Supervised Machine Learning in Multiple Sclerosis

Applications to Clinically Isolated Syndromes

Viktor Wottschel

A dissertation submitted in partial fullfillment of the requirements for the degree of

Doctor of Philosophy

of the

University College London

Department of Neuroinflammation

University College London

2017

Supervisors: Prof. Olga Ciccarelli
Prof. Daniel Alexander

Examiners: Dr. Jonathan Clayden
Prof. Stefan Klöppel

Declaration

I, Viktor Wottschel, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Amsterdam, The Netherlands, April 2017

Viktor Wottschel

Acknowledgements

This work would not have been possible without the help of many people. I will mention a few of them in the following paragraphs but I am very thankful to everyone who had an impact on my academic life.

I would like to thank my supervisors Olga Ciccarelli and Daniel Alexander. Danny gave me the opportunity to come to UCL for a MSc project in collaboration with Olga. Their subsequent encouragement and help with my PhD application were invaluable and brought me to where I am now. Olga and Danny provided great supervision and guidance whilst giving me a lot of freedom to pursue my own ideas.

I would like to thank everyone in the NMR Research Unit for providing a great work environment in the past four years. I truly enjoyed being part of this group and I am grateful for all the discussions, meetings, conferences and social events we had. Special thanks go to Francesco for being a great friend throughout the entire PhD.

I would like to thank everyone in the Microstructure Imaging Group as well as the Progression Of Neurodegenerative Disease initiative for the many fruitful discussions as well as for being great company outside the office. Special thanks to the MAPS committee for making sure we don't take life too serious.

I would like to thank the many people at the Translational Imaging Group who developed and develop great tools for image processing and were of great help to me trying to understand some of the underlying concepts of the imaging world.

I would like to thank the MAGNIMS research group for providing me the data for my PhD experiments and the discussions during the many meetings.

I would like to thank Ifrah simply for being a wonderful person and great influence in my life. Without your constant motivation this project probably would never have finished.

Vielen Dank an meine Familie. Meine Eltern Alwira und Andrei haben mich während meiner akademischen Laufbahn immer unterstützt, auch wenn das bedeutete, dass ich in den letzten fast 10 Jahren nur noch sporadisch zu Hause war. Vielen Dank auch an meine Schwester Olga und meinen Schwager Michael, die mir immer einen Grund gegeben haben nach Dortmund zu kommen. Umso mehr noch seit meine Nichte Greta das Licht der Welt erblickt hat.

Abstract

Multiple sclerosis (MS) is an inflammatory, demyelinating disease that can cause various neurological symptoms. The first episode of this disease is called a clinically isolated syndrome (CIS) and leads to the diagnosis of MS in the majority of patients in the long-term. Fast conversion from CIS to MS is associated with higher disability and more severe disease progression so that it is of high clinical interest to identify risk patients that will convert to MS within a short time. Several risk factors for conversion have been identified but they can only be applied on cohort levels.

In this thesis we provide an overview of supervised machine learning approaches that can be used to distinguish individual CIS-stable patients from those who will experience a second attack within one to five years and consequently will be diagnosed with clinically definite MS. This classification is based on information available at baseline derived from routine MRI scans and complemented by clinical information such as lesion masks, age, gender, disability and CIS type of onset.

We introduce the classification landscape, an overview of supervised classification studies with respect to their method and task complexity, and show that our experiments cover a large range of feature complexities in this landscape for the rather complex task of outcome prediction in CIS patients.

We show that low-level voxel-based information such as tissue density of grey and white matter are not informative and lead to inconclusive results, whereas the introduction of high-level features such as lesion load, age, gender or disability improves accuracies to 71.4 % and 68 % at one- and three-year follow-up respectively in a single-centre data set. Finally, we propose a recursive feature elimination method that is able to identify specific regions that are relevant with respect to disease progression in MS and achieves accuracies of 73.9 % and 74.3 % at one- and three-year follow-up respectively even in a multi-centre setting.

CONTENTS

Part I

GENERAL INTRODUCTION

1

MOTIVATION

Multiple sclerosis (MS) is a chronic inflammatory disease of the central nervous system (CNS) characterised by brain and spinal cord lesions which can cause a broad range of neurological symptoms depending on their location in the brain. The majority of MS patients has an onset with a clinically isolated syndrome (CIS) which is the diagnosis given after a single neurological episode that is caused by demyelination or inflammation and lasts for at least 24 hours. After a second attack (or relapse) the patient is diagnosed with clinically definite multiple sclerosis (CDMS), which is MS diagnosed using purely clinical evidence.

Most patients convert from CIS to CDMS over time and there are different risk factors associated with faster progression, which are described in chapter 2. Of particular interest for this work is that patients have an increased risk of poor prognosis when the time between CIS onset and relapse is short. The identified risk factors are only valid on a cohort level and cannot be applied to individual patients with high confidence.

Machine learning, or more specifically supervised classification, has shown promising results in several neurological disorders in the last decade. Healthy subjects have been successfully distinguished from patients with high accuracy and also the prognosis of outcome e.g. in patients with mild cognitive impairment has been demonstrated with results well above chance level. There have only been few studies using classification in relation to multiple sclerosis, however, and none of them considered the challenging task of predicting outcome in CIS patients.

1.1 PROBLEM STATEMENT

Given a data set consisting of information from a standard neurological examination, which supervised classification approach is suited to predict clinical outcome in patients with clinically isolated syndrome? In this context, we are not only interested in the highest possible accuracy but we want to explore the *classification landscape* using different types of features and classification approaches and compare the insight these methods provide.

1.2 AIMS

1. to develop multivariable models to predict conversion from CIS to CDMS at different follow-up ranges and assess the model performances.
2. to identify features relevant for disease progression in CIS.

1.3 SUMMARY OF CONTRIBUTIONS

1. Introduction of the classification landscape: an overview of machine learning studies based on the complexity of the applied methods and classification tasks as shown in section 1.4.
2. Comprehensive set of machine learning studies on the prediction of second relapse in CIS patients: we place our studies in the proposed landscape and compare the classification performance of the different approaches. We show in chapter 4 that low-level voxel-based features are not able do discriminate between CIS-stable patients and CIS-converters. High-level features are expensive to obtain because they depend on human input and clinical expertise. However, they significantly improve the classification performance as shown in chapters 5 and 6. Medium-level features such as regions of interest containing information derived from Magnetic resonance Imaging (MRI) provide only little insight due to the lack of sensitivity and specificity in the prediction as shown

in chapter 7. Our final model combines medium- and high-level features and is able to identify features known to be relevant for progression in MS, in particular atrophy-related measures in deep grey matter and insula, and uses local information to provide future diagnosis in individual patients with an accuracy of 73.9 % and 74.3 % at one- and three-year follow-up respectively in a multicentre setting using 296 patients.

1.4 CLASSIFICATION LANDSCAPE

The application of machine learning classifiers on neuroimaging data to solve neurological problems has become increasingly popular over the last years. One of the first applications was the automated diagnosis of Alzheimer's disease (AD) in a cohort of cognitively normal subjects (CN) and AD patients [64]. Most publications focus on dementia and only few publications are exploring classifications tasks related to other diseases such as Huntington's disease or multiple sclerosis.

The models in these studies can be roughly divided by complexity of the applied methods and by complexity of the classification task. In this case, methods covers the classifier complexity as well as the applied type of feature. Using voxel information from MR intensities or derived measures such as grey matter density can be considered low-level features as they are directly available from the imaging data. Region-based measures rely on atlases and are considered to have mid-level complexity, and finally we define high-level features as measures that need a person (usually an expert clinician) to obtain them such as clinical scores or manually outlined lesion masks. Defining classification task complexity is slightly more ambiguous but it is generally easier to distinguish patients from healthy controls based on brain pathology than it is to differentiate different sub-types of a certain condition where changes are often more subtle. The most complex task in this context it the prediction of future outcome from baseline data.

In the following paragraphs we will describe a selection of studies in MS and AD that have used supervised classification and we attempt to place them in the proposed classification landscape.

MULTIPLE SCLEROSIS CLASSIFICATION TASKS Only few publications are available on classification tasks related to MS which might be due to the lack of publicly available data sets. This also makes it challenging to compare different studies with each other as they generally use different cohorts where sample size, age, gender ratio and disease stage vary.

The first paper using Support Vector Machines (SVM) to perform binary classification in MS was published by Bendfeldt et al. [7] who looked at three different classification tasks: 1. patients with short disease duration ($<$5 y) vs long disease duration ($>$10 y), 2. low T2 lesion load ($<$1 ml) vs high T2 lesion load ($>$10 ml), and 3. benign MS ((Expanded Disability Status Scale (EDSS) $\leqslant$ 3) vs non-benign MS (EDSS $>$ 3).

The classes in each classification task were balanced in size but rather small ranging from 13 to 20 patients per group. Grey matter segmentations were obtained from T1 scans and used as features in a linear SVM with leave-one-out cross-validation. The results for the three classification tasks were:

1. early vs late MS: 85 % accuracy, 82.3 % sensitivity, 88.2 % specificity,
2. low vs high lesion load: 83 % accuracy, 85 % sensitivity, 80 % specificity,
3. benign vs non-benign MS: 77 % accuracy, 76.9 % sensitivity, 76.9 % specificity.

The SVM kernel was used to highlight brain areas which are relevant with respect to the individual classification tasks and relate these findings to previously published studies. Both the classification tasks and the feature type (i.e. GM density) in this study can be considered of low complexity in the classification landscape as indicated by study (a) in Figure 1.

Weygandt et al. [129] presented a method to distinguish 44 relapsing-remitting MS (RRMS) patients (mean disease duration 80 months $\pm$ 76.3) from 26 healthy controls (HCs) using linear SVMs. Different areas of the brain were examined using a searchlight approach where neighbourhoods around a voxel of interest are explored. Masks were created for normal appearing WM (NAWM), normal appearing GM (NAGM), normal appearing brain tissue (NABT) and lesion tissue (LT) to be applied to T2 MRI scans after registration to MNI space. The highest accuracy of 95 % was obtained using lesion areas of the brain, which is expected as healthy sub-

jects generally do not have lesions and hence the presence of lesions can be used to correctly classify a large proportion of subjects. However, also the NAGM and NAWM maps provided very high accuracies leading to 84 % and 91 % respectively with a p-value of at 10^{-7} or lower. The searchlight approach using the whole NABT did not lead to a statistically significant finding. Particularly informative areas have been reported to be deep GM nuclei and the cerebellum; in the WM areas of dirty white matter were found to be of interest. Although the high accuracies are very encouraging, it must be noted that this classification task is not very sophisticated as MS patients show substantial changes in brain MRI compared to healthy controls. Lesions are much more frequent in MS patients than in healthy controls and also normal appearing tissue can be affected by microstructural changes [109]. The classification task in this study is considered to be of low complexity but the searchlight approach, even though using MRI intensities, is more sophisticated (see study (b) in Figure 1).

Another study tested the performance of the different classifiers VDC (voxel-wise displacement classifier; a classifier based on Fisher's linear discriminant analysis), SVM, Random Forest (RF) and Adaboost when using displacement fields as features [19]. Twenty-nine RRMS, 8 secondary-progressive MS (SPMS), 4 CIS, 1 primary-progressive (PPMS) patients and 36 healthy controls were registered to a study-specific template. The resulting transformation functions or displacement fields describe how much every voxel was morphed to match the template image. This information is downsampled to reduce the dimensionality of features and then applied to the classifiers in order to distinguish young from old subjects (5/11 youngest vs 5/11 oldest) and MS patients (SPMS/RRMS/all) from healthy controls. The proposed VD classifier consistently outperformed the other methods reaching up to 100 % in the 5 vs 5 age classification. The performance was particularly good in the small cohorts for age and SPMS classification (88-100 %) but also reached 76 % in the case of all MS vs HC. It should be noted that the study has a clear focus on promoting the novel VD classifier. The rather low sample sizes, however, do not allow for generalisation of the results. This study is placed as study (c) in Figure 1 due to the medium

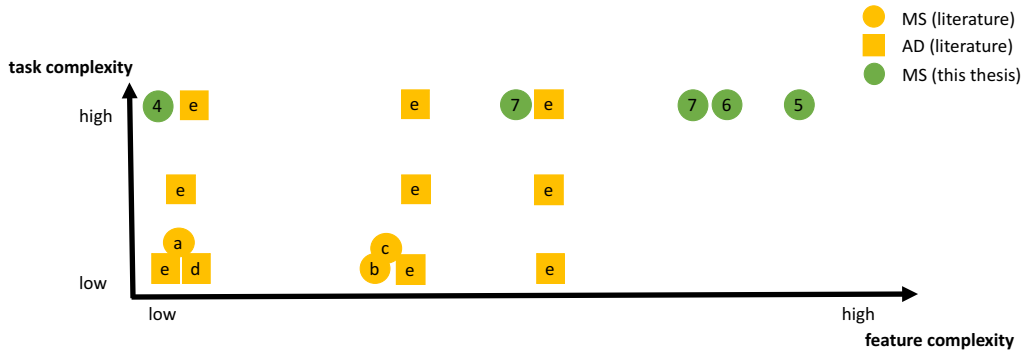


Figure 1: Classification landscape. Overview of existing work and contributions made in this thesis. Letters in yellow squares and circles refer to studies described in this section. Numbers in green circles refer to chapters of this thesis.

complexity of the displacement-voxel-based features and the low complexity of the classification task.

DEMENTIA-RELATED CLASSIFICATION TASKS By far the most neuroimaging classification publications focus on Alzheimer's disease. This has partially to do with the increasing amount of funding for AD research as it is becoming more and more common in ageing societies. Another important factor is the ADNI (AD neuroimaging initiative) dataset, which is publicly available since 2003 and includes imaging (i.e. MRI, PET) and clinical data (i.e. genetic markers, cognitive test scores) of over 1000 subjects which are clinically diagnosed as HC (or CN), mild cognitive impairment (MCI) or AD. The data was obtained at different centres but standardised protocols have been put in place to reduce data heterogeneity. Follow-up data exists in irregular intervals for the majority of the subjects [86]. Early work in AD classification focused on the differentiation between HC and AD (and sometimes MCI). Modern research in this fields targets more predictive tasks using diagnoses from follow-up examinations. In the context of AD, there is particular interest in the transition from MCI to AD so that researchers try to differentiate between MCI-stable patients and MCI-converters. We describe only a few studies here to provide an overview of the methods and how the approaches are distributed in the classification landscape.

Klöppel et al. [64] were one of the first groups to perform SVM classification on structural MRI data. The group used a linear SVM to distinguish AD patients from

HC and AD patients from patients with frontotemporal lobar degeneration (FTLD). In total they performed 7 classification experiments on different data sets in order to show generality of their approach:

1. 20 AD patients vs 20 HC, patients' diagnosis was histopathologically confirmed
2. 14 AD patients vs 14 HC, different centre from 1, patients' diagnosis was histopathologically confirmed
3. 1. and 2. combined
4. train classifier on 1 and test on 2
5. train classifier on 2 and test on 1
6. 33 AD patients vs 57 HC, clinical diagnosis of probable early AD for all patients
7. 18 AD patients vs 19 FTLD patients, AD patients are from same cohort as 2 and FTLD were diagnosed clinically based on consensus criteria.

T1-weighted MRI was obtained for all subjects and segmented into WM, GM and CSF using SPM5. The images were then registered to a study-specific template created from groups 1 and 2. Only the grey matter areas of the brain were used for classification and a leave-one-out cross-validation was used. The obtained accuracies from task 1 and 2 are 95 % and 92.9 % with sensitivity/specificity of 95/95 % and 100/85.7 % respectively. A combination of data sets yielded an accuracy of 95.6 % and the centre exchange in task 4 and 5 resulted in accuracies of 96.4 % and 87.5 % respectively. The classification of mild AD and controls correctly assigned 81.1 % of the subjects to the correct group when using the whole brain's white matter. A restriction to medial temporal lobe improved the result to 85.6 %/. Finally, 89.2 % of the cases in task 7 were correctly classified. The method is completely automated and does not need any manual intervention except for quality control after registration and segmentation steps. The authors report that the obtained accuracies are better than the diagnostic accuracy of clinicians using standardised criteria. This study uses rather simple features (i.e. GM density) and only compares patients and

controls leading to a low complexity on both axes of our proposed classification landscape (see study (d) in Figure 1).

Cuingnet et al. [28] offer a review that compares a large variety of approaches to different classification tasks using structural information from the ADNI data set. This is of particular interest because even though most studies use ADNI data, they usually only use a subset of it which then differs between studies. The review, however, repeats all experiments as explained in the respective papers using the same set of patients. All patients with available preprocessed MRI scans were included in the review study leading to a cohort of 509 subjects from 41 centres with 162 cognitively normal subjetcs, 137 AD patients, 76 MCI patients who convert to AD within 18 months (MCI converter or MCIc), and 134 MCI patients who did not convert within 18 months (MCI stable or MCIs). The presented methods include whole-brain voxel-based approaches, parcellation and selection of ROIs, different registration methods, cortical thickness measures, and hippocampal features. All tested algorithms were tested on three classification tasks:

1. AD vs HC, 2. MCI vs HC, and 3. MCI stable vs MCI converter.

All methods perform very well on task 1 reaching balanced accuracies (= mean of sensitivity and specificity) between 74 and 88 %. The comparison shows that rather simple features such as GM probabilities (whole brain or mean over ROIs) perform significantly better than more sophisticated features like measures of hippocampal volume or shape. Accuracies for task 2 are noticeably lower ranging between 72.5 and 81.5 % as the differences between the two groups are smaller. Task 3 is known to be the most challenging because patients from both classes have the same diagnosis at the time of the MRI scan and the classification aims to predict future outcome. In line with this, the obtained accuracies again are lower and, in multiple approaches, even assign all patients in to one single group (i.e. sensitivity = 0 %, specificity = 100 %) leading to a balanced accuracy of 50 % which is equivalent to a random assignment. The other methods' accuracies range between 56.6 and 67.5 %. This review article also attempts to group the different approaches by their feature types but does not follow the same order as the proposed classification landscape. In our overview where the classification tasks cover low (AD vs HC), medium (MCI

vs AD), and high complexity (MCI stable vs MSI converter). However, the features are only of low and medium complexity due to the fact that no expert knowledge was needed to obtain any of them. It must be noted, though, that results from clinical and cognitive tests such as the mini-mental state examination (MMSE), which would be considered of high complexity, cannot be included as features using the ADNI cohort because they were already used to diagnose the subjects. Due to the large number of compared approaches, we place this study multiple times in the classification landscape in Figure 1 indicated by (e) at the appropriate positions.

SUMMARY It can be seen from Figure 1 that the classification landscape is well covered by literature regarding classification tasks in dementia and especially AD. The main exception is the usage of high-level features, which is due to the fact that clinical examinations and cognitive test scores are important factors for antemortem diagnosis and therefore cannot be used again to predict said diagnosis as it would create a circular argument. With regard to MS there is an obvious lack of classification studies leading to only few data points with low task complexity and covering only low- and medium-complexity features. In this thesis we present a set of experiments exploring a high-complexity task and a large range of method complexities.

Part II

BACKGROUND

2

MULTIPLE SCLEROSIS

Multiple sclerosis (MS) is an inflammatory demyelinating disease of the central nervous system (CNS) with a strong neurodegenerative component [23]. The cause of the disease is unknown but it appears predominantly in genetically susceptible patients and is triggered by combinations of environmental factors [70].

2.1 EPIDEMIOLOGY

The global prevalence of MS is estimated to be around 30 per 100,000 people but the rates show strong variations between different regions with strong gradients from equatorial regions towards the northern and southern hemispheres [81, 132]. The average incidence rate in Europe is 80 per 100,000 [132] but can go up to 200 per 100,000 in northern European populations [81]. Generally, the prevalence increases with increasing latitude so that the highest rates are observed in northern Europe (65 ° to 45 ° north), the northern United States of America, southern Canada, New Zealand and southern Australia. Areas in southern Europe, northern Australia and South America have an intermediate prevalence, and the disease rate in Africa and Asia is comparably low [72].

The disease onset is usually in peoples' late twenties or early thirties but can be observed in children and people over 50 as well. Due to its early onset, MS is considered the most common cause of disability of young adults in the developed world [23]. It affects women twice as often as men [132] and this ratio has increased in the past decades alongside a general increase of the incidence rate of MS [5, 68, 69, 74, 90].

2.2 ENVIRONMENTAL INFLUENCES ON CAUSE OF DISEASE

The correlation between latitude and MS prevalence suggests an environmental contribution to the cause of MS. This correlation has been shown for both the northern and the southern hemisphere [77, 100, 126] and it was also observed that migration from high to low latitudes reduces the risk [46]. A possible cause of this correlation might be vitamin D, which is produced in the skin after exposure to sunlight. A higher exposure to sun in the populations closer to the equator would then increase vitamin D production and reduce the risk of MS. This theory is supported by the fact that populations with vitamin-D-rich diets were shown to have a reduced risk of MS compared to other populations in similar latitudes [1]. However, clinical trials using vitamin D interventions lacked statistical power and results have been inconclusive or even negative [60, 113, 131]. Other possible causes are viral infections from e.g. the Epstein Barr virus (EBV), which has a seropositivity of 100 % in MS patients compared to 90 % in reference groups [10, 88, 127]. Tobacco smoking is associated with a higher risk of conversion from clinically isolated syndrome to clinically definite MS [33], higher EDSS, higher lesions count and load in T2 and T1 weighted MRI, a faster transition from RRMS to SPMS [52, 66] and an accelerated decline in cognitive functions [94].

2.3 GENETIC INFLUENCES ON CAUSE OF DISEASE

Several studies have examined patients' genetic susceptibility to MS by comparing familial cohorts to the non-related population. It was consistently shown that first degree relatives have an increased risk of 3-5 % of developing MS [16, 35, 85, 105]. In cases with monozygotic twins, the risk is approximately 30 % [36, 106], whilst dizygotic twins have show the same rates as other first degree relatives [57, 61, 87]. The strong difference between monozygotic twins and other first degree relatives strongly suggests that there is a polygenetic component to the disease. Various studies found gene loci and antigens that contribute to MS susceptibility, however,

these contributions are generally very small and populations in different areas of the world are affected by different areas of the genome [12, 45, 107, 108].

2.4 PATHOLOGY AND CLINICAL COURSE

As explained before, the cause of MS is not yet understood but it is believed to be influenced a combination of genetic and environmental factors that contribute to an immune attack of the nervous system [22, 23].

In MS the myelin sheaths around the axons are damaged due to a reduction of oligodendrocytes, which are cells that create this myelin. These myelin sheaths are electrical insulators that support the propagation of action potentials through the axons. A reduction of this myelin leads to a reduced signal transmission between grey matter areas and eventually to a breakdown of information flow [22, 23]. Scar-like lesions occur after repeated attacks when the body's remyelination attempts become less effective [18] (see Figure 2). These lesions and the associated immune responses are responsible for the symptoms arising from an attack. MR imaging can be used to visualise lesions as they appear as hyperintense (e.g. in T2 and PD weighting) or iso/hypointense (e.g. in T1 weighting) spots (see Figure 3).

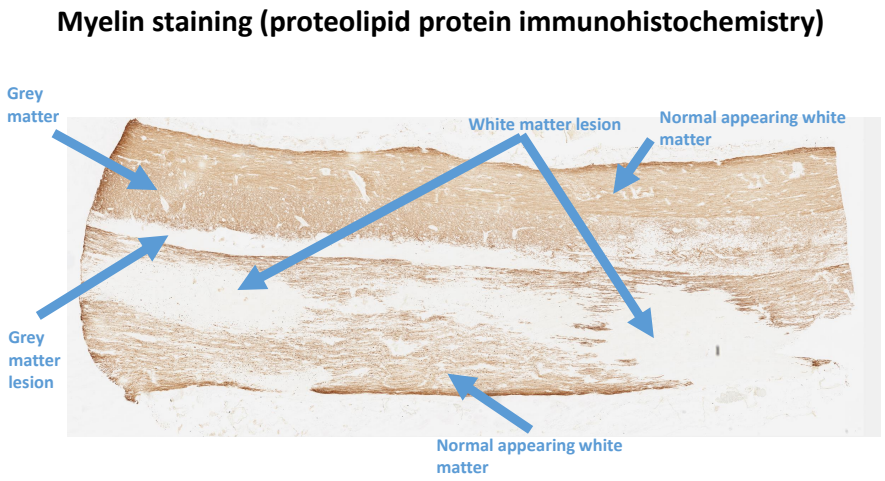
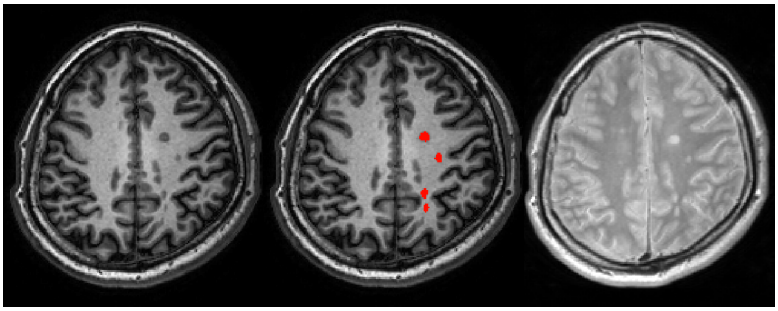


Figure 2: Histological image of human spinal cord tissue with MS lesions. The myelin staining clearly shows reduced myelin content in lesion areas. Image courtesy of F. Grussu.



(a) (b) (c)

Figure 3: MS patient's brain MRI scan with (a) T1 weighting, (b) T1 weighting overlaid with binary lesion mask in red, and (c) PD weighting. Note the hypo- and hypterintense lesions in (a) and (c) respectively.

T cells usually have an important function in the body's defence system but a failure in their regulation can cause inflammation. Attacks which occur after a viral or bacterial infection weakened the blood-brain barrier and made it permeable for T cells [22]. Inflammatory processes are then triggered so that other immune cells, cytokines and antibodies are released and the transmission of information in neurons is significantly reduced [23]. This inflammation causes transections of the axons and leads to acute axonal loss from Wallerian degeneration within 18 months. Additional axonal loss is seen in chronically demyelinated axons due to the lack of trophic support of the myelin [22].

In summary, acute lesions cause acute symptoms of relapses, whereas axonal loss is assumed to be responsible for progressive accumulation of disability.

There are three main phenotypes of MS as described by the National Multiple Sclerosis Society of the United States [73]: relapsing-remitting (RRMS), secondary-progressive (SPMS) and primary-progressive (PPMS). These phenotypes are used to describe the most common patterns of progression and have been defined empirically based on the past course of the disease (see Figure 4). Approximately 80 % of MS patients have an initial course described by RRMS [23]: attacks occur at unpredictable times and locations of the CNS. The deficits suffered from the attacks become permanent in about 40 % of the cases and even more likely with increasing disease duration [23, 120]. RRMS courses with only non-permanent deficits are referred to as benign MS [97]. Most patients with a RRMS disease course will eventually convert to SPMS after an average of 19 years [102]. SPMS is characterised

by a RRMS onset that converts to a continuous increase of disability without definite periods of remission [23, 73]. 10-20 % of MS patients suffer from the PPMS subtype which is characterised by a continuous neurologic decline with none (or rare and minor) remissions and also show fewer attacks [73, 79]. PPMS onset is usually at approximately 40 years, which is a similar age as the mean age for conversion from RRMS to SPMS [23].

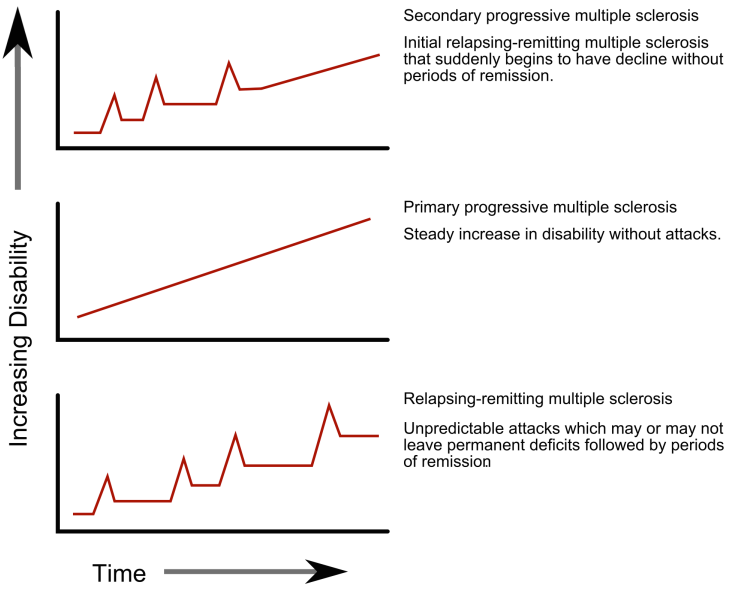


Figure 4: Illustration of progression in the three main phenotypes of MS. Image adapted from [130].

2.5 DISABILITY IN MS

Patients with MS can suffer from a wide range of neurological conditions and the symptoms of these conditions are related to the location of lesions, inflammatory processes and especially axonal loss within the central nervous system. Common problems are loss of sensitivity, numbness, muscle weakness or spasms, difficulties with moving, coordination and balance, tiredness, pain, and visual problems. Cognitive and emotional problems (i.e. depression) can increase throughout the course of the disease [23]. Approximately 75 % of all MS patients are still able to walk independently 15 years after disease onset but the majority is wheelchair-bound by the time of death. The average life expectancy is about 30 years from disease onset and

almost 70 % of deaths can be directly related to the disease (i.e. infections in weaker, non-ambulatory patients) [75].

The expanded disability status scale (EDSS) is the most common measure to describe disability of patients with MS [71]. The score ranges from 0 to 10 in steps of 0.5 with a higher score indicating higher disability (see Table 1). Especially the first steps of the score are determined by the degree of dysfunction in eight functional systems (FS) which are assessed during a neurological examination: pyramidal, cerebellar, brainstem, sensory, bowel and bladder, visual, cerebral or mental, and other. Each function is scored on the basis of the extent of impairment.

Although EDSS is widely used and accepted it has several disadvantages as it increases in constant steps of 0.5 but is not a linear scale (i.e. an increase from 1.0 to 1.5 [increase in number of affected functional systems but no disability] is not as severe as an increase from 6.5 to 7 [transition from walking support to wheelchair]). Also, EDSS is strongly biased towards ambulation of a patient which means that the inability to walk may mask disability or improvement in other areas. Furthermore, studies show that the vague description of the lower scores leads to poor inter-rater agreement [110] which can limit the reliability in multi-centre studies.

In addition to EDSS, the MS functional composite (MSFC) score has been introduced and is often used in more recent studies. It combines tests of physical and cognitive function and consists of the following tasks [103]:

- timed walk test (TWT): patients have to walk a certain distance (i.e. 25 ft) while the needed time is stopped. It measures the mobility and can be performed with assisting devices such as crutches.

- nine-hole peg test (NHPT): patients have to put pegs of different shapes into the right hole in a box. This test measures the function of arms and hands as well as basic cognitive functions. The task has to be performed with both hands consecutively.

- paced auditory serial addition test (PASAT): patients hear a number every 3 seconds and have to add the last two numbers they heard. This assesses working memory, information processing and calculation ability.

- self-assessment with basic questions about the previously described tasks.

Both NHPT and PASAT are sometimes criticised as measures because the results can be improved through practice. The value of the TWT can be questionable due to use of walking aids.

2.6 CLINICALLY ISOLATED SYNDROME

Most patients who develop MS present initially with a clinically isolated syndrome (CIS). This is a term used for acute or sub-acute neurological symptoms at onset which are characteristic of demyelination of the CNS [78]. CIS is only diagnosed if the neurological attack lasts for at least 24 hours and is not associated with fever, infections or clinical features of encephalopathy. The disorder is usually clinically isolated in space but can also occur with multifocal onset. Common presentations are lesions in the spinal cord (46 %), inflammations of the optic nerve (21 %) and brainstem syndromes (10 %). Approximately 23 % have a multifocal CIS onset [78, 80]. The median age of onset is 29-31 years [24, 128] and the disease prevalence for women is 250 % higher for women than for men [78].

Several features have been reported to influence the prognosis of patients with CIS. Generally, the prognosis is good when symptoms are isolated, attacks rare and benign and no lesions can be seen in MRI, whereas multifocal symptoms, high relapse rate, disability and lesions are factors for a poorer prognosis [80] (see Table 2). Even though lesion load is considered an important predictor for disease progression [70, 78], lesion load and distribution is actually very similar between patients who convert to MS within one year and those who convert later or not at all as shown in Figure 5 [50].

Table 2: CIS features that have been reported to affect patients' prognosis [70, 80].

Good prognosis	**Poor prognosis**
• optic neuritis	• multifocal CIS
• isolated sensory symptoms	• different systems affected
• long interval to second relapse	• high relapse rate in the first 5 years
• no disability after 5 years	• substantial disability after 5 years
• normal MRI	• abnormal MRI with large lesion load
• oligoclonal bands negative	• oligoclonal bands positive
• male sex	• female sex

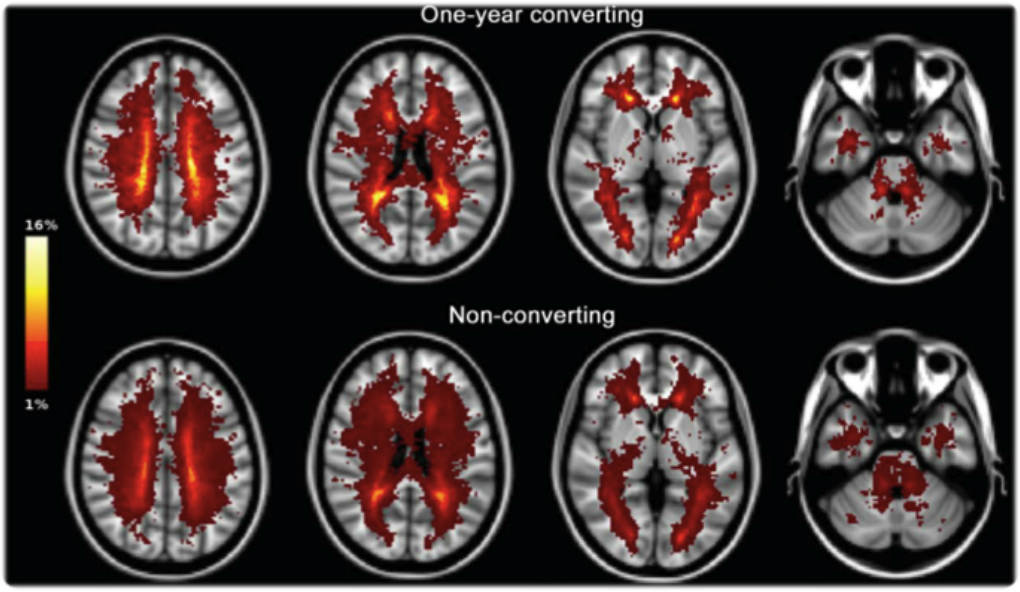


Figure 5: Distribution of MS lesions in one-year converters and non-converters. Source: Giorgio et al. [50].

