Alzheimer's Disease Prediction

A multi-classification approach for predicting the moment of progression from Mild Cognitive Impairment subjects to Alzheimer's disease with MRI biomarkers

Julia van Veen
ANR: 514474
Master Thesis DSBG

Thesis committee:
Dr. W. Huijbers
Dr. M. Postma

Tilburg University
School of Humanities
Tilburg, The Netherlands
July 2017

Preface

This thesis has been written to complete my Master's program Data Science: Business and Governance at Tilburg University. I was engaged in researching and writing this thesis from January 2017 to July 2017. First and foremost I would like to thank Willem Huijbers for supervising me during this process. I am also grateful to Nanne van Noord, for the times he engaged in the discussions between me and Willem Huijbers.

Julia van Veen
Tilburg, July 217

Abstract

The current study opens up a new field of not only predicting '*if*' a subject progresses to Alzheimer's disease but also '*when*' this moment of progression likely will take place, since research has not yet addressed this multiclass classification problem. Predicting the moment of progression is of interest because this could contribute to finding therapies that modify the disease. The main purpose of the current study is to investigate to the extent to which predicting the progression of Mild Cognitive Impairment (*MCI*) to Alzheimer's disease and this progression's corresponding moment based on MRI biomarkers is possible. To classify subjects from the ADNI database to either *MCI* non-converters or *MCI* converters in 6, 12, 18, 24, 30 or 48 months in the future, the author conducted three experiments that led to a final model based on the Support-Vector classifier. This model performed significantly better than the Dummy classifier on the multiclass classification problem. However, these results must be interpreted with caution since results from a follow-up experiment suggest that the intervals of the moment of progression are too small. Future research should investigate the impact of different time intervals on the performance of the classifier to improve the model.

Keywords: Alzheimer's disease, moment of progression, multiclass classification, imbalanced data, Support-Vector classifier, Decision-Tree classifier, Logistic Regression, Stochastic Gradient Descent, Perceptron, linear Support-Vector classifier

Contents

Section 1: Introduction

This section provides a general introduction to Alzheimer's disease prediction. The introduction is divided into four parts. Subsection 1.1 provides the general context of the current study. Subsection 1.2 presents the problem statement and the corresponding research questions, followed by the scientific and practical relevance in Subsections 1.3 and 1.4, respectively.

1.1 Context

The current study was conducted in the field of Alzheimer's disease prediction. The prediction of Alzheimer's disease is a relatively new field because not that long ago, the classification problem targeted by researchers shifted from mainly distinguishing Alzheimer's disease subjects from controls to the far more challenging problem of predicting Alzheimer's progression (Weiner et al., 2015). In this new field, findings may have clinical application. A physician can use this prediction to identify subjects who will progress to Alzheimer's disease. This will then allow them to take suited actions for the subject, such as making psychological interventions and prescribing proper medication.

Alzheimer's disease is most notable by (short-term) memory loss and other behavioral changes, which are caused by degeneration of brain cells (Moini, 2015). Although "Alzheimer's disease" is often used interchangeably with "dementia", it is important to note that these two terms are not the same. According to the Diagnostic and Statistical Manual of Mental Disorders, dementia is a general term for a decline in mental ability, which has to be severe enough to interfere with daily life (American Psychiatric Association, 2013). Alzheimer's disease is the most common form of dementia. Explaining other forms of dementia, such as vascular dementia, frontotemporal dementia and Lewy body dementia, is beyond the scope of the current study and are, therefore, not discussed further.

To predict Alzheimer's disease, it is important to know how this disease develops over time. For this, some general terminology has to be introduced. Throughout the current study, the terms *progression* and *conversion* refer to the process of moving towards another stage. The first stage is Clinical Normal (from here on referred to as *CN*). A *CN* subject has no signs of cognitive decline. When there are cognitive changes that are significant enough to be noticed by the subject or by others surrounding him or her, but not large enough to interfere with daily life, the subject converts from *CN* to Mild Cognitive Impairment (from here on referred to as *MCI*). Mild Cognitive Impairment is an intermediate stage between age-related cognitive decline and Alzheimer's disease. It is sometimes divided into early *MCI* and late *MCI*, which are denoted as *EMCI* and *LMCI*, respectively. The difference between *EMCI* and *LMCI* is that *LMCI* subjects have a lower score on the Logical Memory II subscale from the Wechsler Memory Scale (ADNI Manual, 2017). Mild Cognitive Impairment subjects are at risk of converting to Alzheimer's disease. Mitchell and Shiri-Feshki (2009) indicated

that annually, 7% of *MCI* subjects convert to Alzheimer's disease (*AD*), which is the final stage. The authors also indicated that the remaining subjects remained stable in *MCI*, developed other forms of dementia, or converted back to *CN*.

The current study concerns the prediction of which subjects at the *MCI* stage will progress to *AD*. A distinction is made between subjects diagnosed with *MCI* who remain stable and subjects diagnosed with *MCI* who progress to *AD* over time. For classification, subjects who remain stable are defined as *stableMCI*, and subjects who progress to *AD* are defined as *progressionMCI*. The development of *AD* over time is comprehensively explained in Section 2.1.

To predict which *MCI* subjects will progress to *AD*, machine-learning approaches have been adopted (for review, see Weiner et al. (2015)). Machine learning is a field of pattern recognition that focuses on automatically detecting patterns in example data or through past experiences. These patterns are the basis for predicting.

There are different types of machine learning, namely supervised learning, unsupervised learning and reinforcement learning. The current study focusses on supervised learning. Supervised learning is a machine-learning task of inferring a model from supervised-training data. The training data consist of a set of training examples. Each example consists of an input object and its label or target. In the case of predicting *AD*, the input object consists of medical information of the subject and the label of whether the subject is likely to progress to *AD*. The learning algorithm analyzes the training data and produces a model that should predict the correct label for any valid input object. This requires the learning algorithm to generalize from the training data to unseen data in a reasonable manner.

The medical information that is the basis for the input object could be vast. However, for the clinical application, gathering information based on many different methods to make a prediction is not ideal (Bauer, Rosendaal & Heit, 2012). In an ideal world, for every diagnosed *MCI* subject, a prediction is made on whether the subject is likely to progress to *AD*. To do so, the methods used for gathering medical data for the input object should be executed on a large scale. Thus, gathering all the information on a subject using all available methods would be costly.

To consider the clinical application of findings of the current study, the author compares the three most common techniques that provide insight into the health of a subject, namely Positron Emission Tomography (PET) scans, Magnetic Resonance Imaging (MRI) scans and lumbar puncture (for cerebrospinal fluid data), and subsequently selects one method. Regarding MRI's applicability on a large scale, MRI scans are already performed on subjects who face troubles with their memory, which make MRI already accessible. Besides that, MRI scans are less expensive than PET scans and safer than PET scans and lumbar puncture.

Data from MRI scans are also of significant importance in the prediction of *AD*. MRI scans

provide insight into the volumes of different parts of the brain. Changes in volume of certain brain structures can be measured by MRI scans (Douglas, 1995). According to Jack et al. (2011), these changes, especially in the medial temporal structures, are considered to be a valid biomarker for conversion to *AD* at the *MCI* stage (Jack et al., 2011). Most prior studies relating to the prediction of *AD* included MRI biomarkers in the input object. Including MRI biomarkers in the input object resulted in higher accuracies in comparison with those achieved by excluding them when predicting *AD* progression.

Since MRI scans are accessible at the *MCI* stage and have proven to be of significant importance for the prediction of *AD*, the selected method in the current study is MRI scans. As a result, the input object in the current study is based on MRI biomarkers. The aforementioned leads to a trade-off. Excluding highly informative features from PET scans, lumbar puncture or other methods not investigated in the current study could lower the performance of the learning algorithm. However, this exclusion is made to consider its clinical application.

An interesting aspect of predicting whether a subject will progress to *AD* is not only knowing '*if*' but also '*when*' this progression is likely to take place. For clinical application, this information is of importance for both the subject and medical treatment. In terms of medication and psychological interventions, it could be that an effective treatment differs for a subject who is likely progress to *AD* in four years in comparison with that of subjects who are likely to progress in six months. The practical and scientific importance is further discussed in Sections 1.3 and 1.4, respectively.

To predict the moment of progression, the task shifts from binary classification to multiclass classification. Almost all studies concerning *AD* prediction are binary-classification problems (Huang, Yang, Feng & Cheng, 2017; Trezpacs, Sun, Schuh, Case & Witte, 2014; Chupin et al., 2009; Devanand et al., 2007). This initially means that the learning algorithm could classify each training example into two classes: progression to *AD* and no progression to *AD*. In the case of multiclass classification, there are more than two classes into which a training example could be classified. For predicting the moment of progression, possible classes are progression in one year, progression in two years, progression in three years and no progression.

Despite the importance of predicting the moment of progression, no research has investigated it yet. The reason why no researchers have investigated this before might be because multiclass classification is much harder than binary classification. Also, the differences between the classes when the progression point is more in the future may be too subtle and complex for a learning algorithm to find their pattern. Nevertheless, considering that predicting the moment of progression could benefit both the subject and the medical treatment, the extent to which it is possible to predict the moment of progression is worth investigating and discovering.

1.2 Problem Statement and Research Questions

The goal of the current study is to investigate whether it is possible to predict not only *if* a subject will progress to *AD* but also *the moment* this progression will take place. This is examined by using supervised learning based on data from MRI scans. Therefore, the problem statement (PS) for the current study is as follows:

PS: *To what extent can a classifier predict the progression of subjects from MCI to AD and the progression's corresponding moment, based on MRI biomarkers?*

Three research questions are set up to find an answer to the problem statement. The first research question (RQ) examines a classifier that performs best in distinguishing *MCI* and *AD* subjects at baseline. The second research question investigates the extent to which the classifier from experiment 1 is able to make a distinction between *stableMCI* and different *progressionMCIs* in a binary classification task. The third research question examines the extent to which the classifier from experiment 1 is able to make a multiclass classification between *stableMCI* and different a *progressionMCIs*.

The first research question is formulated as follows:

RQ1: *What classifier and in combination with which pre-processing method performs best in distinguishing MCI subjects from Alzheimer's disease subjects at baseline*

The first research question examines, in a proof-of-concept study, a classifier that performs best in distinguishing the following two classes: *MCI* subjects and *AD* subjects at baseline. These two classes are easier to distinguish than *stableMCI* and *progressionMCI*, and this ease makes it easier to verify that this classifier has practical potential. Different classifiers are investigated: Decision-Tree classifier, linear Support-Vector classifier, Logistic Regression, Perceptron, Stochastic Gradient Descent and Support-Vector classifier. Besides that, the best pre-processing step for each classifier will also be investigated. The pre-processing methods that are investigated are standardizing, normalizing and adding feature interactions. The classifier that performs best is used throughout the remainder of the study.

The second research question is as follows:

RQ2: *To what extent can the optimized classifier make a binary distinction between* stable *MCI subjects and MCI subjects who progress to Alzheimer's disease within 6, 12, 24, 30 and 48*

months?

To predict progression and its corresponding moment, the extent to which the classifier from experiment 1 is able to make a distinction between *stableMCI* and different *progressionMCIs* in a binary-classification task is made. The moment of progression for the *progressionMCI* subjects is after 6, 12, 24, 30 and 48 months. These time intervals are chosen in such a manner that every subset consists of at least 50 subjects before its pre-processing steps take place. This research question will gain insight into the complexity of making a distinction between the different classes. The hypothesis that will be tested is that classification is more complex when the progression moment is further away in time. Also, the performance achieved in this experiment can be compared to the performance of classifiers from previous studies to verify that the performance of this experiment's classifier is in line with previously achieved performances. From now on, *progression6MCI* refers to the data of the subjects available 6 months before the progression takes place, *progression12MCI* represents the data of the subjects available 12 months before the progression takes place, and so on.

The third and last research question is formulated as follows:

RQ3: *To what extent can the optimized classifier predict progression from MCI subjects to AD and its corresponding moment in a multiclass classification task?*

From here on, multiclass classification is investigated. The six classes are the same as those used for experiment 2, namely *stableMCI, progression6MCI, progresion12MCI, progression24MCI, progression30MCI and progression48MCI*. Since multiclass classification is a different task from binary classification, the optimal parameters are investigated again. Finally, based on these findings and the results of the previous research questions, the optimized classifier is compared to a Dummy classifier to investigate the difference in performance of predicting progression and its corresponding moment.

1.3 Scientific Relevance

Since no previous studies tackled the multiclass problem regarding *AD*, the current study can be seen as a gateway to a new field in which the question of not only '*if*' subjects progress to *AD* but also '*when*' subjects progress to *AD* is an important question to be answered.

The findings could make an important contribution to finding *AD*-modifying therapies. Currently, there is no *AD*-modifying therapy available (Sperling, Jack & Aisen, 2011). Series of disappointing clinical trials over the past decade have raised concerns about the current strategy for the development of *AD*-modifying therapies. The lack of clinical benefit could be caused by attempting

interventions at the wrong stage of the disease. This suggests that like other diseases, such as cancer, HIV/AIDS and cardiovascular diseases, the best opportunity to modify the course of *AD* is before extensive and permanent damage has occurred (Sperling et al., 2014). The ability to predict the moment of the disease's progression opens up a new world of finding the right window for intervention. It could be possible that no *AD*-modifying therapy has been found, because the moment of progression has not been predicted yet. The accurate prediction of the moment of progression to *AD* could be a key aspect of this disease's treatment.

1.4 Practical Relevance

Alzheimer's disease affects the ageing population. The risk of getting *AD* increases as one's age increases: worldwide, the percentage of people who have *AD* is 10% in over 65 years, 20% in over 80 years, and over 40% of people in over 90 years (Alzheimer's Association, 2016). It is estimated that worldwide, more than 46 million people are suffering from *AD*, and this number is likely to increase to 131.5 million by 2050, since the life expectancy increases (World Alzheimer's Report, 2015).

The findings could make an important contribution to lowering the disease burden for the subject and reducing healthcare costs. For the subject, knowing when he/she is likely to progress to *AD* might be hard to process. By contrast, depending on how far this moment is away, fitted psychological and psychosocial interventions can take place. It is known that *AD* is one of the largest burdens of diseases for the aging population, since the disease has the largest impact on the quality of life. The aforementioned interventions have the potential to improve cognitive function, delay institutionalization, reduce care strain and improve the quality of life. These interventions already take place, but researchers stated that their effectiveness depends on how far a subject is in the process of developing *AD* (Mueller, Weiner, Thal, Petersen, Jack, Jagust, Trojanowski, Toga & Beckett, 2005). A psychological intervention for someone who will progress over 4 years could be completely different from that for someone who will progress over 6 months.

Currently, the cost of *AD* is already one of the highest of health care. In 2014, the total cost was 5 billion, which represented 5% of the total healthcare cost, in the Netherlands. Finding disease-modifying therapies by predicting the moment of progression can reduce the cost of these therapies. It could be possible that lowering the dose of some medication in combination with memory training could be more effective for someone who will progress to *AD* in four years, and having a higher dose with no memory training could be more effective for someone who will progress in six months. Drug dosage and applying it in a more efficient and personal manner is a step toward more effective treatment and reducing its cost.

1.5 Outline

The remainder of the current study is structured as follows: Section 2 contains a review of work on *AD* prediction. Section 3 describes the experimental setup, the data set and methods in detail. Subsequently, Section 4 presents the results of the experiments. Algorithms are compared to one another based on their performances, and results of binary- and multiclass-classification problems are also presented. Section 5 provides a discussion and recommendations for further research. Finally, Section 6 presents the conclusions of the current study.

Section 2: Related work

This section presents related work regarding *AD* prediction. It starts with an explanation of the development of *AD* in Subsection 2.1. Thereafter, in Subsection 2.2, some relevant research regarding the current state of *AD* prediction is discussed. This will be followed by a discussion of classifiers that are used in the current study (Subsection 2.3). Subsequently, Subsections 2.4 and 2.5 discuss multiclass classification and imbalanced data, respectively.

2.1 Development of Alzheimer's disease

A large and growing body of literature has investigated the development of *AD*. It is important to know how this disease develops over time, to gain insight into its related abnormalities that take place in the brain. Awareness of these abnormalities is necessary for selecting features for predicting progression to *AD*.

As discussed in the introduction, *AD* is caused by the degeneration of brain cells. Prior research indicates that this degeneration is likely to start 20 to 30 years before an individual is diagnosed with *AD* (Sperling et al., 2011). This means that the *AD* pathology develops while the individual is still cognitively normal (Dubois et al., 2010). At some point in time, this degeneration damages the brain in such a manner that it results in cognitive impairment. This is when an individual is diagnosed with *MCI*.

The aforementioned can be seen in Figure 2.1.1, which illustrates that different biomarkers indicate increasing abnormalities in the stage of *CN* and *MCI*. A point of interest is the line that represents brain structures, as brain structures are a biomarker that is visible on MRI scans. When the subject is diagnosed with *MCI*, this biomarker abruptly appears to be more abnormal and still increases even when someone is diagnosed with dementia.

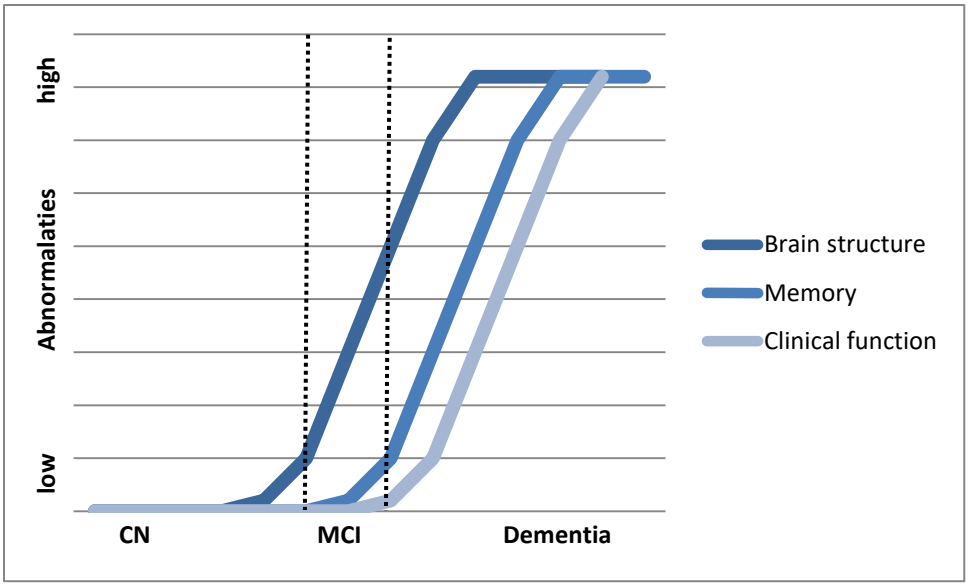


Figure 2.1.1. Dynamic biomarkers of the Alzheimer's disease development over time. On the x-axis, the clinical disease stage is represented, which starts with cognitively normal, followed by *MCI* and then dementia. On the y-axis, the biomarkers' magnitudes are illustrated from normal to abnormal. Since the current study is based on MRI data, the line representing brain structure is of importance because this biomarker is measured using MRI. This Figure demonstrates that when a subject is diagnosed with *MCI*, the brain structure is increasingly abnormal, with a higher level of abnormality at the end of *MCI* in comparison with that at the beginning of *MCI* (Jack et al., 2010). This suggests that predict the moment of progression may be possible. Abbreviations: *MCI* = Mild Cognitive Impairment.

The aforementioned brain structure consists of partial brain structures. These partial brain structures allow the demonstration of abnormalities at different times (Jack et al., 2010). Figure 2.1.2 provides insight into the order of presenting abnormalities for different parts of the brain. It demonstrates that as the disease progresses, the medial temporal lobe first changes. The medial temporal lobe includes the hippocampus, along with the surrounding regions that consist of parts of the ventricles, fusiform and entorhinal regions. The medial temporal lobe is known for its functions in the long-term memory. This explains why memory loss is often the first symptom of *AD* (Burns, Page & Winter, 2005).

Shortly after changes in the medial temporal lobe, the lateral temporal lobe will be affected. The lateral temporal lobe is located in the outer part of the brain and plays an important role in hearing, verbal and language functions as well as visual recognition. Examples of brain structures of which the lateral temporal lobe is composed are the fusiform, ventricles and entorhinal regions.

After this, the next abnormalities appear in the frontal lobe. The frontal lobe is located at the front of the brain and involves movement, reasoning, planning, certain speech functions, and problem solving. Part of the ventricles belongs to the frontal lobe.

