter 2.1* and *2.2* respectively.

ing statistical analyses. This study concluded that, in schizophrenia, total brain (TB) volume and grey matter (GM) volume of patients with metabolic syndrome is decreased in comparison to that of patients without metabolic syndrome. Moreover, the study found that the summed volume of ten cortical and subcortical brain areas (regions of interest/ROIs) that are 'reward related' is also smaller in subjects with metabolic syndrome (de Nijs et al., 2018).

1.2 Current study

Although De Nijs' finding is valuable, it is not sufficient for clinical use, since it reports differences at group level and does not analyse the effect of the features on individual subjects. Compared to group difference, individual predictions are considered a much harder task (Arbabshirani, Plis, Sui, & Calhoun, 2017).

The study in this thesis will make an attempt to individual predict which of the patients with schizophrenia suffer from metabolic syndrome based on certain brain measures. In this thesis, a machine learning (ML) approach is used to make subject-oriented prediction of MetS possible. The dataset of De Nijs' study is used in the current study to develop and test the model.

Note the different purposes of machine learning classifications on one side and statistical analyses like De Nijs did in her research on the other side. The main purpose of machine learning diagnostic models is to get the best prediction per subject. In contrast, the purpose of a statistical analysis is to acquire knowledge about the relationship between variables in a group of subjects. A highly significant group difference does not always translate into a well classification result or the other way around: a feature that is used in a well performing classification model, does not mean that there is a significant difference at group level (Arbabshirani et al., 2017).

However, with this in mind, machine learning can also be used to gain insight into relationships between variables, which is done in this thesis. Particularly, one of the major advantages of ML is the ability to analyse various variables, or features, simultaneously and therefore discover underlying associations between several of them (Falahati, Westman, & Simmons, 2014). In this case, the associations between the volume, cortical thickness and area of different regions of interest and the presence of metabolic syndrome will be analysed. The purpose of the current study is to get a better understanding of these associations and the way machine learning can contribute to the medical field (especially in brain-image analyses).

In this study, a support vector machine model is built to classify subjects with and without metabolic syndrome in a group of schizophrenic patients. Both cortical and subcortical regions of interest are used as features.

1.3 Thesis structure

In the next two chapters, the theoretical background of this thesis is set out. Chapter 2 starts with a literature review on metabolic syndrome and schizophrenia. In chapter 3, the different machine learning approaches used in this thesis are described and analysed. The exact use of these machine learning approaches is described in chapter 4, along with a detailed description of the used sample. In chapter 5 of this thesis, the results of the trained models are reported. Afterwards, in chapter 6, the results are discussed and in chapter 7, a conclusion is drawn.

Chapter 2

Literature Review

In this chapter, literature about metabolic syndrome and schizophrenia is reviewed. The origin, current state of affairs with regard to diagnoses (with and without use of machine learning) and possible challenges of both diseases are set out.

2.1 Schizophrenia

The first mention of schizophrenia was more than a century ago (Kraepelin, 1893) and a lot of research has been done since. Schizophrenia is found to be a serious mental disorder characterized by its complexity. The lifetime prevalence of schizophrenia is around 0.4% (Bhugra, 2005).

Currently, diagnosis of schizophrenia is made from criteria of the DSM-V (Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition). However, until 2013, subject were diagnosed based on a previous version of the DSM. Because the dataset used in this study was collected before 2013, the subjects were diagnosed through this version (DSM-IV) which consist of the following criteria:

Two (or more) of the following, each present for a significant portion of time during a 1-month period (or less if successfully treated):

1. *delusions*
2. *hallucinations*
3. *disorganized speech (e.g., frequent derailment or incoherence)*
4. *grossly disorganized or catatonic behavior*
5. *negative symptoms, i.e. affective flattening, alogia (poverty of speech), or avolition (lack of motivation)*

These criteria are quite broad, which results in diverse clinical manifestations. However, schizophrenia is not only clinically heterogeneous, but also the etiology of schizophrenia is known to be heterogeneous. This is, multiple pathways can lead to schizophrenia (Schnack, 2019). One or more of those might be found in the volume of the brain. A major difficulty of heterogeneity is the fact that a straightforward linear separation between patients and healthy controls is impossible.

Diagnosing subjects based on brain measures instead of clinical criteria could help, because schizophrenia is known to be associated with volume reductions in specific brain areas. Such a association was found in white matter and grey matter, in which the last reduction was found to be considerably larger (Haijma et al., 2013). Moreover, grey matter volume was found to decrease during the first year of illness, while white matter volume did not (Cahn et al., 2002). However, another study found that specific areas of white matter (i.e. frontal, parietal and temporal lobe) were associated with a volume reduction over time (Olabi et al., 2011). In addition to white and grey matter, also third and lateral ventricle volume (both larger in patients) (Cahn et al., 2002), nucleus accumbens (smaller) (van Erp et al., 2016) and hippocampal volume (smaller) (Koolschijn et al., 2010) were found to relate with brain abnormalities.

2.2 Metabolic syndrome

Metabolic syndrome (MetS) is a cluster of five health problems (increased blood pressure, central obesity, hypertriglyceridemia, low HDL-cholesterol and hyperglycemia) that often occur together and increase the risk of several health complication such as cardiovasculair disease (CVD) and diabetes type 2 (Stern, Williams, González-Villalpando, Hunt, & Haffner, 2004). Furthermore, individuals with the syndrome have a three times higher risk to have a heart attack or stroke (Alberti, Zimmet, & Shaw, 2006).

The concept of a cluster of risk factors that often occur together and is associated with an increased risk with regards to diseases mentioned before is first described by Raeven in 1988. By that time, this cluster of risk factors was called *syndrome X*, which is currently known as metabolic syndrome. Remarkably, Raeven did not include central obesity as a risk factor, whereas nowadays that is regarded as one of the most important risk factor of MetS (Ritchie & Connell, 2007). In fact, by definition of the International Diabetes Federation, a person is diagnosed with MetS whenever central obesity is determined, in addition to any of two additional risk factors (Alberti et al., 2006). The exact criteria can be found in Tabel 2.1.

Central obesity	Waist circumference (ethnicity specific)	
Any two of the following	**Raised triglycerides**	$\geq$ 1.7 mmol/l (150 mg/dl) *or specific treatment for this lipid abnormality*
	Reduced HDL cholesterol	$<$ 1.03mmol/l (40 mg/dl) in males $<$ 1.29 mmol/l (50 mg/dl) in females *or specific treatment for this lipid abnormality*
	Raised blood pressure	Systolic: $\geq$ 130 mmHg Diastolic: $\geq$ 85 mmHg *or treatment of previously diagnosed hypertension*
	Raised fasting plasma glucose (FPG)	FPG $\geq$ 100 mg/dL (5.6 mmol/L) *or previously diagnosed type 2 diabetes*

Table 2.1: International Diabetes Federation metabolic syndrome criteria

To complicate things, the International Diabeters Federation definition is not the only definition that is commonly used. For example, the World Health Organization (WHO) and European Group for the Study of Insulin Resistance (EGIR) use slightly different criteria which are focussed on Insuline Resistance instead of central obesity (Kassi, Pervanidou, Kaltsas, & Chrousos, 2011).

Since the first mention of MetS by Raeven, a great amount of studies that describe the risk factors and prevalence of MetS has been published. However, the prevalence of MetS vary considerably among studies (Cameron, Shaw, & Zimmet, 2004; Ford, Li, & Zhao, 2010). Of course, one of the reasons for this is the use of various diagnosis criteria. Yet, although the exact prevalence in the worldwide population is unknown, all studies describe a vast proportion of the population suffering from MetS (Kassi et al., 2011).

The last couple years, machine learning has become an innovative and promising method to diagnose MetS. Because it is unclear which of the risk factors exactly contribute to MetS and to what extent, machine learning could be helpful. In a recent literature review, forty studies are found that use several machine learning techniques to identify MetS. In this review, artificial neural networks and decision tree are indicated as the techniques with the highest predictive performance, followed by Support Vector Machine (SVM) (Kakudi, Loo, & Moy, 2020). In a study comparing diagnosis of MetS using SVM and decision tree, the first is found to have higher sensitivity, specificity and accuracy (Karimi-Alavijeh, Jalili, & Sadeghi, 2016). In studies using SVM usually only features with a clinical origin (e.g. body mass index, blood pressure and age) are used (Karimi-Alavijeh et al., 2016; van Schependom et al., 2015; Gutiérrez-Esparza, Infante Vázquez, Vallejo, & Hernández-Torruco, 2020). However, one study is found that includes genetic information as features in addition to clinical features (Choe et al., 2018). No machine learning study based on MRI scans is found.

However, using brain data could be very interesting, because subjects with metabolic syndrome are known to have abnormalities in specific brain areas. For example, metabolic syndrome is known to be associated with a smaller hippocampal volume, in both diabetic and non-diabetic subjects (Yau, Castro, Tagani, Tsui, & Convit, 2012). Also, the volume of grey and white matter is found to be decreased in subjects with metabolic syndrome (Sala et al., 2014). Furthermore, decreased volume of the right nucleus accumbens was found in a group with metabolic syndrome in comparison with a healthy control group (Song et al., 2015). Different cortical regions of the brain of subjects with metabolic syndrome are found to have a significantly decreased cortical thickness (Song et al., 2015).

2.3 Brain connection between schizophrenia and MetS

Research has conclude that metabolic syndrome and schizophrenia often occur together (see chapter 1.1). In addition, as described in the previous sections, brain abnormalities exist in both schizophrenia and metabolic syndrome. It seems like there is a certain similarity in these abnormalities, especially in some areas of the subcortical brain (hippocampus, nucleus accumbens), the grey and white matter. The question whether possible brain abnormalities in schizophrenia cause metabolic syndrome is not that easily answered, because also external factor may play a role.

For example, the use of anti-psychotic medication may have an influence on the occurrence of metabolic syndrome. Studies have shown that especially schizophrenia patients that use lifelong medication against schizophrenia are likely to develop metabolic syndrome (Papanastasiou, 2013). Antipsychotic medication has a influence on appetite control, body composition and metabolic regulation, which all contribute to the development of metabolic syndrome (Ringen et al., 2014). Furthermore, in patients with schizophrenia, the use of antipsychotic medication is found to be associated with a progressive reduction of the grey matter volume (Haijma et al., 2013).