Although, patients with CIS have a very high risk of developing MS, some are affected by other neurological conditions, which can have a similar onset but a very different disease progression such neuromyelitis optica (NMO) [39, 54]. However, 43 % of all CIS patients convert to MS after 5 years, 59 % after 10 years and 68 % after 14-20 years [38, 43]. Patients with abnormal baseline MRI (i.e. presence of T2 lesions) convert in 70-80 % of the cases, whereas patients with radiologically normal imaging convert in only 20-25 % of all cases.

Diagnosis of MS is standardised by the 2001 McDonald criteria, which combine characteristics from MR imaging, such as the occurrence of new lesions, with clinical measures such as positive CSF tests or further clinical attacks [76]. These criteria have been revised in 2005 [98] and 2010 [99]. The most recent 2010 criteria are shown in Table 3. Generally, MS can only be diagnosed when dissemination in time and space is proved or, clinically, when the patient has had a second neurological attack after a period of well-being. Latter is the definition of clinically definite multiple sclerosis (CDMS) which is used as gold standard in the scope of this work. The accuracy of the (original) McDonald criteria has been reported to be 80-83 % with a sensitivity and specificity of 74-83 % and 83-86 % respectively [31, 115].

There is currently no clinical standard to predict conversion from CIS to MS at baseline. Diagnosis of MS can only be made by showing dissemination in time and space. If dissemination in space and time cannot be proven clinically, then MRI can be used for follow-up assessments or MRI scans using a gadolinium tracer, which can show both active (new) and inactive (old) lesions and subsequently fulfil the criteria in a single scan.

Table 1: EDSS score for measurement of disability in MS.

Score	Description
0.0	Normal neurological Exam
1.0	No disability, minimal signs in 1 FS
1.5	No disability, minimal signs in more than one FS
2.0	Minimal disability in 1 FS
2.5	Minimal disability in 2 FS
3.0	Moderate disability in 1 FS; or mild disability in 3 - 4 FS, though fully ambulatory
3.5	Fully ambulatory but with moderate disability in 1 FS and mild disability in 1 or 2 FS; or moderate disability in 2 FS; or mild disability in 5 FS
4.0	Fully ambulatory without aid, up and about 12 hours a day despite relatively severe disability. Able to walk without aid for 500 m
4.5	Fully ambulatory without aid, up and about much of day, able to work a full day, may otherwise have some limitations of full activity or require minimal assistance. Relatively severe disability. Able to walk without aid for 300 m
5.0	Ambulatory without aid for about 200 m. Disability impairs full daily activities and ability to work full day without special provisions.
5.5	Ambulatory for 100 m, disability precludes full daily activities
6.0	Intermittent or constant unilateral assistance (cane, crutch or brace) required to walk 100 m with or without resting
6.5	Constant bilateral support (cane, crutch or braces) required to walk 20 m without resting
7.0	Unable to walk beyond 5 m even with aid, essentially restricted to wheelchair, wheels self, transfers alone; active in wheelchair about 12 hours a day
7.5	Unable to take more than a few steps, restricted to wheelchair, may need aid to transfer; wheels self, but may require motorised chair for full day's activities
8.0	Essentially restricted to bed, chair, or wheelchair, but may be out of bed much of day; retains self care functions, generally effective use of arms
8.5	Essentially restricted to bed much of day, some effective use of arms, retains some self care functions
9.0	Helpless bed patient, can communicate and eat
9.5	Unable to communicate effectively or eat/swallow
10.0	Death due to MS

FS: functional system

Table 3: Revised McDonald criteria for the diagnosis of MS as defined in 2010 by an international panel in association with the National Multiple Sclerosis Society of America [99].

Clinical presentation	**Additional Data Needed**
* 2 or more attacks (relapses) * 2 or more objective clinical lesions	None; clinical evidence will suffice (additional evidence desirable but must be consistent with MS)
* 2 or more attacks * 1 objective clinical lesions	Dissemination in space (DIS), demonstrated by: * 1 or more T2 lesions in at least 2 of 4 of the following MS-typical areas of the CNS: periventricular, juxtacortical, infratentorial, or spinal cord * or further clinical attack involving different site
* 1 attack * 2 or more objective clinical lesions	Dissemination in time (DIT), demonstrated by: * simultaneous presence of asymptomatic gadolinium-enhancing and non-enhancing lesions at any time * or a new T2 and/or gadolinium-enhancing lesion(s) on follow-up MRI, irrespective of its timing with reference to a baseline scan * or await a second clinical attack
* 1 attack * 1 objective clinical lesion (CIS)	Dissemination in space and time, demonstrated by: - For DIS: * 1 or more T2 lesion in at least 2 of 4 MS-typical regions of the CNS (periventricular, juxtacortical, infratentorial, or spinal cord) * or await a second clinical attack implicating a different CNS site - For DIT: * simultaneous presence of asymptomatic gadolinium-enhancing and nonenhancing lesions at any time * or a new T2 and/or gadolinium-enhancing lesion(s) on follow-up MRI, irrespective of its timing with reference to a baseline scan * or await a second clinical attack
Insidious neurological progression suggestive of MS (primary progressive MS)	One year of disease progression (retrospectively or prospectively determined) and two or three of the following: a) evidence for DIS in the brain based on 1 or more T2 lesions in the MS-characteristic (periventricular, juxtacortical, or infratentorial) regions b) evidence for DIS in the spinal cord based on 2 or more T2 lesions in the cord c) positive CSF (isoelectric focusing evidence of oligoclonal bands and/or elevated IgG index)

3

MACHINE LEARNING

Machine learning (ML) is a part of artificial intelligence and describes a set of algorithms where performance improves through learning from previously seen cases. The general concept has been introduced several decades ago but the number and complexity of applications increased rapidly in over last decade. Recent advances include mastering Atari video games solely based on the knowledge of controls and the resulting scores [83] or AlphaGo, an algorithm that is able to beat the world's best players of Go, a game more complex than chess [49]. Similarly spectacular results do not exist yet in medical applications even though several projects in the medical sector exist such as IBM's Watson which is a programme that performs evidence-based diagnoses and aims to reduce the amount of misdiagnosis in clinical practice (see also https://www.ibm.com/watson/health/).

In this thesis we focus on more conventional supervised classification which is used to distinguish two or more groups within a data set based on common patterns of these labelled groups. A common application of this is handwriting recognition, where an algorithm can learn the shape of individual letters from many examples and then be applied to 'translate' previously unseen hand-written texts into digital text. Similarly, email filters can detect spam messages based on predefined rules (i.e. learned from examples) that are continuously improved through user interaction (i.e. when spam is incorrectly marked).

In this chapter, we will explain some technical aspects of the supervised classification algorithms used in this thesis. Since, this is well described in literature already, we will focus on the more fundamental points relevant to this thesis. Detailed information on Support Vector Machines can be found in [26], on Random Forests in [27] and [11], and on general concepts of machine learning and pattern recognition in [8].

3.1 SUPERVISED CLASSIFICATION

The motivation for supervised classification is that we want to automate e.g. a diagnosis process based on available historical data about a certain condition. This could be medical records regarding a certain disease, which contains a set of measures as well as a diagnose given by an expert in the clinical field. The same information can now be obtained from healthy subjects so that a classifier can be used to find patterns in the measurements that distinguishes patients from healthy subjects. This in itself does not provide new insight though, since we already know the clinical status of the two groups. However, a well-trained classifier can also be generalised and applied to unseen data. This means that if we obtain the same measures for a new patient the classifier model will be able to provide a diagnosis without the need to of a clinical expert.

More technically, the training data $\mathcal{D}$ for a classification algorithm consists of n tuples $(\mathbf{x}_i, l_i)$ of the form $\mathcal{D} = \{(\mathbf{x}_i, l_i)|\mathbf{x}_i \in \mathbb{R}^p, l_i \in \{C_k\}\}_{i=1}^{d}$ where $\mathbf{x}_i = (x_i^1, x_i^2, \ldots, x_i^d)$ is a d-dimensional feature vector describing e.g. an individual patient's measures and $l_i \in C_k$ is an associated label that could e.g. describe the patient's diagnosis. In this thesis, two disjoint classes C_1 and C_2 were used such that patients were labeled either CIS-MS-converters or non-converters. The classification algorithm can then be described as a function $f(\mathcal{D})$ which uses the information stored in $\mathcal{D}$ to map all $\mathbf{x}$ to labels l.

3.2 SUPPORT VECTOR MACHINE

Support vector machines (SVMs) are the most commonly used classifiers, which can be partially attributed to the comprehensive toolbox LibSVM [17] but also to its simplicity and efficacy. The algorithm was designed to solve binary classification tasks and to look at classification as a geometric problem that can be solved by separating the two classes spatially using a hyperplane (i.e. a straight line in a two-dimensional space). In contrast to other classifiers such as a perceptron, a SVM has

a unique solution where the margin between the two classes is maximised using the points closest to the decision boundary, these points are called support vectors (see Figure 6 (a)).

Since real-world data is often not linearly or not perfectly separable (e.g. because of noise, misdiagnosis or outliers), many extensions of the original formulation have been proposed. Soft-margin SVMs use a cost function which allows for a certain degree of misclassification in the training data, which may arise from mislabelled data or noise. Kernel SVMs were introduced to solve classification tasks which are not linearly separable in the space originally spanned by the features vectors (see Figure 7).

LINEAR SVM A hyperplane is a geometric construct that has one dimension less than the space it is described in. In the case of a two-dimensional feature space the hyperplane would be a (one-dimensional) line, in three-dimensional spaces it is a (two-dimensional) plane, while in higher dimensions it is generalised to the term hyperplane. In a linear SVM the hyperplane of interest is the one that separates the two classes such that the margin between them is maximised.

This hyperplane is defined by the data points $\mathbf{x}$ that satisfy

$$y(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} - b = 0 \tag{1}$$

where $\mathbf{w}$ is a normal vector to the hyperplane, $\mathbf{x}$ contains the feature vectors, and b determines the offset of the hyperplane from the origin. Both $\mathbf{w}$ and b are chosen such that the margin, the perpendicular distance between the separating hyperplane and the closest data points of each class, is maximised. The closest data points are called support vectors and lie on two parallel hyperplanes

$$y(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} - b = 1 \tag{2}$$

and

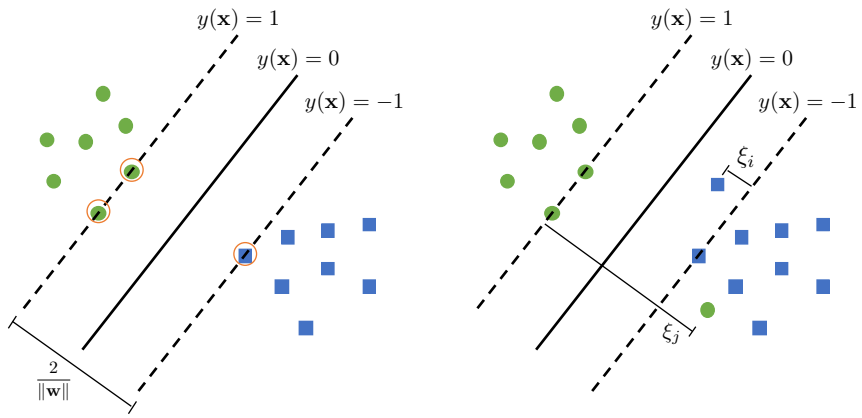
$$y(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} - b = -1 \tag{3}$$

which have a distance of $\frac{2}{\|\mathbf{w}\|}$ from each other as shown in Figure 6 (a). This distance can be maximised by minimising $\|\mathbf{w}\|$ using the constrained optimisation

$$\min_{\mathbf{w},b} \frac{\mathbf{w}^T\mathbf{w}}{2} \tag{4}$$

$$\text{subject to } l_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geqslant 1.$$

The constraint ensures that all support vectors are on or outside the supporting hyperplanes and not inside the margin.



(a) (b)

Figure 6: Illustration of support vectors and hyperplanes. (a) Two classes (green circles and blue squares) separated by a hyperplane (continuous line). Two supporting hyperplanes are spanned by the data points closest to the decision surface (circled in orange). In (b) slack variables ξ_i and ξ_j allow data points inside the margin and misclassification.

SOFT-MARGIN SVM Since many real-world applications do not have data that is linearly separable or perfect labels, certain degrees of misclassification can be allowed by the introduction of slack variables $\blacksquare_i$. The SVM algorithm then finds a trade-off between a maximal margin as described before and the degree of misclas-

sification as shown in Figure 6 (b). This is controlled by the cost or penalty parameter C such that the constrained optimisation (4) is extended to

$$\min_{\mathbf{w},\blacksquare,b} \left\{ \frac{\mathbf{w}^T\mathbf{w}}{2} - C\sum_{i=1}^{N} \xi_i \right\} \tag{5}$$

subject to $l_i(\mathbf{w}\cdot\mathbf{x}_i - b) \geqslant 1 - \xi_i$ and $\xi_i \geqslant 0$.

NON-LINEAR SVM In some cases, the data cannot be linearly separated in the space spanned by the available features. Non-linear kernels $k(\mathbf{x})$ can be used to map each feature vector into a higher-dimensional space where linear separation may become possible. The illustration in Figure 7 shows an example where one-dimensional data becomes separable after a simple mapping into two-dimensional space using $k(\mathbf{x}) = \mathbf{x}\cdot\mathbf{x}$.

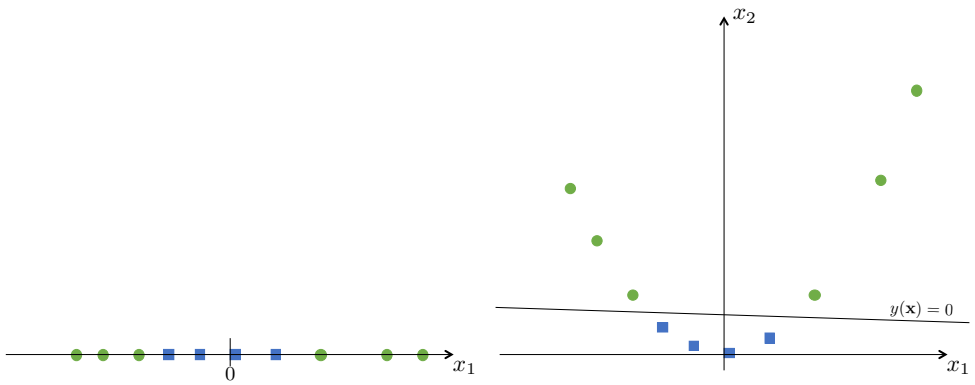


Figure 7: A not-linearly-separable one-dimensional data set (a) is mapped into a two-dimensional space (b) where linear classification is possible using $y(\mathbf{x})$.

Common kernel functions are polynomials of degree d

$$k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)^d$$

or the Gaussian radial basis function (RBF)

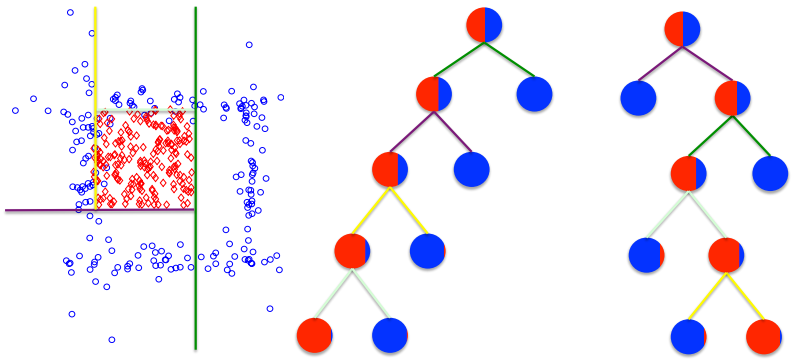
$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2), \text{ with } \gamma > 0.$$

3.3 RANDOM FOREST

Random forests are ensemble classifiers which consist of multiple classification trees. Each of these trees can be used individually to classify new cases. However, single trees tend to overfit the data which means that they work extremely well on the data set they were trained on but they generalise very poorly to new data. Combining a larger number of trees reduces the risk of overfitting and instead increases the generality of the classifier model.

CLASSIFICATION TREES Classification trees are directed graphs, where nodes and edges are hierarchically organised. The algorithm starts at the root node where a split function is applied that separates the data into two or more child nodes based on feature thresholds that decrease the heterogeneity of class labels in the child nodes (see Figure 8 (a) and (b)). In the scope of this thesis, we will only work with binary classification tasks and two child nodes, and limit the following description to these cases.

In each parent node the features are parsed in order to find the feature and its associated cutoff-threshold that reduces the heterogeneity in the resulting child nodes most. Since the computational expense is proportional to the number of parsed features, it is not desirable to look through all features at each node, especially when the number of features is very high. As a solution for this, the algorithm selects a random subset of features at each parent node. The number of these randomly selected features can be selected by the user and is often set to the square root of all features as it is in the work presented in this thesis. Reducing the number of parsed features, however, results in a tree that does not necessarily find the optimal features and thresholds but only local maxima. We will explain later in this chapter why this is advantageous for Random Forest classifiers anyway.



(a) (b) (c)

Figure 8: A data set with two classes 'red' and 'blue' can be separated (a) using a tree. Depending on the axis (or feature) in the first step, multiple trees achieve results that are different but very similar to each other (b) and (c). The heterogeneity of the data (sub-) sets is indicated by the colour-coding of the nodes in (b) and (c).

SPLIT FUNCTION A split function measures the heterogeneity of labels in two child nodes that have been created from thresholding a certain feature. Usually, multiple features and thresholds are tested and the combination with the highest reduction of heterogeneity is selected. In Figure 8 (a) we show a distribution of red and blue circles with an equal number of instances in each class. The first selected feature for the tree shown in (b) is the x-axis and the threshold is indicated by the green line. This results in two child nodes, which represent two subsets of the data space (see second level of the tree in (b) and the data points left and right of the green line in (a) respectively). One of these subsets is now completely homogeneous and represents the blue class, while the other subset is still heterogeneous but the distribution shifted towards the red class. Subsequent applications of this split function will create more child nodes and reduce the heterogeneity of the subsets further as indicated by the purple, yellow and light blue lines in Figure 8 (a) and (b).

A common measure for a split function is information gain

$$I_{IG} = H(S) - \sum_{i \in \{L,R\}} \frac{|S^i|}{|S|} H(S^i) \tag{6}$$

which uses the (discrete) Shannon entropy

$$H(S) = -\sum_{C_k \in \{-1,1\}} p(C_k) \log\left(p(C_k)\right) \tag{7}$$

to calculate the decrease of heterogeneity from the parent node S to the two child nodes S^i based on the class proportions $p(C_k)$.

Another option would be gini impurity

$$I_g = \sum_{C_k \in \{-1,1\}} p(C_k)(1 - p(C_k)), \tag{8}$$

which measures the relative proportion of labels $p(C_k)$ in each child node. It is minimal ($I_g = 0$) when both nodes only consist of instances with the same class and is maximal ($I_g = 0.5$) when both classes are equally distributed in the both nodes.

RANDOMNESS The name Random Forest arises from the fact that an ensemble of classification trees is used and these trees are randomly different from each other (see Figure 8 (b) and (c)). This difference arises from the random selection of feature subsets as described above. If the algorithm is always presented with the entire set of features it would always pick the same feature and cutoff-threshold at each respective node and, consequently, each tree will be the same. With a reduced number of features, the correlation between trees is getting reduced as well because the algorithm will not be able to pick the global best feature and threshold but only a local optimum from the presented subset. As a result, the trees will be less correlated and will have much better generalisation properties because the features and associated cutoff-thresholds are less rigid.

An example of this behaviour is given in Figure 9 where the algorithm is presented with two clearly distinct clusters of yellow and red dots (see (a)). The features are defined by the axes x_1 and x_2 and there are many threshold options that would perfectly separate the two classes (see (b)). A single tree, however, creates very rigid

classification boundaries as shown in the left side of Figure 9 (c) where all new instances in the yellow area would be assigned the label 'yellow' even if they are in the lower right corner of the field and therefore are much closer to the red cluster. If we create a larger number of trees where both the x_1 and the x_2 axis have been selected as features and the thresholds have been varied, the classification boundaries become more smooth and only the areas directly around the data clusters will still be dominantly yellow or red (see middle of Figure 9 (c)). The areas in between become more uncertain with increasing distance from the data as one would expect intuitively. With a further increase of the number of trees, this becomes even more evident in the right side of Figure 9 (c) where we can see a continuous transition from a high probability of 'yellow' in the top left corner to a high probability of 'red' in the bottom right corner and a high degree of uncertainty in between.

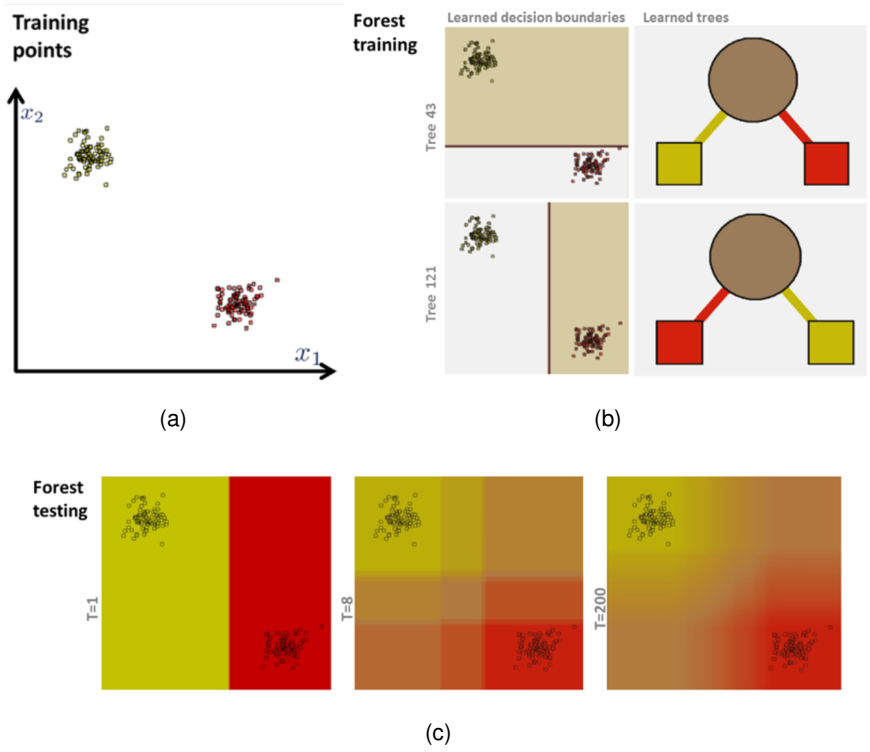


Figure 9: Illustration of the influence of forest size. Two classes (red and yellow circles) (a) can be separated in many different ways along both axes x_1 and x_2 (b). Single trees tend to overfit and generalise less well (c, left side) while an increasing number of trees T reduces these effects and allows for a more stable probabilistic classification (c, middle and and right side). Image source: Criminisi et al. [27].

OVERFITTING Individual classification trees tend to overfit the training data unless certain precautions are taken into account. It can be seen that the final separation in Figure 8 is not perfect and could be further improved by adding more levels to the trees. This would optimise the result for the given data set but it is likely that the model would not perform well on new data close to the border between red and blue instances.

The misclassification in the example above is only due to individual outliers so an 'improvement' of the tree could only be achieved by isolating single instances in a node. To avoid this, we can simply ensure that a parent node will not be split if the resulting child nodes will contain only a very small number of instances. Similarly, we can ignore potential parent nodes if they contain only very small numbers of features. Usually, these two measures are already sufficient to limit overfitting. In very large data sets, however, it might be useful to also limit the total number of nodes or levels in the tree since overly large trees are more likely to overfit and their exponential growth makes trees computationally expensive after a certain point.

COMBINING TREES The proportions of instances from the training set that end up in each terminal or leaf node determine the outcome of a new data point that has to be classified. The bottom left leaf node in Figure 8 (b) for instance was predominantly filled with 95 % blue training data points. A new instance that is propagated through the tree and ends up in that leaf will therefore be associated with a probability of 95 % for belonging to the blue class and 5 % for the red class.

It is then possible to define a threshold (e.g. 50 %) and assign a definite label to the data point. Outcomes from different trees can subsequently be combined using a majority vote. This means that a data point gets the class label that has been assigned to it by most of the trees. Alternatively, we could take advantage of the probabilistic nature of classification trees and calculate an average probability over the outcomes from each single tree for all instances. The resulting mean probability can then be thresholded if a binary label is needed.

3.4 PERFORMANCE ESTIMATION

Classifiers do not work perfectly in real world applications due to many reasons such as small or unrepresentative data or large variations in the population. Therefore, it is necessary to make certain precautions that allow to estimate the performance of the classifier and ensure that the findings from a limited small data set can be generalised to a larger population.

ACCURACY MEASURES The accuracy of a binary classifier is generally defined as the proportion of instances that have been assigned to the correct class:

$$\text{accuracy (acc)} = \frac{\#\text{true positive} + \#\text{true negative}}{\#\text{subjects}}.$$

In this thesis, we defined the conversion from CIS to CDMS as the positive event and non-conversion as the negative event which results in a confusion matrix as shown in Table 4.

Consequently, we can define additional performance measures such as

$$\text{sensitivity (sens)} = \frac{\#\text{true positive}}{\sum \text{converters}},$$

$$\text{specificity (spec)} = \frac{\#\text{true negative}}{\sum \text{non-converters}},$$

$$\text{positive predictive value (PPV)} = \frac{\#\text{true negative}}{\sum \text{predicted converters}}, \text{ and}$$

$$\text{negative predictive value (NPV)} = \frac{\#\text{true negative}}{\sum \text{predicted non-converters}}.$$

Table 4: Confusion matrix.

	predicted converter	**predicted non-converter**
converter	# true positive	# false negative
non-converter	# false positive	# true negative

In the case of imbalanced class sizes (i.e. one class has much more subjects than the other), accuracy can be highly misleading as a performance measure. If we assume such an imbalanced cohort of 100 subjects with 80 subjects having condition A (positive) and 20 subjects having condition B (negative), it is evident that we can achieve an overall accuracy of 80 % with a classifier that diagnoses all subjects with condition A. Using additional measures, however, reveals that we have a sensitivity of 100 % and a specificity of 0 %. This can be combined to a single measure called

balanced accuracy which is defined as $\text{acc}_{\text{bal}} = \frac{\text{sens} + \text{spec}}{2}$ and would be 40 % in the above example. It now becomes clear that the classifier does not perform well as indicated by the accuracy of 80 % but in fact is worse than a completely random assignment of diagnoses, which would be expected to yield a (balanced) accuracy of 50 %.

CROSS-VALIDATION Clinical data is always limited in the number of available patients. At the same time it is necessary to train a classifier with as much data as possible and also evaluate how the model performs on unseen data. A possible solution to this dilemma is k-fold cross-validation (CV) where the available data is split into k disjunct subsets. To create the classifier model k-1 subsets are used and the remaining one is employed for testing how well the model generalises to 'new' data. The k subsets are consequently permuted k times so that all subsets have been used for testing once. The definitions for the accuracy measures remain valid.

An important parameter is the selection of an appropriate k where the two most extreme cases would be $k = 2$ and $k =$ #subjects. In the case of $k = 2$ the two resulting models are completely independent as there is no overlap between the subjects in each set. As a result, the variance in model performance can be expected to be high when the data has a high variance. When $k =$ #subjects, the resulting models are highly correlated because they share all but one subject when compared pairwise. As a result, findings using $k =$ #subjects tend to be positively biased and hence yield higher accuracies compared to real unseen data [67]. This approach is also known as leave-one-out (LOO) cross-validation.

A value of $k = 10$ is often suggested as a good compromise as it reduces the correlation between models and also does not introduce too much bias to the estimate [67]. Additionally, a small k reduces the computational expense because less models have to be created compared to LOO.

3.5 DATA SAMPLING

The available clinical data is likely to be not completely representative of the entire population with a certain disorder. This becomes even more evident when the sample size is small or when the data was collected at only one single centre. This effect is known as sampling bias and cannot be avoided completely but only reduced by collecting large sample sizes from different locations. In retrospective analyses, however, there is no way to change the available data.

RESAMPLING As mentioned before, unbalanced class sizes lead to a potential bias in the accuracy estimate, which can be reduced with alternative performance measures. In addition, the use of imbalanced class sizes often have a negative impact on the classifier performance. A common solution for this is resampling, which can be divided in oversampling and undersampling [37].

Oversampling means that subjects from the smaller class are randomly selected with replacement such that final number of selected subjects matches the number of the larger class. This, however, only solves the problem formally without adding any new information that the classifier could learn from. In fact it could even increase the bias when outlier patients are selected multiple times.

When using undersampling or downsampling, subjects from the larger cohort are randomly selected without replacement in order to match the number of subjects in the smaller cohort. This has the advantage that no data is artificially added but it has the clear disadvantage that a large portion of the data set is ignored for the analysis and this effect increases with increasing imbalance. Since downsampling is the more conservative variant it was used for the experiments in this thesis where appropriate.

PERMUTATION Resampling can introduce a second type of sampling bias, due to the random selection of subjects in both over- and downsampling. In both cases it is possible that the resulting cohort is not representative of the original data set. A good way to overcome this is permutation where the experiment is not only performed once but repeatedly with new randomly selected sets of subjects. This leads

to a variance in the model performance because non-representative cohorts will result in models that perform overly well or overly poorly with respect to the classification task [53]. Performing a large number of repetitions or permutations, however, shows that the set of obtained accuracies follows a Gaussian distribution so that the mean value of this accuracy distribution can be assumed to be indicative of the real performance.

Part III

MACHINE LEARNING EXPERIMENTS

4

VOXELS AS FEATURES

A common approach in supervised classification of neuroimaging data is to use arrays of voxel information [62, 64]. This can be direct intensity information or, more commonly, derived information such as grey matter or white matter density. In simple classification tasks it can be sufficient to apply linear SVMs in order to distinguish subject groups such as healthy controls and patients with Alzheimer's disease, or patients with early MS and patients with late MS. Both diseases are characterised by increasing brain atrophy with longer disease duration, so that distinct patterns become visible that can be automatically detected and used for classification. Studies, which were successful using this approach were placed in the lower-left corner of our proposed classification landscape (see also Figure 1) indicating low feature and low classification task complexity.

In this chapter we show the first experiments on the classification of CIS patients who will convert to CDMS which is a very challenging task. We do not expect simple patterns to exist for patients with early CIS since MS is mainly characterised by inflammation rather than atrophy in early stages [92] and appearance of MS lesions is rather random. However, there is a possibility of other, less obvious patterns, which could already form at CIS onset and help discriminate between progressive and non-progressive CIS types. We run pattern analysis experiments using linear SVMs exploratory on voxel information from T1-, T2- and PD-weighted MR images, as well as grey and white matter density maps. This chapter will explore the classification landscape in MS using low-complexity features in a high-complexity task.

4.1 DATA

We included data from two independent MS centres in Barcelona, Spain and London, UK, which are both part of the European 'Magnetic Resonance in Multiple Sclerosis' (MAGNIMS) research group. The centres provided MRI scans with T1-weighting, as well as dual echo PD-T2 images with a resolution of 0.975x0.975x3 mm^3, which were obtained according to the centres' local acquisition protocols. All MRI scans had been used previously for a different study [50] but were re-examined for strong artefacts or distortions. No problems were found so that all 189 patients scanned in Barcelona and all 73 patients from London were included for this study. More information on the cohorts can be found in Table 5.

Table 5: Demographic and clinical characteristics of the cohorts used for experiments in chapter 4. Information shown for both data sets.

	Barcelona data	**London data**
Age (mean, median, range)	31.7, 31, 16-50	33.1, 34, 19-49
Gender	134F/55M	49F/25M
EDSS (median, [range])	2 [0-6.5]	1 [0-8]
Type of onset	brainstem/cerebellum:	brainstem/cerebellum:
(number, [converters])	55 [12]	6 [1]
	spinal cord 52 [9]	spinal cord: 4 [4]
	optic neuritis: 50 [7]	optic neuritis: 64 [17]
	other: 32 [6]	other: 0 [0]
Converters at follow up	34 (18 %)	22 (30 %)

4.2 IMAGE PROCESSING

All MR images were initially corrected for bias fields using the N3 algorithm [111] and brain masks were obtained with FSL `bet`. The T2- and PD-weighted MRI scans were affinely registered to T1-space using `reg_aladin` from the NiftyReg toolbox [84]. The T1-weighted images were used to obtain tissue probability maps using the `seg_LoAd` algorithm [14] from the NiftySeg toolbox. Here, five priors were used for white matter, grey matter, external CSF, deep grey matter and internal CSF.