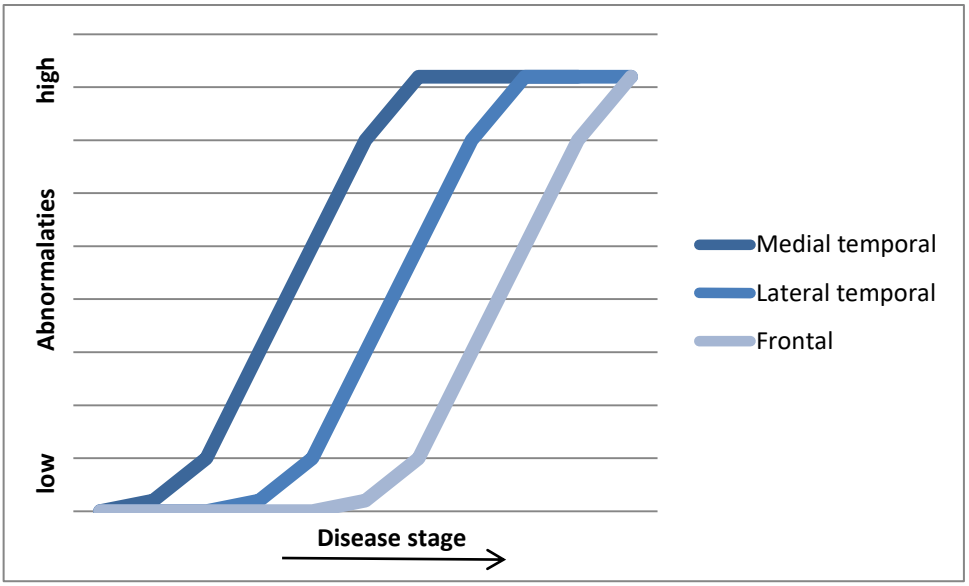


Figure 2.1.2. Dynamic biomarkers of the Alzheimer's disease stage. On the x-axis, the clinical disease stage is represented. On the y-axis, the biomarkers' magnitudes are illustrated from normal to abnormal. This Figure demonstrates that abnormalities appear in the following order: medial temporal, lateral temporal, and frontal lobe (Jack et al., 2010).

Taken together, Figure 2.1.1 indicates that abnormal changes take place way before the disease is physically evident. These abnormalities make *AD* prediction suitable for learning algorithms. Because the medial temporal lobe illustrates the first abnormalities according to Figure 2.1.2, including the structures that belong to this lobe would make sense in the input object for machine learning, such as hippocampus, ventricles, the fusiform and entorhinal regions.

2.2 The current state of Alzheimer's disease prediction

As mentioned in the introduction, only recently, the classification problem targeted by researchers has shifted from mainly distinguishing *AD* subjects from controls to the more challenging and more clinical applicable classification problem of distinguishing *MCI* converters from non-converters (Weiner et al., 2015). Not surprisingly, distinguishing *MCI* converters from non-converters is a more complex task in comparison with distinguishing *AD* subjects from controls, because the differences between the two classes are smaller.

This is illustrated by the accuracies of the different classifying tasks. Prior studies indicated that the best classifiers reached accuracies in the mid-90% range for distinguishing between *AD* subjects and control subjects. The best classifier for classifying *MCI* converters and non-converters achieved accuracies in the low 80% (for review, see Weiner et al. (2015)). These best accuracies are achieved when cognitive measures, genetic, CSF, MRI and PET biomarkers are combined (Weiner et al., 2015).

Huang, Yang, Feng and Cheng (2017) demonstrated that 79% accuracy can be achieved when one uses only MRI biomarkers as the input object when distinguishing *MCI* converters from *MCI* non-converters. The authors did not distinguish in when the conversion from *MCI* to *AD* takes place.

When a specific moment in time is selected for the progression to *AD*, accuracies are slightly lower. As an example, Trezpacs, Sun, Schuh, Case and Witte (2014) reported that logistic regression based on MRI biomarkers achieved an accuracy of 67% for classifying *MCI* converters and *MCI* non converters with the moment of progression 24 months in the future. Chupin et al. (2009) achieved a higher accuracy of 71% for classifying *MCI* controls and *MCI* converters, but their *MCI* converters would convert in 18 months instead of 24. Another example is the study conducted by Wolz et al. (2010). The authors achieved a lower accuracy (64%) of distinguishing *MCI* controls from *MCI* subjects who would convert to *AD* in 12 months.

One would expect that a learning algorithm would perform better when the moment of progression was closer. This is because, as can be seen in Figure 2.1.2, the brain is more distinctive and thus makes it easier to find boundaries and patterns for the learning algorithm (Jack et al., 2010). These results suggest otherwise. However, because these researchers used different learning algorithms, a sufficient comparison cannot be made.

One of the first studies that thoroughly examined the prediction of progression from *MCI* to *AD* through multiple future time points using the same learning algorithm was conducted by Westman, Muehlboeck, and Simmons (2012). Their goal was to further investigate how well the model that could distinguish between *AD* subjects and controls could also predict the conversion at different time points (12, 18, 24 and 36 months). The approach was the following: based on different time points after the first visit, namely after 12, 18, 24 and 36 months, the class to which the subjects were assigned is determined based on those time points: *stableMCI* or *progressionMCI*. For example, at the time point of 36 months, subjects who were assigned to *MCI* converters consist of subjects who converted at 12, 18, 24 and 36 months. The input object for the classification task between *MCI* converters and non-converters was MRI scans of the subjects that were made on the objects' first visits. The Support-Vector classifier achieved an accuracy of 59%, 66%, 66% and 66% for 12, 18, 24 and 36 months, respectively.

There are two key limitations of this approach. Even though the researchers examined the difference in performance on different time points, they did not separate the moment of progression for subjects assigned to *progressionMCI*. As a result, when the future time point is further away in time, there is more variety in the moment subjects progressed in *progressionMCI*. This means that for the time point of 36 months, *progressionMCI* consists of all subjects who progressed to *AD* in 12, 18, 24 and 36 months after the first visit. This is because they are all assigned to the same group. Only for the time point of 12 months is it certain that the subject assigned to *progressionMCI* progressed within

12 months. The consequences of this drawback of their approach is that no conclusions can be made about the reason why the model performed better, worse or the same in comparison with other time points.

Another limitation is that there is no distinction made between subjects who remained stable during the study and subjects who remained partly stable and progressed to *AD* later on in the study. For example, some subjects who are predicted as *MCI* non-converters at 12 months after the first visit are likely to convert to *AD* at time point 18, 24 or 36 months after the first visit. The problem with this is that subjects who progressed later in the study could already demonstrate an abnormal pattern at baseline, since abnormalities take place inside the brain long before a subject is diagnosed with *AD*, as discussed in Section 2.1. Despite this, these subjects are assigned to *stableMCI*, whereas they will progress to *AD* later on in the study; thus, they are not so stable after all. As a result, this could have, especially, a negative effect on the subset based on the future time point of 12 months. This is because *stableMCI* consists of subjects who converted at the point of 18, 24 and 36 months. It is likely that the difference of performance for the moment of 12 months after the first visit (59% in comparison with 66% for the other time points) is due to the impurity of the *stableMCI* class.

The study of Westman, Muehlboeck, and Simmons (2012) would have been more interesting if these authors had adopted another approach for assigning subjects to the *stableMCI* and *progressionMCI* classes.

The general trend based on all the studies reviewed here is that the accuracies of classification based on MRI scans are comparable with the accuracies achieved when using the best feature combinations. When one selects a precise moment of progression, the accuracies are somewhat lower. Very few publications can be found on examining the prediction of progression from *MCI* to *AD* through multiple moments of progression using the same learning algorithm. However, the approach taken by Westman, Muehlboeck, and Simmons (2012) has its limitations, which are caused by the impurity of the *stableMCI* class.

2.3 Classifiers

Plentiful learning algorithms have been adopted for *AD* classification. This is in line with the observations of Caruana and Niculescu-Mizil (2006), who stated that there is no learning algorithm that performs best on all tasks. Some models with a high average performance have a poor performance on some problems, whereas some models with a low average performance have a high performance on other problems. Thus, it is common practice to try out different machine-learning algorithms to find the algorithm that suits a particular task best. Hence, six learning algorithms that have proved to work for either classifying *AD* from controls or *MCI* non-converters from *MCI* converters are reviewed. The algorithms are as follows: Decision-Tree classifier (DTC), linear

Support-Vector classifier (LSVC), Logistic Regression (LR), Perceptron (PER), Stochastic Gradient-Descent classifier (SGD) and Support-Vector classifier (SVC).

In the remaining part of this subsection, these six classifiers are discussed, together with their implementation according to the literature on *AD* prediction.

The first classifier to be reviewed is the DTC. The decision tree is used in prior research for distinguishing *AD* subjects from *CN* subjects based on CSF biomarkers (Bombois et al., 2013). A decision tree is built in a top-down manner, with the aim of finding an attribute through which to split the classes at each stage and then recursively processing the subproblems that result from the split. The best split represents a separation of two classes that is as pure as possible; in this case, purity means that instances are all from the same class (Murty & Raghava, 2016). This generates a decision tree. One of the strengths of the decision tree is that it can handle both continuous and categorical data; this is useful for predicting *AD*. A weakness of the decision tree is that it can only search for decision boundaries that are parallel to the axis. This means that when the decision boundaries are not parallel, the decision tree performs poorly.

The LSVC is the second classifier to be reviewed. It was used in prior research for separating *AD* subjects from no-*AD* subjects (Martínez-Murcia, Ortiz, Górriz, Ramírez, & Illán, 2015). This classifier fits the data that are provided and returns a best-fit hyperplane that makes a distinction in the data. The LSVC is useful in many cases and has a high average performance. A disadvantage of the LSVC is that it only performs well when the data are linearly separable.

The third classifier to be reviewed is LR. Logistic Regression is used in prior research for distinguishing *MCI* converters from *MCI* non-converters based on metabolomic data (Orešič et al., 2011), demographic and genetic information, baseline cognitive scores, lab tests and MRI data (Li, Liu, Gong & Zhang, 2014). Logistic Regression gives an output that represents the probability that a certain input belongs to a certain class. Advantages of LR are that it is inherently simple, it has a low variance, and it is less prone to over-fitting in comparison with other classifiers such as the decision tree. Disadvantages are that all variables need to be relevant for prediction, whereas other classifiers can make a distinction themselves about which variables are informative or not. Therefore, subsection 3.1.3 verifies whether this is the case, to be able to apply logistic regression.

The fourth classifier to be discussed is the PER. It is used in prior research for distinguishing *AD* subjects from controls based on different cognitive tests, physical examinations, age, neuropsychiatric assessments, mental examinations and laboratory investigations (Joshi, Simha, Shenoy, Venugopal & Patnaik, 2010). The PER is the simplest form of a neural network, which has two layers: the input and

the output layers. The PER learns how to transform input into a desired outcome; therefore, it is most used for classification tasks. An advantage is that the PER can be used for complex problems, which are for a part of multiclass classification. A weakness is that it can only deal with linearly separable data.

The fifth classifier to be examined is the SGD. The SGD is used in prior research for distinguishing *AD* subjects from controls (Sarraf, Anderson & Tofighi, 2016). This classifier updates a particular set of parameters in such a manner that the error function is minimized. Based on only one training sample at a time, the parameters are updated. An advantage of the SGD is its efficiency to handle large amounts of data. One disadvantage is that SGD is sensitive to feature scaling. Therefore, it is highly recommended to scale the data. This is considered in selecting the best pre-processing method.

The last classifier to be discussed is the SVC. The SVC is used in prior research to distinguish *MCI* converters from *MCI* non-converters based on MRI (Westman, Muehlboeck, & Simmons, 2012; Huang, Yang, Feng, & Chen, 2017). The SVC tries to find a hyperplane that separates two classes to a maximum extent. It is similar to the LSVC, except for the fact that data for the SVC do not have to be linearly separable. This is one of the classifier's advantages over others. One other strength is that it performs well in a high-dimension feature space. One of the weaknesses is that it is easy to overfit, and one method of minimizing overfitting is using cross-validation, which is used in the current study and explained in Subsection 3.6.

2.4 Multiclass classification

So far, no research has been found regarding multiclass classification of *AD* progression. All studies regarding *AD* prediction have been binary-classification tasks, either classifying *AD* from controls or *MCI* converters from *MCI* non-converters. A binary-classification task means that the model predicts whether a subject is in class A or B. When there are more than two classes, the prediction is called a multiclass-classification task. A multiclass-classification task is more complex than a binary one because the classifier has to learn constructing a larger number of separation boundaries or relations (Hsu & Lin, 2002). In general, the classification error rate is higher in multiclass problems compared to binary problems, as there can be an error in any of the decision boundaries or relations (Bala & Agrawal, 2010).

There are two types of multiclass-classification algorithms: algorithms that deal directly with multiple classes and algorithms that divide a multiclass problem into sets of binary problems and then combine them. The decision tree is an example of an algorithm that deals directly with multiple

classes. All other classifiers decompose this problem into sets of binary ones. Most learning algorithms include a parameter on whose basis the method can be chosen.

2.5 Imbalanced Data

In recent years, there has been an increasing amount of literature published on class-imbalance classification (Ali, Shamsuddin & Relescu, 2015). Class imbalance occurs when one class is significantly larger than another class. This can be observed in various domains, including medical diagnosis. For a medical diagnosis, such as predicting *AD*, the class of interest is often underrepresented. However, recognizing this class is important because errors in diagnostics bring further complications to the subjects' treatment.

The problem with imbalanced data is that most classification algorithms assume that the training set is equally distributed. Imbalanced data hinder a classification task, thus resulting in lower performance. This is due to the lack of data of the minority class. This small sample size leads to difficulties in discovering patterns within these data. In an analysis of the effect of sample size on error rate, Japkowicz and Stephen (2002) indicated that when training sample size increases, the error rate of the imbalanced-class classification reduces. This is because the classifier is able to build better patterns for classes since there is more information available. To determine the effect of the degree of imbalanced-data distribution on performance, Sun, Wong and Kamel (2009) compared various degrees, but the effect is not yet explicitly known.

Besides that imbalanced data decrease the performance of the classifier, there is another problem with imbalanced data. When classes are highly imbalanced, accuracy is not a sufficient evaluation method. The explanation is that when the minority class is underrepresented by the data, this minority class will be ignored by the classifier, and the model can still achieve a high accuracy. For example, when the majority class consists of 95% of the data, an accuracy of 95% can be achieved by only classifying each example as the majority class. This is called the accuracy paradox.

To overcome the problem of imbalanced data, there are two methods of making the training data balanced. One method is applying undersampling on the training data, which means that examples from the majority class are removed until there are equal numbers of examples in both classes. Removing data has the downside of losing potentially important information about the class.

The other method is called oversampling. This means that examples from the minority class get duplicated. One oversampling technique is the Synthetic Minority Oversampling Technique (SMOTE), and it is becoming more and more popular with imbalanced-class classification problems (Borowska & Topczewska, 2016). The manner in which SMOTE works is as follows: it adds new examples to the minority class, by computing a probability distribution to model the minority class and performing certain operations on the original data. This increases new examples, but important to note

is that it does not provide new information about the class. Besides that, in highly unbalanced data sets, too much oversampling may result in overfitting (Agrawal et al., 2015). It is important to note that SMOTE is not applied to the test set because that would synthesize the test set and thus make the prediction worthless (Borowska & Topczewska, 2016).

Prior research investigated how to improve the performance of a learning algorithm on imbalanced binary-class data sets, in which there was one majority class and one minority class. The researchers suggested using a combination of oversampling and undersampling to reach the best performance of a classifier (Kotsiantis, Kanellopoulos & Pintelas, 2006; Chawla, 2005; Weiss, McCarthy & Zabar, 2007). Therefore, this suggestion is followed for the binary-classification problem in the current study.

While imbalanced data are a problem for binary classification, they are even more problematic in multiclass classification (Singh & Ade, 2015). There are multiple ways in which class imbalance can appear in multiclass data. One common way is that there is a so-called 'super majority' class that contains most of the instances in the data set. Another possibility is that there is a class that is significantly small in comparison with other classes (Hoens, Qian, Chawla & Zhou, 2012).

One issue arises for multiclass data sets: most existing solutions for oversampling are applicable to binary-class problems only. These solutions, such as SMOTE, cannot be applied directly to a multiclass imbalanced data set (Sun, Kamel & Wang, 2006). To date, not many methods have been developed and introduced to deal with imbalanced data for multiclass classification (Agrawal et al., 2015).

To address this problem, Agrawal et al. (2015) proposed an algorithm called SCUT (SMOTE and Clustered Undersampling Technique). The basis of this technique is the following: the average number of examples of all classes is calculated. Oversampling is applied to classes that contain fewer examples than the average. This will be done through a one-versus-all method (OVA). Undersampling is applied to classes that contain more examples than the average. Oversampling and undersampling are applied to all classes in such a way that all classes contain as many examples as the average number of examples.

Even though this algorithm is not widely adopted, Agrawal et al. (2015) indicated that SCUT provides a sufficient remedy for the multiclass problem; therefore, this suggestion is followed for the multiclass-classification problem in the current study.

Section 3: Method

This section provides a description of the methods used in the current study. It starts with a description of the data in Subsection 3.1. This includes an explanation of the manner in which the subsets are made for each research question in the section, a description of the features used for making these subsets, and features used for object information on the learning algorithms. Thereafter, an explanation of how this study dealt with missing values and imbalanced data is provided in Subsections 3.2 and 3.3, respectively. This is followed by the evaluation method (Section 3.4) and software used in the current study (Subsection 3.5). Finally, this section provides an extensive explanation of the procedures for the current study's experiment in Subsection 3.6.

3.1 Data Description

The data used for the current study are from the ADNImerge file obtained from the Alzheimer's disease Neuroimaging Initiative database (ADNI, 2017). The ADNImerge data set includes a total of 1.737 adults who were recruited from over 50 locations across the United States and Canada, with ages ranging from 55 to 92 years. These subjects were followed over time. The data is contained in 12.734 rows and 201 columns, of which each row represents one visit from each subject. On every visit, PET scans, MRI scans, cognitive test and other tests were conducted. In this file, all data from MRI scans were converted into MRI voxels.

At baseline, when each of them visited the study for the first time, the 1.737 subjects were diagnosed with the following: 417 subjects with *CN*, 106 subjects with Subjective Memory Concerns (*SMC*), 310 subjects with *EMCI*, 562 subjects with *LMCI* and 342 subjects with *AD*. As can be seen in Figure 3.1, not every subject visited the study at every planned follow-up visit. Only 46% of all subjects at the start visited the study after 3 months, but 93% showed up on the follow-up visit, i.e. 6 months after the first visit. This indicates that some subjects remained in the study, without being present at every follow-up meeting. Also, there was a significant drop after the 24th month as well as after the 36th and the 48th months. This indicates that over time, fewer data are available.

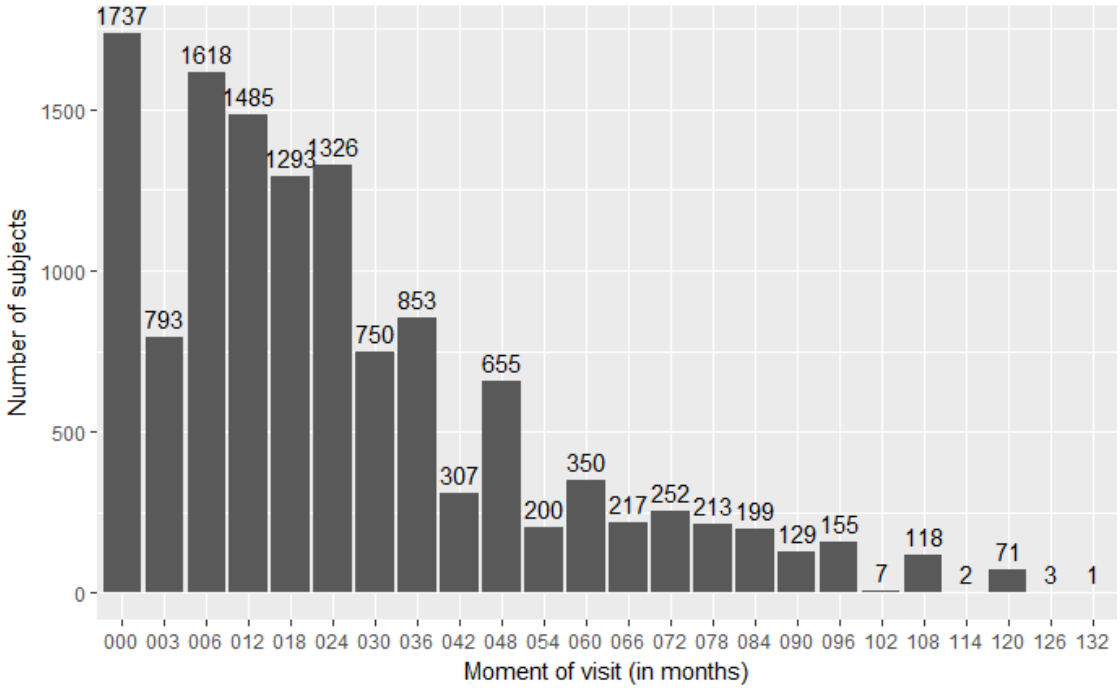


Figure 3.1. An overview of the number of subjects per visit. The x-axis represents the moment of visit, with 000 representing the first visit (baseline), 003 representing three months after the first visit, and so on. The y-axis illustrates a unique number of subjects on each visit, which is equal to the number of observations on that visit. A total of 12,734 visits were made by 1.737 subjects. This figure demonstrates that as the time continues, fewer data are available.