Chapter 3

Machine Learning Approaches

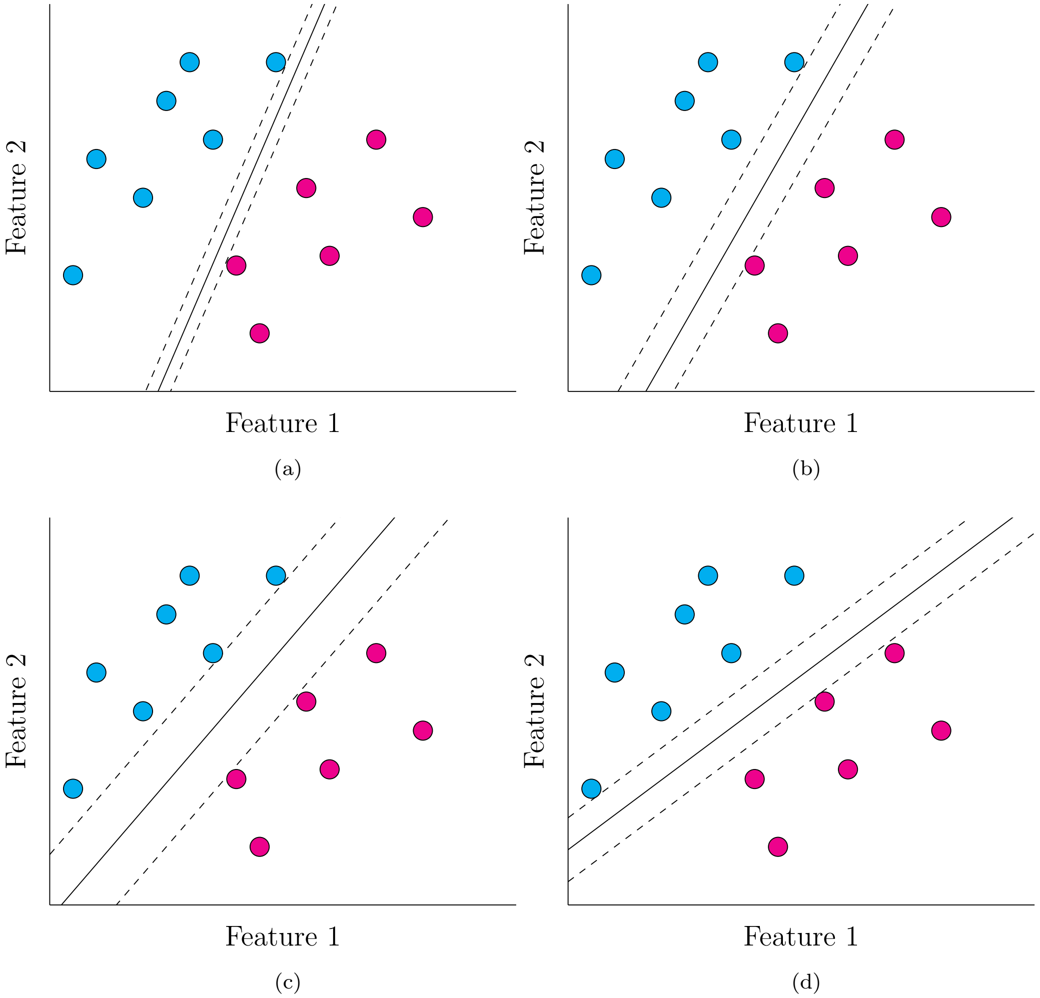
(Supervised) machine learning model development usually goes through two phases: the training phase, in which a model is trained using a machine learning algorithm and the testing phase, in which the trained model is validated. In this chapter, the approach used for training (support vector machine) and testing (cross-validation) are set out. After that, some issues that occur frequently when machine learning is used in the medical field are described, as well as possible solutions for these issues.

3.1 Support vector machine

Support vector machine (SVM) is a machine learning technique that is first mentioned in 1995 (Cortes & Vapnik, 1995) and is the most popular technique in neuro imaging (Orru, Pettersson-Yeo, Marquand, Sartori, & Mechelli, 2012). The goal of a SVM is to classify subjects into two or more different classes while using a hyperplane as a decision boundary (Luts et al., 2010). In this study, we focus on support vector machine models that classify subjects into two classes.

In a two-dimensional feature space a hyperplane is a line separating the classes, while in higher dimensions a hyperplane is a higher dimension generalization of a line. There can be multiple lines that exactly separate two classes, but the SVM selects the optimal separating hyperplane (OSH), that is the hyperplane with the largest margin (i.e distance between hyperplane and nearest datapoint) (Orru et al., 2012). In Figure 3.1 several lines separating a two-dimensional dataset are shown. The line in Figure 3.1 (c) has the largest margin and is therefore considered optimal. The data points lying on the margin are called support vectors. The support vectors determine the orientation of the OSH and influence the width of the margin.

The decision boundary of a SVM consist of a set of weights, one attached to each feature. The optimal set of weights for a certain dataset results in the largest margin and is the decision boundary of the OSH. Let $\mathbf{w}$ be a vector including the weights of each feature and $\mathbf{x}$ be a vector describing a data point. Both are of length n, the amount of features in the dataset. In addition, let $y \in \{-1, 1\}$ be the actual class label of data point $\mathbf{x}$. The



(a) (b)

(c) (d)

Figure 3.1: A two-dimensional dataset is linearly separated by four lines. All lines classify the data perfectly in the right class. The line in (c) has the largest margin and so is the OSH.

output of the SVM is given by:

$$sign(\mathbf{w}^T\mathbf{x} + b)$$

Intercept b is added and is, as well as weight vector **w,** optimized by the support vector machine. Note that for a perfect separation every datapoint x_n should be classified correctly and so:

$$\forall x_n \ (y_n \dot{(}\mathbf{w}^T\mathbf{x}_n + b) \geq 1)$$

In order to get the optimal separating hyperplane, the margin of the decision boundary must be maximized. The margin is given by:

$$\frac{2}{||\mathbf{w}||} = \frac{2}{\sqrt{\mathbf{w}^T\mathbf{w}}}$$

Maximizing the margin could be done by minimizing the inverse:

$$\text{minimise}\ \frac{\mathbf{w}^T\ \mathbf{w}}{2} = \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

This optimization problem could be easily solved using for example gradient descent.

Soft-margins

The described procedure works with precisely separable datasets, which in reality are not that common. Instead, most datasets have overlapping classes. However, in this case a slightly adjusted version of SVM could be used. In soft-margin SVM, a certain penalty could be added to allow for some misclassifications. In doing so, big margin could still be used, but data points lying inside the margin or on the wrong side are allowed. A correctly classified data point does not contribute to a penalty and the further a data point lies from its own margin, the bigger the penalty becomes.

A cost parameter C determines the amount of error given by a data point in the margin. A high C results in a smaller margin. Let ξ be the error and C the cost parameter. Our optimization problem becomes:

$$\text{minimise}\ \frac{1}{2}\mathbf{w}^T\ \mathbf{w} + C.\sum_{n=1}^{N}\xi_n$$

3.2 Cross-validation

After the training phase, model evaluation is done during the test phase. It is important that evaluation is done with unseen data, since the use of new data gives an unbiased estimate of the capacity of the model when used in real-world situations (Vabalas, Gowen, Poliakoff, & Casson, 2019). One method to make sure that unseen data is used, is to simply split the dataset into a train set and a test set. The test set mimic the real-world samples for which class labels have to be predicted (R. Simon, 2003).

However, when dealing with small datasets, this so called *split-sample method* could cause overfitting, because the train test is too small. A model that is perfectly trained on only a few data points, is not able to generalise well on new data. On the other side, a larger train set, results in a small test set, which is not able to estimate the performance of the model reliable.

K-fold cross-validation (CV) could be used as an alternative method to split test and train set. In k-fold cross-validation all data can be used for training and be reused for testing, which ensures that less data is needed (Vabalas et al., 2019). This can be achieved by dividing the data into k non-overlapping folds and using every created fold as a test fold once. For each fold, a model is trained using all $k - 1$ folds (except the test fold) and validated using the test fold. The performance of the model is estimated with the mean of the performance of the test folds.

It is important that stratification is used when the data is divided into folds. Stratification ensures that in all folds the class proportions reflect that of the entire sample (Berrar, 2019). For example, if a sample consists of ten subjects in the positive class and thirty subject in the negative class, all folds should contain around three times as many samples in the negative class as in the positive class. Hence, all folds are an as good estimate of the real-world as the entire sample.

In a soft-margin support vector machine, the cost parameter C has to be tuned. A lack of parameter tuning could decrease the model performance significantly (Arbabshirani et al., 2017). However, Varma and Simon (2006) showed that tuning parameters using the train data could cause a biased estimate of the model performance . Therefore, the optimal value of C should be chosen in another cross-validation loop. The method in which a double CV-loop is used, is called *nested cross-validation* and reduces the bias of the model properly (Varma & Simon, 2006).

In nested cross-validation, the training fold is again split up into l (*inner) folds* in which the cost parameter C is optimized. Just like the outer folds, one of the folds is used as test fold, while the other $l - 1$ folds are used as training fold. The value of C with the best performance in the inner fold is used to train a model based on the corresponding outer fold and validated on the outer fold test set. The nested cross-validation procedure is schematically shown in Figure 3.2.

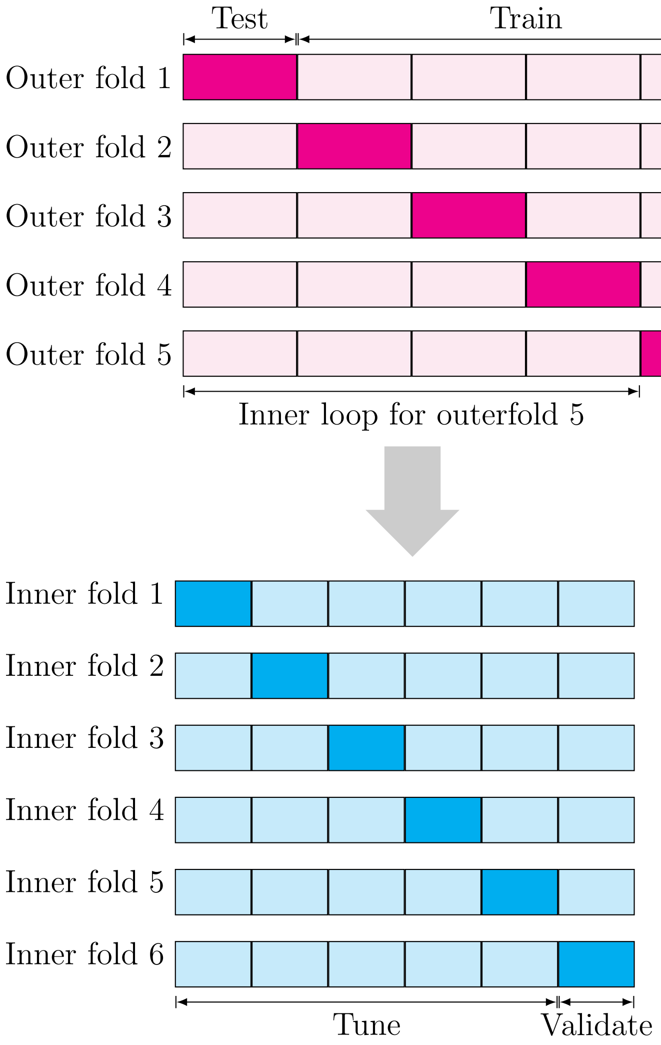


Figure 3.2: Illustration of nested cross-validation with k = 5 and l = 6

3.3 Challenges in the medical field

Despite a great amount of studies about machine learning in the medical field and the potentials of diagnostic classifiers reported by that studies, (Arbabshirani et al., 2017; Gunčar et al., 2018; Kakudi et al., 2020) ML models are not often deployed into clinical practice. The purpose of this paragraph is to get a better understanding of the way machine learning could contribute to the medical field (especially in brain-image analyses) and to gain insight into two main challenges in this field: small sample size and class imbalance.