Since we are aiming to find voxelwise patterns, the MRI scans have to be transformed to a common space so that anatomical structures overlap as closely as possible. To this end, the T1-weighted MR images were diffeomorphically registered to the MNI-ICBM-152 template with a resolution of 1x1x1 mm^3 using `reg_aladin` for affine and `reg_f3d` for non-linear transformation consecutively. In the following step, the probabilistic tissue density maps as well as the co-registered T2 and PD images were resampled using the control point parameter maps from the T1-to-MNI registration. The individual tissue maps for grey and deep grey matter were added to create a unified map containing all grey matter voxels, and the GM and WM maps were thresholded at a probability of 50 % to avoid an overlap of tissue classes in border regions.

Due to physiological variation, the location of grey matter voxels in the brain varies between subjects so there is no GM mask that fits all patients. Therefore, all individual GM masks were overlaid to ensure that all GM voxels will be included for the analysis. However, this means that the combined mask will include WM, CSF or even non-brain voxels for some patients. To avoid this bias, the non-GM voxels were set to zero for individual patients. The same procedure was applied for the white matter tissue respectively.

4.3 EXPERIMENT DESIGN

Due to the exploratory nature of this study, many different combinations of centres and features types were analysed. Initially, we used T1-, T2- and PD-weighted intensity values as well as GM and WM densities independently in both centres respectively. For the London data set, we then looked for performance changes arising from combining the three imaging weightings by concatenating the respective feature vectors. Finally, we looked at GM and WM features in a data set consisting of all included patients from both centres. A detailed list of the performed experiments is given in Table 6.

SVMs are more stable when classes of similar size are used [58]. However, our two data sets are highly imbalanced with 155 non-converters versus 34 converters

Table 6: Overview of the experiments performed in chapter 4.

Data set	Feature type	#features (voxels)	#subjects
Barcelona	T1 intensity	2132133	34 vs 34
Barcelona	T2 intensity	2132133	34 vs 34
Barcelona	PD intensity	2132133	34 vs 34
Barcelona	GM probability	2057429	34 vs 34
Barcelona	WM probability	1389843	34 vs 34
London	T1 intensity	2132133	22 vs 22
London	T2 intensity	2132133	22 vs 22
London	PD intensity	2132133	22 vs 22
London	T1+T2+PD intensity	6396399	22 vs 22
London	GM probability	1995315	22 vs 22
London	WM probability	1307015	22 vs 22
Barcelona + London	GM probability	2225051	56 vs 56
Barcelona + London	WM probability	1447246	56 vs 56

and 51 non-converters versus 22 converters respectively. To overcome this problem, we randomly sampled patients from the large non-converter group to match the number of converters, a method known as subsampling or bootstrapping without replacement. To reduce the probability of a spuriously well-performing selection of patients and get a more generalizable outcome, this bootstrapping procedure was repeated 100 times. Furthermore, we used a leave-one-out (LOO) cross-validation. This means that all but one subject is used to train the classifier and the left out subject is then used to test the classifier's performance. Training and test patients are then systematically permuted until each patient has been used for testing once.

For these experiments, linear SVMs were used with a cost parameter varying between 2^{-1} and 2^{5}. In this study, we used the LibSVM library [17] within MATLAB 2012a.

The accuracy outcome was measured as the proportion of correctly classified patients with respect to the total number of patients. We report both the mean accuracies over all 100 bootstraps as well as the accuracy range and 95 % confidence interval (CI).

4.4 INTENSITY NORMALISATION

MRI intensities are not standardised as they depend on the scanner, the MRI coil, the subject and many more parameters that cannot be kept constant. This can have a crucial impact on algorithms that perform decisions based on intensity thresholds. Therefore, an additional experiment was designed to compare the classification outcome from 'native' (N3-corrected and registered) and intensity-normalised scans.

This intensity normalisation was performed as a piecewise linear transformation based on histogram information. In the histogram of a standard MRI scan of the brain with good contrast there are two clearly distinct peaks right next to each other (and one at the beginning of the low intensity spectrum for the dark background). These two peaks represent white and grey matter and are shown in Figure 10 (a). If the contrast is low, however, the distributions of the two tissue types strongly overlap and it is no longer possible to separate them (see Figure 10 (b)). Since we have some low-contrast data, we decided to create an algorithm that reduces the influence of this effect. It is based on the location of the highest peak, which is usually WM but can be a combination of WM and GM in the case of low contrast. Firstly, the location of all main peaks in the patients' histograms were estimated, and then the average over all these peak locations was calculated. This mean peak location was used as an anchor point for a piecewise linear intensity transformation such that intensities between zero and a patient's main peak location were stretched/squeezed into a new range between zero and the overall mean peak location. The intensities between the individual patients' main peaks and one were transformed respectively. The histograms of the cohorts are shown in Figure 10 (c) before and in (d) after the normalisation.

Additional experiments have been run with normalised MRI intensities on the London data set. The three imaging weightings were firstly used independently and then all three of them concatenated (see Table 7).

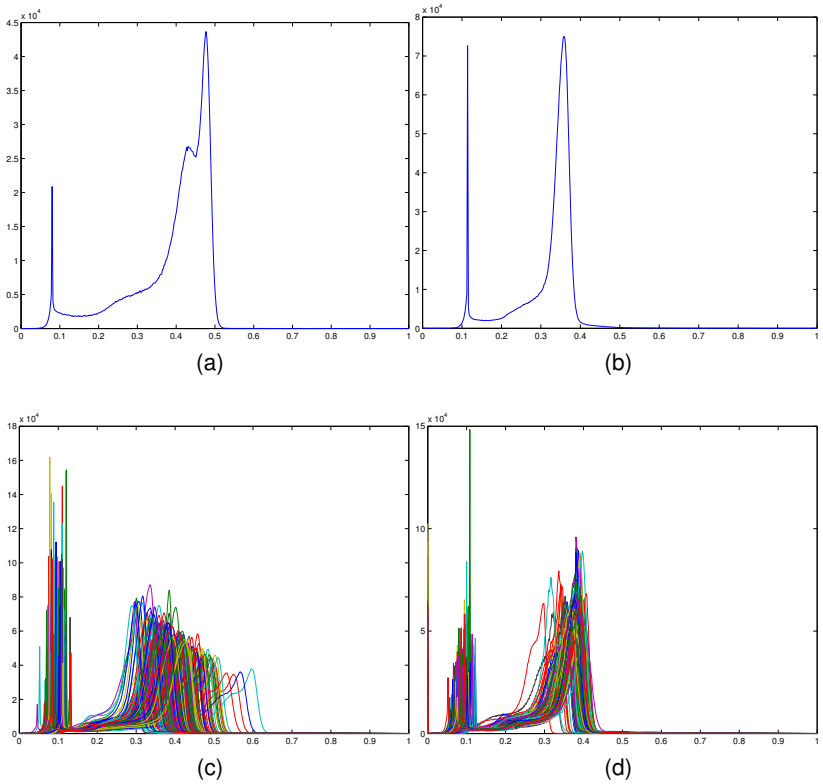


Figure 10: Illustration of intensity normalisation. (a) two main peaks for white and grey matter between 0.4 and 0.55 (and a background peak around 0.1). (b) grey and white matter peaks merged due to low contrast in image. (c) histograms of all subjects before normalisation. (d) same histograms after normalisation.

Table 7: Overview of the performed experiments on normalised intensity values.

Data set	Feature type	#features (voxels)	#subjects
London	T1 intensity normalised	2132133	22 vs 22
London	T2 intensity normalised	2132133	22 vs 22
London	PD intensity normalised	2132133	22 vs 22
London	T1&T2&PD intens. norm.	6396399	22 vs 22

4.5 RESULTS

The accuracies of predicting CDMS in the voxel-intensity-based experiments range between 47.2 % and 48.5 % in the Barcelona data set and between 46.7 % and 55.8 % in the London data set. The individual bootstrap samples of these experiments range between 20.5 % and 75 %. The highest accuracy here was obtained using the combination of all three MRI contrasts in the London data. Detailed results for all experiments are shown in Table 8.

Intensity normalisation of the London MRI scans lead to accuracies noticeably lower compared to the original data with a range from 43.5 % to 45.9 %.

The SVMs with tissue probability maps as features provide an accuracy of 48.7 % and 47.8 % for grey and white matter respectively in the Barcelona data set. The bootstrap accuracies range from 30.9 % to 63.2 % with 95 % confidence intervals (CI) of 47.3 % to 50.1 % and 46.6 % to 48.9 % for the two tissue types respectively. In the London data, these results are slightly lower with 45.7 % (CI: 43.5 %-47.9 %) using GM and 38.1 % (CI: 36.6 %-39.6 %) using WM with a range from 11.4 % to 70.5 %.

Combining both centre's data and apply the respective GM and WM densities to SVMs showed only little changes compared to single centre experiments. The accuracy using mixed GM masks is 50.6 % (range: 39.3 %-64.2 %, CI: 49.6 %-51.7 %) and using WM it is 59.1 % (range: 36.6 %-60.7 %, CI: 48.2 %-50.0 %).

4.6 DISCUSSION

All mean accuracies of the performed experiments are between 38.1 % and 59.1 % which is not a strong deviation from a random finding of 50 %. Hence, it can be concluded that - if used without further information - neither voxel intensities nor tissue probability masks are informative enough to find a difference between CIS patients who will convert to CDMS within one year and those who will not. This is very different from findings of other groups performing classification of CN vs AD

Table 8: Results of the performed experiments of section 4. Performance is presented as mean accuracy, 95 % confidence interval (CI) and range over 100 repetitions.

Data set	Feature type	Accuracy (CI) (%)	Range (%)
Barcelona	T1 intensity	47.2 (45.8-48.7)	27.9-64.7
Barcelona	T2 intensity	48.1 (46.6-49.6)	27.9-75.0
Barcelona	PD intensity	48.5 (47.1-49.8)	33.8-69.1
Barcelona	GM probability	48.7 (47.3-50.1)	30.9-63.2
Barcelona	WM probability	47.8 (46.6-48.9)	30.9-60.3
London	T1 intensity	46.7 (45.0-48.4)	20.5-65.9
London	T2 intensity	54.4 (53-55.8)	34.1-70.5
London	PD intensity	54.1 (52.6-55.6)	27.3-70.5
London	T1+T2+PD intensity	55.8 (55.5-57.1)	36.4-72.7
London	GM probability	45.7 (43.5-47.9)	11.4-70.5
London	WM probability	38.1 (36.6-39.6)	20.5-59.1
London	T1 intensity normalised	43.5 (42.0-45.0)	18.2-56.8
London	T2 intensity normalised	45.0 (43.6-46.4)	22.7-63.6
London	PD intensity normalised	45.9 (44.5-47.3)	22.7-63.6
London	T1+T2+PD intens. norm.	43.8 (42.3-45.2)	20.5-59.1
Barcelona+London	GM probability	50.6 (49.6-51.7)	39.3-64.2
Barcelona+London	WM probability	49.1 (48.2-50.0)	36.6-60.7

[64] or several subgroups of patients with MS [7]. It should be noted though that this classification task is extremely challenging as there is currently no known method to clinically assess a CIS patient's outcome from baseline data.

Due to the rather random outcome of the classification experiments we did not examine the SVM kernel weights in order to identify regions of particular interest as it has been done in other studies [7, 64]. The assumption that the strong difference of MRI intensity values between subjects (or different scans in general) causes a reduced classification performance could not be verified since the classification accuracies actually decrease after an intensity normalisation.

Tissue probability maps are expected to be widely independent from differences between scans and scanners because they are usually calculated from T1-weighted MRI scans and mostly rely on a good prior and sufficient contrast between grey and white matter structures rather than specific intensity ranges. Therefore, a combination of tissue probability maps from different centres should not significantly influence the outcome. In our experiments, it could be shown that the accuracy of the com-

bined maps is 50.6 % and 59.1 % for GM and WM respectively which is slightly higher than what we observe in single centres. Since we would expect a lower or similar accuracy due to the higher complexity in a multi-centre classification task, this is an additional indictor that this is in fact a random finding.
Generally, it must be noted that all results are at - or close to - 50 % and therefore has to be considered random so that no strong conclusions about the possibility of multi-centre classification experiments should be drawn from this study.

We showed in section 3.2 that it can become necessary to use kernel functions in order to map the data points into a higher-dimensional space when no linear separation is possible in the original feature space. Even though a linear separation is indeed not possible for the experiments in this chapter, it is very unlikely that this arises from a lack of dimensionality. The voxelwise approach creates a very high-dimensional feature space (see column #features (voxels) in Table 6), which is very sparsely populated with data points. Thus, it can be assumed that a further increase of feature-space dimensionality will not have any influence on the outcome and that the bad classification performance actually has to arise from a lack of differences between the two classes in this study.

Generally, it can be concluded that low-level features do not contain sufficient information to solve highly complex classification tasks such as the prediction of outcome in CIS patients.

5

HIGH-LEVEL FEATURES

In chapter 4 we showed that voxelwise information from MRI intensities, GM and WM densities, or combinations of these does not provide sufficient predictive power to classify CIS converters and non-converters with an accuracy above chance level. Expert knowledge can help overcome this problem by introducing high-level features that are described in literature and are known to play an important role in MS. Clinical studies usually look at single measures of interest - corrected for effects arising from differences in age, gender, etc - and present findings on a cohort level. Here, we explore how combinations of a set of high level-features perform in a supervised classification setting using SVMs. The aim is to use baseline data to predict conversion at 1- and 3-year follow-up.

It must be noted that these high-level features generally require input from clinical experts who e.g. outline lesions or perform clinical assessments, which can then be included in the classification analysis. This information was available for all patients in our retrospective study but if this work were to be generalised to a larger and more general cohort the method could not be fully automated and relies significantly on manual input. Due to this need for a clinical expert, we classify our task as high feature complexity and high task complexity in the classification landscape shown in Figure 1.

Parts of this work have been published in NeuroImage:Clinical [134].

5.1 DATA

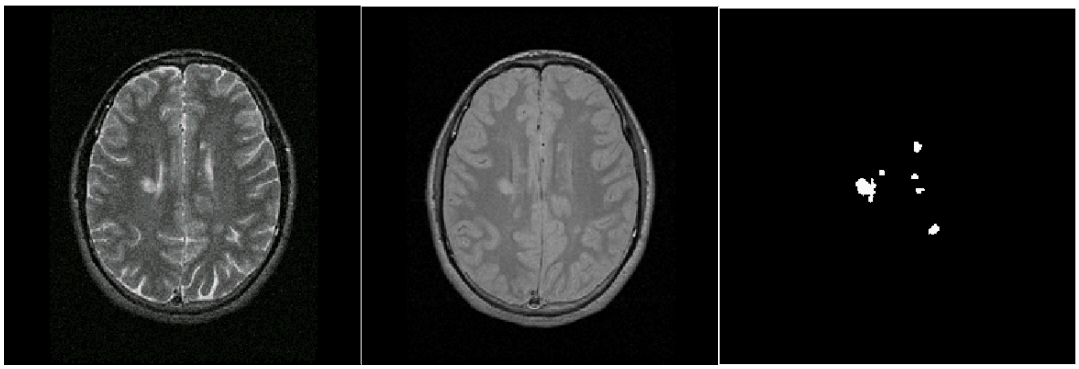
In the scope of this retrospective study 74 patients were included who were scanned at the UCL Institute of Neurology as part of a larger cohort recruited between 1995 and 2004. The time between disease onset and examination was on average 6.15

weeks (std 3.4), and all included patients were clinically reviewed after one year. Seventy patients were also included with a follow-up at 3 years. The inclusion criteria for our study were 1. presence of at least one demyelinating lesion visible in baseline MRI scans, 2. artefact-free MRI, 3. availability of baseline EDSS and onset location (i.e. spinal cord, optic nerve, brainstem or multifocal), 4. knowledge of age and sex, 5. knowledge of clinical outcome at one- and three-year follow-up. The clinical outcome noted at each follow-up was clinical conversion to MS due to the occurrence of a second clinical attack attributable to demyelination with a duration of more than 24 hours (see also first McDonald criterion in Table 3). In this cohort, 22 patients had a second relapse within one year (30 %) and 31 patients (44 %) fulfilled the criteria for CDMS after three years. None of the patients was on disease modifying treatment since they were at disease onset. Informed consent and approval by the local ethics committee was obtained prior to the study.
Detailed information on the cohort characteristics can be found in Table 9.

The baseline MRI protocol was undertaken using a 1.5 T GE Signa MRI scanning system. PD- and T2-weighted brain images were obtained using a FSE dual echo sequence with a repetition time (TR) = 3200 ms, echo time (TE) = 5/90 ms. The contiguous axial slices have a thickness of 3 mm and the in-plane resolution is 0.9375x0.9375 mm^2. Binary lesion masks were created by the same experienced neurologist for all patients with a semi-automated in-house software. Lesions were outlined in the PD-weighted images using the corresponding T2 images as a reference (see also Figure 11).

5.2 FEATURE DEFINITIONS

Expert knowledge can be included in the analysis in the form of high-level features. Academic literature reports a large set of features that are associated with good or bad prognosis of patients with CIS (see Table 2). These features have been reported as independent predictors of disease progression comparing large cohorts of patients. To our knowledge, this is the first study looking at a multivariable ap-



(a) (b) (c)

Figure 11: Example of T2- and PD-weighted images and corresponding binary lesion mask. Axial T2-weighted image (a) and proton-sensity-weighted image (b) showing hyperintense white matter lesions; the corresponding binary lesion mask (c) was used to obtain the lesion features.

Table 9: Demographic and clinical characteristics of the cohort used for the experiments in chapter 5. Information shown for both one- and three-year follow-up. Lesion count bins correspond to low, medium and high lesion count.

	1 year	**3 years**
Age (mean, median, range)	33.1, 34, 19-49	33.2, 34, 19-49
Gender	49F/25M	47F/23M
EDSS [range]	1 [0-8]	1 [0-8]
Type of onset	brainstem/cerebellum:	brainstem/cerebellum:
(number, [converters])	6 [1]	5 [1]
	spinal cord: 4 [4]	spinal cord: 4 [4]
	optic neuritis: 64 [17]	optic neuritis: 61 [26]
	other: 0 [0]	other: 0 [0]
Lesion count	$\leqslant$3: 14	$\leqslant$3: 13
(number of patients	4-10: 23	4-10: 23
per bin)	$\geqslant$11: 37	$\geqslant$11: 34
Converters at follow-up	22 (30 %)	31 (44 %)

proach combining these independent predictors and applying them to the outcome prediction of individual patients.

Lesions are the most dominant and most obvious features at early stages of MS such as CIS or RRMS. According to the literature abnormal MRI with large lesion load is predictive of poor prognosis [70]. Most of the features included in this study were derived from binary lesion masks. The idea was to collect information not only about the total volume but also to capture the number and distribution of a patient's lesions. MRI intensity in the PD- and T2-weighted images in lesion areas were used as approximate measures of lesion activity. Additionally, we included gender, EDSS and type of onset, which are known to correlate with disease progression in MS, as well as the patients' age since MS mostly affects younger people. The CIS type was coded according to $1 \mathrel{\widehat{=}}$ optic neuritis, $2 \mathrel{\widehat{=}}$ spinal cord, $3 \mathrel{\widehat{=}}$ brainstem, and $4 \mathrel{\widehat{=}}$ other. This coding was arbitrarily chosen but a permutation of this numbering, however, has little effect and reduces the accuracies of the best feature combinations by a maximum of 1.7 %.

A full list of features and their description can be found below:

1. Lesion count: this feature reflects the total number of lesions in the brain, extracted from the native lesion masks; it was computed using the original binary lesion masks and an 18-neighbourhood for voxel connectivity. This means that only lesion voxels that are connected by their faces are treated as the same lesion. Pure edge connections are not considered since there is a high chance of them being separate lesions where the apparent connections are due to partial volume effects.
2. Lesion load: this feature reflects the total lesion volume, in voxels, extracted from the native lesion masks
3. Average lesion PD intensity: this feature reflects the average PD intensity of the lesional voxels marked in the native lesion masks.
4. Average lesion T2 intensity: this feature reflects the average T2 intensity of the lesional voxels included in the native lesion masks.
5. Average distance of lesions from the centre of the brain: this feature gives the average distances between all lesional voxels and the centre of the brain

(defined as the central voxel of the ICBM-MNI 152 template). It provides information on how spread out the lesions were on the registered images.

6. Presence of lesions in proximity of the centre of the brain: this binary feature is 1 if there are lesions within a cube of 1 cm^3 centred around the central voxel of the ICBM-MNI 152 template, or 0 if no lesions were in the central box. This feature was selected because of the evidence that lesions located in the corpus callosum, which is a midline brain structure, are useful in predicting conversion to CDMS in addition to Barkhof criteria [56].
7. Shortest horizontal distance of a lesion from the vertical axis of the brain: this feature measures the shortest distance of a lesion's centroid (centre of mass) from the intersection of the midsagittal and midcoronal planes of the image. This feature represents an additional way of reflecting the distance of the lesions from the centre of the image.
8. Lesion size profile: this feature reflects the distribution of lesion sizes. All lesions were sorted according to their size in native space and divided in three groups of equal length representing small (1-15 voxels), medium-sized (16-36 voxels) and large (37+ voxels) lesions which give similar numbers in each category over the whole data set.
9. Age: MS affects mostly young people in their twenties and thirties.
10. Gender: Woman are three times more likely to develop MS compared to men but men have a worse prognosis.
11. Onset location: location of disease onset is controversially discussed in literature as a sole predictor. However, it is likely to be informative when combined with other features.
12. EDSS: fast progression from CIS is associated with higher disability when developing MS. High baseline EDSS might be indicative of faster progression.

5.3 EXPERIMENT DESIGN

The aim of this study is to predict the conversion from CIS to CDMS at 1- and 3-year follow-up using information available at CIS onset. Firstly, we want to examine

how individual features compare to feature combinations, and secondly, we want to identify the most predictive feature combination.

Since the proportion of converters at such short follow-up intervals is relatively small, the unbalanced group sizes of 22 converters vs. 52 non-converters and 31 converters vs. 39 non-converters for one and three years respectively can lead to a bias of the hyperplane weighting towards the larger group, and, in addition, often results in a high sensitivity and a low specificity or vice versa [67]. Therefore, as in section 4.3, we performed a bootstrapping without replacement. This means that patients were randomly selected from the larger non-converter group in order to match the size of the smaller converter group. This was repeated 100 times to reduce the effect of sampling bias (i.e. a coincidental selection of non-representative patients), provide a confidence interval of the prediction, and give a better idea of how well the model generalises to the whole cohort. Overall, we perform 100 experiments with 22 converters vs 22 non-converters for the 1-year follow-up and 100 experiments with 31 converters vs 31 non-converters for the 3-year follow-up in a leave-one-out cross-validation (LOO-CV).

In a LOO-CV for our 1-year follow-up 43 out of 44 patients are used in the training phase to calculate an optimal separating hyperplane (OSH). The remaining patient is then classified based on this OSH. The training and testing samples are permuted until every patient was used for testing once. The nature of LOO-CV implies that in each individual training step the classes are slightly imbalanced (i.e., 21 vs. 22 or 30 vs. 31) as one patient is always left out of the training cohort. This procedure, however, is performed for both classes in the exact same way so that this effect can be neglected.

Each experiment is performed using the built-in functions `svmtrain` and `svmclassify` from the MATLAB 2012a statistics toolbox. We used a polynomial kernel $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + c)^d$ with degrees d varying from 1 to 5. This includes the widely used linear kernel, which is a polynomial kernel of degree one but also allows the classifier to build more complex models. The maximum degree of 5 was chosen as a tradeoff between model complexity and overfitting. Parameter optimisation was

performed with an inherent sequential minimal optimisation (SMO) with 10 million iterations to allow for convergence.

Initially, all possible feature combinations were tested starting from every individual feature itself, pairs of features, triplets of features, etc. up to a concatenation of all 12 features. This leads to a total of $2^{12} - 1 = 4095$ feature combinations that are tested with each of the polynomial degrees yielding $4095 \times 5 = 20475$ experiments per follow-up. This ensures that we find the most predictive combination of features but has the risk of finding spurious effects arising from the application of a large set of models on a relatively small data set. This can be accounted for using statistical methods for multiple comparisons such as Bonferroni correction. This would mean that the significance level of the p-value - the probability of our result arising from random distribution - needs to be divided by the number of tested models. Using the default threshold for statistical significance of $\mathrm{p} = 0.05$, the adjusted significance level would be $\mathrm{p} = \frac{0.05}{20475} = 0.00000244$.

Alternatively, it is possible to perform a more systematic approach to feature selection known as forward Recursive Feature Elimination (fRFE) to create the model. This is an iterative method where features are added if they improve the accuracy of the previously applied feature set. It starts with performing the classifications using all individual features by themselves and identifying the feature providing the highest accuracy. Then, the remaining 11 features are subsequently combined with the most predictive single feature in order to find the most predictive pair, etc. This procedure is continued until adding a new feature does not increase the obtained accuracy anymore (see Figure 13). It should be noted that this method does not necessarily find the best combination because the most predictive individual feature does not have to be part of the overall-best combination of features. In these cases, fRFE will only find a local-maximum solution.

5.4 RESULTS

The highest average accuracy over all 100 bootstraps at 1-year follow-up was 60.6 % using individual features, and 71.4 % for both the fRFE approach and the exhaustive

search through all possible feature combinations. At three-year follow-up the average accuracy was 63.6 % for the most predictive single feature, 73.5 % for the overall best feature combination, and 68 % using the fRFE. An overview of the accuracies obtained with the individual features can be seen in Figure 12. Detailed results for the fRFE approach and the exhaustive search can be found in Tables 10 and 11 respectively.

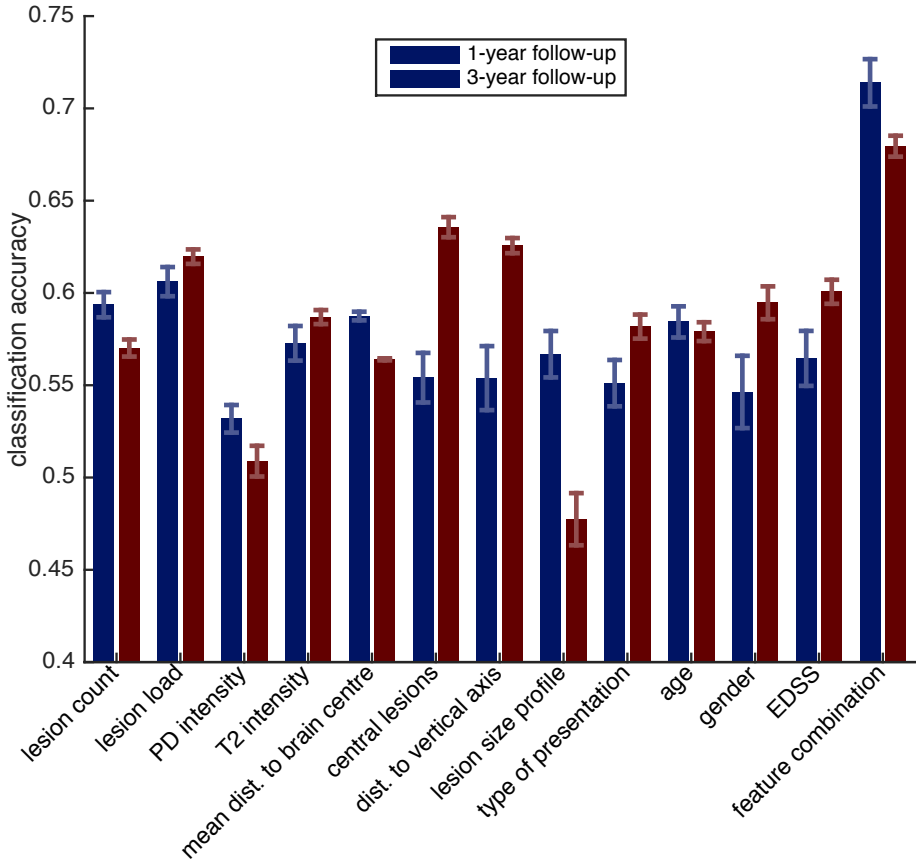


Figure 12: Accuracies obtained using individual features and the best fRFE combination for predicting the conversion from CIS to CDMS at one- and three-year follow-up. Error bars indicate the 95 % confidence interval.

Using fRFE at one-year follow-up, the accuracy increased from 60.6 % using only lesion count as feature to 71.4 % (sensitivity/specificity 77 %/66 %) by adding two additional features and using a polynomial kernel of degree 4. For the linear kernel (polynomial degree of one), the accuracy decreased when adding additional features. With degrees 2, 3 and 5, we can observe an increase in accuracy when adding two or three features but the obtained accuracy is not as high as with a kernel of fourth degree (see Figure 13 (a)).

Respectively, at three-year follow-up the accuracy increased from 63.6 % using only the average distance to the centre of the brain to 68 % (sensitivity/specificity 60 %/76 %) using fRFE. The linear kernel performed best here utilising six features. It can be seen in Figure 13 (b) that the polynomial kernels of degree 1 and 4 reached similar accuracies using 4 features but additional features did not increase the accuracy any further at degree 4 whereas the use of a linear kernel improved the result by another 0.8 %.

The same set of features provided the highest accuracy both with the fRFE and the exhaustive search approach on the one-year follow-up: location of onset, gender and lesion load. For the three-year follow-up the feature sets were different for the two methods so that lesion count, PD intensity, mean distance from lesions to centre of the brain, shortest distance from lesions to the vertical axis of the brain, EDSS, and age were most predictive using fRFE, and lesion count, lesion load, shortest distance from lesions to the vertical axis of the brain, age, gender, and EDSS were most predictive using the exhaustive search. The detailed results can be seen in Tables 10 and 11.

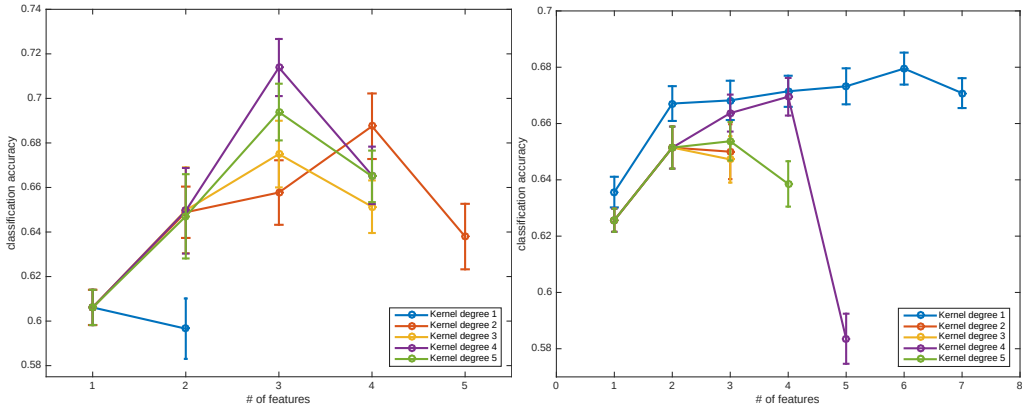


Figure 13: Accuracies obtained with forward RFE using an increasing number of features. The graphs show the progression of mean accuracies for the different polynomial degrees after recursively adding features in order to find the most predictive combination for conversion at one-year (a) and three-year (b) follow-up. Error bars indicate 95 % confidence intervals.

Table 10: Most predictive feature combination for classification of CIS converters and non-converters at one- and three- year follow-up as given by the exhaustive search across all possible feature combinations. Classification outcomes are presented using accuracy, range, 95 % CI, sensitivity, specificity, PPV and NPV.

	1 year	**3 years**
MRI features		
Lesion count		•
Lesion load	•	•
Average lesion PD intensity		
Average lesion T2 intensity		
Average distance of lesions from the centre of the brain		
Presence of lesions in proximity of the centre of the brain		
Shortest horizontal distance of a lesion from the vertical axis		•
Lesion size profile		
Clinical features		
Type of onset	•	
Age		•
Gender	•	•
EDSS at onset		•
SVM-based classification		
Polynomial degree	4	1
Accuracy (%)	71.4	73.5
Range (%)	52-84	65-81
95 % CI (%)	70-73	73-74
Sensitivity (%)	77	67
Specificity (%)	66	80
PPV (%)	70	77
NPV (%)	74	71

CI = confidence interval; PPV = positive predictive value; NPV = negative predictive value.

Table 11: Best feature combinations for classification of CIS converters and non-converters at a one- and three-year follow-up as selected through fRFE. Classification outcomes are presented using accuracy, range, 95 % CI, sensitivity, specificity, PPV and NPV.

	1 year	**3 years**
MRI features		
Lesion count		•
Lesion load	•	
Average lesion PD intensity		•
Average lesion T2 intensity		
Average distance of lesions from the centre of the brain		•
Presence of lesions in proximity of the centre of the brain		
Shortest horizontal distance of a lesion from the vertical axis		•
Lesion size profile		
Clinical features		
Type of onset	•	
Age		•
Gender	•	
EDSS at onset		•
SVM-based classification		
Polynomial degree	4	1
Accuracy (%)	71.4	68.0
Range (%)	52-84	61-74
95 % CI (%)	70-73	67-69
Sensitivity (%)	77	60
Specificity (%)	66	76
PPV (%)	70	72
NPV (%)	74	65

CI = confidence interval; PPV = positive predictive value; NPV = negative predictive value.

5.5 DISCUSSION

SVMs were utilised to correctly classify CDMS (or the absence of clinical conversion) at one and three years in 71.4 % and 68 % of CIS patients respectively in a cohort of 74 patients. Information was derived from individually labelled brain scans and associated clinical information and results were averaged over 100 bootstraps with balanced training data sets using leave-one-out cross-validation. Patients who present with CIS in a neurological clinic today are told that they have a long-term risk for CDMS of 60-80 % when white matter lesions are seen on the brain scans. Depending on the number and location of brain lesions they have a low, medium and high conversion risk to MS [116]. The relative risk of developing CDMS for female patients compared to males is 1.20 (95 % CI 0.98-1.46) [34]. It must be noted however, that these findings come from group studies and there are strong limitations in accuracy (sensitivity and specificity) when extrapolating radiological and clinical predictors to individual cases in routine clinical practice. Machine-learning-based classification has the strong potential that it can overcome these limitations and be used for a single subject (or individualised) prediction of clinical conversion to MS, which may lead to a more tailored prognosis, which, in turn, would translate into more timely and better-informed treatment choices. Further benefits could be expected for the preparation of clinical trials and research studies where the accurate prediction of prognosis from individual subjects' scans could help selecting patients.