3.1.1 Subsets

To answer the research questions, different subsets are made from the original data. For RQ1, the goal was to find the best classifier that could make a distinction between *MCI* and *AD* at baseline. Hence, subjects who were diagnosed with *EMCI* or *LMCI* (together representing *MCI*) and *AD* on the first visit were selected. Table 3.1.1.1 gives insight into the frequencies and percentage of the target variable.

Table 3.1.1.1

Target variable RQ1: Frequency and percentage of MCI and AD diagnosis at baseline

Category	Description	Frequency	Percentage
1	*MCI*	872	72
2	*AD*	342	28
Total		1214	100

Frequency stands for the number of subjects in a class. In comparison with the total number of subjects, 1214, a percentage is made to gain insight into the ratio between the two classes. Abbreviations: MCI = Mild Cognitive Impairment; AD = Alzheimer's disease.

The goal for RQ2 was to investigate the extent to which this classifier is able to make a binary distinction between *stableMCI* and *progressionMCI*, on different time intervals. First, two groups were made from the original data set: *stableMCI* and *progressionMCI*.

The criteria for selecting subjects for *stableMCI* were as follows: the subjects were diagnosed with *EMCI* or *LMCI* on the first visit, and on every follow-up visit, they received no other diagnosis except *EMCI* or *LMCI*. Thus, each subject assigned to *stableMCI* remained truly stable during the course of the study.

The criteria for selecting subjects for *progressionMCI* were as follows: the subjects were diagnosed with *EMCI* or *LMCI* on the first visit, and at a certain point in time, they progressed to *AD* and were diagnosed with *AD* on every follow-up meeting. Therefore, 25 subjects were excluded from the study because they were diagnosed with *MCI* after an *AD* diagnosis. See Table 3.1.1.2 for the frequencies and percentages of these two groups the subset for RQ2.

Table 3.1.1.2

Target variables RQ2 and RQ3: Frequency and percentage of stableMCI and progressionMCI at baseline

Category	Description	Frequency	Percentage
1	*stableMCI*	484	61
2	*progressionMCI*	303	39
Total		787	100

Frequency stands for the number of subjects in a class. In comparison with the total number of subjects, 787, a percentage is made to gain insight into the ratio between the two classes. Abbreviations: stableMCI = subjects who remained stable; progressionMCI = subject who progressed to Alzheimer's disease.

After this, the *progressionMCI* group was further divided. It was first divided according to the moment in time the subject made the progression to *AD*. Based on that moment of progression, five subsets were made, containing the subject's data 6, 12, 24, 30 and 48 months before the subject made the progression to *AD*. Table 3.1.1.3 presents the number of subjects at different moments in time. Important to note is that for categories 2 to 6, the number of subjects is equal to the number of observations, since these categories represent a particular moment in time. However, for category 1, this is not the case since this group is stable and does not convert to *AD*, and it should not matter whether observations are selected in the beginning, middle, or end of the study. Therefore, no particular moment in time is chosen. The *MCI* stable group consists of 484 subjects with a total of 2.469 observations that are divided among different time points. To prevent bias towards a particular subject, because multiple observations of the same subject are compared against the other class in

which every subject is unique, only one observation per subject is used. This is selected randomly, and this selection for *stableMCI* is used for all comparisons.

As a result, five subsets were made for RQ2, namely *stableMCI* and *progression6MCI*, *stableMCI* and *progression12MCI*, *stableMCI* and *progression24MCI*, *stableMCI* and *progression30MCI* as well as *stableMCI* and *progression48MCI*.

Table 3.1.1.3

Target variables RQ2 and RQ3: Frequency and percentage of stableMCI, progression6MCI, progression12MCI, progression24MCI, progression30MCI and progression48MCI

Category	Description	Frequency	Percentage
1	*stableMCI*	484	41
2	*Progression6MCI*	180	15
3	*Progression12MCI*	205	17
4	*Progression24MCI*	173	15
5	*Progression30MCI*	79	7
6	*Progression48MCI*	52	4
Total		1173	100

Frequency stands for the number of subjects who are available for each class separately. In comparison with the total, 787, a percentage is made to gain insight into the ratio. Important to note is that for categories 2 to 6, the frequency is equal to the number of observations, whereas for category 1, this is not the case.

For RQ3, the goal was to investigate the extent to which an optimized classifier that is able to predict whether and when a subject will progress to *AD*. The same classes from RQ2 were used. The difference is that instead of a binary classification, all classes were merged together, thereby changing the classification task to a multiclass classification.

3.1.2 Variables

To make these subsets, four variables were necessary: research ID (RID), visit (VISCODE), diagnosis at first visit (DX.bl) and diagnosis on follow-up visit. The variable DX.bl is a categorical value that represents the baseline diagnosis. This variable was needed for all subsets. The feature DX is also a categorical value that represents the diagnosis on a follow-up visit. Important to note is that the labels from DX differ from those from DX.bl. In DX.bl, there is a distinction between *EMCI* and *LMCI*, and in DX, this is not the case. Therefore, when subsets are made, *EMCI* and *LMCI* are taken together. DX also has a label: "*MCI* to Dementia". This label represents the moment of progression and is used for subsets for RQ2 and RQ3.

Research ID is unique for every subject and it makes it possible to track a subject over time, by combining different rows that belong to the same *RID*. *VISCODE* stands for the moment subjects visit the ADNI study. For example, m03 stands for 3 months after the first visit, and m12 stands for 12 months after the first visit. This variable is represented by a string, but to make it possible to track a subject over time by *VISCODE* and *RID*, *VISCODE* is transformed to a factor in R.

3.1.3 Feature Selection

From the remaining 197 features, features regarding the volume of brain structures were selected. It is important that these brain structures are located in the medial temporal lobe since this lobe indicates the first abnormalities. As a result, the following six features were selected: Hippocampus, Ventricles, WholeBrain, Entorhinal Cortex, the Fusiform and MidTemp.

A correlation matrix is made from these features (Table 3.1.3.1) to see whether there are any highly correlated features. Generally, when the correlation between two features is higher than or equal to 0.75, those features should be removed. However, the correlation between MidTemp and WholeBrain is 0.76. This can be explained because the WholeBrain exists partially of MidTemp.

Table 3.1.3.1

Pearson correlation matrix of selected features

	Hippoc.	Ven.	WholeB.	Ent.	Fusi.	MidT.
Hippoc.	1					
Ven.	-0.25	1				
WholeB.	0.59	0.11	1			
Ent.	0.68	-0.14	0.51	1		
Fusi.	0.53	-0.06	0.73	0.56	1	
MidT.	0.58	-0.01	**0.76**	0.51	0.71	1

All correlations higher than 0.75 are printed in bold. Abbreviations: Hippoc. = Hippocampus, Ven. = Ventricles, WholeB. = WholeBrain, Ent. = Enthornihal Cortex, Fusi. = Fusiform, MidT. = MidTemp.

To use LR, all features have to be informative as discussed in subection 2.3. To test whether MidTemp and/or WholeBrain has to be excluded, recursive feature elimination with cross-validation is executed. This method makes a selection of the best features. When it turns out that this method does not include MidTemp and/or WholeBrain in the selection of best features, these/this feature(s) is left out. All six features are tested, and all of them are ranked as a valuable feature for predicting progression. Therefore, MidTemp and WholeBrain are used in the current study regardless of their correlation being 0.76. All features are discussed in subsections 3.1.2.1 to 3.1.2.6. See Table 3.1.3.2 for an

overview of the features that are relevant and used as a starting point for different subsets.

Table 3.1.3.2

An overview of selected features

Feature	Range	Type and measure	Description
RID		Label	Participant ID (unique for every subject)
VISCODE	(bl, m03, m06, m12, m18, …, m132)	Label	Visit code
DX.bl	*CN*, SMC, E*MCI*, L*MCI*, *AD*	Categorical	Baseline Diagnosis
DX	NL, *MCI*, Dementia, NL to *MCI*, *MCI* to Dementia, NL to Dementia, Dementia to *MCI*, Dementia to NL	Categorical	Diagnosis on visit
Hippocampus	2.219–11.207 mm^3	Continuous	Hippocampus
Ventricles	5.650–162.729 mm^3	Continuous	Ventricles
WholeBrain	649.091–1.486.036 mm^3	Continuous	WholeBrain
Entorhinal	1.041–6.711 mm^3	Continuous	Entorhinal Cortex
Fusiform	7.739–29.950 mm^3	Continuous	Fusiform
MidTemp	8.044–32.189 mm^3	Continuous	MidTemp

Abbreviations DX.bl: CN =Clinical Normal, SMC = Significant Memory Concern, MCI =Mild Cognitive Impairment, EMCI= Early MCI, LMCI = Late MCI, AD = Alzheimer's disease,
Abbreviations DX: NL = Normal Controls, MCI =Mild Cognitive Impairment
CN and NL are similar; there is no difference in criterion. SMC subjects indicated that they had concerns about their memory, but their scores are within the normal range for cognition. NL to MCI, MCI to Dementia, NL to Dementia, Dementia to MCI and Dementia to NL indicate moments of progression.

3.1.3.1 Hippocampus

The hippocampus feature gives insight into the volume in voxels of the hippocampus of the subject. The hippocampus is important for storing information in memory, orientation in space and controlling behavior that is important for survival. A damaged hippocampus can lead to a reduced ability to store

new information in memory.

As can be seen in Table 3.1.3.2, hippocampus is a continuous variable that ranges from 2.219 to 11.207 mm³. In the following figures, every black line represents a unique subject within a group over time. The regression line in blue denotes the average course over time of all subjects together. Comparing A with B in Figure 3.1.3.1, where A represents subjects within *stableMCI* and B represents all subjects within *progressionMCI*, the regression line in blue indicates that the volume of the hippocampus of subjects in *progressionMCI* shrinks, whereas the volume of the hippocampus of subjects in *stableMCI* remains the same.

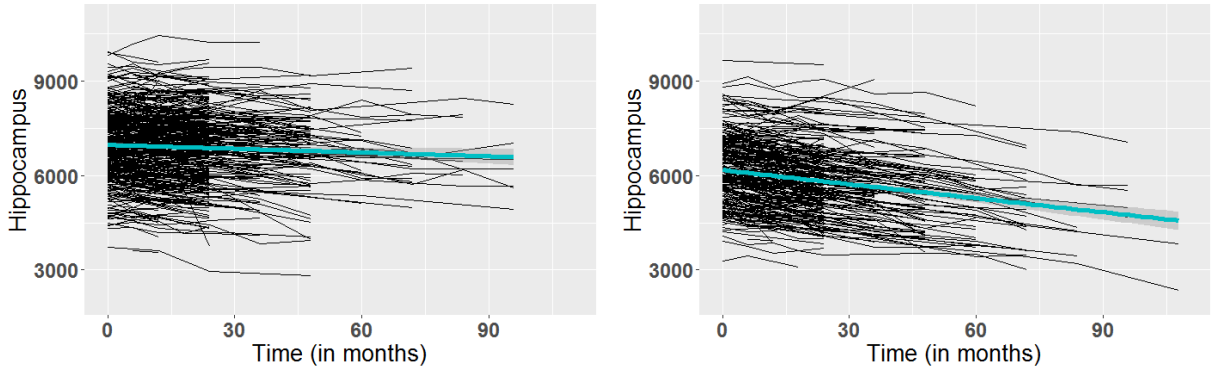


Figure 3.1.3.1. The volume of hippocampus over time for subjects in *stableMCI* (A) and that for subjects in *progressionMCI* (B). In both graphs, the x-axis represents the time in months, and the y-axis represents the volume of the hippocampus in voxels. All black lines represent a unique subject within the class. The blue regression line illustrates the average course over time of all subjects together within a class. This Figure demonstrates that the volume of the hippocampus shrinks over time for *progressionMCI*, whereas it remains the same for *stableMCI*.

3.1.3.2 Ventricles

The Ventricles feature represents the volume in voxels of the Ventricles. Inside the brain, the Ventricles are four interconnected cavities, filled with cerebrospinal fluid (CSF). The Ventricles are also connected with the subarachnoid space, which is the space between the inner and middle meninges. The CSF that flows through the ventricles cleans the brain and also helps the brain with maintaining the right temperature.

This feature is also a continuous variable that ranges from 5.650 to 162.719 mm³, according to Table 3.1.3.2. Figure 3.1.3.2 indicates that the volume of the Ventricles increases over time for *progressionMCI* (B) steeper in comparison with *stableMCI* (A), which also exhibits a subtle increase of the volume of the Ventricles over time. This can be explained as follows: when the brain shrinks, the space within the cavities increases.

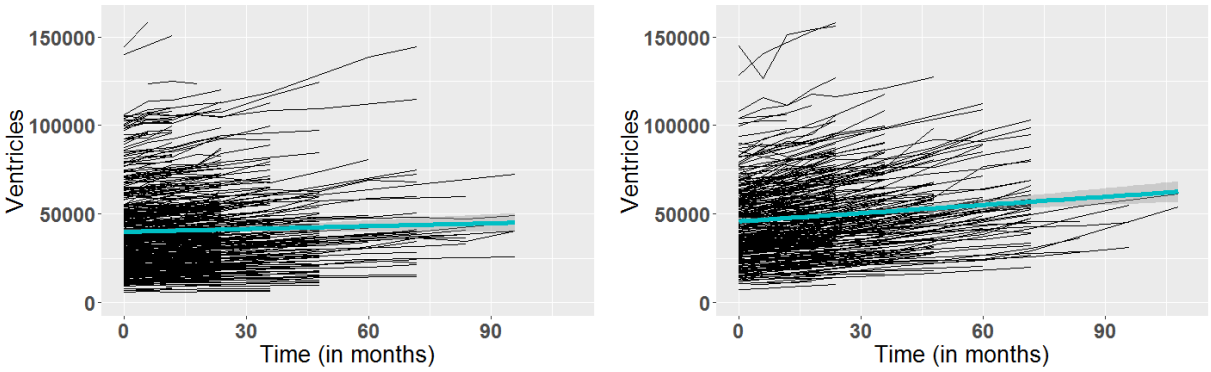


Figure 3.1.3.2. The volume of Ventricles over time for subjects in *stableMCI* (A) and that for subjects in *progressionMCI* (B). In both graphs, the x-axis represents the time in months, and the y-axis represents the volume of the hippocampus in voxels. All black lines represent a unique subject within a class. The blue regression line illustrates the average course over time of all subjects together within that class. This Figure demonstrates that the volume of the ventricles increases more for a subject in *progressionMCI* than for a subject in *stableMCI*

3.1.3.3 WholeBrain

The WholeBrain feature indicates the volume in voxels of a subject's whole brain. The volume of the whole brain ranges from 649.091 to 1.486.036 mm³ (Table 3.1.3.2), and, as expected, the volume of the whole brain shrinks for subjects in *progressionMCI*. The volume of the whole brain in subjects in *stableMCI* also illustrates a subtle decline over time, as can be seen in Figure 3.1.3.3.

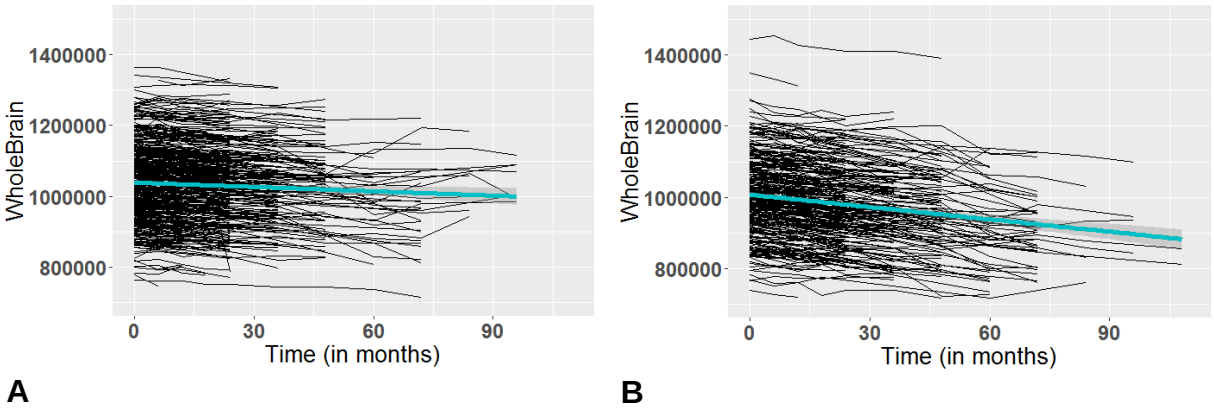


Figure 3.1.3.3. The volume of WholeBrain over time for subjects in *stableMCI* (A) and that for subjects in *progressionMCI* (B). In both graphs, the x-axis represents the time in months, and the y-axis represents the volume of the hippocampus in voxels. All black lines represent a unique subject within a class. The blue regression line denotes the average course over time of all subjects together within that class. This Figure demonstrates that the volume of the ventricles shrinks more for a

subject in *progressionMCI* than for a subject in *stableMCI*

3.1.3.4 Entorhinal Cortex

The Entorhinal Cortex feature indicates the volume in voxels of the Entorhinal cortex of the subject. The Entorhinal Cortex connects the temporal bark through the subiculum with circuits within the hippocampus. The Entorhinal Cortex is important for coding and storage of long-term information.

The range of this feature is from 1.041 to 6.711 mm³ (Table 3.1.3.2). Figure 3.1.3.4 indicates that there is a decline of the volume of Entorhinal Cortex in subjects in *progressionMCI* over time, whereas the volume remains the same for subjects in *stableMCI*.

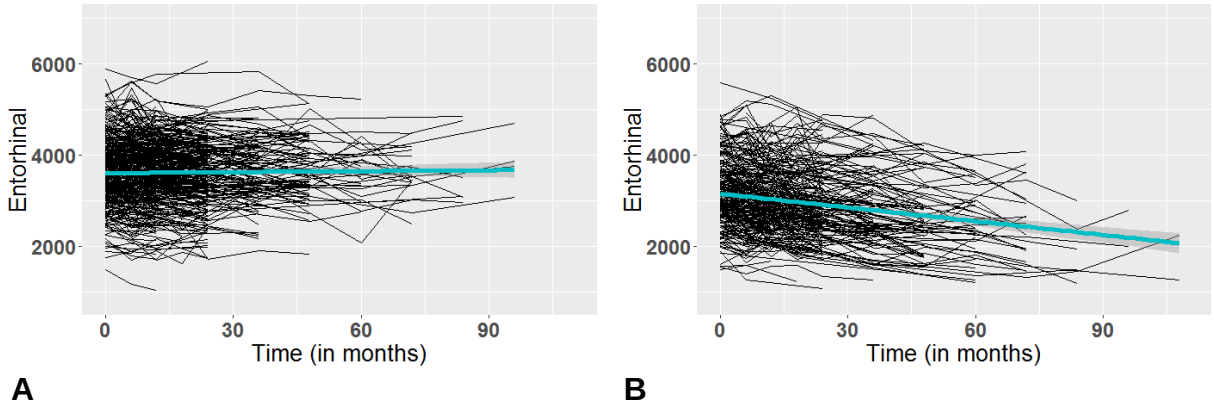


Figure 3.1.3.4. The volume of Entorhinal Cortex over time for subjects in *stableMCI* (A) and that for subjects in *progressionMCI* (B). In both graphs, the x-axis represents the time in months, and the y-axis represents the volume of the hippocampus in voxels. All black lines represent a unique subject within a class. The blue regression line denotes the average course over time of all subjects together within that class. This Figure demonstrates that the volume of the ventricles decreases for a subject in *progressionMCI*, whereas it seems to remain the same for a subject in *stableMCI*.