3.3.1 Small sample size

One of the major challenges in the medical field and especially in neuroimaging studies is a relatively small sample size. Small sample size is a commonly seen problem in neuroimaging studies, because MRI scans and other types of data collection using human subjects are quite expensive and time consuming (Kononenko, 2001). Small sample sizes pose multiple problems, for example it could cause an unstable model, that is varying performance measures in different runs. In a study that used a SVM to classify schizophrenia patients and healthy controls, it was found that a training set of at least 130 subject was required for a stable model (Nieuwenhuis et al., 2012). Besides that, a model is less likely to generalise well on unseen data, whenever the ratio between amount of features and sample size becomes higher (Vabalas et al., 2019). It appears that the reason for this is that small sample sizes do not represent the entire patient group (Arbabshirani et al., 2017).

Possible solution: feature selection with machine learning

Because sample size and amount of features are associated numbers, a possible solution of a small sample size could be to also reduce the amount of features. If less features are used to train a linear support vector machine model, the ratio between the amount of features and the sample size will be lower. Therefore, the variance of the model will be lower, which ensures a better performance. Especially when multiple features contain overlapping information, removing features could improve the performance. The question is, however, which features contain noise or overlap with other features? A selection could be knowledge-based, but another option is to use machine learning to select the best possible subsets of features. In the next paragraphs, some of the commonly used feature selection methods are described.

There are many different methods for feature selection. These methods roughly can be classified into three different groups: filter methods, wrapper methods and embedded methods. Filter methods select a feature subset regardless of the machine learning algorithm used. Features are for example ranked or a chosen subset is evaluated. Only after selecting the optimal features, a machine learning model is trained and evaluated (Jović, Brkić, & Bogunović, 2015). Because filter methods select features independent from a specific machine learning method, they avoid overfitting, which is, of course, a great advantage of this kind of feature selection (Chandrashekar & Sahin, 2014). However, the fact that feature selection is done before model training has its disadvantages too. Because the estimated accuracy of a machine learning algorithm is the best measure to evaluate the values of a certain feature, the use of this algorithm usually leads to a more optimal feature selection (Das, 2001).

Wrapper method: step forward algorithm

Methods that use a machine learning algorithm to select features are called wrapper methods. In such methods a machine learning model is used as a black box and its performance as a function to evaluate feature subsets. The main drawbacks of wrapper methods are that they are more likely to overfit and are much slower compared to filter methods (Chandrashekar & Sahin, 2014).

An example of a wrapper method is called *step forward algorithm*, which is a iterative method that starts with an empty subset and add features one at a time. The *step forward algorithm* works as follows: In the first step, all features are evaluated individually using a certain performance measure and the feature which results in the best performance is selected. Afterwards, all possible combinations of the selected feature with a another feature are tested and the pair that produces the best performance is chosen. In this way, features are selected until a certain stopping criteria is met. A possible stopping criteria could be that none of the features does increase the performance of the model when added to the feature set (Aha & Bankert, 1996). All possible machine learning algorithms could be chosen in the step forward algorithm, but because of the slow speed of wrapper algorithms, fast modelling algorithms such as SVM work out the best (Jović et al., 2015). The *step forward algorithm* is visually shown in Figure 3.3.

A variation of the *step forward algorithm* allows for deletion of added features, which could increase the chance of selecting the optimal subset (Pudil, Novovičová, & Kittler, 1994). The opposite of the algorithm also exists. This, so-called *step backward algorithm* is similar to the *step forward algorithm* but starts from the set of all features and removes one feature at a time (Chandrashekar & Sahin, 2014).

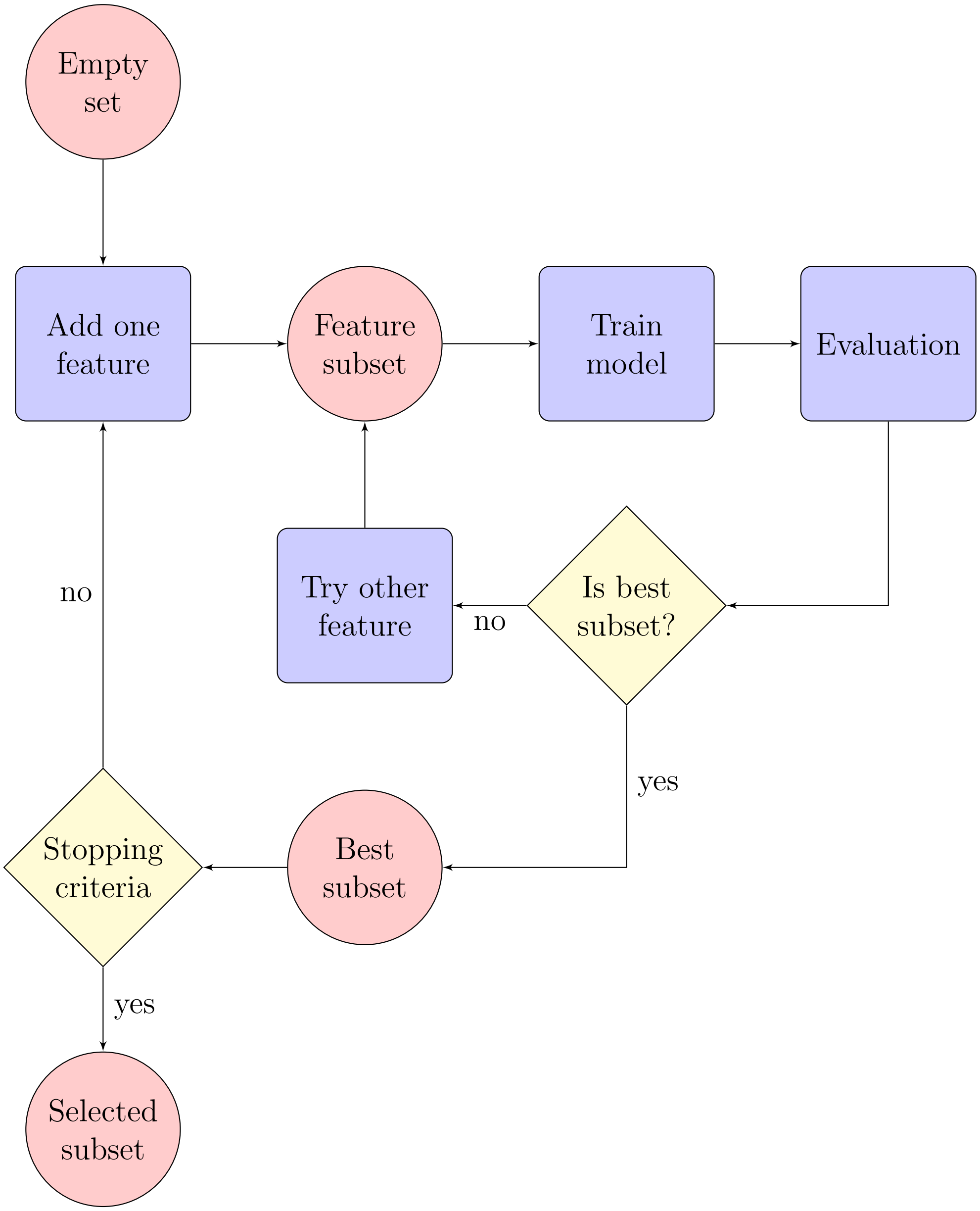


Figure 3.3: Step forward algorithm

Embedded method: LASSO

Both filter and wrapper methods have advantages and disadvantages and work better in particular situations. A third group of feature selection methods named embedded methods and are in between filter and wrapper methods. They use a machine learning algorithms to select feature (an advantage of wrapper methods), but are also relatively fast due to the embedded nature of these methods (an advantage of filter methods) (Tang, Alelyani, & Liu, 2014). Unlike other methods embedded methods do not split the feature selection and training steps, but the feature selection is embedded in the machine learning algorithm (Lal, Chapelle, Weston, & Elisseeff, 2006).

Many embedded models for feature selection are regularization models that add a penalty to the loss function of the model and eventually eliminate certain features which coefficients have shrunken to zero. The features with non-zero coefficents are selected to be part of the model (Fonti & Belitser, 2017).

An example of an regularised feature selection method is LASSO (Least Absolute Shrinkage and Selection Operator.) The loss function without penalty of LASSO is the same as the ordinary least squares (OLS) regression (linear regression), that is the sum of the squared residuals (distance between predicted point and real point). So the loss function LASSO can be defined by (Tang et al., 2014; van der Kooij & Meulman, 2008) :

$$Loss(\mathbf{w}) = ||\mathbf{y} - \sum_{i=1}^{m} \mathbf{w}_i \mathbf{x}||^2 + \lambda \, penalty(\mathbf{w})$$

In which $\mathbf{w}$ is a vector including the weights of each feature and λ is the regularization or tuning parameter that controls the strength of the penalty. The penalty used in LASSO regression is defined by:

$$penalty(\mathbf{w}) = \sum_{i=1}^{m} |\mathbf{w}_i|$$

So the penalty is ≥ 0 and the higher the regularization parameter, the higher the penalty, which results in more features become equal to zero. Actually, if λ is high enough, all features will be zero.

3.3.2 Class imbalance

Class imbalance means that one class is over- or under-represented compared to other classes. Most machine learning classifiers are biased towards the biggest class, which leads to poor results in the smallest class or even to a model that classifies all subjects into the biggest class (Longadge & Dongre, 2013).

Due to the class imbalance, even a model that classifies all subjects in one class can reach a pretty high accuracy. Accuracy is the ratio of the number of correct classifications and the total number of subjects and is commonly used in evaluation of classification problems. The sample used in the current study includes 54 subject in the negative class and 19 in the positive class. The accuracy of the model that classifies all subjects as negative will be pretty high, which could give a disorted view of the model performance.

$$\text{Accuracy} = \frac{0+54}{0+54+0+19} \approx 0.7397$$

Not only in final evaluation class imbalance causes problems. If cross-validation is used to tune parameters, accuracy is often used to tune parameters, which easily leads to a model that is biased toward the biggest class.

Most real-world datasets are somehow imbalanced and often the most interesting class (for example the class of subjects with a certain illness) is the smaller one (Liu, Yu, Huang, & An, 2011). In the medical field, the cost of misclassifying a subject in the smaller class is higher than that of misclassifying a subject in the bigger class, that is misclassifying an ill subject as healthy is way worse than misclassifying a healthy control as ill. Therefore, class imbalance in the medical field is a serious issue and there is need for good sampling techniques in the medical field (Rahman & Davis, 2013).