The average accuracies of 71.4 % and 68 % obtained with SVMs in this study are lower than those reported in previous applications of SVMs to other neurological diseases such as Alzheimer's diseases, Huntington's disease or depression [63, 65, 89]. However, it is important to note that the classification of patients into those who will develop MS within a short-term follow-up and those who will not is a more challenging problem than classifying patients vs. healthy subjects [51, 129], since some of the patients in the non-converter group may still develop MS in the long-term. Studies focussing on a similar classification task on patients with mild cognitive impairment (MCI) obtained lower or similar accuracies with a range from 62 % to 75 % for distinguishing between MCI-stable patients and MCI patients who convert

to Alzheimer's disease [136]. This behaviour is also described in the classification landscape in section 1.4.

MOST RELEVANT FEATURES

Looking at the results shown in Tables 10 and 11, it can be seen that the preferred lesion measures using both methods in this study are lesion load and lesion count rather than other features such as lesion size. Similar results were found in previous clinical studies [78]. For the prediction of CDMS at three-year follow-up, we found that the distance of lesions from the centre or the vertical axis of the brain were selected as predictive features, which suggests that lesion location is indeed associated with CIS conversion as suggested in several studies regarding the corpus callosum [56], the brainstem [118], or the corona radiata, optic radiation, and splenium of the corpus callosum (periventricularly) [30]. In our study, a shorter distance of the lesions to the vertical axis of the brain was seen more often in converters than non-converters. Lesion probability maps have been recently used to correlate high lesion frequency in specific white matter regions with conversion to MS [50].

Clinical and demographic features such as age or gender are known to play an important role in the conversion from CIS to MS on a cohort level [34, 104]. We have shown that these biomarkers are also informative when predicting individual patients as they are present in the combinations of features associated with the highest accuracy for classification at three and one year respectively. In particular, it can be seen that younger, female patients convert to MS more often than older, male patients. The type of CIS is relevant at short-term conversion of one year, where patients with a spinal cord type are more likely to convert to MS. Generally, it can be seen that the performance in predicting clinical outcome is considerably higher when using combinations of both MRI and clinical/demographic features rather than individual features. This indicates that it is crucial to combine information from various sources to obtain the highest possible accuracy for classification of individual patients.

Complex models (i.e. models with higher degree polynomial kernels) with a high-dimensional features space (i.e. many features) should always classify training data better than simpler models but are likely to overfit the data. The cross-validation test

in out study, however, reduces this effect and allows for generalisation to unseen test data as well as for identification of the best feature combinations. Consequently, the models leading to the highest average accuracies contain only a small number of features (three and six respectively at 1- and 3-year follow-up) and do not use the highest possible kernel degree of five, even though more complex models with up to 12 features were possible. This suggests that the obtained prediction models are indeed based on the intrinsic structure of the data and are not the result of an overfitted model.

LIMITATIONS

It is important to ensure that no subject that has been used to train the classifier is used again for testing when performing supervised learning because the classifier is expected to perform particularly well on previously seen data. Having completely independent data sets for training and test would be ideal to avoid any bias. In practice, however, this is often not possible because available data sets are usually relatively small as it was in the case of this study. Leave-one-out cross-validation is one approach to partially overcome this problem, but it has the disadvantage that it generally introduces a positive bias in the accuracy estimate [67]. Due to this bias, the actual values for the accuracy of our models are likely to be lower on actual unseen data. However, it can be expected that the comparison and ranking of the feature combinations remains valid since all feature combinations in this study were tested with the exact same methods.

The choice of features that need to be selected to perform the experiments with machine learning techniques is absolutely crucial [20]. In this study we only used a small set of high-level features that were selected a priori and were associated with white matter lesions (visible on T2-weighted scans). These measures are known to be of value in the development of MS [78] and discriminate between MS and healthy subjects [51]. However, the feature set is mainly based on lesion masks, which are manually created by an observer, rather than the outputs of automated image analysis and pattern recognition methods. This means that expensive human interaction is necessary to first create the lesion masks by an expert rater and then select ap-

propriate features to be included in the analysis. However, there are advances in research on automated lesion segmentation [48], which could be incorporated in future image analysis pipelines. Similarly, recent advances in machine learning research using depth sensors indicate that it might soon be possible to automatically assess EDSS in patients [29] so that future work can focus on more automated approaches of combining high-level information.

There are many more potential biomarkers that have shown promising results with respect to prediction of MS prognosis but were not available in the data sets used here. Non-standard MRI such as magnetisation transfer imaging (MTR) has been founds to reflect damage outside of MS lesions [2] but results as an independent predictor are inconclusive [47, 119] so that a combination with other measures would be of interest. Para-clinical measures such as intrathecal synthesis of oligoclonal bands [117], grey matter atrophy [13], and genetic factors [59] are discussed in the context of MS but have never been used in a machine learning setting. This is also true for more clinical features such as the presence gadolinium-enhancing lesions, which allow the diagnosis of MS in CIS patients without a follow-up MRI scan or a second attack [99, 101], spinal cord lesions, which could add predictive value for patients with a non-spinal-cord type of CIS [55, 112] or cortical lesions, which indicate GM damage but need additional double-inversion recovery (DIR) or phase-sensitive inversion recovery (PSIR) MRI acquisition [42].

On the other hand, it can be seen as an advantage that the available data set was limited to conventional MRI acquisitions and very basic demographic and clinical features. This information can be obtained in any clinical centre even if they lack specialist research expertise so that a machine learning model such as ours could support the local physicians in their patient management.

It should be noted that even though the recursive feature elimination algorithm is a useful method to identify relevant features, it is possible that it only finds a local maximum solution rather than the actual most predictive combination of features. The alternative option of exploring all possible combinations of features exhaustively is very tedious, computationally expensive and leads to a multiple comparisons problem. Testing $2^{12} - 1 = 4095$ different models on the same classification task is likely

to identify a combination of features that spuriously performs well on this relatively small data set but would not generalise well to unseen data. The fRFE algorithm, on the other hand, is more likely to perform well on unseen data and also inherently controls for redundant features as it only adds the one feature with the highest information gain at each iteration. If two features contain the same information only one of them will be selected. This resulting feature set is not necessarily the only one informative about the classification task, since some highly correlated features may have been rejected. In the case of our study, however, we found that this is not the case and there is indeed only one combination of the 12 analysed features that leads to the reported accuracy values.

This work can be extended to confirm our findings in a larger data set, which could divide the data into separate training and testing sets. It would also be of interest to combine data from multiple MS centres in order to reduce centre effects and increase the generality of our method. Additionally, it is possible to evaluate and compare the classification rates for progression of disability or clinical outcome from novel algorithms, such as the event-based model recently applied to Alzheimer's and Huntington's disease cohorts [44, 135].

6

HIGH-LEVEL FEATURES II

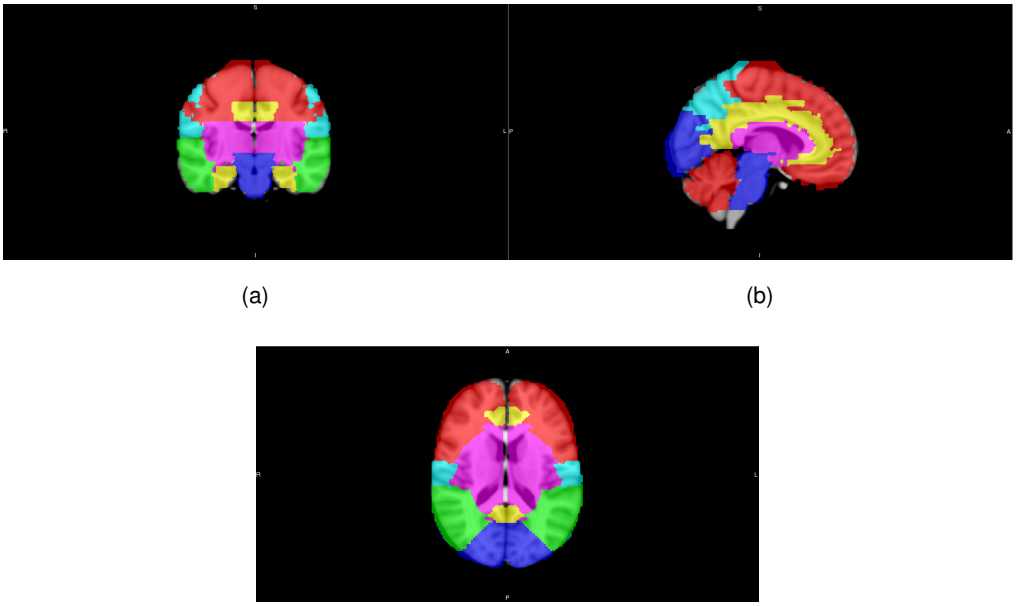
Parts of this work have been published in Magnetic Resonance Materials in Physics, Biology and Medicine [133].

It is evident form the experiments presented in chapter 5 that many of the tested features were not selected for the most informative feature combination using our SVM approach and therefore can be assumed to contain little predictive power. A modified set of features was therefore examined with regard to predictive power in the task to predicting clinical conversion to MS at 1-year follow-up. In particular, it can be seen that features regarding MRI voxel intensity and lesion location were not contributing to the final classification model. We removed these features, and instead increased the number of lesion-based measures and performed a rough parcellation of the brain in order to capture lesion load and count in different brain areas (see Figure 14). To this end, the Talairach atlas [114] was used at second-level detail as described in [http://www.talairach.org/labels.html]. To reduce the number of features the 12 Talairach ROIs were merged into larger coherent structures resulting in the following ROIs that were finally utilised: brainstem, cerebellum, frontal lobe, temporal lobe, occipital lobe, parietal lobe, limbic lobe and sub-lobar structures (see also Figure 14). The atlas was diffeomorphically registered from ICBM-MNI-152-space to the individual patients' native spaces. Furthermore, we obtained brain masks and GM/WM probability maps and used these to quantify the volume of WM, GM and intracranial structures. The revised feature set is as follows:

1. Global lesion count: total number of lesions in the brain.
2. Global lesion load: total lesion volume.
3. Lobar lesion count: number of lesions in the eight ROIs.
4. Lobar lesion load: lesion volume in the eight ROIs.
5. Mean lesion size.

6. Standard deviation of lesion size: measure of lesion size variability.
7. Size of the smallest lesion.
8. Size of the largest lesion.
9. Intracranial volume.
10. White matter volume.
11. Grey matter volume.
12. Age.
13. Gender.
14. Onset location.
15. EDSS.

The resulting feature types are similar to the ones used in chapter 5, so again we place this study in the classification landscape as a high-complexity task with medium- to high-complexity features due to the usage of ROIs and high-level measures.



(a) (b)

(c)

Figure 14: Parcellation of the brain in eight areas following the Talairach atlas. Coronal (a), sagittal (b) and transversal (c) view of the ICBM-MNI 152 brain overlaid with the 8 ROIs.

6.1 EXPERIMENT DESIGN

In this study, the data set described in section 5.1 was used again to keep results comparable. However, we focussed on the one-year follow-up only since short-term conversion is generally of higher interest. All possible combinations of the 15 features ($2^{15} - 1 = 32767$) were tested to avoid local-maximum results. As explained in section 5.3, the number of patients in each class needs to be balanced for the SVM classification which has been done with 100 bootstraps without replacement as described previously. The data was cross-validated using the LOO method. Both the polynomial and the RBF kernel have been tested in order to find the most predictive model. Results are reported as averages over all 100 bootstraps.

6.2 RESULTS

The highest obtained average accuracies were 73.5 % (sensitivity/specificity 73 %/ 73 %) with the polynomial kernel and 71.6 % (sensitivity/specificity 75 %/69 %) with the RBF kernel. In fact, two slightly different feature combinations provided the same accuracy with the polynomial kernel and three different combinations provided the exact same mean accuracy using the RBF kernel.

The selected features for the final model using the polynomial kernel are global lesion count, lobar lesion load, lobar lesion count, GM volume, onset location, and age. The second model providing the same mean accuracy also included brain volume. The first model has a relatively large spread of accuracies in the 100 bootstraps ranging from 59 % to 91 %, whereas the second model is slightly more stable and ranges from 62 % to 84 %.

Using the RBF kernel, the selected features are lobar lesion count, lobar lesion load, mean lesion size, WM volume and all clinical and demographic features, as well as brain volume, GM volume or both. All three combinations provide the same average accuracy of 71.6 % with a range of 55 % to 86 % across the bootstraps. More detailed results can be found in Table 12

Table 12: Features used in the most predictive feature combinations for classification of CIS converters and non-converters at a one-year follow-up. Results are shown for two different kernel types: polynomial and RBF. For each kernel function we show multiple combinations as they respectively provide equal accuracy. Classification outcomes are presented using accuracy, sensitivity, specificity, PPV and NPV.

MRI features	**Polynomial**		**RBF**		
Global lesion count	•	•			
Global lesion load					
Lobar lesion count	•	•	•	•	•
Lobar lesion load	•	•	•	•	•
Mean lesion size			•	•	•
Std of lesion size					
Smallest lesion					
Largest lesion					
Brain volume		•		•	•
WM volume			•	•	•
GM volume	•	•	•		•
Clinical features					
Type of onset	•	•	•	•	•
Age	•	•	•	•	•
Gender			•	•	•
EDSS at onset			•	•	•
SVM-based classification					
Accuracy %	73.5	73.5	71.6	71.6	71.6
Range %	59-91	62-84	55-86	55-86	55-86
Sensitivity %	73	74	75	75	75
Specificity %	73	72	69	69	69
PPV %	73	73	71	71	71
NPV %	73	73	73	73	73

PPV = positive predictive value; NPV = negative predictive value.

6.3 DISCUSSION

The modified feature set leads to a 2 % higher accuracy for the one-year follow-up compared to the experiments in chapter 5 which indicates that the classification is very sensitive to the included measures. Comparing the different models shown in Table 12, it can be seen that some features appear in all top-performing combinations, which suggests a particularly important role of lobar lesion count, lobar lesion load, CIS type and age. The lobar lesion measures show a clear benefit over global measures suggesting that a more refined parcellation of the brain might improve findings even further.

7

ROI-BASED FEATURES

In the previous chapters we showed that the choice of features is the most important part in a classification project. Specific feature combinations seem to perform better than a collection of all features because classifiers are able to reduce the weights of certain 'noise' features but cannot eliminate their influence completely. In this chapter we present a study comparing different types of features, which are combined into coherent groups, and apply them to the three common classifiers Linear SVM, RBF SVM and Random Forest. The aims of these experiments are a) to identify predictive feature groups and b) to compare classifier performance.

The classification experiments will be again performed on baseline data of patients with CIS. We use three different data sets both independently and in a multi-centre setting. The features used in this study include high-level features such as EDSS, CIS onset type or lesion masks but the majority of the included measures are local measures automatically derived from MRI. The two parts of this study are placed as medium and medium-to-high complexity in the feature space in a highly complex task in the classification landscape shown in Figure 1.

7.1 DATA

This is a retrospective study performed on data that was obtained by three centres in Barcelona/Spain, London/UK and Siena/Italy. It is part of a larger cohort acquired by the Magnetic Resonance in Multiple Sclerosis (MAGNIMS, www.magnims.eu) research group. The total number of included patients is 296, and 66 (22.3 %) of them converted from CIS to CDMS within one year. Additional follow-ups were available for the centres Barcelona (3 and 5 years) and London (3 years). The available data varies slightly between centres since local protocols were used for the origi-

nal studies. For all patients T1-weighted MRI, PD/T2-weighted MRI, manually outlined, binary lesion masks, demographics (age and gender), and clinical information (type of CIS and EDSS) was obtained. In-plane MRI resolution was approximately $0.95 \times 0.95\,\mathrm{mm}^2$ with a slice thickness of 3 mm. Lesion masks were manually drawn from PD/T2-weighted MRI based on the centres' internal protocols. The same MRI scanner has been used for all scans at each respective centre but scanner types vary between centres. Ethics approval and patient consent was obtained prior to the study. Inclusion criteria for this study were the availability of the previously mentioned data, and the presence of WM brain lesions (i.e. non-empty lesion masks). Detailed information on the data is given in Table 13 and Table 14.

Table 13: Demographic and clinical characteristics of the cohort used for experiments in chapter 7.

Centres	**Barcelona**	**London**	**Siena**
# patients	176	72	48
# MS converters at 1y	34 (19.3 %)	22 (30.6 %)	10 (20.8 %)
# MS converters at 3y	78 (44.3 %)	29 (40.3 %)	NA
# MS converters at 5y	95 (54 %)	NA	NA
Gender	51M/126F	28M/44F	22M/26F
MRI data	—- T1, T2, PD —-		
Clinical data	—- EDSS, CIS type —-		
Demographic data	—- age, gender —-		
Median EDSS (range)	2 (0-6.5)	1 (0-8)	2.5 (0-2.5)
Mean age (range)	32 (16-50)	34 (19-50)	32.5 (21-54)
Onset type (brainstem, optic nerve, hemispheric, spinal cord, multifocal)	52/45/30/ 49/0	6/62/0/ 4/0	10/8/9/ 18/3

Table 14: Overview of class characteristics for all centres at 1-year follow-up.

group	**converters**	**non-converters**	**all**
Gender	49F/17M (74.2 %/25.8 %)	146F/84M (63.5 %/36.5 %)	195F/101M (65.9 %/34.1 %)
Mean age (range)	32.3 (16-50)	32.7 (16-54)	32.6 (16-54)
Median EDSS (range)	1.5 (0-8)	1.5 (0-5.5)	1.5 (0-8)

7.2 IMAGE PROCESSING

For this study, a comprehensive image processing pipeline was created to calculate the various features used in the for classification experiments. The individual steps are described below.

1. N4 correction: all MRI scans were initially corrected for bias field inhomogeneities using the N4 algorithm [121].

2. Registration: lesion masks were created from PD/T2-weighted images whereas most other image processing is performed in T1 space. Therefore, the PD/T2-weighted MRI scans were affinely registered to T1 space using `reg_aladin` from the NiftyReg toolbox [84]. Lesion masks were subsequently resampled using the obtained transformation parameters.

3. Brain parcellation: we showed in chapter 6 that even a rough brain parcellation has a beneficial effect on the classification outcome. Therefore, we perform a more refined brain parcellation here using the GIF (geodesic information flows) algorithm [15]. This tool segments the brain into 143 ROIs, of which most are cortical areas as shown in Figure 15. To evaluate the effect of the level of detail in the brain parcellation, the GIF-ROIs were merged into nine larger areas. Most of these areas correspond to the anatomical brain lobes, which is why we refer to all of them as 'lobes' in the context of this work. These 'lobes' were limbic, insular, frontal, parietal, temporal, occipital, cerebellum, GM and WM. In addition to the 143 ROIs, the algorithm also provides a segmentation of GM and WM, as well as binary masks of intracranial volume and all brain tissue. A complete list of the GIF-ROIs can be found in Appendix A.1.

4. Cortical thickness: cortical thickness (CT) was calculated using DiReCT, a registration-based algorithm [32]. It has been shown to have the same degree of reproducibility as the more commonly used Freesurfer method [122] but is much faster once WM and GM density maps are available. The algorithm works reliably as long as the WM and GM density maps used as input are of

good quality. Here, we used the state-of-the-art probabilistic segmentations from GIF.

5. ROI masking. We used the ROIs from 3 to calculate local information from GM/WM probability maps, T1, T2, PD MRI intensities, CT maps, and the lesion masks.

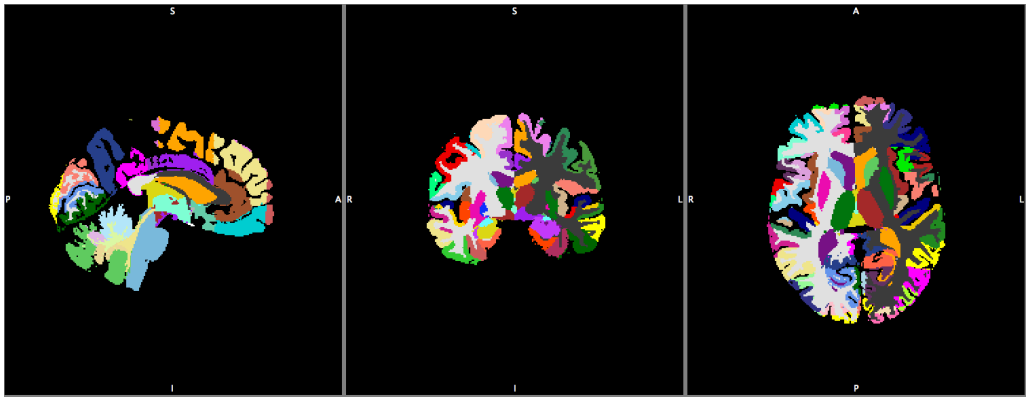


Figure 15: T1-weighted MRI scans were parcellated using the GIF algorithm. This figure illustrates the resulting ROIs used in this study.

7.3 FEATURE DEFINITIONS

Following the image processing, an extensive list of features has been defined on different ROI scales as follows.

Global features The following nine non-local features have been included:
global lesion count, global lesion load, brain volume, GM volume, WM volume, age, gender, EDSS, CIS onset type.

GIF-ROI features and lobar features The following 9 features were calculated both on the level of the 143 ROIs from the GIF parcellation and on the level of the 9 lobe ROIs (resulting in 1287 and 81 features respectively):
lesion count, lesion load, CT, WM, GM, volume, T1 intensity, T2 intensity, PD intensity.

Some of the listed features actually do not make sense from an anatomical perspective and are merely the result of the ROI masking described in section 7.2 step 5. There is no cortical thickness to be measured in white matter structures or no white

matter lesions in grey matter ROIs for instance. As a result, there are 36 features that have a value of zero for all analysed patients and were consequently removed. It should be noted that one would expect more than 36 features to have a value of zero (e.g. due to the high number of cortical areas, which should not contain WM and therefore also no lesions). However, the tissue class segmentation provides a probabilistic atlas, which can have non-zero values for WM density in GM areas. In addition to this, it is possible that WM lesions close to the cortex are incorrectly assigned to cortical ROIs due to small errors in the registration. This has not been corrected for because it affects both patient groups in the same way seeing that the lesion distribution is comparable in both classes (see also section 2.6 and Figure 5).

The feature matrix has been normalised across the individual feature axes. This centres the data to zero mean with unit variance following $x' = \frac{x-\bar{x}}{\sigma}$, where x' is the normalised vector, $\bar{x}$ the mean and σ a feature's standard deviation.

7.4 EXPERIMENT DESIGN

This is an exploratory study with the aim of identifying feature types that are predictive of conversion from CIS to CDMS. We explore multiple follow-up durations in three single centres as well as a combination of these in a multi-centre setting.

Due to the large number of 1341 included features it is not possible or advisable to look at every single combination of features as it has been done in chapters 5 and 6 since this would lead to $2^{1341} - 1$ experiments, which firstly is not feasible to perform within a reasonable time frame and secondly and more importantly would give rise to a multiple comparisons problem and consequent spurious findings. For similar reasons, it is also not advisable to run a recursive feature elimination in the way proposed in chapter 5.

Instead, we decided to follow two separate approaches:

a) Manual feature combination: we pool the features into coherent groups and analyse them separately using a set of three common classifiers: Linear SVM, RBF SVM and Random Forest.

b) Automated feature selection: we perform a recursive feature elimination (RFE) with a modified algorithm compared to chapter 5.

A) MANUAL FEATURE COMBINATION

We grouped the features as follows:

1. all: a vector containing all 1341 features for each patient.
2. global: all global features as defined in section 7.3.
3. ROI: all 1287 features based on GIF-ROIs.
4. lobes: 81 lobe ROIs consisting of merged GIF-ROIs as described in 7.3.
5. lesions: features containing lesion count and lesion load at different ROI size levels.
6. nonlesion: all non-lesion features.
7. derived: all automatically derived measures such as cortical thickness and GM/WM density.
8. CT: all cortical thickness measures at different ROI size levels.
9. GMWM: GM/WM density maps at different ROI size levels.
10. imaging: T1-, T2-, and PD-weighted MR intensities at different ROI size levels.
11. global_derived: a combination of feature groups 2 and 7.
12. global_imaging: a combination of feature groups 2 and 10.
13. global_lesions: a combination of feature groups 2 and 5.
14. global_lobes: a combination of feature groups 2 and 4.

SVM CLASSIFIER The experiments have been performed using a linear SVM and a RBF-kernel SVM independently. The SVM classifiers were used with nested cross-validation [123]. The data was first divided into a training and a testing set, then the training data was divided again to estimate the cost and scaling parameters within a nested loop. This procedure is necessary to reduce the overfitting bias in the parameter optimisation. Afterwards, a SVM was trained using the complete training set and the performance was evaluated on the left out test set.

RANDOM FOREST CLASSIFIER For the Random Forest classifiers, a normal cross-validation was used because no parameters were optimised for each data set. At each node 37 features were randomly sampled and 500 classification trees were created for each forest. 37 is the square root of the total number of features, which is often used as a default setting as it is a good compromise between computational expense and feature space exploration. We did not restrict the size of leaf nodes for splitting but the employed tree implementation from scikit-learn avoids overfitting using an online bootstrapping approach [96].

All experiments in a) were repeated 1000 times to reduce sampling bias and perform stable statistics.

B) AUTOMATED FEATURE SELECTION

The RFE algorithm is defined as follows:

We create initial models using all features in a Linear SVM with 250 repetitions. Each of these 250 models provides weights for the individual features, where the value of the weights indicates the importance of the feature in the model. These weights can vary strongly between models and overfit to the individual data sets. In particular, it is possible that a feature contributes positively in one model and negatively in another one, and hence no strong evidence can be derived from the individual models. By averaging the coefficients over all 250 repetitions, we obtain mean weights for all features, which allow for better generalisation to the whole data set. Now, it is possible to identify features, which have only a very small contribution to the model (i.e. are associated with small average weights). The lowest 25 % of the features are then removed, and the process repeated with the remaining 75 %. In each subsequent iteration 25 % of the features are removed until only 5 features are left after 20 steps.

Due to higher model complexity, we did not use an RBF-kernel SVM for this set of experiments. Similarly, we did not perform the RFE in combination with Random Forests because this type of classifier has an inherent feature selection, which should be sufficient if the number of trees in the forest is high enough. This condition can be assumed to be satisfied with the 500 trees used in experiment set a) [93].

The performance of the classifiers was evaluated using balanced accuracy, accuracy, sensitivity, and specificity. Statistical significance was estimated from the balanced accuracy. Results were considered significant if less than 5 % of the repetitions had a random result (i.e. $\leqslant$50 % balanced accuracy). This is illustrated as a histogram for one set of features in Figure 16.

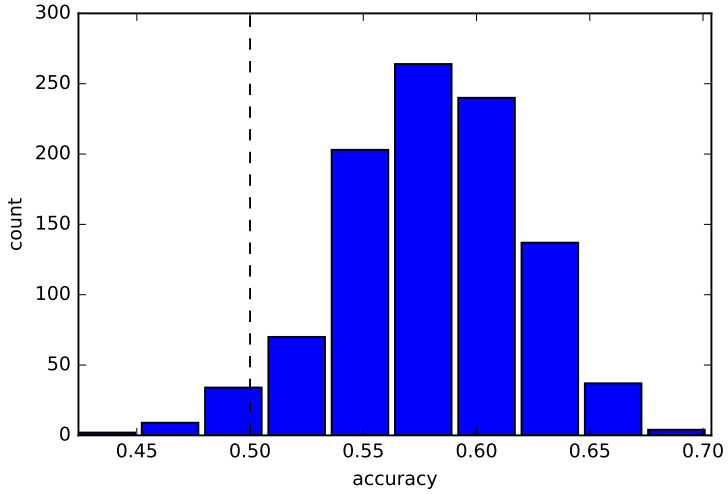


Figure 16: p-values are calculated as the proportion of sampling repetitions below the threshold of 50 %. They were considered statistically significant if $\mathrm{p} \leqslant 0.05$.

7.4.1 *Patient sampling*

An imbalance in class sizes can lead to a bias in the model and make it more likely for the classifier to map unseen data to the larger group because it has learned a higher variability there [67]. In these cases, a high overall classification accuracy is not necessarily a sign of a good model but merely the result of a dominating large group.

Downsampling can be used to avoid this imbalance bias. As in previous chapters, we sampled as many subjects from the larger class as there are subjects in the smaller class. For most experiments in this chapter there are more CIS-stable patients than converters due to the short follow-up duration so we randomly sampled patients from the non-converter group to match the size of the smaller converter group. The only exception is the 5-year follow-up in the Barcelona data set where the subsampling was performed vice versa. It is obvious now that this random sampling introduces a bias because it is not guaranteed that the downsampled group is representative of the original cohort because it might, by chance, only contain outliers for instance. However, it can be assumed that this bias can be averaged out by repeating the sampling procedure many times. For the manually combined features we repeated all experiments (and samplings) 1000 times and the automated feature selection experiments were repeated 250 times.

7.5 RESULTS

7.5.1 *Manually grouped features*

We present the classification results as balanced accuracy averaged over 1000 bootstraps for the three classifiers Linear SVM, RBF SVM and Random Forest as well as an average of the three classifiers. In this section, we focus on the findings with 2-fold and 10-fold cross-validation because they provide better generalisation to

unseen data. Leave-one-out cross-validation has been performed as well and the complete results can be found in the Appendix Tables 18 to 41.

Depending on the data set, follow-up, cross-validation and classifier, the highest mean balanced accuracies range from 41.4 % to 69.7 % but for most cases the accuracy is approximately 60 %.

For each data set and follow-up, there are specific feature types that are quite consistently associated with the highest classification performance regardless of the classifier or the type of cross-validation. We describe the main findings for each data set separately. All result tables can be found in Appendix Tables 18 to 41. Results that are non-random at a significance level of $\mathrm{p} = 0.05$ are indicated with a *.

BARCELONA, 1-YEAR FOLLOW-UP The highest accuracy using a 2-fold CV was 59.7 % using a combination of global and lesion features with a linear SVM. Similarly, lesion features led to an accuracy of 63.2 %* in a 10-fold CV with a linear SVM. On average, global and lesion features provided accuracies of 57.8 % and 59.4 % using 2-fold and 10-fold CV respectively. In the case of 10-fold CV and averaged classifiers outcomes, both lesion features and a combination of global and lesion features provided the same mean accuracy.

BARCELONA, 3-YEAR FOLLOW-UP Global features dominate the classification results in this set of experiments along all classifiers and cross-validation types. The highest accuracies were 61.9 %* and 62.5 %* at 2-fold and 10-fold CV respectively obtained with Random Forests. The average accuracies across classifiers were 60.9 % and 61.2 % for the two CV types respectively.

BARCELONA, 5-YEAR FOLLOW-UP As for the 3-year follow-up, the global features provide the highest mean balanced accuracies with 59.4 %* and 60.8 %* for 2-fold and 10-fold CV respectively when using the linear SVM classifier. Also when averaged over classifiers, global features are most predictive with 58.2 % for 2-fold and 59.3 % for 10-fold CV.

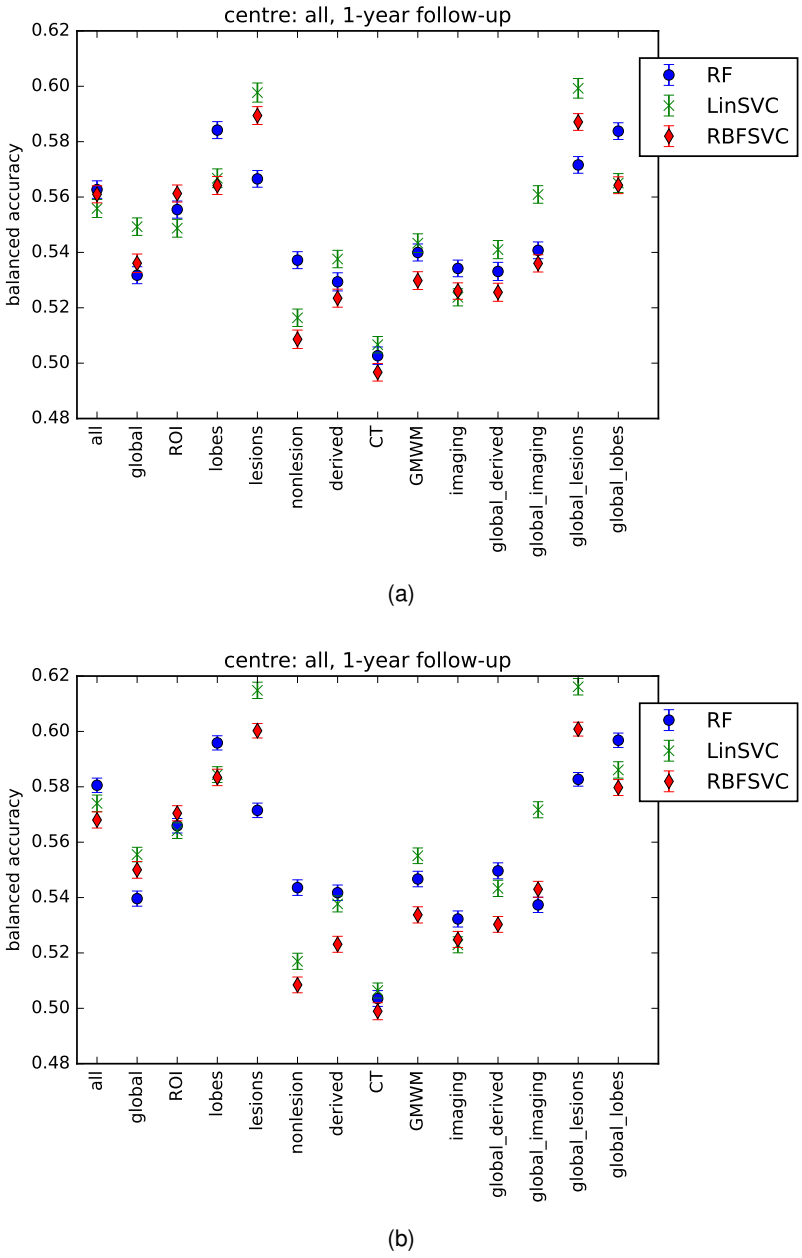


Figure 17: Qualitative overview of balanced accuracy using different types of features. Plots shown for multi-centre data at 1-year follow-up using 2-fold cross-validation in (a) and 10-fold CV in (b). Error bars indicate the 95 % confidence intervals over 1000 repeated samplings.