3.1.3.5 The Fusiform

The Fusiform feature indicates the volume in voxels of a subject's fusiform. The function of the Fusiform gyrus is not fully understood, but it has been linked with various neural pathways related to recognition. Additionally, it has been linked to various neurological phenomena such as synesthesia, dyslexia and prosopagnosia.

The range of this feature is between 7.739 and 29.950 mm³ (Table 3.1.3.2). As can be seen in Figure 3.1.3.5, the volume of the Fusiform shrinks over time in subjects who are in *progressionMCI*

and seems to remain the same in those in *stableMCI*.

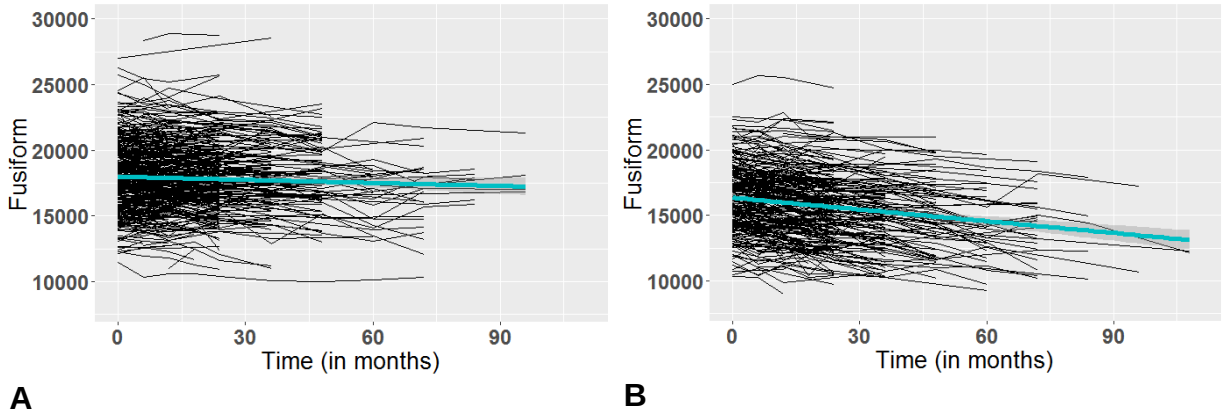


Figure 3.1.3.5. The volume of the Fusiform over time for subjects in *stableMCI* (A) and the volume of that for subjects in *progressionMCI* (B). In both graphs the x-axis represents the time in months, and the y-axis represents the volume of the hippocampus in voxels. All black lines represent a unique subject within a class. The blue regression line illustrates the average course over time of all subjects together within that class. This Figure demonstrates that the volume of the ventricles decreases for a subject in *progressionMCI*, whereas it seems to remain the same for a subject in *stableMCI*.

3.1.3.6 MidTemp

The MidTemp feature indicates the volume in voxels of the medial temporal lobe of a subject. Though the function of the Middle Temporal Gyrus is not known, it is linked with the recognition of known faces and accessing the meaning of words while one is reading.

The range of this feature's volume is from 8.044 to 32.189 mm³, according to Table 3.1.3.2. Figure 3.1.3.6 illustrates a decline in volume over time in subjects in *MCI* to *AD*, and the volume of the MidTemp in subjects in *MCI* stable indicates no decline.

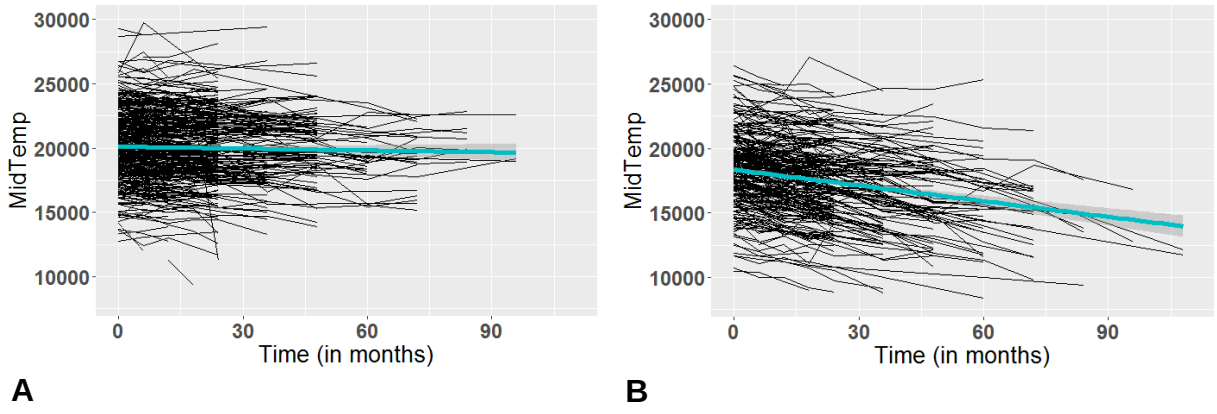


Figure 3.1.3.6. The volume of MidTemp over time for subjects in *stableMCI* (A) and that for subjects

in *progressionMCI* (B). In both graphs, the x-axis represents the time in months, and the y-axis represents the volume of the hippocampus in voxels. All black lines represent a unique subject within a class. The blue regression line denotes the average course over time of all subjects together within that class. This Figure demonstrates that the volume of the ventricles decreases for a subject in *progressionMCI,* whereas it seems to remain the same for a subject in *stableMCI*.

3.2 Missing Values

An important step in data processing is data cleaning. The original data set is also used for other studies, and all clinical data are cleaned (ADNI Protocol, 2017). However, the number of missing values is high. This is due to several reasons, such as high measurement cost (PET scans), poor data quality and unwillingness of the patients to receive the invasive test (lumbar puncture) (Thung, Wee, Yap & Shen, 2015). As a result, all 12.734 rows from the original data set contain at least one missing value.

Most learning algorithms cannot deal with missing values (Billsus & Pazzani, 1998). Thung et al. (2015) suggested that there are two approaches to handling missing data: removing missing data and inputting missing data. In general, the data-inputting approach is more preferable as it enables using as many samples as possible for machine learning. A condition for this algorithm to recover a large proportion of missing values is that the missing data are distributed randomly and uniformly (Candès & Recht, 2012). However, in the ADNI data, this is not the case. Therefore, the list-wise deletion approach is used, which is also the default method for handling missing values (Allison, 2012).

For the data set used in the current study, this means that all rows with at least one missing value in the subset that is denoted as Not Available (*NA*) were removed completely, according to the list-wise deletion approach, to prepare the data for the learning algorithm. This is done after feature selection because when the list-wise deletion approach is applied before feature and subset selection, there will be no rows left. After applying this approach, the subset contains no missing values. For the process of creating a subset, see Figure 3.2.1.

Figure 3.2.1. The process of making a subset from a complete data set

For the subset made for RQ1, this means that 100 *AD* subjects and 182 *MCI* subjects are excluded,

which leaves us with 242 *AD* subjects and 690 *MCI* subjects at first visit. For RQ2 and RQ3, this means that for *stableMCI, progression6MCI, progression12MCI, progression24MCI, progression30MCI and progression48MCI* classes, 38, 22, 80, 0, 18 and 17 subjects are excluded, respectively. This leaves us with the following subjects: 446, 158, 125, 123, 61 and 35 for *stableMCI, progression6MCI, progression12MCI, progression24MCI, progression30MCI and progression48MCI* classes, respectively. For an overview, see Table 3.2.1.

Table 3.2.1

An overview of original, deleted and final subjects for the different classes

RQ	Class	Original subjects	Excluded subjects	Included subjects
RQ1	*MCI*	872	182	690
	AD	342	100	242
RQ2/RQ3	*stableMCI*	484	38	446
	Progression6MCI	180	41	139
	Progression12MCI	205	80	125
	Progression24MCI	173	50	123
	Progression30MCI	79	18	61
	Progression48MCI	52	17	35

Abbreviations: MCI = Mild Cognitive Impairment, AD = Alzheimer's disease

For every class, except stableMCI, the number of subjects is equal to the number of observations.

3.3 Imbalanced Data

For the binary-classification task, the imbalanced data problem is tackled as suggested by Kotsiantis, Kanellopoulos and Pintelas (2006), Chawla (2005) as well as Weiss, McCarthy and Zabar (2007). The authors stated that a combination of oversampling and undersampling achieved the best performance of a classifier, as discussed in subection 2.5. By considering the caution made by Agrawal et al. (2005) that too much oversampling may result in overfitting, no more than 100% of new instances are created.

For the multiclass-classification task, the imbalanced data problem is tackled using the SCUT algorithm suggested by Agrawal et al. (2005), as discussed in subection 2.5.

3.4 The Evaluation Method

To avoid the accuracy paradox, since there is large class imbalance, the classifiers used in the current study are evaluated by their F1 score (Chawla, 2005). The F1 score is a harmonious mean between precision and recall.

There is an important difference between precision and recall. Having high precision means, in the current study, that when a model predicts that a subject will convert into *AD*, the model is usually right about that. This is about how many predicted converters are actually converters. In the current study, having high recall means that a model is able to identify most of the converters out there. This is about how many converters the model was able to predict out of all the converters out there. For example, the model predicted that 20 subjects would convert. All these 20 subjects are actually converters, meaning that the precision is 100%. However, the model missed 500 other subjects who are also converters. In this case, the score for recall would be very low.

Ideally, the model should predict all converters out there while being careful not to predict that a subject will convert when it actually will not. This model should have high precision and high recall. Since it is more useful to have a single number to describe the performance, the mean of precision and recall is used, which is the F1 score. The F1 score is between 0 and 1: the closer the score is to 1, the better. See Table 3.4.1 for how the F1 score is calculated.

Table 3.4.1

Calculation of the F1 score

Concepts	Definitions and calculations
True Positives (TP):	the number of positive examples, labeled as such.
False Positives (FP):	the number of negative examples, labeled as positive.
True Negatives (TN):	the number of negative examples, labeled as such.
False Negatives (FN):	the number of positive examples, labeled as negative.
Precision	TP / (TP + FP)
Recall	TP / (TP + FN)
F1 score	2 × (precision × recall) / (precision + recall)

There are three different methods of calculating the F1 score: micro, macro and weighted. F1-micro score is computed using the global count of true positives and false negatives, so no distinction between classes is made. To calculate the F1-macro score, the average of F1 scores for each class is computed. To compute the F1 weight, the average of the F1 score for each class is computed, but the weight is attached to the F1 score by the support of a class: the more elements in the class, the more important the F1 score will be in computing the average for this class.

Taken together, the F1-micro score is a measure of effectiveness of the majority of classes in a test collection. To gain insight into the effectiveness of all classes, the F1-macro or F1-weighted score should be computed (Manning et al., 2008). Since the data used for the current study is imbalanced, and the majority class is not more important than the minority class, the F1-macro score is computed and evaluates the classifiers.

3.5 Software

All data pre-processing, including making subsets, selecting features, deleting missing values, and applying SMOTE, was done using the programming language R in Rstudio (version 1.0.136). The following R packages were used: Hmisc for loading the data set, Plyr and Dplyr for performing manipulations for the subsets, ggplot for making plots and the DMwR package for applying SMOTE. All steps that belong to machine learning were taken using the programming language Python in PyCharm (edition 2016.3.2). The following modules were used: numpy for converting the features into the right format; Train_test_split and cross_Val_score, StratifiedKfold from sklearn model selection for cross-validation; RFECV from sklearn feature selection for ranking the most informative features; StandardScaler, MinMaxScaler and PolynomialFeatures from sklearn preprocessing for applying pre-processing methods on the data; Accuracy_score, f1_score, recall_score, precision_score and consufsion_matrix from sklearn metrics for evaluating the performance of a classifier; Stats from scipy for conducting the mcnemar test; DummyClassifier from the sklearn dummy for comparing the performance of the classifier with that of a Dummy classifier; as well as SGDClassifier, Logistic Regression and Perceptron from the sklearn linear model, LinearSVC and SVC from sklearn svm, and DecisionTreeClassifier from the sklearn tree for the classifiers used in the current study.

3.6 Experimental Procedure

This section gives per research question a description of the experiments done in order to answer the question.

3.6.1 Experimental Procedure for Answering RQ1: Model Selection

This experimental procedure was set up to find the classifier that performs best in distinguishing *MCI* from *AD* at baseline and eventually reuse this learning algorithm throughout the current study.

Step 1: Splitting a subset into a training set and a test set

Prior research indicated that the most common splits for training, validation and test data are ratios such as 50/25/25, 60/20/20, 70/15/15 and 80/10/10 (Raykar & Saha, 2015). According to Tan and Wong (2017), most researchers have applied the ratio of 70/15/15. The current study follows the majority by also dividing the data at the ratio of 70/15/15.

The subset created for RQ1 in Subsection 3.1.1 consists of 932 subjects in total, which are divided into 242 *AD* subjects and 690 *MCI* subjects. The proportion of the two classes in the subset for RQ1 is 26:74 for *AD* to *MCI*, which is highly unbalanced. Since it is important that the test set contains the same ratio between *MCI* and *AD*, as it is in the original data, a fixed number of *AD* and *MCI* subjects

are randomly selected for the test set to ensure this ratio.

This subset is split into a training (85%) and a test set (15%). This means that the test set contains a total of 139 subjects (15% of total subset), with 103 *MCI* and 36 *AD subjects*, to keep the proportion at 26:74. This test set is set apart and is not used until step 55.

The subjects who are not selected for the test set are automatically in the training set. For steps 2 and 3, this training set is divided into training and validation sets to find the best pre-processing method and optimal parameters for each classifier.

Step 2: SMOTE on training set

As mentioned in Subsection 3.2, SMOTE is applied to the training set. The parameters for SMOTE are set to 100 for *perc.over*, to make the quantity of the minority class twice as much. The parameter *perc.under* is set to 200 to remove at most half of the majority class and make the data balanced. These settings of *perc.over* and *perc.under* are used in prior research (Hao, Wang & Bryant, 2014). The latter step is a form of undersampling.

Important to note is that each time SMOTE is applied, the synthesized data calculated for the minority class by SMOTE and the samples removed from the majority class are different (Mashayekhi & Gras, 2012). To discover whether this has any effect, SMOTE is applied five times to the training set. As a result, five different training sets are developed. After SMOTE is applied, the training set is split into training and validation data. Table 3.6.1.1 gives insight into the number of observations for training and the test set.

Table 3.6.1.1

An overview of subjects for RQ1 in the test set and the training set before and after SMOTE

	Test set		Training set (before SMOTE)		Training set (after SMOTE)	
	MCI	*AD*	*MCI*	*AD*	*MCI*	*AD*
Subjects	103	36	587	206	412	412

Abbreviations: MCI = Mild Cognitive Impairment, AD = Alzheimer's disease

Step 3: Selecting a pre-processing method for each classifier

The goal of this step is to prepare the data in such a manner that they best support the classifier. The classifiers that the current study investigates are as follows: DTC, LSVC, LR, PER, SGD and SVC. In this step, all parameters of the classifier are set to default, since the tuning is in step 4. The pre-processing strategies that are tested for all classifiers can be seen in Table 3.6.1.2.

Table 3.6.1.2

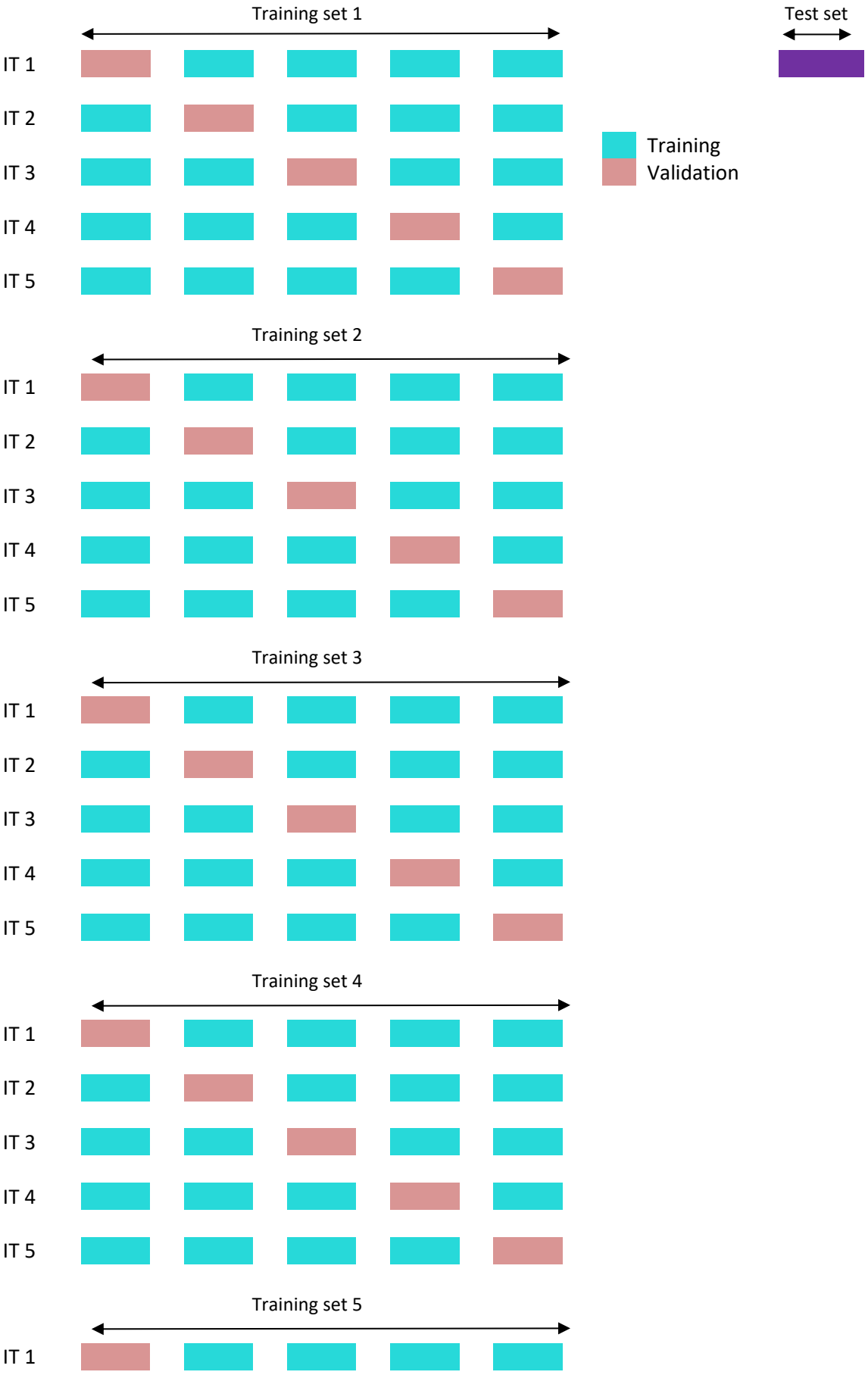
Pre-processing methods investigated by the current study

Number	Pre-processing method	Description	Used by
1.	No pre-processing	No transformation is performed	
2.	Z-score / Standardization	Transforming features by centering them, i.e. by removing the mean value of each feature.	
3.	Normalization	Scaling the features between 0 and 1	
4.	Adding feature interactions	Adding polynomial feature interactions in which new features are defined*	Li et al. (2014)

** This is a different sort of a pre-processing method in comparison with standardization and normalization*

To limit overfitting, cross-validation is used on the training set. The cross-validation method used is StratifiedKFold, which is a cross-validation method that splits data into *k* folds, with an equal number of examples in each fold. What makes StratifiedKFold different from the standard K fold is that each subject is exactly one time in the validation set. In the standard K fold, this process is random, and it could occur that one subject is in the validation set multiple times. Prior research indicated that the cross-validation method StratifiedKFold performs better than standard cross-validation methods in terms of bias and variance (Sechidis, Tsoumakas & Vlahavas, 2011).
Since the current study strives towards a ratio of 70/15/15 for training, validation and the test set, respectively, the number selected for k is 5. In this case, 85% of the total data are split into 5 equal folds, where for each iteration's 4 folds represent the training data, and 1 fold represents the validation data. The exact ratio between training, validation and test set is 68/17/15. As can be seen in Figure 3.6.1.1, the classifiers are trained with cross-validation. StratifiedKFold = 5 is used for each of the five training sets. The test set is held apart from the training set and is not used in this part of the experiment.

A boxplot of all the F1 scores together is created for all five training sets to gain insight into the effect of SMOTE on performance. Thereafter, for each classifier, a boxplot with corresponding pre-processing methods is created, since it easily gives insight into pre-processing types and classifiers they fit best. The pre-processing method with the highest median for the F1 score is selected. Once the pre-processing step is selected for each classifier, this will remain the same for the rest of this study.