Possible solution: different performance metrics

One way to deal with class imbalance is by calculating not only accuracy scores of the model, but also use other performance metric scores. As mentioned before, the accuracy score could give an overly optimistic view of the model performance. Measures like balanced sensitivity and specificity are more desirable than accuracy for evaluation of classification problems with imbalanced classes (Arbabshirani et al., 2017).

Performance metrics often are calculated based on a confusion matrix which labels a subject in a binary model in one of four groups based on its predicted and actual value. The Confusion Matrix can be seen in Table 3.1.

		Actual value	
		Postive	Negative
Pred. value	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Table 3.1: Confusion Matrix

The accuracy of a model is the proportion of correct predicted subjects and could be calculated by:

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

In an imbalanced dataset, the balanced accuracy is a way more accurate metric, since it ensures that both positive and negative classification classes contribute equally to the final score. Balanced accuracy is the mean of two other performance metrics, sensitivity and specificity.

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}$$

Sensitivity, also known as recall, tells what proportion of the subjects in the positive class is classified as positive by the algorithm and is commonly used to evaluate medical algorithms, since mislabelling someone ill as healthy is worse than vice versa.

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

The counterpart of sensitivity is specificity, that tells what proportion of healthy people are classified as so. The specificity can be calculated by:

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

Another way to evaluate machine learning algorithms is the ROC (Receiver Operating Characteristics) curve, which could compare two classifiers across the entire range of class distribution (Ling, Huang, Zhang, et al., 2003). It expresses the trade off between sensitivity and the false positive ratio. The false positive ratio is calculated by:

$$\text{False positive ratio} = \frac{\#FP}{\#FP + \#TN} = 1 - \text{Specificity}$$

The ROC curve gives a lot of information about a classifier, but comparing two models could be difficult. Fortunately, the area under the ROC curve (AUC score) gives a simple score of the performance of the model. The higher the score, the higher the performance of the model.

An ideal model has completely separable classes and consequently a sensitivity and specificity score of 1. As a result, Area Under the Curve (AUC) will be 1. In Figure 3.4, ROC curves of 4 different models are shown. A model with an AUC score of around 0.5 performs as good as a random classifier and is therefore not so valuable.

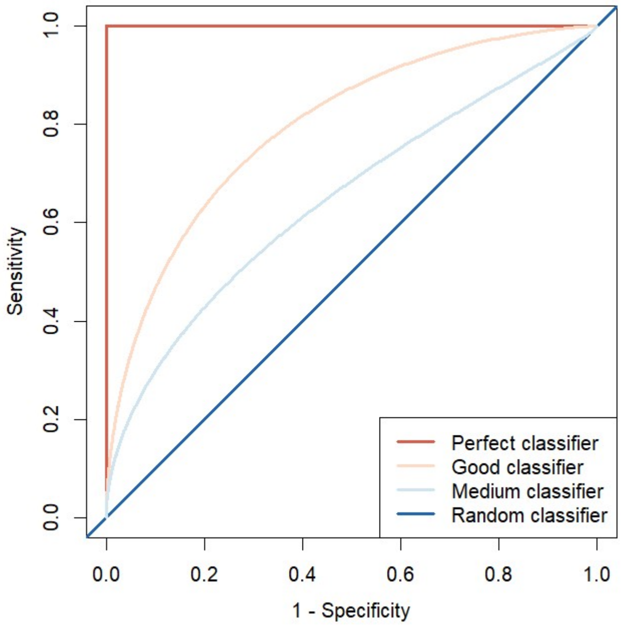


Figure 3.4: An ROC curve of 4 different models with AUC ranging from 0.5 (random classifier) to 1 (perfect classifier)

Another metric that could be used to evaluate a machine learning model is Odds Ratio (OR). In the medical field, OR is the ratio of the likelihood a subject will be ill given a positive prediction relative to the likelihood a subject will be ill given a negative prediction (Szumilas, 2010).

The odds ratio is calculated by:

$$\text{Odds ratio} = \frac{\#TP/\#FP}{\#FN/\#TN} = \frac{\#TP \times \#FN}{\#FP \times \#TN}$$

If the odds ratio for an event become substantially higher than 1, the odds ratio for the non-occurrence of the event will become substantially lower than 1 (S. D. Simon, 2001).

Possible solution: class weights

The use of various performance metrics gives a more complete picture of the capacity of models than accuracy does. However, it does not completely solve the problem of class imbalance. Therefore, class weights could be added, which penalise misclassifications of the positive (smaller) class heavier then misclassifications of the negative (larger) class.

Chapter 4

Methods

4.1 Sample

The sample is part of a longitudinal study named GROUP (Genetic Risks and Outcome of Psychosis) project, a Dutch research that conducts research on psychotic disorders (Korver et al., 2012).

The sample contains information of 84 subjects with schizophrenia or other schizophrenia-like diagnoses. All subjects are fluent Dutch speakers and have given written permission to participate in the study. Furthermore, of all the subjects a diagnosis of a non-affective psychotic disorder according to the criteria of DSM-IV (see chapter 2.2) is present. The sample includes both clinical information (i.e. sex, age, IQ) and MRI scans of the subjects. This study only uses MRI data (Korver et al., 2012).

Information about a metabolic syndrome diagnosis is available from 73 subjects of the same sample. The patients diagnosed with metabolic syndrome together are identified as one group (MetS+). The other group consists of patients without metabolic syndrome (MetS-). Patients are categorized into the MetS+ group using the diagnosis criteria of the International Diabetes Federation (see chapter 2.1). The MetS+ group consists of 19 subject, that is 26% of all 73 subjects. All were in the age range between 16 and 43 years at time of the study (mean ± SD = 27.14 ± 5.54) and a large majority of the subjects is male (89%).

4.2 Neuro-imaging

Structural 3-dimensional T1-weighted scans with $1 \times 1 \times 1.2$ mm^3 voxels were acquired on a 1.5 tesla MRI Philips scanner and were preprocessed at the Department of Psychiatry at the UMC in Utrecht. Freesurfer software was used to automatically subdivide both left hemisphere (LH) and right hemisphere (RH) into 7 subcortical and 34 cortical regions of interest (ROIs) (Desikan et al., 2006). All ROIs used in this study can be found in Appendix A.

Cortical thickness and surface area of the cortical ROIs were calculated using Freesurfer Software. Brain volumes of cortical ROIs were computed using the thickness by area of each vertex. Brain volumes of subcortical ROIs were automatic calculated using image intensity, probabilistic atlas location and spatial relationships between subcortical structures. In addition to information about ROIs, also information about larger brain areas (e.g. total brain volume, grey matter volume) was automatic calculated and manually corrected, if necessary after inspection (Kubota et al., 2015).

A quality check was done with regard to the MRI scans. Around a quarter of the subjects was excluded, because the quality of the MRI scans was too low. This quality check was done at the start of the study, therefore all 84 subjects mentioned before did have a well enough quality. More information about the MRI scans and preprocessing can be found in Kubota et al (2015).

4.3 Models to classify metabolic syndrome

The brain measures from MRI data were used to train a linear soft-margin support vector machine algorithm using the 73 subject with a schizophrenia-like diagnosis and available MetS-data. The goal was to separate the MetS+ group and the MetS- group as accurate as possible.

After training, the performance of all models was estimated using nested cross-validation with 5 inner folds and 5 outer folds. To find the optimal cost value C, a range of possible values of C was reviewed: $C \in \{0.01, 0.025, 0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 1, 2.5, 5\}$. The value of C with the best balanced accuracy in the inner folds was used to train a model based on the corresponding outerfold.

The nested cross-validation as described above was repeated 100 times. This is, the data set was randomly divided into 5 outer folds 100 times. The outer test fold was used to evaluate the models using several performance metrics, which are described in section 4.5.

4.3.1 Feature selection

As mentioned before, one of the major difficulties of this study is the relatively small sample size alongside a large quantity of features. As described in chapter 3, one way to deal with this challenge is by reducing the amount of features. This method is used in the present study.

First, all clinical features such as IQ and sex were removed, since only brain measures are used in this study. Afterwards, a model was trained with

	Feature selection	**No. features**
All features	Without merging LH and RH	224
	With merging LH and RH	114
Knowledge-based	Cortical volume + area + thickness	102
	Cortical + subcortical volume	41
	Cortical volume	34
	Subcortical volume	7
	Hippocampus + accumbens volume	4
Machine Learning	StepForward	48
	LASSO	21

Table 4.1: Feature selections used to train a MetS classifier with number of features

the remaining 224 features. Because no research is found which shows that ROIs in the left or right hemisphere individually contribute to the developing of metabolic syndrome, the amount of features was further reduced by merging the same features in LH and RH. Merging was done by taking the mean of the LH and RH values of one brain area. All remaining features were used to train a second model (114).

In this study, several methods to further reduce the amount of features were used. These methods are described below and all feature selections, with the amount of features are shown in Table 4.1.

Knowledge-based feature selection

One method of feature selection is a knowledge-based method, that makes selections based on previous research. A possible feature selection could be to train the model on GM and TB volume of the patients. As described before, de Nijs (2018) found that TB volume, GM volume and a group of ten reward-related brain structures are smaller in the MetS+ group compared to the MetS- group. However, de Nijs' research used the same sample as the current study, which poses bias (Arbabshirani et al., 2017).

However, we could use the information of de Nijs' research that TB volume is associated with MetS. TB volume is based on cortical and subcortical volume, so instead of the TB volume feature, all volume cortical and subcortical ROI volume features (41) were used to train a model. Moreover, two additional models were trained on just volume cortical ROI volume features (34) or subcortical ROI volume features (7).

Alternatively, previous studies with a different data set can be used to select features. As described before, hippocampal volume (Yau et al., 2012) and volume of the nucleus accumbens (Song et al., 2015), both areas of the subcortical brain, are associated with a smaller brain in metabolic syndrome. Therefore a model is trained on the small feature set consisting of hippocampus and nucleus accumbens volumes (4).

Selecting features with machine learning

Another method for feature selection could be to use a machine learning algorithm to select a subset of features. In this study, *step forward algorithm* and *LASSO* were used to select a subset of the features. Step forward feature selection can be used with every machine learning algorithm possible and in this study a soft-margin SVM with 5-fold nested cross-validation was used. Balanced accuracy was used to test which combination performs best. The *step forward algorithm* had selected 48 features.

In the *LASSO regression* model also nested cross-validation was used, with 5 outer folds and 5 inner folds, to determine the value of the penalty. To extract the selected features, after cross-validation, a final model using all subjects is trained. This was the model with the optimal regularization parameter lambda. The final model that was trained using *LASSO regression* had selected 21 features.