LONDON, 1-YEAR FOLLOW-UP In this data set, global features were most dominant with accuracies of 63.6 % for 2-fold CV and 66.3 %* for 10-fold CV using linear SVMs. The other classifiers had slightly worse performance leading an average accuracy of 59.5 % and 62 % for the two CV types respectively when using global features.

LONDON, 3-YEAR FOLLOW-UP The best performance for the longer follow-up is similar to the 1-year results with an average balanced accuracy of 63.8 %* with a 2-fold CV and 68.7 %* with a 10-fold CV and a linear SVM. The highest classifier averages using global features were 59 % and 62.9 % for 2-fold and 10-fold CV respectively.

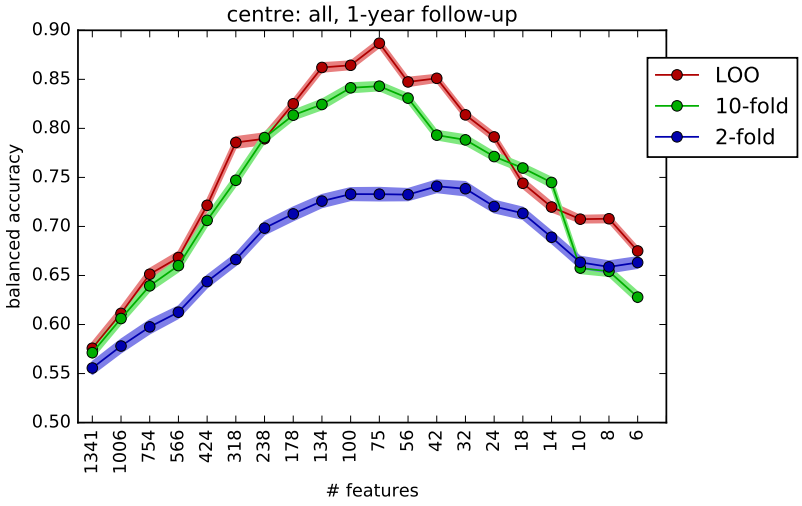
SIENA, 1-YEAR FOLLOW-UP In this data set, the GMWM features were most dominant leading to mean accuracies of 66.2 % and 69.7 % for 2-fold and 10-fold CV with Random Forest classifiers. The other classifiers performed similarly so that the averages are 64.9 % and 69.7 % respectively for the two types of CV.

MULTI-CENTRE, 1-YEAR FOLLOW-UP The most predictive feature set in this multi-centre setting is the combination of global and lesion features for both cross-validation types with an accuracy of 59.9 % and 61.6 %* respectively when using a linear SVM. The highest balanced accuracies averages over all three classifiers are also achieved with the combination of global and lesion features with 58.6 % for 2-fold and 60 % for 10-fold CV. The obtained accuracies for all feature types are exemplified in Figure 17 (a).

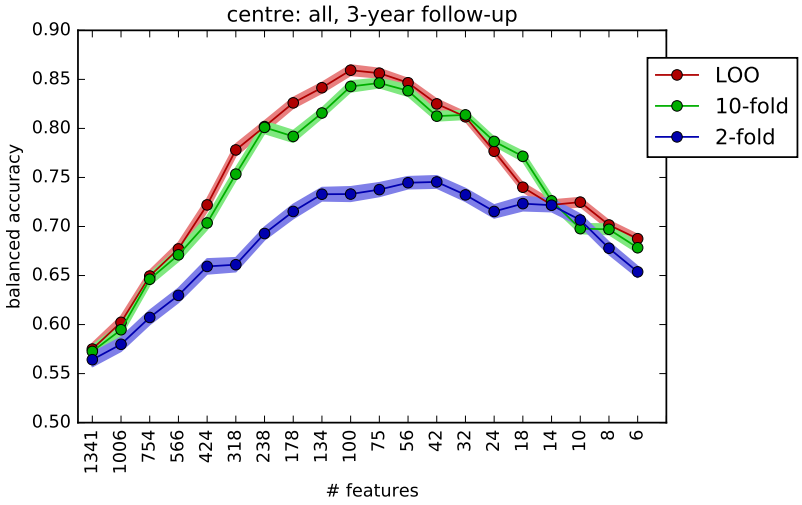
MULTI-CENTRE, 3-YEAR FOLLOW-UP The highest mean accuracies here are slightly lower compared to the 1-year follow-up. An outcome of 58.3 % was achieved with a combination of global and lesion features for the 2-fold CV using linear SVMs. Both lesion features as well as the combination of global and lesion features led to an accuracy of 60.4 %* for the linear SVM model with 10-fold CV. Lobe features (and the combination of global and lobe features) also provided overall high accuracies so that the highest average accuracies across all classifiers were 57.4 % and 58.9 % respectively achieved with the combination of global and lobe features using a 2-fold CV and with both lobe features and the combination of global and lesion features using a 10-fold CV. An overview of all accuracies obtained with the different feature types and classifiers can be found in 17 (b).

7.5.2 *Automated feature selection*

The RFE algorithm repeatedly removes the 25 % of features that contribute least to the classifier model. Generally it can be observed that the mean balanced accuracy increases in the first iterations and reaches its peak value when using between 30 and 150 features. For lower numbers of features, the accuracy decreases again. This behaviour is exemplified in Figure 18.



(a)



(b)

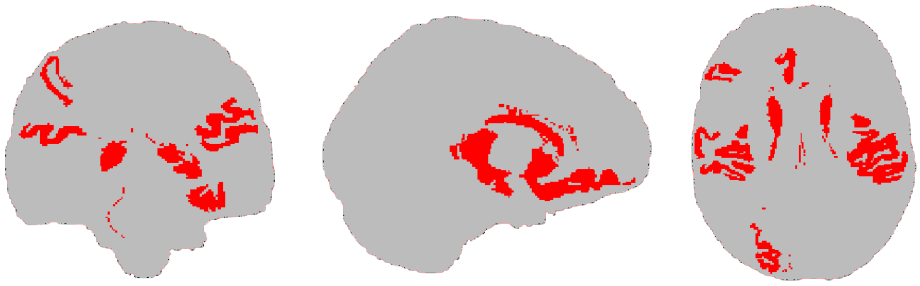
Figure 18: Balanced accuracy plotted against the number of used features at each iteration of the RFE algorithm. Shown for multi-centre data at one-year (a) and three-year follow-up (b). Shaded areas indicate the 95 % confidence intervals.

Furthermore, it can be observed that 2-fold cross-validation provides the lowest accuracies, followed by 10-fold CV and leave-one-out CV, which generally provides the highest accuracies.

Since 2-fold cross-validation is the most conservative approach and should provide the most generalizable outcome, we will only report those findings to avoid positive bias. Complete results for 10-fold and leave-one-out cross-validation can be found in Appendix Tables 42 and 43.

MULTI-CENTRE DATA The highest obtained mean balanced accuracy in the multi-centre data with one-year follow-up is 74.1 % (95 % CI: 73.5 %-74.7 %) and was achieved using 42 features after 13 iterations of the RFE algorithm. The selected features are listed in Table 16 and include features of all types except global. Selected ROIs include deep grey matter structures such as basal ganglia, thalamus, putamen and palidum, as well as insula and operculum, and the orbital gyri as shown in Figure 19. The same parameters resulted in an accuracy of 74.5 % (95 % CI: 73.9 %-75.1 %) for the three-year follow-up using similar feature groups and ROIs. The most striking difference being the selection of EDSS as a predictive feature for the longer follow-up as shown in Table 17.

SINGLE-CENTRE DATA The application of data from the individual centres results in higher mean accuracies compared to the multi-centre data. The exact results are 80.2 % (95 % CI: 79.4 %-80.9 %) in the Barcelona data set for the one-year follow-up, 83.9 % (95 % CI: 83.5 %-84.3 %) for the three-year follow-up and 86.0 % (95 % CI: 85.6 %-86.4 %) in the five-year follow-up. The classifier performance is even higher in the London and Siena data sets with 92.0 % (95 % CI: 91.3 %-92.6 %) using data from London with a one-year follow-up, 92.2 % (95 % CI: 91.6 %-92.7 %) with a three-year follow-up and 96.6 % (95 % CI: 96.2 %-97.0 %) using the Siena data set. An overview of these findings is given in Table 15.



(a) (b) (c)

Figure 19: Illustration of ROIs selected by RFE classifier at one-year follow-up (excluding background, which is considered an artefact).

Table 15: Results of RFE experiments using a 2-fold cross-validation. Findings are shown as balanced accuracy, range over all repetitions and and 95 % confidence interval (CI) for the included centres.

centre	follow-up	bal. acc. (%)	range (%)	CI (%)	# features
multi-centre	1	74.1	61.4-87.9	73.5-74.4	42
multi-centre	3	74.5	58.0-87.5	73.9-75.1	42
Barcelona	1	80.2	58.8-94.1	79.4-80.9	56
Barcelona	3	83.9	72.4-92.3	83.5-84.3	56
Barcelona	5	86.0	77.8-93.2	85.6-86.4	100
London	1	92.0	72.7-100	91.3-92.6	75
London	3	92.2	74.1-100	91.6-92.7	75
Siena	1	96.6	80.0-100	96.2-97.0	56

7.6 DISCUSSION

MANUALLY GROUPED FEATURES The classifier models predicted the conversion or non-conversion from CIS to CDMS with mean balanced accuracies between 57.9 % and 61.6 % in multi-centre settings using global features in combination with local lesion and lobar measures. Single-centre results are slightly higher with mean accuracies between 59.7 % and 69.7 %. We used linear SVMs, RBF-kernel SVMs and Random Forests on all data sets using feature derived from MRI and clinical information. Data sets were balanced with respect to class sizes and all experiments were repeated 1000 times to reduce sampling bias.

Generally, it can be observed that data sets with smaller sample sizes show increased accuracy in prediction and a higher number of folds for the cross-validation

also leads to higher accuracies. The latter is a well-known effect, which arises from the fact that the classifier models are highly correlated when the number of folds is high. In the extreme case of leave-one-out cross-validation, the data sets used create the models are almost identical when compared pair-wise with only one patient being different each time [67]. As a result, the decision boundary will barely move between classifiers and hence produce a positively biased outcome.

The increasing accuracy arising from smaller data sets may seem surprising because one would expect a worse classifier performance if there is only little data available to train the model. However, it must be noted that the data from single centres is much more homogeneous, which makes it more likely for global features such as EDSS or lesion count to be selected since they can show high variability between centres [41, 91] (see e.g. results from Barcelona at 3- and 5-year follow-up or London at both follow-ups). Using the heterogeneous multi-centre data, the model needs to identify features that are informative for 132 or 214 patients ($2 \times$ #converters) at the same time while the smallest data set in this study (Siena data set with one-year follow-up) only contains 48 subjects with 10 converters, so only 20 patients have to be fitted, which is much more likely to result in a well-performing model. However, such a 'small' model is less likely to generalise well to the general population of CIS patients.

It can be observed that the different classifiers' prediction accuracies only differ by very few percent when the same set of features is applied. The choice of features however, can lead to very strong differences of 10 % or more. This indicates that indeed the choice of features is considerably more important than the choice of classifier.

AUTOMATED FEATURE SELECTION The classification approach using recursive feature elimination increases the mean prediction accuracies in all applied data sets and follow-ups compared to the manual grouping of features. The optimised models can correctly predict 73.9 % of the cases in a multi-centre setting at one-year follow-up and 74.3 % in a three-year follow-up. In the single-centre experiments, the obtained accuracies are even higher. As for the manually grouped features, it can be

observed that accuracies increase with decreasing sample size so that the accuracies are 79.9 %, 91.7 % and 96.1 % for the one-year follow-ups in the Barcelona, London and Siena data sets respectively. Additionally, the classification models achieve better results when using longer follow-ups in all data sets, which can be explained with the higher proportion of converters at later disease stages.

It can be observed that an increasing number of folds k in a cross-validation scheme introduces a positive bias in the accuracy estimate due to an increasing correlation of the individual classifiers at each fold as mentioned in section 3.4. When k is very high but only few data points are used as support vectors it is possible (and increasingly likely with increasing k) that a permutation of the folds will not result in a change of the decision boundary when the permutation does not affect any of the support vectors. Subsequently, the permuted test subject will be perfectly classified because the classifier model is identical to a model where the test subject was part of the training group. Thus, it is expected to see the pattern of increasing accuracy with increasing k in Figure 18.

The accuracy changes with varying number of features have a distinct parabolic pattern as shown in Figure 18. This pattern is a direct result of the proposed feature elimination approach. When using all features at the beginning of the experiments, the feature vector contains many entries that are not informative. The SVM algorithm has an internal weighting of the applied features to ensure an ?optimal? hyperplane. However, this weighting usually does not apply a weight of exactly zero (i.e. ignore a certain uninformative feature). Therefore, uninformative or noisy features always contribute to the calculation of the decision boundary and this can significantly reduce the classification accuracy. Using the proposed feature elimination scheme, these low-weight features are iteratively removed, which results in an increase of accuracy in the first iterations. Since our algorithm always excludes the lowest 25 % of features but does not apply an upper threshold on the weights, it eventually excludes features that have a significant positive contribution to the classification so that the accuracy decreases at lower numbers of features.

FEATURES SELECTED BY RFE Both multi-centre follow-ups reached their highest accuracy after 13 iterations of the RFE using 42 features. Looking at the actually utilised features for the multi-centre classification experiments in Table 16 and 17, it can be seen that many features are shared between the two sets of experiments. Unfortunately, atrophy cannot be measures from MRI scans obtained at a single time point. Several measures such as GM density, cortical thickness or ROI volume, however, can be used as a less-strong indicator of atrophy. These feature types were indeed selected as predictors in the operculum and insula, which is in line with previous research that showed atrophy in the insula to be correlated with MS [95] and GM density to be correlated with disability in MS [9]. Additionally, our analysis shows that lesion count and load in the operculum are predictive features as well. Even though this would be supported by studies like [124] it must be noted that we only included lesion measures in WM while the operculum is a cortical structure so that this finding in more likely to be explained by (mis-registered) juxtacortical lesions, which are also known to be associated with early diagnosis of MS [4]. Other regions of interest are the basal ganglia where WM/GM density, cortical thickness and ROI volume were selected by the classifier. Again, these measures are indicative of atrophy, which is indeed associated with MS in literature [6, 40]. Cortical thickness and volume have been predictive in the thalamus and the ventral diencephalon, which is consistent with other studies that identified atrophy in CIS patients [3] and showed that atrophy is stronger in MS patients [21]. Generally, deep grey matter atrophy and inflammation are associated with CIS [3, 125] but we were only able to utilise atrophy-related features in our study due to the limited data set, which did not include DIR or PSIR MRI scans that would allow for identification of GM lesions. Our classifier model selected GM density in the orbital gyrus, where atrophy has been related to different types of MS [95]. Previous studies have shown that short progression time is associated with higher disability [25] and indeed EDSS has been included by the RFE for the three-year follow-up but not for the one-year follow-up.

The areas involved in the prediction of CIS-conversion are similar for the single-centre data sets as well. Especially, deep grey matter structures and the insula and operculum are selected in all data sets even though there are minor variations in the

actually selected features. However, most of them can be linked to atrophy so that a coherent result can be observed. Due to the increased homogeneity in the smaller data sets, we can also find more global features such as age, sex or onset type. All of these features are associated with MS but their selection in our classification model could also be due to a slight overfitting to the specific data set. A definite conclusion could only be given if we repeated the experiments with cohorts that were matched for age and gender at each follow-up. This matching was not done here to avoid a reduction in sample sizes.

LIMITATIONS The feature vector at each vector size is most likely not optimal but only represents a local solution. The only way to find the actual optimum is to perform an exhaustive search over all combinations of features (at a given feature vector size) as presented in chapter 5. This, however, is not feasible with a large number of features as the number of possible combinations increases exponentially. Using recursive feature elimination, it is always possible that a potentially useful feature is eliminated at one of the first iterations because it had a low weight there. But that same feature could have had a positive effect on the outcome at a later stage with shorter feature vectors. This behaviour can be observed in chapter 5 where the exhaustive search and the RFE lead to different outcomes at three-year follow-up, which result in a difference in classification accuracy of 5.5 %.

It must be noted that the RFE also included some features such as lesion count in the skull and pial tissue or T1 intensities in the pial tissue. These are likely to be spurious findings and are caused by mis-registration or noise. In fact, most cases where MR intensities have been selected as predictive features they are located in rather irrelevant random areas such as ventricles, vessels, non-ventricular CSF, the cerebral exterior (pial tissue) or even the background and skull (compare Tables 44 to 48). Since we only excluded the least contributing 25 % of features in each RFE iteration, these features survived the feature selection process by chance without actually contributing much to the classification model.

The fact that we did not explicitly exclude features that are known to only add noise and not have a meaningful contribution to the model can be seen as a rather strong

limitation of our study because it can introduce overfitting and bias the result. However, at the same time, our results show that only very few of these noise features have been selected even though a large number of them was present. In particular, all intensity-based features can be considered noise features judging by the findings of the previous chapters 4 and 5 where we showed that MR intensities do not contain information with respect to the classification of CIS patients. This indicates that our RFE approach using a large number of repetitions is rather stable and mostly detects specific features that are known to be associated with multiple sclerosis and its disease progression. Clinical studies identifying risk factors for MS aim to generalise to large populations but their findings are not necessarily applicable to single patients. Our proposed method, however, is able to predict the clinical outcome of individual patients with an accuracy of approximately 74 %, which has not been possible using previous methods.

The features included in this study are a combination of medium- and high-level complexity measures of which only the former can be derived automatically. This is a clear limitation of this study if it was to be generalised to larger cohorts, which don't necessarily have EDSS or lesion masks available. However, it can be seen in the RFE experiments that the multi-centre results show a dominance of medium-level features containing local information from MRI, which are driving the classification in the most generalizable setting. Measures such as EDSS that can only be obtained by a clinical expert were not selected by the model for one-year follow-up and could potentially be left out in future studies. Its correlation with future disability, however, indicates increasing importance for longer follow-ups as shown in our three-year results. Lesion segmentation has been performed by expert neurologists in the data available in this study but there is also extensive research ongoing trying to derive this information automatically from MRI scans [48]. Even though these methods do not yet achieve the same accuracy as human raters, it might be sufficient to perform an automated lesion segmentation on the available MRI data for use in a machine learning setting. This would have the benefit of a consistent objective algorithm for lesion marking and consequently would avoid bias arising from inter- and intra-rater variability that is present in manual segmentation [82].

Table 16: Selected features in best result from RFE experiments using multi-centre data at 1-year follow-up with 2-fold cross-validation:

1. WM_Left_Accumbens_Area
2. WM_Right_Caudate
3. WM_Right_MFC_medial_frontal_cortex
4. WM_Right_OFuG_occipital_fusiform_gyrus
5. GM_Right_Caudate
6. GM_Left_Caudate
7. GM_Right_Pallidum
8. GM_Left_Basal_Forebrain
9. GM_Left_MCgG_middle_cingulate_gyrus
10. GM_Left_MOrG_medial_orbital_gyrus
11. GM_Right_POrG_posterior_orbital_gyrus
12. GM_Left_PP_planum_polare
13. CT_Right_Thalamus_Proper
14. CT_Left_Thalamus_Proper
15. CT_Cerebellar_Vermal_Lobules_I-V
16. CT_Right_Basal_Forebrain
17. CT_Right_ACgG_anterior_cingulate_gyrus
18. CT_Left_AIns_anterior_insula
19. CT_Left_OrIFG_orbital_part_of_the_inferior_frontal_gyrus
20. CT_Right_PIns_posterior_insula
21. CT_Right_PO_parietal_operculum
22. CT_Left_PT_planum_temporale
23. CT_Right_SPL_superior_parietal_lobule
24. volume_Right_Accumbens_Area
25. volume_Left_Putamen
26. volume_Right_Ventral_DC
27. volume_Left_CO_central_operculum
28. volume_Left_FuG_fusiform_gyrus
29. volume_Left_GRe_gyrus_rectus
30. volume_Left_MCgG_middle_cingulate_gyrus
31. volume_Right_PoG_postcentral_gyrus
32. volume_Left_PP_planum_polare
33. volume_Left_SCA_subcallosal_area
34. volume_Right_SOG_superior_occipital_gyrus
35. volume_Right_TrIFG_triangular_part_of_the_inferior_frontal_gyrus
36. lesionCount_Background_and_skull
37. lesionCount_Right_Cerebellum_Exterior
38. lesionCount_Left_Hippocampus
39. lesionCount_Left_Ventral_DC
40. lesionCount_Right_CO_central_operculum
41. lesionCount_Left_CO_central_operculum
42. lesionLoad_Left_PO_parietal_operculum

Table 17: Selected features in best result from RFE experiments using multi-centre data at 3-year follow-up with 2-fold cross-validation:

1. EDSS
2. WM_4th_Ventricle
3. WM_Left_Accumbens_Area
4. WM_Right_FRP_frontal_pole
5. WM_Left_FRP_frontal_pole
6. GM_Left_Caudate
7. GM_Right_Pallidum
8. GM_Left_Basal_Forebrain
9. GM_Left_LOrG_lateral_orbital_gyrus
10. GM_Left_MCgG_middle_cingulate_gyrus
11. CT_Right_Basal_Forebrain
12. CT_Right_CO_central_operculum
13. CT_Right_MTG_middle_temporal_gyrus
14. CT_Right_PIns_posterior_insula
15. CT_Right_PO_parietal_operculum
16. CT_Left_TTG_transverse_temporal_gyrus
17. T1_Right_Cerebral_Exterior
18. volume_Right_Accumbens_Area
19. volume_Left_Accumbens_Area
20. volume_Right_Putamen
21. volume_Right_Ventral_DC
22. volume_Right_vessel
23. volume_Left_Basal_Forebrain
24. volume_Left_CO_central_operculum
25. volume_Left_MCgG_middle_cingulate_gyrus
26. volume_Right_MOrG_medial_orbital_gyrus
27. volume_Left_MPoG_postcentral_gyrus_medial_segment
28. volume_Right_MTG_middle_temporal_gyrus
29. volume_Right_PoG_postcentral_gyrus
30. volume_Left_PP_planum_polare
31. volume_Right_SOG_superior_occipital_gyrus
32. volume_Right_TrIFG_triangular_part_of_the_inferior_frontal_gyrus
33. lesionCount_Background_and_skull
34. lesionCount_Right_Cerebellum_Exterior
35. lesionCount_Left_Ventral_DC
36. lesionCount_Right_AOrG_anterior_orbital_gyrus
37. lesionCount_Right_CO_central_operculum
38. lesionCount_Left_CO_central_operculum
39. lesionCount_Left_LOrG_lateral_orbital_gyrus
40. lesionCount_Left_MOrG_medial_orbital_gyrus
41. lesionLoad_Left_Cerebellum_White_Matter
42. lesionLoad_Left_Hippocampus

8

CONCLUSION

Machine learning models and in particular supervised classification in the sense of predicting future outcome have become very popular in the field of neuroscience but the majority of studies has been performed on types of dementia, especially on Alzheimer's disease utilising the ADNI data set. In this thesis we presented the first comprehensive overview of supervised classification models applied to clinically isolated syndromes (CIS) with the aim of predicting future conversion to clinically definite multiple sclerosis (CDMS) and put our work into context with existing classification studies by introducing the classification landscape as a visualisation of feature and task complexity.

In chapter 4 we adapted a method that has been used in other classification studies. Voxel-wise measures of grey matter and white matter density as well as MR intensity were used in single- and multi-centre settings using linear support vector machines (SVM). Even though this approach has been successful when applied to dementia cohorts or some lower-complexity tasks in MS subgroups, it did not prove to be able to distinguish between CIS-stable patients and those converting to CDMS within one year of disease onset above chance level. A correction for differences in the intensity distribution of the MRI scans did not change the results significantly. It can be concluded that there are no simple pattern present in voxel information that could be used to predict conversion to CDMS.

High-level features were introduced in chapter 5 where we used a set of twelve measures known to be related to MS. The aim was to compare the performance of individual features and combinations of them with respect to CIS-conversion at one- and three-year follow-up. In order to identify the most informative set of features we performed an exhaustive search through all 4095 possible combinations of the twelve features as well as a more systematic recursive feature elimination (RFE). We

found that certain feature combinations provide an accuracy of up to 71.4 % for the one-year follow-up and 73.5 % for the three-year follow-up which is approximately 10 % higher than what we obtained with individual features. In particular, the exhaustive search through the feature space highlighted the features lesion count and load as well as the clinical features EDSS, type of onset, age and gender as important markers.

Since MR intensity and our measures of lesion distribution did not contribute strongly to the classifier models, we modified the set of features in chapter 6. This new set included measures of volume of the brain, grey matter and white matter, as well as more local measures of lesion distribution arising from a brain parcellation following the Talairach atlas. We performed classification experiments on a single-centre data set using the one-year follow-up as a label and achieved up to 73.5 % accuracy, which is slightly higher than the 71.4 % obtained in chapter 5. The newly introduced features describing volume and local lesion distribution were indeed selected in the most predictive combination of features, indicating an important role of regional measures.

This led to a final set of experiments in chapter 7 using a very fine grained brain parcellation into 143 regions of interest and subsequently to a large set of features. We grouped features into coherent groups such as lesion, non-lesion, ROI, lobes, WM/GM density or global features and tested the performance of two SVM classifiers and a Random Forest model using these feature groups. While we could show that the best performing features came mostly from the groups global, lesion and lobes (as well as combinations of them) there was actually very little difference in classification accuracy as they all ranged between approximately 50 % and 61.6 % in the most generalizable multi-centre setting and only reached up to 69.7 % in small single-centre data sets which are likely to overfit. Furthermore, we showed that the difference in classification performance is much larger between different types of features than between different classifier models.

In addition to the manually grouped features, we presented another application of recursive feature elimination in chapter 7, which represents a more automated data-driven method for feature selection. Using this approach, we applied all 1341

features on all ROI scales (i.e. global, lobe and local ROIs) to linear SVMs and iteratively removed features that have the smallest weight and therefore the least contribution to the classification model. This increased the accuracy to 73.9 % in the multi-centre data set at one-year follow-up and to 74.3 % at three-year follow-up. The results obtained using individual centres were much higher but are likely to be biased due to the smaller sample size.

The features and ROIs selected by the RFE such as changes in the deep grey matter or insula are in line with previous research but our proposed model allows for the prediction of outcome in individual patients which has not been possible with previous models.

FUTURE WORK A strong limitation of this work is the lack of novel biomarkers as discussed in chapter 5. Future work could potentially improve the classification model by utilising a richer data set containing measures such as magnetic transfer imaging, which can be used to indicate damage outside of lesions, diffusion MRI, which can show changes in normal appearing white matter, DIR or PSIR MRI to detect grey matter lesions or even genetic information and CSF measures such as oligoclonal bands. It must be noted, however, that it is very difficult to obtain such a diverse data set longitudinally in a large enough cohort to perform meaningful and generalizable analyses. We emphasise the importance of collaborations and research consortia such as MAGNIMS with the aim of pooling together data from various centres and studies as well as sharing expertise in their specific fields.

Physical and cognitive disability are prevalent in MS patients at later disease stages. Similar to the progression from CIS to MS, there are many risk factors identified and associated with disability but it is currently not possible to predict the outcome for individual patients reliably. A possible modification of our proposed model from chapter 7 could use measures of disability as labels and perform classification or even regression analyses to predict patients' long-term disability.

The extension of studies, such as the ones presented in this thesis, to larger cohorts is very expensive and difficult due to the necessary MRI scans. Furthermore, our studies in chapters 5 to 7 rely heavily on human interaction and clinical expertise

to segment lesions or to perform clinical assessments for the EDSS scores, which have been shown to be beneficial for the classification of CIS patients. Current ongoing work on automated lesion segmentation as well as machine-learning-based automated assessment of EDSS could overcome or at least reduce these limitations and also lead to a more automated process. This would not only reduce cost but also create more objective scores since both lesion marking and EDSS scoring are known to have a high variability between different raters, which currently makes it difficult to combine this type of data coming from different centres.

Even though there has been substantial progress in the understanding of MS over the lest decades, there is still little information available about the individual processes and their interactions. Data-driven models such as the event-based model proposed by Fonteijn et al. [44, 135] for Alzheimer's and Huntington's disease can be used to identify a temporal order of pathological changes in cross-sectional or longitudinal cohorts. Adaptations of this model for MS could significantly improve our understanding of underlying disease mechanisms.

A

SUPPLEMENTARY MATERIAL

A.1 LIST OF GIF-ROIS

The ROIs used by the GIF algorithm are defined by the Neuromorphometrics atlas (http://neuromorphometrics.com) and include the following areas:

1. Background and skull
2. Non-ventricular CSF
3. 3rd Ventricle
4. 4th Ventricle
5. 5th Ventricle
6. Right Accumbens Area
7. Left Accumbens Area
8. Right Amygdala
9. Left Amygdala
10. Pons
11. Brain Stem
12. Right Caudate
13. Left Caudate
14. Right Cerebellum Exterior
15. Left Cerebellum Exterior
16. Right Cerebellum White Matter
17. Left Cerebellum White Matter
18. Right Cerebral Exterior
19. Left Cerebral Exterior
20. Right Cerebral White Matter
21. Left Cerebral White Matter
22. 3rd Ventricle (Posterior part)
23. Right Hippocampus
24. Left Hippocampus
25. Right Inf Lat Vent
26. Left Inf Lat Vent
27. Right Lateral Ventricle
28. Left Lateral Ventricle
29. Right Lesion
30. Left Lesion
31. Right Pallidum
32. Left Pallidum
33. Right Putamen
34. Left Putamen
35. Right Thalamus Proper
36. Left Thalamus Proper
37. Right Ventral DC
38. Left Ventral DC
39. Right vessel
40. Left vessel
41. Optic Chiasm

42. Cerebellar Vermal Lobules I-V
43. Cerebellar Vermal Lobules VI-VII
44. Cerebellar Vermal Lobules VIII-X
45. Left Basal Forebrain
46. Right Basal Forebrain
47. Right ACgG anterior cingulate gyrus
48. Left ACgG anterior cingulate gyrus
49. Right AIns anterior insula
50. Left AIns anterior insula
51. Right AOrG anterior orbital gyrus
52. Left AOrG anterior orbital gyrus
53. Right AnG angular gyrus
54. Left AnG angular gyrus
55. Right Calc calcarine cortex
56. Left Calc calcarine cortex
57. Right CO central operculum
58. Left CO central operculum
59. Right Cun cuneus
60. Left Cun cuneus
61. Right Ent entorhinal area
62. Left Ent entorhinal area
63. Right FO frontal operculum
64. Left FO frontal operculum
65. Right FRP frontal pole
66. Left FRP frontal pole
67. Right FuG fusiform gyrus
68. Left FuG fusiform gyrus
69. Right GRe gyrus rectus
70. Left GRe gyrus rectus
71. Right IOG inferior occipital gyrus
72. Left IOG inferior occipital gyrus
73. Right ITG inferior temporal gyrus
74. Left ITG inferior temporal gyrus
75. Right LiG lingual gyrus
76. Left LiG lingual gyrus
77. Right LOrG lateral orbital gyrus
78. Left LOrG lateral orbital gyrus
79. Right MCgG middle cingulate gyrus
80. Left MCgG middle cingulate gyrus
81. Right MFC medial frontal cortex
82. Left MFC medial frontal cortex
83. Right MFG middle frontal gyrus
84. Left MFG middle frontal gyrus
85. Right MOG middle occipital gyrus
86. Left MOG middle occipital gyrus
87. Right MOrG medial orbital gyrus
88. Left MOrG medial orbital gyrus
89. Right MPoG postcentral gyrus medial segment
90. Left MPoG postcentral gyrus medial segment
91. Right MPrG precentral gyrus medial segment
92. Left MPrG precentral gyrus medial segment

93. Right MSFG superior frontal gyrus medial segment
94. Left MSFG superior frontal gyrus medial segment
95. Right MTG middle temporal gyrus
96. Left MTG middle temporal gyrus
97. Right OCP occipital pole
98. Left OCP occipital pole
99. Right OFuG occipital fusiform gyrus
100. Left OFuG occipital fusiform gyrus
101. Right OpIFG opercular part of the inferior frontal gyrus
102. Left OpIFG opercular part of the inferior frontal gyrus
103. Right OrIFG orbital part of the inferior frontal gyrus
104. Left OrIFG orbital part of the inferior frontal gyrus
105. Right PCgG posterior cingulate gyrus
106. Left PCgG posterior cingulate gyrus
107. Right PCu precuneus
108. Left PCu precuneus
109. Right PHG parahippocampal gyrus
110. Left PHG parahippocampal gyrus
111. Right PIns posterior insula
112. Left PIns posterior insula
113. Right PO parietal operculum
114. Left PO parietal operculum
115. Right PoG postcentral gyrus
116. Left PoG postcentral gyrus
117. Right POrG posterior orbital gyrus
118. Left POrG posterior orbital gyrus
119. Right PP planum polare
120. Left PP planum polare
121. Right PrG precentral gyrus
122. Left PrG precentral gyrus
123. Right PT planum temporale
124. Left PT planum temporale
125. Right SCA subcallosal area
126. Left SCA subcallosal area
127. Right SFG superior frontal gyrus
128. Left SFG superior frontal gyrus
129. Right SMC supplementary motor cortex
130. Left SMC supplementary motor cortex
131. Right SMG supramarginal gyrus
132. Left SMG supramarginal gyrus
133. Right SOG superior occipital gyrus
134. Left SOG superior occipital gyrus
135. Right SPL superior parietal lobule
136. Left SPL superior parietal lobule
137. Right STG superior temporal gyrus
138. Left STG superior temporal gyrus

139. Right TMP temporal pole
140. Left TMP temporal pole
141. Right TrIFG triangular part of the inferior frontal gyrus
142. Left TrIFG triangular part of the inferior frontal gyrus
143. Right TTG transverse temporal gyrus
144. Left TTG transverse temporal gyrus

A.2 MANUALLY GROUPED FEATURES

Collection of result tables from experiments a) in chapter 7 (Tables 18 to 41).