Training set 1
Test set
IT 1
IT 2
IT 3
IT 4
IT 5
Training
Validation
Training set 2
IT 1
IT 2
IT 3
IT 4
IT 5
Training set 3
IT 1
IT 2
IT 3
IT 4
IT 5
Training set 4
IT 1
IT 2
IT 3
IT 4
IT 5
Training set 5
IT 1

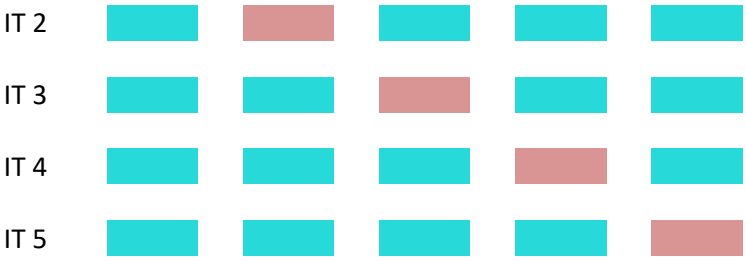


Figure 3.6.1.1. *Overview of training fase on each training set, by means of cross-validation. Training set consist of 85% of the data and the test set consist of the remaining 15%. The test set held apart and is not used.*
Abbreviation: IT = iteration.

Step 4: Tuning parameters of each classifier

Once the pre-processing step is selected, the parameters of the classifiers will be tuned. In this step, the goal is to find the best parameters of the classifiers. The parameters that will be adjusted can be seen in Table 3.6.1.3, which also shows how many combinations are tried for each classifier. To be able to reproduce the experiments, all classifiers are given a random state of 0. Once the parameters are selected for each classifier, they will remain the same for RQ2 and a part of RQ3.

Table 3.6.1.3

Overview of parameters and corresponding values to be tuned for each classifier

Classifier	Parameters	Number of combinations
DTC	Criterion: 'gini', 'criterion' Splitter: 'best', 'random'	4
LSVC	C: 0.001, 0.5, 1, 10, 100, 1000 Multi_class: 'ovr', 'crammer_singer' Fit_intercept: True, False Intercept_scaling: 0.5, 1, 2, 10 Loss: 'hinge', 'squared_hinge' Penalty: 'l1', 'l2' Tol: 0.01, 0.001, 0.0001, 0.0001	1536
LR	Penalty: 'l1', 'l2' C: 0.001, 0.5, 1, 10, 100, 1000 Class_weight: None, 'balanced' Solver: 'liblinear', 'newton-cg', 'lbfgs', 'sag' Multi_class: 'ovr', 'multinomial'	192
PER	Penalty: 'none', 'l2', 'l1', 'elasti*CN*et' Alpha: 0.0001, 0.0003, 0.0005, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3 Fit_intercept: True, False N_inter: 5, 10, 15, 20, 25, 50 eta0: [[10 ** x for x in range(-5, 00)]]	2160
SGD	Loss: 'hinge', 'log', 'modified_huber', 'squared_hinge', 'perceptron', Penalty: 'none', 'l2', 'l1', 'elasti*CN*et' Alpha: 0.0001, 0.0005, 0.001, 0.005, 0.01 Fit_intercept: False, True eta0: [[10 ** x for x in range(-5, 00)]]	3000
SVC	C: 0.001, 0.5, 1, 10, 100, 1000 Kernel: 'rbf', 'linear'	576

Tol: 0.01, 0.001, 0.0001, 0.0001

Decision: ‘ovo’, ‘ovr’, None

Gamma: 0.01, 0.001, 0.0001, ‘auto’

Abbreviations: DTC = Decision-Tree classifier, LSVC = linear Support-Vector classifier, LR = Logistic Regression, PER = Perceptron, SGD = Stochastic Gradient-Descent classifier, SVC = Support-Vector classifier.

As for step 3, problems such as overfitting need to be limited. For this step, training set 1 is reused from the previous step. StratifiedKFold is applied again on this training set (see Figure 3.6.1.2). For each iteration, parameters that provide the highest F1 score are stored. When all 5 iterations are completed, the parameters that appear the most are chosen for the classifiers.

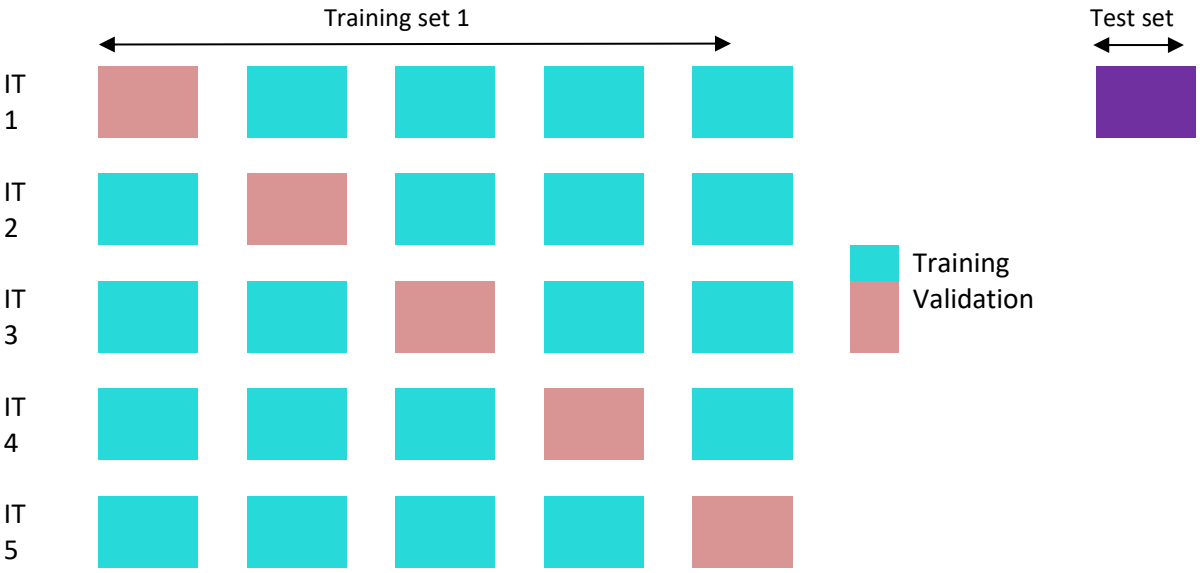


Figure 3.6.1.2. An *overview of cross-validation on training data. The training set consists of 85% of the data, and the test set consists of the remaining 15%. The test set held apart and is not used. Abbreviation: IT = iteration.*

Step 5: Selecting the final classifier for following research questions

The classifier with the highest F1 score in step 4 is the classifier that is used for the coming research questions. To gain insight into how this model performs on unseen data, the model is used on the test set which have been isolated in step 1. Now, the model is trained on the full training data (see Figure 3.6.1.3). After training, the model is applied on the test set. To evaluate the performance on separating *MCI* subjects from *AD* subjects, the accuracy, F1 score, recall and precision are calculated. This gives insight into how this model performs on new data in terms of overfitting and compared to prior research.

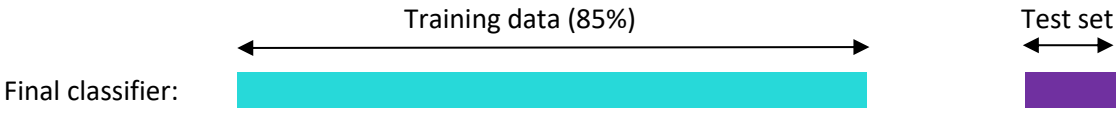


Figure 3.6.1.3. *An overview of the training phase on all training data. The model is tested on the test set.*

3.6.2 Experimental Procedure for Answering RQ2: Binary Classification (Predicting Progression)

This experimental procedure is set up to examine how this classifier performs on five binary-classification problems, by distinguishing *stableMCI* from different *progressionMCIs*, to gain insight into the potential of increasing complexity to distinguish these classes as the moment of progression is further away in time.

Step 1: Splitting all subsets in the training set and the test set

For answering RQ2, five subsets are made as described before in Subsection 3.1.1: *stableMCI* and *progression6MCI*, *stableMCI* and progression12*MCI*, *stableMCI* and *progression24MCI*, *stableMCI* and *progression30MCI* as well as *stableMCI* and *progression48MCI*.

The ratio of the training set to the test set is set at 85%:15%, for the same reason as that described in Subsection 3.6.1. Besides that, the ratio between the two classes presented in the original data is also be the ratio of the training set to the test set. Therefore, a fixed number of subjects from each group is randomly selected for the test set. See Table 3.6.2.1 for the split between the training and test sets, with the same ratio in both training and test set.

Table 3.6.2.1.

An overview of subjects for RQ2 in the test set and training set before SMOTE

	Test set		Training set (before SMOTE)	
#	s*MCI*	p#*MCI*	s*MCI*	p#*MCI*
6	67	21	379	118
12	67	19	379	106
24	66	19	380	104
30	68	8	378	53
48	67	5	379	30

Abbreviations:sMCI = stableMCI, p#MCI = progression#MCI, where # stands for a number in the column with name '#'.

Step 2: SMOTE on training set

As mentioned in Subsection 3.2, SMOTE is applied to the training set. As for experiment 1, the parameters for SMOTE are set to 100 for *perc.over*, to make the size of the minority class twice as large. The parameter *perc.under* is set to 200, to remove at most half of the majority class and make the data balanced. These settings of perc.over and perc.under are used in prior research (Hao, Wang & Bryant, 2014). The latter step is a form of undersampling. See Table 3.6.2.2 for an overview of subjects in training set before and after SMOTE.

Table 3.6.2.2.

Overview of subjects for RQ2 training set before and after SMOTE

	Training set (before SMOTE)		Training set (after SMOTE)	
#	s*MCI*	p#*MCI*	s*MCI*	p#*MCI*
6	379	118	236	236
12	379	106	212	212
24	380	104	208	208
30	378	53	106	106
48	379	30	60	60

Abbreviations:sMCI = stableMCI, p#MCI = progression#MCI, where # stands for a number in the column with name '#'.

Step 3: Determining F1 scores

For each subset made for RQ2, the optimized classifier chosen in the last step of experiment 1 is trained on the training data and tested on the test set, as can be seen in Figure 3.6.2.1. To evaluate the performance of classifying *stableMCIs* and *progressionMCIs*, the accuracy, F1 score, recall and precision are calculated. This gives insight into how this model performs on new data and the complexity of different classification problems.

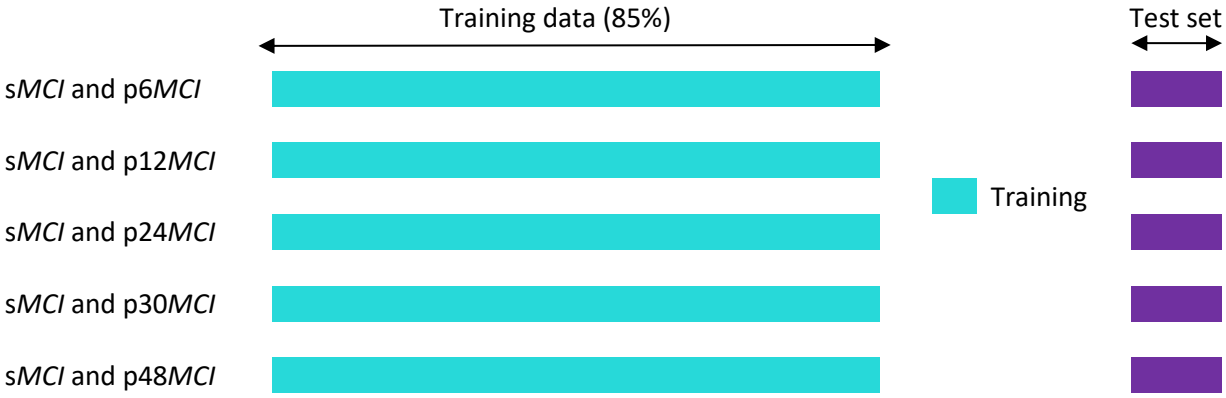


Figure 3.6.2.1. *An overview of the training phase on all training data. The model is tested on the*

corresponding test set.

3.6.3 Experimental Procedure for Answering RQ3: Multiclass Classification I (Predicting Progression and Its Corresponding Moment)

In this experiment, a multiclass classification is set up. The classes consist of *stableMCI*, *progression6MCI*, *progression12MCI*, *progression24MCI*, *progression30MCI* and *progression48MCI*.

Step 1: Splitting all subsets in the training and test set

As for RQ1, 15% of the data are the test set, and the remaining part is training set (85%), which will be divided into validation and training sets. See Table 3.6.3.1 for an overview of the subjects.

Table 3.6.3.1

An overview of subjects for RQ3 in the test and training sets before SMOTE

	s*MCI*	p6*MCI*	p12*MCI*	p24*MCI*	p30*MCI*	p48*MCI*	Total
TOTAL	446	139	125	123	61	35	929
Percentage	0.48	0.15	0.13	0.13	0.07	0.04	1
Training (85%)	379	118	107	106	51	29	790
Test (15%)	67	21	18	17	10	6	139

Abbreviations: sMCI = stableMCI, p6MCI = progression6MCI, p12MCI = progression12MCI, p24MCI= progression24MCI, p30MCI = progression30MCI, p48MCI = progression48MCI.

Step 2: SMOTE on the training set

Because this is a multiclass-classification task, the SCUT algorithm is applied. The average of all six classes in the training set is 132. This means that oversampling is first applied to *progression6MCI*, *progression12MCI*, *progression24MCI*, *progression30MCI* and *progression48MCI* in such a manner that a class contains 132 subjects after SMOTE has been applied. Undersampling is applied to *stableMCI*. In the end, each class in the training set contains exactly 132 subjects.

Step 3: Determining F1 scores

The goal of this step is to gain insight into the performance of the model from RQ1 and RQ2 on this multiclass-classification task. This is obtained by conducting a StratifiedKFold cross-validation on the training set with $k = 5$ (see Figure 3.6.1.1).

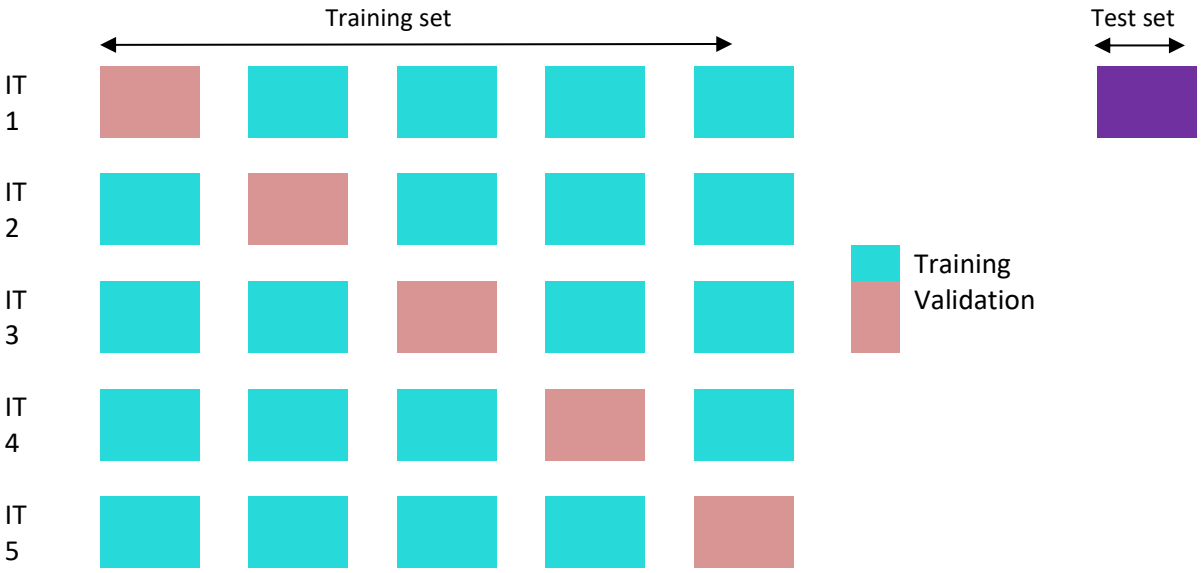


Figure 3.6.1.1. *An overview of the cross-validation of the training set and test set. The training set consists of 85% of the data, and the test set consists of the remaining 15%.*
Abbreviation: IT = iteration.

Step 4: Parameter tuning

Since multiclass classification is a different problem from binary classification, the optimal parameters for the final classifier are sought after. The parameters that are tuned are the same parameters as those seen in Table 10. This is done by conducting StratifiedKFold cross-validation on the training set, with $k = 5$, to minimize overfitting. These F1 scores are compared to the F1 scores from step 3.

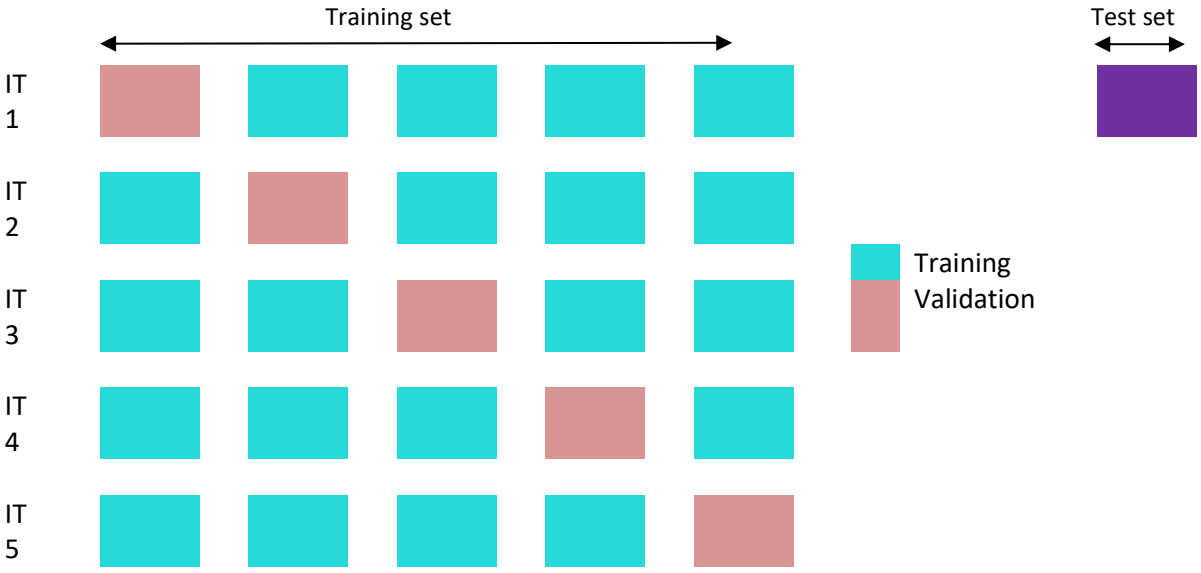


Figure 3.6.3.2. *An overview of the cross-validation of training data and the test set. The training set consists of 85% of the data, and the test set consists of the remaining 15%.*
Abbreviation: IT = iteration.

Step 5: Comparing an optimized classifier to a Dummy classifier

Now, the classifier is optimized, and the performance of the model on unseen data is tested. The model is trained on all training data. A Dummy classifier, which generates predictions uniformly at random, is also trained on the same training data, and the difference in performance is compared (see Figure 17). The difference in performance is tested for significance according to the McNemar Test. Comparing the performance of both methods may indicate whether the final classifier has failed to achieve a reasonable performance on predicting *AD* (Musafa, Kraft & Krüger, 2015).

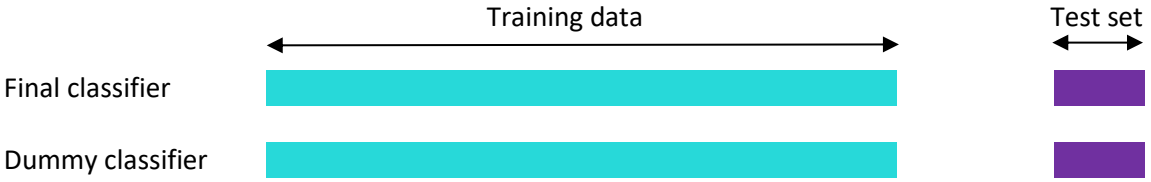


Figure 3.6.3.3. *An overview of the training phase on all training data. The model is tested on the test set. Training data and the test set are the same for both classifiers.*

3.6.4 Experimental Procedure's Follow-up 1: Multiclass Classification II (Predicting the Moment of Progression)

Based on the results from Experimental Procedure 3, a follow-up study is conducted to investigate the extent to which the classifier could make a multiclass distinction among *progression6MCI, progression12MCI, progression24MCI, progression30MCI* and *progression48MCI*. This allows for an investigation of whether a classifier is able to predict the moment of progression.