4.4 Models to classify schizophrenia

Classification of subjects with MetS- and MetS+ in a group with schizophrenic patients based on brain measures is quite complex and is, as far as we know, never done before. Therefore, it is difficult to predict how well the classification will work. So, first of all, two models were trained with a dataset that was expected to be easier to classify. In this case, a SVM model separated patients with schizophrenia from healthy controls. These models used 288 subjects, which included all 84 subjects of the sample diagnosed with schizophrenia and a control group of 204 healthy subjects of whom MRI data is available.

A first model was trained using all features (after merging RH and LH). A second model used only the cortical volume ROIs as features to train the model. To keep this model the same as the other trained model, as far as possible, again a soft-margin version of SVM with 5-fold nested cross-validation was used, along with the same performance metrics (section 4.5).

The expectation was that the schizophrenia models would perform better than the MetS models. Not only is classification of schizophrenic patients considered easier, but also is the sample size used in the schizophrenia models almost 4 times as big as the sample used in the MetS models. As described in chapter 3.1, a model with a larger sample size is able to better generalise on new data (Vabalas et al., 2019), which will increase the model performance.

4.5 Performance Metrics

All models were compared on various performance metrics. More than one is chosen to get a more realistic view of the capacity of the algorithm.

First of all, the accuracy of all models was computed. Accuracy is, despite some failures (see chapter 3.3) a great measure to give a first view of the performance of the model.

Also balanced accuracy, sensitivity and specificity were calculated, which are more accurate metrics for an imbalanced dataset. In this case, sensitivity tells what proportion of people with metabolic syndrome are classified as positive by the algorithm. Because mislabelling someone with metabolic syndrome as healthy is worse than mislabelling a healthy subject as someone with metabolic syndrome, sensitivity can be seen as a more important metric than specificity that tells what proportion of healthy controls are classified as so. Moreover, AUC scores and odds ratio of the models were calculated.

Balanced accuracy scores were used to determine the optimal cost value in the Suport Vector Machine models. The other performance metrics were mainly used to compare the different models.

Finally, a permutation significance test was performed to obtain how big the chance of a random model that is performing as well as the original model. The class labels of the data set were randomly permuted 1000 times and the permutation significance score was calculated. This score was calculated by the ratio of the number of permutation models performing better than the original model and the number of permutations (1000). A threshold of p < 0.05 was chosen, that is a model with a permutation significance score smaller than 0.05 was considered significant.

4.6 Class weights

As mentioned before, the GROUP dataset is quite unbalanced. The dataset used in this study includes 73 patients with schizophrenia, of which 19 patients (around 26%) are diagnosed with metabolic syndrome and 54 patients are not (around 74%).

To penalise misclassifications of the positive (smaller) class heavier then misclassifications of the negative (larger) class, class weights are added. The class weights are opposed to the class size, that is the positive subjects are penalised 54 times, while the negative subjects are penalised 19 times.

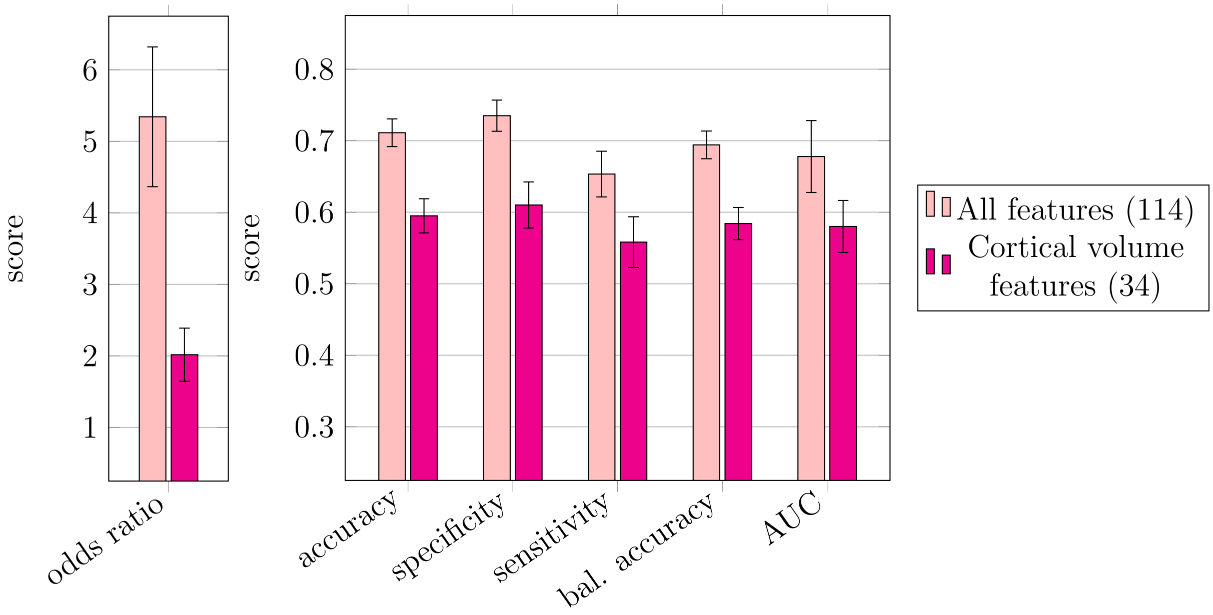
Chapter 5

Results

In this section, the performance of the classification models are presented and compared.

5.1 Schizophrenia classifier

Firstly, the support vector machine algorithms that classify subjects with and without schizophrenia are evaluated. As can be seen in Figure 5.1, the model using all features clearly has better performance scores on all performance metrics than the model using the cortical volume features.



(a) Odds ratio

(b) Accuracy, specificity sensitivity, balanced accuracy and AUC

Figure 5.1: Average performance scores over 100 runs of the schizophrenia classifier trained on two different feature selections with number of features and standard deviation

Figure 5.1 also shows that balanced accuracy and AUC score lie close to each other, with balanced accuracy around 69% and AUC score around 67%

for the model using all features. Both balanced accuracy and AUC score of the model using the cortical volume features are around 58%.

For both models, specificity score is higher than sensitivity score, indicating that the chance that a patient with schizophrenia is misclassified as a healthy control is higher, than the chance that a healthy control is misclassified as ill.

In addition, odds ratio of the model using all features (5.343) is more than two times as high as that of the cortical volume feature model (2.017).

A permutation significance test was done, of which the results are shown in Tabel 5.1. Both models have a permutation significance score smaller than 0.05, which means that both results are considered significant.

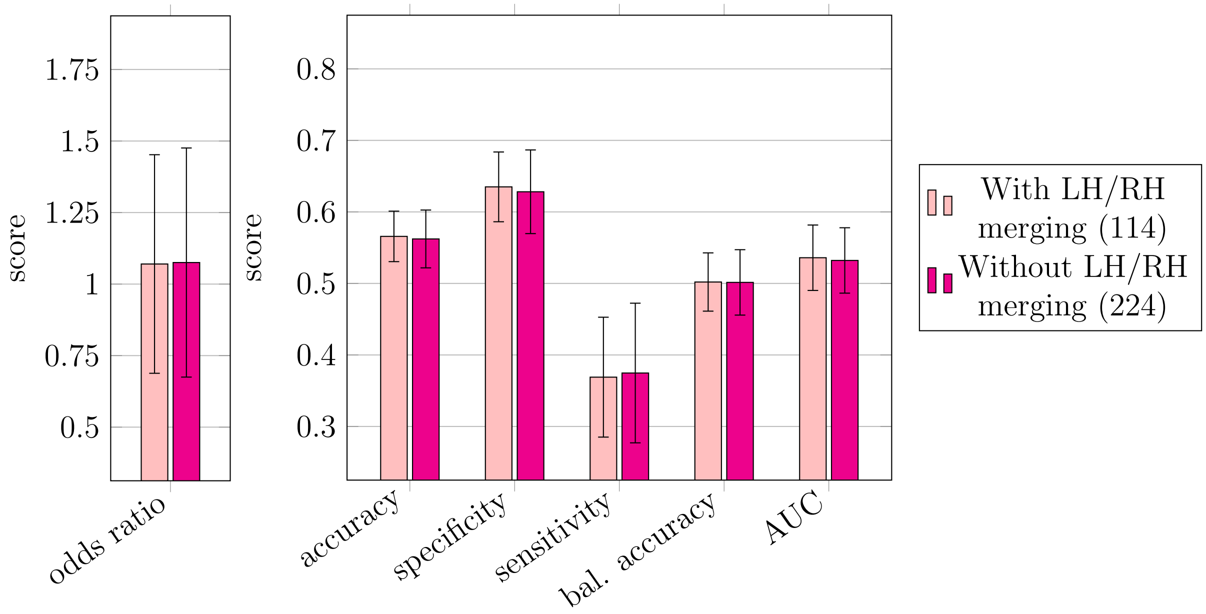
	Permutation significance score	Significant?
All features	< 0.001	Yes
Cortical volume features	0.014	Yes

Table 5.1: Permutation significance score over 1000 permutations of the schizophrenia classifier trained on two different feature selections

5.2 MetS classifiers

All features

The support vector machine algorithms that classify subjects with and without metabolic syndrome are evaluated. First, the models using all data with and without merging RH and LH are compared. The result is shown in Figure 5.2.



(a) Odds ratio

(b) Accuracy, specificity sensitivity, balanced accuracy and AUC

Figure 5.2: Average performance scores over 100 runs of the MetS classifiers trained with SVM on featuresets with and without merged LH/RH values with number of features and standard deviation

The model trained with merged features is performing not significantly better than the model without merged features on all performance metrics. For both models, specificity score (around 63%) is almost two times as big as sensitivity score (around 37%), which means that the chance that a MetS+ subject is misclassified as a MetS- subject is much higher, than the chance that a MetS+ subject is misclassified as a MetS- subject.

Note that the standard deviation of the sensitivity score and odds ratio is quite high, which could indicate an unstable model.

Knowledge-based feature selection models

The models using smaller knowledge-based selected feature subsets are compared. The model using all features (with merging) is added as a reference. The results are shown in Figure 5.3 and Figure 5.4.