Table 18: Results of multi-centre experiments at 1-year follow-up with 2-fold cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	55.6	56.1	56.3	56
	(39.3-71.3)	(34.8-69.2)	(38.9-69.8)	
global	54.9	53.6	53.2	53.9
	(37.9-68.3)	(37.1-72.3)	(39.4-68.2)	
ROI	54.9	56.1	55.5	55.5
	(35.4-70)	(38.2-72.4)	(38.5-68.2)	
lobes	56.7	56.4	**58.4***	57.2
	(35.6-72.9)	(24.2-69.3)	(39-75.9)	
lesions	59.8	**58.9**	56.7	58.5
	(38.6-75.1)	(36-72.8)	(41.2-70.6)	
nonlesion	51.6	50.9	53.7	52.1
	(37-65.3)	(24.4-67.2)	(36-67.5)	
derived	53.8	52.3	52.9	53
	(34.8-67.8)	(25-75.6)	(37.8-66.4)	
CT	50.7	49.7	50.3	50.2
	(31-63.6)	(25-75.2)	(35.6-65.5)	
GMWM	54.3	53	54	53.8
	(35.3-70)	(25-67.9)	(34.7-69)	
imaging	52.4	52.6	53.4	52.8
	(36.7-68.2)	(32.9-66.5)	(36.4-67.5)	
global_derived	54.1	52.6	53.3	53.3
	(34.8-70.9)	(24.6-66.5)	(36.4-64.7)	
global_imaging	56.1	53.6	54.1	54.6
	(37.8-69.7)	(24.8-66.4)	(37.1-68.6)	
global_lesions	**59.9**	58.7	57.2	**58.6**
	(38.6-73.5)	(36.2-70.6)	(36.7-72.8)	
global_lobes	56.5	56.4	**58.4**	57.1
	(34.1-72.1)	(40.2-71.9)	(40.1-71.5)	

* p-value$\leqslant$0.05

Table 19: Results of multi-centre experiments at 1-year follow-up with 10-fold cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	57.4	56.8	58.1*	57.4
	(39.2-74.8)	(40.7-69.8)	(43.9-69.9)	
global	55.6	55	54	54.8
	(40.3-69.3)	(40.6-69.8)	(40.1-69)	
ROI	56.4	57	56.6	56.7
	(43-75.3)	(42.1-68.1)	(42.4-69)	
lobes	58.4*	58.3	59.6*	58.8
	(43.1-74.3)	(41.2-72.9)	(48.5-74.3)	
lesions	61.5*	60*	57.2	59.6
	(44.7-72.2)	(39.4-72.2)	(44.7-68.5)	
nonlesion	51.7	50.8	54.4	52.3
	(37.8-65)	(37.8-64.3)	(39.4-70.6)	
derived	53.8	52.3	54.2	53.4
	(39.1-68.5)	(37.8-65)	(33.3-68.3)	
CT	50.6	49.9	50.4	50.3
	(35.5-64.5)	(35.6-66.7)	(37.1-64.4)	
GMWM	55.5	53.4	54.7	54.5
	(38.3-68.8)	(38.6-67.7)	(40-67.8)	
imaging	52.3	52.5	53.2	52.7
	(37.8-66.1)	(35.5-69)	(33.3-66.8)	
global_derived	54.3	53	55	54.1
	(37-67.1)	(39-66)	(36.2-69.7)	
global_imaging	57.2	54.3	53.7	55.1
	(42.4-71.2)	(35.4-67.9)	(41.6-69)	
global_lesions	**61.6***	**60.1***	58.3*	**60**
	(43.5-73.3)	(43.2-74.4)	(42.4-70.4)	
global_lobes	58.6*	58*	**59.7***	58.8
	(41.5-72.9)	(43.4-73.2)	(43.9-72)	

* p-value$\leqslant$0.05

Table 20: Results of multi-centre experiments at 1-year follow-up with LOO cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	57.7	55.4	57.2	56.8
	(41.7-72.2)	(28.1-71.5)	(40.8-73.9)	
global	55.7	54	52.9	54.2
	(41.5-68.4)	(23.4-71.6)	(31.8-65.7)	
ROI	57.1	53.8	55.8	55.6
	(43.1-71)	(24.9-70.4)	(35.6-68.5)	
lobes	59.3*	58.9	**59.1***	59.1
	(45.3-72.7)	(37.7-72.9)	(46.2-72.1)	
lesions	61.4*	**60.9***	56.9	59.7
	(44.5-74)	(35.7-74.9)	(40.9-69.9)	
nonlesion	51.6	46.4	53.2	50.4
	(34.8-67.4)	(8.1-64)	(38.3-66.7)	
derived	54.1	48.7	52.7	51.8
	(37.1-66.6)	(10.3-67.2)	(35.5-66.6)	
CT	51.1	43.6	48.7	47.8
	(34.6-64.7)	(4.5-64.4)	(32.6-63.1)	
GMWM	55.8	51.9	53.7	53.8
	(39.9-67)	(14.6-69.3)	(40.1-66)	
imaging	52	49.7	52.2	51.3
	(33.8-68.2)	(24.5-67.1)	(37.1-66.7)	
global_derived	54.6	49.6	53.8	52.7
	(37.1-69.3)	(19.2-69.7)	(39.4-65.9)	
global_imaging	57	53.4	53.2	54.5
	(41.6-70.5)	(27.2-67.5)	(36.3-67.1)	
global_lesions	**62***	**60.9***	57.7	**60.2**
	(45-73.5)	(42.3-73.2)	(41.6-71.7)	
global_lobes	59.2*	58.5	**59.1***	58.9
	(43.9-72.9)	(30.3-71.9)	(47-74.3)	

* p-value$\leqslant$0.05

Table 21: Results of multi-centre experiments at 3-year follow-up with 2fold cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	56.1	57	56.2	56.4
	(37.5-73.3)	(39.8-73.5)	(37.7-71.5)	
global	54	53.3	53	53.4
	(37.5-74)	(34.9-72.5)	(35.7-69.5)	
ROI	56.3	57	56.1	56.5
	(35.5-70.9)	(36.5-73.8)	(38.4-70.7)	
lobes	57	56.8	57.6	57.1
	(35.4-72.9)	(36.5-72.1)	(37.7-71.5)	
lesions	58.2	**57.9**	54.8	57
	(34.4-73.4)	(35.5-76.2)	(37.1-68.8)	
nonlesion	53	51.9	54.2	53
	(34.8-67.9)	(24.8-72.1)	(30.5-69)	
derived	53.1	52.1	52.2	52.5
	(31.9-67.7)	(24.8-75.7)	(34.7-70.3)	
CT	51	49.9	50.1	50.3
	(33-67.1)	(24.1-67.4)	(32.1-66.2)	
GMWM	53.5	52.5	53.5	53.1
	(33-66.4)	(25-66.7)	(32.1-67.2)	
imaging	53.7	54.6	55.7	54.7
	(37.5-70.7)	(32.4-71.9)	(40-71.5)	
global_derived	53.6	52.8	52.7	53
	(34.8-69)	(25-68.1)	(34.4-67.1)	
global_imaging	57	56.1	56.4	56.5
	(36.5-71.5)	(24.5-70.9)	(39.1-71.4)	
global_lesions	**58.3**	**57.9**	55.4	57.2
	(38.1-75)	(35.2-75.9)	(36.6-73.2)	
global_lobes	57.6	56.6	**57.9**	**57.4**
	(35.7-71.7)	(35.7-77.2)	(41.9-70.5)	

* p-value$\leqslant$0.05

Table 22: Results of multi-centre experiments at 3-year follow-up with 10-fold cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	57.5	57.2	58.4*	57.7
	(40.9-70.9)	(40-73.9)	(44.5-75.1)	
global	54.2	53.9	54.2	54.1
	(37.9-66.3)	(38.1-69.1)	(38.4-67.9)	
ROI	56.9	57.5	57.7	57.4
	(41.1-71.1)	(40.2-71.9)	(41.8-73.2)	
lobes	59.3*	58.6*	**58.8***	**58.9**
	(43.7-73.2)	(41-75.1)	(43.7-72.5)	
lesions	**60.4***	59.1*	55.7	58.4
	(37.7-76.2)	(40.9-71.4)	(40.1-71.5)	
nonlesion	53.4	52.9	55.7	54
	(37.9-67.9)	(37.4-69.8)	(38.4-70.5)	
derived	54	52.8	53.1	53.3
	(36.4-70.9)	(34.8-72.4)	(35.7-70.7)	
CT	51.7	50.4	49.6	50.6
	(29-67.2)	(24.5-67.6)	(32.1-63.4)	
GMWM	54.7	53.2	54.3	54.1
	(35.7-68.7)	(31.2-67.5)	(35.6-67)	
imaging	54.2	55.2	55.9	55.1
	(37.5-68.9)	(39.9-69.9)	(37.4-72.7)	
global_derived	54.4	53.2	53.9	53.8
	(38.4-71.7)	(37.1-67.1)	(37.5-66.1)	
global_imaging	57.2	56.8	56.9	57
	(41.8-73.5)	(43.6-70.6)	(41-74.2)	
global_lesions	**60.4***	**59.6***	56.5	**58.9**
	(40.3-75)	(40.3-71.4)	(37.5-70.3)	
global_lobes	59.4*	58.3	58.7*	58.8
	(40.2-72.9)	(41.1-71.8)	(43.7-70.5)	

* p-value$\leqslant$0.05

Table 23: Results of multi-centre experiments at 3-year follow-up with LOO cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	57.8	53.6	56.5	55.9
	(40.2-71.1)	(8.8-73.7)	(41.1-69.7)	
global	54.4	48	53.1	51.8
	(38.3-67.2)	(11.6-67.1)	(37.5-65.3)	
ROI	57.4	52.3	55.8	55.2
	(41.1-72.9)	(12.1-68.4)	(40.1-71.5)	
lobes	59.6*	58.8	57.6	58.7
	(40.1-72.7)	(23.2-73.2)	(42.8-71.7)	
lesions	**60.6***	59.4	54.2	58.1
	(37.7-74.7)	(24.5-72.5)	(31.9-68.8)	
nonlesion	53.6	48	54.6	52.1
	(29.5-69.5)	(2.7-69.7)	(38.4-68.8)	
derived	53.7	46.2	51.9	50.6
	(37.5-68.3)	(2.7-67.9)	(33-65.3)	
CT	51.5	36.1	47.6	45
	(32.1-66.1)	(0-68.8)	(31.2-62.6)	
GMWM	55.1	49.4	52.6	52.3
	(36.6-70.6)	(9.8-67.5)	(35.7-66.6)	
imaging	54.1	53.8	55.4	54.5
	(33.8-69.6)	(13.1-70.1)	(39.2-68.1)	
global_derived	54.5	47.9	52.3	51.6
	(39.2-67.9)	(8-67.7)	(36.5-69.9)	
global_lesions	**60.6***	**59.7**	55.3	58.5
	(38.9-75.6)	(27.4-73.8)	(40-67.9)	
global_lobes	59.7*	58.9	**57.7**	**58.8**
	(41.9-75.8)	(34-72.5)	(39.3-72.5)	

* p-value$\leqslant$0.05

Table 24: Results of Barcelona experiments at 1-year follow-up with 2-fold cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	56.7	54.9	53.2	54.9
	(30.9-73.9)	(23.4-75.8)	(30.9-74.9)	
global	55.8	55	54.8	55.2
	(35.8-73)	(23.8-77.9)	(25.1-74.3)	
ROI	56.6	54.3	53.4	54.8
	(33.1-77.3)	(23.8-72.7)	(26.4-72.5)	
lobes	57.9	55.3	55.6	56.3
	(27.5-76.1)	(25-74.3)	(35.1-76.6)	
lesions	59.6	**56.7**	56.5	57.6
	(32.1-79.4)	(32.4-78.3)	(30.9-73.9)	
nonlesion	53.8	50.6	52	52.1
	(25-72.2)	(23-75.4)	(29.1-73.9)	
derived	48.9	47.7	47.7	48.1
	(22-69.5)	(23.4-75.8)	(23.2-66.9)	
CT	50.7	48.2	48.7	49.2
	(26.3-68.2)	(23.4-68)	(25-72.1)	
GMWM	49	46.8	48.2	48
	(30.9-72.2)	(24.2-75.8)	(27.5-73.6)	
imaging	56	53.5	53.1	54.2
	(30.9-75.2)	(24.2-75.8)	(31.3-70.9)	
global_derived	49.6	47.7	48.7	48.7
	(26.5-69.3)	(23.8-68.9)	(26.5-68.2)	
global_imaging	56.9	54.2	53.9	55
	(29.3-75)	(24.6-77.9)	(30.7-74.3)	
global_lesions	**59.7**	55.9	**57.8**	**57.8**
	(24.4-83.3)	(24.2-79.3)	(33.5-73.6)	
global_lobes	57.5	55.1	55.7	56.1
	(35.1-76.6)	(24.2-75.6)	(30.7-75.6)	

* p-value$\leqslant$0.05

Table 25: Results of Barcelona experiments at 1-year follow-up with 10-fold cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	58	55.6	54.7	56.1
	(35.2-76.8)	(37.5-76.1)	(33.7-73)	
global	57	57.4	55.2	56.5
	(39.3-72.4)	(40.9-79.8)	(34.4-73)	
ROI	57.8	54.9	54.2	55.6
	(36.7-72.2)	(31.1-74.3)	(27.5-76.8)	
lobes	59	57	57	57.7
	(37.9-77.9)	(39.3-76.1)	(33.8-81.1)	
lesions	**63.2***	**57.7**	57.3	**59.4**
	(35.1-80.2)	(28.6-77)	(29.1-76.5)	
nonlesion	54.6	52	51.3	52.6
	(33.7-74.9)	(32.1-78)	(20.5-73.6)	
derived	49.7	47.1	48.2	48.3
	(29.1-70)	(28.2-66.3)	(26.1-72.2)	
CT	51	48.7	49.2	49.7
	(27.9-72.1)	(23-67.6)	(20.8-70.7)	
GMWM	48.4	46.2	48.7	47.8
	(23.9-70.7)	(26.6-66.9)	(27.8-69.1)	
imaging	57.4	54.8	53.8	55.3
	(38.2-72.5)	(35.1-77.4)	(27.9-71.2)	
global_derived	50.3	47.1	49.1	48.8
	(24.8-78.2)	(22.1-69.5)	(29.4-67.6)	
global_imaging	57.7	55.4	54.3	55.8
	(38.2-79.5)	(34.4-75.5)	(32.1-72.2)	
global_lesions	63*	57.4	**57.9**	**59.4**
	(33.5-77.4)	(29.8-79.3)	(33.7-75.6)	
global_lobes	58.9	57.1	57.3	57.7
	(33.7-75)	(36.7-73.7)	(36.8-75.2)	

* p-value$\leqslant$0.05

Table 26: Results of Barcelona experiments at 1-year follow-up with LOO cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	58.6	30.2	52.9	47.2
	(33.1-78.6)	(1.4-64.8)	(32.1-73.6)	
global	56.8	29.5	54.3	46.9
	(37.9-76.8)	(5.9-68.2)	(27.8-78.6)	
ROI	58.2	29.6	52.2	46.6
	(33.7-78.6)	(4.1-68.1)	(29.3-73.7)	
lobes	59.2	39.7	55.3	51.4
	(38.9-75)	(5.7-73.9)	(32.1-73.9)	
lesions	63.7*	**41.4**	56.2	53.8
	(33.8-80.2)	(2.8-74.5)	(33.8-77.3)	
nonlesion	55.5	22.7	49.2	42.5
	(35.2-75)	(0-70.9)	(23.2-70.9)	
derived	50.6	20.3	44.2	38.4
	(23.9-73.6)	(0-67.9)	(21.8-66.2)	
CT	50.9	21.3	45.9	39.4
	(21.8-70.7)	(0-60.5)	(20.6-66.2)	
GMWM	48.9	21.4	44.5	38.3
	(27.9-69.3)	(0-69.3)	(26.4-64.8)	
imaging	57.2	29.9	52.7	46.6
	(36.5-75.6)	(1.4-71.8)	(30.9-76.1)	
global_derived	51	21.9	45.4	39.4
	(27.9-73.5)	(0-63.3)	(18.9-63.3)	
global_imaging	57.5	34.8	53.5	48.6
	(36.7-75.2)	(8.7-69.3)	(23.4-73.9)	
global_lesions	**63.8***	41.3	**57**	**54**
	(38.1-77.4)	(0-80.9)	(35.2-72.5)	
global_lobes	58.7	33.6	56.3	49.5
	(39.7-77.3)	(7.3-79)	(38.1-76.6)	

* p-value$\leqslant$0.05

Table 27: Results of Barcelona experiments at 3-year follow-up with 2-fold cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	56	54.3	55.3	55.2
	(41.4-67.3)	(24.2-69.4)	(39.5-68.1)	
global	**61.1***	**59.8***	**61.9***	**60.9**
	(44.2-70.5)	(44.5-71.8)	(50.6-72.2)	
ROI	55.6	54	54.6	54.7
	(40.4-68)	(38.2-66.5)	(41.6-66.7)	
lobes	55.5	54.8	56.6	55.6
	(41.6-67.3)	(40.7-64.4)	(44.2-69.3)	
lesions	56.2	54.4	54.7	55.1
	(39.6-67.5)	(38.2-68.1)	(39.8-66.2)	
nonlesion	54.2	52.6	53.4	53.4
	(40.2-66)	(24.8-66.9)	(39.5-64.1)	
derived	52.5	52.9	55.3	53.6
	(36.8-64.9)	(25-64.3)	(40.4-66.8)	
CT	52.7	52.5	55.1	53.4
	(37-62.8)	(33.8-75.2)	(40.3-68.1)	
GMWM	52.3	52.6	55.5	53.4
	(38.4-64.5)	(38.3-67.1)	(36.5-66.7)	
imaging	55.6	53.9	52.6	54
	(37.8-68.1)	(41-66.8)	(39.7-65.5)	
global_derived	53.3	53.6	56	54.3
	(40.3-66.1)	(38.4-67.3)	(42.1-67.4)	
global_imaging	56.6	54.4	53.4	54.8
	(42.2-66.8)	(25-66)	(37-65.2)	
global_lesions	58.2*	55.4	57.6*	57
	(40.9-69.6)	(37.8-68.7)	(46.8-68)	
global_lobes	56.5	55.9	57.6*	56.7
	(43.5-68.1)	(41-68.2)	(42.9-68.7)	

* p-value$\leqslant$0.05

Table 28: Results of Barcelona experiments at 3-year follow-up with 10-fold cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	58.1*	55.6	56.1*	56.6
	(46.8-68.6)	(44.8-67.3)	(45.5-66.2)	
global	**61.2***	**59.9***	**62.5***	**61.2**
	(52.6-68.7)	(50-70.6)	(55.1-70)	
ROI	58.2*	55.4	55.4	56.3
	(46.8-67.3)	(44.2-64.1)	(46.1-64.2)	
lobes	56.5*	54.9	57.4*	56.3
	(47.4-66)	(44.7-64.4)	(47.4-66)	
lesions	58*	55.2	55.3	56.2
	(47.1-68.8)	(45.4-66.4)	(46.1-64.8)	
nonlesion	57.5*	55.3	54.8	55.9
	(47.4-66)	(44.9-65.4)	(44.2-63.5)	
derived	54.2	54.4	58.3*	55.6
	(42.7-62.4)	(43.2-66.2)	(48.7-68.1)	
CT	52.9	53.6	58.3*	54.9
	(41.7-62.9)	(39.1-63.5)	(49.4-67.9)	
GMWM	53.8	53.2	56.9*	54.6
	(41.7-63.5)	(42.1-64.8)	(48.1-65.4)	
imaging	59.2*	54.4	54	55.9
	(44.2-66.9)	(45.3-64.2)	(44.2-62.8)	
global_derived	55	55.7	58.9*	56.5
	(44.2-64.2)	(42.3-67.5)	(48.7-67.9)	
global_imaging	58*	54.4	54.5	55.6
	(46.2-67.1)	(44.5-63.6)	(43.6-64.1)	
global_lesions	58.4*	56.7*	58.1*	57.7
	(46.4-68.6)	(38.9-70.1)	(49.4-65.4)	
global_lobes	57.9*	55.9*	57.9*	57.2
	(49.4-68.6)	(44.1-65.6)	(48.7-66.8)	

* p-value$\leqslant$0.05

Table 29: Results of Barcelona experiments at 3-year follow-up with LOO cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	59.2*	57*	54.7*	57
	(48.7-68)	(48.1-65.5)	(44.8-63.6)	
global	**60.8***	**59.7***	**62.1***	**60.9**
	(53.2-68)	(49.3-69)	(54.5-70.1)	
ROI	58.9*	56.9*	53.9	56.6
	(50-66.7)	(48.1-66)	(45.5-61.6)	
lobes	56.8*	55.3	56.5*	56.2
	(47.4-66)	(44.9-63.6)	(48.7-64.2)	
lesions	58.2*	56.9*	54.8	56.6
	(49.3-67)	(47.2-67.7)	(45.4-63.5)	
nonlesion	58.3*	58.2*	54.1	56.9
	(48.1-64.7)	(50-66.7)	(44.2-63.5)	
derived	54.9	57.4*	57.2*	56.5
	(45.5-63)	(49.4-65.4)	(48.7-66.1)	
CT	53	55.7*	58*	55.6
	(42.3-62.3)	(46.8-65.4)	(48.7-66.7)	
GMWM	54.5	55.5*	55.9*	55.3
	(41.7-63.4)	(44.9-64.8)	(46.8-64.3)	
imaging	60.3*	53.6	53.2	55.7
	(50.6-66.2)	(41-63.5)	(42.3-62.9)	
global_derived	55.5*	58*	58.1*	57.2
	(45.5-63.6)	(49.4-66.9)	(50-66.7)	
global_imaging	58.6*	54.5	53.6	55.6
	(48.7-66.7)	(41.3-64.3)	(44.2-61.5)	
global_lesions	58.6*	58.4*	57.4*	58.1
	(48.6-67.4)	(45.8-69)	(50-63.5)	
global_lobes	58.5*	56*	57.1*	57.2
	(49.3-66.4)	(45.5-65.5)	(48.7-66.7)	

* p-value$\leqslant$0.05

Table 30: Results of Barcelona experiments at 5-year follow-up with 2-fold cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	56.3	55.4	55.7	55.8
	(42.6-67.9)	(40.1-66.7)	(38.8-67.9)	
global	**59.4***	57.2	**58***	**58.2**
	(42.6-69.9)	(43.2-68.5)	(44.4-70)	
ROI	56.1	55.1	55.6	55.6
	(41.3-68.5)	(40.1-67.4)	(43.2-66.1)	
lobes	55	54.3	54.4	54.6
	(40.7-67.9)	(40.6-64.9)	(42.6-65.4)	
lesions	50.8	49.6	53.3	51.3
	(35.7-61.5)	(34.6-62.5)	(41.1-66.4)	
nonlesion	56.2	55.4	55.2	55.6
	(42.6-67.9)	(40.7-66.7)	(42-67.9)	
derived	55.2	54.5	55.4	55
	(41.4-66.1)	(25-66.9)	(40.1-68.1)	
CT	52.7	51.5	53.4	52.5
	(39.4-63.6)	(25-67.3)	(37-65.4)	
GMWM	53.1	53.4	55.9	54.1
	(36.4-64.2)	(41.1-64.2)	(42.5-66.1)	
imaging	57.2	57.2	54.2	56.2
	(43.2-69.8)	(43.8-67.2)	(38.2-66.1)	
global_derived	55.6	55.3	55.9	55.6
	(40.1-68)	(34.5-67.9)	(39.5-67.9)	
global_imaging	57.9*	**57.4***	54.7	56.7
	(43.2-68.6)	(45.3-69.3)	(42-66.1)	
global_lesions	52.2	51	55.1	52.8
	(40.1-65.4)	(24.7-63.5)	(41.3-65.5)	
global_lobes	56.1	55.4	55.1	55.5
	(41.2-69.1)	(42.9-68.3)	(40.3-65)	

* p-value$\leqslant$0.05

Table 31: Results of Barcelona experiments at 5-year follow-up with 10-fold cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	58.4*	56.9*	57*	57.4
	(46.9-66.8)	(46.3-66.5)	(48.8-64.2)	
global	**60.8***	**59.2***	57.9*	**59.3**
	(51.9-68.1)	(48.8-67.2)	(49.4-66.1)	
ROI	57.9*	56.8*	56.4*	57.1
	(48.1-66.8)	(46.9-65.1)	(48.8-64.8)	
lobes	56.3*	56*	54.4	55.6
	(43.8-66.1)	(41.9-64.6)	(45.6-63)	
lesions	50.1	49.3	53.3	50.9
	(40.4-59.6)	(36.5-60.5)	(43.1-61.3)	
nonlesion	57.2*	57.2*	55.7*	56.7
	(45.6-67.3)	(46.9-65.6)	(47.5-63.6)	
derived	56.9*	57.7*	58.8*	57.8
	(46.3-67.4)	(45.1-66.8)	(50-67.4)	
CT	54.2	53.5	55.9*	54.6
	(43.8-63.6)	(40.7-62.5)	(46.9-63.6)	
GMWM	54.1	54.4	58.6*	55.7
	(43.1-63.6)	(44.9-65.1)	(48.1-66.7)	
imaging	59.5*	58.5*	54.9*	57.6
	(47.5-67.4)	(49.4-67.3)	(45.7-63)	
global_derived	57.4*	58.5*	**59.3***	58.4
	(46.9-66.2)	(45.7-68.5)	(50.6-67.9)	
global_imaging	60.1*	58.5*	55.3*	58
	(48.8-69.8)	(46.3-66.9)	(45.1-63.6)	
global_lesions	51.3	50.5	54.8*	52.2
	(40-60.7)	(35.6-62.6)	(45.7-62.4)	
global_lobes	58*	57.1*	55.4*	56.8
	(48.8-66.1)	(48.7-66.8)	(46.9-62.4)	

* p-value$\leqslant$0.05

Table 32: Results of Barcelona experiments at 5-year follow-up with LOO cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	58.7*	56.8*	56*	57.2
	(47.5-67.9)	(48.1-66.1)	(48.1-62.3)	
global	**61***	59.4*	57*	**59.1**
	(54.3-67.9)	(50.6-66.8)	(50.6-67.4)	
ROI	58.1*	56.7*	55.5*	56.7
	(48.1-65.5)	(49.4-64.3)	(46.9-63.6)	
lobes	56.3*	56.3*	53.1	55.3
	(44.4-66.1)	(46.3-65.3)	(45.1-61.7)	
lesions	50	53.5	52.1	51.9
	(40.5-58.5)	(44.3-65.6)	(45-60.6)	
nonlesion	57.3*	57.4*	54.3	56.3
	(48.1-66.1)	(49.4-66.1)	(46.3-63)	
derived	57.5*	58.9*	58.1*	58.2
	(46.2-68.5)	(50-66.9)	(50.6-65.5)	
CT	54.4	55.7*	55.3*	55.1
	(46.9-64.9)	(46.9-65.4)	(48.1-61.8)	
GMWM	54.6	55.4*	58*	56
	(44.4-64.6)	(45-65.4)	(50-65.6)	
global_derived	58.2*	**59.7***	**58.5***	58.8
	(48.8-67.3)	(51.2-67.4)	(48.8-67.3)	
global_imaging	60.8*	58.7*	54.5*	58
	(50.6-68.5)	(50-66.1)	(46.9-62.3)	
global_lesions	51	53	53.7	52.6
	(38.7-60.7)	(42.3-63.4)	(46.3-62.4)	
global_lobes	58.5*	57.3*	54.6*	56.8
	(48.1-68.7)	(49.4-66.2)	(44.4-61.8)	

* p-value$\leqslant$0.05

Table 33: Results of London experiments at 1-year follow-up with 2-fold cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	57.1	55.7	56.9	56.6
	(24.9-77.5)	(21.1-78.2)	(24.5-77.5)	
global	**63.6**	**58.7**	56.3	**59.5**
	(27.2-84.7)	(24.4-80.1)	(23.6-79.5)	
ROI	57.8	55.8	56.8	56.8
	(24.5-76.4)	(25-77.8)	(20.4-77.5)	
lobes	58.4	55.5	56.6	56.9
	(27.1-81.4)	(24.4-78.2)	(27.1-77.5)	
lesions	54.2	49.7	51	51.6
	(27.1-78.6)	(23.2-78.2)	(27.1-73.5)	
nonlesion	57.5	55.7	56.6	56.6
	(24.5-76.4)	(23.2-78.2)	(24.5-81.8)	
derived	48.3	48.1	48.7	48.4
	(17.9-70.5)	(22.5-76.2)	(22.2-73.5)	
CT	48.4	48.2	48.6	48.4
	(21.8-72.9)	(20.4-78.2)	(25.4-71.2)	
GMWM	51.6	49.8	50.7	50.7
	(27.1-75.1)	(20.3-78.2)	(26.5-73.5)	
imaging	57.5	57.2	58.4	57.7
	(26.5-75.5)	(29.5-78.2)	(25.4-75.1)	
global_derived	49.6	48.7	49	49.1
	(23.6-72.7)	(23.8-76.8)	(22.7-70.8)	
global_imaging	58.3	57.4	**59.1**	58.3
	(27.3-75.5)	(22.5-75.6)	(28.8-75.5)	
global_lesions	57.3	51.7	52.3	53.8
	(29.2-81.4)	(23.2-76.2)	(23.8-75.1)	
global_lobes	60.2	57.4	57	58.2
	(27.1-93.3)	(27.1-78.9)	(29-78.2)	

* p-value$\leqslant$0.05

Table 34: Results of London experiments at 1-year follow-up with 10-fold cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	59.4	58.3	57.8	58.5
	(31.2-75.5)	(25.6-82.9)	(29.2-77.3)	
global	**66.3***	**62.5**	57.3	**62**
	(43.2-88.7)	(34.3-87.9)	(31.7-78.2)	
ROI	59.3	57.4	57.5	58.1
	(24.9-75.5)	(21.8-76.4)	(29.2-78.2)	
lobes	59.9	58.5	56.1	58.2
	(39.5-81.8)	(32.5-78.2)	(22.5-77.5)	
lesions	54.9	49	50	51.3
	(33.8-77.3)	(22.5-77.5)	(23.6-70.8)	
nonlesion	59.2	58.4	57.3	58.3
	(22.5-77.5)	(22.5-76.4)	(26.5-78.2)	
derived	48.3	46.5	48.2	47.7
	(20.4-72.7)	(20.3-69.4)	(23.8-70.8)	
CT	47.9	47.6	47	47.5
	(24.9-73.5)	(25.4-75.1)	(22.5-68.2)	
GMWM	51.4	48.3	50.5	50.1
	(27.1-70.8)	(22.5-69.6)	(29.2-77.5)	
imaging	59.5	58.4	58.2	58.7
	(31.7-77.3)	(34.1-76.4)	(28.4-75.1)	
global_derived	50	46.5	47.8	48.1
	(22.7-75.5)	(20.3-71.6)	(21.8-77.5)	
global_imaging	59.6	58.7	**58.8**	59
	(33.8-75.5)	(32.3-80.1)	(38.6-75.1)	
global_lesions	58.3	52.4	53	54.6
	(33.8-81.4)	(24.1-77.8)	(27.1-74.6)	
global_lobes	61.7	59.6	56.5	59.2
	(38.4-82.9)	(31.2-82.1)	(31.8-78.2)	

* p-value$\leqslant$0.05

Table 35: Results of London experiments at 1-year follow-up with LOO cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	61.1	53.9	54.4	56.5
	(38.6-77.5)	(2.2-80.1)	(29.2-76.4)	
global	**66.6***	**61.4**	55.6	**61.2**
	(45.5-84.2)	(27.1-82.1)	(33.8-75.5)	
ROI	61.2*	52.9	54.8	56.3
	(38.4-75.5)	(2.2-78.2)	(26.5-72.9)	
lobes	59.8	55.2	53	56
	(38.4-79.6)	(4.2-78.2)	(33.8-75.1)	
lesions	56	37.6	46.5	46.7
	(29.2-78.6)	(4.2-70.8)	(17.1-68.3)	
nonlesion	60.7	54.4	53.9	56.3
	(22.5-77.5)	(0-78.2)	(27.3-77.5)	
derived	52.4	21.1	41.8	38.5
	(29.5-77.3)	(0-66.2)	(17.9-68.2)	
CT	48.2	26.8	40.9	38.6
	(18.8-70.5)	(0-68.3)	(13.6-64.1)	
GMWM	51.3	29.3	45.6	42.1
	(23.6-78.2)	(0-68.2)	(22.5-68.3)	
imaging	59.7	56.3	57	57.7
	(34.1-77.5)	(2.2-75.5)	(35.3-72.9)	
global_derived	54	24.2	41.3	39.8
	(30.4-77.5)	(0-65.9)	(17.1-70.5)	
global_imaging	59.2	56.5	**57.5**	57.8
	(38.6-77.5)	(2.2-80.1)	(38-80.1)	
global_lesions	59.1	44.6	50.8	51.5
	(35.3-83.3)	(2.2-80.1)	(29.2-71.6)	
global_lobes	61.3	59.1	54.4	58.3
	(38.6-81.2)	(11.3-80.1)	(31.7-77.3)	