Step 1: Splitting all subsets in the training and test sets

As for the previous experimental procedures, the subset is split into a training set (85%) and a test set (15%). A total of 483 observations are in this subset. Seventy-two observations are selected for the test set, in the same ratio as in the total set, which can be seen in Table 3.6.4.1.

Table 3.6.4.1

An overview of subjects for RQ3 in the test and training sets before SMOTE

	p6*MCI*	p12*MCI*	p24*MCI*	p30*MCI*	p48*MCI*	Total
Subjects	139	125	123	61	35	483
Percentage	0.29	0.26	0.25	0.13	0.07	1
Training (85%)	118	106	105	52	30	411
Test (15%)	21	19	18	9	5	72

Abbreviations:p6MCI = progression6MCI, p12MCI = progression12MCI, p24MCI= progression24MCI, p30MCI = progression30MCI, p48MCI = progression48MCI.

Step 2: SCUT on training data

Because this is a multiclass-classification task, the SCUT algorithm is applied. The average size of all five classes in the training data is 82. This means that oversampling is first applied to *progression30MCI* and *progression48MCI* in such a manner that the classes contain 82 subjects after SMOTE has been applied. Undersampling is applied to *progression6MCI*, *progression12MCI* and *progression24MCI*. In the end, each class in the training set contains exactly 82 subjects.

Step 2: Determining F1 scores

The optimized model from experiment 3 is also used in the follow-up experiment. This model is trained on the training set, and the learning algorithm will then be tested on the test set. This is also done with a Dummy classifier (see Figure 3.6.3.4). Comparing the performance of both methods may indicate whether the classifier has failed to achieve a reasonable performance in predicting *AD* (Musafa, Kraft & Krüger, 2015).

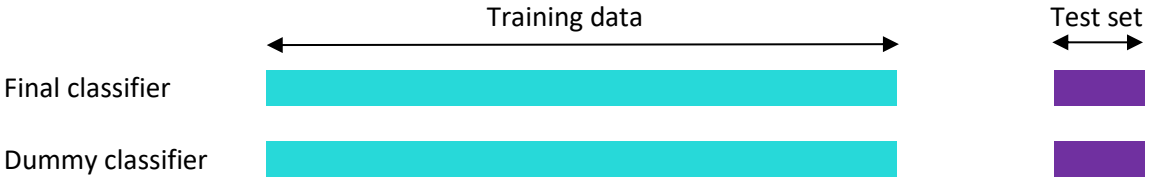


Figure 3.6.3.4. *An overview of the training phase on all training data. The model is tested on the test set. Training data and the test set are the same for both classifiers.*

Section 4: Results

This section provides the results of the conducted experiments. It starts with the results of experiment 1 in Subsection 4.1, which seeks for the classifier that performs best in distinguishing *MCI* subjects from *AD* subjects at baseline. After this, in Subsection 4.2, the results of experiment 2 are provided, which examines the extent to which this classifier is able to make a binary distinction between *stableMCI* and *progressionMCI*, as the moment of progression differs. Subsection 4.3 provides the results of experiment 3, which investigates the extent to which the optimized classifier is able to predict whether and when a subject is likely to progress to *AD*. Finally, Subsection 4.4 presents the result of the follow-up experiment, which is an investigation of the extent to which the optimized classifier is able to predict when a subject is likely to progress to *AD*.

4.1 Result of Experiment 1: Model Selection

To select the best-performing classifier on the classification task between *MCI* and *AD* subjects, the extent to which the random factor in SMOTE has an effect on the F1 score is first examined. Figure 4.1.1 compares the average F1 scores of all five training sets for each classifier. Overall, the random factor in SMOTE did not affect the F1 score to a large extent, except for LSVC and SGD with a pre-processing method that added feature interactions. From here on, a random state for SMOTE is chosen (seed = 1) to reproduce the events.

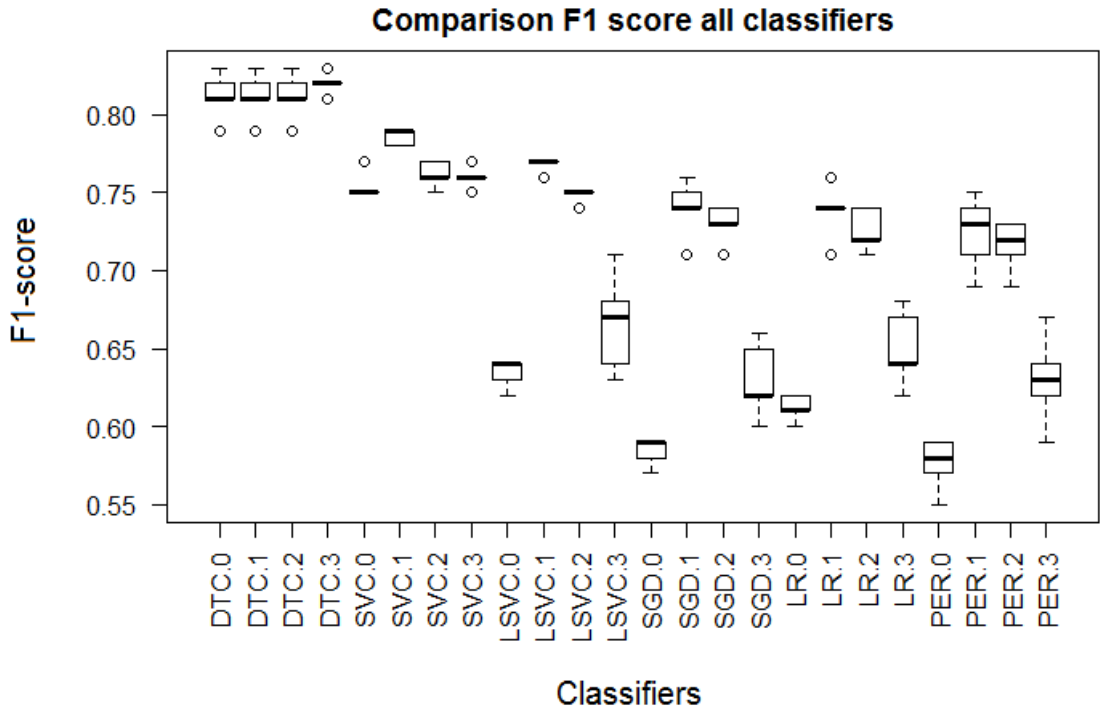


Figure 4.1.1. *Comparison of F1 scores of all classifiers.*

Abbreviations: DTC = Decision-Tree classifier, SVC = Support-Vector classifier, LSVC = linear Support-Vector classifier, SGD = Stochastic Gradient Descent, LR = Logistic Regression, PER = Perceptron, 0 = no pre-processing method, 1 = normalizing, 2 = standardizing, 3 = adding feature interactions.

In the next part of the experiment, the pre-processing steps were selected for each classifier. The selected pre-processing method for the DTC added feature interactions, because the median was the highest for this method, as can be seen in Figure A.1 in the appendices. The selected pre-processing method for the SVC, LSVC, SGD, LR and PER is normalizing, since this method had the highest median for these classifiers, as can be seen in Figures A.2 to A.6 in the appendices.

The following part of the experiment sought for the optimal parameters of each classifier with the selected pre-processing method, according to Table 3.6.1.1. Table 4.1.1 presents the optimal parameters for each classifier and the average F1 score achieved based on cross-validation.

Table 4.1.1

Tuned parameters for each classifier and its F1 score based on cross-validation

Classifier	Averaged F1 Score (out of 5)	Best parameters
Decision-Tree classifier	0.82	‘gini’, ‘best’
Support-Vector classifier	0.83	1000, ‘rbf’, 0.01, ‘ovo’, ‘auto’
Linear Support-Vector classifier	0.75	10, ‘ovr’, True, 2, ‘squared_hinge’, ‘l2’, 0.01
Logistic Regression	0.75	‘l1’, 1, None, ‘liblinear’, ‘ovr’
Perceptron	0.75	‘l2’, 0.0001, True, 5, 1e-05
Stochastic Gradient Descent	0.78	‘hinge’, ‘none’, False, 5, 1e-05, 0.005

Figure 4.1.2 gives an overview of the aforementioned averaged F1 score of each classifier (Table 4.1.1) with the selected pre-processing method and tuned parameters. All classifiers perform well, but the DTC and the SVC are worth investigating since they have achieved the highest scores. Because the SVC has a higher upper quartile and has a better averaged F1 score in comparison with the DTC, the SVC is selected for use throughout the current study.

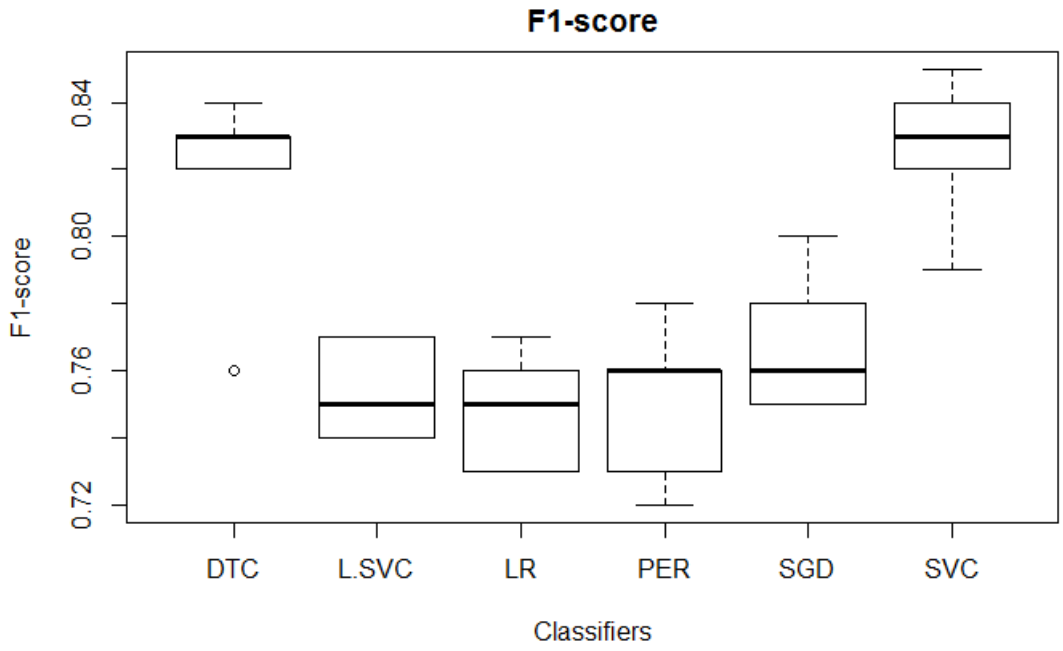


Figure 4.1.2. Boxplot comparing F1-macro results of classifiers, based on cross-validations. *Abbreviations: DTC = Decision-Tree classifier, SVC = Support-Vector classifier, LSVC = linear Support-Vector classifier, SGD = Stochastic Gradient Descent, LR = Logistic Regression, PER = Perceptron.*

The SVC, with parameters tuned according to Table 4.1.1, is then tested on the test set. The results are presented in Table 4.1.2.

Table 4.1.2

F1 score on the test set for RQ1

F1 score	Accuracy	Recall	Precision
0.76	0.79	0.80	0.74

4.2 Result of Experiment 2: Binary Classification (Predicting Progression)

The purpose of this experiment was to examine how the SVC performs on five binary-classification problems, by distinguishing *stableMCI* from different *progressionMCIs.* The SVC with tuned parameters (1000, 'rbf', 0.01, 'ovo', 'auto') and pre-processing method normalization was trained on each training set. After the training had been completed, the model was tested on the test set. As can be seen in Table 4.2.1, the closer to the moment of progression, the higher the accuracy. Concerning the F1 score, there seems to be a drop of performance from *progression24MCI* to *progression30MCI*.

Table 4.2.1

RQ2: Overview of F1 scores, accuracy, recall and precision of different subsets on test set

Classes	F1 score	Accuracy	Recall	Precision
s*MCI* and p6*MCI*	0.66	0.77	0.65	0.68
s*MCI* and p12*MCI*	0.68	0.76	0.71	0.67
s*MCI* and p24*MCI*	0.63	0.71	0.66	0.62
s*MCI* and p30*MCI*	0.48	0.66	0.53	0.51
s*MCI* and p48*MCI*	0.54	0.72	0.57	0.52

Abbreviations: sMCI = stableMCI, p6MCI = progression6MCI, p12MCI = progression12MCI, p24MCI = progression24MCI, p30MCI = progression30MCI, p48MCI = progression48MCI

4.3 Result of Experiment 3: Multiclass Classification I (Predicting Progression and Its Corresponding Moment)

The goal of this experiment was to investigate the extent to which the optimized classifier performed on multiclass classification to be able to predict whether and when a subject is likely to progress to *AD*. First, the SVC with tuned parameters (1000, 'rbf', 0.01, 'ovo', 'auto', None) and pre-processing method normalizing was trained on the training set and evaluated on the validation set. The results can be seen in Table 4.3.1.

Table 4.3.1

An overview of F1 score, accuracy recall, and precision for the SVC with cross-validation

Kfold	F1 score	Accuracy	Recall	Precision
1	0.35	0.35	0.35	0.36
2	0.36	0.37	0.37	0.36
3	0.33	0.32	0.32	0.35
4	0.37	0.41	0.41	0.38
5	0.39	0.41	0.41	0.38
Average	0.36	0.37	0.37	0.37

Abbreviation: SVC = Support-Vector classifier.

Parameters for SVC are 1000, 'rbf', 0.01, 'ovo', 'auto', and None

Hereafter, through cross-validation on the training set, the optimal parameters for this multiclass-classification task are examined. These optimal parameters for each cross-validation can be seen in Table 4.3.2. The most common settings are 100, 'rbf', 0.01, 'ovo', and 'auto' for the parameters c,

kernel, tol, decision_function_shape and gamma, respectively. These optimized parameters are slightly different from the parameters used in RQ1 and RQ2, and the model achieved a slightly better F1 score on average.

Table 4.3.2.

An Overview of best parameters for the SVC with cross-validation

Kfold	F1 score	C	Kernel	Tol	DFS	gamma
1	0.36	100	'rbf'	0.01	'ovo'	'auto'
2	0.36	0.5	'rbf'	0.01	'ovo'	'auto'
3	0.33	1000	'rbf'	0.01	'ovo'	'auto'
4	0.39	0.001	'rbf'	0.01	'ovo'	'auto'
5	0.39	100	'rbf'	0.01	'ovo'	'auto'
Average / most common	0.37	100	'rbf'	0.01	'ovo'	'auto'

Abbreviation: DFS = decision_function_shape

This optimized model (SVC with optimized parameters (100, 'rbf', 0.01, 'ovo', 'auto') and pre-processing method normalization) is trained on all the training data and tested on the test set. To be able to compare this performance, a Dummy classifier is also trained on the same training set and tested on the same test set. The differences in performance between the optimized model and the Dummy classifier can be seen in Table 4.3.3. The McNemar test is conducted to compare both performances. This table demonstrates that the SVC has higher scores in terms of F1 score, accuracy, recall and precision in comparison with a Dummy classifier, which is significant at the $p < 0.001$ level.

Table 4.3.3

The final performance of predicting progression and the moment it occurs in comparison with Dummy classifier

	F1 score	Accuracy	Recall	Precision
SVC*	0.26	0.35	0.26	0.29
Dummy classifier (strategy = "uniform")*	0.12	0.16	0.13	0.16

** difference is significant (McNemar test: p = 0.0002)*

Table 4.3.4 compares the F1 score, recall and precision for each class separately. This table is revealing in several ways. First, it supports the idea that it is harder for the classifier to classify subjects of whom the progression moment is further away in time. Second, for *stableMCI* and *progression6MCI*, the model performs better on precision. For the other classes, it performs better on recall. Third, most notable is that the highest score for both recall and precision is achieved by the

stableMCI class. The F1 score for the *stableMCI* class is higher than those of other classes.

Table 4.3.4

RQ3: F1 score, recall, and precision for each class on the test set

	s*MCI*	p6*MCI*	p12*MCI*	p24*MCI*	p30*MCI*	p48*MCI*
F1 score	0.59	0.39	0.23	0.21	0.14	0.10
recall	0.48	0.33	0.17	0.23	0.20	0.17
precision	0.78	0.47	0.10	0.18	0.11	0.07

Abbreviations: sMCI = stableMCI, p6MCI = progression6MCI, p12MCI = progression12MCI, p24MCI = progression24MCI, p30MCI = progression30MCI, p48MCI = progression48MCI

4.4 Results of Follow-up 1: Multiclass Classification II (Predicting the Moment of Progression)

Since the F1 score for the *stableMCI* class was considerably higher in comparison with those for the other classes, these results suggest that the SVC performed better than the Dummy classifier because it could classify *stableMCI* subjects well. A follow-up experiment is conducted to investigate the extent to which the optimized classifier performs on a multiclass-classification task when the classifier only predicts the moment of progression. The *stableMCI* class, which was included in experiment 3, is left out.

To compare the performance of the classifier, the optimized SVC is compared to a Dummy classifier. As illustrated in Figure 4.4.1, the optimized SVC performs similarly to the Dummy classifier on F1 score, recall and precision.

Table 4.4.1

Follow-up 1: F1 score, accuracy, recall and precision on the follow-up on the test set

	F1 score	Accuracy	Recall	Precision
SVC*	0.15	0.18	0.15	0.13
Dummy Classifier (strategy = "uniform")*	0.18	0.18	0.18	0.17

* Difference is not significant (McNmar test: p-value = 0.17)

The F1 score, recall and precision for each separate class are compared in Table 4.4.2. It is apparent from this table that the classifier performs poorly in predicting *progression24MCI* and *progression30MCI* subjects.

Table 4.4.2

F1 score, recall, and precision for each class's follow-up test

	p6*MCI*	p12*MCI*	p24*MCI*	p30*MCI*	p48*MCI*
F1 score	0.29	0.24	0.06	0	0.15
recall	0.33	0.21	0.07	0	0.13
precision	0.26	0.27	0.07	0	0.13

Abbreviations: p6MCI = progression6MCI, p12MCI = progression12MCI, p24MCI = progression24MCI, p30MCI = progression30MCI, p48MCI = progression48MCI

To gain more insight into how this result was achieved for the different classes, a confusion matrix was calculated (Table 4.4.3). Note that when the classifier made a wrong prediction, the prediction was frequently made in the direction of the surrounding classes. As an example: the classifier accurately classified *progression12MCI* subjects four times. When the predicted subjects of the surrounding classes (*progression6MCI* and *progression24MCI*) are added, this number increases to 18 (10 + 4 + 4) for recall and 12 (2 + 4 + 6) for precision.

Table 4.4.3

Confusion matrix follow-up

		Predicted label				
		p6*MCI*	p12*MCI*	p24*MCI*	p30*MCI*	p48*MCI*
	p6*MCI*	**7**	2	6	3	3
True	p12*MCI*	10	**4**	4	0	1
label	p24*MCI*	5	6	**1**	4	2
	p30*MCI*	3	3	2	**0**	1
	p48*MCI*	2	0	1	1	**1**

Abbreviations: p6MCI = progression6MCI, p12MCI = progression12MCI, p24MCI = progression24MCI, p30MCI = progression30MCI, p48MCI = progression48MCI
A correct predicted label is in bold.

To investigate the aforementioned pattern, a post-hoc analysis is conducted. In this *post hoc* analysis, two new confusion matrices are calculated to investigate how much the accuracy would increase when the surrounding classes are added to the target class. These new confusion matrices are presented in Tables 4.4.4 and 4.4.5, respectively.

Table 4.4.4

Confusion matrix: combining p6MCI + p12MCI and p24MCI + p30MCI + p48MCI

		Predicted label	
		p6*MCI* + p12*MCI*	p24*MCI* + p30*MCI* + p48*MCI*
True label	p6*MCI* + p12*MCI*	23 (TP)	17 (FN)
	p24*MCI* + p30*MCI* + p48*MCI*	19 (FP)	13 (TN)

Abbreviations: TP = True Positive, FN = False Negative, FP = False Positive, TN = True Negative.

Tables 4.4.4 and 4.4.5 indicated that the True Positive and False Negative increase when one combines classes, which is expected since there are now only two classes instead of five.