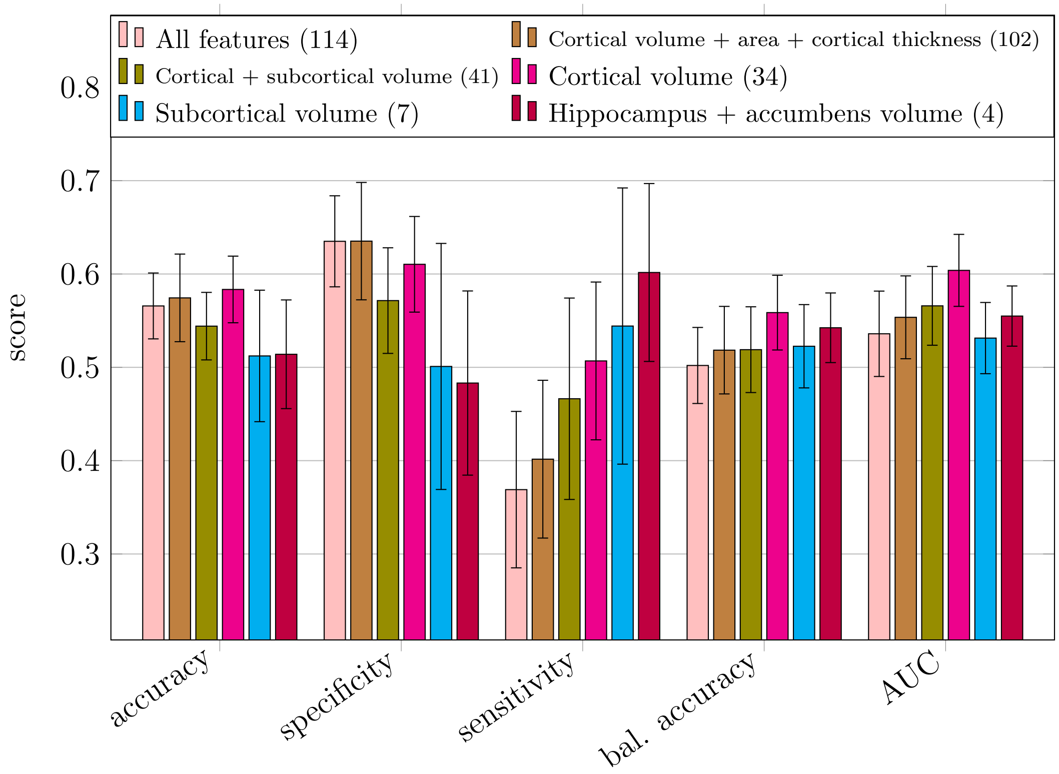


Figure 5.3: Average performance scores over 100 runs of the MetS classifier trained with SVM on different feature selections with number of features and standard deviation

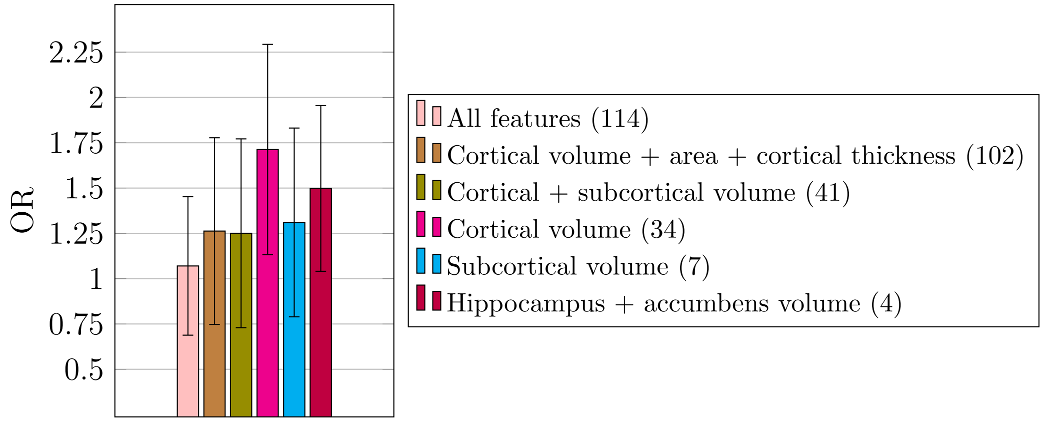


Figure 5.4: Average odds ratio over 100 runs of the MetS classifier trained on different feature selections with number of features and standard deviation

All models perform poorly, with balanced accuracy scores ranging from 50% to 56%. The best performing model measured on accuracy, balanced accuracy, AUC scores and OR is the model trained with the cortical volume features. The model with the highest sensitivity is the model using hippocampus and accumbens volume features, which is the model with the smallest number of features.

In fact, with regard to sensitivity, the smaller the amount of features, the higher the score. The opposite is true for specificity: the higher the amount of features, the higher the score, except for the specificity score of the cortical volume model (around 60%), which is higher than the cortical and subcortical volume model (around 56%).

Again it is striking that the standard deviation of the odds ratio and sensitivity score, and to a lesser extent the specificity score, is quite high. This could indicate an unstable model.

Moreover, permutation significance test are done. Table 5.2 shows the permutation significance scores of all knowledge based feature selection models.

	Permutation significance	Significant?
All features	0.499	No
Cortical volume + area + cortical thickness	0.520	No
Cortical + subcortical volume	0.370	No
Cortical volume	0.192	No
Subcortical volume	0.275	No
Hippocampus + accumbens volume	0.236	No

Table 5.2: Permutation significance score over 1000 permutations of the MetS classifier trained on six different feature selections

The permutation significance scores of all models are quite high, ranging from 0.192 for the model with cortical and subcortical volume features to 0.52 for the model using volume, area and cortical thickness of all cortical regions of interest.

Machine learning feature selection models

The models using smaller feature subsets selected by machine learning algorithms are evaluated. Table 5.3 shows the features selected by LASSO and StepForward algorithm.

Feature	**LASSO**	**StepForward**
parsorbitalis_VOL_freesurfer	✓	✓
bankssts_VOL_freesurfer	✓	✓
frontalpole_VOL_freesurfer	✓	✓
bankssts_area_freesurfer		✓
entorhinal_area_freesurfer	✓	✓
rostralmiddlefrontal_VOL_freesurfer		✓
transversetemporal_CT_freesurfer	✓	✓
supramarginal_VOL_freesurfer		✓
parstriangularis_area_freesurfer		✓
caudalanteriorcingulate_area_freesurfer	✓	✓
lingual_VOL_freesurfer		✓
cuneus_VOL_freesurfer		✓
inferiorparietal_area_freesurfer		✓
precentral_area_freesurfer		✓
parsorbitalis_area_freesurfer		✓
parsorbitalis_CT_freesurfer	✓	✓
supramarginal_area_freesurfer	✓	✓
precentral_VOL_freesurfer		✓
inferiortemporal_VOL_freesurfer		✓
inferiorparietal_VOL_freesurfer		✓
superiortemporal_area_freesurfer		✓
parahippocampal_CT_freesurfer		✓
superiorfrontal_VOL_freesurfer		✓
totalgray_VOL_freesurfer		✓
paracentral_CT_freesurfer		✓

medialorbitofrontal_area_freesurfer		✓
fusiform_area_freesurfer		✓
inferiorparietal_CT_freesurfer		✓
frontalpole_area_freesurfer	✓	✓
lingual_CT_freesurfer		✓
rostralanteriorcingulate_area_freesurfer		✓
precentral_CT_freesurfer		✓
totalbrain_VOL_freesurfer		✓
frontalpole_CT_freesurfer	✓	✓
bankssts_CT_freesurfer		✓
rostralanteriorcingulate_VOL_freesurfer		✓
fusiform_VOL_freesurfer		✓
inferiortemporal_area_freesurfer	✓	✓
caudate_VOL_freesurfer		✓
lingual_area_freesurfer		✓
middletemporal_area_freesurfer		✓
cuneus_CT_freesurfer		✓
superiorparietal_CT_freesurfer		✓
superiortemporal_CT_freesurfer		✓
superiorfrontal_area_freesurfer		✓
temporalpole_CT_freesurfer		✓
parsopercularis_CT_freesurfer		✓
caudalmiddlefrontal_CT_freesurfer		✓
pallidum_VOL_freesurfer	✓	
inferiortemporal_CT_freesurfer	✓	
middletemporal_CT_freesurfer	✓	
parstriangularis_CT_freesurfer	✓	
precuneus_CT_freesurfer	✓	
parstriangularis_VOL_freesurfer	✓	
caudalmiddlefrontal_area_freesurfer	✓	
pericalcarine_area_freesurfer	✓	
posteriorcingulate_area_freesurfer	✓	
parsopercularis_area_freesurfer	✓	

Table 5.3: Selected features of the LASSO model and StepForward model

As shown in Table 5.3, eleven features are selected by both algorithms. LASSO selected twenty features of the cortical brain and one feature of the subcortical brain, whereas StepForward selected 45 features of the cortical brain, one feature of the subcortical brain and two features based on larger brain volumes.

Since LASSO does not use SVM, an additional soft-margin support vector machine model is trained using the non-zero features selected by the LASSO model. The result of this model is compared to the StepForward model and the original LASSO model, which is visualised in Figure 5.5. The best of the knowledge-based feature selection models (using cortical volume), is also shown as a reference.

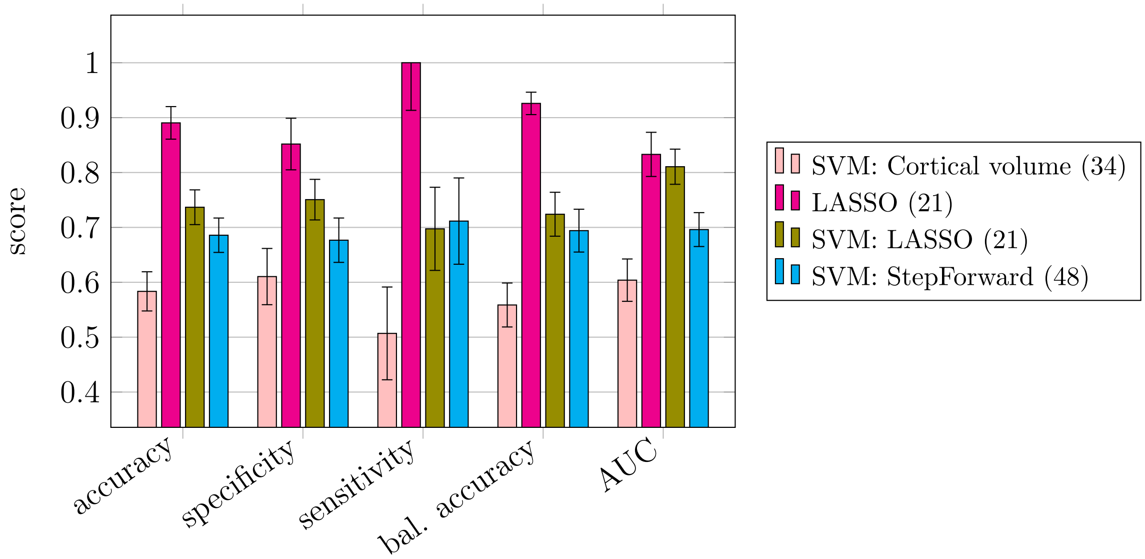


Figure 5.5: Average performance scores of the MetS classifier trained on different feature selections with number of features

Figure 5.5 shows that Step Forward and both variants of LASSO perform better than the knowledge based model, whereby the original LASSO model is performing best with performance metrics between 85% and 100%.

The SVM model using the non-zero LASSO features is also performing better on all performance metrics, except sensitivity, than the StepForward model with scores ranging between 70% and 81%. However, the scores of this model are clearly lower than the scores of the original LASSO model.