* p-value$\leqslant$0.05

Table 36: Results of London experiments at 3-year follow-up with 2-fold cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	56.9	56.3	56.4	56.5
	(31.7-71.3)	(24.6-77.9)	(27.2-76.4)	
global	**63.8***	**58.2**	55	**59**
	(35.8-82.8)	(24.6-79.9)	(27.7-74.2)	
ROI	56.5	56.7	56.7	56.7
	(24-70.9)	(25-77.4)	(29.9-72.5)	
lobes	58.1	56.4	54.6	56.4
	(28-74.9)	(25.7-78.4)	(32.2-73.4)	
lesions	54.2	50.2	53.4	52.6
	(32.6-74.3)	(21.6-75.9)	(30.9-71.3)	
nonlesion	56.2	55.7	55.7	55.9
	(29.3-70.9)	(25-75.9)	(33.8-72)	
derived	50.2	48.9	50.4	49.8
	(24-72.8)	(24.1-77.9)	(29.1-69.3)	
CT	50.9	50.9	52.5	51.5
	(24.4-72.5)	(25-75.4)	(25.7-71.3)	
GMWM	51.2	48.6	50	49.9
	(30.9-74.4)	(24.1-75.4)	(27.1-69.8)	
imaging	54.7	57.6	56.5	56.3
	(25.8-69.3)	(32.2-74.2)	(33.2-76.4)	
global_derived	51.3	49.2	50.1	50.2
	(32.7-79.3)	(25-75.9)	(27.2-70.7)	
global_imaging	55.3	57.1	**57.2**	56.5
	(30.7-69.3)	(32.2-74.3)	(29.3-74.4)	
global_lesions	57	52.8	54.3	54.7
	(34.2-76.4)	(23.6-75.9)	(32.4-72.4)	
global_lobes	59.4	56.9	55.3	57.2
	(30.9-76.4)	(31.3-77.6)	(29.3-73.4)	

* p-value$\leqslant$0.05

Table 37: Results of London experiments at 3-year follow-up with 10-fold cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	58	58.6	56.9	57.8
	(41.3-72.5)	(38.8-73.4)	(33.8-72)	
global	**68.7***	**63.4***	56.5	**62.9**
	(48.2-81.4)	(43-77.9)	(39.6-72.5)	
ROI	57.9	58	**57.5**	57.8
	(37.9-70.9)	(37.4-72.9)	(39.6-71.3)	
lobes	59	60.6	54.4	58
	(39.5-79.5)	(39.3-79.3)	(37.7-72)	
lesions	55.9	51.2	52.3	53.1
	(37.4-72)	(34.4-69.1)	(32.4-65.6)	
nonlesion	57.6	58.8	56.3	57.6
	(38.6-69.8)	(39.6-75.6)	(39.5-72.5)	
derived	52	48.6	51.7	50.8
	(32.2-72.5)	(25.8-70.7)	(25.6-70.9)	
CT	51.2	51.4	53.9	52.2
	(32.6-69)	(32.7-72.8)	(32.2-71.3)	
GMWM	50.2	46.8	50.8	49.3
	(30.2-67.8)	(25.8-66.1)	(28.7-69.1)	
imaging	56	59.5	56	57.2
	(34.2-70.9)	(34.7-72)	(34.4-74.9)	
global_derived	53.2	49.6	51.6	51.5
	(36.1-70.7)	(29-70.9)	(32.7-70.9)	
global_imaging	55.6	59.5	56.7	57.3
	(36.2-69.8)	(37.4-73.4)	(39.5-75.4)	
global_lesions	58.3	53	52.6	54.6
	(36.9-78)	(34.2-75.6)	(35.8-67.4)	
global_lobes	60.7*	60.4	56	59
	(43.1-79.3)	(41.2-77.6)	(37.9-74.9)	

* p-value$\leqslant$0.05

Table 38: Results of London experiments at 3-year follow-up with LOO cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	59.8*	56	54.8	56.9
	(42.7-72.8)	(21.8-72.9)	(35.8-70.7)	
global	**69.5***	**64.8***	54.8	**63**
	(51.7-82.9)	(44.4-81.4)	(39.5-69.1)	
ROI	59.1*	55.2	55	56.4
	(41.2-71.3)	(10.2-72.9)	(35.4-71.3)	
lobes	59.9*	60.6	52	57.5
	(43.1-77.6)	(37.9-85.5)	(31-71.3)	
lesions	56.4	45.5	48.8	50.2
	(37.9-73.5)	(14.5-69.8)	(30.2-63.9)	
nonlesion	58	56.4	53.8	56.1
	(39.6-70.7)	(17.1-73.4)	(29.1-70.7)	
derived	54.8	34.2	47	45.3
	(36.1-70.9)	(1.7-69.1)	(26.6-65.6)	
CT	51.5	36.8	50.2	46.1
	(34.4-69.1)	(6.1-67.3)	(36.1-67.3)	
GMWM	50.3	32.9	46.7	43.3
	(32.6-65.5)	(4.7-67.3)	(24.4-65.5)	
imaging	56.3	58.3	54.4	56.3
	(41.2-69.8)	(18-72)	(36.2-72.5)	
global_derived	55.8	37.1	47.7	46.9
	(37.9-74.4)	(3.4-65.6)	(25.8-65.6)	
global_imaging	55.1	58.8	**55.7**	56.5
	(39.3-72.5)	(17-72.9)	(39.5-72.8)	
global_lesions	58.7	51	49.4	53
	(42.7-79.3)	(20.5-69.8)	(30.2-63.9)	
global_lobes	61.3*	60.6	53.3	58.4
	(46.5-76)	(37.9-79.9)	(36.2-73.4)	

* p-value$\leqslant$0.05

Table 39: Results of Siena experiments at 1-year follow-up with 2-fold cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	55.7	52.2	60.1	56
	(18.8-95.5)	(14.6-95.5)	(10-95.5)	
global	51	50	49	50
	(11.5-83.3)	(11.5-88.5)	(11.5-88.5)	
ROI	55.2	52	59	55.4
	(10-91.7)	(14.6-100)	(11.5-95.5)	
lobes	61.1	56	63.5	60.2
	(22.5-100)	(18.8-91.7)	(18.8-95.5)	
lesions	62.7	56.4	56.1	58.4
	(22.5-88.5)	(10-91.7)	(14.3-85.7)	
nonlesion	54.8	50.9	60.3	55.3
	(11.5-100)	(14.6-85.4)	(14.6-95.5)	
derived	64.8	60.6	65.5	63.6
	(14.6-95.5)	(14.3-91.7)	(14.3-91.7)	
CT	61.6	56.8	63.4	60.6
	(10-91.7)	(14.6-90)	(22.5-91.7)	
GMWM	**65.8**	**62.8**	**66.2**	**64.9**
	(18.8-95.5)	(16.7-91.7)	(16.7-91.7)	
imaging	44.8	44.5	44.2	44.5
	(8.3-81.2)	(10-88.5)	(8.3-80)	
global_derived	64.9	59.9	65.7	63.5
	(14.6-95.5)	(16.7-95.5)	(16.7-91.7)	
global_imaging	45.1	44.4	43.6	44.4
	(8.3-85.4)	(10-81.2)	(4.5-88.5)	
global_lesions	61.6	57.4	56	58.3
	(22.5-85.7)	(18.8-85.7)	(14.3-88.5)	
global_lobes	59.4	56	63.1	59.5
	(10-95.5)	(16.7-88.5)	(20-95.5)	

* p-value$\leqslant$0.05

Table 40: Results of Siena experiments at 1-year follow-up with 10-fold cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	59.5	52.9	64.4	58.9
	(10-95.5)	(16.7-91.7)	(14.6-95.5)	
global	50.3	49.4	49.1	49.6
	(8.3-88.5)	(14.3-88.5)	(8.3-85.7)	
ROI	59.1	53.1	63.4	58.5
	(14.6-95.5)	(16.7-95.5)	(8.3-95.5)	
lobes	61.8	59.1	65.2	62
	(18.8-95.5)	(16.7-95.5)	(20-91.7)	
lesions	65.1	59.8	57.4	60.8
	(18.8-88.5)	(20.6-85.7)	(22.5-85.7)	
nonlesion	57.2	51.6	64.8	57.9
	(4.5-91.7)	(16.7-85.4)	(14.6-91.7)	
derived	67.3	61.1	68.1	65.5
	(24.7-95.5)	(18.8-95.5)	(26.2-91.7)	
CT	62.1	58.6	65.6	62.1
	(34.8-91.7)	(16.7-85.7)	(29.2-95.5)	
GMWM	**68.5**	**64.8**	**69.7**	**67.7**
	(14.6-95.5)	(16.7-91.7)	(24.7-91.7)	
imaging	42.1	44.2	45	43.8
	(4.5-81.2)	(11.5-85.7)	(4.5-85.4)	
global_derived	67.4	61.2	68.7	65.7
	(29.2-95.5)	(18.8-95.5)	(34.8-91.7)	
global_imaging	42.4	44.3	44.6	43.7
	(4.5-77.5)	(14.3-90)	(4.5-85.4)	
global_lesions	63.4	58.4	58.3	60
	(14.3-85.7)	(18.8-85.7)	(11.5-85.7)	
global_lobes	59.8	57.1	63.4	60.1
	(14.6-91.7)	(18.8-91.7)	(24.7-90)	

* p-value$\leqslant$0.05

Table 41: Results of Siena experiments at 1-year follow-up with LOO cross-validation. Averaged balanced accuracies (and range over all repetitions) shown for individual classifiers as well as average across all classifiers.

Features	LinSVC	RBFSVC	RF	avg
all	66.7	73.1*	56	65.3
	(8.3-95.5)	(43.3-100)	(0-95.5)	
global	50.3	78*	43.6	57.3
	(8.3-88.5)	(33.5-100)	(4.5-85.7)	
ROI	67.8	72.6*	54.1	64.8
	(18.8-100)	(39.6-100)	(0-90)	
lobes	61.9	77.6*	60.6	66.7
	(14.6-95.5)	(44.9-100)	(20-91.7)	
lesions	69.4	77*	51.4	65.9
	(18.8-88.5)	(34.8-100)	(8.3-85.7)	
nonlesion	64.6	72.1*	57.1	64.6
	(14.6-95.5)	(34.8-100)	(4.5-90)	
derived	71.3*	**83.8***	60.9	72
	(44.5-95.5)	(44.5-100)	(8.3-90)	
CT	62.4	83.5*	60.9	68.9
	(34.8-85.4)	(44.9-100)	(18.8-91.7)	
GMWM	**71.7***	82.1*	**64.6**	**72.8**
	(18.8-95.5)	(44.9-100)	(11.5-91.7)	
imaging	41.6	77.3*	38.2	52.3
	(0-77.5)	(39.6-100)	(0-75.3)	
global_derived	70.6*	83.4*	61.3	71.8
	(34.8-95.5)	(34.8-100)	(20-91.7)	
global_imaging	41.7	78*	38.6	52.8
	(0-75.3)	(33.5-100)	(0-75.3)	
global_lesions	66	78*	53.9	66
	(14.3-85.7)	(44.9-100)	(8.3-85.7)	
global_lobes	60.8	78.4*	60.7	66.6
	(14.3-91.7)	(40-100)	(14.6-91.7)	

* p-value$\leqslant$0.05

A.3 AUTOMATED FEATURE SELECTION

Accuracies of 10-fold and leave-one-out experiments using RFE (Tables 42 and 43), and lists of selected features (Tables 44 to 49).

Table 42: Results of RFE experiments using a 10-fold cross-validation. Findings are shown as balanced accuracy, range over all repetitions and and 95 % confidence interval (CI) for the included centres.

centre	follow-up	bal. acc. (%)	range (%)	CI (%)	# features
multi-centre	1	84.3	75.0-93.2	83.9-84.7	75
multi-centre	3	84.6	73.2-95.5	84.2-85.1	75
Barcelona	1	86.7	73.5-97.1	86.1-87.3	100
Barcelona	3	94.9	89.7-98.7	94.7-95.0	134
Barcelona	5	95.7	92.6-98.8	95.1-95.4	75
London	1	98.9	90.9-100	98.7-99.1	75
London	3	99.1	93.1-100	98.9-99.2	75
Siena	1	98.4	85.0-100	98.1-98.8	32

Table 43: Results of RFE experiments using a leave-one-out cross-validation. Findings are shown as balanced accuracy, range over all repetitions and and 95 % confidence interval (CI) for the included centres.

centre	follow-up	bal. acc. (%)	range (%)	CI (%)	# features
multi-centre	1	88.7	77.3-97.7	88.3-89.1	75
multi-centre	3	85.9	75.0-98.0	85.4-86.5	100
Barcelona	1	89.4	76.5-100	88.9-89.9	75
Barcelona	3	96.6	91.7-99.4	96.5-96.8	178
Barcelona	5	96.9	93.2-99.4	96.8-97.0	100
London	1	99.2	95.5-100	99.0-99.3	100
London	3	99.6	94.8-100	99.5-99.7	75
Siena	1	97.8	85.0-100	97.5-98.2	24

Table 44: Selected features in best result from RFE experiments using Barcelona data at 1-year follow-up with 2-fold cross-validation.

1. lesionLoad_WM
2. WM_Left_Accumbens_Area
3. WM_3rd_Ventricle_(Posterior_part)
4. WM_Optic_Chiasm
5. WM_Right_Basal_Forebrain
6. WM_Left_LOrG_lateral_orbital_gyrus
7. WM_Left_MTG_middle_temporal_gyrus
8. WM_Left_OCP_occipital_pole
9. WM_Right_POrG_posterior_orbital_gyrus
10. WM_Right_SOG_superior_occipital_gyrus
11. GM_Non-ventricular_CSF
12. GM_Cerebellar_Vermal_Lobules_VIII-X
13. GM_Left_Basal_Forebrain
14. GM_Right_MFC_medial_frontal_cortex
15. GM_Left_PP_planum_polare
16. CT_Left_Pallidum
17. CT_Right_Thalamus_Proper
18. CT_Left_Thalamus_Proper
19. CT_Right_Basal_Forebrain
20. CT_Right_GRe_gyrus_rectus
21. CT_Right_LiG_lingual_gyrus
22. CT_Right_MTG_middle_temporal_gyrus
23. CT_Right_PO_parietal_operculum
24. T2_Left_Cerebral_Exterior
25. PD_Left_AOrG_anterior_orbital_gyrus
26. PD_Right_OCP_occipital_pole
27. volume_Left_Accumbens_Area
28. volume_Optic_Chiasm
29. volume_Left_Basal_Forebrain
30. volume_Left_CO_central_operculum
31. volume_Right_Cun_cuneus
32. volume_Left_GRe_gyrus_rectus
33. volume_Right_LiG_lingual_gyrus
34. volume_Left_MCgG_middle_cingulate_gyrus
35. volume_Right_MOG_middle_occipital_gyrus
36. volume_Left_MPoG_postcentral_gyrus_medial_segment
37. volume_Right_PCgG_posterior_cingulate_gyrus
38. volume_Right_PO_parietal_operculum
39. volume_Left_PP_planum_polare
40. volume_Right_PT_planum_temporale
41. volume_Right_SOG_superior_occipital_gyrus

42. volume_Right_TrIFG_triangular_part_of_the_inferior_frontal_gyrus
43. volume_Left_TTG_transverse_temporal_gyrus
44. lesionCount_Right_Cerebellum_Exterior
45. lesionCount_Right_Cerebral_White_Matter
46. lesionCount_Right_Lateral_Ventricle
47. lesionCount_Left_Ventral_DC
48. lesionCount_Right_CO_central_operculum
49. lesionCount_Left_FuG_fusiform_gyrus
50. lesionCount_Left_MOrG_medial_orbital_gyrus
51. lesionCount_Right_OFuG_occipital_fusiform_gyrus
52. lesionCount_Right_PCgG_posterior_cingulate_gyrus
53. lesionCount_Right_PHG_parahippocampal_gyrus
54. lesionLoad_Left_Cerebral_White_Matter
55. lesionLoad_Right_FuG_fusiform_gyrus
56. lesionLoad_Left_FuG_fusiform_gyrus

Table 45: Selected features in best result from RFE experiments using Barcelona data at 3-year follow-up with 2-fold cross-validation.

1. onset
2. lesionLoad_WM
3. WM_Right_vessel
4. WM_Right_FO_frontal_operculum
5. WM_Right_MFC_medial_frontal_cortex
6. WM_Left_OpIFG_opercular_part_of_the_inferior_frontal_gyrus
7. WM_Left_PHG_parahippocampal_gyrus
8. WM_Left_PT_planum_temporale
9. WM_Left_SMC_supplementary_motor_cortex
10. WM_Right_SOG_superior_occipital_gyrus
11. WM_Left_TTG_transverse_temporal_gyrus
12. GM_Non-ventricular_CSF
13. GM_Brain_Stem
14. GM_Left_Hippocampus
15. GM_Right_AIns_anterior_insula
16. GM_Right_Calc_calcarine_cortex
17. GM_Left_Calc_calcarine_cortex
18. GM_Left_IOG_inferior_occipital_gyrus
19. GM_Left_OCP_occipital_pole
20. GM_Right_SCA_subcallosal_area
21. GM_Left_TMP_temporal_pole
22. CT_Right_Pallidum
23. CT_Left_vessel
24. CT_Cerebellar_Vermal_Lobules_I-V
25. CT_Left_Calc_calcarine_cortex
26. CT_Right_LiG_lingual_gyrus
27. CT_Left_MOrG_medial_orbital_gyrus
28. CT_Left_OCP_occipital_pole
29. CT_Right_PO_parietal_operculum
30. CT_Left_PoG_postcentral_gyrus
31. PD_Left_OCP_occipital_pole
32. volume_4th_Ventricle
33. volume_Left_Pallidum
34. volume_Right_vessel
35. volume_Right_AIns_anterior_insula
36. volume_Right_Ent_entorhinal_area
37. volume_Right_FO_frontal_operculum
38. volume_Left_FRP_frontal_pole
39. volume_Left_GRe_gyrus_rectus
40. volume_Left_MPoG_postcentral_gyrus_medial_segment
41. volume_Right_OpIFG_opercular_part_of_the_inferior_frontal_gyrus

42. volume_Right_PIns_posterior_insula
43. volume_Left_PO_parietal_operculum
44. volume_Right_SOG_superior_occipital_gyrus
45. volume_Right_SPL_superior_parietal_lobule
46. lesionCount_Right_Cerebellum_White_Matter
47. lesionCount_Right_Lateral_Ventricle
48. lesionCount_Right_CO_central_operculum
49. lesionCount_Right_ITG_inferior_temporal_gyrus
50. lesionCount_Left_MTG_middle_temporal_gyrus
51. lesionCount_Right_PCgG_posterior_cingulate_gyrus
52. lesionCount_Right_PT_planum_temporale
53. lesionCount_Left_PT_planum_temporale
54. lesionLoad_Right_Cerebral_White_Matter
55. lesionLoad_Left_Inf_Lat_Vent
56. lesionLoad_Right_SMC_supplementary_motor_cortex

Table 46: Selected features in best result from RFE experiments using Barcelona data at 5-year follow-up with 2-fold cross-validation.

1. age
2. sex
3. onset
4. lesionLoad_global
5. lesionLoad_WM
6. WM_Brain_Stem
7. WM_Right_Pallidum
8. WM_Right_vessel
9. WM_Optic_Chiasm
10. WM_Cerebellar_Vermal_Lobules_VIII-X
11. WM_Right_FO_frontal_operculum
12. WM_Left_LOrG_lateral_orbital_gyrus
13. WM_Right_MFC_medial_frontal_cortex
14. WM_Right_MFG_middle_frontal_gyrus
15. WM_Left_OpIFG_opercular_part_of_the_inferior_frontal_gyrus
16. WM_Left_PCgG_posterior_cingulate_gyrus
17. WM_Right_PCu_precuneus
18. WM_Left_PHG_parahippocampal_gyrus
19. WM_Right_PIns_posterior_insula
20. WM_Left_PP_planum_polare
21. WM_Left_PrG_precentral_gyrus
22. WM_Left_PT_planum_temporale
23. WM_Right_SFG_superior_frontal_gyrus
24. WM_Left_SMC_supplementary_motor_cortex
25. GM_Non-ventricular_CSF
26. GM_3rd_Ventricle
27. GM_Right_Amygdala
28. GM_Brain_Stem
29. GM_Left_Cerebral_Exterior
30. GM_3rd_Ventricle_(Posterior_part)
31. GM_Left_Hippocampus
32. GM_Right_Pallidum
33. GM_Left_Calc_calcarine_cortex
34. GM_Left_OrIFG_orbital_part_of_the_inferior_frontal_gyrus
35. CT_Right_Pallidum
36. CT_Left_vessel
37. CT_Cerebellar_Vermal_Lobules_I-V
38. CT_Left_Basal_Forebrain
39. CT_Right_Calc_calcarine_cortex
40. CT_Left_Calc_calcarine_cortex
41. CT_Right_FuG_fusiform_gyrus

42. CT_Right_GRe_gyrus_rectus
43. CT_Right_LiG_lingual_gyrus
44. CT_Left_MCgG_middle_cingulate_gyrus
45. CT_Right_MOG_middle_occipital_gyrus
46. CT_Left_MOG_middle_occipital_gyrus
47. CT_Right_MPoG_postcentral_gyrus_medial_segment
48. CT_Left_MSFG_superior_frontal_gyrus_medial_segment
49. CT_Right_OrIFG_orbital_part_of_the_inferior_frontal_gyrus
50. CT_Left_PIns_posterior_insula
51. CT_Right_PO_parietal_operculum
52. CT_Left_PP_planum_polare
53. CT_Right_PT_planum_temporale
54. CT_Right_SCA_subcallosal_area
55. CT_Right_SFG_superior_frontal_gyrus
56. CT_Right_SMC_supplementary_motor_cortex
57. T2_Right_vessel
58. PD_Background_and_skull
59. PD_Left_OCP_occipital_pole
60. volume_Left_Caudate
61. volume_Right_Pallidum
62. volume_Left_Putamen
63. volume_Right_vessel
64. volume_Right_AIns_anterior_insula
65. volume_Right_AOrG_anterior_orbital_gyrus
66. volume_Right_CO_central_operculum
67. volume_Right_Ent_entorhinal_area
68. volume_Left_FRP_frontal_pole
69. volume_Right_GRe_gyrus_rectus
70. volume_Left_LOrG_lateral_orbital_gyrus
71. volume_Right_MCgG_middle_cingulate_gyrus
72. volume_Right_MFC_medial_frontal_cortex
73. volume_Right_MOG_middle_occipital_gyrus
74. volume_Right_MOrG_medial_orbital_gyrus
75. volume_Left_MPoG_postcentral_gyrus_medial_segment
76. volume_Left_OFuG_occipital_fusiform_gyrus
77. volume_Right_OrIFG_orbital_part_of_the_inferior_frontal_gyrus
78. volume_Left_PHG_parahippocampal_gyrus
79. volume_Right_PIns_posterior_insula
80. volume_Right_SFG_superior_frontal_gyrus
81. volume_Left_SMG_supramarginal_gyrus
82. volume_Right_SOG_superior_occipital_gyrus
83. volume_Right_SPL_superior_parietal_lobule
84. volume_Left_TTG_transverse_temporal_gyrus

85. lesionCount_Left_Cerebellum_White_Matter
86. lesionCount_Left_Pallidum
87. lesionCount_Right_CO_central_operculum
88. lesionCount_Right_ITG_inferior_temporal_gyrus
89. lesionCount_Left_ITG_inferior_temporal_gyrus
90. lesionCount_Left_OpIFG_opercular_part_of_the_inferior_frontal_gyrus
91. lesionCount_Right_PCgG_posterior_cingulate_gyrus
92. lesionCount_Left_POrG_posterior_orbital_gyrus
93. lesionCount_Left_PT_planum_temporale
94. lesionCount_Right_SMC_supplementary_motor_cortex
95. lesionLoad_Right_Cerebral_White_Matter
96. lesionLoad_Right_Inf_Lat_Vent
97. lesionLoad_Left_POrG_posterior_orbital_gyrus
98. lesionLoad_Left_PT_planum_temporale
99. lesionLoad_Right_SMC_supplementary_motor_cortex
100. lesionLoad_Right_TTG_transverse_temporal_gyrus

Table 47: Selected features in best result from RFE experiments using London data at 1-year follow-up with 2-fold cross-validation.

1. age
2. EDSS
3. onset
4. mean_CT_insular
5. mean_CT_WM
6. volume_insular
7. WM_Brain_Stem
8. WM_Right_Caudate
9. WM_Left_Caudate
10. WM_Right_Cerebral_Exterior
11. WM_Right_Hippocampus
12. WM_Right_Inf_Lat_Vent
13. WM_Right_Pallidum
14. WM_Right_vessel
15. WM_Left_AIns_anterior_insula
16. WM_Right_Ent_entorhinal_area
17. WM_Left_Ent_entorhinal_area
18. WM_Right_FRP_frontal_pole
19. WM_Left_FRP_frontal_pole
20. WM_Right_MCgG_middle_cingulate_gyrus
21. WM_Left_MCgG_middle_cingulate_gyrus
22. WM_Right_OCP_occipital_pole
23. WM_Left_SCA_subcallosal_area
24. GM_Right_Caudate
25. GM_Left_Caudate
26. GM_Right_Pallidum
27. GM_Right_AnG_angular_gyrus
28. GM_Right_Ent_entorhinal_area
29. GM_Left_FO_frontal_operculum
30. GM_Left_FRP_frontal_pole
31. GM_Right_GRe_gyrus_rectus
32. GM_Left_MCgG_middle_cingulate_gyrus
33. GM_Left_MPrG_precentral_gyrus_medial_segment
34. GM_Right_OCP_occipital_pole
35. GM_Left_OCP_occipital_pole
36. GM_Right_TrIFG_triangular_part_of_the_inferior_frontal_gyrus
37. CT_Right_Accumbens_Area
38. CT_Right_Cerebral_White_Matter
39. CT_Right_vessel
40. CT_Left_vessel
41. CT_Left_CO_central_operculum

42. CT_Left_LiG_lingual_gyrus
43. CT_Right_MOrG_medial_orbital_gyrus
44. CT_Right_OFuG_occipital_fusiform_gyrus
45. CT_Left_OFuG_occipital_fusiform_gyrus
46. CT_Right_PIns_posterior_insula
47. T1_Left_Lateral_Ventricle
48. T2_Right_OCP_occipital_pole
49. T2_Right_SOG_superior_occipital_gyrus
50. T2_Left_TTG_transverse_temporal_gyrus
51. volume_Right_Accumbens_Area
52. volume_Right_Caudate
53. volume_Cerebellar_Vermal_Lobules_VI-VII
54. volume_Right_AIns_anterior_insula
55. volume_Left_AIns_anterior_insula
56. volume_Right_FO_frontal_operculum
57. volume_Right_OFuG_occipital_fusiform_gyrus
58. volume_Left_SMC_supplementary_motor_cortex
59. lesionCount_Background_and_skull
60. lesionCount_Right_Cerebellum_White_Matter
61. lesionCount_Right_AOrG_anterior_orbital_gyrus
62. lesionCount_Left_AOrG_anterior_orbital_gyrus
63. lesionCount_Right_Calc_calcarine_cortex
64. lesionCount_Left_CO_central_operculum
65. lesionCount_Left_LOrG_lateral_orbital_gyrus
66. lesionCount_Left_MTG_middle_temporal_gyrus
67. lesionCount_Left_PO_parietal_operculum
68. lesionCount_Left_PoG_postcentral_gyrus
69. lesionCount_Left_TMP_temporal_pole
70. lesionLoad_Right_Caudate
71. lesionLoad_Right_Cerebellum_White_Matter
72. lesionLoad_Right_ACgG_anterior_cingulate_gyrus
73. lesionLoad_Right_MOrG_medial_orbital_gyrus
74. lesionLoad_Left_MTG_middle_temporal_gyrus
75. lesionLoad_Left_PoG_postcentral_gyrus

Table 48: Selected features in best result from RFE experiments using London data at 3-year follow-up with 2-fold cross-validation.

1. EDSS
2. onset
3. WM_4th_Ventricle
4. WM_Right_Amygdala
5. WM_3rd_Ventricle_(Posterior_part)
6. WM_Right_Hippocampus
7. WM_Right_Inf_Lat_Vent
8. WM_Left_Inf_Lat_Vent
9. WM_Left_Lateral_Ventricle
10. WM_Right_Putamen
11. WM_Right_Ventral_DC
12. WM_Right_vessel
13. WM_Left_AOrG_anterior_orbital_gyrus
14. WM_Right_Calc_calcarine_cortex
15. WM_Right_FRP_frontal_pole
16. WM_Left_GRe_gyrus_rectus
17. WM_Left_MCgG_middle_cingulate_gyrus
18. WM_Right_OCP_occipital_pole
19. WM_Right_SCA_subcallosal_area
20. GM_Right_Inf_Lat_Vent
21. GM_Right_Putamen
22. GM_Right_vessel
23. GM_Right_AnG_angular_gyrus
24. GM_Right_Calc_calcarine_cortex
25. GM_Right_Ent_entorhinal_area
26. GM_Right_PIns_posterior_insula
27. GM_Right_SCA_subcallosal_area
28. CT_Right_Accumbens_Area
29. CT_Right_vessel
30. CT_Left_vessel
31. CT_Right_AIns_anterior_insula
32. CT_Right_Calc_calcarine_cortex
33. CT_Left_CO_central_operculum
34. CT_Left_FO_frontal_operculum
35. CT_Right_OFuG_occipital_fusiform_gyrus
36. CT_Left_OFuG_occipital_fusiform_gyrus
37. CT_Left_PoG_postcentral_gyrus
38. CT_Left_SCA_subcallosal_area
39. T1_Left_Lateral_Ventricle
40. T2_Non-ventricular_CSF
41. T2_Right_Cerebral_Exterior

42. PD_Right_Cerebral_Exterior
43. volume_Right_Accumbens_Area
44. volume_Right_Cerebral_Exterior
45. volume_Left_Lateral_Ventricle
46. volume_Left_vessel
47. volume_Right_Cun_cuneus
48. volume_Right_FO_frontal_operculum
49. volume_Left_GRe_gyrus_rectus
50. volume_Right_ITG_inferior_temporal_gyrus
51. volume_Right_OCP_occipital_pole
52. volume_Right_OFuG_occipital_fusiform_gyrus
53. volume_Left_PO_parietal_operculum
54. volume_Right_PoG_postcentral_gyrus
55. volume_Right_PT_planum_temporale
56. volume_Left_SMC_supplementary_motor_cortex
57. volume_Right_SOG_superior_occipital_gyrus
58. volume_Right_STG_superior_temporal_gyrus
59. lesionCount_Right_Cerebellum_White_Matter
60. lesionCount_Right_Pallidum
61. lesionCount_Right_AOrG_anterior_orbital_gyrus
62. lesionCount_Left_CO_central_operculum
63. lesionCount_Right_GRe_gyrus_rectus
64. lesionCount_Left_MFG_middle_frontal_gyrus
65. lesionCount_Left_MTG_middle_temporal_gyrus
66. lesionCount_Left_PO_parietal_operculum
67. lesionCount_Right_PP_planum_polare
68. lesionCount_Left_TMP_temporal_pole
69. lesionLoad_Right_Cerebellum_White_Matter
70. lesionLoad_Right_Ventral_DC
71. lesionLoad_Right_GRe_gyrus_rectus
72. lesionLoad_Right_MOrG_medial_orbital_gyrus
73. lesionLoad_Left_MTG_middle_temporal_gyrus
74. lesionLoad_Right_PoG_postcentral_gyrus
75. lesionLoad_Left_TMP_temporal_pole

Table 49: Selected features in best result from RFE experiments using Siena data at 1-year follow-up with 2-fold cross-validation.

1. mean_CT_temporal
2. WM_4th_Ventricle
3. WM_Right_Accumbens_Area
4. WM_Right_Caudate
5. WM_Left_Basal_Forebrain
6. WM_Left_AOrG_anterior_orbital_gyrus
7. WM_Left_LOrG_lateral_orbital_gyrus
8. WM_Left_MCgG_middle_cingulate_gyrus
9. WM_Right_OFuG_occipital_fusiform_gyrus
10. WM_Left_PrG_precentral_gyrus
11. WM_Right_SMC_supplementary_motor_cortex
12. WM_Right_TMP_temporal_pole
13. GM_Right_Accumbens_Area
14. GM_Cerebellar_Vermal_Lobules_VI-VII
15. GM_Left_Basal_Forebrain
16. GM_Left_AOrG_anterior_orbital_gyrus
17. GM_Left_LOrG_lateral_orbital_gyrus
18. GM_Right_PCgG_posterior_cingulate_gyrus
19. GM_Left_PHG_parahippocampal_gyrus
20. GM_Right_POrG_posterior_orbital_gyrus
21. GM_Left_PP_planum_polare
22. CT_Right_Caudate
23. CT_Right_Cerebellum_Exterior
24. CT_Left_AIns_anterior_insula
25. CT_Left_AOrG_anterior_orbital_gyrus
26. CT_Left_CO_central_operculum
27. CT_Right_GRe_gyrus_rectus
28. CT_Left_LOrG_lateral_orbital_gyrus
29. CT_Right_MFC_medial_frontal_cortex
30. CT_Left_OpIFG_opercular_part_of_the_inferior_frontal_gyrus
31. CT_Left_OrIFG_orbital_part_of_the_inferior_frontal_gyrus
32. CT_Left_PHG_parahippocampal_gyrus
33. CT_Left_PoG_postcentral_gyrus
34. CT_Right_PP_planum_polare
35. CT_Left_PP_planum_polare
36. CT_Left_PrG_precentral_gyrus
37. CT_Left_PT_planum_temporale
38. CT_Left_TrIFG_triangular_part_of_the_inferior_frontal_gyrus
39. volume_Right_Caudate
40. volume_Right_Cerebellum_White_Matter
41. volume_Left_Hippocampus

42. volume_Left_vessel
43. volume_Left_AIns_anterior_insula
44. volume_Left_Ent_entorhinal_area
45. volume_Right_MCgG_middle_cingulate_gyrus
46. volume_Left_OFuG_occipital_fusiform_gyrus
47. lesionCount_Left_Cerebellum_Exterior
48. lesionCount_Right_Ventral_DC
49. lesionCount_Left_IOG_inferior_occipital_gyrus
50. lesionCount_Left_MOG_middle_occipital_gyrus
51. lesionCount_Left_PT_planum_temporale
52. lesionLoad_Left_Cerebellum_Exterior
53. lesionLoad_Left_Thalamus_Proper
54. lesionLoad_Left_MOG_middle_occipital_gyrus
55. lesionLoad_Left_PCu_precuneus
56. lesionLoad_Left_PO_parietal_operculum

BIBLIOGRAPHY

[1] Alberto Ascherio, Kassandra L Munger, and K Claire Simon. “Vitamin D and multiple sclerosis.” In: *Lancet neurology* 9.6 (2010), pp. 599–612.