Table 4.4.5

Confusion matrix: combining p6MCI + p12MCI + p24MCI and p30MCI + p48MCI

		Predicted label	
		p6*MCI* + p12*MCI* + p24*MCI*	p30*MCI* + p48*MCI*
True label	p6*MCI* + p12*MCI* + p24*MCI*	45 (TP)	13 (FN)
	p30*MCI* + p48*MCI*	11 (FP)	3 (TN)

Abbreviations: TP = True Positive, FN = False Negative, FP = False Positive, TN = True Negative.

Based on these confusion matrices, the F1 score, accuracy, recall and precision are calculated, which are presented in Table 4.4.6. This table needs to be interpreted with caution, since this is a *post hoc* analysis and the classifier is not trained on these new intervals. Both combinations give higher F1 scores (0.56 and 0.79) than the dummy variable (F1 score: 0.18). Notably, when combining p6*MCI* + p12*MCI* + p24*MCI and* p30*MCI* + p48*MCI*, the F1 score is even higher than the F1 scores achieved in all previous experiments.

Table 4.4.6

Follow-up 1: F1 score, accuracy, recall and precision calculated from confusion matrices in Tables 4.4.4 and 4.4.5.

Classes together:	F1 score	Accuracy	Recall	Precision
p6*MCI* + p12*MCI and* p24*MCI* + p30*MCI* + p48*MCI*	0.56	0.50	0.58	0.55
p6*MCI* + p12*MCI* + p24*MCI and* p30*MCI* + p48*MCI*	0.79	0.67	0.78	0.80

Section 5: Discussion

This section discusses the findings of the present study. For each experiment that was conducted, the findings are discussed in Subsection 5.1. In Subsection 5.2, the answers to the problem statement are provided, followed by the limitations of the current study and recommendations for future research in Subsections 5.3 and 5.4, respectively.

5.1 Answers to Research Questions

The goal of the current study was to investigate the extent to which a classifier is able to predict subjects' progression from *MCI* to *AD* and its corresponding moment using MRI data. In reviewing the literature, no researchers investigated predicting the moment of progression but only predicting progression based on different features. Three research questions were set up to achieve this goal:

RQ1: *What classifier and in combination with which pre-processing method performs best in distinguishing MCI subjects from Alzheimer's disease subjects at baseline*

RQ2: *To what extent can the optimized classifier make a binary distinction between* stable *MCI subjects and MCI subjects who progress to Alzheimer's disease within 6, 12, 24, 30 and 48 months?*

RQ3: *To what extent can the optimized classifier predict progression from MCI subjects to AD and its corresponding moment in a multiclass classification task?*

Each research question is answered in the remaining part of this section.

RQ1: *What classifier and in combination with which pre-processing method performs best in distinguishing MCI subjects from Alzheimer's disease subjects at baseline*

Since there is no classifier that performs best on every problem, the first question in the current study sought to determine a classifier that performs best in distinguishing *MCI* subjects from *AD* subjects in a proof-of-concept study. The best-performing classifier is used throughout the current study. It is beyond the scope of the study to examine which classifier performs best on all different experiments. Performances of the following classifiers are compared: DTC, LSVC, LR, PER, SGD and SVC. First, different pre-processing methods that work best for each classifier separately are evaluated. These findings, as outlined in Subsection 4.1, indicate that in this case, adding feature interactions works best for the DTC, and for all the other classifiers, standardizing the data works best. After this, the best parameters for all classifiers with their pre-processing methods are investigated. Using StratifiedKFold, the DTC and the SVC perform better than other classifiers, as presented in Table 4.1.1. Since the F1 score based on cross-validation on the training set was the highest for the SVC, this

classifier was chosen to be used for the remaining part of the current study. To see how this classifier performs on unseen data, it was tested on the test set. The classifier performed slightly worse on the test set (F1 score: 0.76) in comparison with its performance based on cross-validation on the training set (F1 score: 0.83). This indicates that the model seems to slightly overfit the training data.

The answer to the first research question, therefore, is as follows: out the set of classifiers tried, the SVC performs best with the normalizing pre-processing method and tuned parameters, but it slightly overfits the training data.

RQ2: *To what extent can the optimized classifier make a binary distinction between* stable *MCI subjects and MCI subjects who progress to Alzheimer's disease within 6, 12, 24, 30 and 48 months?* The second research question sought to investigate how well the SVC performed in distinguishing *stableMCI* from *progressionMCI*, on different moments before progression. It was hypothesized that the closer the moment of progression, the better the performance of the classifier, since the differences between the classes probably increase as the moment of progression is closer. Therefore, it would be relatively easier to make a distinction between the two classes, as suggested in Subsection 2.1. These results in Subsection 4.3 further support the idea that it is possible to predict progression from *MCI* to *AD*. Also, the hypothesis was confirmed, since the classifier performs with an F1 score of 0.63 or higher on *progression6MCI* to *progression24MCI (*Table 4.2.1). This F1 score is higher in comparison with *progression30MCI* and *progression48MCI*, where the performance drops to 0.5 and below. One interesting finding is that this is not a straight descending line: the F1 score for *progression48MCI* and *stableMCI* is higher than that for *progression30MCI* and *stableMCI*, even though having a subtle difference of 0.02, which can be seen in Subsection 4.3. This is also the case for *progression12MCI* and *stableMCI*, which exhibits a higher F1 score than *progression6MCI* and *stableMCI*.

The performances in terms of accuracy observed in the current study are comparable with those observed by other researchers who used MRI data with a specific moment of progression (Trezpacs et al., 2014; Chupin et al., 2009; Wolz et al., 2010). For distinguishing *stableMCI* from *progression12MCI*, experiment 2.1 achieved a higher accuracy than that by Wolz et al. (2010), i.e. 76% to 64%. This was also the case for distinguishing *stableMCI* from *progression24MCI*: experiment 2.3 achieved an accuracy of 71%, whereas Trezpacs et al. (2014) achieved an accuracy of 67%. Chupin et al. (2009) achieved an accuracy of 71% in distinguishing *stableMCI* from *progression18MCI*, but *progression18MCI* is not concluded in the current study; this makes comparing the results more complex. Since experiments 2.2 and 2.3 achieved an accuracy of 76% and 71% for *progression12MCI* and *progression24MCI*, respectively, it is believed that similar results for *progression18MCI* would be achieved. All these results together indicate that the model performs in line with prior research.

The answer to the second research questions, therefore, is as follows: the SVC is able to make a binary distinction between *stableMCI* and the different *progressionMCIs*, with similar performances

in comparison with prior studies. The classifier performs better when the moment of progression is closer, which confirms the hypothesis made in Subsection 1.2.

RQ3: *To what extent can the optimized classifier predict progression from MCI subjects to AD and its corresponding moment in a multiclass classification task?*
The third research question sought to investigate the extent to which the optimized classifier is able to predict progression and its corresponding moment. First, the difference in performance of the classifier used in RQ1, RQ2, and step 3 of RQ3 (F1 score: 0.36, see Table 4.3.1) as well as the optimized classifier are addressed (F1 score: 0.37, see Table 4.3.2). The difference in F1 score is 0.01, which is in favor of the optimized classifier. However, according to the McNemar Test, this difference in performance is not significant ($p = 1$). However, the current study continued for the remaining part using the 'optimized' classifier.

This model was tested on the test set. The classifier performed worse on the test set (F1 score: 0.26) than on cross-validation of the training set (F1 score: 0.37). This is a drop of 30% in performance, which indicates that this model overfitted the training data.

However, even though this classifier seems to overfit the training data, when one compares the performance of the optimized classifier on the test set (F1 score: 0.26) with a Dummy classifier (F1 score: 0.15), the SVC performs significantly better according to the McNemar Test ($p = 0.0002$; for a contingency table, see Appendix B.1). In comparison of the F1 scores achieved in experiment 2, there is a significant drop of performance. However, this was expected since multiclass classification is much more complex than binary classification. The F1 score of 0.26 is constructed by the average of all classes. The F1 score per class, which is mentioned in Section 4.3, indicated that the F1 score of 0.26 is higher than the Dummy classifier mainly because of *stableMCI* and *progression6MCI*, with F1 scores of 0.59 and 0.39 respectively. The high F1 score for *stableMCI* (0.59) stands out, since it is more than as twice as the average of all classes (F1 score: 0.26). All other classes scored below average and reached the lowest performance for *progression48MCI* (0.1). These results together suggest that the reason why the SVC beat the Dummy classifier is that it could classify *stableMCI* (and, to a less extent, *progression*6*MCI*) very well. Based on these results, it is not clear to what extent the Support-Vector classifier can predict when a subject is likely to convert to *AD*.

The answer to the third research question, therefore, is as follows: the SVC is able to make a multiclass prediction better than the Dummy classifier. However, this seems to be due to the *stableMCI* (and, to a less extent, the *progression6MCI*) class, which achieved an F1 score that was out of proportion. To gain more insight into this problem, a follow-up study was conducted, which excluded *stableMCI* in the multiclass-classification task.

Follow-up Question: *To what extent can the optimized classifier make a multiclass distinction among MCI subjects who progress to Alzheimer's disease within 6, 12, 24, 30 and 48 months?*

The follow-up experiment sought to examine how this classifier performs on the same multiclass-classification task, except that the class *stableMCI* was excluded. This gives insight into the extent to which the classifier is able to make a distinction among the different *progressionMCI*s and conclude whether the classifier is able to predict the moment in which subjects will progress to *AD*.

One unanticipated finding was that the performance of the optimized classifier (0.15) was not better than the Dummy classifier (0.18), which was not significant according to the McNemar Test ($p = 0.17$; for a contingency table see Appendix B.2). These findings, which are provided in Subsection 4.4, are rather disappointing. It can, therefore, be assumed that it is difficult for the classifier to make a sufficient distinction between those classes. However, when one looks into the performance more closely, an interesting pattern occurs. When the classifier makes a wrong prediction, the prediction is often made in the direction of the class *before* or *after* the accurate class. This may suggest that it is too complex to make a distinction among *progression6MCI*, *progression12MCI*, *progression24MCI*, *progression30MCI* and *progression48MCI*, because the time intervals are too close to predict the moment of progression.

There are several possible explanations for this result. As presented in Table 4.2.1, predicting progression in a binary-classification task is more complex as the moment of progression is further away into the future. A possible explanation is that that the decision boundaries, especially when the moment of progression is further away into the future, are not sufficient for predicting the moment of progression.

As stated in Subsection 2.1, the course of *AD* progresses slowly. Therefore, another possible explanation is that MRI biomarkers are not sensitive enough to make a distinction among the relatively close time intervals chosen in the current study. This idea is supported by the *post hoc* analysis. When one adjusts the time intervals in a *post hoc* analysis, the accuracy increases, as can be seen in Table 4.4.6. Combining the *progression6MCI*, *progression12MCI* and *progression24MCI*, as well as *progression30MCI* and *progression48MCI* classes gives the highest F1 score (0.79). This suggests that predicting the moment of progression is possible, but only when broader time intervals are chosen. Further research is needed to confirm this idea.

Therefore, the answer to the follow-up questions is as follows: because the SVC did not perform better than a Dummy classifier, the SVC is insufficient for making a multiclass distinction among the different *progressionMCI* classes. Therefore, the classifier cannot predict when a subject will progress to *AD* with time intervals chosen for the current study. Thus, the results of experiment 3 must be interpreted with caution, as the SVC performed significantly better than the Dummy classifier because it could only classify *stableMCI* well. Investigating the extent to which predicting the moment of progression it is possible with other time intervals is beyond the scope of the current study.

5.2 The Answer to the Problem Statement

The current study addressed the problem statement: *To what extent can a classifier predict the progression of subjects from MCI to AD and the progression's corresponding moment, based on MRI biomarkers?* The answer to the problem statement is as follows. The SVC can predict progression and its corresponding moment better than the Dummy classifier. The SVC performs better at predicting progression when the moment of progression is in the near future. However, this finding should be interpreted with caution because the classifier is insufficient for predicting the moment of progression.

In practice, when this model is applied on a subject - from the ADNI database or relatable database; when a subject is in *stableMCI*, the chance that this model predicts this class right is 48%. This means that the chance is 52% that the model will predict that the subject will progress to *AD*, whereas the subject will remain stable. This is an example of a type I error. In terms of healthcare costs, type I errors are expensive because medicines and treatment would be provided when this is unnecessary. Since this will happen more than half of the time, these results indicate that it is not a sufficient model in terms of reducing healthcare costs. By contrast, given the fact that a random subject is predicted to be in the *stableMCI group*, this model is right 78% of the time. This means that the chance is 22% that the classifier predicts that a subject will not progress to *AD* when the subject will actually progress. This is an example of a type II error. In healthcare-cost terms, the cost of type II error is lower than the cost of a false alarm (type I error). However, a type II error leads to inappropriate and inadequate treatment of both the subject and his or her disease. Even though the percentages of both type I and type II errors seem to be already high for the *stableMCI* class, type I and type II errors of the other classes are even higher: 67% and 53% for *progression6MCI*, 83% and 90% for *progression12MCI*, 77% and 90% for *progression24MCI*, 80% and 89% for *progression30MCI* as well as 83% and 93% for *progression48MCI*, respectively.

With such high error rates, this model should not be used in practice. As stated in Subection 1.3, no disease-modifying therapy has been found for *AD*, which could be due to the fact that the moment of progression to the disease has not yet been predicted. With this model, the moment of progression still cannot be predicted. As stated in Subsection 1.4, *AD* is an illness with one of the highest costs of health care. With this model, the healthcare cost would increase even more because many subjects would be treated for *AD* when they do not actually suffer from the disease. Besides that, many subjects who are predicted to remain stable are likely to develop *AD* later on and receive inappropriate treatment.

Therefore, the conclusion is that the model can predict progression to *AD* better than the Dummy classifier. However, this is in all likelihood a consequence of the *stableMCI* class. Because of the high rates of type-I error and type-II error for all classes, the performance is insufficient for clinical application.

5.3 Limitations

The findings of the current study are subject to at least three limitations. The first limitation concerns the generalizability. The generalizability of these results to all *MCI* subjects worldwide is limited. The population consisted of North Americans. There could also be variances within the population of North America as compared to those of Europe and other continents.

The second limitation concerns the approach to assigning subjects to the *stableMCI* class. Even though another approach was taken for assigning subjects to the *stableMCI* class in comparison with the approach by Westman, Muehlboeck, and Simmons (2012), to be sure that the subjects assigned to *stableMCI* are stable during the complete study, one cannot guarantee that *stableMCI* subjects will not progress to *AD* after the current study. These subjects could already have more abnormal volumes of some brain structures, which could cause noise in the data. As a result, the decision boundaries for a classifier to be found are more complex. If it was known that a subject converted to *AD* after the current study, the subject would be removed from the *stableMCI* class. However, not knowing what happens to a subject after the study is a limitation of almost all observational studies.

The third limitation concerns the number of subjects in the test set. Due to the limited number of data available for the current study, the number of subjects in each class was small. As a consequence, the number of subjects in the test set was also small. For the classes in which the moment of progression is further away in time, such as *progression30MCI* and *progression48MCI*, there were less than 10 subjects in the test set. As a result, the classifier's performance on these classes is not sufficiently reliable.

5.4 Future Research

The current study is expected to serve as a gateway to a new research field within the *AD* prediction, where not only the question '*if*' a subject's progression to *AD* but also '*when*' this progression is likely to take place will be answered. Three directions are proposed in which further research can improve.

The first direction of research concerns the input object. For the current study, six features from MRI scans are included. In the current study, no investigation has been done on the effect of an increase in the number of features on performance, nor has the effect of choosing another method, such as PET scan or lumbar puncture, on performance been investigated. Since MRI biomarkers are not sensitive enough for the short time intervals chosen in the current study, research into the other feature combination for multiclass classification could provide insight into the possibility of predicting the moment of progression with short time intervals.

The second direction of research concerns the learning algorithms used in the current study. In Subsection 3.6.1, algorithms were trained of which was known that they could classify *MCI* converters

from stable *MCI* or *AD* from *MCI*. In addition to DTC, LSVC, LR, PER, SGD and SVC, other learning algorithms could be tested. There are numerous unexplored learning algorithms when it comes to *AD* prediction. Besides this, since the DTC and the SVC performed well in Experiment 1, investigating similar models, such as the Decision-Tree Forrest, is recommended.

The third direction of research concerns the time intervals for *progressionMCI* classes. As the results indicate, the SVC is insufficient for making a multiclass distinction among *progression6MCI*, progresion12*MCI*, *progression24MCI*, *progression30MCI* and *progression48MCI*. The results of the follow-up experiment indicate that this could be due to the time intervals used in the current study. Research into predicting the moment of progression with broader time intervals could be of great value for the early detection of *AD*. Based on the confusion matrix in Subection 4.4, making a distinction between subjects who progressed within two years and those who progressed over two years is recommended.

The fourth direction of research concerns oversampling and undersampling to the training set with multiclass classification. Section 2.5 stated that most solutions for oversampling and undersampling are only applicable to binary-class problems. The current study followed the SCUT algorithm proposed by Agrawal et al. (2015). Research into other solutions for applying oversampling and undersampling techniques to multiclass-classifications problems could benefit the classifier's performance.

Section 6: Conclusion

This section concludes the findings of the present study and provides recommendations for further research.

The present study was designed to investigate the extent to which it is possible to predict a subject's progression from *MCI* to *AD* and this progression's corresponding moment. The current study has indicated that the SVC performs significantly better than a Dummy classifier. The most interesting finding to emerge from the current study is that the model is able to predict progression, but the model's performance regarding predicting the moment of progression is insufficient. This seems to be due to the time intervals chosen for the moment of progression, which might be too close to each other. This suggests that the model could improve if broader time intervals for the moment of progression were chosen.

Even though the model is insufficient for practical application, the present study makes several noteworthy contributions to the field of *AD* prediction. This is the first study that investigated multiclass classification while predicting the moment of progression. However, the current study did not succeed in indicating whether it is possible to predict the moment of progression. The current study offers some insight into the importance of selecting the right time intervals, which contributes to the process of finding better models in the future that could eventually be used in practice.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

ADNI. (2017). Sharing Alzheimer's Research Data with the world. Retrieved from http://adni.loni.usc.edu/

Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2015). Classification with class imbalance problem: A Review. *Int. J. Advance Soft Compu. Appl*, *7*(3).

Allison, P. D. (2012). *Missing data*. Thousand Oaks, CA: Sage.

Alzheimer's Association. (2016). Alzheimer's Disease Facts and Figures. *Alzheimer's & Dementia 2016*, *12*(4).

Agrawal, A., Viktor, H. L., & Paquet, E. (2015). SCUT: Multi-class imbalanced data classification using SMOTE and cluster-based undersampling. *Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on* (1), 226-234.

American Psychiatric Association. (2013). Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). Retrieved from https://books.google.nl

Bala, M., Agrawal, R.K. (2010). Kernel Parameter Selection for SVM Classification: Adaboost Approach. *Strategic Pervasive Computing Applications: Emerging Trends*. Retrieved from https://books.google.nl

Bauer, K. A., Rosendaal, F. R., & Heit, J. A. (2002). Hypercoagulability: too many tests, too much conflicting data. ASH Education Program Book, 2002(1), 353-368.

Billsus, D., & Pazzani, M. J. (1998). Learning Collaborative Information Filters. *Icml*, 98, 46-54.

Bombois, S., Duhamel, A., Salleron, J., Deramecourt, V., Mackowiak, M. A., Deken, V., ... & Schraen-Maschke, S. (2013). A new decision tree combining Abeta 1-42 and p-Tau levels in Alzheimer's diagnosis. *Current Alzheimer Research*, *10*(4), 357-364.

Borowska, M., Topczewska, M. (2016). New Data Level Approach for Imbalanced Data Classification Improvement. In Proceedings of the 9th International Conference on Computer Recognition. Retrieved from https://books.google.nl

Burns, A., Page, S., Winter, J. (2005). Diagnosis and assessment of dementia. In Alzheimer's Disease and Memory Loss Explained. Retrieved from https://books.google.nl

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd international conference on Machine learning*,161-168.

Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, 853-867. Springer US.

Chupin, M., Gérardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehéricy, S. and Alzheimer's Disease Neuroimaging Initiative. (2009). Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. *Hippocampus*, *19*(6), 579.

Devanand, D.P. Pradhaban, G., Liu, X., Khandji, A., De Santi, S., Segal, S., Rusinek, H., Pelton, G.H., Honig, L.S., Mayeux, R., Stern, Y., Tabert & M.H. de Leon, M.J. (2007). Hippocampal and entorhinal atrophy in mild cognitive impairment prediction of Alzheimer disease. *Neurology*, *68*(11), 828-836.

Douglas, J. (1995). MRI-based measurement of hippocampal volume in patients with combat-related posttraumatic stress disorder. The American journal of psychiatry, 152, 973-998.