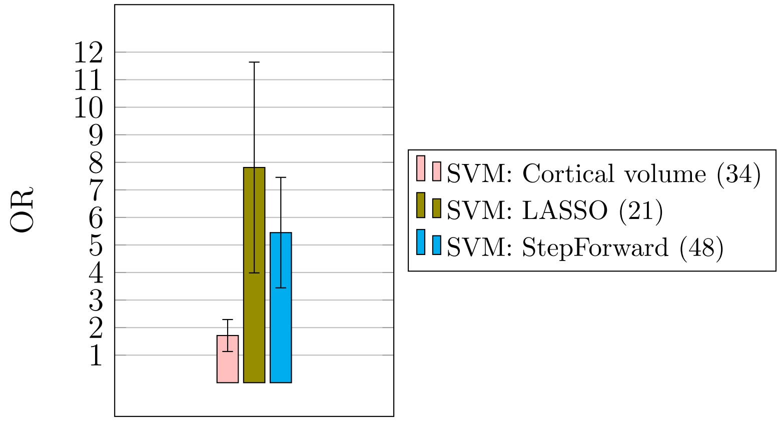


Figure 5.6: Average odds ratio of the MetS classifier trained on different feature selections with number of features and standard deviation

The odds ratios of the models are calculated. Because in the LASSO model the amount of False Negatives is equal to 0, the odds ratio of the model could not be calculated (dividing by zero is not allowed). The other three models are shown in Figure 5.6. The SVM model using the LASSO features has the highest OR (7.812), however the standard deviation is quite high.

	Permutation significance	Significant?
LASSO	0.13	No
SVM: LASSO	< 0.001	Yes
SVM: StepForward	0.006	Yes

Table 5.4: Permutation significance score over 1000 permutations of the MetS classifier trained on two feature selections selected by machine learning algorithms: LASSO and StepForward

Moreover, the permutation significance scores of the feature selections are calculated. As shown in Table 5.4, the odds ratio both SVM models are smaller than 0.05 and are therefore considered significant. The LASSO model is not considered significant, yet LASSO is an embedded model. Therefore, the feature selection process is done in the training phase of the machine learning model. Therefore, the feature selection process is done in the training phase of the machine learning algorithm. If the class labels are randomly permuted, different features might be selected while training, which causes a some what unfair comparison.

Chapter 6

Discussion

In this study, classifiers are built that tend to divide patients with and without schizophrenia. Inside the schizophrenia group classification of subjects with and without metabolic syndrome is attempted. Different feature are used to train the models and those features are selected based on written literature (knowledge-based) and machine learning approaches.

Schizophrenia models

Two feature sets were used to train schizophrenia classifiers. The model using all features was the best of the two models, with an estimated balanced accuracy score around 69%. With a permutation significance of <0.001 and relatively small variance among different runs, the result of the model is considered stable.

The other model, that was trained on cortical volume features, is also significant, with a permutation significance score of 0.0014. It is notable that this model performs worse than the model using all features. However, a direct explanation cannot be given. More research with different feature combinations should be done to draw a conclusion about which features have the greatest link with schizophrenia.

Knowledge-based feature selection models

The metabolic syndrome classifier with a knowledge-based features selection that is performing best, is the model using cortical volume features with a balanced accuracy of 56% and a AUC score of 60%. However, based on De Nijs et al. (2018), the expectation was that the model using cortical and subcortical volume features would perform best. De Nijs' statistical study found that total brain volume of subjects with metabolic syndrome is smaller. Because total brain volume is based on both cortical and subcortical volume, the expectation was that the model trained on both cortical and subcortical volume features would be the best performing model. A possible explanation for our different result is that the decreased total brain volume De Nijs found is mainly caused by abnormalities in cortical brain features.

In fact, none of the knowledge-based model is yielding good performance results. Balanced accuracies range from 50% to 54%, while none of the permutation significance scores are lower than 0.05. Moreover, the sensitivity and odds ratio score of all models have a large standard deviation, meaning that the variance among different runs is big. With regard to sensitivity, the small positive class contributes to the large standard deviation, since even 1 True Positive more or less causes a relatively big difference in sensitivity score. However, even with this in mind, the standard deviation is still relatively large, which could indicate an unstable model.

Remarkably, sensitivity and specificity score seems to be associated with the amount of features used to train the model. The more features are used, the higher the specificity and the lower the sensitivity score becomes. More research should be done to investigate this relationship.

Comparison MetS and schizophrenia models

The expectation was that the schizophrenia models would perform better than the MetS models. With regard to the model using all features, this clearly was the case. Nevertheless, the model using the cortical volume features yields similar results for both classifiers. Accuracy, odds ratio and balanced accuracy were higher for the schizophrenia classifier, while AUC score was higher for the MetS classifier. However, the model trained on cortical volume features was the best performing model of the knowledge-based MetS models, while the model with the same feature selection was the worst of the schizophrenia models. This makes the comparison not entirely fair, but it indicates that metabolic syndrome somehow is associated with cortical volume features. Also, the permutation significance score of the MetS model was not significant, while the score of the schizophrenia model was significant.

Machine learning feature selection models

Both LASSO and StepForward models are performing clearly better than all knowledge-based feature selection model, wherby the first is performing best with a balanced accuracy of 93%. This result suggest that in classification of metabolic syndrome a machine learning feature selection out-performs knowledge-based feature selection, despite a great amount of studies to brain abnormalities in subjects with metabolic syndrome. Nevertheless, more research should be done using more samples.

An additional model is trained using the selected features by the LASSO algorithm in a support vector machine. This model is also having pretty good results (balanced accuracy: 73%), but not as good as the LASSO model. A possible explanation is that the performances of the LASSO model might be over-optimistic given the fact that after the nested cross-validation procedure one model is trained to obtain the features selected by the algorithm. Another explanation of the different performances could be the choice of the machine learning algorithm, in this case LASSO, that could effect the performance of the models. Further research should be done that compare different machine learning algorithms. A different validation set is preferred to avoid bias.

The permutation significance scores of both the StepForward and the SVM: LASSO model are <0.05, which suggest that the chance the performance of the models is yield by luck is small.

Amount of features in feature selection

From our literature review about using a small sample size in machine learning, we expected that merging features of the left and right hemisphere would result in a better performing model, since a lower ratio between the amount of features and sample size usually causes a model that better generalises to new data (Vabalas et al., 2019). In our study, the model with merged features of both hemispheres indeed performs a little better than the model without merged features, yet the difference is not as striking as expected. A possible reason for this could be that features of ROIs in the left and right hemisphere individually influence the occurrence of metabolic syndrome. A study is found that indicates a certain asymmetry between volume of the left and right hemisphere (Goldberg et al., 2013). More research on the impact of the hemispheric volume difference on metabolic syndrome should be done to conclude if merging features of left and right hemisphere is a good step. Based on the present study, it seems not.

Furthermore, if the models with smaller knowledge-based features selections are considered, models with only a few features not necessarily are the best performing models. However, it could also be the case that in those models the features that are selected are not the features that have the greatest link to the target and therefore the performance of these models is lower. Also, the amount of features does not seem to play a role. Perhaps, a certain threshold exist with regard to the ratio between amount of features and sample size. A possible further research could be to investigate whether there is such a threshold.

The problem of an unstable model, of which the expectation was that it is associated with amount of features, is seen in the knowledge-based models more than in the machine learning feature selection models. This suggest that the stability of a model over different runs does not depend on amount of features but rather on which features are selected.

Clinical utility

The major reason that the sample size of this study is relatively small, is that neuro-imaging data collection (i.e. MRI scans) is quite expensive (Kononenko, 2001). Therefore, a model based on MRI data should perform really well to be interesting for clinical use. If the results of the machine learning feature selection models are compared to a model that is trained using clinical features, both LASSO (balanced accuracy: 93%, AUC: 83%) and StepForward models (balanced accuracy: 69%, AUC: 70%) out-perform the clinical model (balanced accuracy: 61%, AUC: 62%) (van de Poppe, 2018). The clinical model uses a soft-margin SVM, as well as the StepForward model. The sample of the clinical model is part of the GROUP project, yet the sample size of the clinical model is higher (N = 1973). The result suggest that further research with MRI scans could be worth it, despite the high costs, also with a higher sample size.

Moreover, models used in a clinical setting should be explainable, which means that it must be able to explain why a certain diagnosis is made (Kononenko, 2001). Models trained on features selection made by machine learning algorithms are not always explainable and therefore not clinical usable. However, those models could inspire further (statistically) research. For example, eleven features are selected by both LASSO model and StepForward model. A further research could be done to look whether there is an association between one or more of these features and the occurence of metabolic syndrome.

Note that a large majority of the subjects (89%) is male, which is a major limitation of this study. Furthermore, there are indications that volumes of specific brain areas differ between men and women (Schlaepfer et al., 1995), which could effect the presence of metabolic syndrome. Further research should be done with a more balanced dataset and sex as feature before a model could be of clinical utility.

Chapter 7

Conclusion

The purpose of this study was to get a better understanding of the association between metabolic syndrome and the cortical and subcortical regions of interest in the brain and to gain insight into the way machine learning can contribute to neuro-imaging field.

This study shows that despite a small sample size and class imbalance, machine learning can be useful in the classification of subjects with and without metabolic syndrome. Feature selection plays an important role in how well a model performs. In this study, machine learning feature selections models perform significantly better than models trained with knowledge-based selected features, which suggest that especially in feature selection machine learning could be helpful.

Models using machine learning for feature selection cannot directly be of clinical utility, because such models should be understandable for clinicians. Therefore, it should be understood how selected features relate to certain diseases. This can be achieved by additional clinical research to the features selected by a machine learning algorithm. Machine learning could help to find associations between disaeses and brain features, which seems the most important way machine learning can contribute to the neuro-imaging field.

References

Aha, D. W., & Bankert, R. L. (1996). A comparative evaluation of sequential feature selection algorithms. In *Learning from data* (pp. 199–206). Springer.

Alberti, K. G. M. M., Zimmet, P., & Shaw, J. (2006). Metabolic syndrome—a new world-wide definition. a consensus statement from the international diabetes federation. *Diabetic medicine*, *23*(5), 469–480.

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage*, *145*, 137–165.

Berrar, D. (2019). Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology*, *1*, 542–545.

Bhugra, D. (2005). The global prevalence of schizophrenia. *PLoS Med*, *2*(5), e151.

Cahn, W., Pol, H. E. H., Lems, E. B., van Haren, N. E., Schnack, H. G., van der Linden, J. A., ... Kahn, R. S. (2002). Brain volume changes in first-episode schizophrenia: a 1-year follow-up study. *Archives of general psychiatry*, *59*(11), 1002–1010.

Cameron, A. J., Shaw, J. E., & Zimmet, P. Z. (2004). The metabolic syndrome: prevalence in worldwide populations. *Endocrinology and Metabolism Clinics*, *33*(2), 351–375.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28.