[2] Bertrand Audoin, Kryshani T M Fernando, Josephine K Swanton, Alan J Thompson, Gordon T Plant, and David H Miller. “Selective magnetization transfer ratio decrease in the visual cortex following optic neuritis.” In: *Brain: a journal of neurology* 129.Pt 4 (2006), pp. 1031–9.

[3] Bertrand Audoin, Wafaa Zaaraoui, Françoise Reuter, Audrey Rico, Irina Malikova, Sylviane Confort-Gouny, Patrick J Cozzone, Jean Pelletier, and Jean-Philippe Ranjeva. “Atrophy mainly affects the limbic system and the deep grey matter at the first stage of multiple sclerosis.” In: *Journal of neurology, neurosurgery, and psychiatry* 81.6 (2010), pp. 690–5.

[4] F Barkhof, M Filippi, D H Miller, P Scheltens, A Campi, C H Polman, G Comi, H J Adèr, N Losseff, and J Valk. “Comparison of MRI criteria at first presentation to predict conversion to clinically definite multiple sclerosis.” In: *Brain: a journal of neurology* 11 (1997), pp. 2059–69.

[5] M.H Barnett, D.B Williams, S Day, P Macaskill, and J.G McLeod. “Progressive increase in incidence and prevalence of multiple sclerosis in Newcastle, Australia: a 35-year study.” In: *Journal of the Neurological Sciences* 213.1-2 (2003), pp. 1–6.

[6] Sonia Batista, Robert Zivadinov, Marietta Hoogs, Niels Bergsland, Mari Heininen-Brown, Michael G. Dwyer, Bianca Weinstock-Guttman, and Ralph H. B. Benedict. “Basal ganglia, thalamus and neocortical atrophy predicting slowed cognitive processing in multiple sclerosis.” In: *Journal of Neurology* 259.1 (2012), pp. 139–146.

[7] Kerstin Bendfeldt et al. “Multivariate pattern classification of gray matter pathology in multiple sclerosis.” In: *NeuroImage* 60.1 (2012), pp. 400–8.

[8] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 2006, p. 738.

[9] Benedetta Bodini, Zhaleh Khaleeli, Mara Cercignani, David H. Miller, Alan J. Thompson, and Olga Ciccarelli. “Exploring the relationship between white matter and gray matter damage in early primary progressive multiple sclerosis: An in vivo study with TBSS and VBM.” In: *Human Brain Mapping* 30.9 (2009), pp. 2852–2861.

[10] P. F. Bray, L. C. Bloomer, V. C. Salmon, M. H. Bagley, and P. D. Larsen. “Epstein-Barr Virus Infection and Antibody Synthesis in Patients With Multiple Sclerosis.” In: *Archives of Neurology* 40.7 (1983), pp. 406–408.

[11] Leo Breiman. “Random Forests.” en. In: *Machine Learning* 45.1 (2001), pp. 5–32.

[12] William S Bush, Stephen J Sawcer, Philip L de Jager, Jorge R Oksenberg, Jacob L McCauley, Margaret A Pericak-Vance, and Jonathan L Haines. “Evidence for polygenic susceptibility to multiple sclerosis–the shape of things to come.” In: *American journal of human genetics* 86.4 (2010), pp. 621–5.

[13] M Calabrese, F Rinaldi, I Mattisi, V Bernardi, A Favaretto, P Perini, and P Gallo. “The predictive value of gray matter atrophy in clinically isolated syndromes.” In: *Neurology* 77.3 (2011), pp. 257–63.

[14] M. Jorge Cardoso, Matthew J. Clarkson, Gerard R. Ridgway, Marc Modat, Nick C. Fox, and Sebastien Ourselin. “LoAd: A locally adaptive cortical segmentation algorithm.” In: *NeuroImage* 56.3 (2011), pp. 1386–1397.

[15] M. Jorge Cardoso, Marc Modat, Robin Wolz, Andrew Melbourne, David Cash, Daniel Rueckert, and Sebastien Ourselin. “Geodesic Information Flows: Spatially-Variant Graphs and Their Application to Segmentation and Fusion.” In: *IEEE Transactions on Medical Imaging* 34.9 (2015), pp. 1976–1988.

[16] H Carton, R Vlietinck, J Debruyne, J De Keyser, M B D’Hooghe, R Loos, R Medaer, L Truyen, I M Yee, and A D Sadovnick. “Risks of multiple sclerosis in relatives of patients in Flanders, Belgium.” In: *Journal of Neurology, Neurosurgery & Psychiatry* 62.4 (1997), pp. 329–333.

[17] Chih-Chung Chang and Chih-Jen Lin. “LIBSVM.” In: *ACM Transactions on Intelligent Systems and Technology* 2.3 (2011), pp. 1–27.

[18] Divya M Chari. “Remyelination in multiple sclerosis.” In: *International review of neurobiology* 79 (2007), pp. 589–620.

[19] Min Chen, Aaron Carass, Daniel S Reich, Peter A Calabresi, Dzung Pham, and Jerry L Prince. “Voxel-Wise Displacement as Independent Features in Classification of Multiple Sclerosis.” In: *Proceedings of SPIE* 8669 (2013), 86690K–86690K–6.

[20] Carlton Chu, Ai-Ling Hsu, Kun-Hsien Chou, Peter Bandettini, and ChingPo Lin. “Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images.” 2012.

[21] Alberto Cifelli, Marzena Arridge, Peter Jezzard, Margaret M. Esiri, Jacqueline Palace, and Paul M. Matthews. “Thalamic neurodegeneration in multiple sclerosis.” In: *Annals of Neurology* 52.5 (2002), pp. 650–653.

[22] Alastair Compston and Alasdair Coles. “Multiple sclerosis.” In: *Lancet* 359.9313 (2002), pp. 1221–31.

[23] Alastair Compston and Alasdair Coles. “Multiple sclerosis.” In: *Lancet* 372.9648 (2008), pp. 1502–17.

[24] C. Confavreax, G. Aimard, and M. Devic. “Course and Prognosis of Multiple Slerosis Assessed by the Computerized Data Processing of 349 Patients.” en. In: *Brain* 103.2 (1980), pp. 281–300.

[25] Christian Confavreux, Sandra Vukusic, and Patrice Adeleine. “Early clinical predictors and progression of irreversible disability in multiple sclerosis: an amnesic process.” In: *Brain: a journal of neurology* 126.Pt 4 (2003), pp. 770–82.

[26] Corinna Cortes and Vladimir Vapnik. “Support-vector networks.” In: *Machine Learning* 20.3 (1995), pp. 273–297.

[27] A. Criminisi and J. Shotton, eds. *Decision Forests for Computer Vision and Medical Image Analysis*. London: Springer London, 2013.

[28] Rémi Cuingnet, Emilie Gerardin, Jérôme Tessieras, Guillaume Auzias, Stéphane Lehéricy, Marie-Odile Habert, Marie Chupin, Habib Benali, and Olivier Colliot. “Automatic classification of patients with Alzheimer’s disease from structural MRI: a comparison of ten methods using the ADNI database.” In: *NeuroImage* 56.2 (2011), pp. 766–81.

[29] M. D’Souza et al. “Prediction of expanded disability status scale subscores of motor dysfunction in multiple sclerosis using depth-sensing computer vision.” In: *Congress of the European Committee for Treatment and Research in Multiple Sclerosis*. 2015.

[30] C. Dalton, B. Bodini, R. Samson, M. Battaglini, L. Fisniku, A. Thompson, O. Ciccarelli, D. Miller, and D. Chard. “Brain lesion location and clinical status 20 years after a diagnosis of clinically isolated syndrome suggestive of multiple sclerosis.” In: *Multiple Sclerosis Journal* 18.3 (2012), pp. 322–328.

[31] Catherine M Dalton, Peter A Brex, Katherine A Miszkiel, Simon J Hickman, David G MacManus, Gordon T Plant, Alan J Thompson, and David H Miller. “Application of the new McDonald criteria to patients with clinically isolated syndromes suggestive of multiple sclerosis.” In: *Annals of neurology* 52.1 (2002), pp. 47–53.

[32] Sandhitsu R Das, Brian B Avants, Murray Grossman, and James C Gee. “Registration based cortical thickness measurement.” In: *NeuroImage* 45.3 (2009), pp. 867–79.

[33] F Di Pauli et al. “Smoking is a risk factor for early conversion to clinically definite multiple sclerosis.” In: *Multiple sclerosis (Houndmills, Basingstoke, England)* 14.8 (2008), pp. 1026–30.

[34] Ruth Dobson, Sreeram Ramagopalan, and Gavin Giovannoni. “The effect of gender in clinically isolated syndrome (CIS): a meta-analysis.” In: *Multiple sclerosis (Houndmills, Basingstoke, England)* 18.5 (2012), pp. 600–4.

[35] G. C. Ebers. “The natural history of multiple sclerosis: a geographically based study: 8: Familial multiple sclerosis.” In: *Brain* 123.3 (2000), pp. 641–649.

[36] G C Ebers, D E Bulman, A D Sadovnick, D W Paty, S Warren, W Hader, T J Murray, T P Seland, P Duquette, and T Grey. “A population-based study of multiple sclerosis in twins.” In: *The New England journal of medicine* 315.26 (1986), pp. 1638–42.

[37] Bradley Efron. “Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods.” In: *Biometrika* 68.3 (1981), pp. 589–599.

[38] Maja Eriksson, Oluf Andersen, and Björn Runmarker. “Long-term follow up of patients with clinically isolated syndromes, relapsing-remitting and secondary progressive multiple sclerosis.” In: *Multiple Sclerosis* 9.3 (2003), pp. 260–274.

[39] Arman Eshaghi et al. “Classification algorithms with multi-modal data fusion could accurately distinguish neuromyelitis optica from multiple sclerosis.” In: *NeuroImage. Clinical* 7 (2015), pp. 306–14.

[40] M. Filippi. “Magnetic resonance imaging findings predicting subsequent disease course in patients at presentation with clinically isolated syndromes suggestive of multiple sclerosis.” In: *Neurological Sciences* 22.8 (2001), S49–S51.

[41] M Filippi, M A Horsfield, M Rovaris, T Yousry, M A Rocca, C Baratti, S Bressi, and G Comi. “Intraobserver and interobserver variability in schemes for estimating volume of brain lesions on MR images in multiple sclerosis.” In: *AJNR. American journal of neuroradiology* 19.2 (1998), pp. 239–44.

[42] M. Filippi, M. A. Rocca, M. Calabrese, M. P. Sormani, F. Rinaldi, P. Perini, G. Comi, and P. Gallo. “Intracortical lesions: Relevance for new MRI diagnostic criteria for multiple sclerosis.” In: *Neurology* 75.22 (2010), pp. 1988–1994.

[43] L K Fisniku, P A Brex, D R Altmann, K A Miszkiel, C E Benton, R Lanyon, A J Thompson, and D H Miller. “Disability and T2 MRI lesions: a 20-year follow-up of patients with relapse onset of multiple sclerosis.” In: *Brain : a journal of neurology* 131.Pt 3 (2008), pp. 808–17.

[44] Hubert M. Fonteijn et al. “An event-based model for disease progression and its application in familial Alzheimer’s disease and Huntington’s disease.” In: *NeuroImage* 60.3 (2012), pp. 1880–1889.

[45] T. Fukazawa et al. “Both the HLA-DPB1 and -DRB1 alleles correlate with risk for multiple sclerosis in Japanese: clinical phenotypes and gender as important factors.” In: *Tissue Antigens* 55.3 (2000), pp. 199–205.

[46] Catharine R. Gale and Christopher N. Martyn. “Migrant studies in multiple sclerosis.” In: *Progress in Neurobiology* 47.4-5 (1995), pp. 425–448.

[47] Antonio Gallo, Marco Rovaris, Beatrice Benedetti, Maria Pia Sormani, Roberto Riva, Angelo Ghezzi, Vittorio Martinelli, Andrea Falini, Giancarlo Comi, and Massimo Filippi. “A brain magnetization transfer MRI study with a clinical follow up of about four years in patients with clinically isolated syndromes suggestive of multiple sclerosis.” In: *Journal of Neurology* 254.1 (2007), pp. 78–83.

[48] Daniel García-Lorenzo, Simon Francis, Sridar Narayanan, Douglas L. Arnold, and D. Louis Collins. “Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging.” In: *Medical Image Analysis* 17.1 (2013), pp. 1–18.

[49] Elizabeth Gibney. “Google AI algorithm masters ancient game of Go.” In: *Nature* 529.7587 (2016), pp. 445–446.

[50] Antonio Giorgio et al. “Location of brain lesions predicts conversion of clinically isolated syndromes to multiple sclerosis.” In: *Neurology* 80.3 (2013), pp. 234–41.

[51] Kerstin Hackmack, Friedemann Paul, Martin Weygandt, Carsten Allefeld, and John-Dylan Haynes. “Multi-scale classification of disease using structural MRI and wavelet transform.” In: *NeuroImage* 62.1 (2012), pp. 48–58.

[52] Miguel A Hernán, Susan S Jick, Giancarlo Logroscino, Michael J Olek, Alberto Ascherio, and Hershel Jick. “Cigarette smoking and the progression of multiple sclerosis.” In: *Brain : a journal of neurology* 128.Pt 6 (2005), pp. 1461–5.

[53] James J Higgins. *Introduction to Modern Nonparametric Statistics*. 1st ed. Cengage Learning, 2003.

[54] Nicky R Holdeman, Tammy Nguyen, and Rosa A Tang. “Demyelinating optic neuritis presenting as a clinically isolated syndrome.” In: *Optometry (St. Louis, Mo.)* 83.1 (2012), pp. 9–18.

[55] M. Hutchinson. “Spinal cord MRI should always be performed in clinically isolated syndrome patients: Commentary.” In: *Multiple Sclerosis Journal* 20.13 (2014), pp. 1690–1691.

[56] N Jafari, K L Kreft, H Z Flach, A C J W Janssens, and R Q Hintzen. “Callosal lesion predicts future attacks after clinically isolated syndrome.” In: *Neurology* 73.22 (2009), pp. 1837–41.

[57] W H James. “Concordance in twins and recurrence in sibships of multiple sclerosis.” In: *Lancet* 1.8273 (1982), p. 690.

[58] N Japkowicz. “The class imbalance problem: Significance and strategies.” In: *Proc. of the Int’l Conf. on Artificial Intelligence* (2000).

[59] M.Ann Kelly, David A. Cavan, Michelle A. Penny, Catherine H. Mijovic, David Jenkins, Sean Morrissey, David H. Miller, Anthony H. Barnett, and David A. Francis. “The influence of HLA-DR and -DQ alleles on progression to multiple sclerosis following a clinically isolated syndrome.” In: *Human Immunology* 37.3 (1993), pp. 185–191.

[60] Samantha M Kimball, Melanie R Ursell, Paul O’Connor, and Reinhold Vieth. “Safety of vitamin D3 in adults with multiple sclerosis.” In: *Am J Clin Nutr* 86.3 (2007), pp. 645–651.

[61] Esko Kinnunen. “Genetic Susceptibility to Multiple Sclerosis.” In: *Archives of Neurology* 45.10 (1988), p. 1108.

[62] S Klöppel, C Chu, G C Tan, B Draganski, H Johnson, J S Paulsen, W Kienzle, S J Tabrizi, J Ashburner, and R S J Frackowiak. “Automatic detection of preclinical neurodegeneration: presymptomatic Huntington disease.” In: *Neurology* 72.5 (2009), pp. 426–31.

[63] Stefan Klöppel et al. “Accuracy of dementia diagnosis: a direct comparison between radiologists and a computerized method.” In: *Brain : a journal of neurology* 131.Pt 11 (2008), pp. 2969–74.

[64] Stefan Klöppel, Cynthia M Stonnington, Carlton Chu, Bogdan Draganski, Rachael I Scahill, Jonathan D Rohrer, Nick C Fox, Clifford R Jack, John Ashburner, and Richard S J Frackowiak. “Automatic classification of MR scans in Alzheimer’s disease.” In: *Brain : a journal of neurology* 131.Pt 3 (2008), pp. 681–9.

[65] Stefan Klöppel et al. “White matter connections reflect changes in voluntary-guided saccades in pre-symptomatic Huntington’s disease.” In: *Brain : a journal of neurology* 131.Pt 1 (2008), pp. 196–204.

[66] Marcus Koch, Annemarie van Harten, Maarten Uyttenboogaart, and Jacques De Keyser. “Cigarette smoking and progression in multiple sclerosis.” In: *Neurology* 69.15 (2007), pp. 1515–20.

[67] Ron Kohavi. “A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection.” In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI’95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143.

[68] D Kotzamani, T Panou, V Mastorodemos, M Tzagournissakis, H Nikolakaki, C Spanaki, and A Plaitakis. “Rising incidence of multiple sclerosis in females associated with urbanization.” In: *Neurology* 78.22 (2012), pp. 1728–35.

[69] O Krökki, R Bloigu, M Reunanen, and A M Remes. “Increasing incidence of multiple sclerosis in women in Northern Finland.” In: *Multiple sclerosis (Houndmills, Basingstoke, England)* 17.2 (2011), pp. 133–8.

[70] J Kuhle et al. “Conversion from clinically isolated syndrome to multiple sclerosis: A large multicentre study.” In: *Multiple sclerosis (Houndmills, Basingstoke, England)* 21.8 (2015), pp. 1013–24.

[71] J. F. Kurtzke. “Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS).” In: *Neurology* 33.11 (1983), pp. 1444–1444.

[72] John F. Kurtzke, Gilbert W. Beebe, Benedict Nagler, Leonard T. Kurland, and Thomas L. Auth. “Studies on the natural history of multiple sclerosis—8.” In: *Journal of Chronic Diseases* 30.12 (1977), pp. 819–830.

[73] F. D. Lublin and S. C. Reingold. “Defining the clinical course of multiple sclerosis: Results of an international survey.” In: *Neurology* 46.4 (1996), pp. 907–911.

[74] A H Maghzi, H Ghazavi, M Ahsan, M Etemadifar, Sa Mousavi, F Khorvash, and A Minagar. “Increasing female preponderance of multiple sclerosis in Isfahan, Iran: a population-based study.” In: *Multiple sclerosis (Houndmills, Basingstoke, England)* 16.3 (2010), pp. 359–61.

[75] Ruth Ann Marrie, Lawrence Elliott, James Marriott, Michael Cossoy, James Blanchard, Stella Leung, and Nancy Yu. “Effect of comorbidity on mortality in multiple sclerosis.” In: *Neurology* 85.3 (2015), pp. 240–247.

[76] W. Ian McDonald et al. “Recommended diagnostic criteria for multiple sclerosis: Guidelines from the international panel on the diagnosis of multiple sclerosis.” In: *Annals of Neurology* 50.1 (2001), pp. 121–127.

[77] D H Miller, S R Hammond, J G McLeod, G Purdie, and D C Skegg. “Multiple sclerosis in Australia and New Zealand: are the determinants genetic or environmental?” In: *Journal of Neurology, Neurosurgery & Psychiatry* 53.10 (1990), pp. 903–905.

[78] David H Miller, Declan T Chard, and Olga Ciccarelli. “Clinically isolated syndromes.” In: *The Lancet Neurology* 11.2 (2012), pp. 157–169.

[79] David H Miller and Siobhan M Leary. “Primary-progressive multiple sclerosis.” In: *The Lancet. Neurology* 6.10 (2007), pp. 903–12.

[80] David Miller, Frederik Barkhof, Xavier Montalban, Alan Thompson, and Massimo Filippi. “Clinically isolated syndromes suggestive of multiple sclerosis, part I: natural history, pathogenesis, diagnosis, and prognosis.” In: *The Lancet. Neurology* 4.5 (2005), pp. 281–8.

[81] Ron Milo and Esther Kahana. “Multiple sclerosis: geoepidemiology, genetics and the environment.” In: *Autoimmunity reviews* 9.5 (2010), A387–94.

[82] J. Ross Mitchell, Stephen J. Karlik, Donald H. Lee, Michael Eliasziw, George P. Rice, and Aaron Fenster. “The variability of manual and computer assisted quantification of multiple sclerosis lesion volumes.” In: *Medical Physics* 23.1 (1996), p. 85.

[83] V Mnih, K Kavukcuoglu, D Silver, AA Rusu, and J Veness. “Human-level control through deep reinforcement learning.” In: *Nature* (2015).

[84] Marc Modat, Gerard R. Ridgway, Zeike A. Taylor, Manja Lehmann, Josephine Barnes, David J. Hawkes, Nick C. Fox, and Sébastien Ourselin. “Fast free-form deformation using graphics processing units.” In: *Computer Methods and Programs in Biomedicine* 98.3 (2010), pp. 278–284.

[85] Cristina Montomoli et al. “Multiple sclerosis recurrence risk for siblings in an isolated population of Central Sardinia, Italy.” In: *Genetic epidemiology* 22.3 (2002), pp. 265–71.

[86] Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. “The Alzheimer’s disease neuroimaging initiative.” In: *Neuroimaging clinics of North America* 15.4 (2005), pp. 869–77, xi–xii.

[87] C. J. Mumford, N. W. Wood, H. Kellar-Wood, J. W. Thorpe, D. H. Miller, and D.A.S. Compston. “The British Isles survey of multiple sclerosis in twins.” In: *Neurology* 44.1 (1994), pp. 11–11.

[88] M. Munch, K. Riisom, T. Christensen, A. Møller-Larsen, and S. Haahr. “The significance of Epstein-Barr virus seropositivity in multiple sclerosis patients?” In: *Acta Neurologica Scandinavica* 97.3 (2009), pp. 171–174.

[89] Benson Mwangi et al. “Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder.” In: *Brain : a journal of neurology* 135.Pt 5 (2012), pp. 1508–21.

[90] A Nicoletti, F Patti, S Lo Fermo, V Sorbello, E Reggio, D Maimone, M Zappia, and A Reggio. “Possible increasing risk of multiple sclerosis in Catania, Sicily.” In: *Neurology* 65.8 (2005), pp. 1259–63.

[91] J H Noseworthy, M K Vandervoort, C J Wong, and G C Ebers. “Interrater variability with the Expanded Disability Status Scale (EDSS) and Functional Systems (FS) in a multiple sclerosis clinical trial. The Canadian Cooperation MS Study Group.” In: *Neurology* 40.6 (1990), pp. 971–5.

[92] J R Oksenberg and L F Barcellos. “Multiple sclerosis genetics: leaving no stone unturned.” In: *Genes and Immunity* 6.5 (2005), pp. 375–387.

[93] Thais Mayumi Oshiro, Pedro Santoro Perez, and José Augusto Baranauskas. “How Many Trees in a Random Forest?” In: *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2012, pp. 154–168.

[94] Muhammed Emin Ozcan, Bahri Ince, Ayhan Bingöl, Simge Ertürk, Meriç Adil Altınöz, Hasan Hüseyin Karadeli, Abdulkadir Koçer, and Talip Asil. “Association between smoking and cognitive impairment in multiple sclerosis.” In: *Neuropsychiatric disease and treatment* 10 (2014), pp. 1715–9.

[95] Elisabetta Pagani, Maria A Rocca, Antonio Gallo, Marco Rovaris, Vittorio Martinelli, Giancarlo Comi, and Massimo Filippi. “Regional brain atrophy evolves differently in patients with multiple sclerosis according to clinical phenotype.” In: *AJNR. American journal of neuroradiology* 26.2 (2005), pp. 341–6.

[96] Fabian Pedregosa et al. “Scikit-learn: Machine Learning in Python.” In: *The Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[97] Sean J Pittock, Vanda A Lennon, Jerome de Seze, Patrick Vermersch, Henry A Homburger, Dean M Wingerchuk, Claudia F Lucchinetti, Hélène Zéphir, Kevin Moder, and Brian G Weinshenker. “Neuromyelitis optica and non organ-specific autoimmunity.” In: *Archives of neurology* 65.1 (2008), pp. 78–83.

[98] Chris H Polman et al. “Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria".” In: *Annals of neurology* 58.6 (2005), pp. 840–6.

[99] Chris H Polman et al. “Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria.” In: *Annals of neurology* 69.2 (2011), pp. 292–302.

[100] S V Ramagopalan, A E Handel, G Giovannoni, S Rutherford Siegel, G C Ebers, and G Chaplin. “Relationship of UV exposure to prevalence of multiple sclerosis in England.” In: *Neurology* 76.16 (2011), pp. 1410–4.

[101] Maria A Rocca et al. “A three-year, multi-parametric MRI study in patients at presentation with CIS.” In: *Journal of neurology* 255.5 (2008), pp. 683–91.

[102] Marco Rovaris, Christian Confavreux, Roberto Furlan, Ludwig Kappos, Giancarlo Comi, and Massimo Filippi. “Secondary progressive multiple sclerosis: current knowledge and future challenges.” In: *The Lancet. Neurology* 5.4 (2006), pp. 343–54.

[103] RA Rudick, G Cutter, and S Reingold. “The Multiple Sclerosis Functional Composite: a new clinical outcome measure for multiple sclerosis trials.” In: *Multiple Sclerosis* 8.5 (2002), pp. 359–365.

[104] A. Ruet, M. S. Deloire, J.-C. Ouallet, S. Molinier, and B. Brochet. “Predictive factors for multiple sclerosis in patients with clinically isolated spinal cord syndrome.” In: *Multiple Sclerosis Journal* 17.3 (2011), pp. 312–318.

[105] A D Sadovnick, P A Baird, and R H Ward. “Multiple sclerosis: updated risks for relatives.” In: *American journal of medical genetics* 29.3 (1988), pp. 533–41.

[106] A D Sadovnick, H Armstrong, G P Rice, D Bulman, L Hashimoto, D W Paty, S A Hashimoto, S Warren, W Hader, and T J Murray. “A population-based study of multiple sclerosis in twins: update.” In: *Annals of neurology* 33.3 (1993), pp. 281–5.

[107] Stephen Sawcer, Robin J M Franklin, and Maria Ban. “Multiple sclerosis genetics.” In: *The Lancet. Neurology* 13.7 (2014), pp. 700–9.

[108] Stephen Sawcer et al. “Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis.” In: *Nature* 476.7359 (2011), pp. 214–9.

[109] T. Schneider, W. Brownlee, H. Zhang, O. Ciccarelli, D. H. Miller, and C Wheeler-Kingshott. “Application of multi-shell NODDI in Multiple Sclerosis.” In: *Proc. Intl. Soc. Mag. Reson. Med.* Vol. 22. 2014, p. 19.

[110] Basil Sharrack and Richard A.C. Hughes. “Clinical scales for multiple sclerosis.” In: *Journal of the Neurological Sciences* 135.1 (1996), pp. 1–9.

[111] J G Sled, A P Zijdenbos, and A C Evans. “A nonparametric method for automatic correction of intensity nonuniformity in MRI data.” In: *IEEE transactions on medical imaging* 17.1 (1998), pp. 87–97.

[112] M. H. Sombekke, M. P. Wattjes, L. J. Balk, J. M. Nielsen, H. Vrenken, B. M. J. Uitdehaag, C. H. Polman, and F. Barkhof. “Spinal cord lesions in patients with clinically isolated syndrome: A powerful tool in diagnosis and prognosis.” In: *Neurology* 80.1 (2013), pp. 69–75.

[113] M S Stein et al. “A randomized trial of high-dose vitamin D2 in relapsing-remitting multiple sclerosis.” In: *Neurology* 77.17 (2011), pp. 1611–8.

[114] Jean Talairach and Pierre Tournoux. *Co-planar stereotaxic atlas of the human brain*. 1st. Thieme, 1988.

[115] M. Tintore, A. Rovira, J. Rio, C. Nos, E. Grive, J. Sastre-Garriga, I. Pericot, E. Sanchez, M. Comabella, and X. Montalban. “New diagnostic criteria for multiple sclerosis: Application in first demyelinating episode.” In: *Neurology* 60.1 (2003), pp. 27–30.

[116] M Tintoré, A Rovira, J Río, C Nos, E Grivé, N Téllez, R Pelayo, M Comabella, J Sastre-Garriga, and X Montalban. “Baseline MRI predicts future attacks and disability in clinically isolated syndromes.” In: *Neurology* 67.6 (2006), pp. 968–72.

[117] M. Tintore et al. “Do oligoclonal bands add information to MRI in first attacks of multiple sclerosis?” In: *Neurology* 70.Issue 13, Part 2 (2008), pp. 1079–1083.

[118] M Tintore et al. “Brainstem lesions in clinically isolated syndromes.” In: *Neurology* 75.21 (2010), pp. 1933–8.

[119] A. Traboulsee, J. Dehmeshki, P. A. Brex, C. M. Dalton, D. Chard, G. J. Barker, G. T. Plant, and D. H. Miller. “Normal-appearing brain tissue MTR histograms in clinically isolated syndromes suggestive of MS.” In: *Neurology* 59.1 (2002), pp. 126–128.

[120] Benjamin K-T Tsang and Richard Macdonell. “Multiple sclerosis- diagnosis, management and prognosis.” In: *Australian family physician* 40.12 (2011), pp. 948–55.

[121] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. “N4ITK: improved N3 bias correction.” In: *IEEE transactions on medical imaging* 29.6 (2010), pp. 1310–20.

[122] Nicholas J Tustison et al. “Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements.” In: *NeuroImage* 99 (2014), pp. 166–79.

[123] Sudhir Varma et al. “Bias in error estimation when using cross-validation for model selection.” In: *BMC Bioinformatics* 7.1 (2006), p. 91.

[124] Marco Vercellino et al. “Grey matter pathology in multiple sclerosis.” In: *Journal of neuropathology and experimental neurology* 64.12 (2005), pp. 1101–7.

[125] Marco Vercellino et al. “Demyelination, inflammation, and neurodegeneration in multiple sclerosis deep gray matter.” In: *Journal of neuropathology and experimental neurology* 68.5 (2009), pp. 489–502.

[126] Sandra Vukusic, Vincent Van Bockstael, Sophie Gosselin, and Christian Confavreux. “Regional variations in the prevalence of multiple sclerosis in French farmers.” In: *Journal of neurology, neurosurgery, and psychiatry* 78.7 (2007), pp. 707–9.

[127] K.-P. Wandinger, W. Jabs, A. Siekhaus, S. Bubel, P. Trillenberg, H.-J. Wagner, K. Wessel, H. Kirchner, and H. Hennig. “Association between clinical disease activity and Epstein-Barr virus reactivation in MS.” In: *Neurology* 55.2 (2000), pp. 178–184.

[128] B. G. Weinshenker, B. Bass, G. P. A. Rice, J. Noseworthy, W. Carriere, J. Baskerville, and G. C. Ebers. “The Natural History of Multiple Sclerosis: A Geographically based Study.” In: *Brain* 112.1 (1989), pp. 133–146.

[129] Martin Weygandt, Kerstin Hackmack, Caspar Pfüller, Judith Bellmann-Strobl, Friedemann Paul, Frauke Zipp, and John-Dylan Haynes. “MRI pattern recognition in multiple sclerosis normal-appearing brain areas.” In: *PloS one* 6.6 (2011). Ed. by Christoph Kleinschnitz, e21138.

[130] Wikimedia Commons. *Progression types of Multiple sclerosis*.

[131] D M Wingerchuk, J Lesaux, G P A Rice, M Kremenchutzky, and G C Ebers. “A pilot study of oral calcitriol (1,25-dihydroxyvitamin D3) for relapsing-remitting multiple sclerosis.” In: *Journal of neurology, neurosurgery, and psychiatry* 76.9 (2005), pp. 1294–6.

[132] World Health Organization. *Atlas: Multiple Sclerosis in the World*. Tech. rep. Geneva: World Health Organization, 2008, pp. 15–16.

[133] V Wottschel, DC Alexander, DT Chard, DH Miller, and O Ciccarelli. “Prediction of Second Clinical Attack in CIS patients using SVMs.” In: *Magnetic Resonance Materials in Physics, Biology and Medicine* 26.1 (2013), pp. 262–263.

[134] V. Wottschel, D.C. Alexander, P.P. Kwok, D.T. Chard, M.L. Stromillo, N. De Stefano, A.J. Thompson, D.H. Miller, and O. Ciccarelli. “Predicting outcome in clinically isolated syndrome using machine learning.” In: *NeuroImage: Clinical* 7 (2015), pp. 281–287.

[135] Alexandra L Young, Neil P Oxtoby, Pankaj Daga, David M Cash, Nick C Fox, Sebastien Ourselin, Jonathan M Schott, and Daniel C Alexander. “A data-driven model of biomarker changes in sporadic Alzheimer’s disease.” In: *Brain : a journal of neurology* 137.Pt 9 (2014), pp. 2564–77.

[136] Jonathan Young, Marc Modat, Manuel J. Cardoso, Alex Mendelson, Dave Cash, and Sebastien Ourselin. “Accurate multimodal probabilistic prediction of conversion to Alzheimer’s disease in patients with mild cognitive impairment.” In: *NeuroImage: Clinical* 2 (2013), pp. 735–745.