Dubois, B., Feldman, H. H., Jacova, C., Cummings, J. L., DeKosky, S. T., Barberger-Gateau, P., ... & Gauthier, S. (2010). Revising the definition of Alzheimer's disease: a new lexicon. *The Lancet Neurology*, *9*(11), 1118-1127.

Jack, C. R., Barkhof, F., Bernstein, M. A., Cantillon, M., Cole, P. E., DeCarli, C., ... & Hampel, H. (2011). Steps to standardization and validation of hippocampal volumetry as a biomarker in clinical trials and diagnostic criterion for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(4), 474-485.

Jack, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., ... & Trojanowski, J. Q. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*, *9*(1), 119-128.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, *6*(5), 429-449.

Joshi, S., Simha, V., Shenoy, D., Venugopal, K. R., & Patnaik, L. M. (2010). Classification and treatment of different stages of alzheimer's disease using various machine learning methods. International Journal of Bioinformatics Research, 2(1), 44-52.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, *30*(1), 25-36.

Li, H., Liu, Y., Gong, P., Zhang, C., Ye, J., & Alzheimers Disease Neuroimaging Initiative. (2014). Hierarchical interactions model for predicting Mild Cognitive Impairment (*MCI*) to Alzheimer's Disease (*AD*) conversion. *PloS one*, *9*(1), e82450.

Hao, M., Wang, Y., & Bryant, S. H. (2014). An efficient algorithm coupled with synthetic minority over-sampling technique to classify imbalanced PubChem BioAssay data. *Analytica chimica acta*, *806*, 117-127.

Hoens, T.R., Qian, Q., Chawla, N. V., Zhou, Z. (2012). Building Decision Trees for the Multi-class Imbalance Problem. Advances in Knowledge Discovery and Data Mining. Retrieved from https://books.google.nl

Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, *13*(2), 415-425.

Huang, M., Yang, W., Feng, Q., Chen, W., & Alzheimer's Disease Neuroimaging Initiative. (2017). Longitudinal measurement and hierarchical classification framework for the prediction of

Alzheimer's disease. *Scientific reports*, *7*.

Martínez-Murcia, F. J., Ortiz, A., Górriz, J. M., Ramírez, J., & Illán, I. A. (2015). A volumetric radial LBP projection of MRI brain images for the diagnosis of Alzheimer's disease. *International Work-Conference on the Interplay Between Natural and Artificial Computation*, 19-28.

Mashayekhi, M., & Gras, R. (2014). Investigating the effect of spatial distribution and spatiotemporal information on speciation using individual-based ecosystem simulation. *GSTF Journal on Computing (JoC)*, *2*(1).

Mitchell, A. J., & Shiri-Feshki, M. (2009). Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies. *Acta Psychiatrica Scandinavica*, *119*(4), 252-265.

Moini, J. (2015). Anatomy and Psysiology for Health Professionals. Degeneration of the brain. Retrieved from https://books.google.nl

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., Trojanowski, J.Q., Toga, A.W. & Beckett, L. (2005). Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's & Dementia*, *1*(1), 55-66.

Murty, M. N., Raghava, R. (2016). Support Vector Machines and Perceptrons: Learning, Optimization, Classification, and Application to Social Networks. India. SpringersBrief in Computer Science. Retrieved from https://books.google.nl

Orešič, M., Hyötyläinen, T., Herukka, S. K., Sysi-Aho, M., Mattila, I., Seppänan-Laakso, T., ... & Kivipelto, M. (2011). Metabolome in progression to Alzheimer's disease. *Translational psychiatry*, *1*(12), e57.

Raykar, V. C., Saha, A. (2015). Data split Strategies for Evolving Predictive Models. In Machine Learning and Knowledge Discovery in Databases: European part I. Retrieved from https://books.google.nl

Sarraf, S., Anderson, J., & Tofighi, G. (2016). DeepAD: Alzheimer′ s Disease Classification via Deep Convolutional Neural Networks using MRI and fMRI. *bioRxiv*, 070441.

Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). On the stratification of multi-label data. *Machine Learning and Knowledge Discovery in Databases*, 145-158.

Singh, R., & Ade, R. R. (2015). Review on Class Imbalance Learning: Binary and Multiclass. *International Journal of Computer Applications*, *131*(16), 4-8.

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., ... & Park, D. C. (2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia*, *7*(3), 280-292.

Sun, Y., Kamel, M.S., Wang, Y. (2006). Boosting for learning multiple classes with imbalanced class distribution. *IEEE ICDM: Proceedings of the Sixth International Conference on Data Mining*, *6*, 592-602.

Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence, 23*(04), 687-719.

Tan, L. P., & Wong, K. Y. (2017). A Neural Network Approach for Predicting Manufacturing Performance using Knowledge Management Metrics. *Cybernetics and Systems, 48*(4), 348-364.

Trzepacz, P. T., Yu, P., Sun, J., Schuh, K., Case, M., Witte, M. M., ... & Alzheimer's Disease Neuroimaging Initiative. (2014). Comparison of neuroimaging modalities for the prediction of conversion from mild cognitive impairment to Alzheimer's dementia. Neurobiology of aging, 35(1), 143-151.

Thung, K. H., Wee, C. Y., Yap, P. T., Shen, D., & Alzheimer's Disease Neuroimaging Initiative. (2014). Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion. *NeuroImage, 91*, 386-400.

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Cedarbaum, J., ... & Luthman, J. (2015). 2014 Update of the Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. *Alzheimer's & dementia, 11*(6), e1-e120.

Weiss, G. M., McCarthy, K., & Zabar, B. (2007). Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?. *DMIN*, *7*, 35-41.

Westman, E., Muehlboeck, J. S., & Simmons, A. (2012). Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion. *Neuroimage*, *62*(1), 229-238.

Wolz, R., Heckemann, R. A., Aljabar, P., Hajnal, J. V., Hammers, A., Lötjönen, J., ... & Alzheimer's Disease Neuroimaging Initiative. (2010). Measurement of hippocampal atrophy using 4D graph-cut segmentation: application to ADNI. *NeuroImage*, *52*(1), 109-118.

World Alzheimer Report. (2015). The Global Impact of Dementia: an analysis of prevalence, incidence, cost and trends. *Alzheimer's Disease International*.

Appendices

Appendix A: Comparing F1 scores of pre-processing methods for each classifier

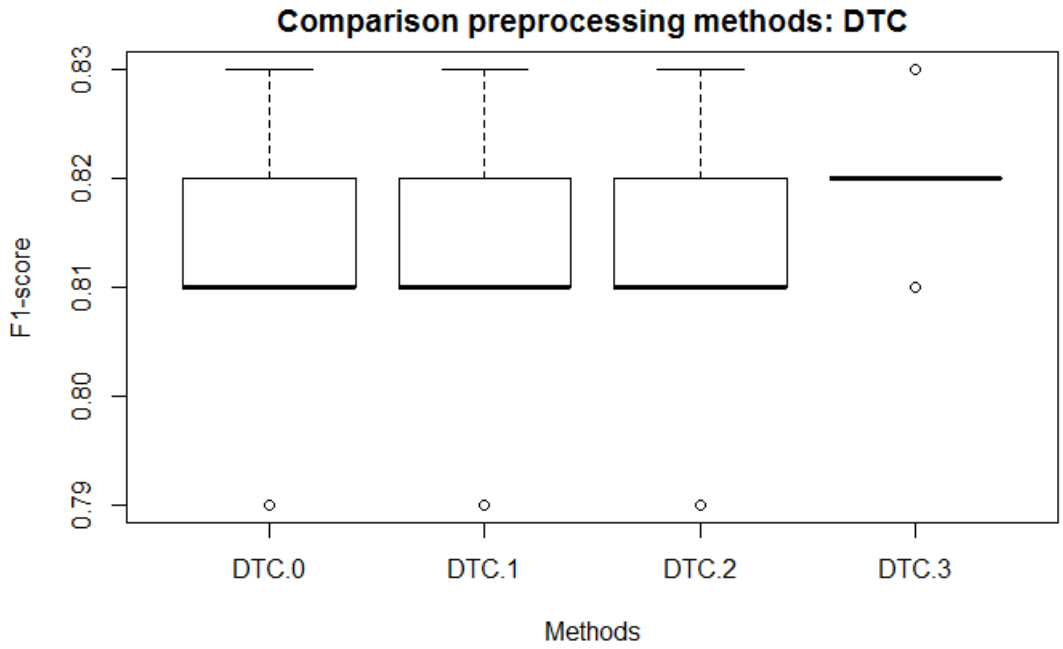


Figure A.1. Comparing F1 scores of pre-processing methods for Decision-Tree classifier.

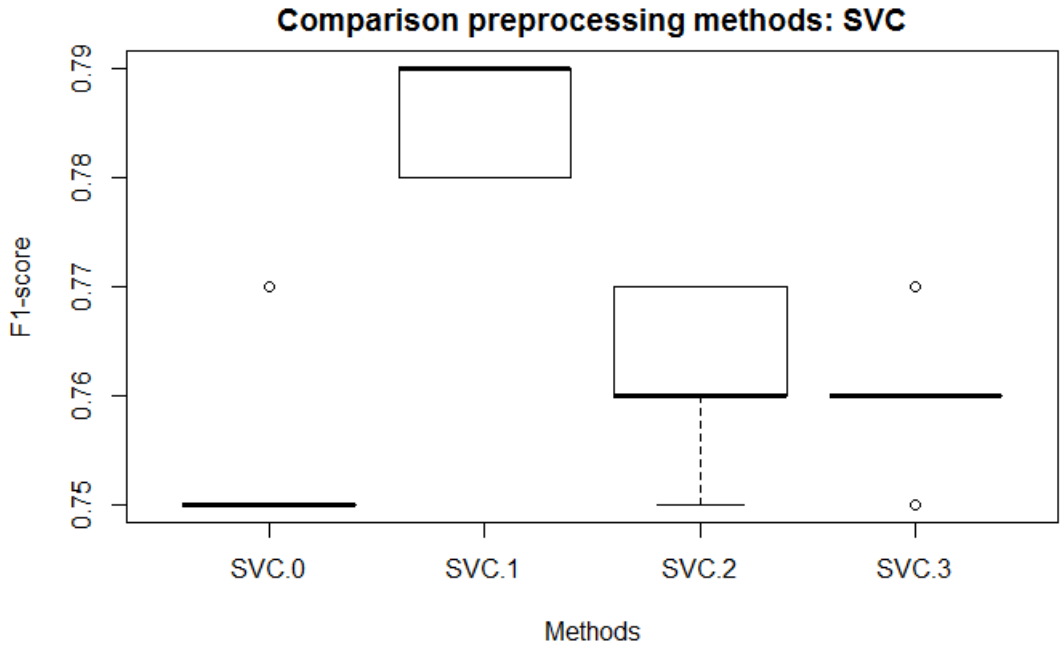


Figure A.2. Comparing F1 scores of pre-processing methods for Support-Vector classifier.

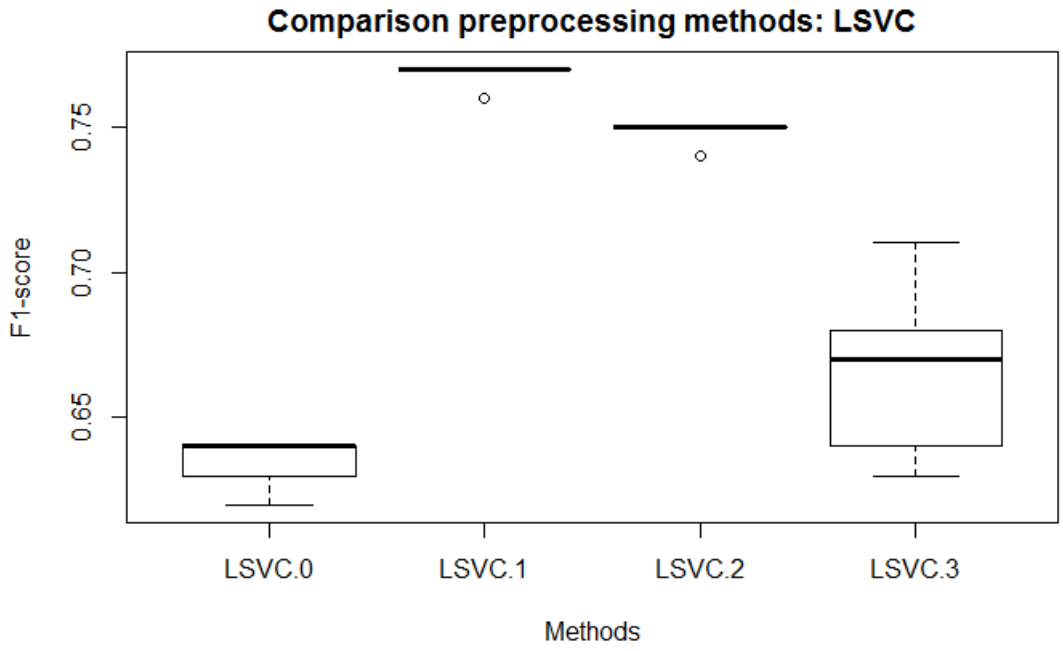


Figure A.3. Comparing F1 scores of pre-processing methods for linear Support-Vector classifier.

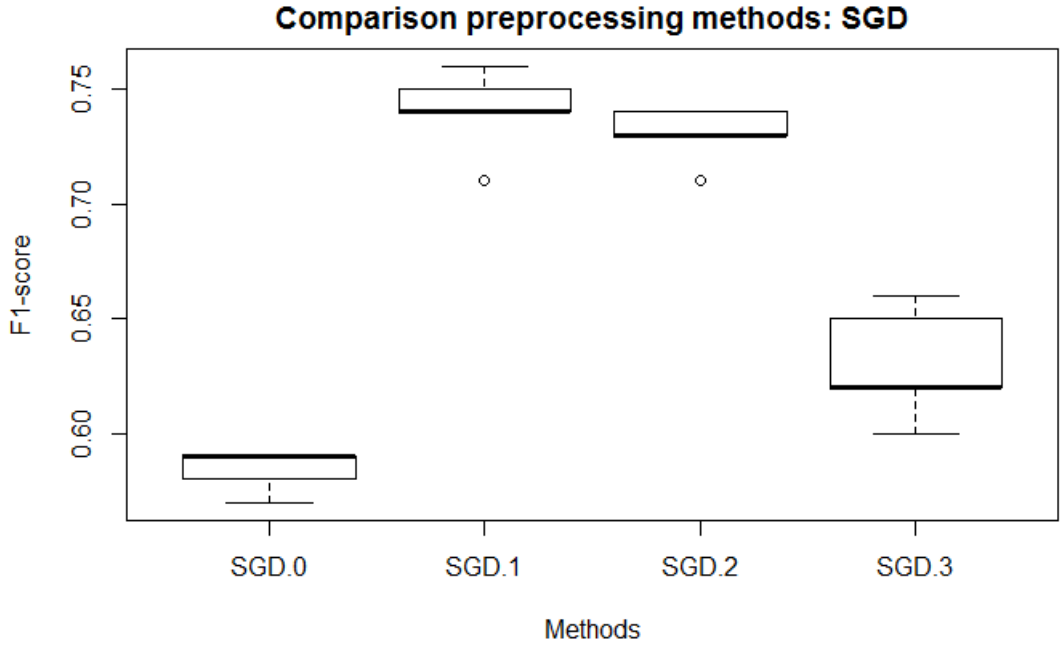


Figure A.4. Comparing F1 scores of pre-processing methods for Stochastic Gradient Descent.

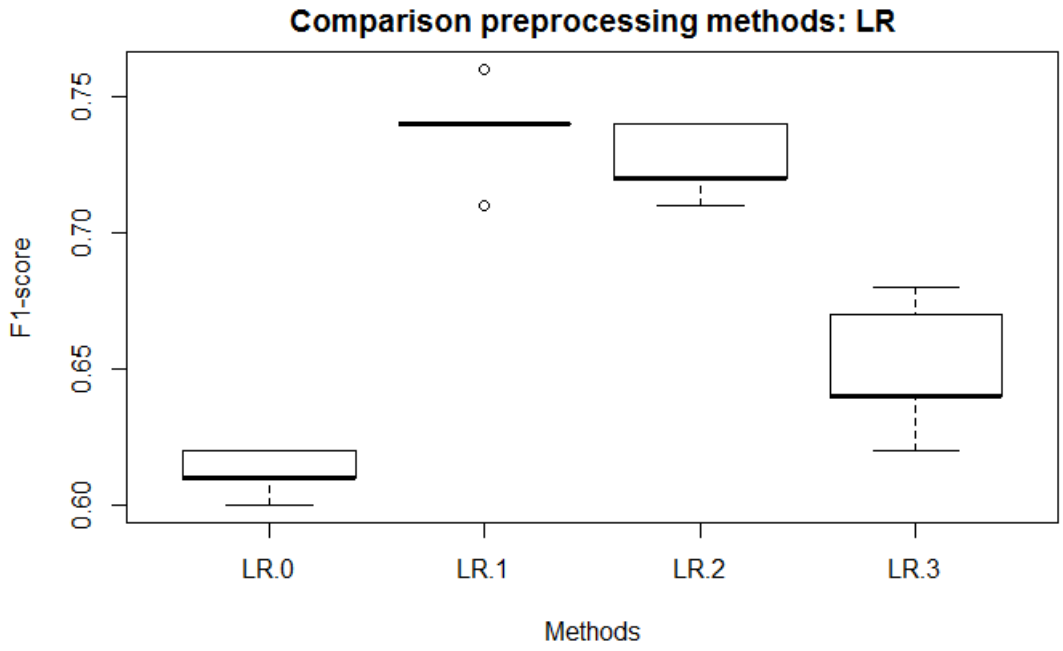


Figure A.5. Comparing F1 scores of pre-processing methods for Logistic Regression.

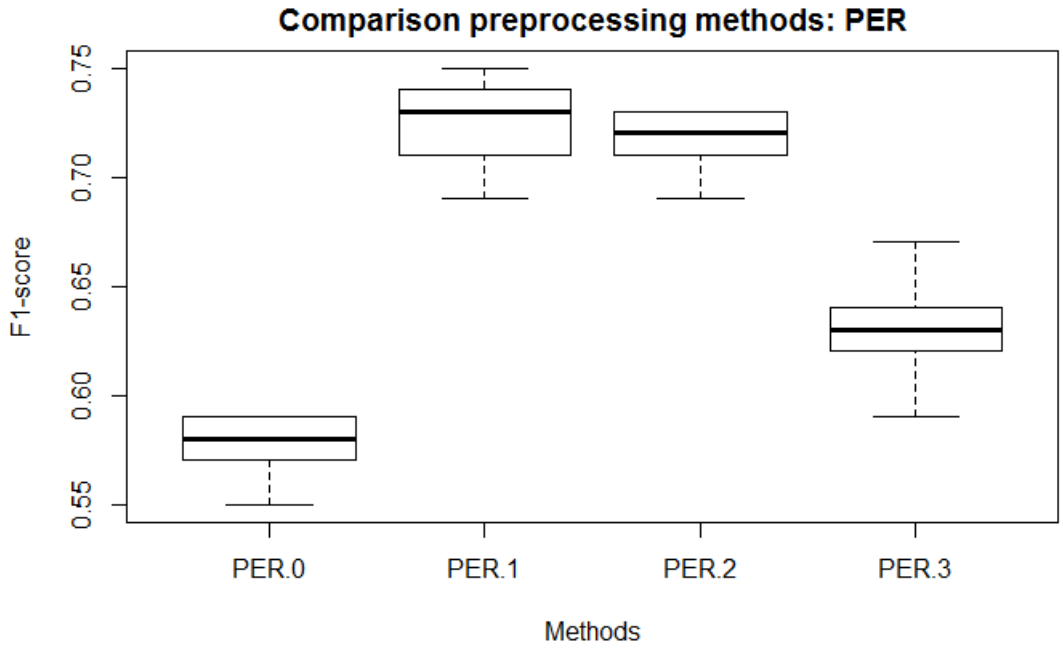


Figure A.6. Comparing F1 scores of pre-processing methods for Perceptron.

Appendix B: Contingency Table needed for McNemar test

Table B.1
Contingency Table for RQ3 (significant)

	Classifier A: wrong	Classifier A: good
Classifier B: wrong	79	36
Classifier B: good	10	13

Classifier A: Support-Vector classifier, Classifier B: Dummy classifier

Table B.2
Contingency Table for Follow-Up Question (not significant)

	Classifier A: wrong	Classifier A: good
Classifier B: wrong	42	9
Classifier B: good	17	3

Classifier A: Support-Vector classifier, Classifier B: Dummy classifier