Choe, E. K., Rhee, H., Lee, S., Shin, E., Oh, S.-W., Lee, J.-E., & Choi, S. H. (2018). Metabolic syndrome prediction using machine learning models with genetic and clinical information from a nonobese healthy population. *Genomics & Informatics*, *16*(4).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Das, S. (2001). Filters, wrappers and a boosting-based hybrid for feature selection. In *Icml* (Vol. 1, pp. 74–81).

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., ... others (2006). An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, *31*(3), 968–980.

van Erp, T. G., Hibar, D. P., Rasmussen, J. M., Glahn, D. C., Pearlson, G. D., Andreassen, O. A., ... others (2016). Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the enigma consortium. *Molecular psychiatry*, *21*(4), 547–553.

Falahati, F., Westman, E., & Simmons, A. (2014). Multivariate data analysis and machine learning in alzheimer's disease with a focus on structural magnetic resonance imaging. *Journal of Alzheimer's disease*, *41*(3), 685–708.

Fonti, V., & Belitser, E. (2017). Feature selection using lasso. *VU Amsterdam Research Paper in Business Analytics*, *30*, 1–25.

Ford, E. S., Li, C., & Zhao, G. (2010). Prevalence and correlates of metabolic syndrome based on a harmonious definition among adults in the us. *Journal of diabetes*, *2*(3), 180–193.

Galassi, A., Reynolds, K., & He, J. (2006). Metabolic syndrome and risk of cardiovascular disease: a meta-analysis. *The American journal of medicine*, *119*(10), 812–819.

Goldberg, E., Roediger, D., Kucukboyaci, N. E., Carlson, C., Devinsky, O., Kuzniecky, R., ... Thesen, T. (2013). Hemispheric asymmetries of cortical volume in the human brain. *cortex*, *49*(1), 200–210.

Gunčar, G., Kukar, M., Notar, M., Brvar, M., Černelč, P., Notar, M., & Notar, M. (2018). An application of machine learning to haematological diagnosis. *Scientific reports*, *8*(1), 1–12.

Gutiérrez-Esparza, G. O., Infante Vázquez, O., Vallejo, M., & Hernández-Torruco, J. (2020). Prediction of metabolic syndrome in a mexican population applying machine learning algorithms. *Symmetry*, *12*(4), 581.

Haijma, S. V., van Haren, N., Cahn, W., Koolschijn, P. C. M., Hulshoff Pol, H. E., & Kahn, R. S. (2013). Brain volumes in schizophrenia: a meta-analysis in over 18 000 subjects. *Schizophrenia bulletin*, *39*(5), 1129–1138.

Heald, A., Pendlebury, J., Anderson, S., Narayan, V., Guy, M., Gibson, M., ... Livingston, M. (2017). Lifestyle factors and the metabolic syndrome in schizophrenia: a cross-sectional study. *Annals of General Psychiatry*, *16*(1), 12.

Jović, A., Brkić, K., & Bogunović, N. (2015). A review of feature selection methods with applications. In *2015 38th international convention on information and communication technology, electronics and microelectronics (mipro)* (pp. 1200–1205).

Kakudi, H. A., Loo, C. K., & Moy, F. M. (2020). Diagnosis of metabolic syndrome using machine learning, statistical and risk quantification techniques: A systematic literature review. *medRxiv*.

Karimi-Alavijeh, F., Jalili, S., & Sadeghi, M. (2016). Predicting metabolic syndrome using decision tree and support vector machine methods. *ARYA atherosclerosis*, *12*(3), 146.

Kassi, E., Pervanidou, P., Kaltsas, G., & Chrousos, G. (2011). Metabolic syndrome: definitions and controversies. *BMC medicine*, *9*(1), 48.

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

van der Kooij, A., & Meulman, J. (2008). Regularization with ridge penalties, the lasso, and the elastic net for regression with optimal scaling transformations. *Submitted for publication*.

Koolschijn, P. C. M., van Haren, N. E., Cahn, W., Schnack, H. G., Janssen, J., Klumpers, F., ... Kahn, R. S. (2010). Hippocampal volume change in schizophrenia. *The Journal of clinical psychiatry*, *71*(6), 737–744.

Korver, N., Quee, P. J., Boos, H. B., Simons, C. J., de Haan, L., & Investigators, G. (2012). Genetic risk and outcome of psychosis (group), a multi site longitudinal cohort study focused on gene–environment interaction: objectives, sample characteristics, recruitment and assessment methods. *International journal of methods in psychiatric research*, *21*(3), 205–221.

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

Kubota, M., van Haren, N. E., Haijma, S. V., Schnack, H. G., Cahn, W., Pol, H. E. H., & Kahn, R. S. (2015). Association of iq changes and progressive brain changes in patients with schizophrenia. *JAMA psychiatry*, *72*(8), 803–812.

Lal, T. N., Chapelle, O., Weston, J., & Elisseeff, A. (2006). Embedded methods. In *Feature extraction* (pp. 137–165). Springer.

Ling, C. X., Huang, J., Zhang, H., et al. (2003). Auc: a statistically consistent and more discriminating measure than accuracy. In *Ijcai* (Vol. 3, pp. 519–524).

Liu, Y., Yu, X., Huang, J. X., & An, A. (2011). Combining integrated sampling with svm ensembles for learning from imbalanced datasets. *Information Processing & Management*, *47*(4), 617–631.

Longadge, R., & Dongre, S. (2013). Class imbalance problem in data mining review. *arXiv preprint arXiv:1305.1707*.

Luts, J., Ojeda, F., Van de Plas, R., De Moor, B., Van Huffel, S., & Suykens,

J. A. (2010). A tutorial on support vector machine-based methods for classification problems in chemometrics. *Analytica Chimica Acta*, *665*(2), 129–145.

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.

de Nijs, J., Schnack, H., Koevoets, M., Kubota, M., Kahn, R., van Haren, N., & Cahn, W. (2018). Reward-related brain structures are smaller in patients with schizophrenia and comorbid metabolic syndrome. *Acta Psychiatrica Scandinavica*, *138*(6), 581–590.

Olabi, B., Ellison-Wright, I., McIntosh, A. M., Wood, S. J., Bullmore, E., & Lawrie, S. M. (2011). Are there progressive brain changes in schizophrenia? a meta-analysis of structural magnetic resonance imaging studies. *Biological psychiatry*, *70*(1), 88–96.

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.

Papanastasiou, E. (2013). The prevalence and mechanisms of metabolic syndrome in schizophrenia: a review. *Therapeutic advances in psychopharmacology*, *3*(1), 33–51.

van de Poppe, J. (2018). Support vector machine based prediction of metabolic syndrome in schizophrenia. *Preprint*.

Pudil, P., Novovičová, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern recognition letters*, *15*(11), 1119–1125.

Rahman, M. M., & Davis, D. N. (2013). Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, *3*(2), 224.

Reaven, G. M. (1988). Role of insulin resistance in human disease. *Diabetes*, *37*(12), 1595–1607.

Ringen, P. A., Engh, J. A., Birkenaes, A. B., Dieset, I., & Andreassen, O. A. (2014). Increased mortality in schizophrenia due to cardiovascular disease–a non-systematic review of epidemiology, possible causes, and interventions. *Frontiers in psychiatry*, *5*, 137.

Ritchie, S., & Connell, J. (2007). The link between abdominal obesity, metabolic syndrome and cardiovascular disease. *Nutrition, Metabolism and Cardiovascular Diseases*, *17*(4), 319–326.

Sala, M., de Roos, A., van den Berg, A., Altmann-Schneider, I., Slagboom, P. E., Westendorp, R. G., ... van der Grond, J. (2014). Microstructural brain tissue damage in metabolic syndrome. *Diabetes Care*, *37*(2), 493–500.

van Schependom, J., Yu, W., Gielen, J., Laton, J., De Keyser, J., De Hert, M., & Nagels, G. (2015). Do advanced statistical techniques really help in the diagnosis of the metabolic syndrome in patients treated with second-generation antipsychotics? *The Journal of clinical psychiatry*, *76*(10), 1292–1299.

Schlaepfer, T. E., Harris, G. J., Tien, A. Y., Peng, L., Lee, S., & Pearlson, G. D. (1995). Structural differences in the cerebral cortex of healthy female and male subjects: a magnetic resonance imaging study. *Psychiatry Research: Neuroimaging*, *61*(3), 129–135.

Schnack, H. G. (2019). Improving individual predictions: machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases). *Schizophrenia research*, *214*, 34–42.

Simon, R. (2003). Supervised analysis when the number of candidate features (p) greatly exceeds the number of cases (n). *ACM SIGKDD Explorations Newsletter*, *5*(2), 31–36.

Simon, S. D. (2001). Understanding the odds ratio and the relative risk. *Journal of andrology*, *22*(4), 533–536.

Song, S.-W., Chung, J.-H., Rho, J. S., Lee, Y.-A., Lim, H.-K., Kang, S.-G., ... Kim, S.-H. (2015). Regional cortical thickness and subcortical volume changes in patients with metabolic syndrome. *Brain imaging and Behavior*, *9*(3), 588–596.

Stern, M. P., Williams, K., González-Villalpando, C., Hunt, K. J., & Haffner, S. M. (2004). Does the metabolic syndrome improve identification of individuals at risk of type 2 diabetes and/or cardiovascular disease? *Diabetes care*, *27*(11), 2676–2681.

Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian academy of child and adolescent psychiatry*, *19*(3), 227.

Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. *Data classification: Algorithms and applications*, 37.

Tiehuis, A. M., van der Graaf, Y., Mali, W. P., Vincken, K., Muller, M., & Geerlings, M. I. (2014). Metabolic syndrome, prediabetes, and brain abnormalities on mri in patients with manifest arterial disease: the smart-mr study. *Diabetes care*, *37*(9), 2515–2521.

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PloS one*, *14*(11), e0224365.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, *7*(1), 91.

Yau, P. L., Castro, M. G., Tagani, A., Tsui, W. H., & Convit, A. (2012). Obesity and metabolic syndrome and functional and structural brain impairments in adolescence. *Pediatrics*, *130*(4), e856–e864.

Appendix A

Regions of Interest cortical and subcortical

Cortical		Subcortical
bankssts	parsorbitalis	thalamus
caudalanteriorcingulate	parstriangularis	caudate
caudalmiddlefrontal	pericalcarine	putamen
cuneus	postcentral	pallidum
entorhinal	posteriorcingulate	hippocampus
fusiform	precentral	amygdala
inferiorparietal	precuneus	accumbens
inferiortemporal	rostralanteriorcingulate	
isthmuscingulate	rostralmiddlefrontal	
lateraloccipital	superiorfrontal	
lateralorbitofrontal	superiorparietal	
lingual	superiortemporal	
medialorbitofrontal	supramarginal	
middletemporal	frontalpole	
parahippocampal	temporalpole	
paracentral	transversetemporal	
parsopercularis	insula	